JYVÄSKYLÄ STUDIES IN COMPUTING
217

Fedor Chernogorov

Advanced Performance Monitoring for Self-Healing Cellular Mobile Networks

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 217

Fedor Chernogorov

Advanced Performance Monitoring for Self-Healing Cellular Mobile Networks

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Delta-salissa
elokuun 17. päivänä 2015 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, hall Delta, on August 17, 2015 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Advanced Performance Monitoring for Self-Healing Cellular Mobile Networks

JYVÄSKYLÄ STUDIES IN COMPUTING 217

Fedor Chernogorov

Advanced Performance Monitoring for Self-Healing Cellular Mobile Networks

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Editors
Timo Männikkö
Department of Mathematical Information Technology, University of Jyväskylä
Pekka Olsbo, Timo Hautala
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-6235-7
ISBN 978-951-39-6235-7(PDF)

ISBN 978-951-39-6234-0 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2015

To my parents, grandparents and my love Katyusha

Fedor

Посвящается моим дорогим родителям, бабуле, деду и любимой Катюше

Федор

Jyväskylä, Finland
1 June 2015

Ювяскюля, Финляндия
1 Июня 2015 года

ABSTRACT

Chernogorov, Fedor
Advanced Performance Monitoring for Self-Healing Cellular Mobile Networks
Jyväskylä: University of Jyväskylä, 2015, 120 p.(+included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; 217)
ISBN 978-951-39-6234-0 (nid.)
ISBN 978-951-39-6235-7 (PDF)
Finnish summary
Diss.

This dissertation is devoted to development and validation of advanced performance monitoring system for existing and future cellular mobile networks. Knowledge mining techniques are employed for analysis of user specific logs, collected with Minimization of Drive Tests (MDT) functionality. Ever increasing quality requirements, expansion of the mobile networks and their extending heterogeneity, call for effective automatic means of performance monitoring. Nowadays, network operation is mostly controlled manually through aggregated key performance indicators and statistical profiles. These methods are are not able to fully address the dynamism and complexity of modern mobile networks. Self-organizing networks introduce automation to the most important network functions, but the opportunity of processing large arrays of user reported performance data is underutilized.

Advanced performance monitoring system developed in the presented research considers both numerical and sequential properties of the MDT data for detection of faults. Network malfunctions analyzed in this study are sleeping cells in either physical or medium access layer. A full data mining cycle is employed for identification of problematic regions in the network. Pre-processing with statistical normalization and sliding window methods, both linear and nonlinear transformation and dimensionality reduction algorithms, together with clustering and classification methods are used in the discussed research. Several post-processing and detection quality evaluation methods are proposed and applied. The developed system is capable of fast and accurate detection of nontrivial network dysfunctions and is suitable for future mobile networks, even in combination with cognitive self-healing. As a result, operation of modern mobile networks would become more robust, increasing quality of service and user experience.

Keywords: quality and performance management, knowledge mining, performance monitoring, self-organizing networks, data mining, anomaly detection, sleeping cell, sequence-based analysis, cellular mobile networks.

Author

Fedor Chernogorov
Department of Mathematical Information Technology
University of Jyväskylä, Finland

Supervisors

Dr. Professor Tapani Ristaniemi
Department of Mathematical Information Technology
University of Jyväskylä, Finland

Dr. Dmitry Petrov
Magister Solutions Ltd.
Jyväskylä, Finland

Reviewers

Dr. Olli Simula
Professor Emeritus
Department of Computer Science
Aalto University, Finland

Dr. Jarno Niemelä
Elisa Ltd.
Helsinki, Finland

Opponent

Dr. Professor Yevgeni Koucheryavy
Department of Communication Engineering
Tampere University of Technology, Finland

GLOSSARY

2G	2^{nd} Generation
3G	3^{rd} Generation
3GPP	3^{rd} Generation Partnership Programme
4G	4^{th} Generation
5G	5^{th} Generation
AA	Anomaly Analysis
AGNES	AGglomerative NESting
A-GNSS	Assisted-Global Navigation Satellite System
ANR	Automatic Neighbor Relations
AUC	Area under Curve
BCR	Blocked Call Rate
BER	Bit Error Rate
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
BLER	Block Error Rate
BS	Base Station
CBLOF	Cluster-Based Local Outlier Factor
CBR	Case-Based Reasoning
CCSR	Call Completion Success Rate
CLIQUE	CLustering In QUEst
CM	Configuration Management
COC	Cell Outage Compensation
COD	Cell Outage Detection
COMMUNE	COgnitive network ManageMent under UNcErtainty
CPICH	Common Pilot Channel
CQI	Channel Quality Indicator
C-RNTI	Cell Radio Network Temporary Identifier

CSI	Channel State Indicator
CSSR	Call Setup Success Rate
DBSCAN	Density-based spatial clustering of applications with noise
DCR	Drop Call Ratio
DIANA	DIvisive ANAlysis
DL	Downlink
DM	Diffusion Maps
DRX	Discontinuous Reception
eNB	E-UTRAN NodeB
EPS	Evolved Packet System
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency Division Duplexing
FER	Frame Error Rate
FM	Fault Management
FRF	Frequency Reuse Factor
GSM	Global System for Mobile Communications
HARQ	Hybrid Adaptive Repeat and reQuest
HDP	Hierarchical Dirichlet Process
HLR	Home Location Register
HO	Handover
HOF	Handover Failure
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HW	Hardware
ID	Identification
ISCP	Interference Signal Code Power
ITU	International Telecommunication Union
KM	Knowledge Mining

K-NN	K-Nearest Neighbors
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LOF	Local Oulier Factor
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
MAC	Medium Access Control
MCA	Minor Component Analysis
MDT	Minimization of Drive Tests
MLN	Markov Logic Networks
MME	Mobility Management Entity
MOS	Mean Opinion Score
MQA	Mobile Quality Agent
MRO	Mobility Robustness Optimization
NE	Network Element
NGMN	Next Generation Mobile Networks
NM	Network Management
ns-3	Network Simulator 3
OAM	Operations, Administration, and Maintenance
OFD	Operational Fault Detection
OFDM	Orthogonal Frequency-Division Multiplexing
OPTICS	Ordering points to identify the clustering structure
OSS	Operations Support System
PCA	Principal Component Analysis
PCI	Physical Cell Identity
PESQ	Perceptual Evaluation of Speech Quality
PGW	Packet Gateway
PHR	Power Headroom

PI	Performance Indicator
PM	Performance Monitoring
PRACH	Physical Random Access Channel
QoE	Quality of Experience
QoS	Quality of Service
QPM	Quality and Performance Management
RA	Recovery Analysis
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RCEF	RRC Connection Establishment Failure
RF	Radio Frequency
RLF	Radio Link Failure
RNC	Radio Network Controller
ROC	Receiver Operating Characteristic
RRC	Radio Resource Control
RRM	Radio Resource Management
RSCP	Received Signal Code Power
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
SDCCH	Stand-alone Dedicated Control Channel
SEMAFOUR	Self-Management for Unified Heterogeneous Radio Access Networks
SGSN	Serving GPRS Support Node
SGW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
SIR	Signal to Interference Ratio

SOCRATES	Self-Optimisation and self-ConfiguRATion in wirelEss networkS
SOM	Self-Organizing Maps
SON	Self-Organizing Network
SORTE	Second ORder sTatistic of the Eigenvalues
SQM	Service Quality Management
STING	STatistical INformation Grid
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SW	Software
TCE	TRACE Collection Entity
TTT	Time to Trigger
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
UTRA	Universal Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
WCDMA	Wideband Code Division Multiple Access
WiFi	Wireless Fidelity

ACKNOWLEDGEMENTS

My deepest appreciation goes to Dr. Prof. Tapani Ristaniemi for his guidance, scientific advice and comprehensive help during the work on this dissertation. Professor Ristaniemi created a wholesome environment for me to stay in Finland and to successfully go through the doctoral studies. He was always motivating, supportive and truly interested in the ongoing research. I would like to sincerely thank my second supervisor Dr. Dmitry Petrov for his friendly help, valuable feedback, meaningful discussions and comments. Gratefulness should be expressed to co-authors of the conference and journal articles Dr. Jussi Turkka, Kimmo Brigatti, Sergey Chernov, Dr. Jani Puttonen, Dr. Timo Nihtilä, for being innovative, open-minded and easy-going during the research process. It is necessary to thank the Department of Mathematical Information Technology and postgraduate school in computing and mathematical sciences of the University of Jyväskylä for the financial support of this study.

I would like to express my gratitude to the CEO of Magister Solutions Ltd., Dr. Janne Kurjenniemi for the creation of perfect conditions and opportunities, which made the dissertation work easier, more efficient and fruitful. It is important to note, that a professional, friendly and warm atmosphere at Magister Solutions created the perfect background for the research work on this thesis. I am very grateful to each and every colleague of mine who directly or indirectly supported me during my scientific work and life in Finland.

It is necessary to thank colleagues from the EU Celtic+ COMMUNE project Dr. Seppo Hämäläinen, Dr. Vilho Räisanen and others for their cooperation and meaningful discussions. Additionally, I would like to show my gratefulness to my teachers from the Faculty of Physics at Demidov Yaroslavl State University. For me, they cultivated an interest in science and some of them are forever in my memory.

I am deeply grateful to my friends in many countries around the world. I truly value all of them for playing an important role in my life, and appreciate memorable time we spent together.

I would like to express my greatest appreciation to my parents Andrey and Marina, for their energy, love, patience, support and understanding, which helped me make the greatest achievements in my life, including the finalization of my doctoral studies. I am deeply grateful to my grandparents, and particularly to Viktor Fedorovich and Valentina Dmitrievna for their participation in my life, valuable advice and support. It is especially important for me to thank my love and my wife Katyusha, who fills every life instant with meaning. Her love and care inspired and motivated me to concentrate and put the last dot in the Ph.D. studies.

Jyväskylä, June 1, 2015
Fedor Chernogorov

LIST OF FIGURES

FIGURE 1 Components of mobile network quality and performance management system. 19
FIGURE 2 Evaluation process of new Quality and Performance Management (QPM) methods. 22
FIGURE 3 Flow of Quality and Performance Management. 28
FIGURE 4 Evolution of mobile network quality and performance management systems. 29
FIGURE 5 Management and signaling based TRACE. 34
FIGURE 6 Radio link failure procedure in Long Term Evolution (LTE). 37
FIGURE 7 Key performance indicator aggregation dimensions. 38
FIGURE 8 Key Performance Indicator (KPI) thresholds. 41
FIGURE 9 Selection of thresholds on the basis of statistical profile of Stand-alone Dedicated Control Channel (SDCCH) success rate. 42
FIGURE 10 Structure of the autonomic element. 43
FIGURE 11 Autonomic manager for mobile networks. 44
FIGURE 12 Network dominance map. 49
FIGURE 13 Knowledge mining process. 52
FIGURE 14 Sliding window transformation. 57
FIGURE 15 Sample data before and after transformation with Principal Component Analysis (PCA). 59
FIGURE 16 Confusion matrix and corresponding example. 66
FIGURE 17 Example of Receiver Operating Characteristic (ROC) curves. 68
FIGURE 18 Structure of research activities for advanced quality and performance management with MDT data. 80
FIGURE 19 Sleeping cell detection in diffusion maps embedded space and k-means clustering from [PIV]. 81
FIGURE 20 Classification results of periodic and event triggered MDT reports from [PV]. 82
FIGURE 21 Results of FindCluster-Based Local Outlier Factor (CBLOF) clustering in the embedded space after 2-gram and PCA transformation. Axes are components of PCA. 83
FIGURE 22 Sleeping cell detection based on anomalous 2-gram: "Handover COMMAND–A2 RSRP Enter". 84
FIGURE 23 Quality of sleeping cell detection in [PI]. 85
FIGURE 24 Detection results and performance from [PII]. 86
FIGURE 25 Quality and performance management architecture for future mobile networks. 89
FIGURE 26 Models in LENA module of Network Simulator 3 (ns-3). 117

LIST OF TABLES

TABLE 1 Measurement triggering events in LTE. 36
TABLE 2 Examples of network KPIs. 40
TABLE 3 Types of performance monitoring variables from perspective of data mining. 55
TABLE 4 Example of N-gram analysis per character, $N = 2$. 57
TABLE 5 Comparative analysis of advanced QPM studies. Part I. 72
TABLE 6 Comparative analysis of advanced QPM studies. Part II. 76
TABLE 7 Comparison of advanced QPM studies, devoted to detection of sleeping cell failures with analysis of sequential characteristics of MDT reports. 114
TABLE 8 Comparison of advanced QPM studies, devoted to detection of sleeping cell failures with analysis of numerical characteristics of MDT reports. 115

CONTENTS

ABSTRACT
GLOSSARY
ACKNOWLEDGEMENTS
LIST OF FIGURES
LIST OF TABLES
CONTENTS
LIST OF INCLUDED ARTICLES

1 INTRODUCTION 17
1.1 Background 17
1.2 Vision and research challenges of Quality and Performance Management 20
1.3 Problem statement 21
1.4 Outline of the dissertation 23
1.5 Main contribution 24

2 QUALITY AND PERFORMANCE MANAGEMENT IN CELLULAR MOBILE NETWORKS 27
2.1 Evolution of Quality and Performance Management systems 28
2.2 Network Failures and Malfunctions 30
2.3 Data collection in traditional performance monitoring systems 32
2.3.1 Data sources 32
2.3.2 Alarms 35
2.3.3 Counters 35
2.3.4 Measurements 36
2.4 Traditional data analysis methods for performance monitoring 37
2.4.1 Key Performance Indicators 37
2.4.2 Thresholds and Profiles 41
2.5 Self-organizing networks for Quality and Performance Management 42
2.5.1 Self-optimization 45
2.5.2 Self-healing 46
2.5.3 Minimization of Drive Testing 47
2.6 Summary 49

3 KNOWLEDGE MINING 51
3.1 Data mining and anomaly detection 53
3.1.1 Data Types in Anomaly Detection 53
3.1.2 Data Types in Mobile Networks 54
3.2 Data pre-processing 54
3.3 Transformation techniques 56
3.3.1 Dimensionality reduction 58
3.3.1.1 Curse of dimensionality 58

3.3.1.2 Principal and Minor Component Analyses 58
3.3.1.3 Diffusion Maps 59
3.4 Pattern recognition approaches 62
3.4.1 Types of learning in pattern recognition 62
3.4.2 Examples of pattern recognition algorithms 64
3.4.2.1 Nearest neighbor algorithms 64
3.4.2.2 K-means clustering 65
3.4.2.3 FindCBLOF algorithm 65
3.4.3 Post-processing methods 65
3.5 Metrics for evaluation of pattern recognition 66
3.6 Summary 68

4 ADVANCED FAULT DETECTION, DIAGNOSIS AND HEALING IN MOBILE NETWORKS 70
4.1 Knowledge mining for quality and performance management: State of the art 70
4.2 Advanced performance monitoring with MDT data 79
4.2.1 Sleeping cell detection based on numerical MDT data 81
4.2.2 Sequence-based analysis of MDT data for sleeping cell detection 83
4.3 Discussion 86
4.3.1 Pros and Cons of Anomaly Detection in Performance Monitoring 87
4.3.2 Architecture of future cognitive QPM systems 88
4.4 Summary 90

5 CONCLUSION 91

YHTEENVETO (FINNISH SUMMARY) 93

REFERENCES 94

APPENDIX 1 COMPARISON OF THE STUDIES IN QUALITY AND PERFORMANCE MANAGEMENT FOR MOBILE NETWORKS . 113

APPENDIX 2 SYSTEM LEVEL SIMULATIONS FOR QUALITY AND PERFORMANCE MANAGEMENT RESEARCH 116

INCLUDED ARTICLES

LIST OF INCLUDED ARTICLES

PI Fedor Chernogorov, Sergey Chernov, Kimmo Brigatti, Tapani Ristaniemi. Sequence-based Detection of Sleeping Cell Failures in Mobile Networks. *Wireless Networks, The Journal of Mobile Communication, Computation and Information (submitted for review, available on arxiv.org), 2015*.

PII Sergey Chernov, Fedor Chernogorov, Dmitry Petrov, Tapani Ristaniemi. Data Mining Framework for Random Access Failure Detection in LTE Networks. *Proc.* 25^{th} *IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2014*.

PIII Fedor Chernogorov, Tapani Ristaniemi, Kimmo Brigatti, Sergey Chernov. N-gram analysis for sleeping cell detection in LTE networks. *Proc.* 39^{th} *IEEE International Conference on Acoustics, Speech and Signal Processing, 2013*.

PIV Fedor Chernogorov, Jussi Turkka, Tapani Ristaniemi, Amir Averbuch. Detection of Sleeping Cells in LTE Networks Using Diffusion Maps. *Proc. 73rd IEEE Vehicular Technology Conference (VTC Spring), 2011*.

PV Jussi Turkka, Fedor Chernogorov, Kimmo Brigatti, Tapani Ristaniemi, and Jukka Lempiäinen. An Approach for Network Outage Detection from Drive-Testing Databases. *Journal of Computer Networks and Communications, Volume 2012 (2012), Article ID 163184*.

PVI Fedor Chernogorov, Ilmari Repo, Vilho Räisänen, Timo Nihtilä, Janne Kurjenniemi. Cognitive Self-Healing for Future Mobile Networks. *Proc.* 11^{th} *IEEE International Wireless Communications & Mobile Computing Conference (IWCMC), 2015*.

1 INTRODUCTION

1.1 Background

Management and performance monitoring in modern cellular mobile networks are complicated, non-trivial tasks. The reason is that mobile networks are highly dynamic and consist of huge and constantly increasing number of elements. The propagation environment includes slow and fast fading effects, what leads to variability in radio signal strength. Users move with variable speeds, which might require different settings for network mobility procedures, such as handover or cell re-selection. Moreover, cells with different coverage areas, e.g. in rural, urban and sub-urban environments would have various amounts of handovers. In addition, there are simply more and less populated regions and areas, and hence amounts of call establishments and mobility related events vary a lot. During different times of the day, days of the week, and even seasons, network in general and individual cells in particular, demonstrate various load patterns. This denotes temporal dependence of network behavior. Another important aspect, is that networks are becoming heterogeneous. According to currently developed requirements [1], 5^{th} Generation (5G) networks should combine previous generations of mobile networks, including such Radio Access Technologies (RATs) as Wireless Fidelity (WiFi), Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS): Wideband Code Division Multiple Access (WCDMA), High Speed Packet Access (HSPA), LTE of UMTS and Long Term Evolution Advanced (LTE-A). Another factor which contributes to network heterogeneity is that 5G networks will consist of different cell types, e.g. macro, micro, pico, femto and relays. Due to such variability in network conditions it may be hard to judge whether the observed network behavior is normal or abnormal using traditional Quality and Performance Management (QPM) systems, e.g based on analysis of absolute values of performance statistics.

Diversity of network failure types also makes maintenance of high service quality more difficult. Problems can be caused by malfunctions in various hardware components, e.g. in antenna amplification or cabling. Another class of fail-

ure are software problems, e.g. after an upgrade of the base station firmware. In many cases, failures are noticed by the operator, due to explicit notification, e.g. an alarm, or because of severe changes in the monitored KPIs. Even then, diversity of a networks setup and cell configuration requires substantial effort to maintain timely and accurate detection and network failures. However, there are such failures, called sleeping cells, which do not trigger any alarm, and cannot always be seen from KPIs, but users are suffering from low Quality of Service (QoS) or even absence of service.

All of the factors outlined above jointly contribute to high complexity of future networks and make manual management tedious and expensive task. In order to maintain an adequate level of provided service quality, it is necessary to use a highly efficient QPM system [2, 3, 4]. Such system includes control of the operational network state, configuration adjustment and handling of emerging failures. Network QPM systems consist of Performance Monitoring (PM) and recovery parts, as it is shown in Figure 1. PM is aimed at controlling the network state through identification of emerging failures. First, performance data is collected and analyzed. Then detection of malfunctions helps to identify network regions, elements or devices, which demonstrate degraded quality of service. The nature of failures and the extent of their impact on network performance, as well as the type of the available data, should be taken into account in development of detection methods. Diagnosis of the detected failures allows for the understanding of the root causes of the occurred malfunction, and provides input to the recovery part. The latter component of QPM is responsible for restoration of the communication services and reduction of the negative effect caused by the failures. Appropriate measures used to maintain a sufficient level of network operational quality are called recovery or healing. These actions can include, e.g. reboot or reconfiguration of related base station(-s) for compensation of network failures. Each step of QPM might require a different amount of human involvement from purely manual to highly automated and even cognitive parts based on machine learning and data mining.

The main disadvantage of the traditional QPM systems – is the large amount of manual work required for data collection and analysis. Healing is mostly done manually, and is inherently related to physical maintenance of network elements in cases of severe breakdowns. Thus, new challenges related to the improvement of network QPM, are faced by the research community and the industry of cellular mobile communications. To address this, an approach towards automation of network maintenance called Self-Organizing Network (SON) [2] has been proposed. It is widely accepted both in research and standardization areas as an enabler for mobile networks automation. Requirements for SON have been first published by Next Generation Mobile Networks (NGMN) [5, 6]. Some of them are further developed by 3^{rd} Generation Partnership Programme (3GPP) for HSPA and LTE/LTE-A networks [7]. On the scientific side many industry driven projects have been organized. For instance, Self-Optimisation and self-ConfiguRATion in wirelEss networkS (SOCRATES) is concentrated on development and improvement of SON features, and Self-Management for Unified Het-

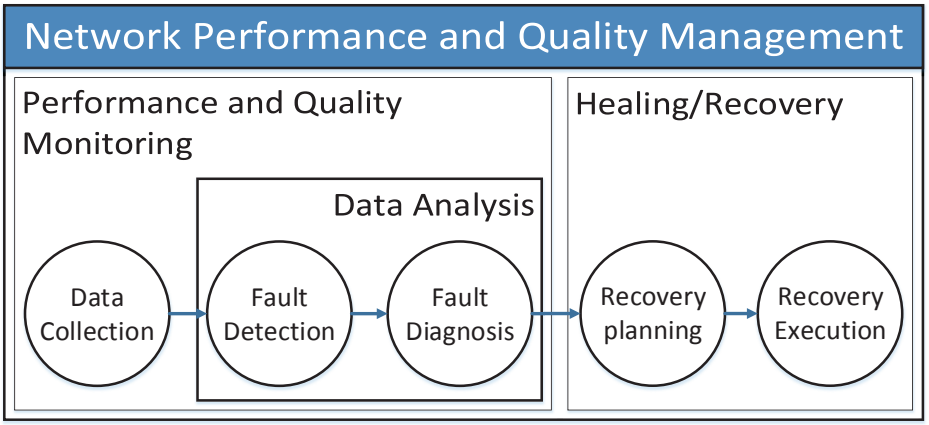


FIGURE 1 Components of mobile network quality and performance management system.

erogeneous Radio Access Networks (SEMAFOUR) is more orientated towards automation of network management and coordination of SON functions.

SON technology is divided into three main parts: self-configuration, self-optimization and self-healing. Nowadays coordination between SON functions is also widely studied, and can be called fourth area of self-organization. The thesis if focused on QPM, thus self-healing and self-optimization are discussed in terms of SON concept. Requirements applied for 3GPP standardization of SON functions are discussed in [7, 8, 9]. Many of 3GPP standards already include a detailed description of SON functions with principles of their operation [10]. Nevertheless, SON is only an initial step towards automation of routine tasks in mobile networks. For that reason, many functions have been the first prototypes in this field, and some simplified assumptions were used for implementation. For instance, triggering condition in self-healing might be based on a single KPI and use a predefined set of fixed thresholds to initiate automatic recovery actions. A simple example of such fault identification approach is utilization of a fixed percentile threshold of cell throughput, or Signal to Interference plus Noise Ratio (SINR) conditions [11, 12].

For further development of QPM systems in general, and SON in particular, more advanced techniques should be applied. To improve PM data collection methods 3GPP included to LTE standards functionality called Minimization of Drive Tests (MDT) [13, 14]. If MDT is configured for periodic reporting, or enabled for large geographical areas with a large number of users, the resulting performance dataset can be multidimensional, and contain from tens of thousands to even millions of entries. Analysis of such data arrays with a traditional approach would require the involvement of a highly qualified network engineer, and most importantly a substantial amount of time. A new way widely studied by the scientific society and also developed in this thesis is to employ methods of knowledge mining [15], such as data mining [16] and anomaly detection [17], which enable efficient and accurate detection and diagnosis of network failures.

A further step can be the improvement of the recovery part of QPM system with advanced data processing methods, such as cognition, based on e.g. machine learning. In dynamically changing conditions traditional and SON-based healing systems are slow and not enough flexible, as the old rules cannot always remain efficient without an update. Cognition addresses this problem, and enables automatic derivation of new recovery solutions.

1.2 Vision and research challenges of Quality and Performance Management

Development of a qualified network engineer takes years of theoretical studies and practical work in the field. Complexity of wireless communication systems is extremely high due to the excessive number of functional elements, constantly evolving radio technologies and the non-uniform nature of network map layouts. Moreover, behavior of users and the corresponding network operation has spatio-temporal dependency, such as day/night, season profiles and also rural, urban or sub-urban environments. In that respect, a network performance analyst has to study in-depth the existing technologies of wireless communication, understand which performance indicators and measurements can represent network behavior. Then, appropriate fault detection thresholds can be set. As it has been discussed earlier, networks are very diverse, and because of that there are only guidelines discussed in books like [3, 18], but there are no universal, ready-made recipes on how to analyze network performance. Thus, every time creation and fine-tuning of an efficient QPM system is a mixture of craftsmanship and art. Naturally, this process is iterative and very slow. For operators, such delays cause increased operational expenditure, and dissatisfaction amongst their customers. Moreover, a traditional approach to QPM, discussed in Chapter 2 is limiting in its nature, as it vastly relies on thresholds, analysis of individual KPIs, and the creation of long-term statistical profiles. Due to these limitations, it is very difficult to create an always up-to-date, flexible and accurate QPM system. All these factors emphasize the need for a new paradigm in performance analysis, based on automation and more intelligent data analysis methods.

Future QPM systems are seen by many researchers as highly adaptive and intelligent [2, 3, 19, 20]. This is motivated by the need to optimize operational costs and efficiency of mobile networks. The key characteristics of QPM system should include the following:

– Accurate detection of network failures. The objective of data analysis in PM is to find failures when they occur. Here it should be taken into account that false alarms (Type I errors) are more critical than miss-detected failures (Type II errors), as they cause unnecessary visits by maintenance personnel, replacement of expensive network equipment and unnecessary configuration changes, which can lead to impaired quality of service. Though, high miss-detection rate makes QPM system useless, as it fails to achieve the goal

of fault detection.

- Timely identification of network failures. Many existing PM systems [20, 21, 22] have demonstrated good results in finding and compensating failures, but the data needed for problem identification has been collected over tens or even thousands of hours of network operation. Such delays reduce subscribers' satisfaction and make perceived Quality of Experience (QoE) smaller. This can lead to income reduction for operators.
- Capability to handle imbalanced performance data, i.e. disproportion in number of normal measurements, if compared to measurements which indicate malfunction, also contributes to complexity of fault identification. For such data, anomaly detection methods should be used for QPM, to take into consideration limitations created by the nature of the data, discussed in [23].

Methodology of research in this area itself is rather complicated. In order to validate the proposed methods for the improvement of QPM, it is necessary to use accurately collected or modeled network performance data and sophisticated network scenarios. One of the most problematic parts is that particular type of failure should be present in the network. This would enable the development of PM fault detection and diagnosis algorithms, based on realistic performance data. In a real network such things are very hard to achieve for several reasons. First, operators aim at a non-interrupted service for their users, and artificially induced problems are not desired, even for future benefits. Secondly, such functions as MDT are not yet broadly taken into use in real networks, as up to the moment mainly Release 8 LTE networks are deployed. Moreover, for certain special kinds of failures, like Random Access Channel (RACH) malfunction, performance monitoring data from real networks are not available for public access. This makes development of corresponding fault detection algorithms very complicated.

To overcome the problems discussed above and propose efficient QPM algorithms and frameworks, their validation and evaluation is done with computer modeling and simulations, prior to deployment in real networks, Figure 2. Link and system level simulations of mobile communication systems like LTE, allow for the collection of the necessary performance statistics and development of the required detection, diagnosis and recovery methods. The key advantage of this approach is that new features or effect of special network failures can be implemented relatively easily. Probably, this is one of the reasons why simulations are widely used for wireless technology development and standardization, e.g. in 3GPP. However, thorough validation and calibration of the developed simulation tools should be done. This proves that the received results are reliable and can be used in future networks.

1.3 Problem statement

In modern networks there are new challenges for QPM systems. Both traditional and SON based PM approaches are not efficient enough to face the problems

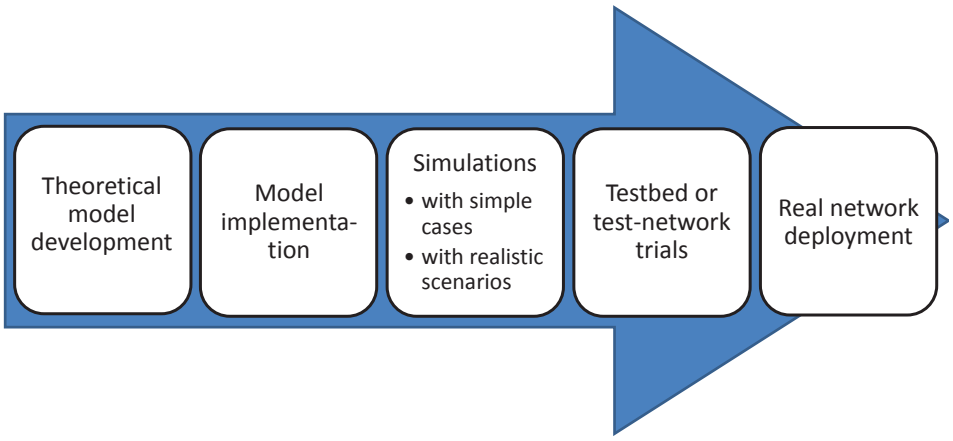


FIGURE 2 Evaluation process of new QPM methods.

of growing heterogeneity and highly dynamic nature of the existing and future networks. Explicit alarms can be used on key hardware elements, but not everywhere. Sometimes these kinds of notifications are not sent in time. In cases of particular non-trivial failures it is difficult, or even impossible to prepare a *predefined* set of rules, which would always enable a correct discrete decision about the state of a faulty or normal network. The potential of collecting and analyzing large performance measurement datasets from different network elements, including user devices is not fully used. The goal of the dissertation is to address the need in timely and accurate performance monitoring through usage of advanced data mining and anomaly detection techniques. Thus, the following problems have been considered:

1. *How to enhance current performance monitoring and self-healing in cellular mobile networks?* In order to address this question a PM framework based on knowledge mining is developed. This implies the application of different data mining and anomaly detection techniques, at pre-processing, transformation, pattern recognition and post-processing phases. Knowledge mining concepts, methods and particular algorithms can be found in Chapter 3. The developed framework and corresponding results are discussed in Chapter 4.
2. *How to detect various types of non-trivial performance malfunctions, such as sleeping cells?* The work described in this dissertation is devoted to the detection of sleeping cell problems. This is a complex type of malfunction and it can be caused by different failures on the network side. A more exact definition of sleeping cell problems can be found in Section 2.2. Particular failures studied in this thesis include: physical layer problem with hardware equipment, and random access channel malfunction. The latter is more sophisticated and complex type of failure from a detection point of view. Approaches to detection of these failures are mainly described in Chapter 4.
3. *How to enrich existing systems of performance monitoring with analysis of new data types, such as periodical and event based user measurement reports?* In at-

tempt to answer this, we analyzed simple event-triggered and periodical measurement MDT data. Special emphasis is put on the analysis of network event sequences, as this information is proved to be useful for revealing complex network failures.

4. *How to integrate performance monitoring based on anomaly detection methods and cognitive recovery planning?* This is discussed in Section 4.2, and is largely based on results described in [PVI]. The developed demonstrator system enables efficient cooperative usage of cognitive recovery analysis and anomaly analysis based on data mining.
5. *What future directions of knowledge mining application to advanced QPM in cellular mobile networks should be studied?* Advanced data mining techniques are largely aimed at the analysis of big datasets, and for that reason flow of data between network elements might be increased, what may cause overhead on the corresponding connection links. Thus, advanced state of the art QPM studies, future prospects and a cognitive self-healing architecture are presented in the end of Chapter 4.

1.4 Outline of the dissertation

The rest of this dissertation is organized in the following way.

Chapter 2 is devoted to overview of the existing QPM systems. First the flow of the QPM process is presented. Then classification of different QPM generations, based on the extent of introduced automation is discussed. After that different network failures and their possible root causes are elaborated and sleeping cell problem is defined. The most common sources of performance monitoring data are then presented. This includes concepts of both cell-level data collection, and user-specific measurement data and reporting based on TRACE and MDT functions. These data collection methods are discussed in detail starting from architectural options to data gathering modes. Also the concept of mobile quality agents is introduced. Then description of the KPIs and statistical profiles for performance monitoring is given. The key issues of traditional PM systems are presented. Next the notion of automation and the SON concept are presented. The role of self-optimization and self-healing in QPM process is discussed. Also the place of these functionalities in 3GPP standardization is outlined. Thorough attention is given to discussion of different MDT modes and use cases.

Chapter 3 describes the notion and the key steps of knowledge mining. This chapter gives motivation for using knowledge mining in general and presents the "curse of dimensionality". Different types with examples from real mobile networks are outlined. Data mining and anomaly detection – standardization, transformation, pattern recognition and post processing are thoroughly discussed. Such methods as sliding window and z-score transformation, and dimensionality reduction by means of diffusion maps, Minor Component Analysis (MCA) and Second ORder sTatistic of the Eigenvalues (SORTE) [24] techniques are pre-

sented. Also algorithms for sequence-based data analysis with N-gram, classification with nearest neighbors, K-means and FindCBLOF clustering are discussed. Different detection quality evaluation techniques are presented, such as confusion matrix, precision, recall, F-score and ROC curve.

Chapter 4 illustrates the application of advanced analysis methods based on knowledge mining in QPM. The most recent research activities in this field are reviewed, compared and classified. Analysis of performance data with an enhanced statistical approach, Bayesian networks, neural networks, clustering, classification techniques and ensemble methods are outlined. The author's contribution and the results of the original research in application of knowledge mining for construction of efficient QPM system are discussed. The presented analysis demonstrates how user reported MDT data can be processed with various data mining and anomaly detection techniques for identification of sleeping cell failures. Additionally, the strengths and weaknesses of knowledge mining for QPM are outlined. The discussion section 4.3 gives prospects of QPM systems' development and concluding remarks regarding the role of cognition in cellular mobile networks.

Chapter 5 concludes the dissertation with a short summary of the current situation and future needs in advanced QPM methods based on knowledge mining, combined with self-organization and development of future mobile networks.

Appendix 1 presents the comparison of the advanced QPM methods for MDT data. Description of the utilized simulation tools, and references to 3GPP validation documents can be found in **Appendix 2** of this dissertation.

1.5 Main contribution

Published articles are devoted to the analysis of MDT data with advanced performance monitoring methods based on knowledge mining. The presented results can be divided into two logical parts – analysis of numerical and sequential characteristics of the collected performance data. Also there is a difference in the root cause of the sleeping cell, which is the main detection object. In one group of studies, a more simple case of hardware failure is considered. Another case analyzes a random access channel failure - which represents a more complex type of sleeping cell. Articles devoted to analysis of numerical characteristics of MDT data for detection of Hardware (HW) failure are the following.

In publication [PIV], detection of a physical hardware antenna gain sleeping cell is studied. Identification is done by means of data mining in event-triggered MDT measurement reports. First, dimensionality reduction with diffusion maps is applied. Then in the embedded diffusion space, iterative k-means unsupervised clustering is done. This paper describes the initial work, which demonstrates that analysis of high-dimensional MDT data with anomaly detection techniques can be used for identification of network failures. The author of this thesis is the first author of the article and is responsible for proposing data

mining clustering framework used for the detection of the failure. Hence, writing the paper and the results analysis are the main author's contributions.

In article [PV], the detected problem is cell outage due to physical malfunction of signal amplification circuitry. The main difference to the analysis described in all other papers ([PI], [PII], [PIII], [PIV], [PVI]) is that MDT reports, which consist of various measured KPI values are both event based and periodic. This significantly extends the MDT database. The first step of the analysis is transformation - reduction of data dimensionality with diffusion maps. Classification with K-Nearest Neighbors (K-NN) algorithm has been done to three classes: periodical, Handover (HO), or Radio Link Failure (RLF). Thus, the main goal is to find samples which resemble RLFs and by that increase reliability of anomaly detection and identification of the sleeping cell. Publication of [PV] has been mainly done by Dr. Jussi Turkka. Author's contribution to this publication was partial data analysis and peer-review of the article at the writing stage.

In publication [PVI], cell outage compensation is done by means of fault detection and cognitive iterative recovery planning and execution. Identification of network failures is based on user measurement reports and combines anomaly detection methods and statistical profiling. The recovery part relies on modified case-based reasoning algorithms, which allows for the prioritizing of efficient solutions and develop new ones if the systems behavior changes. The demonstration system presented in this publication is a result of collaborative work of Magister Solutions Ltd. and Nokia Solutions and Networks during Celtic+ COMMUNE project. Definition of simulation assumptions and development of anomaly analysis entity have been made by the author, as well as writing of the paper. Results of this research have also been presented by the author at Celtic+ Event 2014 and COMMUNE final review meeting.

The articles discussed next are devoted to the analysis of sequential characteristics of the user performance data and the modeled failure is random access channel malfunction. Detection of this problem using the MDT reports is possible in 3^{rd} Generation (3G) and LTE, LTE-A networks. The necessity to do the timely identification is caused by different hidden failures, which negatively affect the networks quality. For instance, a situation when the user is handed over to a different RAT because of malfunction in the original RAT, would lead to a reduction in the quality of service and increased consumption of network resources. Such cases should be avoided.

Publication [PIII] introduces the knowledge mining detection framework based on pre-processing, transformation, clustering and post-processing techniques. With this framework detection of users with abnormal behavior is achieved. The applied method for sequence-based analysis is N-gram, which makes it possible to identify the cell, which causes the anomalous user behavior. This proves that the sequence-based approach is beneficial for detection of network failures, as it is demonstrated in this paper. The author of this thesis proposed a novel approach of sequence-based analysis for network QPM and participated in the work of the research group responsible for data analysis. Also implementation of the sleeping cell failure, configuration and run of the dynamic system level

simulations have been done by the author.

In publication [PI] an approach for sleeping cell detection based on the analysis of mobility related event sequences is further developed. The main improvement in this study is that the duration of user calls, i.e. the number of occurred mobility events, does not negatively affect the detection accuracy anymore. This is achieved with the application of the sliding window approach. N-gram analysis remains to be the core of the proposed advanced PM framework. For actual anomaly detection a combination of transformation and K-nearest neighbor anomaly score is used. Additionally, k-fold cross-validation, and various post-processing techniques are used. Results of detection are compared using both traditional (ROC, F-Score, etc.) and heuristic approaches. Also validation tests are done to demonstrate that false alarms are not triggered. As a result, the random access sleeping cell problem is timely and reliably detected. Similarly to paper [PIII] the author has been configuring and running simulations, and contributed to brainstorming for creation of the data analysis framework. Writing of the paper has been started by the research group, but is mainly done by the author.

Paper [PII] is based on the framework developed in article [PI] and is devoted to the detection of a random access channel sleeping cell problem. The results demonstrate that usage of 1-gram for analysis of mobility related sequences of MDT events, also leads to sufficiently good results of sleeping cell detection, if compared to the 2-gram approach. The result gives a slight increment in reduction of computational complexity, and potentially can be used for preliminary detection of suspicious and problematic cells. The role of the author is limited to peer review of this paper, though the presented results are derived using the framework presented in [PI], and are partly based on the findings from [PIII].

Additionally, the author published a number of articles in international conferences. These papers are devoted to QoS verification for MDT and optimization of radio resource control procedures in LTE. Here is a list of original publications:

- Jani Puttonen, Fedor Chernogorov: The Effect of Discontinuous Reception and RRC Release Timer Parameterization on Mobility, 79^{th} *Vehicular Technology Conference, 2014*.
- Fedor Chernogorov, Jani Puttonen: User Satisfaction Classification for Minimization of Drive Tests QoS Verification, 24^{th} *Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications, 2013*.
- Fedor Chernogorov, Timo Nihtilä: QoS Verification for Minimization of Drive Tests in LTE Networks, 75^{th} *Vehicular Technology Conference), 2012*.

2 QUALITY AND PERFORMANCE MANAGEMENT IN CELLULAR MOBILE NETWORKS

In this chapter we present the phases of quality and performance management in cellular mobile networks, starting from data collection methods and finishing with recovery. The main focus is on traditional and Self-Organizing Network (SON) based QPM methods, with a description of their main weaknesses and strengths. The evolution path of QPM systems is presented. A description of the most common failure types is given.

QPM is a process of controlling the mobile network operability, performance and service quality through collection and analysis of the monitoring statistics, with consecutive optimization or recovery. This definition summarizes the discussion on network management and performance analysis encountered in the literature. Other authors use different terms to denote QPM, for instance: Operational Fault Detection (OFD) [20], Service Quality Management (SQM) [3], or Network Management (NM) based on a combination of FM, PM, CM systems (Fault Management, Performance Monitoring and Configuration Management correspondingly), as it is discussed in [2]. Fault management is also sometimes referred to as incident management. In our terminology, PM combines together functionality of performance and fault management systems defined for SON [2]. Both data collection and data analysis are included in PM. Recovery execution can be treated as configuration management in terms of SON. The flow of QPM process is shown in Figure 3, partly based on [4]. Descriptions of each stage, give an idea which procedures and processing is done, and mainly refer to traditional or SON-based QPM systems. The first step of QPM is collection of performance monitoring data from the network. Usually there are many types of performance indicators collected from different network elements, as it is discussed in more details in Sections 2.3. Because of that, the resulting performance dataset can be large and multidimensional, especially when MDT functionality is enabled (see Section 2.5.3). In order to elicit meaningful information from such dataset, gathered data requires analysis. There might be different objectives for analyzing the data, for instance: optimization of mobility performance, identification of the loaded and unloaded regions for balancing and coverage improvement, or

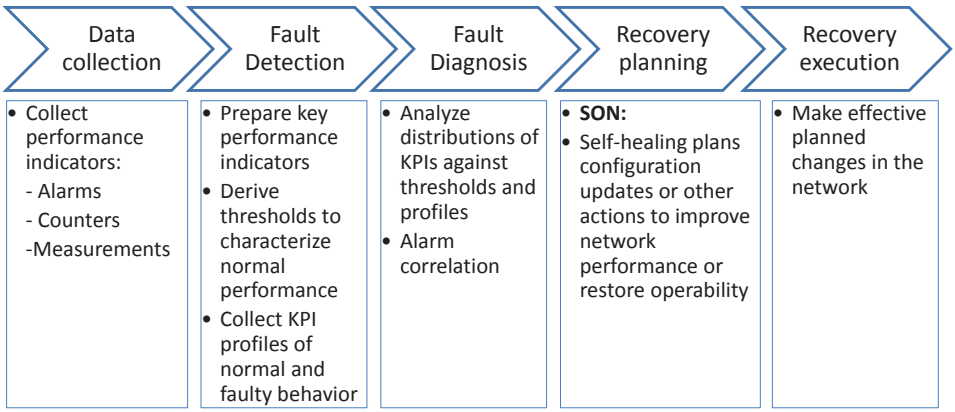


FIGURE 3 Flow of Quality and Performance Management.

detection of failures and malfunctions. However, timely detection and diagnosis of network failures (stages 2 and 3 of QPM process, shown in Figure 3) are especially important, as malfunctioning cells cause significant reduction in network performance quality and lead to dissatisfaction of the subscribers. The last two stages are devoted to recovery analysis, aimed at compensation and healing of the occurred malfunctions.

2.1 Evolution of Quality and Performance Management systems

In order to classify the different approaches and methods of performance data processing and network management, we describe QPM evolution, comprised of 5 generations, shown in Figure 4. This classification is proposed by the author of this dissertation and has never been presented in any literature before. It is partly based on [2]. The main differentiating factors between the generations are the extent of automation introduced into one or several stages of QPM and the amount of time required for the whole cycle - from data collection to recovery actions. In order to make network QPM more autonomous, expert knowledge and intelligent computing methods are employed. Here is the description of the evolution path generation by generation:

Generation A: Traditional methods. In the traditional quality and performance management systems representation of the network operational state is usually based on a collection of Performance Indicators (PIs), discussed in Section 2.3. Detection is semi-automated, as aggregation of PI levels is used, KPI behavior is judged on the basis of thresholds and statistical profiles (see Section 2.4). Diagnosis and recovery are mostly manual and are based on expert knowledge. Engineers responsible for these tasks have to observe a large number of KPIs, to decide on the basis of validity intervals and predefined thresholds, whether the network is in a normal state or not. Thus, in

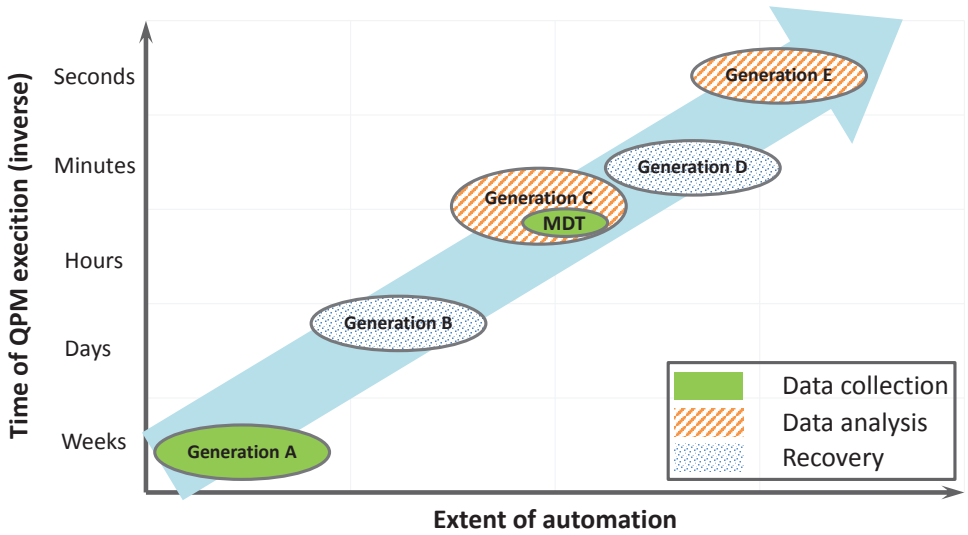


FIGURE 4 Evolution of mobile network quality and performance management systems.

this QPM generation there is a large extent of human involvement.

Generation B: Self-healing. In the second generation, automation is introduced mainly in recovery planning and the execution phases. Self-healing SON concept is used to automate recovery through reconfiguration of radio equipment. Detection and diagnosis functions have been slightly improved: in addition to the analysis of KPIs from the first generation, alarm correlation methods have been developed [25, 26, 27].

Generation C: Advanced performance monitoring. In the third generation, fault detection and diagnosis has been enriched with advanced performance monitoring techniques based on data mining, machine learning, anomaly detection and multivariate statistical analysis. Chapter 3 is devoted to the description of the concept of data mining and anomaly detection, and outlines some of the popular methods. Advanced analysis is necessary to enable faster and more accurate fault detection, and automated fault diagnosis. State of the art QPM systems with advanced performance monitoring are discussed and compared in Chapter 4. In Generation C, recovery planning and execution are based on self-healing functionality, the same as in QPM Generation B. However, a combination of self-healing and advanced PM is not always used.

Generation D: Cognitive network recovery. Cognition, discussed in Section 4.2, is introduced in the recovery planning phase. This allows for the development and selection of more efficient solutions to network failures, by means of learning positive and negative output of the previously taken corrective actions. This means a network creates a knowledge base and uses "experience" to make further decisions.

Generation E: Cognitive data analysis. Cognition is introduced to fault detection

and diagnosis, what makes possible automatic selection of data mining algorithms and their parameters. E.g. for different data types, different anomaly detection or classification techniques could be used. The recovery stage is also fully cognitive.

It is important to emphasize that for the operator, network management practices change from generation to generation, in particular the amount of manual work is declining, what reduces the cost of QPM cycle. Thus, first generations require manual configuration of data collection. Then technical personnel ought to control and react to suspicious changes at all stages of performance monitoring. With the increase of automation and the introduction of cognitive techniques, an operator's role more and more shifts towards development of performance policies, quality targets and high-level economical goals. One aspect which remains to be manual and inevitable is the provision of feedback to the QPM system regarding correctness and quality of the taken decisions and suggested solutions.

2.2 Network Failures and Malfunctions

First, it is necessary to define some key terms related to improper network service. *Fault, failure, malfunction* and *breakdown* are equivalent terms and they denote that element or function in this condition is not capable to maintain its normal operation. *Degradation*, refers to a reduction of network service quality, due to the incapability to execute all operational functions. Obviously, degradation conditions can be caused by mistakes in the configuration of network equipment or failures, and recovery is responsible for handling both of them.

Network failures in cellular communication systems are very diverse, and can be classified in several ways. One approach to grouping of faults is based on their type, as is discussed in [2, 28]:

- hardware;
- software;
- functional, e.g. handover or link power control;
- overload condition;
- communication link or channel deterioration;
- inappropriate configuration.

Knowledge about the type of problem gives partial information about the root cause. For instance, if the encountered failure is related to HW or Software (SW) component, it explicitly denotes the problem. However, functional, overload or channel failures might be caused by different reasons, and additional information is needed to make the final diagnosis decision. Faults related to configuration are related to mistakes in network planning or configuration updates. For instance, there are known trade-offs in network configuration optimization, e.g. in resource allocation (scheduling) [29], parameters related to user mobility or

user battery saving functions (Discontinuous Reception (DRX))) [30]. As a result for certain cells, the same configuration can be appropriate or erroneous. An additional factor, which might make the existing configuration invalid, is a change in the propagation environment, such as construction or demolition of highways, stadiums or other major buildings. One more approach to malfunction typification is based on the failure scope as a grouping factor. The affected entity can be an individual network element, a cell, a site, e.g. with multiple cells, a pair of cells, a domain, and finally, a whole network. This is more like problem location classification, as it describes what, where and how many Network Elements (NEs) are affected. Information regarding root causes is not contained in this classification. The next approach is based on the extent of degradation caused by the failure, and naturally, it does not contain information regarding possible root causes. Degradation can be partial and complete, which is referred to as outage. There are three types of cell degradation, which describe the extent of degradation and service availability (largely based on [31], [32]):

- *Deteriorated* - slightly reduced functionality and corresponding minimal negative effect on QoS. This term in the original classification is "degraded", but in our case degradation is a more generic term.
- *Crippled* - severe shortage in provided service, and noticeable reduction of QoS level. However, to a certain extent, the original functionality is maintained.
- *Catatonic* - complete lack of service, which corresponds to cell outage.

One more, very important classification, is based on the awareness of the operator's QPM system regarding the failure occurrence. As it is discussed in Section 2.3 some faults trigger explicit alarms, e.g. in case of a complete breakdown of certain network elements. Some other malfunctions can be deduced from observation of KPIs, as the reported values are outside the realistic value range. However, for some faults, values of individual KPIs remain nearly normal, and no alarms are reported, but users partially or completely do not get the service. This is called sleeping cell problem [20], [32], [33], [31], [34]. Due to the hidden nature of sleeping cells there is a need for accurate and timely detection, diagnosis and healing of such problems. Complexity in achievement of these goals follows from the diversity of origins of sleeping cells. They can be caused by both hardware, software failures and inappropriate configuration as well. For example, Uplink (UL) or Downlink (DL) antenna gain can be incorrect, due to a fault in the amplifier and that would lead to coverage related sleeping cell. Another type of sleeping cell root causes is SW fault. For instance, malfunctions on Medium Access Control (MAC) layer can impair some logical channels, crucial for cell operation. There can be partial and total sleeping cells, due to SW problems [34]. Partial sleeping cells, can receive the connection request from the user, initiated with random access procedure, but no connection setup message is responded. Total sleeping cells do not even accept connection request message. Such situation can happen at RACH [30] malfunction, which can be caused by both SW fault in base station, incorrect transmit power setting, congestion and misconfig-

uration of e.g. random access response timer or back-off parameter [35]. Random access is critical for operation of mobile networks, as it is used in the key procedures related to user connection. For instance in LTE the following situations require random access [30]:

1. User Equipment (UE) changes it state from Radio Resource Control (RRC) IDLE to RRC CONNECTED, e.g. when the initial UL access is established or tracking area is updated.
2. UE in RRC CONNECTED state tries to hand over from its current serving cell to a target cell.
3. Connection re-establishment procedure after RLF.
4. UE in RRC CONNECTED state, but not UL-synchronized.
5. Periodical positioning update of UE in RRC CONNECTED state.

It can be seen that if random access does not work properly in one of the first four cases described above, this would result in connection breaks, the inability to establish connection or send/receive data. Moreover, in many cases malfunctioning random access is hidden, as it cannot be revealed with the existing measurements [13]. Thus, additional methods for detection of RACH problems are needed.

In general, one can see that origin and consequences of failures in mobile networks are very diverse. Some of the failures are easy to detect, as they are explicit, but some malfunctions are hidden, which would require a great deal of effort to identify before these problems cause a substantial negative impact. In this dissertation we consider detection of both HW breakdowns - problems with transmit power and antenna gain of a cell, and SW errors, which lead to connection issues due to a faulty random access channel. However, in order to achieve accurate detection, it is necessary to collect the necessary performance data. The next section is devoted to the description of the types and sources of the performance data in mobile networks.

2.3 Data collection in traditional performance monitoring systems

Performance monitoring data, which can be collected in cellular mobile networks include: alarms, counters and measurements, discussed in this section. Through analysis of this data, network performance can be observed. A general term which describes different kinds of monitoring data is Performance Indicator, defined as an informative measure regarding the performance of network element, function or process [3]. Here we discuss different types of PIs and outline their role in the performance monitoring process. But first the sources, i.e. collection points of PIs in the network are outlined.

2.3.1 Data sources

In traditional PM systems PIs can be collected from different sources and by means of a variety of tools, such as drive tests, network interface tracing, sig-

naling element counters, transport network statistics [36], mobile quality agents and standardized TRACE functionality. The main idea of drive testing is to measure the network performance from the perspective of a user. To achieve this, specialized radio measurement equipment, placed in a van, travels around the network. The process of driving around the network and collecting various test statistics gave its name to this method - "drive testing" [37]. Equipment used in drive tests are measurement receivers and test engineering phones. The first ones are responsible for capturing the Radio Frequency (RF) picture at the observation point, and provides accurate information regarding network coverage. Instrumented test phones, provide more connectivity and mobility related measurements. Also test phones are utilized to observe user throughput, packet delay and other QoS metrics [38, 39]. Operators carry out drive test campaigns after deployment of new base stations or cells, in case of construction of major objects, like buildings or highways, periodically or due to customers' complaints [13]. Cumulative costs of drive testing is rather high, if we consider expenses for the radio measurement equipment, working hours of qualified network engineers, carrying out drive tests, and analyzing the collected data, and additional expenditures, like the price of gasoline. Moreover, changing environment calls for frequent testing.

Probing systems provide the view on networks performance in static points, and in many cases are passive, unattended entities. The core idea of using probes is to provide an independent view on network operation either in terms of RF measurements or from the perspective of interface performance and traffic volume. The main disadvantage of probe systems is their high cost, what reduces their usage. Signaling element counters are responsible for the collection of performance information related to data. Transport network statistics is more related to core network operation, backhaul, and interfaces between elements. E-UTRAN NodeB (eNB) tracing gives a very detailed picture of signaling which terminates at the considered eNB. It is an isolated view on one particular base station or cell, and for that reason some additional analysis techniques are needed to grasp the idea about overall network performance.

Standardized TRACE deserves more attention, as it lays the grounds for MDT functions, discussed in Section 2.5.3. TRACE is implemented in all generations of digital cellular mobile networks, namely for GSM [40], UMTS and Evolved Packet System (EPS) [41]. This functionality is capable of collecting detailed information with respect to a particular subscriber, mobile terminal equipment, service, ability to collect data about all active calls in a cell or group of cells (cell traffic trace). There are two TRACE modes which are used to trace different elements. Signaling-based mode is aimed at tracing specific subscription or equipment, while management-based mode traces users in a particular area of the network, i.e. in one or several cells [42]. These differences affect how activation and deactivation of trace functions is done in the network. Architectures of signaling and management-based trace are presented in Figure 5. It is visible, that for tracing a particular user, an element manager has to first address Home Location Register (HLR) or Home Subscriber Server (HSS), which provide

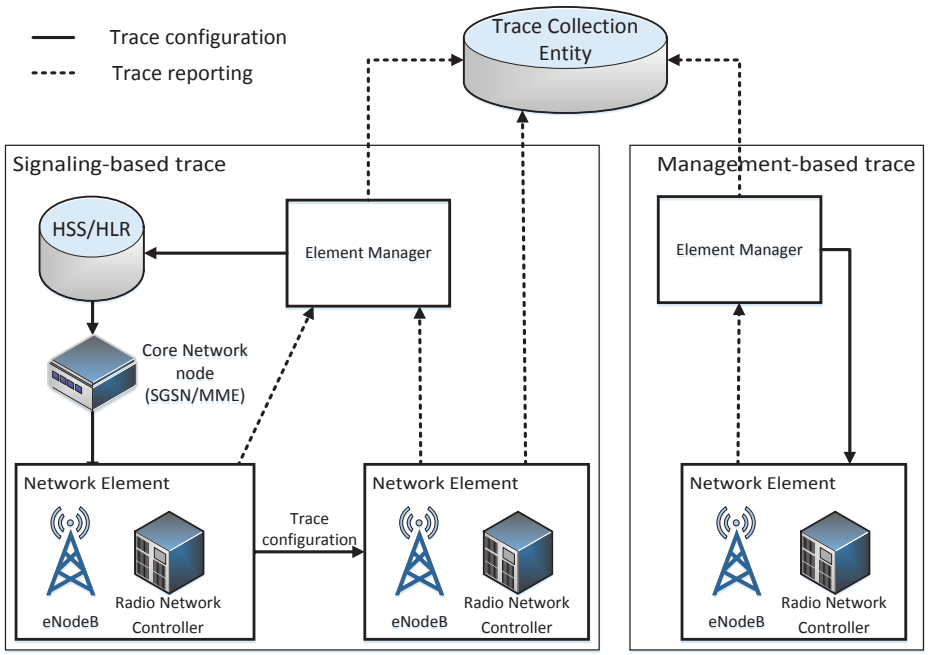


FIGURE 5 Management and signaling based TRACE.

eNB or Radio Network Controller (RNC) with appropriate configuration through a corresponding core network element. For management-based trace it is not necessary to retrieve information regarding a specific user or its equipment, and configuration can be done directly from a trace element manager to the network elements participating in traced data collection. Reporting of the gathered data information can be done through the element manager to the TRACE Collection Entity (TCE), or directly from the network element to TCE, depending on the capability of the network node. Data collected with trace is very diverse and illustrates the state of the user terminal at interfaces (e.g. Uu, Iur, Iub [43, 44]), of various network elements, such as, Serving GPRS Support Node (SGSN), HSS, Packet Gateway (PGW), Mobility Management Entity (MME), Serving Gateway (SGW), to name a few. The data itself is an indication of a particular event, e.g. location update, handover request, handover command, measurement report type, and many more as discussed in [45]. In addition, it is possible to choose the "depth" of the trace, by configuring it to have "Min", "Med" or "Max" level of detail.

Thus, the amount of information which can be elicited by trace campaigns is substantial. However there are no proposed, well established, and efficient ways of analyzing these data sets with conventional methods. Nowadays more attention in performance monitoring is given to processing of KPIs. This makes TRACE functionality a good potential source of data for advanced PM analysis, discussed further in Chapter 4.

Mobile or device quality agents are also widely used by network operators for control of performance quality. This is an intermediate solution for systems

where MDT is not yet available. Mobile Quality Agents (MQAs) can be treated as proprietary implementation of MDT, where the core idea is collection of performance measurement data from the end user equipment. The advantage of device quality agents is that they support multi-RAT performance statistics collection and most importantly, provide the user perspective of the service quality. The drawback of MQAs is that they are not standardized data source, such as MDT, which is capable to provide unified performance information from the whole network. It is highly unlikely that all of the users install MQA in their terminal equipment, and because of that MDT is a preferable source of data with full control from the core network side. However, advanced performance monitoring methods, presented in Chapter 4 are also applicable for the data collected with mobile quality agents.

2.3.2 Alarms

Alarms are PIs, produced by the network elements in cases of critical malfunctions related to hardware or software. Usually alarms are equipment-dependent and because of that are very diverse. Examples of alarms are modulation link failures, high equipment temperature, malfunctioning base band processor, etc [19]. Alarm is a type of fault management data, generated on the basis of configurable thresholds [3]. For further analysis alarms are sent to the network management system, which decides on the service performance. In other words, network alarms can be an indication of failure in a particular element, sent towards Operations, Administration, and Maintenance (OAM) system. Such notifications can also be produced by violation of a threshold for particular KPI [2]. With respect to usability of alarm information for QPM purposes, there are certain limitations. First of all, alarms are isolated from other elements and network functions, and because of that have a very limited scope. Hence, a broader picture of network performance cannot be deduced from a single alarm. That in turn, can lead to an incorrect interpretation of the existing network state, unnecessary actions, like base station site visits, and as a result increased cost of network maintenance. One more possible source of errors in rising alarms is related to KPI threshold settings. As it is discussed in Section 2.4.2, it might be a complicated task to derive a correct universal threshold, taking into account network dynamics and heterogeneity.

2.3.3 Counters

The simplest and most common variant of counter is cumulative value [2, 3]. In this case, counter is a simple natural value, which starts at 0 and is incremented every time the counted event occurs. In mobile networks such events can be very different, for instance the number of handovers or RLFs, data packet retransmissions, connection attempts, rejects, amount of transmitted/received data, etc [46]. Collected counter values, which can be used to evaluate how operational are different network functions state are called PIs. A list of various protocols' coun-

TABLE 1 Measurement triggering events in LTE.

Event name	Intra-Inter-RAT	Event description
A1	Intra	Serving cell becomes better than the absolute threshold
A2	Intra	Serving cell becomes worse than the absolute threshold
A3	Intra	Neighbour cell becomes an offset better than the serving cell
A4	Intra	Neighbour cell becomes better than the absolute threshold
A5	Intra	Serving cell becomes worse than one absolute threshold and the neighbor cell becomes better than another absolute threshold
A6	Intra	Neighbor becomes offset better than the secondary cell
B1	Inter	Neighbor cell becomes better than the absolute threshold
B2	Inter	Serving cell becomes worse than one absolute threshold and the neighbor cell becomes better than another absolute threshold.

ters and corresponding configurations for different technologies, such as GSM, UMTS, along with related definitions can be found in a family of 3GPP specifications [47, 48, 49, 50, 51, 52, 53].

2.3.4 Measurements

Radio-related measurements in mobile networks can be divided according to the measuring element: either a base station or a user terminal. Measurements taken by a base station can be common, i.e. related to a whole cell, or dedicated to a specific connection [54]. For instance, in Universal Terrestrial Radio Access Network (UTRAN) the following common measurements are taken: received total wideband power, transmit carrier power, preambles of Physical Random Access Channel (PRACH), and in Evolved Universal Terrestrial Radio Access Network (E-UTRAN) those are received interference power, DL reference signal transmission power, thermal noise power, average SINR, list of detected preambles, UL Channel State Indicator (CSI), [55, 56]. Dedicated measurements include timing advance, angle of UL signal arrival, Signal to Interference Ratio (SIR), SIR error, transmitted code power in UTRAN. User terminal measurements are Universal Terrestrial Radio Access (UTRA) Common Pilot Channel (CPICH) E_c/N_0, UTRA Frequency Division Duplexing (FDD) CPICH Received Signal Code Power (RSCP), Received Signal Strength Indicator (RSSI) in both GSM and UTRA, in E-UTRAN Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), SINR in form of Channel Quality Indicator (CQI) levels, event Identifications (IDs) [56]. Measurement triggering events define when a particular measurement will be reported to a network. Users periodically make the necessary measurements of the serving cell, and or certain neighbor cells and when the triggering event conditions are met, the report is sent to the base station. A

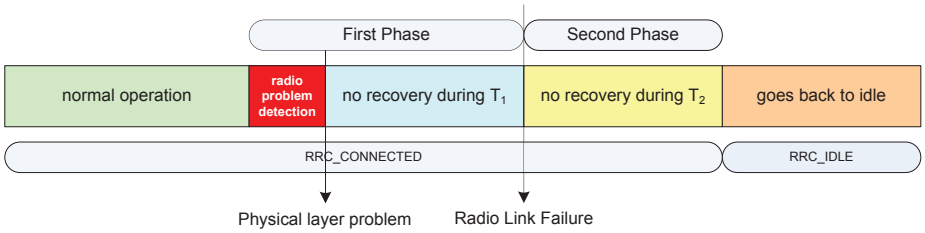


FIGURE 6 Radio link failure procedure in LTE.

list of triggering of events in LTE networks and corresponding descriptions is presented in Table 1. These events, pre-configured in the UE by the network [30], are used for mobility and can partly reveal coverage issues. It is important to mention that signal strength measurements for event triggering can be based on either RSRP (RSCP in UTRAN) or RSRQ, but common practice is to use RSRP. Other types of report, existent in the network, and directly related to channel conditions is Radio Link Failure (RLF). This event is triggered in case of SINR level is under the specified threshold (this point is called physical layer problem) for the duration of a configured timer [57]. The overall procedure of triggering RLF is shown in Figure 6 [58, 59, 60]. A user goes into idle mode if, after a specified timer connection is not re-established. Obviously some periodic measurements are also reported, such as e.g. CQI, used for general Radio Resource Management (RRM). Configuration of measurement collection, with respect to its type, time and frequency of reporting is obligatory and at the same time a complicated task. Standardized classification and approach to configuration of radio measurements can be found in [61] for UTRAN and [57] for E-UTRAN.

2.4 Traditional data analysis methods for performance monitoring

This section describes conventional ways of running data analysis in performance monitoring. The most popular approaches are aggregation of PIs, derivation of new KPIs, selection of thresholds, and analysis of statistical distributions and profiles of KPIs. Also Key Quality Indicator (KQI) are used to evaluate network QoS. In addition, we outline the advantages and disadvantages of the existing traditional methods of data analysis in PM.

2.4.1 Key Performance Indicators

Aggregation of PIs implies representation of their most important characteristics, within a particular scope, also called dimension. The most representative PM data dimensions are shown in Figure 7 (partly based on [3]). Performance indicators can be any data value discussed in Section 2.3. Network element implies a cell, a user terminal, a group of cells, etc. Time is PI aggregation interval. The fourth dimension is some network resource, which can be physical, e.g. carrier

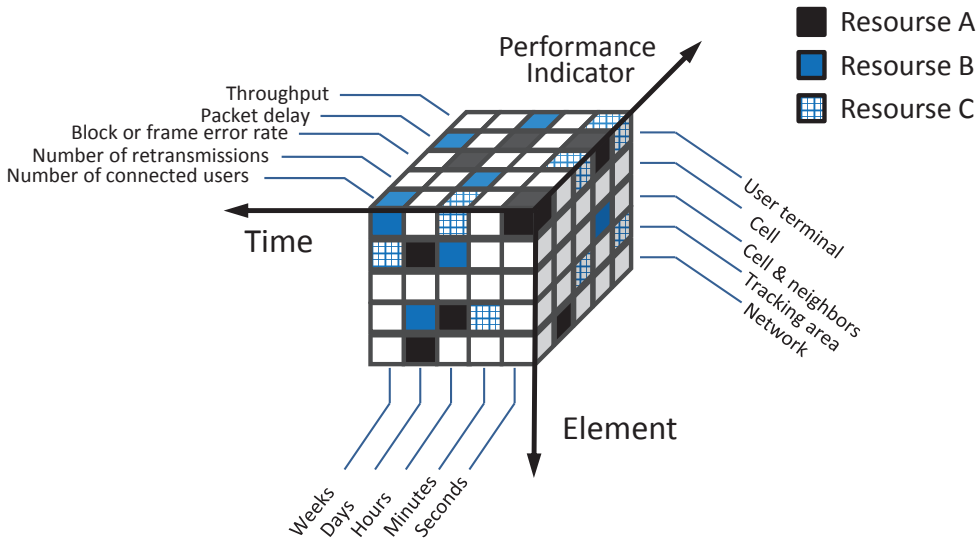


FIGURE 7 Key performance indicator aggregation dimensions.

frequency, channel or scrambling code, or logical, such as a call, call type, transport channel, transmission direction, and so on. Thus, PM data can represent specific PI, say throughput, for a particular element - a cell, and describe a specified period of time, e.g. 15 minutes, and for a particular carrier frequency. Thus, using the four described dimensions, one should control the performance of elements, groups of elements and the network as a whole. In order to achieve this, it is necessary to select the most representative PIs, if necessary convert them, and choose the most informative setup on observation dimensions. Some of PIs can be generalized to represent network performance in a broader scope, e.g. longer time scale or over a large group of elements. Such an approach to data representation is usually calculated according to a predefined formula and called Key Performance Indicator (KPI). According to the general definition, KPI is a measurement and a quantitative measure of the most important performance parameters of operation in the monitored system [62].

KPIs are broadly used in traditional network quality and performance management. From a statistical point of view, KPI can be a representation of multiple PIs samples with one statistical value, such as mean, median, mode, deviation, maximum or minimum [63]. Additionally, various percentiles [64] of the PM data statistical distribution are used for representation, for instance 5^{th} and 95^{th} percentiles are extremely popular, and even a whole section is devoted to usage of the 95^{th} percentile in [18]. Selection of an appropriate statistical metric purely depends on the nature of the PI, and network elements. I.e. representations of PI which give most for understanding of a networks performance should be chosen. An important characteristic of KPIs is that they are more network, or even cell oriented, i.e. aimed at representation of a networks operational state, and not related to any particular service. Hence, from KPIs it is not obvious what QoS level is achieved, and in that way they are different from KQIs. In general Key

Quality Indicator (KQI), is used to represent quality of provided network service. According to definitions in International Telecommunication Union (ITU) standards, quality of service means the degree of user satisfaction with the provided service [65], and a degree of conformance between promised and provided service [66]. Thus, KQIs mostly illustrate non-network related performance, e.g. provision time, repair time and complaint resolution time [65], or represent a broader view of a networks operation using high level aggregated values of KPIs [2]. Such as the selection of thresholds and the creation of statistical profiles for individual KPIs. For cellular mobile networks KPIs have to comply with the following requirements [67]:

- Accessibility of the network, connection and service - which are probabilities to access request accepted, followed by connection establishment, and the ability to obtain the necessary service through this connection, provided by the network [65].
- Retainability - ability of the service provision or connection to be maintained, i.e. guaranty that the number of interruptions will be minor.
- Integrity - the property that data has not been altered in an unauthorized manner [65].
- Availability - physical presence of the network, which is capable of providing the necessary service.
- Mobility - ability of users to receive service and maintain connection during movement. In other words, it is a technical possibility to make handovers between different cells without connection interruption.

Thus, the main goal of KPIs is to represent the technical aspects of the networks performance, from the perspective of requirements listed above. KPIs are widely accepted and for that reason some of them are included into the RAT standards. Description of GSM and UMTS KPIs can be found in [62], while E-UTRAN LTE KPIs are described in [68]. In addition, some of the KPIs are proprietary and equipment dependent, or developed internally by the operators. A sample list of the most common KPIs used to characterize network performance in terms of the described requirements is shown in Table 2 [18, 43, 69, 36]. The natural advantage of KPIs is that network behavior is presented in a condensed manner, as aggregation over the dimensions discussed earlier is done. The downside of such an approach is that collection of information should be done for sufficiently long periods of time to be statistically reliable. Another problem is that due to aggregation, a large part of potentially meaningful data is left out. This raises the question - are there different ways to process non-averaged performance data and find network problems faster and more accurately? Chapter 4 outlines the research activities and the latest results in the area of advanced performance monitoring. However, as it can be seen from Table 2, the number of network KPIs is large, hence they can be collected in the form of a multi-dimensional data array. Traditional approaches, such as thresholds and statistical profiles are mostly capable of individual, rather than mutual analysis of KPIs. This also leaves space

TABLE 2 Examples of network KPIs.

Bit Error Rate (BER)	Frame Error Rate (FER)
Block Error Rate (BLER)	Frequency Reuse Factor (FRF)
Drop Call Ratio (DCR)	Total base station transmission power
Radio signal strength measurement (e.g. RSRP or RSSI)	Radio channel quality - SINR, CQI, Chip energy over noise (Ec/N_0), SIR
SIR error	Cell load ratio: used vs available RBs
Active users ratio	Number of connected users
Connection success ratio	Call blocking ratio
Service specific radio bearer accessibility	SIP success rate
Attach success/failure ratio and setup time	Paging success/failure ratio and setup time
UE context success/failure ratio and setup time	Retainability rate [#drops/session time]
UE context drop rate	Variations in throughput during mobility
Hybrid Adaptive Repeat and reQuest (HARQ) failures	Number of HARQ retransmissions
RLC block acknowledges ratio	Intra-/Inter-RAT handover success/failure rate
off-peak/peak traffic times	Handover success rate
Soft handover overhead,	Number of RLFs
Accepted new calls ratio	UL inter-cell interference
Jitter	Latency
Dropped packets	Cell throughput
Call throughput	Session throughput
Call Setup Success Rate (CSSR)	Call Completion Success Rate (CCSR)
Call set-up time	Speech quality (Mean Opinion Score (MOS) [70], Perceptual Evaluation of Speech Quality (PESQ) [71])

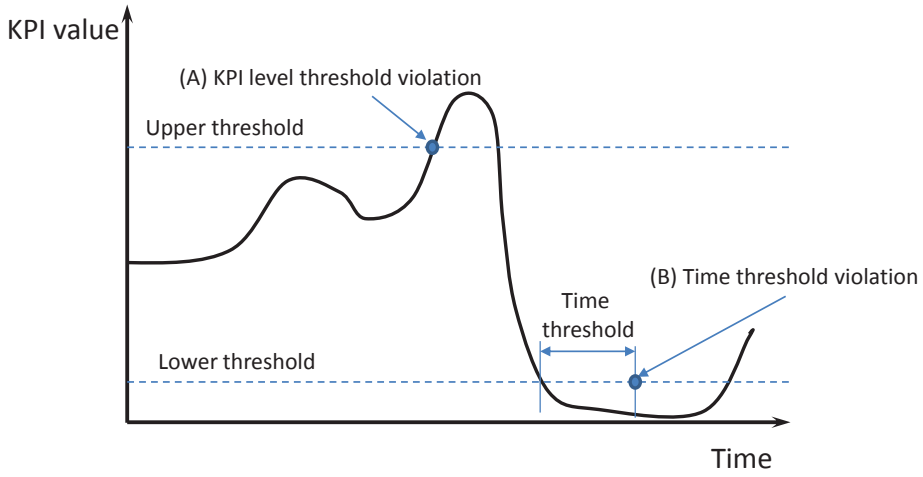


FIGURE 8 Key Performance Indicator (KPI) thresholds.

for enhancement of PM data analysis with advanced methods, which can do processing of multidimensional arrays of PI and KPI values.

Thus, with KPIs it is possible to grasp an idea about the overall network performance, but effective control of the networks operation is a tedious, expensive and hardly achievable task. Nevertheless, the traditional method of KPI analysis is widely used, as it has been developed side by side with the cellular network technologies themselves. Therefore, it is necessary to describe how identification and diagnosis of malfunctions is done on the basis of the observed values KPIs in conventional PM systems. The two main methods of KPIs analysis are thresholds and statistical profiles, discussed in the next section.

2.4.2 Thresholds and Profiles

The most intuitive approach to find a problem in the networks operation is to monitor a set of KPIs and compare their values against a set of predefined thresholds. If the new values are out of range, then an element, or another aggregation level which reported these KPI values has malfunctioned. There are several different types of thresholds. As is shown in Figure 8 [20], a threshold can be set to either an upper or lower KPI value, or both. In that case point "A" would be a violation of a threshold set for KPI level. Another threshold type is when KPI level is combined with the time scale. Thus, KPI values went under the lower bound, but an alarm regarding threshold violation is triggered only at point "B". A logical question here is how to derive the threshold values which would result in accurate detection of malfunction situations and would not lead to a high false alarm rate? This is probably the most complicated task, due to high dynamism, complexity and heterogeneity of mobile networks. Threshold values can become obsolete because of the changes in the network structure, or may be unusable during different periods of time, like seasons, or day and night times. However, in an attempt to come up with some estimate of threshold values, a combination

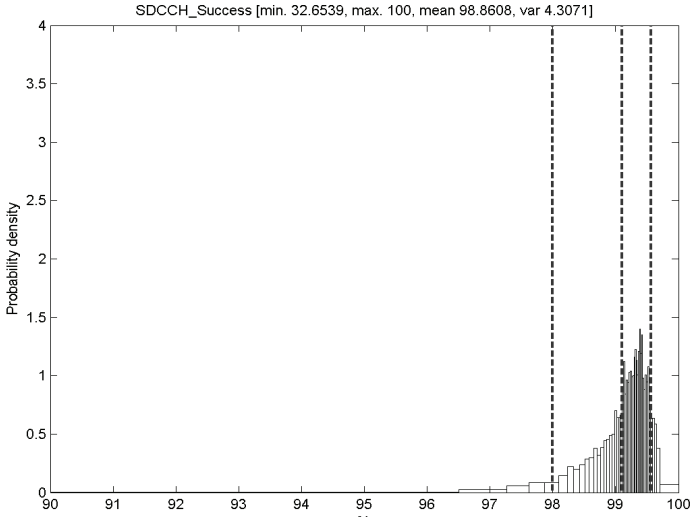


FIGURE 9 Selection of thresholds on the basis of statistical profile of SDCCH success rate.

of *a priori* knowledge and statistical distributions of KPI is used. An example of this approach for SDCCH success rate KPI is shown in Figure 9 [3]. In this figure thresholds are derived for good, normal, bad and unacceptable KPI behaviour, when looking from left to right. The major principle here is to gather KPI values during normal operation state and create a distribution or probability density function [72]. The next step is the comparison of the new KPI sample against the created profiles and a decision whether the system is in its normal state or if there is a problem and some corrective recovery actions should be made. The biggest issue here is that KPIs are estimated separately from each other, and that the creation of profiles and threshold derivation requires substantial time, in the order of days or most likely weeks. The last aspect is especially crucial when network behavior changes. In such situations traditional PM data analysis systems are very ineffective and error prone.

2.5 Self-organizing networks for Quality and Performance Management

In order to overcome the discussed drawbacks of the conventional QPM systems, the concept of SON was developed. The overall idea of SON is to introduce automation into operational network management. The concept of autonomous management, or autonomic computing has already been introduced in the early 2000-s and it implies that a computer system independently controls its operation through sensing its current state and updates the configuration if necessary [73, 74]. Human operators do not have to manually control and update a networks configuration. Instead, high-level objectives have to be provided to the

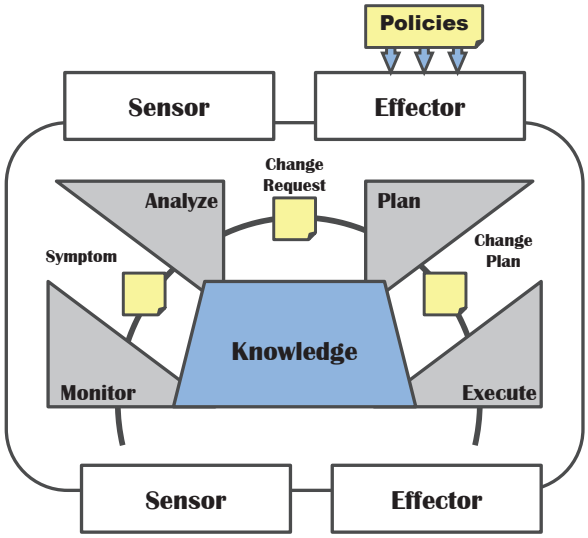


FIGURE 10 Structure of the autonomic element.

autonomic manager. Architecture of an autonomic element is shown in Figure 10 [74]. The flow of actions inside this element forms a closed loop, starting with monitoring the current performance through a sensor, analysis of what is observed, plan changes if necessary, and execute the plan to improve the network operation. In the best case, collection of knowledge is done at all stages of the management process, however this part is probably one of the most complicated and requires consideration of a concrete system. In this early paradigm, the concept of autonomy implied that each element of the network would possess an autonomic structure. This goes with the fact that system-wide self-organization can be reached through local individual interactions [75], just like in a flock of birds or shoals of fish.

In relation to automation in mobile networks NGMN introduced a concept of self-organizing networks, by releasing a set of requirements [5] and use cases [6] in 2008. However, a self-organization concept, adapted form of the concept of automation, described earlier, does not narrow down to a fully distributed approach. In cellular mobile networks there are three architectural options of self-organization [76, 77]:

- Centralized - where execution of SON algorithms is done in the central management entity, such as e.g. OAM in LTE. This option suggests execution of SON algorithms either in the network manager, or on an element manager level - slightly decentralized sub-option.
- Distributed - SON algorithms are run locally in corresponding network elements, for instance in eNBs.
- Hybrid - when a combination of the centralized and distributed SON management is used. This approach is probably the most flexible, versatile, but more complicated than the other two.

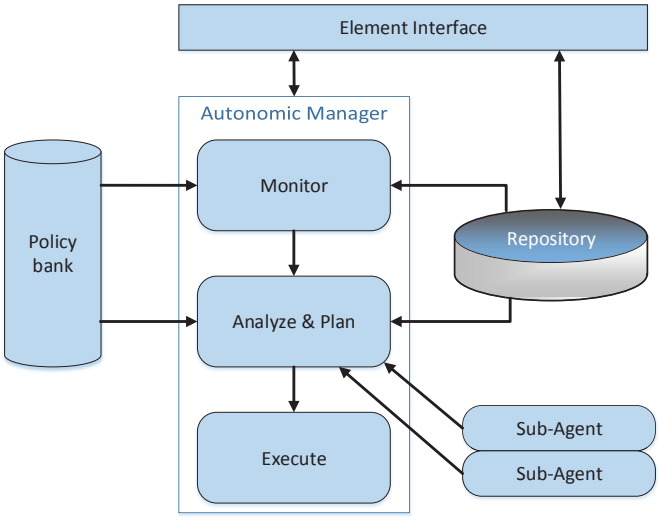


FIGURE 11 Autonomic manager for mobile networks.

Mobile network operators are willing to retain control over self-management functions and for that reason centralized and hybrid options are preferred. Network managers can also look differently if compared to the one suggested for autonomic computing. A mobile network manager is presented in Figure 11 [78]. As it can be seen, operators' policies are stored in a policy bank, and they are used to govern monitoring and analysis functions. Network measurements and parameter configurations are collected to the repository, which can be similar to TRACE TCE, discussed in Section 2.3.1 or an MDT server, as it is further presented in Section 2.5.3. Sub-agents called by *Analyze&Plan* entity are self-organizing functions.

In general 3GPP defines three main SON functions: self-configuration, self-optimization and self-healing [7]. Moreover, research area of coordination between different healing or optimization actions is rather popular, but is not standardized. Coordination is needed to prevent networks from constant configuration adjustments due to conflicting goals of different SON functions, e.g. if they change the value of the same parameter in different directions, as is discussed in [79, 4, 80, 81, 82].

Self-configuration is designed for automatic involvement of a newly deployed cell into operation within an existing network, and a selection of appropriate radio network parameters [10, 83]. In other words, this functionality is responsible for plug-and-play of new elements in the network, such as, e.g. base stations and relays, creation of initial configuration parameter set, and self test. Among the configuration functions, the most important are physical cell ID allocation and automatic neighbor relations [58]. The necessity to carefully and thoughtfully allocate PCI is caused by the limited number of unique identifiers (only 504). If neighboring cells would have the same Physical Cell Identitys (PCIs), there will be collisions and normal network operation would be jeopardized. Automatic Neighbor Relations (ANR) function is very labor intensive and is related to main-

tenance of neighbor tables of each cell, used for handover procedures. Creation of neighbor tables in the new networks and update of these tables upon installation of new base stations is a tedious task if carried out manually. In general, self-configuration covers actions taken during network roll-out or initial cell deployment, while QPM process is more related to monitoring of an existing and operational network. Because of that our main scope is on the other two SON functions: self-optimization and self-healing discussed in the next sections.

2.5.1 Self-optimization

Self-optimization is applied in several areas of network operation [10]: mobility robustness - handover optimization; mobility load balancing - RACH optimization; coverage and capacity optimization - inter-cell interference coordination, P_0 (UL power control parameter [84]) optimization; energy saving [2]. This section discusses some examples of SON functions, which illustrate the core idea of self-optimization.

The main goals of Mobility Robustness Optimization (MRO) is to minimize the number of call drops and RLFs, diminish the amount of unnecessary handovers, and make idle mode mobility more reliable. There are different possible problems related to mobility, such as *too early handover*, *too late handover*, *handover to wrong cell*. All of these three types of problems lead to an increase in RLF rate. The root causes of these problems can be inconsistent mobility parameters, interference, lack of coverage or coverage islands. Additionally, there are unnecessary handovers, which can be divided into several groups: ping-pongs, i.e. a series of handovers between two neighboring cells within a very short time interval. Rapid handovers, also known as short stays, are similar to ping-pongs, but cell changes happen on the way from one cell to another, with a short stay in some intermediate cell. Thus, there is one unnecessary handover. One more type of mobility problem is the handover to a lower priority layer, e.g. another frequency or RAT, and the last type of unnecessary handover is when it happens right after the connection has been established. Parameters used for MRO purposes are L1 averaging, L3 filter coefficient, Time to Trigger (TTT) and handover offset [85]. However, in order to apply configuration changes it is necessary to know exactly which cells should be adjusted. This calls for root cause identification, and in MRO specified by 3GPP, the main tools for this purpose are information on the re-establishment request procedure, available since Release 8, and RLF reports, RLF indication to a problematic cell and handover reports – all Release 9 features. Moreover, RLF reporting extensions in Release 10, and MDT reporting (discussed further in Section 2.5.3) can enrich the amount of information for MRO purposes.

Coverage and capacity optimization is another SON functionality aimed at making mobile networks more flexible and responsive to the fluctuations in the operational conditions. For instance, signal propagation and network load can be affected by daily traffic variance, seasonal changes, and mistakes in radio network planning due to construction or demolition of major objects. In order to automatically control uninterrupted coverage and appropriate capacity level,

several technological enablers, such as adaptive antennas, were needed. Remote electrical tilt antennas, variable electrical tilt and active antenna systems can be used to compensate coverage or capacity issues. Adjustment of antenna azimuth, half-power beamwidth and antenna tilt, are parameters of adaptive antenna systems and their careful adjustment can give a different extent of positive impact on the resulting coverage and capacity. Performance indicators which can be used to judge efficiency of optimization solutions are channel conditions, the number of mobility problems and cell throughput [86]. Another way to adjust the coverage and capacity is reconfiguration of Base Station (BS) transmission power. Coverage and capacity optimization does not consider how the detection of problematic regions is done. Instead, such SON functions as MRO and further discussed MDT are responsible for that.

2.5.2 Self-healing

The main aim of self-healing is to maintain high levels of network service quality through automatic performance control and adjustment of appropriate configuration parameters. This SON function consists of fault detection, diagnosis, recovery action planning/selection and execution [9]. There are several different use cases in self-healing: self-recovery of NE software, self-healing of board faults and cell outage conditions. The latter use case, similarly to coverage and capacity optimization functionality, is related to handling of coverage problems. However, the scope of a self-healing case is broader, as it includes Cell Outage Detection (COD), Cell Outage Compensation (COC), recovery and return from compensation state. Self-healing of cell outage has been prioritized both by 3GPP and the research community, e.g. in SOCRATES and COgnitive network ManageMent under UNcErtainty (COMMUNE) projects. Thus, this case is discussed further in this section as well.

The task of cell outage detection is very important, as it initiates a series of recovery and compensation actions. When detection is inaccurate, a lot of unnecessary expenses are caused. There are various ways to carry out COD, such as alarm correlation, analysis of KPIs with methods discussed in Section 2.4, and more advanced approaches presented in Chapter 4. In order to find out the interrelation between alarms and their actual reasons, an alarm correlation method is employed. This term refers to automatic identification of root causes behind triggered alarms. There are several research activities devoted to the analysis of alarm data [27, 25], even using data mining methods such as, e.g. neural networks [26]. Thus, the ultimate goal of alarm correlation is to reduce the effort needed to manually evaluate alarms, and by that reduce operational costs [2]. However, this approach relies on the alarm base only, and does not take into consideration KPIs, network measurements, or the history of the networks operation, e.g. mobility events. Thus, alarm correlation should be extended with other types of network data and analysis methods, to make PM and fault management more efficient. These goals match the motivation of the described studies and the scope of this thesis.

In COC uses a case where a problematic region with a malfunctioning cell is covered with neighboring cells through the adjustment of appropriate parameters. Several research activities in this area have been carried out [12]. It is demonstrated that using the antennas down tilt, power of reference signal and P_0 parameter [84] configuration, it is possible to achieve fairly good coverage compensation. At the same time, efficiency of COC varies depending on the operators' priorities in the radio network planning. Coverage, capacity or traffic load can be selected as the targets, as it is discussed more elaborately in [87, 88].

The next step of cell outage self-healing is recovery, i.e. restoration of the fully-functional operational state of the network. In many cases it is achieved through base station reset or roll-back to a previous eNB firmware version. If the problem has been caused, e.g. by mistakes of cell configuration, or if the old parameter set has become outdated, a new set of parameters is proposed by a recovery planning entity and conveyed by recovery execution. In the worst case, a site visit of maintenance personnel is needed. As a next step, a return from compensation state is done through reconfiguration, taking into account the latest network state.

2.5.3 Minimization of Drive Testing

In existing networks the majority of KPIs represents performance and quality of service from the networks point of view, not from user perspective. Moreover, in many cases aggregation of PIs is done per cell, or per group of cells over a considerably long period of time, e.g. an hour or a day. This kind of generalized representation does not allow the analysis of the performance and QoS in a detailed manner, as a lot of information is left out. This contradicts the desire of the operators to improve efficiency of QPM process and as a result, to identify network malfunctions in a fast and accurate manner. Among other data sources in the network (see Section 2.3.1), only drive testing approach provides performance measurement information from a users perspective. However, drive testing has several critical disadvantages. It is expensive to carry out regular and detailed tests. Extensive verification of network performance implies both outdoor and indoor measurement campaigns. For common areas with open access, like streets, it is relatively easy to do drive testing, while inside buildings it becomes nearly impossible. Due to these reasons, coverage and QoS maps of the network are incomplete, therefore operators have only a limited view regarding the networks performance. In order to overcome this problem, and automate performance data collection, a study called Minimization of Drive Tests (MDT) has been added to the family of 3GPP SON functions [13].

The core idea of MDT is to collect user-specific measurements and PIs, and if possible, correlate them with a particular geographical location. The resulting database would identify and tackle various network problems. According to 3GPP specifications, MDT can be used to optimize coverage, e.g. by enabling the creation of coverage maps, detection of areas where coverage is weak or unacceptably low, i.e. coverage holes. Moreover, by measuring and reporting SINR

and power headroom, MDT-capable users can help with the identification of excessive interference, overshoot coverage optimization of uplink coverage. Other cases when MDT information can be applied, include optimization of user mobility configuration, network capacity, parameterization of common channels and verification of end-user quality [13].

MDT architecture is based on control plane, i.e. coordination of MDT campaigns, initiated by OAM is done through Radio Access Network (RAN) nodes - RNCs or eNBs [89]. There are two options how MDT can be carried out: signaling-based, aimed at tracing a specific user, and management-based, which is targeted for data collection campaigns in a particular geographical area with multiple users [2]. Activation/deactivation and management procedures for both of these methods rely on TRACE functionality, discussed in Section 2.3.1 and in full detail in [41]. Reporting of MDT data can also be done in several different manners [90]. In connected mode it is called *"immediate"*, as a user terminal does not store any data in the memory, but instead sends a corresponding measurement right away. MDT reporting can be done with pre-configured periodicity or triggered by a certain event, such as e.g. A2 (see Section 2.3.1). In idle user state, terminal makes *"logged"* MDT – logging of the appropriate measurements along with other related information. Upon establishment of a new connection, it informs the eNB regarding the availability of the MDT log, and sends it if requested. There are two sub-types of reporting modes, which are both designed for detection of network failures. One is related to immediate MDT, and it is called RLF reporting [91]. The idea is to create a separate report in case connection failure occurs. There are two types of failures which can trigger such a report: RLF [58] (added to 3GPP specifications in Release 9) and Handover Failure (HOF) [92] (added in Release 10). Another type of MDT report, standardized by 3GPP in Release 11 is RRC Connection Establishment Failure (RCEF). This report is aimed at the detection of problems with network accessibility, i.e. if UE unsuccessfully attempts to establish a data connection, this event would be logged along with some other useful information [14]. The reporting mode is related to logged MDT.

Data collected by MDT functionality include time stamp (either provided by UE or eNB), cell identification information, signal strength measurements from serving and neighboring cells and best-effort location information. Signal strength measurements in UTRAN are RSCP and E_c/N_0, and in E-UTRAN they are RSRP and RSRQ. Location identification methods vary a lot in their accuracy, applicability, impact on the UE, eNB core network system, response time, and availability in particular 3GPP release. A perfect positioning method would have the highest accuracy, and minimum negative impact on all involved entities, e.g. in terms of battery consumption. Additionally, the response time for location estimation should be as short as possible. In reality more accurate methods, such as positioning with Assisted-Global Navigation Satellite System (A-GNSS), have long response time and negatively impact the user [93]. In Release 8, a method called CID (Cell ID) is present, which is basic, the least complicated and the most inaccurate. In Release 9, several new methods have been added: Enhanced Cell ID (E-CID), Observed Time Difference of Arrival (OTDOA), Angle of

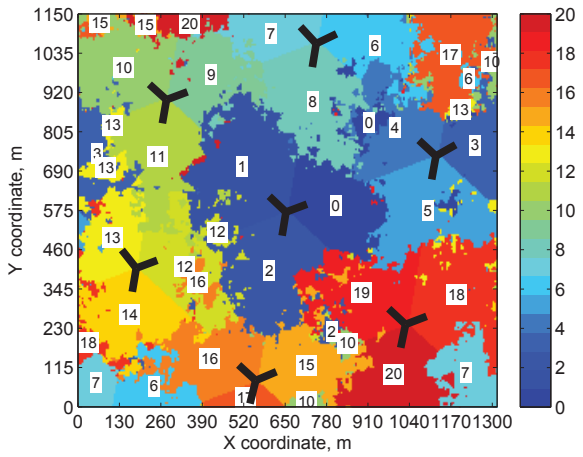


FIGURE 12 Network dominance map.

Arrival (AOA), A-GNSS, mentioned above Radio Frequency-fingerprinting and Adaptive Enhanced Cell Identity (AECID). In Release 11, Uplink Time Difference of Arrival (UTDOA) has been specified. Additionally, standards include a hybrid positioning method, which combines information from several algorithms of location estimation.

Location information can help to create so called dominance maps. Such maps show which particular eNB has the strongest pilot signal in each point of the network, and as a result it would most likely be a serving cell in this area. An example of a dominance map is shown in Figure 12, where cell IDs is marked with both color scale and labels on a network map. Dominance maps are collected with drive testing and MDT reporting, using the best available radio resource location methods. Mapping of a particular cell ID to a location coordinate gives the possibility for more accurate detection of problematic regions and identification of malfunctioning cells.

In the research presented in this dissertation, *management-based MDT* with *immediate and RLF reporting* functions are used for data collection. Mostly *event-triggered* MDT data collection is used, but in [PV] and [PVI], a combination of both event-based and *periodic* MDT data gathering approaches are employed. With respect to location information, serving and target cell IDs together with dominance maps are utilized.

2.6 Summary

This chapter is devoted to the overview of quality and performance management in mobile networks. First the QPM cycle is outlined and classification of different generations of QPM is presented. Then different types of network failures are presented. The next step is discussion regarding the sources of the performance information in the cellular networks, such as:

- Drive tests
- Network interface tracing
- Probing systems
- Signaling element counters
- Transport network statistics
- Trace functionality
- Mobile quality agents
- Minimization of drive tests

The types of the collected information include alarms, counters and measurements. An elaborated overview of the traditional methods of performance analysis is given, which outlines KPIs, statistical profiling and threshold selection. In a nutshell, the main disadvantages of traditional data analysis methods in PM are:

- Routine tasks involve large amount of "low-level" manual work. It means that engineers have to separately monitor a lot of performance indicators.
- Usage of discrete thresholds for performance monitoring. It is hard to derive those thresholds and handling instantaneous violations of the thresholds is not straightforward.
- Hard to maintain thresholds up to date. Mistakes in threshold selection increase false alarm rate.
- Profiles require normal dataset collected over sufficiently long periods of time. Also datasets with no erroneous behavior might be required.
- Fault detection can take a long time.
- Mostly severe problems are noticed, partial degradations are most likely missed.
- Mostly univariate, not multivariate [94].

The concluding part of this chapter is about automation of network functions with self-organizing networks. Self-optimization and self-healing are discussed in some details. Large attention is paid to MDT functionality and partly location estimation methods.

3 KNOWLEDGE MINING

Knowledge Mining (KM) is a generic term, used to represent the sequence of actions for discovery of useful information from databases. The necessity for knowledge mining is caused by the existence of large multidimensional data sets and a constant increase in the number of data sources. With the internet of things [95, 96] the potential amount of data available for analysis is nearly infinite. For instance, Cisco forecasts 24.3 Exabytes (10^{18} bytes or 1 million of terabytes) per month of mobile data traffic by 2019 [97]. According to different estimates by 2020 there will be around 10 billion mobile connections in 2^{nd} Generation (2G), 3^{rd} Generation (3G) and 4^{th} Generation (4G) networks, with around 4.6 billion people with LTE subscriptions [98]. The main purpose of knowledge mining is to elicit the important data, find meaningful patterns inside it, and convert these patterns to knowledge for improvement of the analyzed system or process. Terms data mining and knowledge mining are frequently treated as synonyms. However, many authors consider data mining as one of the steps inside the knowledge mining process [99, 16, 100], shown in Figure 13. In this diagram, the source of the data is a cellular mobile network with base stations, user terminals, etc. However, any process or system which produces data can be a source. At the *step 1* the raw data is gathered into a data base. At the *step 2*, the target data, called *data set*, is selected. Only those data samples which are relevant for future processing are taken. Rows of the data set are called samples, data instances, or valued. For instance, if the user CQI and RSRP measurements are collected to a dataset, each user measurement report will represent one sample. Columns of the data are called features, attributes, variables. In the above example, features are user ID, CQI and RSRP. Thus, a 3 dimensional dataset is constructed with these features. The *step 3* of knowledge mining is *data mining* which includes 3 internal sub-steps: pre-processing, transformation and pattern recognition. Data mining is a broad term, and naturally there are many definitions, but here is one of the most popular [99]:

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful.*

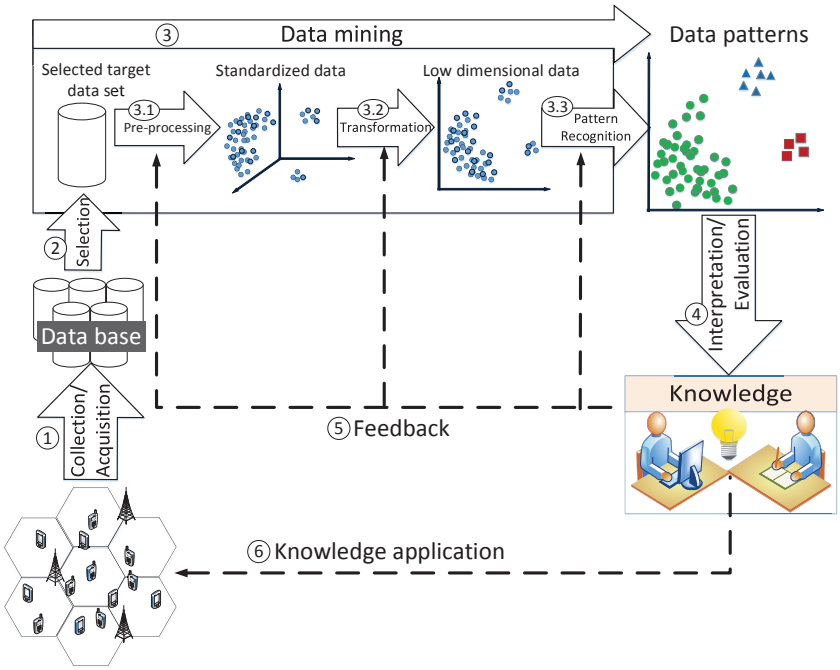


FIGURE 13 Knowledge mining process.

"Relationships" and "summaries" from this definition are frequently called patterns or models. Thus, data mining can be seen as a set of all processing actions needed to convert the selected target data into meaningful patterns. The initial step of data mining (*step 3.1*) is *pre-processing* which is responsible for cleaning, filtering, handling of missing data [101, 102], normalization, scaling and standardization of the input data [16]. After that, at *step 3.2* transformation of the pre-processed data is made. Usually the reason to carry out data transformation is to reduce the number of dimensions in the analyzed data, for reduction of computational complexity and representation of the data in a more beneficial and meaningful scope, e.g. space where interrelations are more obvious. *Step 3.3* - pattern recognition, is the heart of data mining, as at this stage various learning algorithms of e.g. clustering, classification, regression, or decision trees are applied [94]. Information about identified patterns should be then converted into knowledge, i.e. from the presentation in terms of data mining to the actual application domain. In many cases interpretation (*step 4*) is done with the help of visual tools, such as distributions, bar plots, pie charts, graphs, etc. This allows for knowledge creation which is further used to improve the operation of the analyzed system. Another way to apply knowledge is to adjust the processing steps of the data mining itself. In Figure 13 this is *step 5* and is called "Feedback".

In the following sections of this chapter we discuss different parts of data mining according to the process shown in Figure 13. Data mining itself and its sub-class – anomaly detection are defined. Particular algorithms and different data types are presented. Outline of evaluation metrics for pattern recognition al-

gorithms is made. It is also important to emphasize that the goal is not to describe all available methods and algorithms, instead we concentrate on the techniques applied in our research.

3.1 Data mining and anomaly detection

As it is discussed above, the main goal of data mining is to find meaningful patterns in the data. This is done with consequent application of pre-processing, transformation and pattern recognition algorithms. However, in case the goal of the analysis is to find *suspicious* data patterns, this process is called *anomaly detection* [17]. Thus, anomaly detection can be treated as a special case of data mining. Patterns which do not conform with normal behavior in the data are called anomalies or outliers [17]. In the context of anomaly detection these two terms are used most commonly and sometimes interchangeably. Other synonyms are discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants depending on the application domain. In many cases the mass of normal samples is much larger than the number of anomalies. When such disproportion is encountered, the data is called "imbalanced" and additional considerations should be made for successful anomaly detection [23]. The key aspect of any anomaly detection technique is the nature of the input data, which defines what kind of pre-processing and transformation methods (if any at all) should be applied. Origin and the type of data also impacts on the selection of pattern recognition and representation algorithms. The next section introduces the most common data types in the context of data mining and presents how performance monitoring indicators from cellular mobile networks correspond to these types.

3.1.1 Data Types in Anomaly Detection

Hereby we present a classification of data types from the perspective of anomaly detection and data mining.

Numerical data - can be discrete or continuous, but for this data it is possible to calculate the distance between points, apply similarity measures and use statistical analysis methods. The most common example in cellular networks are measurements, such as pilot signal strength, SINR, throughput, call blocking ratio, etc.

Categorical or nominal data is generalization of the binary variable, which can take more than two states. Examples can be the states of some device: ACTIVE, PASSIVE, OFF, or as in weather forecasting: sunny, overcast, and rainy. No relation can be implied among the values of the variable – neither ordering, nor distance measurements are available. This kind of data can also be represented with integer numbers. Such integers are used just for data handling and do not represent any specific ordering. It certainly does not make sense to add the values together, multiply them, or even compare their size. Only equality or inequality

can be checked [103, 104]. Independent data samples, which are called point data, are the most common type of data analyzed in anomaly detection. For instance, reports of measured pilot power strength from different UEs or different cells belong to this type.

Sequence data - instances are linearly ordered, for example, time-series data. Naturally *temporal data* also belongs to the sequence data. A good example of sequence data is a sequence of ACK-NACK messages in TCP transmissions. In case there is a sequential relationship in the categorical data such similarity measures as Hamming distance can also be applied [105].

Spatial data - each data instance is related to its neighboring instances, for example, relations between adjacent cells in the network or vehicular traffic data.

Textual data - is usually contained in a document or collection of documents, which contain for instance records or plain text. It is one of the most complicated types of data for analysis, as it is unstructured and hence, cannot be analyzed with mathematical methods without pre-processing. There is a whole separate area of text mining [106, 107], but it is out of the scope of this dissertation.

It is very important to know the type of data which should be handled by a performance monitoring system, as it defines the choice of algorithms selected for anomaly detection, both at pre-processing, transformation and pattern recognition steps.

3.1.2 Data Types in Mobile Networks

In Section 2.3 we discussed different kinds of performance monitoring statistics and their possible origins. We also gave some examples of each data type from the world of communication technologies. However, here we make a more thorough typification for the most common kinds of mobile network monitoring data. Table 3 presents correspondence of different QPM statistics to particular data types.

3.2 Data pre-processing

Collected network monitoring data should be prepared for further analysis with data mining anomaly detection algorithms. This pre-processing should be done so that the noise is filtered out, missing values are handled and similar value ranges are derived. Also construction of new features and data standardization, normalization is done at this stage. Pre-processing is a very important step of data mining and anomaly detection. For instance, if input data is not properly prepared, the output result of anomaly detection cannot be good (so called "garbage in, garbage out" rule of thumb). On the other hand it is not beneficial to alter the dataset too much before the main analysis (e.g. pattern detection) as meaningful data can be lost, making anomaly detection less accurate.

For data filtering it is assumed that all unnecessary, irrelevant data fields, such as particular samples or features are excluded from the target analysis data

TABLE 3 Types of performance monitoring variables from perspective of data mining.

Monitoring data	Data type	Description/Example
Measurement	Numerical (Continuous)	Pilot signal strength (e.g. RSRP), SINR, user throughput, cell throughput, call blocking ratio, etc.
Counter	Numerical (Discrete)	Number of HOFs, RLFs, etc.
Alarm type	Categorical	Board fault, Power cabling fault, Amplifier fault
MDT event	Categorical (nominal), Sequential	HOF report (e.g. message type 1), RLF report (e.g. message type 2)
UE ID	Categorical	UE1, UE2, etc.
Serving or target cell ID	Categorical, spatial	Cell Radio Network Temporary Identifier (C-RNTI), PCI
UE location coordinates	Numerical, spatial	Geo-positioning coordinates: latitude, longitude, altitude
Coordinates of signal strength measurement in the network	Numerical, spatial	Measurement location attached to MDT report

set. Filtering is also one of the ways for handling missing data, but depending on the nature of the data, a large portion of meaningful information could be lost. Scaling, standardization and normalization implies bringing data variables to the same scale. This is done to avoid the masking of information in low-scale variables, by other large-scale attributes. For instance *z-score* method converts a variable so that it has zero-mean, unit variance [108]. As is shown in (1), new value of i^{th} element of the variable $X = \{x_1, x_2, \dots x_n\}$, is standardized by subtracting μ - the mean of X, and dividing the result by σ - standard deviation.

$$\tilde{x_i} = \frac{x_i - \mu}{\sigma} \tag{1}$$

Other popular normalization methods are max-min or range, max, mean, sum and log, which are selected depending on the nature of the data and the needs of further processing [16, 108].

Construction of new attributes is another area of data pre-processing [109]. For instance, when analyzing data about apartments in the real estate business we have two variables: overall size of the apartment and size of non-living area, e.g. a balcony. An apartment might both have large overall and non-living areas. This would mean that the effective living area ratio between overall size of the apartment and non-living area would represent the "quality" of this apartment as high. But if the overall area is relatively small, while the balcony is huge, relative "quality" of such an apartment can be treated as low. That is why using a new

attribute - "effective living area ratio" can reveal more information regarding the relations within the analyzed data, and can be used to extend the target data set.

In a similar way a density measurement can be used to enrich the data, as it represents the extent of isolation of a particular point from the rest of the data [33, 110]. In order to calculate the density of a data point it is necessary to first define the radius of the proximity region, which is used to construct a n-dimensional sphere around the analyzed data point. Naturally, this radius is normalized according to the scale of each dimension. Value η_m corresponds to the number of points enclosed by the sphere around m^{th} point. By normalizing this value with the sum of all $\eta_i \geq 0,\ i = \overline{1, M}$, where M is the total number of points in the manifold we get the resulting density of m^{th} point.

$$d_m = \frac{\eta_m}{\sum_{i=1}^{M} |\eta_i|}, \tag{2}$$

Density is a non-negative value and it never goes over 1, i.e. has a property: $0 \leq d_m \leq 1$. Thus, for points which are located in a "crowded" region, the density value will be high, while for sparse regions there will be very small density values. Depending on the application area and the nature of the original data either high or low density can be considered as a sign of abnormal behavior. In this thesis, the density measure has been applied in the embedded space constructed by means of transformation and dimensionality reduction with diffusion maps, discussed further in Section 3.3.1.

3.3 Transformation techniques

Transformation of the data implies conversion to a new space, where representation is more beneficial and meaningful. There is a large number of data transformation methods [111, 112], in this thesis, the main scope is on the sliding window technique and N-gram analysis. These algorithms are not related to dimensionality reduction, discussed further in Section 3.3.1.

The Sliding window approach is applied when it is necessary to increase the size of the data set, preserving the interrelations within the data. This is especially beneficial if we have only one variable and a lot of samples, or the other way round, when the number of variables is large, but there are a few samples. Thus, sliding window can be applied to both numerical, categorical and sequential data [113, 114, 115]. Consider that there is a relatively long sequence of events. Sliding window transformation is done so that we take every m samples with step n. Here m is the sliding window size and n is sliding window step. An illustration of the application of this method is show in Figure 14. In case the window size is larger than the step - it overlaps the sliding window, and if the step is the same or larger than the window size - it is non-overlapping. Yet another strong side of the sliding window method is that if entries of different samples in the original data set have a non-equal length, after processing all samples have the same length equal to the window size m.

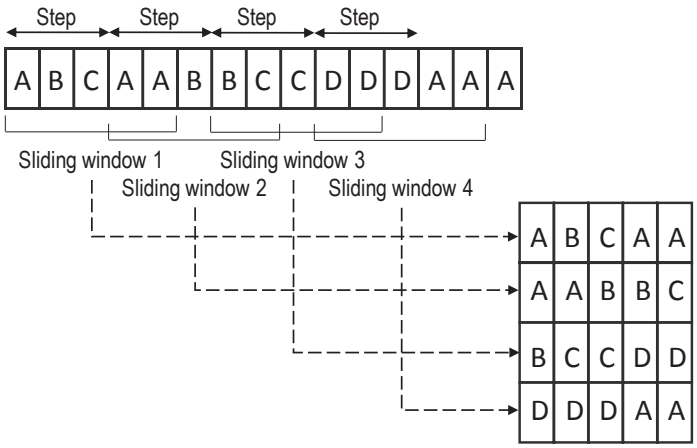


FIGURE 14 Sliding window transformation.

TABLE 4 Example of N-gram analysis per character, $N = 2$.

Analyzed word	pe	er	rf	fo	or	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

N-gram analysis is widely applied in the area of sequential data processing. For instance, this method has been utilized for natural language processing and text analysis applications such as speech recognition, parsing, language recognition and spelling [116, 117, 118, 119, 120]. In addition, N-gram has been applied to the analysis of the whole-genome protein sequences [121] and for computer virus and zero-day attack detection [122, 123]. 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. In general, the recipe is to define a unit in the data, and then take all possible sequences of N units. Therefore, N-gram provides a new way of representation of data by decomposing the original data into small pieces. This new representation is stored in a feature vector. The feature vector contains values or frequencies of how often each particular N-gram sub-sequence occurred in the original data.

Here is an example of N-gram analysis application for two words: 'performance' and 'performer', $N = 2$, and a single unit is considered to be a character, not a word. The resulting frequency matrix after N-gram processing will be as it is shown in Table 4. Thus, the number of possible 2-grams is equal to the number of combinations with repetitions [124]. For instance, consider the English alphabet, there will be $26^2 = 676$ 2-grams, while for simple ASCII characters it will be $128^2 = 16384$. In sequence data mining N-gram analysis is one of the fundamental and at the same time popular methods [125].

3.3.1 Dimensionality reduction

In case of after transformation, the resulting number of data dimensions is lower than in the original data set, it is called dimensionality reduction. The main goals of dimensionality reduction are more of a beneficial representation of the data in the new space, and a decrease of computational complexity for further processing. The latter is related to the "curse of dimensionality".

3.3.1.1 Curse of dimensionality

The necessity to reduce the number of attributes is caused by problems with high computational complexity and usage of common similarity measures encountered when dimensionality of the analyzed data is high. This is referred to as the *curse of dimensionality*, which more formally can be defined as follows: with a linear increase in the number of dimensions, the size of the data needed for an equally accurate representation in the new space grows exponentially [126, 127]. In other words it is an exponential increase of computational complexity with the addition of new dimensions to the data. Let us consider one example which illustrates this phenomenon. Taking as an assumption that a tennis court has only one space dimension, and the ball could take one of 25 positions along the line of this dimension, knowledge about only 25 volley options would have been sufficient for the return hit. But if we consider the movement of the ball on a 2D court, already 625 (25^2) volley options would have been needed (roughly assuming the tennis court is square), and for a 3D imaginary court, a player should have mastered 15625 (25^3) volley options, in order to be able to return the ball. This example shows not only that tennis is a technically sophisticated game, but also that a linear increase of dimensionality requires exponential growth of processing technique complexity. Moreover, some methods such as similarity measures, as e.g. Euclidean distance, or nearest neighbor procedure, become unusable in high dimensional space [128, 129]. One of the accepted ways to overcome the *curse of dimensionality* in data mining is to apply dimentionality reduction methods [16, 130]. There is a large variety of such algorithms which can be logically separated into two main groups: linear and non-linear.

3.3.1.2 Principal and Minor Component Analyses

One of the most popular linear dimensionality reduction algorithms is PCA [131]. The main idea of this method is the creation of a low-dimensional embedding of the original data, preserving as much variance as possible. This is achieved with linear transformation to a set of uncorrelated variables - principal components. The property of the principal components is that in several of them, the most variance of all original variables is contained. The illustration of data transformation with PCA is presented in Figure 15.

The formal and elaborate derivation of components is given in [131]. In a nutshell, eigenvectors corresponding to the highest eigenvalues are referred to as

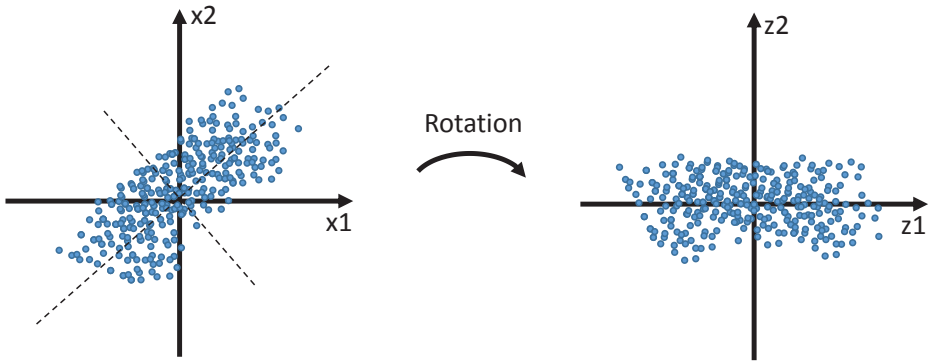


FIGURE 15 Sample data before and after transformation with PCA.

principal components. If X is the initial data matrix given, its covariance matrix is Σ. It is shown [99] that the first principal component is the first eigenvector of the covariance matrix Σ, and it projects the largest variance of the original matrix X. The second principal component corresponds to the second eigenvector and projects the second largest amount of variance. A common way to calculate the basis of orthogonal eigenvectors is to solve Singular Value Decomposition (SVD) for covariance matrix Σ. PCA can be used in the preparation of the data for further analysis, helps to tackle the "curse of dimensionality", or allows one to avoid overfitting, or simplify the interpretation of the data. Among the main advantages of PCA is that no configuration parameters are needed to carry out the processing and that despite its simplicity, PCA demonstrates very good results in a large number of datasets even compared to non-linear methods [130, 132, 133]. However, applicability of PCA is limited by the following factors [134, 135]:

– Inherent linearity, i.e. we assume that the observed dataset can be represented as a linear combination of some basis.
– Directions with the most variance do not necessarily possess the best features for representation.

MCA is another method of dimensionality reduction, aimed at the selection of new representation of the initial data matrix X. This method is largely based on PCA. Similarly, MCA extracts components of the covariance matrix of the initial dataset. However, MCA algorithm utilizes *minor components*, which corresponds to the *smallest* eigenvalues λ of covariance matrix Σ of the original data X. Thus, MCA proposes a new coordinate system where the base vectors project the most stable directions in the original data, i.e. preserve the lowest variances. Same as PCA, MCA basis is orthogonal. The example utilization of both MCA and PCA can be found in [136].

3.3.1.3 Diffusion Maps

In some cases the key aspects of the data are not related to the variance within the data, or even more commonly are non-linear. Then PCA or MCA cannot

be efficiently used for reliable representation of the data in a low dimensional space [137]. Such datasets require more sophisticated - non-linear dimensionality reduction methods. One example also used in our study is the diffusion maps algorithm [138], which demonstrated its efficiency in a wide range of applications such as network security, traffic analysis, machinery stability analysis [139], mobile network performance monitoring and medicine [113, 123, 132, 133, 139, 140, 141, 142].

Diffusion Maps (DM) is a non-linear dimensionality reduction method based on geometric structures in datasets [138]. Derivation of the embedded space is done using a weighted probability graph constructed with random walk [143]. Consider a dataset $X = \{x_1 \ldots, x_n\} \in \mathbb{R}^M$, where n is the number of data points, M is the dimensionality, and μ is a distribution of points on X. The original dataset goes through the following processing steps to be converted into a new low-dimensional embedded space [138, 123, 144].

Step 1. Construct a symmetric complete graph $G = (V, E)$ on X, where vertexes (V) are data points and edges (E) are all possible connections between points.

Step 2. In order to define weights in E, a kernel is selected. For instance, if x and y are the considered vertexes, the corresponding kernel would be:

$$W_\epsilon \triangleq w(x, y) \tag{3}$$

Weight corresponds to the edge between x and y, and defines the affinity between these points. The Kernel function has two properties: it is symmetric: $w(x, y) = w(y, x)$, and positive: $w(x, y) \geqslant 0$. Depending on the task, different weight functions should be selected. For example, Euclidean distance is a good choice for numeric datasets, while Hamming distance is a better choice for categorical data. In the studies presented in papers [PIV] and [PV] we employed the Gaussian kernel with Euclidian distance:

$$W_\epsilon = e^{-\frac{\|x_i - x_j\|^2}{\epsilon}}, \tag{4}$$

where $\epsilon > 0$ is a metaparameter of the DM algorithm discussed further in more details.

Step 3. By means of a random walk on X a Markov probability matrix P is constructed. This is done with normalized graph Laplacian, presented in (5), [145].

$$p(x, y) = \frac{w(x, y)}{d(x)}, \tag{5}$$

where $d(x)$ is the degree of node x:

$$d(x) = \int_X w(x, y) d\mu(y), \tag{6}$$

and μ is a distribution of points on X.

Here $p(x,y)$ is a probability of random walk, starting from point x to come to point y in one step, as far as $p(x) \geqslant 0$ and $\int_X p(x,y)d(\mu) = 1$. Therefore, the Markov matrix of transition probabilities P for one transition step is obtained. This matrix is also known as the stochastic matrix [146]. According to the properties of the Markov matrices, the probability of transition between any two points of the dataset in t time steps is matrix P^t. It is shown in [147] that if the graph is connected, then for $t = +\infty$ this Markov chain has a unique stationary distribution ϕ_0:

$$\phi_0(y) = \frac{d(y)}{\sum_{l \in X} d(z)}. \tag{7}$$

Step 4. After construction of random walk on the data graph, eigenvalues $\{\lambda_l\}$ and corresponding right and left eigenvectors ψ_l and ϕ_l of Markov matrix P_t are computed. This is also called eigen decomposition.

$$p_t(x,y) = \sum_{l \geqslant 0} \lambda_l^t \psi_l(x) \phi_l(y), \tag{8}$$

Usually this is done by means of SVD algorithm. One valuable property of the eigenvectors is that they are orthogonal to each other. Thus, the first m largest eigenvalues and corresponding eigenvectors are used to form the basis of the new Euclidean space:

$$\Psi_t(x) \triangleq \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \cdots \\ \lambda_m^t \psi_m(x) \end{pmatrix} \tag{9}$$

Step 5. Hence, to represent the original data points in the new space, using the derived basis, it is necessary to find a connection between the spectral properties of the Markov matrix and geometry of the dataset X. For that purpose the notion of *diffusion distance* is introduced.

$$D_t^2(x,z) = \sum_{y \in X} \frac{(p(y,t|x) - p(y,t|z))^2}{\phi_0}, \tag{10}$$

where $D_t(x,z)$ is a diffusion distance between points x and z at time t. Expression $p(y,t|x)$ is the conditional probability of random walk started in x would come to y after t time steps. In [138] it is shown that the diffusion distances can be computed using eigen decomposition of stochastic matrix P.

$$D_t^2(x,z) = \sum_{l \geqslant 1} \lambda_l^{2t} (\psi_l(x) - \psi_l(z))^2. \tag{11}$$

Thus, eigenvectors can be used for the calculation of diffusion distances. This fact is utilized for the location of data points in the embedded space, i.e. defines applicability of diffusion maps for dimensionality reduction [138]:

The diffusion map $\Psi_t : X \to Y$*, embeds the data from space* ^{M} *into the Euclidean space* $Y = \mathbb{R}^m$ *in which the distance is equal to the diffusion distance with relative accuracy* $O(t, m)$*.*

$$\|\Psi_t(x) - \Psi_t(y)\| = D_t(x, y) + O(t, m), \tag{12}$$

One of the most important properties of diffusion distance is that it defines connectivity in the high-dimensional data space. Practically it means that, $D_t(x, y)$ is small in case if the large number of short edges connects x and y, what corresponds to a large transition probability between these two data points in both directions. For that reason, if the distances of the group of points in the same region of the manifold are small, this area is likely to be a cluster. Another valuable property of diffusion distance $D_t(x, y)$, is that it is very noise-robust, due to summation over all the paths of the length t, connecting x and y.

Parameter ϵ, is used to define kernel function in (4), is used to achieved a more accurate representation of the data in the embedded space by means of a smaller number of dimensions. More precisely, proper selection of parameter ϵ helps to achieve fast decay of eigenvalues $\{\lambda_l\}$ from (8). Discussion about methods for selection of the optimal value of parameter ϵ can be found in [139, 148].

3.4 Pattern recognition approaches

In this section a classification of learning algorithms used for pattern recognition is discussed. Additionally, some particular examples of learning methods is given with a brief overview of the principles of their operation.

3.4.1 Types of learning in pattern recognition

Pre-processing and transformation, described in the previous sections, are aimed at the preparation of the data for pattern recognition. Before this step data is a collection of points with potentially existing inlaid structures. Pattern recognition, which can be called the heart of data mining, is supposed to identify these structures. In order to achieve this, it is necessary to learn the data and its properties. Learning itself is the process of construction of a mathematical model for knowledge elicitation on the basis of the available data. The ultimate goal of the learning process is to build such an algorithm which would minimize the error of pattern recognition when identifying interrelations within the analyzed data. There are two important terms in relation to learning: training and testing data. Training set denotes the data used for the creation of the mathematical model. Dataset on which this model is applied is called the testing set, which labels are never known and should be found. There are different classes of problems and corresponding types of learning, defined by the availability of the training data itself and labels in it [94, 103, 104].

Supervised learning implies that there is a fully labeled training data available.

Thus, analysis of the testing data is made after a pattern recognition model

is crafted on the basis of the training set. Usually these techniques are called *classification* [16], as there are predefined known classes to which data samples can belong. This approach means learning by example. There is a large variety of classification principles, such as K-NN, Support Vector Machine (SVM), decision trees, neural networks, regression, rule-based classifiers, probabilistic Bayesian classification, etc. [103]. The nearest neighbor algorithm is explained in further detail in Section 3.4.2.1.

Unsupervised learning – no dataset with available labels, i.e. training data does not exist. However, in some cases training data without labels is also employed. The most common application of unsupervised learning is *clustering* problem - grouping of similar objects to the same cluster and dissimilar objects to a different cluster, or marked as outlier without any cluster assigned. This depends on the selected algorithm. In a more general case, the number of clusters is not known and interrelations within the data should be derived automatically. In contrast to classification, this approach is based on learning by observation. Clustering algorithms can be divided into several groups based on the principle of their operation. Partitioning methods use mostly distance similarity measures to assign a data point to a particular cluster. One of the most popular algorithms in this group is k-means, described in Section 3.4.2.2. The next type of clustering algorithms is based on the notion of density within the dataset and areas which are denser and considered to belong to one cluster. Density-based spatial clustering of applications with noise (DBSCAN) and Ordering points to identify the clustering structure (OPTICS) are common examples of density-based clustering. The hierarchical principle is based on the division of the available data to hierarchical structures either by merging small clusters to larger ones (agglomerative approach, e.g. AGglomerative NESting (AGNES) method), or by separating larger clusters into a set of smaller ones (divisive approach, e.g. DIvisive ANAlysis (DIANA)). Model-based clustering attempts to find a mathematical representation of the data. A popular example of a model-based approach is Kohonen's Self-Organizing Maps (SOM) algorithm which constructs an artificial neural network. Grid-based methods represent the data in a quantized grid space with multiple cells, and clustering itself is made within this grid. Such algorithms as STatistical INformation Grid (STING) and CLustering In QUEst (CLIQUE) represent this approach.

Semi-supervised learning : training data is partly labeled. As can be understood from the name, semi-supervised learning is in between the supervised and unsupervised approaches. These types of algorithms, for instance can correct clustering and make it more efficient [149]. This type of learning finds its application in the natural language and text process.

Reinforcement learning , or a "learning with a critic" implies that feedback on the correctness of the tentative label assigned to a data point is provided [94]. For instance, consider an image recognition task. If a submarine is classified as a car, in supervised learning the feedback would be "this is not a car, it is a submarine", but in reinforcement learning similar feedback would be

"this is not a car". I.e. only simple binary feedback can be used for further improvement of a classification algorithm.

3.4.2 Examples of pattern recognition algorithms

In this section a description of the selected classification and clustering algorithms used for anomaly detection is presented.

3.4.2.1 Nearest neighbor algorithms

The principle of the nearest neighbor algorithm [150] is rather simple. It is based on the calculation of the distances between every point in the dataset and a predefined number of neighbors, which is given as a parameter to this algorithm. The assumption needed for the operation of this algorithm is that similarity measure between points can be calculated. The most common option is Euclidean distance [151], which, however, is not the only possible choice [152, 153, 154]. Considering classification with discrete class labels, in order to assign a certain label to a point in a testing set, K-NN algorithm examines the distances from this point to k closest its neighbors from the training set. The assigned label of the considered testing point would be equal to the most popular label in the neighborhood. In case of real-valued prediction, the predicted value of the testing point can be equal to the average of the neighboring points from the training set.

Another class of problems where the nearest neighbor algorithm is applied is anomaly detection. Let us consider the algorithm named K-NN anomaly score. It is assumed that there are no labels in the training set, and the algorithm can be treated as unsupervised. In general, there are two distance-based approaches concerning the implementation of this algorithm; an anomaly score assigned to each point is either the sum of distances to k nearest neighbors [155] or distance to k^{th} neighbor [156]. Points having the largest anomaly scores are referred to as outliers. Usually, data points which have anomaly scores larger than certain percentile in the overall anomaly score distribution are marked as outliers.

There is a density-based Local Oulier Factor (LOF) nearest neighbor algorithm [157] – points which belong to more dense regions are considered as normal, and sparsely placed points are marked as anomalies. Also there are variations of LOF aimed at the improvement of the baseline algorithm [158, 159, 160].

It is the same for the majority of data mining algorithms, the nearest neighbor method also requires configuration parameters: distance measure and the number of nearest neighbors. It is a very important task to set parameters properly, as the resulting performance is heavily dependent on correct configuration. The most common distance measure is Euclidean, and the number of nearest neighbors is a square root of the number of samples in the input data. However, empirical testing is still needed to find the correct configuration.

3.4.2.2 K-means clustering

K-means is a classical distance-based clustering algorithm [94, 103, 161, 162, 163, 164, 165]. Input parameter k denotes the number of clusters to which data must be separated. It is an iterative algorithm based on the calculation of cluster centroids. The algorithm itself works in the following way:

1. Randomly selects k data points as cluster centroids;
2. Points with minimum Euclidean distance to centroids are allocated to a corresponding cluster;
3. Calculates the new mean values of the coordinates in each updated cluster to get the new centroids of clusters;
4. Repeat steps 2 and 3 until centroids are stabilized - mean values in the clusters do not change;
5. Clustering labels are found.

The known issue of this algorithm is that processing may have different clustering results at every run of the algorithm. This happens due to random initialization of the algorithm. In case the clusters are well separated, the actual clusters would not change, may be only cluster IDs. However if the data is not fully separable, there will most likely be different solutions depending on where the algorithm started the calculation of the centroids. In order to overcome this issue in [PII] a cluster separation measure based on the Davies-Bouldin index [166] has been used to select the best clustering solution of k-means algorithm.

3.4.2.3 FindCBLOF algorithm

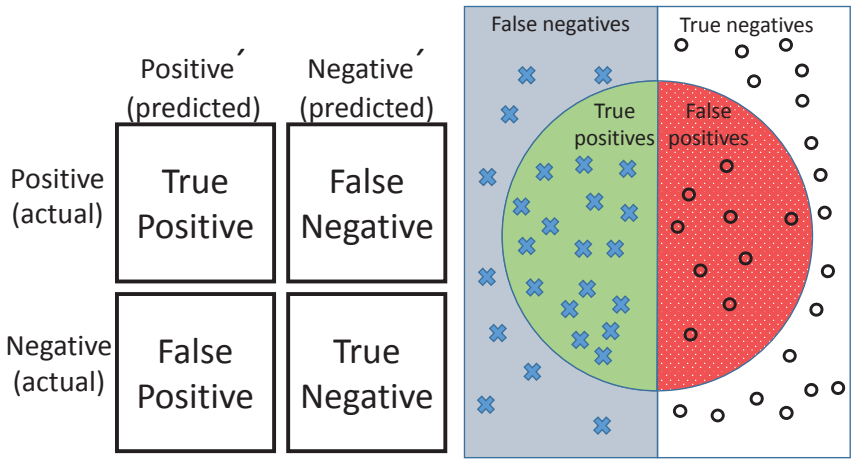
FindCBLOF is a density-based clustering algorithm [167]. In addition to clustering, this algorithm marks the possible outlier points. In other words the algorithm has an embedded notion of anomaly. The first processing step is clustering with a "Squeezer" clustering algorithm [167, 168, 169]. After that the clusters are marked as small or large using the input parameters α and β. CBLOF score is calculated for each point in the analyzed dataset:

- If point t belongs to a large cluster,
  $CBLOF = cluster\_size \times distance(t, cluster)$.
- If point t belongs to a small cluster,
  $CBLOF = cluster\_size \times distance(t, closest\_large\_cluster)$.

Anomaly detection is done using a percentile of CBLOF score distribution and specified as an input parameter of the algorithm.

3.4.3 Post-processing methods

Most commonly post-processing methods serve for interpretation of the results of pattern recognition in a meaningful and most convenient way [99, 170]. Visualization is widely used, e.g. distributions, bar-plots, Q-Q plots, and various



(a) Confusion matrix for 2 classes.

(b) Illustration to confusion matrix for 2 classes.

FIGURE 16 Confusion matrix and corresponding example.

histograms. For the purpose of QPM very commonly cell-wise histograms are employed. An example of such a visualization method is a sleeping cell histogram, which indicates the cumulative level of abnormality of every particular cell in the analyzed network region.

3.5 Metrics for evaluation of pattern recognition

In order to evaluate the quality of the classification algorithm, there is a set of commonly applied metrics. Consider that an algorithm of classification of tumors is malignant and benign has been prepared using the training dataset, and it is necessary to evaluate its performance. The following terms are used to further define efficiency of evaluation metrics. *True positive (TP)*- correct prediction of malignant tumor (disease is present and found by the test). *False positive (FP)* - incorrect prediction of malignant cell (disease is not present, but predicted to exist) - Type I error, or "false alarm". *True negative (TN)* - correct prediction of benign tumor (disease is not present, and predicted that there is no disease). *False negative (FN)* - incorrect prediction of benign tumor (disease is present, but not predicted by the test) - Type II error, or "miss-detection".

Taking into account the definitions given above it is possible to define the actual metrics which characterize classification quality.

$$Accuracy, [\%] = 100\% * \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

Accuracy denotes how many points have been correctly recognized in the overall data set. This is a very commonly used metric, but the problem is that accuracy

can be applied only if the analyzed data is balanced. In case of imbalanced data, i.e. when the size of one class is significantly larger than the other one, the cost of the error is much higher [23]. For instance, referring back to the tumor classification for cancer diagnostics, if there are 1000 considered cases - data points, and only 5 of those truly belong to the anomalous class, i.e. have malignant tumors on the X-ray, magnetic resonance or microscope image. In this case in order to achieve 99.5% accuracy it is necessary to mark all images as benign. Naturally this is not desired behavior of the classifier and a false alarm might be more tolerable than miss detection. Thus, accuracy is not suitable for evaluation of classification for imbalanced data [171, 172, 15, 173].

In order to overcome this issue some precision (14) and recall (15) have been introduced. Precision is a measure of prediction exactness, i.e. the fraction of correctly predicted positives within the overall number of the predicted positives. Hence precision goes down if the amount of false-positives or false alarms increases. Recall is also known as *sensitivity* - a measure of completeness, and indicates the portion of the truly anomalous samples classified as anomalous.

$$Precision = \frac{TP}{TP + FP}, \tag{14}$$

$$Recall = \frac{TP}{TP + FN} = TP_{rate}. \tag{15}$$

Consider a situation where two classifiers are compared against each other using precision and recall. Naturally, the one which has both of these metrics closer to 1 is better. However, in case one classifier precision is larger, and for another recall is larger, it is hard to say which classifier performs better. There is a combined metric called *F-score* (or $F_\beta score$) [16], aimed to represent the algorithm classification quality with one value [174, 175].

$$F_\beta score = \frac{(1 + \beta)^2 \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{16}$$

Most commonly $\beta = 1$. This coefficient is a non-negative real number and it is used to emphasize either precision or recall. For instance, if $\beta = 0.5$, this is $F_{0.5}score$, which weights precision two times higher than recall. Also, the other way round, if $\beta = 2$, the recall is twice as important than precision in the resulting F_2score. Thus, F-score is a flexible one-number measure for evaluation of classification efficiency.

The metrics discussed above give a numerical representation of classification efficiency. A method called ROC curve is a visual tool for performance comparison of several algorithms [16]. ROC curve is created in a space formed by a two-dimensional plane $TruePositive_{rate} - FalsePositive_{rate}$. TP_{rate} is effectively the same as recall, defined earlier in (15). FP_{rate} is shown as follows.

$$FP_{rate} = \frac{FP}{FP + TN}. \tag{17}$$

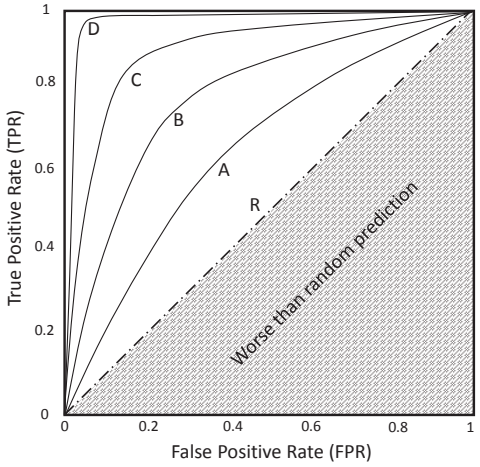


FIGURE 17 Example of ROC curves.

Thus, one classification outcome gives one point in the $TP_{rate} - FP_{rate}$ plane. Getting pairs of $[TruePositive_{rate}, FalsePositive_{rate}]$ for different portions of a testing set, it is possible to draw a curve in the ROC space. Consider that there are 4 algorithms, and the task is to compare their performance, Figure 17 shows corresponding ROC curves. According to this graph, algorithm D has the highest performance among others, while algorithm A is the closest to a random prediction, which is marked as R. Any algorithm which has ROC curve under line R demonstrates very poor performance. A numeric way to compare algorithms using ROC curves is to calculate corresponding Area under Curve (AUC) values [176]. The perfect classifier would have $AUC = 1$. Anyway, AUC values should be used together with the ROC curve graphs as in individual regions of ROC space algorithm with high AUC value can perform worse than an algorithm with lower AUC value. In order to evaluate classification performance with conventional metrics discussed above, it is necessary to have a confusion matrix, i.e. knowledge about actual positive (abnormal) and negative (normal) data points. Having this information calculation of precision, recall, etc. becomes possible. In a situation when original labels are not known it is possible to employ a heuristic approach. It should be based on the knowledge about the application area and ideal expected solution. Then it is possible to calculate the distance between the clustering or classification solution and this expected ideal case. An example of this approach is given in more detail in Section 4.2.

3.6 Summary

In this chapter an introduction to the knowledge mining process is given. Definitions of data mining and anomaly detection are presented. The description starts from an overview of the most common data types and their properties. Data

can be numerical, categorical, sequential, spacial, textual. Then, pre-processing, transformation and pattern recognition - the main steps of data mining, are discussed. Classification of the algorithms at each step is presented. In pre-processing these are:

1. Standardization, normalization - e.g. z-score method;
2. Feature construction, e.g. density calculation.

In transformation most attention is paid to the sliding window and *N-gram analysis* methods. Additionally, examples of both linear, e.g. PCA and MCA, and non-linear, such as diffusion maps, dimensionality reduction algorithms are described.

Next a typology of pattern recognition methods is presented, where classification and clustering take most of the attention. The principles of such algorithms as k-means clustering and nearest neighbor classifier are used as examples for each pattern recognition category. FindCBLOF algorithm is described, to give an example of an algorithm aimed for both clustering and anomaly detection.

Last, but not the least, various quality measures of data mining are discussed, including accuracy, precision, recall, F-score, ROC curve, as well as a more application specific heuristic approach. In the next chapter an overview of studies where knowledge mining methods are applied in the field of quality and performance monitoring in cellular mobile networks is made.

4 ADVANCED FAULT DETECTION, DIAGNOSIS AND HEALING IN MOBILE NETWORKS

The traditional QPM systems have been described in Chapter 2, and the concepts of knowledge and data mining were presented in Chapter 3. The goal of this chapter is to present how knowledge mining techniques are applied to the QPM systems. Thus, it begins with an overview of the state of the art in the current research in this area, discussed in Section 4.1. Results related to the application of anomaly detection based on user-specific MDT reports in QPM systems are presented in Section 4.2. Section 4.3 is devoted to the discussion of the main results of this dissertation, strengths, weaknesses and future prospects of QPM systems in cellular mobile networks.

4.1 Knowledge mining for quality and performance management: State of the art

To make a structured overview of the existing research activities devoted to advanced mobile network QPM, we identify the dimensions which can be used to characterize and compare the analyzed studies. These properties are the following:

- *Type of the network malfunction* – which failure situation is studied. The range of possible problems is rather broad, including hardware, software and configuration errors, as it is discussed in Section 2.2.
- *Origin of the data and validation environment* – real or simulated network, empirical or theoretical model.
- *Data collection time* – the amount of time needed for collection of a sufficient amount of data to carry out performance analysis, e.g. anomaly detection.
- *Analysis target* - can be either detection, diagnosis or healing. Very commonly a combination of detection and diagnosis is done, but only a few studies investigate the whole QPM cycle.

- *Level of analysis and collection entity* – can be related to one network element, e.g. a BS or a cell - a distributed approach. Data can be gathered from a tracking area with performance reports from tens or hundreds of cells - this is a centralized level. If analysis is made on several levels - it is a hybrid approach. UE-specific analysis can be classified as a separate level, when a user reported measurements or events are processed.
- *Type of statistics* – RAN level (e.g. signaling, measurements, network events, call blocking and success ratios), core network level, or TCP/IP level (traffic load, volume, delays, which are higher level statistics). This also includes information about the type of data from a knowledge mining perspective - categorical, numerical, sequential, etc. as is discussed in Section 3.1.
- *Periodicity of the reported data* – periodic, e.g. cell-level PI measurements, event-based, i.e. triggered by some predefined condition. User reports can be both event-triggered and periodic.
- *Data resolution* – how frequently monitored statistics are collected, e.g. milliseconds, seconds, hours, days. For instance, in traditional QPM systems collection samples could come more sparsely, while such methods as TRACE and MDT, discussed in Chapter 2, enable more frequent data collection.
- *Applied knowledge mining algorithms* – the type data mining or anomaly detection, presented in Chapter 3. The choice of the algorithm(-s) is defined by the type of the input performance data.
- *Availability of the training data and labels* – allows for the construction of a more reliable and accurate advanced QPM system. Labeled training data is difficult and expensive to get, as it is tedious and mostly a manual task. That is why availability of an unlabeled training set is more probable. In many cases an unlabeled training set contains only problem-free data from a normal network operation period.

Thus, the latest studies in the area of advanced QPM are compared in tabulated form, along the dimensions described above. A more traditional PM approach which relies on statistical means of analysis is presented in [11]. Detection of malfunctions is based on the usage of correlation coefficient between traffic loads of neighboring base stations. This approach is univariate - only cell load is considered as an input feature. The sought malfunctions are fast or slow in degradation in one of the cells. A big advantage of this study is that the performance data is collected from a real 3G network. The downside is that there is a necessity for a relatively long collection of cell load statistics and manual threshold setting for a correlation level. Moreover, the underlying assumption in degradation detection is that the cells are linearly correlated with each other. The latter is especially critical, as according to the authors correlation of only 5 % of neighboring cell pairs has been higher than 0.7, and for 86 % of cells correlation is under 0.5. Yet another issue is that degradation is modeled as a reduction of the number of users in one cell. However, there are real life scenarios which can lead to a shortage in activity of the users, e.g. end of the business day, and such behavior is absolutely normal.

TABLE 5 Comparative analysis of advanced QPM studies. Part I.

Article	[21]	[22]	[177]	[178, 179, 180]	[181, 182, 183]
Malfunction	User at cell edge; high shadowing; TxP degradation; faulty hardware	Special UE mobility; Partial coverage hole; Slow fading peak	Low circuit and packet switched call success rates	Unspecified service degradation	Unspecified service degradation
Scenario & Environment	Homogeneous, 3-sector directional antennas, LTE RAN dynamic simulator	Homogeneous, Macro 57 cell, LTE simulator, CQI from real HSPA network	70 cells, real data 3G network	70 cells, real data 3G network	2000-4000 cells, real 3G network
Analysis target	Detection & diagnosis	Detection & diagnosis	Detection & diagnosis	Detection	Detection & diagnosis
Analysis level	Cell level	Cell level	Cell level	Cell level	Cell level
Collected PM data	CQI, Call drop rate, HO timing advance, HW alarm	CQI, Call drop rate, Outgoing HO timing advance	Number of successful circuit and packet switched calls	Cell level traffic and call control KPIs	Cell level traffic and call control KPIs
Collection time and frequency	Not specified	14 days, per hour average	180 days, per hour average	125 days, per hour average	75 days, per hour average
Knowledge mining algorithm(-s) **A = Detection;** **B = Diagnosis;** **C = Healing**	A: Statistical profiling and correlation; B: Manual	A: Statistical profiling, z-score preprocessing; B: Likelihood based scoring function	A: Statistical profiling, Kolmogorov-Smirnov test; B: Likelihood based scoring function	A: Ensemble of supervised classification algorithms (time series analysis)	A: Topic modeling and HDP clustering; B: Probabilistic reasoning with MLN.
Training data	Unlabeled faultless data	Unlabeled faultless data	Unlabeled faultless data	Labeled faultless training	Unlabeled normal training data
Num. of features	4	3	2	12	11
KM architecture	Distributed	Distributed	Distributed	Distributed	Centralized

A complete fault detection and diagnosis system is proposed in [21]. The analyzed KPIs include CQI, call drop ratio, HO timing advance and hardware alarms. During the training phase, cell specific statistical distributions of each KPI are created using normal faultless data. Detection with testing data is done through assessment of a cumulative deviation of KPIs from the normal profiles. A diagnosis system is a knowledge base which should be manually filled by an expert, so that it is able to match particular anomalous combination of KPI values and failure root cause. The possible reasons for abnormal behavior distinguished by the diagnosis system are: "user at cell edge", "high shadowing", "transmission power degradation" and "faulty hardware". Among the strong sides of the described fault detection and diagnosis approach is that KPIs are analyzed jointly. Moreover, the processed data is user specific, and even though the analysis operates with distributions, individual values are still much more of an accurate approach if compared to traditional systems. The problem of the diagnosis part is that it cannot operate without a manually created base of failure cases. A significant drawback of the proposed approach is that a long time is needed to train the system and build statistically reliable KPI profiles. Even more problematic might be the creation of comparable profiles for failure situations. Thus, the described system of detection and diagnosis is an improvement against traditional PM systems, but it still has serious issues, which might become obstacles on the way to applying it to real networks.

The study presented in [22] is a continuation of the research done in [21]. It exploits the same idea of KPI profiling, but diagnosis part is significantly improved. KPI profiles themselves are built using averaged values of KPIs over a relatively small number of samples if compared to the overall number of samples in the dataset. Faultless data is used to train the detection system. Testing is carried out online, i.e. sliding window of the last n samples is input to the anomaly detection system. The output of the detection is a set of KPI specific abnormality levels per cell, in the range $[0, 1]$, where 0 is no deviation from normal behavior, and 1 denotes the strongest abnormal behavior. One of the issues in the detection system is that even though it is claimed to derive KPI thresholds automatically, it still contains some calibration constant which significantly impacts the resulting abnormality level. Diagnosis system can be filled by an expert, similarly to the previous study. In addition, and this is the improvement, the new diagnosis targets (failure cases) can be proposed by the system, and under guidance of the expert extend the diagnosis base. Using the abnormality level of each KPI provided by the detection part, the diagnosis calculates the likelihood that a particular combination of KPIs indicates this or that failure case. The target with the largest likelihood value is considered to be the diagnosed failure or normal case (target zero denotes no failure). Thus, the proposed system is guided by the operator, and with time it can be effective in detection and diagnosis of failures, however it still requires substantial training time (in the range of several days to weeks). Because of that in case of normal behavior change there is a large probability of false alarms. Moreover, the operator has to manually identify all normal profiles of the network operation, which is a tedious and time-consuming task.

The study presented in [177] is aimed to address the problem of manual selection of normal network state periods in [22]. Having historical data, the proposed profile learning system is able to identify all normal modes of operation. The operator still has to input the periods of faultless network operation, but without explicit specification of behavioral profiles. In order to measure the distance between statistical distribution of cell KPI profiles a Kolmogorov-Smirnov two-sample test is used [184, 185]. The ultimate goal of profile learner based on the Kolmogorov-Smirnov test is to measure the distance between statistical distribution of cell KPI profiles and derive a minimal set of such profiles. A fault detector then identifies combinations of KPIs from normal behavior and then diagnosis entity classifies the problem using the same logic, as it is presented in [22]. The evaluation of the method is done with the PM data from a real 3G network, with observation from a number of successful packet and circuit switched calls over 180 days. The developed system is capable of identifying significant falls in the number of successful packet-switched calls, and distinguish that the increase of this KPI is not an anomaly. Similarly to previously discussed studies training period took significant amount of time - about 4 days of statistics from faultless operation. The detection and diagnosis part was able to identify errors only about 100 hours after the failure actually occurred. The obvious benefits of the proposed system are the ability to automatically distinguish between different patterns of normal behavior and carry out the diagnosis through classification of faulty cases. However, the time scale for training and then fault detection/diagnosis is too large if the goal of the QPM system is to secure the end-user from poor QoS and QoE.

Another study employs a time series analysis, with classification of KPI behavior to normal and abnormal [178, 179]. This is a supervised anomaly detection approach with manually labeled training dataset. In this study several univariate and multivariate methods were applied to a large database of real measurements collected over 4 months of network operation are exercised. Altogether 12 KPIs related to both application-specific statistics, e.g. UL & DL data volume and throughput, and call control parameters such as call drop and setup success rates. Analysis is done per cell. Trained classification methods are ranked according to their performance. The process of algorithm training required statistics from nearly 2 months of network operation. Poorly performing methods are removed from the ensemble after some time. The advantages are the usage of real data, carefully compared classification performance, and employment of a large number of diverse anomaly detection techniques. Application of ensemble for prioritization of the most efficient classifiers is very beneficial. Moreover the developed system is ready to be joined with a healing process – if the configuration changes, the ensemble triggers the creation of a new model. Computational complexity and detection delay of the algorithms are also evaluated in the ensemble [179]. Some discussion regarding deployment of the proposed system, and corresponding demonstrator system described in [180]. The drawbacks of the proposed ensemble approach include the need for manually prepared labeled data, and the necessity to collect data over a significant amount of time - in the order

of several months for reliable detection. It might also be challenging to come up with exact weights of the ensemble system, and requires an intelligent algorithm for selection of weights and penalty scores.

Another framework for detection and diagnosis of network failures, anomalies and degradations is presented in [181]. Analysis is first performed in a centralized manner using a massive dataset, comprised of hourly statistics from 4000 cells, monitored for about 2.5 months. The number of analyzed KPIs is 11 and they include drop call and setup success rates, UL and DL throughput and data volume, handover success rate, and cell availability information. Topic modeling [186] is applied to categorized groups of cells according to their KPI values to several clusters. In order to detect abnormally behaving groups of cells and distinguish them from normal, a Hierarchical Dirichlet Process (HDP) algorithm [187, 188] is used. Using centroids of KPIs calculated for abnormal clusters, a simple classifier based on anomaly score thresholds is applied to mark the considered cluster as either *GOOD, BAD, VERY GOOD, VERY BAD or RELATIVELY BAD*. Diagnosis is done using Markov Logic Networks (MLN) algorithm [189] which is capable of creating a probabilistic knowledge base even with noisy or incomplete input data. In this study the authors managed to create a complete system for network quality and performance management based on data mining and anomaly detection algorithms. To achieve this, a considerably large dataset is used – collection time was over 2 months of network operation. One of the most critical questions is scalability in the time domain of the presented detection and diagnosis framework. For instance, if the number of samples is much smaller, will the system still be able to identify normal and abnormal clusters of cells? What is the minimal set of cells needed for convergence in the geographical scope? The answers to these questions define practical value and applicability of the system presented in [181].

A continuation study [182] goes into more detail about the contents of normal and anomalous clusters of cells with similar KPI states. Another goal of this study is to demonstrate how incremental learning [190] can be used for the analysis of ever-increasing network performance datasets. This is an initial attempt to address changes in network behavior reflected in KPIs through periodic updates of detection and diagnosis algorithms' parameters. A corresponding demonstration system based on the framework proposed in [181] is outlined in [183]. Comparison of the advanced QPM studies discussed earlier in this section is presented in Table 5.

TABLE 6 Comparative analysis of advanced QPM studies. Part II.

Article	[31]	[191]	[114]	[192, 193, 194, 195]	[196, 197]
Malfunction	HW failure. Catatonic sleeping cell.	High RRC failure rate, due to poor network planning	Unspecified	UL/DL interference, Coverage hole, HW, TxP, Transcoder failures	UL quality problems
Scenario & Environment	57 cells, simulated GSM-like network	335 cells, real 3G network	72 cells, real 2G/3G network	Real GSM, GPRS and 3G (in [192]) network	46 cells, Simulated, WCDMA
Analysis target	Detection	Detection & diagnosis	Detection	Diagnosis	Detection & visualization
Analysis level	User level, generalized to cell level	User level	Cell level	Cell level	Cell level
Collected PM data	Neighbor cell relations statistics	RRC specific categorical data (events)	Num. users, data & traffic stats, UL/DL signal power stats.	HO, call and transmission cell-level statistics	Num users, UL average noise rate, UL FER
Collection time and frequency	100 s, per second	8 minutes, 42215 samples	30 days, per hour	6 months	30 minutes
Knowledge mining algorithm(-s) A = Detection; B = Diagnosis; C = Healing	A: Decision tree and linear discriminant function classifiers	A: Pre-processing, diffusion maps, Density based clustering. B: Manual	A: Pre-processing, diffusion maps, Fisher polynomial classification	B: Bayesian networks	A: Self-organizing map, k-means clustering
Training data	Unlabeled faultless data	No training data needed	No training data needed	Labeled training data	Unlabeled training data
Num. of features	7	72	25	10 or more	3
KM architecture	Centralized	Centralized	Centralized and distributed	Centralized	Centralized

A different approach for the detection of problematic or sleeping cells is presented in [31]. In order to identify errors in network operation a visibility graph based on neighbor cell list reports is constructed. Vertices are cells and edges are relations between neighbors. Temporal changes in this graph are used to construct a vector of 7 numerical features. For detection of malicious behavior, a decision tree and linear discriminant classifiers are applied [94]. The results demonstrate that the problem in the base station high frequency equipment can be successfully detected with the proposed advanced PM methods. Also the impact of cell load on the detection quality is studied – with more than 4 users per cell, the achieved detection accuracy is 100 %. However, the downside of the proposed approach is a high false alarm rate, which complicates practical usage of this detection system. Though, this is most likely related to the classification algorithms, rather than the data representation approach based on the neighbor cell relations information.

There is also a pool of studies where Bayesian networks are used for detection and diagnosis of operational failures in mobile networks [192],[198], [199], [200],[193], [201],[194], [195], [202], [203]. The authors concentrate mainly on automatic diagnosis using cell level statistics, such as Blocked Call Rate (BCR), DCR, DCR during HO, throughput, frequency of active set updates, RRC establishment fail rate, HO fails, average number of radio links per cell, relocation failures during inter-RAT HOs, number of inter-RAT HO attempts, failure to add cell to active set and RSSI. Network failures and faults are considered as conditions which trigger the diagnosis process. Root cause analysis is done through observation of the symptoms – KPIs. To achieve this a network monitoring database of root cause probabilities $P(c)$ and conditional probabilities of symptoms $P(S|c)$ is created. Here S represents all possible symptoms, and c represent all root causes. Having these probabilities, it is possible to go the other way round and calculate the actual probability of the root cause c_i, given the symptoms $E = S_1 \dots S_M$, applying the Bayesian rule:

$$P(c_i|E) = \frac{P(c_i) \cdot \prod_{j=1}^{M} P(S_j|c_i)}{\sum_{n=1}^{K} P(c_n) \cdot \prod_{j=1}^{M} P(S_j|c_n)}, \tag{18}$$

Available types of Bayesian networks, such as simple (naïve) Bayesian networks and independence of casual influence, which can be used for diagnosis of network failures are discussed in [200, 201]. One of the biggest challenges of this approach is related to the estimation of probabilities $P(c)$ and $P(S|c)$. There are two different approaches to solve this problem, called *model construction* are: knowledge-based and data-based. The first one implies that expert knowledge is used to create the probability database. I.e. during their daily work, network monitoring engineers have to add every resolved failure as a root cause and then describe the probabilities of the corresponding symptoms (KPI values). Using the knowledge-based approach, taking into account inaccuracies of the experts' estimates and non-uniform network structure and behavior has been shown that the probability of correct root cause diagnosis is around 60 % [194]. Data-based approach does not need involvement of experts, but requires training data for

elicitation of the root cause and symptom probabilities. Accuracy of this model construction approach is very much dependent upon the size of the training set and can vary from 45 % to nearly 80 %. Yet another challenge is how to represent the different symptoms. Most of the KPIs are continuous variables, however the Bayesian network approach requires an estimation of probabilities for different KPI states. Both continuous [199] and discrete KPIs [195] have been used for parameterization of the probability database. The strong side of model construction on the basis of continuous symptoms is that diagnosis accuracy can be close to 90 %, something which is never reached by models based on discrete symptoms [193]. However, on the other hand, usages of continuous KPIs as symptoms demonstrated that the resulting diagnosis quality is heavily affected by inaccuracies in model parameters [199]. Moreover, for data-driven model construction with small training sets diagnosis accuracy drops to 40 %, which is worse than random [193]. Because of that, usage of discrete KPIs as symptoms is more popular. Approaches to derivation of discrete symptom values is discussed in [195].

The advantage of the discussed approach for automatic performance monitoring is that Bayesian networks are suitable for probabilistic representation of different states and their relations to the symptoms, which are illustrating these states. Among the challenges are the selection of model parameters, discretization of symptoms, unstable performance and overall diagnosis accuracy around 80 %, which is not so high. Additionally, construction of the model itself is a tedious and time consuming task.

Neural networks have also been applied in quality and performance monitoring automation [197, 204, 205, 206]. For instance in [197], a SOM algorithm, which is a type of artificial neural network method has been employed for the detection of UL quality problems. Simulated data from 3G network and real geographical areas are used as a scenario. The 3 main KPIs from all cells measured over time, are input to the SOM algorithm and the resulting points are clustered with the k-means algorithm, discussed in Section 3.4.2. As a result it is possible to identify the isolated cells, i.e. the ones which have noticeably different behavior, and to see what the range of KPI values are. Thus, the proposed technique can be used to visualize the behavior of different cells and potentially identify problematic cells or network regions.

In some of the studies non-linear dimensionality reduction techniques, like diffusion maps have been used. For instance, in [191] user specific call traces, i.e. sequences of network events are considered. TRACE functionality, discussed in Section 2.3, is enabled for 8 minutes in a real 3G network, and provides a dataset of about 42 thousand samples from 335 cells. This is an impressive amount if compared to the previously discussed studies, taking into account the collection speed. To carry out the detection of abnormal base stations, input features are converted from categorical to numeric format. Then dimensionality of the dataset is reduced with diffusion maps, and on the basis of the point densities in the new embedded space the anomaly detection is done. A simple statistical threshold is used to distinguish between normally and abnormally behaving cells. Despite simplicity, the cells marked as anomalous, truly demonstrate malicious behavior.

The fault diagnosis is done manually, on the basis of expert knowledge in performance monitoring.

More extensive utilization of non-linear transformation techniques for advanced PM is presented in [114]. Real network data collected over 4 weeks of monitoring time included such statistics as the number of active users, UL/DL data, percentage of idle time, throughput, DL SINR, UL power back-off. First standardization of the input data is done. The next step is dimensionality reduction with the diffusion map algorithm. This is needed because there are 25 input PIs. Several methods for clustering and classification are applied to detect network anomalies. It is worth noting, that the authors analyzed data both in distributed (one cell at a time) and centralized (all cells at once) manner. Also analysis of the data during the day and night are presented.

4.2 Advanced performance monitoring with MDT data

Most of the studies discussed above are concentrated on the analysis of cell level KPIs. The common problem of this approach is that the collection of a sufficient amount of performance monitoring data takes a substantial amount of time, e.g. weeks or months. The gathered dataset should be statistically reliable and large enough for detection of problematic network regions. This section is devoted to the presentation of results of the studies in advanced performance monitoring based on user level statistics. Also it is shown how enhanced fault detection methods can be applied in the future self-healing mobile networks.

Network functions used for collection of subscriber measurement data are TRACE and MDT, discussed in Section 2.3. TRACE exists since 3G UMTS networks, and is aimed at gathering periodic or event-triggered user specific measurement statistics. MDT is a successor of TRACE added in LTE and later releases of UMTS networks. It extends the capabilities of TRACE with more accurate location information and PM oriented user reports and events. The general purpose of MDT is an improvement of the network performance through an enhanced collection of user statistics on experienced QoS, radio conditions and partly signaling. There is a possibility to flexibly collect data from specific geographic areas, or particular users or groups of users. Data gathering can be done either periodically, or on the basis of certain pre-defined network events. The resulting MDT performance log can reach thousands of samples from just a few minutes of data collection. This leaves a lot of space for the application of knowledge mining techniques for fault detection, diagnosis and healing. The structure of studies devoted to analysis and anomaly detection in MDT logs, which form the basis of this dissertation, is presented in Figure 18.

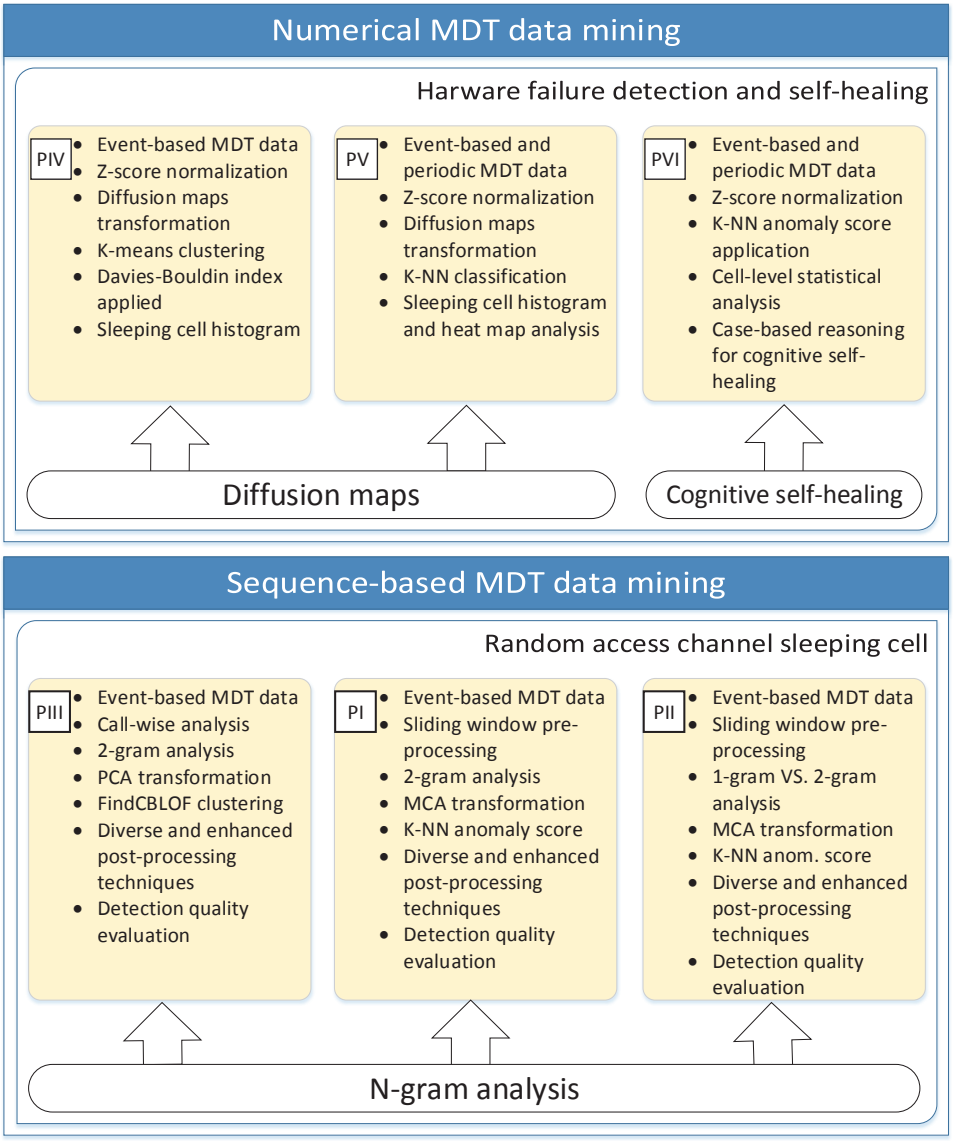
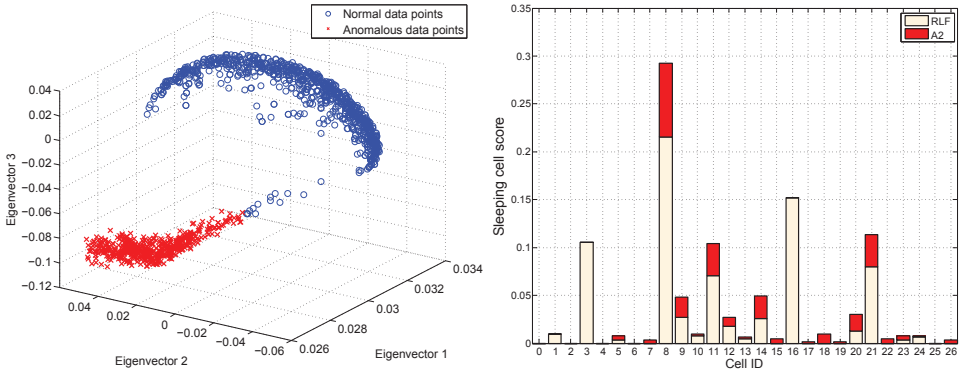


FIGURE 18 Structure of research activities for advanced quality and performance management with MDT data.



(a) Results of k-means clustering in the embedded space.

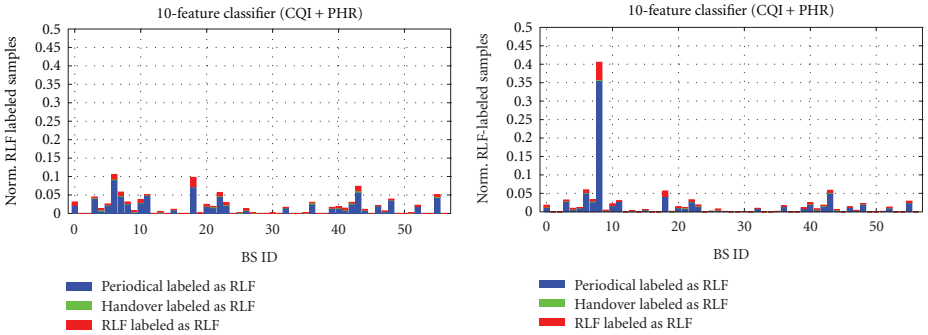
(b) Sleeping cell histogram for anomalous RLF and A2 events.

FIGURE 19 Sleeping cell detection in diffusion maps embedded space and k-means clustering from [PIV].

4.2.1 Sleeping cell detection based on numerical MDT data

The initial study, described in [PIV], is devoted to the analysis of *event-triggered MDT reports* in LTE network by means of diffusion maps transformation and k-means clustering techniques. The analyzed data includes serving and neighboring cell RSRPs and SINR, reported at the moment of certain pre-defined MDT event. Altogether 9 PI variables are measured and hence the input data set has 9 dimensions. The role of the diffusion maps transformation is two-fold: the number of dimensions is reduced from 9 to 6, and the data is presented in the new space in a more meaningful way. The resulting data manifold in the embedded space is used for calculation of point densities and corresponding separation to normal and abnormal clusters with the k-means algorithm. The cloud of data points along with the clustering result in the embedded space, along 3 selected dimensions, is shown in Figure 19a. In order to achieve reliability in the clustering decision, k-means is run multiple times and the best result, in terms of Davies-Bouldin separation measure, is used. The sleeping cell histogram, shown in Figure 19b, demonstrates that an artificially induced hardware failure in cell 8 can be successfully detected with the proposed algorithm, as it has the highest anomaly score. Also the instances of anomalous behavior are observed in the neighbors of eNB 8: cells 9 and 16. More results on this study can be found in [33]. The drawback to this approach is that a false alarm rate in the worst case rises up to almost 15 %, which is rather high. In an attempt to improve the sleeping cell detection quality, numerical analysis of *periodic MDT reports* is done in addition to the event triggered.

In [PV], similarly to [PIV], a hardware failure is introduced in one of the network cells. MDT campaign is enabled for the whole LTE network areas with 57 cells, both event-triggered and periodic reports are collected from the users. The same type of PIs are considered with the addition of Power Headroom (PHR)



(a) RLF-labeled samples per base station in training data.

(b) RLF-labeled samples per base station in testing data.

FIGURE 20 Classification results of periodic and event triggered MDT reports from [PV].

measurements. Data standardization is done with *z-score* and transformation with diffusion maps. However, in the low dimensional space K-NN classification is applied instead of clustering. To make use of K-NN, a training set is labeled into three classes: periodic report, HO-triggered report and RLF report. The result of labeling to RLF class per cell is shown in Figure 20a. The trained K-NN classifier divides the testing set into 3 classes, correspondingly. In case of a certain cell in the testing set has an excessive amount of periodic MDT reports labeled as RLFs, this cell is identified as anomalous. The results of the testing data classification is shown in Figure 20b. A cell with ID 8 has a much larger number of periodical MDT reports if compared to its neighbors, or its own behavior in the error free training case. Hence, it can be concluded that cell 8 contains a failure. The advantage of this method is that the necessary amount of periodic MDT reports for fault detection can be collected very quickly - within several minutes. Compared to the results presented in [PIV], periodical measurements add statistical reliability to detection of the sleeping cell.

In another study, presented in [PVI], the goal was to build a complete self-healing system with fault detection based on knowledge mining and a cognitive healing part. Detection of failures is based on K-NN anomaly score algorithm applied to user-level MDT reports collected from the overall network. The next step of anomaly detection is done cell-wise, and the resulting per-KPI abnormality score is produced. Cognition is implemented by means of case-based reasoning algorithm, which consists of the *retrieve, revise and retain functions* [207]. The desired behavior is that the system selects the solution for the detected failure and learns from the output and how effective it was. The most efficient solutions are then prioritized, while mistaken ones are penalized. To achieve this, first the *retrieve* process is called responsible for the selection of the appropriate solution depending on the input anomaly report. The second step is *revise* process, which is used to verify the effect of the selected healing action. The last step, which concludes the cognitive learning process, is the *retain* process. It extends the case base

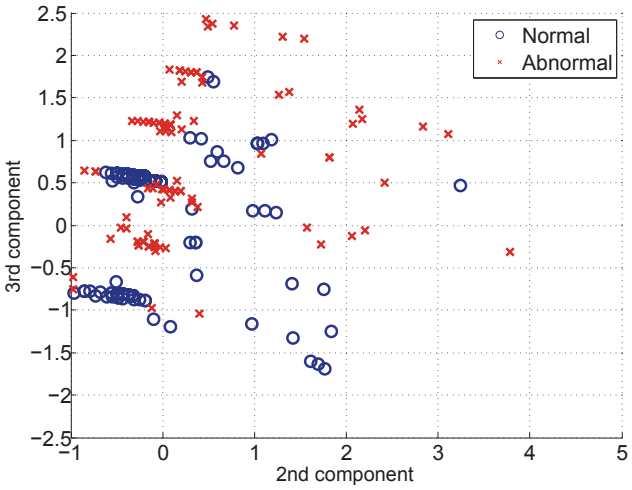
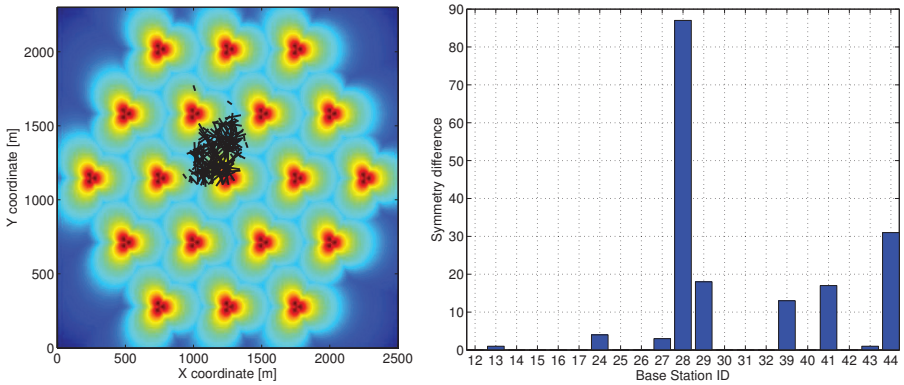


FIGURE 21 Results of FindCBLOF clustering in the embedded space after 2-gram and PCA transformation. Axes are components of PCA.

with a confirmed solution - efficient for solving problem case defined by the input anomaly report. Thus, if next time a similar anomaly report appears, the solution will already be known. More details on the applied case based on the reasoning algorithm can be found in [207]. Application of the described QPM approach leads to the reduction of the number of RLFs by 15 %, and improvement of RSRP level in the area of the coverage hole, as it is discussed in [PVI]. A detailed comparison of the studies [PIV], [PV] and [PVI] devoted to the detection of hardware failures on the basis of numerical MDT data can be found in Appendix 1, Table 8. The studies discussed above concentrate on the analysis of numerical characteristics of MDT data, and require 6 to 10 reported measurements. MDT functionality available in LTE networks and the latest releases of UMTS makes it possible to collect these measurements, but in other networks like GSM, WCDMA or early releases of HSPA not all of them are available, if any. In order to create a more flexible and reliable fault detection system the research direction is switched from processing of numerical data to analysis of sequences of network events. In the studies discussed below, N-gram analysis of the MDT data is done, but the core idea can be extended to fault detection in both TRACE and control message logs.

4.2.2 Sequence-based analysis of MDT data for sleeping cell detection

The first attempt to analyze the series of events which comprise calls of individual users is presented in [PIII]. The heart of fault detection is the N-gram analysis described in Chapter 3. Using $N = 2$, the MDT event sequences are transformed from sequential to the numerical format. There is a trade-off between computational complexity and detection performance when selecting the number of grams N . The tests have shown that $N = 2$ keeps the number of dimensions in the new feature space relatively low, and at the same time the output allows for successful identification of the problematic cell. After conversion of the event



(a) Locations instances of anomalous 2-gram.

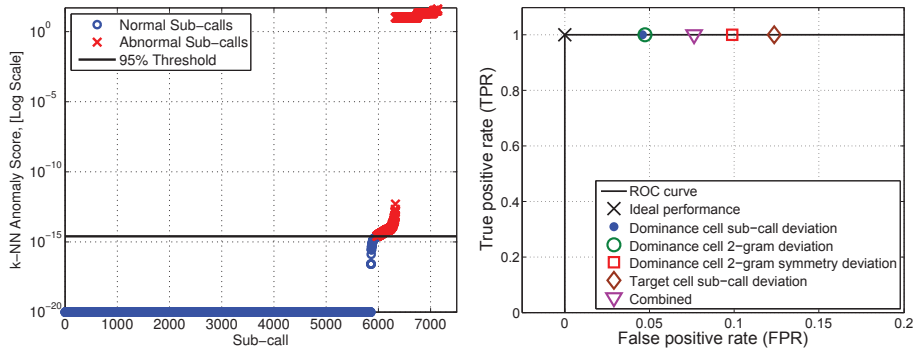
(b) Sleeping cell histogram for anomalous 2-gram.

FIGURE 22 Sleeping cell detection based on anomalous 2-gram: "Handover COMMAND–A2 RSRP Enter".

sequences to a numeric format the data is transformed with PCA and analyzed with FindCBLOF clustering algorithm. The resulting separation to normal and abnormal points is shown in Figure 21.

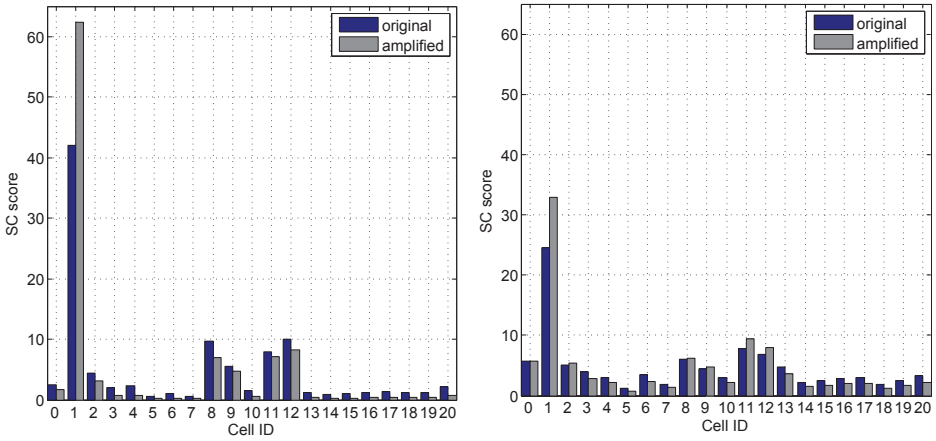
Using information about anomalous user calls it is possible to identify the most suspicious 2-grams and create sleeping cell histograms for each of them. Locations of the anomalous 2-grams are shown on the network map in Figure 22a, and corresponding sleeping cell histogram is depicted in Figure 22b. The introduced sleeping cell failures is a *random access channel failure* in one of the cells. In [PIII] failure occurred in cell 28, so the detection was supposed to identify that cell as problematic. From the sleeping cell histogram it is visible that cell 28 indeed has the highest anomaly score, as well as its neighboring cells 24, 27, 29, 39, 41 and 44. Thus, the achieved detection accuracy is very high.

As it is mentioned above, sleeping cell detection in [PIII] is based on the analysis of individual calls. But calls are variable in duration and in case of very long calls users are able to visit a number of cells, what introduces noise and results in degradation of detection accuracy. This is especially crucial for fast moving users, which are able to reside in tens of cells in a few minutes. To overcome the negative influence of the user behavior in detection quality, the next study has been carried out. In [PI] the original calls have been pre-processed using the sliding window approach. This results in higher detection reliability and makes it independent from the user velocity and call duration. While the core of the fault detection remained the same - N-gram analysis ($N = 2$) for processing of MDT event sequences, corresponding changes were needed in transformation and anomaly detection methods. PCA is substituted with MCA, and a selection of the number of components for embedded space is done using SORTE method. For anomaly detection K-NN anomaly score algorithm is applied instead of FindCBLOF clustering. The results demonstrate that the applied process is capable of representing the data in such a form that anomalous points are fully



(a) Sorted K-NN anomaly scores in problematic dataset.

(b) ROC curve of anomaly detection algorithm, and performance of different post-processing methods.



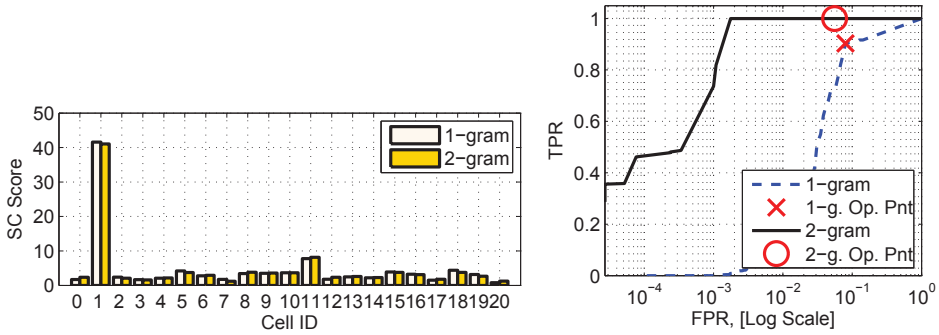
(c) Sleeping cell detection histogram for Dominance Cell 2-Gram Symmetry Deviation post-processing method.

(d) Sleeping cell detection histogram for Target Cell Sub-Calls post-processing method.

FIGURE 23 Quality of sleeping cell detection in [PI].

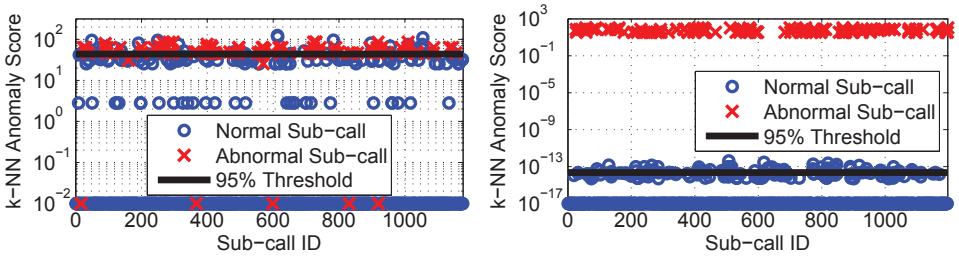
separable from the normal points, as it can be seen from Figure 23a and from the ROC curve in Figure 23b. It is still necessary to mention that the developed post-processing methods are not able to fully avoid a false alarm rate. These methods employ different kinds of information about anomalous sub-calls or 2-grams for mapping to a particular base station. The result is a sleeping cell detection histogram, such as the ones shown in Figures 23c and 23d. An additional method, used to improve the detection quality in post-processing is called amplification, and it takes into account the mutual abnormality impact of neighboring cells. The presented sleeping cell histograms contain both original and amplified cases.

The continuation study, fully based on the framework developed in [PI] investigates the possibility to use 1-gram analysis instead of 2-gram. Target cell deviation post-processing method is applied. The results demonstrate that in the best case the detection quality 2-gram analysis is only slightly higher than for 1-



(a) Sleeping cell detection histogram for Target cell sub-call deviation post-processing method.

(b) ROC curve for comparison of sleeping cell detection with 1-gram and 2-gram analyses.



(c) Anomaly scores in 1-gram analysis.

(d) Anomaly scores in 2-gram analysis.

FIGURE 24 Detection results and performance from [PII].

gram, as it can be seen from sleeping cell histogram and ROC curves in Figure 24. However, the anomaly scores demonstrate that 2-gram analysis (see Figure 24d) provides much clearer separation of the normal and abnormal points, than the 1-gram analysis (see Figure 24c). The reason for the similar performance is that the 95^{th} anomaly score percentile rule for separation to normal and abnormal classes is not perfect, and can be improved with a more intelligent approach. Comparison of studies devoted to sleeping cell detection based on analysis of sequences of MDT events is presented in Appendix 1, Table 7.

4.3 Discussion

Advanced QPM systems vary greatly in approach which they use to monitor and manage network performance. Some analyze the whole network using cell-level statistics, some go to the user level and look into the performance of specific cells. In certain QPM studies, performance statistics are gathered over months with sample resolution of hours, while in others there are only minutes of measurements, but with granularity of milliseconds. Thus, the different extent of centralization/distribution, statistics collection frequency and type serve the same

purpose - detection and diagnosis of problematic situations in network operation.

Even though there are diverse approaches, the lion's share of research activities build the advanced performance monitoring systems on the basis of the cell level statistics. In contrast, the performance monitoring system, developed within the research described in this dissertation, is based on the user data. The following two key concepts form the foundation of the proposed advanced PM system. First, it is the innovative idea of processing user specific MDT reports with knowledge mining techniques, which has never been done before. Second, it is a new fault detection approach of analyzing user event sequences instead of separate samples of numerical measurements. Development and evaluation of the PM framework, based on these two concepts, demonstrates that different types of network malfunctions are accurately and timely detected. Moreover, the selected validation environment, which is a simulated LTE network, included the modeling of non-trivial failures, also known as sleeping cells. At first, the efficacy of anomaly detection for fault identification is demonstrated in a less complicated case – hardware failure in one of the networks base stations. Analyzed features have numeric format, and the data is collected with periodic and event-triggered MDT functionality. Next, the detection framework was built for a more sophisticated MAC layer failure – malfunction of a random access channel. An important characteristic of this logical fault is that no coverage holes are introduced in the network, and hence regular channel quality measurements do not reveal the problem. To address this complication the developed framework employs *sequential knowledge mining analysis* of MDT data. The amount of the statistics necessary for the detection is very limited – only instances of network events, but the measurements themselves are not needed. Notable effort has been done for the achievement of correct mapping of the detected problem with a particular base station or cell. Various post-processing methods have been developed to achieve this, and the results are presented as sleeping cell histograms and heat maps. The detection quality is measured with conventional and heuristic metrics, which demonstrate that the constructed system can be extremely efficient for detection of non-trivial sleeping cell problems. Moreover, the concept of advanced performance monitoring is shown to be applicable in self-healing networks. A self-healing system, enhanced with knowledge mining techniques, can benefit from the automatic malfunction detection and diagnosis, and iteratively reconfigure cellular network parameters according to the identified problems.

4.3.1 Pros and Cons of Anomaly Detection in Performance Monitoring

One of the key advantages of knowledge mining applied to network quality and performance management is the ability to process large, high-dimensional datasets in an automated or semi-automated manner. It is especially important that mutual – multivariate analysis of performance metrics can be done. This differs from the traditional PM systems, where PIs are mostly analyzed independently from each other. In many cases a combination of acceptable deviations in

several KPIs, and unnoticed by a traditional PM system, actually indicates the existence of a problem. Advanced performance monitoring based on knowledge mining can achieve higher accuracy because less or no aggregation of PI values is needed. As it has been shown in Section 2.3 there are 3 dimensions of averaging, and as a result of it, a tremendous amount of potentially useful data is left out from the analysis. The benefit of data aggregation is the burden of transferring this information to the analysis unit is significantly reduced, but the resulting accuracy of fault detection and diagnosis is compromised. Another, and possibly the main cause, is that manual processing of the aggregated KPIs is much simpler. Averaging in time domain has a negative side effect – fault detection takes much longer, as the collection of a reliable statistics base is slow. To build a long term profile of network behavior using averaged KPIs, weeks or even months are required. In contrast, the ability to handle massive datasets by data mining algorithms enables faster identification of network problems as they appear. This time is needed to create the profile of normal behavior. However, in case unsupervised anomaly detection is used, even short term data with sufficiently high granularity might be enough to carry out accurate fault detection with data mining techniques.

Among the complications of advanced QPM systems is that careful selection and configuration of algorithms should be done, taking into account the nature of the data and types of problems which are solved. However, the same is required for the correct operation of traditional PM systems. The main difference is that for knowledge mining algorithms there is no need to define all possible cases of problems in advance. On the other hand, in a properly built advanced PM system the necessity to derive thresholds or build profiles for individual KPIs is eliminated. Instead parameters of anomaly detection algorithms should be carefully selected and tuned. In some cases it is a challenging task to find proper parameters, and this adds complexity of building and maintaining the QPM system based on knowledge mining. This also might be seen as a disadvantage, as operators are willing to have simple solutions for complex problems. However, complexity of the cellular mobile networks would inevitably grow, especially taking into account the trend towards 5G, and this will push the usage of intelligent QPM techniques in such networks. One of the largest challenges in the development of a reliable advanced PM system is the need for addressing network behavior changes. Mobile networks are highly dynamic and normal behavior may drift or even drastically change. Knowledge mining systems should be built so that they distinguish these changes and avoid excessive false alarm rates.

4.3.2 Architecture of future cognitive QPM systems

Advanced QPM aimed at future mobile networks should be able to combine the ideas developed during research and practical work in the field of performance monitoring. In that respect, a certain extent of distribution together with power of centralization should be used. Data should be gathered from both cells and individual users to result in a superior detection and diagnosis quality. Timely

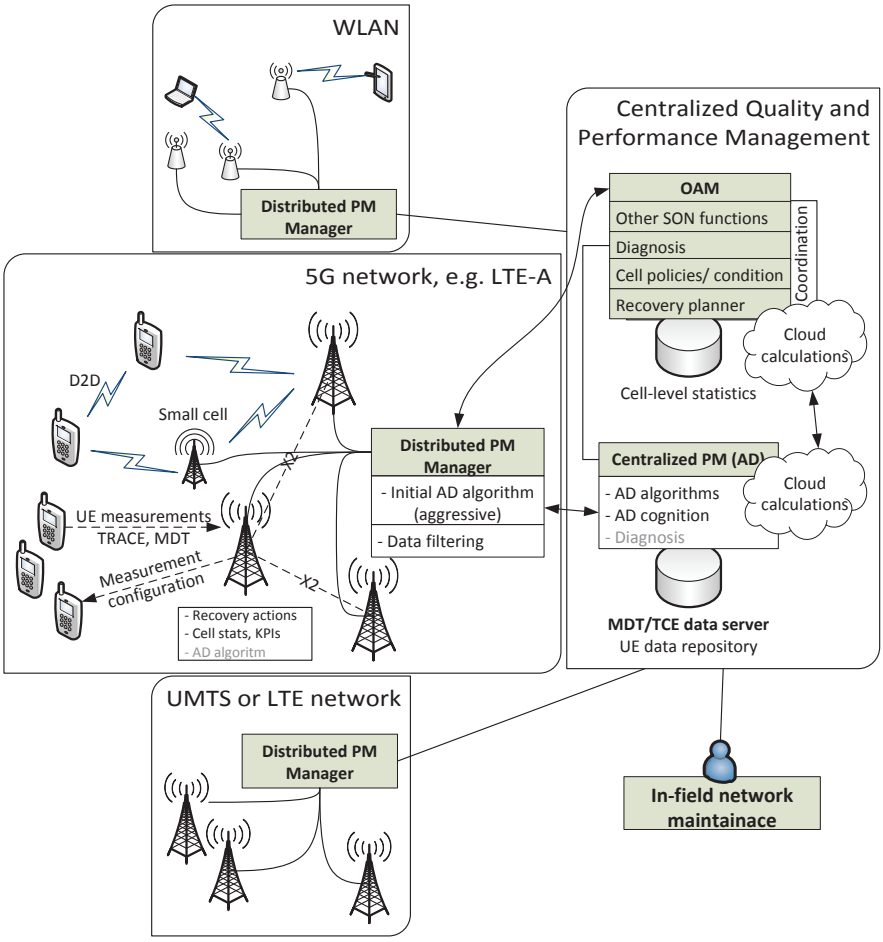


FIGURE 25 Quality and performance management architecture for future mobile networks.

and accurate identification of failures would allow for the compensation and healing actions to minimize the negative impact on the user QoS and QoE. Employment of the cognitive technologies can enable self-healing systems to handle the unexpected network behavior, taking into account heterogeneity of the future wireless communication networks. As a quintessence of ideas gathered during the carried out research and analyzing sources of literature, it is possible to foresee that in the future, mobile networks QPM systems will possess all of the qualities discussed above. Architecture of such networks may be as it is shown in Figure 25. The most important architectural elements of this system are monitored network(-s), distributed PM manager, assigned to a corresponding network or region in the network, and centralized quality and performance management system. Each network can belong to a different technology, such as e.g. LTE, WiFi, LTE-A or any future network. Users and cells inside this network carry out performance measurements and reports which are analyzed by *preliminary*

advanced PM, which resides in the distributed PM manager. This initial anomaly detection algorithm may have very aggressive settings and does not require a low false alarm rate. Also computational complexity of the preliminary analysis should not be high. Its main goal is to find the suspicious network regions and transfer a notification to the centralized QPM unit. In turn, the centralized unit is responsible for running a more thorough analysis of the networks performance, and can even involve cloud computations to achieve a high quality of fault detection and diagnosis. Ensemble of knowledge mining methods can be applied. Additionally, a centralized QPM unit can trigger TRACE, MQA or MDT measurement campaigns to achieve the desired fault detection in a short time. This is where the results of the research presented in this dissertation can be applied. After the problem is found and understood, the self-healing process is triggered. As far as possible, the central unit should also manage other self-organizing network functions, a coordination with self-healing solutions should be achieved relatively easily. Again, the power of cloud computations can be employed for coordination and solution development.

4.4 Summary

This chapter presents an overview of quality and performance management methods based on knowledge mining. First, an extensive review of the most notable research activities is made. A structural comparison of different studies, along the key aspects of PM, is presented. The analysis of cell-level statistics in both centralized and distributed manner forms the foundation of the advanced QPM methods. Next, research devoted to the processing of user level MDT statistics is discussed. The presented results are the core of this dissertation. A description of the analysis for both numerical and sequential features for the identification of complex network failures is given. The employed methods include normalization and standardization techniques, various transformation methods, clustering and classification approaches, along with different post-processing and quality evaluation methods. A combination of advanced PM and cognitive self-healing is outlined. The concluding part is devoted to a discussion about the main dissertation results, strengths and weaknesses of applying knowledge mining algorithms to quality and performance management in mobile networks. Possible architecture of future cellular networks is also presented.

5 CONCLUSION

During the past decade standardization work and practical deployment are booming in the world of wireless communications. Cellular mobile networks are converging with time to a unified system with many integrated radio access technologies. Within some of the advanced systems, cells of different sizes, on different layers, shape heterogeneous environment with myriads of nodes and active units. Thus, it is hard to foresee the scale of the future networks. Nevertheless, both current and future mobile systems have to be managed, and the provided QoS should be guaranteed to meet the demands of the users. Some positive changes have happened in the area of network quality and performance management with the introduction of self-organizing networks concept. SON automates some of the most important functions to reduce and keep at an appropriate level the expenses of the mobile network operators. However, self-healing SON function is least developed if compared to self-configuration and self-optimization. Automation of fault detection and diagnosis still requires substantial improvement. Taking into account that in the future 5G networks quality requirements are much stricter than in any of the previously developed systems, it is obvious that self-healing should also be enhanced to guarantee adequate robustness of mobile networks.

Study of minimization of drive tests and corresponding standardization activities enabled the collection of large and detailed performance datasets, filled with user level measurement statistics. A series of studies is done to make use of capabilities of knowledge mining in analyzing large data and employ them for enhancement of quality and performance management systems in mobile networks. There are two main directions in this research: analysis of the numerical properties of user measurement logs, and consideration of the sequential features with respect to MDT data. Different evaluation setups are studied, however the important assumption for development of the fault detection system, is that nontrivial network malfunctions, such as a sleeping cell problem, are considered. The largest potential in practical application of the proposed performance monitoring method is related to the analysis of sequential properties of user reports. Even though the evaluation is done on the basis of MDT data, the presented ap-

proach is extendable to analysis of TRACE and mobile quality agents statistics as well. This is possible because sequence-based PM requires a very limited amount of user-related quality data, which is similar in TRACE, MDT and MQA. Thus, network performance data processing based on knowledge mining enables maintenance of high service quality by addressing the complexity of mobile networks and dynamism of the changes in radio environment and user behavior. Integration of the advanced performance monitoring system with cognitive automatic recovery can lead to the creation of an intelligent QPM, capable of identifying malfunctions quickly, apply corrective actions, and by that, substantially improve the user quality of experience.

Future work in relation to advanced performance monitoring based on knowledge mining should be related to the integration of these systems to the real networks. Creation of a flexible setup, capable of addressing network changes even without a given notion of normal behavior, and reduced false alarm rate, should be the targets.

YHTEENVETO (FINNISH SUMMARY)

Tässä väitöskirjassa, jonka nimi on Edistynyt suorituskyvynvalvontajärjestelmä itsekorjaantuville mobiileille soluverkoille, kehitetään ja vahvistetaan järjestelmä, jolla voidaan valvoa nykyisten ja tulevien mobiilien soluverkkojen suorituskykyä. Väitöskirjassa käytetään tiedonlouhinnan tekniikoita MDT-toiminnallisuudella kerättyjen käyttäjäkohtaisten mittausraporttien analysointiin. Yhä kasvavat laatuvaatimukset, mobiiliverkkojen laajeneminen ja niiden monimuotoistuminen vaativat yhä tehokkaampia automatisoituja suorituskyvyn valvontakeinoja. Nykyään verkon toimintaa hallitaan enimmäkseen käsin ja hallinta perustuu koostettuihin suorituskyvyn mittareihin. Neljännen sukupolven itseorganisoituvat verkot tarjoavat mahdollisuuden useimpien verkon toimintojen automatisoinnille, mutta käyttäjien raportoimien mittausten analysoinnin mahdollisuus on jäänyt vähälle huomiolle.

Tässä tutkielmassa kehitetty edistynyt suorituskyvynvalvontajärjestelmä ottaa huomioon käyttäjien mittausraporttien numeeriset ja jaksolliset ominaisuudet virhetilanteita tunnistettaessa. Tutkielmassa keskitytään passiivisten solujen analysoitiin sekä fyysisellä että siirtokerroksella. Verkon ongelma-alueet tunnistetaan käyttämällä täyden kierron tiedonlouhintaa. Tiedot esikäsitellään tilastollisella normalisoinnilla ja liukuvan ikkunan menetelmillä, lineaarisilla ja epälineaarisilla muunnoksilla, sekä ulottuvuuksien vähentämisellä. Jatkokäsittelyssä käytetään kokoavia ja luokittelevia menetelmiä. Lisäksi väitöskirjassa ehdotetaan ja käytetään useita jälkikäsittelyn ja tunnistuksen arviointimenetelmiä. Näin muodostunut valvontajärjestelmä on kykenevä nopeaan ja tarkkaan epätriviaalien verkon virhekäyttäytymisien tunnistamiseen ja sopii tuleville mobiiliverkoille, myös yhdistettynä oppiviin itsekorjaantuviin verkkoihin.

REFERENCES

[1] A. Osseiran, V. Braun, T. Hidekazu, P. Marsch, H. Schotten, H. Tullberg, M. Uusitalo, and M. Schellman, "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, June 2013.

[2] S. Hämälainen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. Wiley Publishing, 1st ed., 2012.

[3] J. Laiho, A. Wacker, and T. Novosad, *Radio Network Planning and Optimisation for UMTS, 2nd Edition*. New York, NY, USA: John Wiley & Sons, Inc., 2006.

[4] C. Frenzel, H. Sanneck, and B. Bauer, "Automated rational recovery selection for self-healing in mobile networks," in *Wireless Communication Systems (ISWCS), 2012 International Symposium on*, pp. 41–45, Aug 2012.

[5] NGMN, *Recommendation on SON and O&M Requirements*, 2008.

[6] NGMN, *Use Cases related to Self Organising Network, Overall Description*, 2008.

[7] 3GPP TS 32.500, *TSG Services and System Aspects; Telecommunication Management; Self-Organizing Networks (SON); Concepts and requirements (Release 12)*, October 2014.

[8] 3GPP TS 32.521, V11.1.0, *TSG Services and System Aspects; Telecommunication Management; Self-Organizing Networks (SON) Policy Network Resource Model (NRM) Integration Reference Point (IRP); Requirements (Release 11)*, December 2012.

[9] 3GPP TS 32.541 V12.0.0, *Self-Organizing Networks (SON); Self-healing concepts and requirements (Release 12)*, October 2014.

[10] 3GPP TR 36.902, V9.3.1, *TSG Radio Access Network; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (Release 9)*, March 2011.

[11] M. Z. Asghar, S. Hämälainen, and T. Ristaniemi, "Self-healing framework for LTE networks," in *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2012 IEEE 17th International Workshop on*, pp. 159–161, Sept 2012.

[12] M. Amirijoo, L. Jorguseski, T. Kürner, R. Litjens, M. Neuland, L. C. Schmelz, and U. Türke, "Cell outage management in LTE networks," in *Proceedings*

of the 6th international conference on Symposium on Wireless Communication Systems, ISWCS'09, pp. 600–604, IEEE Press, 2009.

[13] 3GPP TR 36.805 V9.0.0, *TSG Radio Access Network; Study on Minimization of drive-tests in Next Generation Networks; (Release 9)*, December 2009.

[14] 3GPP TS 37.320 V12.2.0, *TSG Radio Access Network; Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2 (Release 12)*, September 2014.

[15] N. Chawla, A. Lazarevic, L. Hall, and K. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003* (N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, eds.), vol. 2838 of *Lecture Notes in Computer Science*, pp. 107–119, Springer Berlin Heidelberg, 2003.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 3rd ed., 2011.

[17] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15:1–15:58, July 2009.

[18] R. Kreher, *UMTS Performance Measurement: A Practical Guide to KPIs for the UTRAN Environment*. Wiley, 2006.

[19] J. Ramiro and K. Hamied, *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*. Wiley Publishing, 1st ed., 2012.

[20] B. Cheung, G. N. Kumar, and S. A. Rao, "Statistical algorithms in fault detection and prediction: Toward a healthier network," *Bell Labs Technical Journal*, vol. 9, no. 4, pp. 171–185, 2005.

[21] S. Novaczki and P. Szilagyi, "Radio channel degradation detection and diagnosis based on statistical analysis," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pp. 1–2, May 2011.

[22] P. Szilagyi and S. Novaczki, "An automatic detection and diagnosis framework for mobile communication systems," *IEEE Transactions on Network and Service Management*, vol. 9, pp. 184–197, June 2012.

[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, pp. 1263–1284, Sept. 2009.

[24] F. Cong, A. Nandi, Z. He, A. Cichocki, and T. Ristaniemi, "Fast and effective model order selection method to determine the number of sources in a linear transformation model," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 1870–1874, Aug 2012.

[25] P. Fröhlich, W. Nejdl, K. Jobmann, and H. Wietgrefe, "Model-based alarm correlation in cellular phone networks," in *Proceedings of the 5th International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems*, MASCOTS '97, (Washington, DC, USA), pp. 197–, IEEE Computer Society, 1997.

[26] H. Wietgrefe, K. dieter Tuchs, K. Jobmann, G. Carls, P. Frohlich, W. Nejdl, and S. Steinfeld, "Using neural networks for alarm correlation in cellular phone networks," in *In Proc. International Workshop on Applications of Neural Networks in Telecommunications*, 1997.

[27] S. Wallin, V. Leijon, and L. Landen, "Statistical analysis and prioritisation of alarms in mobile networks," *Int. J. Bus. Intell. Data Min.*, vol. 4, pp. 4–21, May 2009.

[28] 3GPP TS 32.111-1 V12.0.0, *TSG Services and System Aspects; Telecommunication Management; Fault Management; Part 1: 3G fault management requirements*, June 2013.

[29] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA Based Radio Access*. John Wiley & Sons, 2009.

[30] S. Sesia, I. Toufic, and M. Baker, *LTE - The UMTS Long Term Evolution from Theory to Practice, Second Edition*. John Wiley & Sons, 2011.

[31] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, "A cell outage detection algorithm using neighbor cell list reports," in *Self-Organizing Systems* (K. Hummel and J. Sterbenz, eds.), vol. 5343 of *Lecture Notes in Computer Science*, pp. 218–229, Springer Berlin Heidelberg, 2008.

[32] B. Cheung, S. G. Fishkin, G. N. Kumar, and S. A. Rao, "Method of monitoring wireless network performance," March 2006. US Patent 2006/0063521 A1, CN1753541A, EP1638253A1.

[33] F. Chernogorov, "Detection of sleeping cells in long term evolution mobile networks," Master's thesis, University of Jyväskylä, Finland, 2010.

[34] S. Vadlamudi, "System and method of detecting a sleeping cell and remedying detected conditions in a telecommunication network," January 2012. US Patent 8095075.

[35] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. New York, NY, USA: Cambridge University Press, 1st ed., 2009.

[36] M. Rumney, *LTE and the Evolution to 4G Wireless: Design and Measurement Challenges*. Agilent Technologies, John Wiley & Sons, 2009.

[37] I. Kostanic, N. Mijatovic, and S. Vest, "Measurement based qos comparison of cellular communication networks," in *Communications Quality and Reliability, 2009. CQR 2009. IEEE International Workshop Technical Committee on*, pp. 1–5, May 2009.

[38] F. O. Alomary and I. Kostanic, "Evaluation of quality of service in 4th generation (4g) long term evolution (LTE) cellular data networks," *Universal Journal of Communications and Network, 1 , 110 - 117*, vol. 1, pp. 110–117, November 2013.

[39] F. O. Alomary, *A Methodology for Quality of Service Evaluation in 4th Generation (4G) Long Term Evolution (Lte) of Cellular Data Networks*. Melbourne, FL, USA: Florida Institute of Technology, 2013.

[40] 3GPP TS 52.008 V12.0.0, *TSG Services and System Aspects; Telecommunication management; GSM subscriber and equipment trace (Release 12)*, October 2014.

[41] 3GPP TS 32.422 V12.3.0, *TSG Services and System Aspects; Telecommunication management; Subscriber and equipment trace; Trace control and configuration management (Release 12)*, September 2014.

[42] 3GPP TS 32.421 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Subscriber and equipment trace; Trace concepts and requirements (Release 12)*, October 2014.

[43] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE, 5th edition*. Wiley, 2010.

[44] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*. Wiley Publishing, 2nd ed., 2011.

[45] 3GPP TS 32.423 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Subscriber and equipment trace; Trace data definition and management (Release 12)*, September 2014.

[46] 3GPP TS 32.425 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 12)*, June 2013.

[47] 3GPP TS 32.401 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Concept and requirements (Release 12)*, October 2014.

[48] 3GPP TS 32.403 V7.1.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements - UMTS and combined UMTS/GSM (Release 7)*, December 2005.

[49] 3GPP TS 32.404 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements; Definitions and template (Release 12)*, October 2014.

[50] 3GPP TS 32.405 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Universal Terrestrial Radio Access Network (UTRAN) (Release 12)*, October 2014.

[51] 3GPP TS 32.406 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements; Core Network (CN) Packet Switched (PS) domain (Release 12)*, October 2014.

[52] 3GPP TS 32.406 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements; Core Network (CN) Circuit Switched (CS) domain; UMTS and combined UMTS/GSM (Release 12)*, October 2014.

[53] 3GPP TS 52.402 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Performance Management (PM); Performance measurements - GSM (Release 12)*, October 2014.

[54] F. Lesher, "Use cases related to self organizing network, overall description," December 2008.

[55] 3GPP TS 25.215 V12.0.0, *TSG Radio Access Network; Physical layer; Measurements (FDD) (Release 12)*, September 2014.

[56] 3GPP TS 36.214 V12.0.0, *TSG Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements (Release 12)*, September 2014.

[57] 3GPP TS 36.133 V12.5.0, *TSG Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management (Release 12)*, September 2014.

[58] 3GPP TS 36.300 V12.3.0, *TSG Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 12)*, September 2014.

[59] J. Puttonen, J. Turkka, O. Alanen, and J. Kurjenniemi, "Coverage optimization for minimization of drive tests in LTE with extended RLF reporting," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on*, pp. 1764–1768, Sept 2010.

[60] J. Puttonen, E. Virtej, I. Keskitalo, and E. Malkamaki, "On LTE performance trade-off between connected and idle states with always-on type applications," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2012 IEEE 23rd International Symposium on*, pp. 981–985, Sept 2012.

[61] 3GPP TS 25.133 V12.5.0, *TSG Radio Access Network; Requirements for support of radio resource management (FDD) (Release 12)*, September 2014.

[62] 3GPP TS 32.410 V12.0.0, *TSG Services and System Aspects; Telecommunication management; Key Performance Indicators (KPI) for UMTS and GSM(Release 12)*, October 2014.

[63] G. Bulmer, *Principles of Statistics*. Dover Books on Mathematics Series, Dover Publications, 1979.

[64] R. Kirk, *Statistics: An Introduction*. International student edition, Cengage Learning, 2007.

[65] ITU-T E.800, *Telecommunication standardization sector of ITU Series E: Overall network operation, telephone service, service operation and human factors: Quality of telecommunication services: concepts, models, objectives and dependability planning – Terms and definitions related to the quality of telecommunication services. Definitions of terms related to quality of service*, September 2009.

[66] ITU-T E.860, *Telecommunication standardization sector of ITU Series E: Overall network operation, telephone service, service operation and human factors: Quality of telecommunication services: concepts, models, objectives and dependability planning – Use of quality of service objectives for planning of telecommunication networks. Framework of a service level agreement*, June 2002.

[67] 3GPP TS 32.451 V12.0.0, *TSG Services and System Aspects; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Requirements (Release 12)*, October 2014.

[68] 3GPP TS 32.451 V12.0.0, *TSG Services and System Aspects; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Definitions (Release 12)*, October 2014.

[69] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*. Wiley, 2004.

[70] ITU-T P.800, *Telecommunication standardization sector of ITU Series P: Telephone transmission quality, telephone installations, local line networks: Methods for subjective determination of transmission quality*, August 1996.

[71] ITU-T P.862, *Telecommunication standardization sector of ITU Series P: Telephone transmission quality, telephone installations, local line networks: Methods for objective and subjective assessment of quality Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, November 2005.

[72] K. Krishnamoorthy, *Handbook of Statistical Distributions with Applications*. Statistics: A Series of Textbooks and Monographs, Taylor & Francis, 2006.

[73] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, pp. 41–50, Jan. 2003.

[74] I. Corp., "An architectural blueprint for autonomic computing," Oct. 2004.

[75] C. Prehofer and C. Bettstetter, "Self-organization in communication networks: principles and design paradigms," *Communications Magazine, IEEE*, vol. 43, pp. 78–85, July 2005.

[76] O. N. Østerbo and O. Grondalen, "Benefits of self-organizing networks (SON) for mobile operators," *Journal Computer Networks and Communications*, vol. 2012, 2012.

[77] K. N. Premnath and S. Rajavelu, "Challenges in self organizing networks for wireless telecommunications," in *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*, pp. 1331–1334, June 2011.

[78] D. Soldani, G. Alford, F. Parodi, and M. Kylväjä, "An autonomic framework for self-optimizing next generation mobile networks," in *IEEE World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, June 2007.

[79] Z. Altman, M. Amirijoo, *et al.*, "On design principles for self-organizing network functions," in *11th international Symposium on Wireless Communication Systems (ISWCS)*, August 2014.

[80] T. Bandh, H. Sanneck, and R. Romeikat, "An experimental system for SON function coordination," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pp. 1–2, May 2011.

[81] T. Kurner, M. Amirijoo, *et al.*, "D5.9 final report on self-organisation and its implications in wireless access networks," tech. rep., The SOCRATES (Self-Optimisation and self-ConfiguRATion in wirelEss networkS) FP7 project (INFSO-ICT-216284), 2010.

[82] L. Schmelz, M. Amirijoo, A. Eisenblaetter, R. Litjens, M. Neuland, and J. Turk, "A coordination framework for self-organisation in LTE networks," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, pp. 193–200, May 2011.

[83] 3GPP TS 32.501 V12.1.0, *Technical Specification Group Services and System Aspects; Telecommunication management; Self-configuration of network elements; Concepts and requirements (Release 12)*, December 2013.

[84] A. Simonsson and A. Furuskar, "Uplink power control in lte - overview and performance, subtitle: Principles and benefits of utilizing rather than compensating for sinr variations," in *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pp. 1–5, Sept 2008.

[85] I. Viering, B. Wegmann, A. Lobinger, A. Awada, and H. Martikainen, "Mobility robustness optimization beyond doppler effect and wss assumption," in *Wireless Communication Systems (ISWCS), 2011 8th International Symposium on*, pp. 186–191, Nov 2011.

[86] O. Yilmaz, J. Hämälainen, and S. Hämälainen, "Optimization of adaptive antenna system parameters in self-organizing LTE networks," *Wireless Networks*, vol. 19, no. 6, pp. 1251–1267, 2013.

[87] M. Amirijoo, L. Jorguseski, R. Litjens, and L. Schmelz, "Cell outage compensation in LTE networks: Algorithms and performance assessment," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pp. 1–5, May 2011.

[88] M. Amirijoo, L. Jorguseski, R. Litjens, and R. Nascimento, "Effectiveness of cell outage compensation in LTE networks," in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pp. 642–647, January 2011.

[89] 3GPP TR 32.827 V10.1.0, *TSG Services and System Aspects; Telecommunication management; Integration of device management information with Itf-N (Release 10)*, June 2010.

[90] W. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, and B. Sebire, "Minimization of drive tests solution in 3GPP," *Communications Magazine, IEEE*, vol. 50, pp. 28–36, June 2012.

[91] J. Johansson, W. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3GPP release 11," *Communications Magazine, IEEE*, vol. 50, no. 11, pp. 36–43, 2012.

[92] 3GPP TS 36.331 V12.3.0, *TSG Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 12)*, September 2014.

[93] Ericsson White Paper, *Positioning with LTE: Maximizing performance through integrated solutions*, September 2011.

[94] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[95] O. Vermesan and P. Friess, *Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems*. Rajeev Ranjan Prasad, 2013.

[96] H. Chaouchi, *The Internet of Things: Connecting Objects*. ISTE, Wiley, 2013.

[97] C. Systems, "Cisco visual networking index: Global mobile data traffic forecast update 2014–2019 white paper." https://gsmaintelligence.com/research/2014/12/understanding-5g/451/, February 2014.

[98] G. Intelligence, "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, December 2014.

[99] D. J. Hand, P. Smyth, and H. Mannila, *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001.

[100] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and T. Widener, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27–34, 1996.

[101] C. Enders, *Applied Missing Data Analysis*. Methodology in the social sciences, Guilford Press, 2010.

[102] J. Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. SAGE Publications, 2012.

[103] J. Han and M. Kamber, *Data Mining: Concepts and Techniques, Second edition*, vol. 54. Morgan Kaufmann, 2006.

[104] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[105] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J*, vol. 29, no. 2, pp. 147–160, 1950.

[106] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. ITPro collection, Cambridge University Press, 2006.

[107] C. Aggarwal and C. Zhai, *Mining Text Data*. Springer, 2012.

[108] M. Zaki and W. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[109] T. Runkler, *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. SpringerLink : Bücher, Vieweg+Teubner Verlag, 2012.

[110] J. Turkka, *Aspects of Knowledge Mining on Minimizing Drive Tests in Self-organizing Cellular Networks*. PhD thesis, Tampereen teknillinen yliopisto - Tampere University of Technology, Tampere, Finland, August 2014.

[111] C. Burges, *Dimension Reduction: A Guided Tour*. Foundations and trends in machine learningl, Now Publishers, 2010.

[112] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st ed., 2007.

[113] N. Rabin, *Data mining dynamically evolving systems via Diffusion methodologies*. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel, April 2010.

[114] E. Kassis, "Anomaly-based error detection in base station data," Master's thesis, Tel-Aviv University, Israel, 2010.

[115] F. Chernogorov, K. Brigatti, T. Ristaniemi, and S. Chernov, "N-gram analysis for sleeping cell detection in LTE networks," in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[116] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[117] Nagao, Makoto, Mori, and Shinsuke, "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*, COLING '94, (Stroudsburg, PA, USA), pp. 611–615, Association for Computational Linguistics, 1994.

[118] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175, 1994.

[119] M. Haidar and D. O'Shaughnessy, "Topic n-gram count language model adaptation for speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pp. 165–169, 2012.

[120] A. Islam and D. Inkpen, "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, 24-27 2009.

[121] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman, "Comparative n-gram analysis of whole-genome protein sequences," in *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, (San Francisco, CA, USA), pp. 76–81, Morgan Kaufmann Publishers Inc., 2002.

[122] J. Choi, H. Kim, C. Choi, and P. Kim, "Efficient malicious code detection using n-gram analysis and svm," in *NBiS* (L. Barolli, F. Xhafa, and M. Takizawa, eds.), pp. 618–621, IEEE Computer Society, 2011.

[123] G. David, *Anomaly Detection and Classification via Diffusion Processes in Hyper-Networks*. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel, March 2009.

[124] J. Riordan, *Introduction to Combinatorial Analysis*. Dover Books on Mathematics, Dover Publications, 2002.

[125] G. Dong and J. Pei, *Sequence Data Mining*. Advances in Database Systems, Springer, 2007.

[126] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st ed., 2011.

[127] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2008.

[128] C. C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory — ICDT 2001* (J. Van den Bussche and V. Vianu, eds.), vol. 1973 of *Lecture Notes in Computer Science*, pp. 420–434, Springer Berlin Heidelberg, 2001.

[129] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?," in *Proceedings of the 22Nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, (Berlin, Heidelberg), pp. 482–500, Springer-Verlag, 2010.

[130] L. Van der Maaten, E. Postma, and H. Van den Herik, "Dimensionality reduction: A comparative review," *Technical Report TiCC TR 2009-005*, 2009.

[131] I. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, Springer, 2002.

[132] C. Bartenhagen, H.-U. Klein, C. Ruckert, X. Jiang, and M. Dugas, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data," *BMC Bioinformatics*, vol. 11, no. 1, p. 567, 2010.

[133] F. Tsai, "Comparative study of dimensionality reduction techniques for data visualization," *Journal of Artificial Intelligence*, vol. 3, pp. 119–134, 2010.

[134] S. Sadkhan Al Maliky, *Multidisciplinary Perspectives in Cryptology and Information Security*. IGI Global, 2014.

[135] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control, Wiley, 2004.

[136] M. ling Shyu, S. ching Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," in *in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03*, pp. 172–179, 2003.

[137] K. Q. Weinberger and L. K. Saul, "An introduction to nonlinear dimensionality reduction by maximum variance unfolding," in *Proceedings of the 21st*

National Conference on Artificial Intelligence - Volume 2, AAAI'06, pp. 1683–1686, AAAI Press, 2006.

[138] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5 – 30, 2006.

[139] T. Sipola, A. Juvonen, and J. Lehtonen, "Anomaly detection from network logs using diffusion maps.," in *EANN/AIAI (1)* (L. S. Iliadis and C. Jayne, eds.), vol. 363 of *IFIP Advances in Information and Communication Technology*, pp. 172–181, Springer, 2011.

[140] G. David, A. Averbuch, and R. Coifman, "Hierarcial clustering via localized diffusion folders," in *Association for the Advancement of Artificial Intelligence (AAAI)*, November 2010.

[141] T. Sipola, A. Juvonen, and J. Lehtonen, "Anomaly detection from network logs using diffusion maps," in *Engineering Applications of Neural Networks* (L. Iliadis and C. Jayne, eds.), vol. 363 of *IFIP Advances in Information and Communication Technology*, pp. 172–181, Springer Berlin Heidelberg, 2011.

[142] A. Juvonen and T. Sipola, "Adaptive framework for network traffic classification using dimensionality reduction and clustering," in *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2012 4th International Congress on*, pp. 274–279, Oct 2012.

[143] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 113 – 127, 2006. Special Issue: Diffusion Maps and Wavelets.

[144] G. David and A. Averbuch, "Spectralcat: Categorical spectral clustering of numerical and nominal data," *Pattern Recogn.*, vol. 45, pp. 416–433, Jan. 2012.

[145] F. Chung, *Spectral Graph Theory*. No. no. 92 in CBMS Regional Conference Series, Conference Board of the Mathematical Sciences, 1997.

[146] R. Horn and C. Johnson, *Matrix Analysis*. Matrix Analysis, Cambridge University Press, 2012.

[147] S. Lafon, Y. Keller, and R. Coifman, "Data fusion and multicue data matching by diffusion maps," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 1784–1797, Nov 2006.

[148] A. Schclar, "A diffusion framework for dimensionality reduction," in *Soft Computing for Knowledge Discovery and Data Mining* (O. Maimon and L. Rokach, eds.), pp. 315–325, Springer US, 2008.

[149] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[150] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, pp. 21–27, January 1967.

[151] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[152] R. D. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *Information Theory, IEEE Transactions on*, vol. 27, pp. 622–627, Sep 1981.

[153] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *In Proceedings of the eighth SIAM International Conference on Data Mining*, pp. 243–254, 2008.

[154] C. Jin, X. Li, Y. Lee, G. Pok, and K. Ryu, "A new approach for calculating similarity of categorical data," in *Convergence and Hybrid Information Technology* (G. Lee, D. Howard, and D. Slezak, eds.), vol. 206 of *Communications in Computer and Information Science*, pp. 584–590, Springer Berlin Heidelberg, 2011.

[155] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '02, (London, UK, UK), pp. 15–26, Springer-Verlag, 2002.

[156] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Rec.*, vol. 29, pp. 427–438, May 2000.

[157] M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *PROCEEDINGS OF THE 2000 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA*, pp. 93–104, ACM, 2000.

[158] J. Tang, Z. Chen, A. W.-C. Fu, and D. W.-L. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, (London, UK, UK), pp. 535–548, Springer-Verlag, 2002.

[159] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," in *19th International Conference on Data Engineering*, pp. 315–326, 2003.

[160] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, (Washington, DC, USA), pp. 430–433, IEEE Computer Society, 2004.

[161] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2007.

[162] D. Banks and I. F. of Classification Societies. Conference, *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15-18 July 2004*. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, 2004.

[163] J. Wu, *Advances in K-means Clustering: A Data Mining Thinking*. Springer Theses, Springer, 2012.

[164] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[165] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[166] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-1, pp. 224–227, April 1979.

[167] Z. He, X. Xu, and S. Deng, "Discovering cluster based local outliers," *Pattern Recognition Letters*, vol. 2003, pp. 9–10, 2003.

[168] H. Zengyou, X. Xiaofei, and D. Shengchun, "Squeezer: An efficient algorithm for clustering categorical data," *J. Comput. Sci. Technol.*, vol. 17, pp. 611–624, Aug. 2002.

[169] A. Pathan, *The State of the Art in Intrusion Prevention and Detection*. CRC Press, 2015.

[170] J. Han, M. Kamber, and J. Pei, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2006.

[171] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The databoost-im approach," *SIGKDD Explor. Newsl.*, vol. 6, pp. 30–39, June 2004.

[172] G. M. Weiss, "Mining with rarity: A unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, pp. 7–19, June 2004.

[173] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *International Conf. on Machine Learning*, 2003.

[174] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), pp. 475–480, ACM, 2002.

[175] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, (New York, NY, USA), pp. 169–178, ACM, 2000.

[176] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.

[177] S. Novaczki, "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, March 2013.

[178] G. Ciocarlie, U. Lindqvist, S. Novaczki, and H. Sanneck, "Detecting anomalies in cellular networks using an ensemble method," in *Network and Service Management (CNSM), 2013 9th International Conference on*, pp. 171–174, Oct 2013.

[179] G. Ciocarlie, U. Lindqvist, K. Nitz, S. Novaczki, and H. Sanneck, "On the feasibility of deploying cell anomaly detection in operational cellular networks," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pp. 1–6, May 2014.

[180] G. Ciocarlie, U. Lindqvist, K. Nitz, S. Novaczki, and H. Sanneck, "Dcad: Dynamic cell anomaly detection for operational cellular networks," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, pp. 1–2, May 2014.

[181] G. Ciocarlie, C.-C. Cheng, C. Connolly, U. Lindqvist, K. Nitz, S. Novaczki, H. Sanneck, and M. Naseer-ul Islam, "Anomaly detection and diagnosis for automatic radio network verification," in *6th International Conference on Mobile Networks and Management, MONAMI 2014*, September 2014.

[182] G. Ciocarlie, C.-C. Cheng, C. Connolly, U. Lindqvist, S. Novaczki, H. Sanneck, and M. Naseer-ul Islam, "Managing scope changes for cellular network-level anomaly detection," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pp. 375–379, Aug 2014.

[183] G. Ciocarlie, C.-C. Cheng, C. Connolly, U. Lindqvist, K. Nitz, S. Novaczki, H. Sanneck, and M. Naseer-ul Islam, "Demo: SONver: SON verification for operational cellular networks," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pp. 611–612, Aug 2014.

[184] M. Kac, J. Kiefer, and J. Wolfowitz, "On tests of normality and other tests of goodness of fit based on distance methods," *The Annals of Mathematical Statistics*, vol. 26, no. 2, pp. pp. 189–211, 1955.

[185] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[186] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, pp. 77–84, Apr. 2012.

[187] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[188] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in *In Advances in Neural Information Processing Systems*, pp. 1385–1392, MIT Press, 2005.

[189] M. Richardson and P. Domingos, "Markov logic networks," *Mach. Learn.*, vol. 62, pp. 107–136, Feb. 2006.

[190] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pp. 752–760, 2011.

[191] J. Turkka, T. Ristaniemi, G. David, and A. Averbuch, "Anomaly detection framework for tracing problems in radio networks," in *The 10th International Conference on Networks, ICN 2011*, 2011.

[192] R. Khanafer, B. Solana, J. Triola, R. Barco, L. Moltsen, Z. Altman, and P. Lazaro, "Automated diagnosis for umts networks using bayesian network approach," *Vehicular Technology, IEEE Transactions on*, vol. 57, pp. 2451–2461, July 2008.

[193] R. Barco, P. Lazaro, L. Diez, and V. Wille, "Continuous versus discrete model in autodiagnosis systems for wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 7, pp. 673–681, June 2008.

[194] R. Barco, P. Lazaro, V. Wille, L. Diez, and S. Patel, "Knowledge acquisition for diagnosis model in wireless networks," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 4745 – 4752, 2009.

[195] R. Barco, V. Wille, L. Diez, and M. Toril, "Learning of model parameters for fault diagnosis in wireless networks," *Wireless Networks*, vol. 16, no. 1, pp. 255–271, 2010.

[196] K. Raivio, O. Simula, J. Laiho, and P. Lehtimaki, "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, March 2003.

[197] J. Laiho, K. Raivio, P. Lehtimaki, K. Hatonen, and O. Simula, "Advanced analysis methods for 3g cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 4, pp. 930–942, May 2005.

[198] R. Barco, L. Nielsen, R. Guerrero, G. Hylander, and S. Patel, "Automated troubleshooting of a mobile communication network using bayesian networks," in *Mobile and Wireless Communications Network, 2002. 4th International Workshop on*, pp. 606–610, 2002.

[199] R. Barco, V. Wille, and L. Diez, "System for automated diagnosis in cellular networks based on performance indicators," *European Transactions on Telecommunications*, vol. 16, no. 5, pp. 399–409, 2005.

[200] R. Barco, V. Wille, L. Diez, and P. Lazaro, "Comparison of probabilistic models used for diagnosis in cellular networks," in *Vehicular Technology Conference, 2006. VTC 2006-Spring. IEEE 63rd*, vol. 2, pp. 981–985, May 2006.

[201] R. Barco, L. Diez, V. Wille, and P. Lazaro, "Automatic diagnosis of mobile communication networks under imprecise parameters," *Expert Systems with Applications*, vol. 36, no. 1, pp. 489 – 500, 2009.

[202] J. You, "Research of wireless network fault diagnosis based on bayesian networks," in *Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on*, vol. 3, pp. 59–64, Nov 2009.

[203] A. Coluccia, F. Ricciato, and P. Romirer-Maierhofer, "Bayesian estimation of network-wide mean failure probability in 3g cellular networks," in *Proceedings of the 2010 IFIP WG 6.3/7.3 International Conference on Performance Evaluation of Computer and Communication Systems: Milestones and Future Challenges*, PERFORM'10, (Berlin, Heidelberg), pp. 167–178, Springer-Verlag, 2011.

[204] G. Barreto, J. Mota, L. Souza, R. Frota, L. Aguayo, J. Yamamoto, and P. Macedo, "Competitive neural networks for fault detection and diagnosis in 3g cellular systems," in *Telecommunications and Networking - ICT 2004* (J. de Souza, P. Dini, and P. Lorenz, eds.), vol. 3124 of *Lecture Notes in Computer Science*, pp. 207–213, Springer Berlin Heidelberg, 2004.

[205] G. A. Barreto, J. C. Mota, L. G. Souza, R. A. Frota, L. Aguayo, J. S. Yamamoto, and P. E. Macedo, "A new approach to fault detection and diagnosis in cellular systems using competitive learning," in *Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN04)*, 2004.

[206] P. Kumpulainen and K. Hatonen, "Local anomaly detection for mobile network monitoring," *Information Sciences*, vol. 178, no. 20, pp. 3840 – 3859,

2008. Special Issue on Industrial Applications of Neural Networks 10th Engineering Applications of Neural Networks 2007.

[207] E. Barrett *et al.*, "Deliverable D4.1: Specification of knowledge-based reasoning algorithms," tech. rep., The COMMUNE (COgnitive network ManageMent under UNcErtainty) Project (Celtic/2011-2014 CP08-004), 2012.

[208] S. Caban, C. Mehlführer, M. Rupp, and M. Wrulich, *Evaluation of HSDPA and LTE: From Testbed Measurements to System Level Performance*. UK: John Wiley & Sons, 2012.

[209] M. Chuah and Q. Zhang, *Design and Performance of 3G Wireless Networks and Wireless LANs*. Springer, 2005.

[210] F. Laakso, *Studies on High Speed Uplink Packet Access Performance Enhancements*. PhD thesis, Jyväskylän Yliopisto - University of Jyväskylä, Jyväskylä, Finland, December 2014.

[211] S. Hämälainen, *WCDMA Radio Network Performance*. PhD thesis, Jyväskylän Yliopisto - University of Jyväskylä, Jyväskylä, Finland, January 2003.

[212] ns 3, "Network simulator 3," 2014.

[213] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji, "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, sept. 2005.

[214] 3GPP contribution R1-070674, *TSG RAN WG1: LTE Physical Layer Framework for performance verification*,, February 2007.

[215] 3GPP contribution R1-071952, *TSG RAN WG1: DL Performance Evaluation for E-UTRA*, April 2007.

[216] 3GPP contribution R1-071956, *TSG RAN WG1: E-UTRA Performance Checkpoint: Downlink*, April 2007.

[217] 3GPP contribution R1-071960, *TSG RAN WG1: LTE Performance Benchmarking*, April 2007.

[218] 3GPP contribution R1-071961, *TSG RAN WG1: LTE Downlink Performance Evaluation Results*, April 2007.

[219] 3GPP contribution R1-071967, *TSG RAN WG1: DL E-UTRA Performance Checkpoint*, April 2007.

[220] 3GPP contribution R1-071969, *TSG RAN WG1: E-UTRA Performance Verification: DL Throughput*, April 2007.

[221] 3GPP contribution R1-071976, *TSG RAN WG1: LTE Downlink System Performance Verification Results*, April 2007.

[222] 3GPP contribution R1-071976, *TSG RAN WG1: LTE Downlink Performance*, April 2007.

[223] 3GPP contribution R1-071981, *TSG RAN WG1: Downlink Best Effort Performance*, April 2007.

[224] 3GPP contribution R1-071987, *TSG RAN WG1: LTE Downlink Performance Verification for Frame Structure Type 2*, April 2007.

[225] N. Kolehmainen, "Downlink packet scheduling performance in evolved universal terrestrial radio access network," Master's thesis, University of Jyväskylä, Finland, 2007.

[226] P. Kela, "Downlink channel quality indication for evolved universal terrestrial radio access network," Master's thesis, University of Jyväskylä, Finland, 2007.

[227] 3GPP TR 36.839 V11.1.0, *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)*, December 2012.

[228] M. Karlstedt, B. Herman, *et al.*, "Deliverable D5.2: Cognitive network simulator," tech. rep., The COMMUNE (COgnitive network ManageMent under UNcErtainty) Project (Celtic/2011-2014 CP08-004), 2013.

APPENDIX 1 COMPARISON OF THE STUDIES IN QUALITY AND PERFORMANCE MANAGEMENT FOR MOBILE NETWORKS

TABLE 7 Comparison of advanced QPM studies, devoted to detection of sleeping cell failures with analysis of sequential characteristics of MDT reports.

Article	[PIII]	[PII]	[PI]
Malfunction	RACH sleeping cell	RACH sleeping cell	RACH sleeping cell
Scenario & Environment	57 cells, LTE system level dynamic simulations, non wrap-around	21 cells, LTE system level dynamic simulations, wrap-around	21 cells, LTE system level dynamic simulations, wrap-around
Analysis target	Detection	Detection	Detection
Analysis level	User level	User level	User level
Collected PM data	Event triggered MDT reports	Event triggered MDT reports	Event triggered MDT reports
Collection time and frequency	142 s	142 s	572 s $\approx$ 10 minutes
Knowledge mining algorithm(-s) **A = Detection;** **B = Diagnosis;** **C = Healing**	A: 2-gram analysis, PCA transform, FindCBLOF, post processing	A: Sliding window pre-processing, 2-gram analysis, MCA transform, K-NN anomaly scoring, enhanced post-processing	A: Sliding window pre-processing, 1-gram vs. 2-gram analysis, MCA transform, K-NN anomaly scoring, enhanced post-processing
Training data	Normal training data	Normal training data	Normal training data
Num. of features	1	1	1
KM architecture	Centralized	Centralized	Centralized

TABLE 8 Comparison of advanced QPM studies, devoted to detection of sleeping cell failures with analysis of numerical characteristics of MDT reports.

Article	[PIV]	[PV]	[PVI]
Malfunction	HW failure - antenna gain malfunction	HW failure - antenna gain malfunction	HW failure - TxP circuitry malfunction
Scenario & Environment	27 macro cells, LTE system level dynamic simulations, non wrap-around	57 macro cells, LTE system level dynamic simulations, non wrap-around	57 macro cells, LTE system level dynamic simulations, non wrap-around
Analysis target	Detection	Detection	Detection
Analysis level	User level	User level	User and cell levels
Collected PM data	MDT event-triggered reports: RSRP, RSRQ, CQI	MDT event-triggered and periodic reports: RSRP, RSRQ, PHR, CQI	MDT event-triggered and periodic reports: RSRP, RSRQ, SINR
Collection time and frequency	100 s	142 s	100 s
Knowledge mining algorithm(-s) **A = Detection;** **B = Diagnosis;** **C = Healing**	A: Transformation with diffusion maps, k-means clustering	A: Transformation with diffusion maps, K-NN classification	A: K-NN anomaly score, cell-specific statistical profiling
Training data	Normal training data	Normal training data	Normal training data
Num. of features	9	10	6
KM architecture	Centralized	Centralized	Hybrid

APPENDIX 2 SYSTEM LEVEL SIMULATIONS FOR QUALITY AND PERFORMANCE MANAGEMENT RESEARCH

Computer simulations are widely used for the development and optimization of new algorithms in cellular mobile networks. Modeling of QPM is not an exception, and many of SON studies in 3GPP are based on simulations. Depending on the scope, simulators can be divided into two major groups: link level and system level [208, 209]. The first type implies modeling of one or several communication links between a user terminal and base station, with detailed implementation of the generated electromagnetic waves and properties of the propagation channel. System level simulations are mostly used for performance evaluation of the whole network. Usual practice is to input the link level results into the system level simulations.

From the perspective of user mobility, system level simulators can be divided into three main groups: static, quasi-static and dynamic as it is discussed in [210, 211]. Static simulations model non-moving users and for that reason are mostly aimed at the evaluation of link budget, coverage estimation, etc. Quasi-static simulations imply no mobility, but the users have traffic models and exist in a time domain. Dynamic system level simulations are closer than others to the real life networks, due to the existence of moving users. Additionally, more realistic propagation conditions are considered, as such effects as time dependent slow and fast fading are taken into account. Usually there is also a broad range of propagation models, radio resource management mechanisms, mobility and network control signaling is taken into account. Thus, dynamic simulators are the most complex, if compared to other simulator types, and provide the most accurate results regarding network performance. However, due to high computational complexity, dynamic simulations may last very long and the simulated network operation time is usually in extent of minutes. Thus, consideration of the long term effects in the network calls for either simplification in modeling of the simulator, or utilization of demo systems or emulators, partly based on real equipment.

To gather the necessary performance monitoring data for the research activities discussed in this dissertation, two LTE dynamic system level simulators have been used: FREAC and Network Simulator 3 (ns-3) [212]. FREAC is a proprietary step-based simulator, which models E-UTRAN system with a resolution of 1 Orthogonal Frequency-Division Multiplexing (OFDM) symbol in the time domain. Both uplink and downlink directions can be simulated with this simulator. Mapping of the link level SINR to the system level is done according to the framework presented in [213]. Thus, the appropriate modulation and coding schemes can be selected to guarantee the correct block error rate. FREAC implements a large variety of radio resource management algorithms, scheduling, link adaptation, HARQ, different mobility and traffic models. The most important measurements, such as RSRP, RSRQ, RSSI and CQI are modeled. More information regarding assumptions for implementation of these measurements

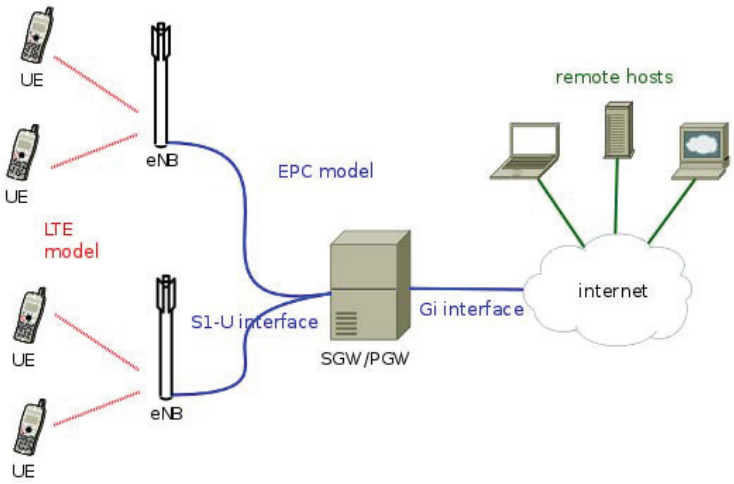


FIGURE 26 Models in LENA module of ns-3.

and FREAC in general can be found in [110]. One of the key features in FREAC for QPM research presented in this thesis is MDT. A measurement log collected from dynamic simulations, can be filled with both event-triggered and periodic reports. Moreover, RRC signaling during mobility and connection establishment procedures are modeled and the corresponding messages, e.g. HO command and complete, A2, A3, etc., can be gathered.

Results produced by FREAC are validated with several calibration campaigns in 3GPP, where companies provide certain statistics for the agreed scenarios [214]. Throughput and spectral efficiency is verified using the results from [214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224]. Comparative analysis of FREAC and simulators from other companies can be found in [225, 226]. Mobility related performance of LTE networks evaluated with FREAC and calibrated against results from simulators of other companies can be found in [227].

LTE model in system level simulator ns-3 is also used in this thesis, for collection of user specific measurement statistics and the validation of the cognitive self-healing function. So called LENA module of ns-3 implements both LTE and enhanced packet core system of E-UTRAN, as it is show in Figure 26. User mobility models, handover procedure and a user measurement log have been implemented [228] to gather the necessary data for the advanced QPM research, carried out with COMMUNE project.

ORIGINAL PAPERS

PI

SEQUENCE-BASED DETECTION OF SLEEPING CELL FAILURES IN MOBILE NETWORKS

by

Fedor Chernogorov, Sergey Chernov, Kimmo Brigatti, Tapani Ristaniemi

Wireless Networks, The Journal of Mobile Communication, Computation and Information (submitted for review, available on arxiv.org), 2015

http://arxiv.org/abs/1501.03935v2

Noname manuscript No.
(will be inserted by the editor)

Sequence-based Detection of Sleeping Cell Failures in Mobile Networks

Fedor Chernogorov · Sergey Chernov · Kimmo Brigatti · Tapani Ristaniemi

Received: date / Accepted: date

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 (UMTS) 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 (MDT) functionality standardized in LTE. Further processing applies dimensionality reduction, anomaly detection with K-Nearest Neighbors (K-NN), cross-validation, post-processing 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 (ROC) 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 Technologys (RATs), and their cor-

F. Chernogorov
Magister Solution Ltd., Sepankatu 14 C, FIN-40720, Jyvaskyla, Finland
University of Jyvaskyla, Department of Mathematical Information Technology, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland E-mail: fedor.chernogorov[at]magister.fi, fedor.chernogorov[at]jyu.fi

S. Chernov, K. Brigatti and Tapani Ristaniemi
University of Jyvaskyla, Department of Mathematical Information Technology, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland E-mail: sergey.a.chernov[at]jyu.fi, kimmobrigatti[at]gmail.com, tapani.e.ristaniemi[at]jyu.fi

responding 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 5G networks there will be much more end-user devices, served by different technologies, and connected to cells of different types. 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) [34], which is a sequence of fault detection, diagnosis and healing, should be developed and applied in the network in addition to other optimization functions.

Concept of Self-Organizing Network (SON) [52, 53] 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, 34, 60]. 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 [59], [31] 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 [15, 14, 34]. 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 Section 4 simulation environment, assumptions and random access channel problem are presented. Also Section 4 describes the generated and analyzed performance MDT data. Section 5 concentrates on the suggested sleeping cell detection knowledge mining framework. It includes overview of the applied anomaly detection methods: K-NN anomaly outlier scores, N-gram, minor component analyses, post-processing 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 post-processing methods is compared. In Section 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 thought 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 3^{rd} Generation Partnership Programme (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 [36]. 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 [34]. 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 [36]. 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

Table 1 Network events triggering MDT measurements and reporting

PL PROBLEM - Physical Layer Problem [30].
RLF - Radio Link Failure [61].
RLF REESTAB. - Connection reestablishment after RLF.
A2 RSRP ENTER - RSRP goes under A2 enter threshold.
A2 RSRP LEAVE - RSRP goes over A2 leave threshold.
A2 RSRQ ENTER - RSRQ goes over A2 enter threshold.
A3 RSRP - A3 event, according to 3GPP specification.
HO COMMAND - handover command received [61].
HO COMPLETE - handover complete received [61].

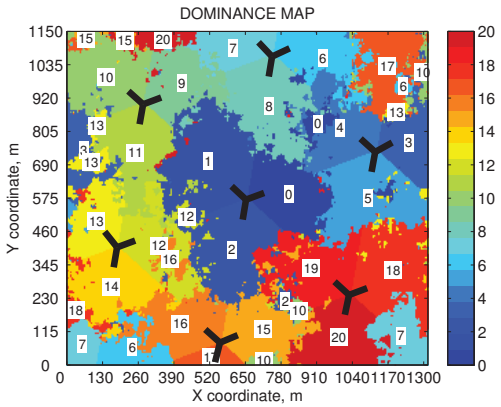


Fig. 1 Wrap around Macro 21 slow faded dominance map

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 NodeB (eNB)[1] with dominating, i.e. strongest radio signal in each point of the network, see Fig. 1. Creation of dominance map requires information about path loss and slow fading. 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 dominance maps approach is that it requires a lot of detailed input measurement information. Though, MDT functionality is one of the ways to create such maps fast and relatively simple. Additionally, more accurate user location information is going to be available with deployment of newer releases of mobile networks [23].

The last method for cell ID and UE report location association uses target cell ID feature. The main advantage of this approach is that it does not require serving

[1] Evolved Universal Terrestrial Radio Access Network (E-UTRAN)

cell ID, user geo-positioning location or knowledge about network dominance areas. This eases the requirements for MDT data collection in amount of details regarding user location. The problem of mapping on the basis of target cell ID, is that it might be useful for detection of only particular types of network problem, such as random access Sleeping Cell (SC). Efficiency of this method for detection of other malfunctions is subject for further verification.

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.

2.2 Advanced data analysis approaches in QPM

Studies in advanced data analysis for QPM can be divided to several groups. In certain studies, the data reported by the users is used for the analysis. For instance, in [50] authors suggest a method for detection of sleeping cells, caused by transmitted signal strength problem, on the basis of neighbor cell list information. Application of non-trivial pre- processing and different classification algorithms allowed to achieve relatively good accuracy in detection of cell hardware faults. However, the proposed anomaly detection system is prone to have relatively high false rate. In [63] a method based on analysis of TRACE-based user data with diffusion maps is presented. More extensive application of diffusion maps for network performance monitoring can also be found in [44].

Even though, user level statistics is more detailed, still majority of studies devoted to improvement of QPM rely on cell-level data. The first proposals of sleeping cell detection automation using statistical methods of network monitoring are presented in [14, 15]. Preparation of normal cell load profile and evaluation of the deviation in observed cell behavior is suggested as a way for identification of problematic cells. The idea of statistical approach has been further studied in [55], [62], [54], where a profile-based system for fault detection and diagnosis is proposed. Bayesian networks have also been applied for diagnosis and root cause probability estimation, given certain KPIs [46, 4, 5, 6]. The complications here are preparation of correct probability model and appropriate KPI threshold parameters. More advanced data mining methods are applied to analysis of cell-level performance statistics, and novel ensemble methods of classification algorithms is proposed [18, 21]. In [19, 20] application of classification and clustering methods for detection and diagnosis of strangely behaving network regions is presented. Some studies also consider neural network algorithms for detection of malfunctions [57, 48].

The largest drawback of processing cell level data is that collection of appropriate statistical base takes substantial amount of time, and can vary from days to months. This increases reaction time in case of outages and does not completely solve the problems of operators in optimization of their QPM. In order to overcome weaknesses of analysis based on cell KPIs, our studies are concentrated at the analysis of the user-level data, collected with immediate MDT functionality [37, 42]. In the early works cell outage detection caused by signal strength problems (antenna gain failure) is studied [11, 12, 64]. This area matches the 3GPP use case called “cell outage detection” [31]. Identification of the cell, in malfunction condition is done by means of analysis of numerical properties of multidimensional

dataset. Each data point represents either periodic or event-triggered user measurement. Such methods as diffusion maps dimensionality reduction algorithm, k-means clustering and k-nearest neighbor classification methods are applied.

To increase robustness of the proposed solutions in MDT data analysis and make the developed detection system suitable for application in real networks, a more sophisticated experimental setup is considered. Sleeping cell caused by malfunction of random access channel, discussed in Section 3, does not produce coverage holes from perspective of radio signal, but still makes service unavailable to the subscribers. This problem is considered to be one of the most complex for mobile network operators, as detection of such failures may take days or even weeks, and negatively affects user experience [34]. To make fault detection framework more flexible and independent from user behavior, such as variable mobility and traffic variation, analysis of numerical characteristics of MDT data is substituted with processing of *network event sequences* with N-gram method. Network events can include different mobility or signaling related nature, such as A2, A3 or handover complete message [40]. Initial results in this area are presented in [13].

3 Sleeping Cell Problem

Sleeping cell is a special kind of cell service degradation. It means malfunction resulting in network performance decrease, invisible for a network operator, but affecting user Quality of Experience (QoE). On one hand, detection of sleeping cell problem with traditional monitoring systems is complicated, as in many cases KPI thresholds do not indicate the problem. On the other hand fault identification can be very sluggish, as creation of cell behavior profile requires long time, as it is discussed in the previous section. Regular, less sophisticated types of failures usually produce cell level alarms to performance monitoring system of mobile network operator. In contrast, for sleeping cells degradation occurs seamlessly and no direct notification is given to the service provider.

In general, any cell can be called degraded in case if it is not 100% functional, i.e. its services are suffering in terms of quality, what in turn affects user experience. Classification of sleeping cells, depending on the extent of performance degradation from the lightest, to the most severe [14],[16]: *impared* 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 cell 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 [2], [65]. RACH malfunction leads to inability of the affected cell to serve any new users, while earlier connected UEs still get served, as pilot signals are transmitted. This problem can be classified to crippled sleeping cell type, and with time it 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) [61]. 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. Thus, whenever UE tries to initiate random access to cell 1, 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 is validated against results from other simulators of several companies in 3GPP [1, 47, 45]. 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 [8]. 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 [10, 35].

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

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	Simulation without sleeping cell
Problematic case	Simulation with RACH problem in cell 1		
A2 RSRP Threshold	-110	A2 RSRP Hysteresis	3
A2 RSRQ Threshold	-10	A2 RSRQ Hysteresis	2

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 the cases are 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 [35]. 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 [25], [35]: data cleaning, integration from different sources, feature selection and extraction, transformation, pattern recognition, pattern evaluation and knowledge presentation. The constructed data analysis framework for sleeping cell detection is semi-supervised, because unlabeled error-free data is used for

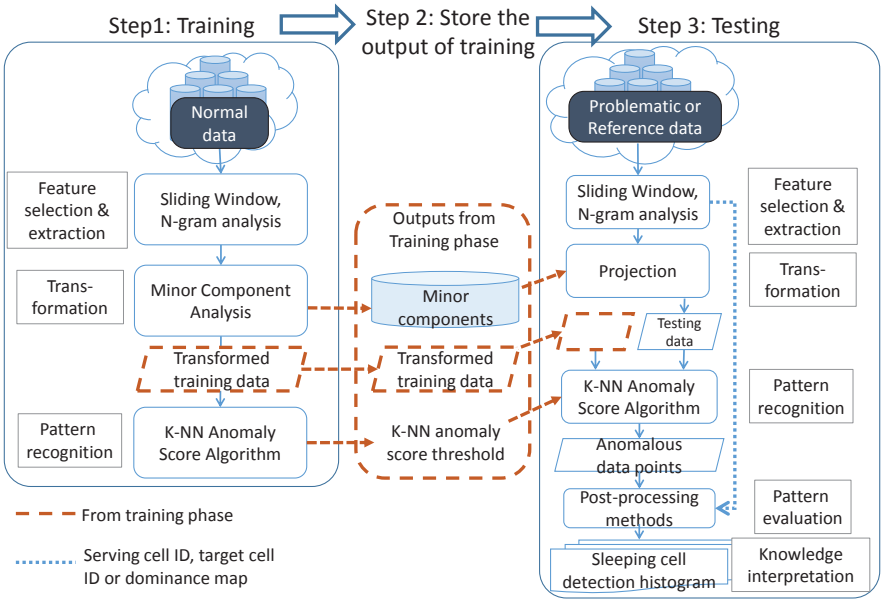


Fig. 2 Sleeping Cell Detection Framework

training of the data mining algorithms. In testing phase problematic data is analyzed to detect abnormal behavior. Reference data is used for testing in 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. Pre-processing 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 Pre-processing

Sliding window approach [56] 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 [44]. The number of calls and sub-calls for all three data sets are shown in Table 3.

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	or	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

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. [7, 51, 9, 33, 41]. In addition, N-gram is applied for whole-genome protein sequences [26] and for computer virus detection [17, 24].

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.

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 [49]. This algorithm has been selected on the basis of comparison with other dimensionality reduction methods such as Principal Component Analysis (PCA) [43] 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 (SORTE) method [38, 39].

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

Table 5 Parameters of algorithms in sleeping cell detection framework

Parameter	**Value**
Number of chunks in K-fold method per dataset	6
Sliding window size	15
Sliding window step	10
N in N-gram algorithm	2
Number of nearest neighbors (k) in K-NN algorithm	35
Number of minor components	6

of distances to k nearest neighbors [3] or distance to k-th neighbor [58]. 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 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 95^{th} 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 post-processing of the anomalous data samples through analysis of their correspondence to particular network elements, such as UEs and cells. 4 post-processing methods are developed for this purpose. 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 post-processing. 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 post-processing 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 SC scores of all non-neighboring cells. Sleeping cell scores, received after post-processing 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 3σ (standard deviation of sleeping cell scores) from the mean SC of 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 post-processing methods enables application of such data mining performance metrics as accuracy, precision, recall, F-score, True Negative Rate (TNR) and False Positive Rate (FPR) [32]. On the basis of these scores ROC curves are plotted.

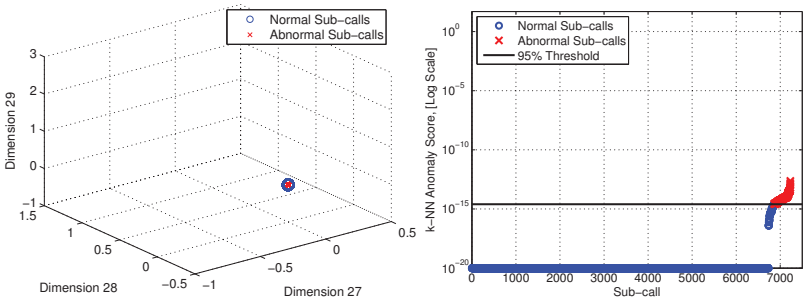
In addition to the conventional performance evaluation metrics described above, a heuristic method is applied to complement the analysis. This approach measures how far is the achieved performance from the *a priori* known ideal solution. Performance of the sleeping cell detection algorithm can be described by a point in the space *"sleeping cell magnitude"*-*"cumulative standard deviation"*. *"Sleeping cell magnitude"* is the highest sleeping cell score, and a sum of all sleeping cell scores is *"cumulative standard deviation"*. This plane contains two points of interest: in case of malfunctioning network, the ideal sleeping cell detection algorithm would have coordinate $[0; 100]$. In case of error-free network, the ideal performance is point $[0; 100/N_{\text{cells in the network}}]$. 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 Pre-processing and K-NN Anomaly Score Calculations

After pre-processing with sliding window and N-gram methods, and transformation with MCA, training MDT data is processed with K-NN anomaly score algorithm. As it is discussed in section 5.3, the anomaly score threshold, used for



(a) Normal training dataset in the embedded space.

(b) Sorted outlier scores of normal training dataset.

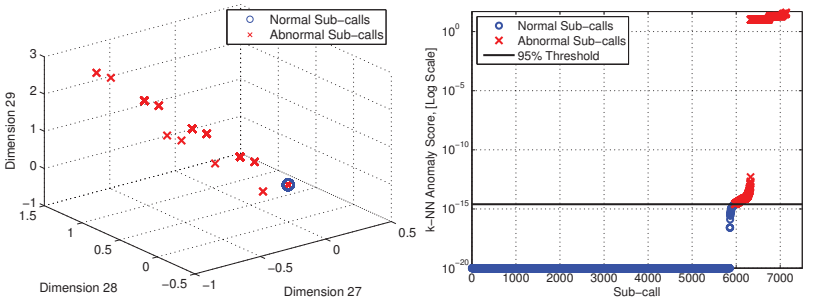
Fig. 3 Normal dataset used for training of the sleeping cell detection framework

separation of data points to normal and abnormal classes, is selected to be 95^{th} percentile of outlier score in training data. Shape of normal training dataset in the embedded space is shown in Fig. 3a, and 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 pre-processing 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 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 Section 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 is low for all points, except 2 outliers.

6.2 Application of Post-Processing Methods for Sleeping Cell Detection

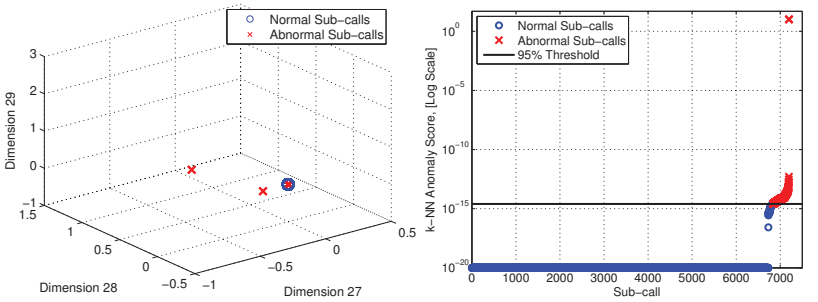
After training and testing phases certain sub-calls are marked as anomalies. The next step is conversion of this information to knowledge about location of malfunctioning cell or cells, and this is done through post-processing described in Section 5.4.



(a) Problem testing dataset in the embedded space

(b) Sorted outlier scores of problem testing dataset

Fig. 4 Problematic dataset used at the testing phase of the sleeping cell detection framework



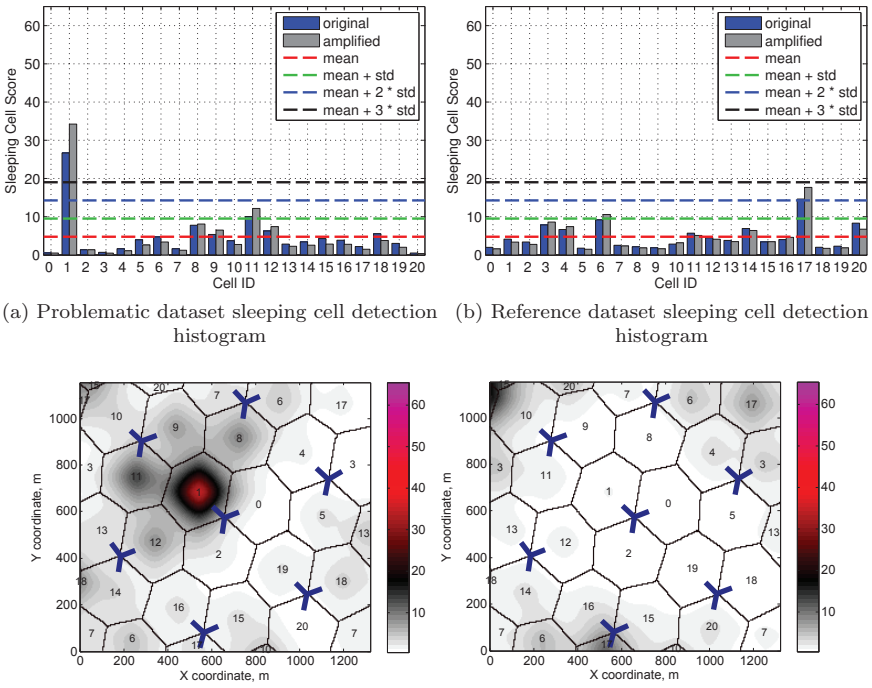
(a) Reference testing dataset in the embedded space

(b) Sorted outlier scores of reference testing dataset

Fig. 5 Reference dataset used at the testing phase of the sleeping cell detection framework

6.2.1 Detection based on Dominance Cell Sub-Call Deviation

In our earlier study [13] post-processing 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 Section 5.1.1. Majority of 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 post-processing step - amplification, described in Section 5.4. In addition to cell 1, its neighboring



(a) Problematic dataset sleeping cell detection histogram

(b) Reference dataset sleeping cell detection histogram

(c) Problematic dataset heat map

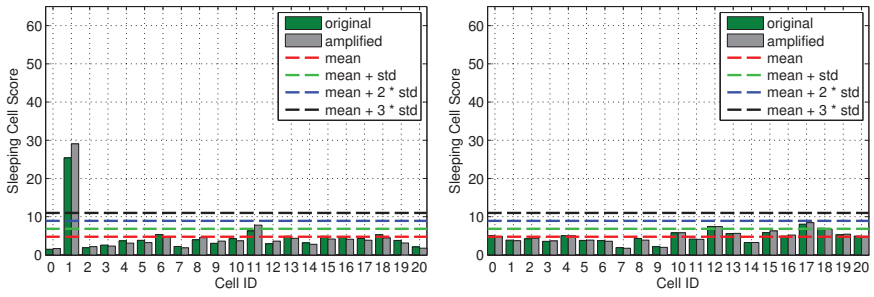
(d) Reference dataset heat map

Fig. 6 Results of sleeping cell detection for Dominance Cell Sub-Call Deviation method

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 and 6d 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.

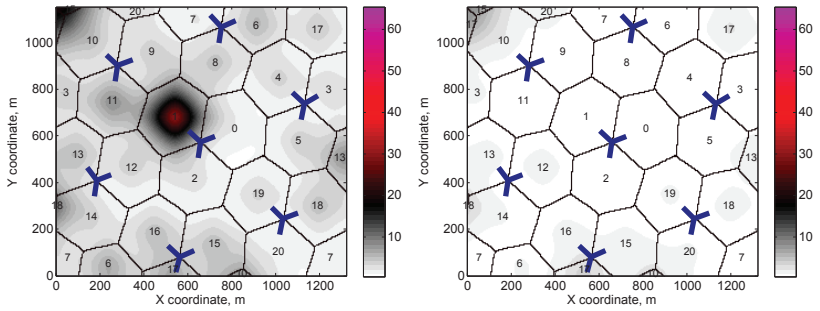
6.2.2 Detection based on Dominance Cell 2-Gram Deviation

In this method problematic network regions are found through comparison of occurrence frequencies, normalized by the total number of users, in training and testing datasets. 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 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 non-



(a) Problematic dataset sleeping cell detection histogram

(b) Reference dataset sleeping cell detection histogram



(c) Problematic dataset heat map

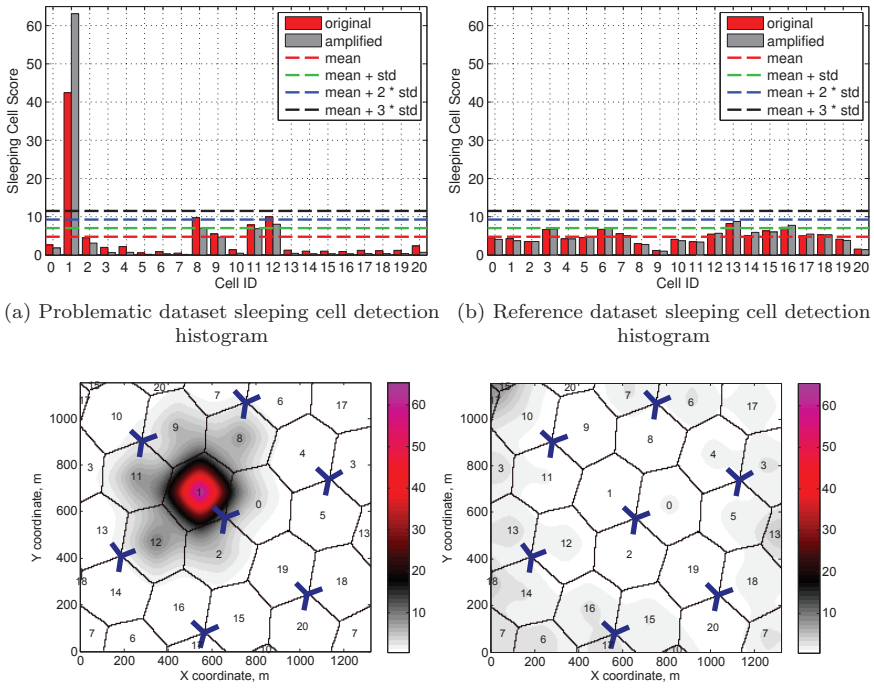
(d) Reference dataset heat map

Fig. 7 Results of sleeping cell detection for Dominance Cell 2-Gram Deviation method

successful 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 post-processing 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 and 7d.

6.2.3 Detection based on Dominance Cell 2-Gram Symmetry Deviation

This post-processing method analyzes the symmetry imbalance of network event 2-grams. 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 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.



(a) Problematic dataset sleeping cell detection histogram

(b) Reference dataset sleeping cell detection histogram

(c) Problematic dataset heat map

(d) Reference dataset heat map

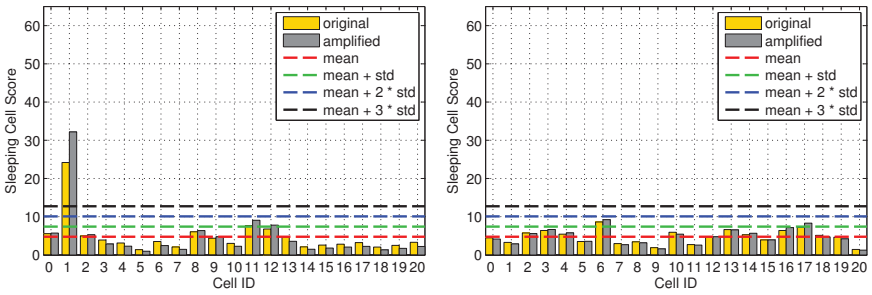
Fig. 8 Results of sleeping cell detection for Dominance Cell 2-Gram Symmetry Deviation method

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 post-processing 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.

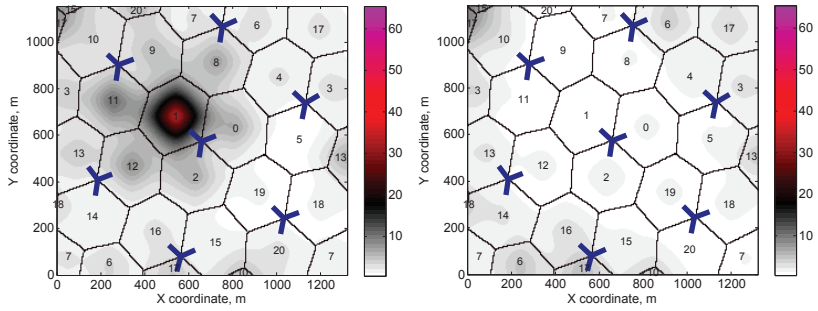
6.2.4 Detection based on Target Cell Sub-Calls

As it is discussed in Section 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 post-processing methods, based on dominance cell deviation. Another shortcoming



(a) Problematic dataset sleeping cell detection histogram

(b) Reference dataset sleeping cell detection histogram



(c) Problematic dataset heat map

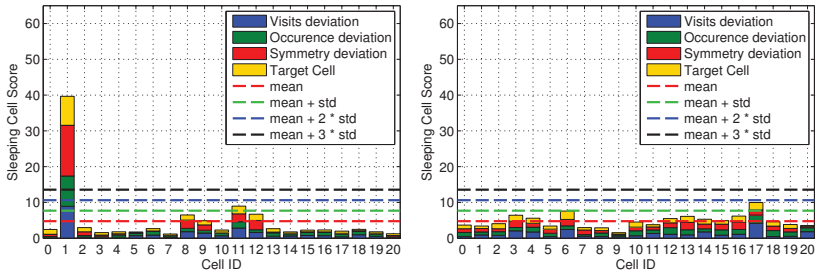
(d) Reference dataset heat map

Fig. 9 Results of sleeping cell detection for Target Cell Sub-Calls method

is that target cell sub-call method is more prone to trigger false alarms. This can be seen from the results when reference data is used as testing, Fig. 9b. Sleeping cell score of cell 6 is reaching threshold of mean plus 2 standard deviations. For cells 16 and 17 SC scores are also quite high, as it can also be noticed from Fig. 9d. On the other hand, target cell sub-call method is much simpler, and requires significantly less information about user event occurrence location.

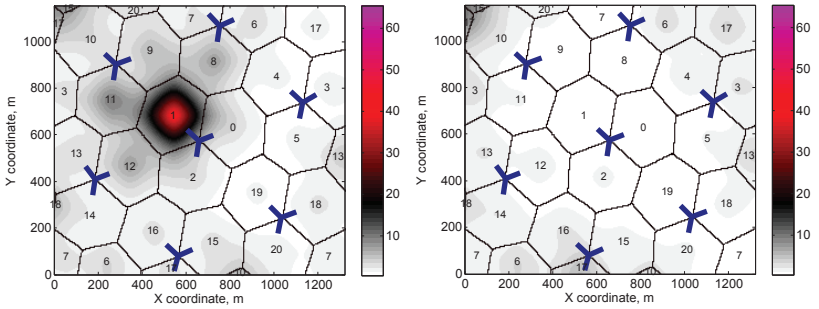
6.3 Combined Method of Sleeping Cell Detection

The idea of this method is to create a cumulative sleeping cell detection histogram based on the results from all 4 post-processing 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. 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 post-processing 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



(a) Problematic dataset sleeping cell detection histogram

(b) Reference dataset sleeping cell detection histogram



(c) Problematic dataset heat map

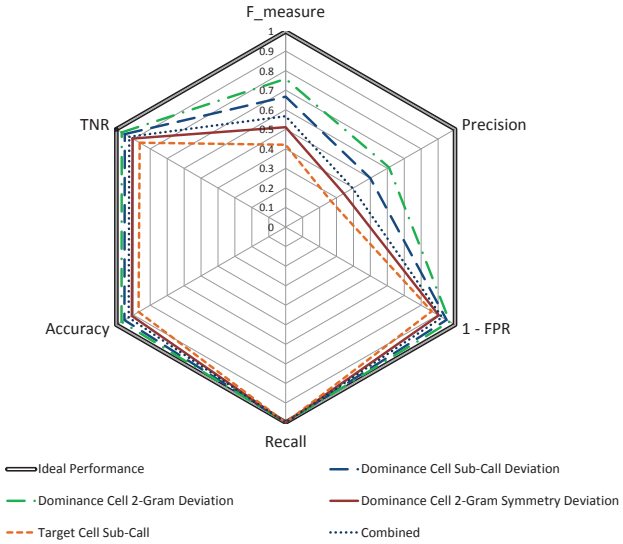
(d) Reference dataset heat map

Fig. 10 Results of sleeping cell detection for amplified combined method

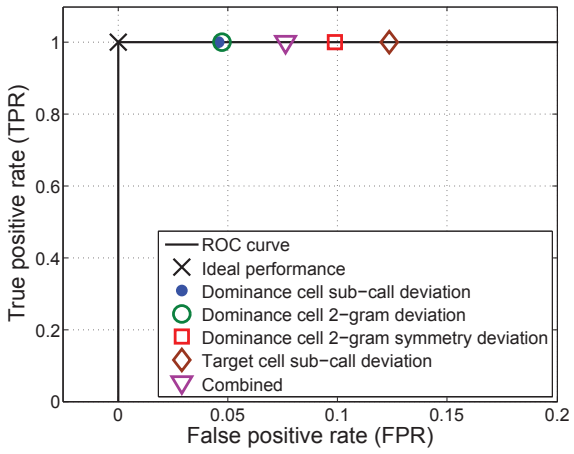
of optimal weights is a matter of a separate study and is not discussed in this article.

6.4 Comparison of Algorithms and Performance Evaluation

The post-processing methods discussed above have their own advantages and disadvantages. Traditional data mining metrics, discussed in Section 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. 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 post-processing 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 neighboring cells of cell 1 exceed the 3σ 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 observed behavior emphasizes that amplification should be complemented by some other ways to to take network topology into



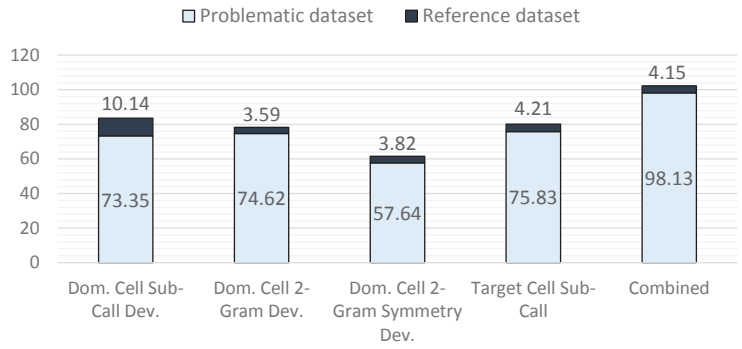
(a) Performance Measures of Algorithms



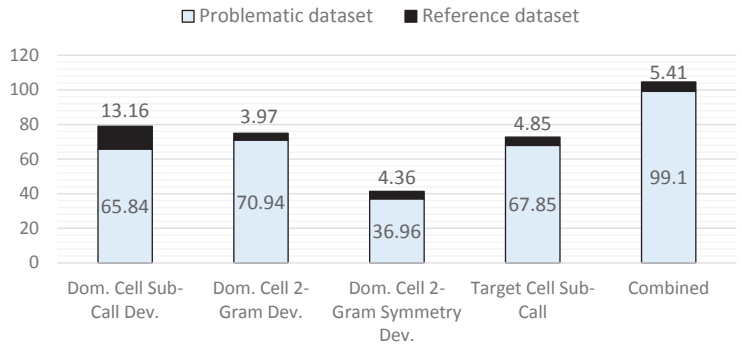
(b) ROC curve of sleeping cell detection framework

Fig. 11 Performance measures for comparison of sleeping cell detection algorithms

account. However this is a subject for further study. ROC curve of of the designed sleeping cell detection algorithm is presented in Fig. 11b. The proposed framework is able to create such a projection of the MDT data, that in the new space normal data and anomalous data points are fully separable and do not overlap. Hence, the suggested data mining framework for sleeping cell detection is successful, and for reduction of false alarm rate it is necessary to invent a better separation rule, than 3σ deviation from mean SC score.



(a) Distances in original - non amplified approach



(b) Distances in amplified approach

Fig. 12 Heuristic performance comparison of algorithms

Another method for comparison of post-processing algorithms is a heuristic approach described in Section 5.5.1. According to this method, more accurate post-processing algorithm is the one, which has the smallest distance to the ideal solution point for either problematic or error-free case. Cumulative distances for different algorithms in non-amplified and amplified cases are presented in Fig. 12a and Fig. 12b correspondingly. It can be seen that Dominance Cell 2-Gram Symmetry Deviation method has the smallest distance from the ideal detection case. Thus, from perspective of the heuristic performance evaluation approach this method outperforms other post-processing 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 type

of sleeping cell problem is rather complex, and detection of this problem has never been studied before.

The designed knowledge mining framework is semi-supervised and has centralized architecture from perspective of self-organizing networks. 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 pre-processing 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 and using this information, anomalous data points are converted with post-processing to the knowledge about location of the problematic regions in the network. Comparison of different location mapping post-processing methods is done, additionally, so called amplification is used to take into account neighbor relations between cells and network topology, for improvement of sleeping cell detection performance.

Results demonstrate, that the suggested framework allows for efficient detection of the random access sleeping cell problem in the network. Evaluation shows that post-processing method named Dominance Cell 2-Gram Symmetry Deviation demonstrates the best combination of performance results. Amplification also proves to be the very efficient approach for improvement of the detection quality. 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 especially 5^{th} Generation (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. (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). 3GPP TR 36.839 V11.1.0
2. 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
3. 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
4. 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
5. 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
6. 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
7. Brown PF, deSouza PV, Mercer RL, Pietra VJD, Lai JC (1992) Class-based n-gram models of natural language. Computational Linguistics 18:467–479
8. Brueninghaus K, Astely D, Salzer T, Visuri S, Alexiou A, Karger S, Seraji GA (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
9. Cavnar WB, Trenkle JM (1994) N-gram-based text categorization. In: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp 161–175
10. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput Surv 41(3):15:1–15:58
11. Chernogorov F (2010) Detection of sleeping cells in long term evolution mobile networks. Master's thesis, University of Jyväskylä, Finland
12. 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
13. 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)
14. Cheung B, Kumar GN, Rao SA (2005) Statistical algorithms in fault detection and prediction: Toward a healthier network. Bell Labs Technical Journal 9(4):171–185
15. Cheung B, Fishkin SG, Kumar GN, Rao SA (2006) Method of monitoring wireless network performance. Tech. rep., Los Angeles, CA, uS Patent 2006/0063521 A1, CN1753541A, EP1638253A1

16. Cheung B, Fishkin SG, Kumar GN, Rao SA (2006) Method of monitoring wireless network performance. US Patent 2006/0063521 A1, CN1753541A, EP1638253A1
17. Choi J, Kim H, Choi C, Kim P (2011) Efficient malicious code detection using n-gram analysis and svm. In: Barolli L, Xhafa F, Takizawa M (eds) NBiS, IEEE Computer Society, pp 618–621
18. 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
19. Ciocarlie G, Cheng CC, Connolly C, Lindqvist U, Nitz K, Novaczki S, Sanneck H, Naseer-ul Islam M (2014) Anomaly detection and diagnosis for automatic radio network verification. In: 6th International Conference on Mobile Networks and Management, MONAMI 2014
20. Ciocarlie G, Cheng CC, Connolly C, Lindqvist U, Novaczki S, Sanneck H, Naseer-ul Islam M (2014) 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
21. Ciocarlie G, Lindqvist U, Nitz K, Novaczki S, Sanneck H (2014) 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
22. Coifman RR, Lafon S (2006) Diffusion maps. Applied and Computational Harmonic Analysis 21(1):5 – 30
23. Commission FC (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
24. David G (2009) Anomaly detection and classification via diffusion processes in hyper-networks. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel
25. 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
26. Ganapathiraju M, Weisser D, Rosenfeld R, Carbonell J, Reddy R, Klein-Seetharaman 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 (2009) 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 (2009) 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 (2014) Self-organizing networks (SON); self-healing concepts and requirements (release 12). Tech. rep., 3GPP TS 32.541 V12.0.0
32. Guillet F, Hamilton HJ (eds) (2007) Quality Measures in Data Mining, Studies in Computational Intelligence, vol 43. Springer
33. Haidar 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
34. Hämälainen S, Sanneck H, Sartori C (2012) LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency, 1st edn. Wiley Publishing
35. Han J, Kamber M (2006) Data Mining: Concepts and Techniques, Second edition, vol 54. Morgan Kaufmann
36. Hapsari W, Umesh A, Iwamura M, Tomala M, Gyula B, Sebire B (2012) Minimization of drive tests solution in 3GPP. Communications Magazine, IEEE 50(6):28–36
37. Hapsari W, Umesh A, Iwamura M, Tomala M, Gyula B, Sebire B (2012) Minimization of drive tests solution in 3GPP. Communications Magazine, IEEE 50(6):28–36
38. He Z, Cichocki A, Xie S (2009) Efficient method for tucker3 model selection. Electronics Letters 45:805
39. He Z, Cichocki A, Xie S, Choi K (2010) Detecting the number of clusters in n-way probabilistic clustering. IEEE Trans Pattern Anal Mach Intell 32(11):2006–2021
40. Holma H, Toskala A (2011) LTE for UMTS: Evolution to LTE-Advanced, 2nd edn. Wiley Publishing
41. 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
42. Johansson J, Hapsari W, Kelley S, Bodog G (2012) Minimization of drive tests in 3GPP release 11. Communications Magazine, IEEE 50(11):36–43
43. Jolliffe I (2002) Principal Component Analysis. Springer Series in Statistics, Springer
44. Kassis E (2010) Anomaly-based error detection in base station data. Master's thesis, Tel-Aviv University, Israel
45. Kela P (2007) Downlink channel quality indication for evolved universal terrestrial radio access network. Master's thesis, University of Jyväskylä, Finland
46. Khanafer R, Solana B, Triola J, Barco R, Moltsen L, Altman Z, Lazaro P (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
47. Kolehmainen N (2007) Downlink packet scheduling performance in evolved universal terrestrial radio access network. Master's thesis, University of Jyväskylä, Finland
48. 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

49. Luo FL, Unbehauen R, Cichocki A (1997) A minor component analysis algorithm. Neural Networks 10(2):291–297
50. Mueller CM, Kaschub M, Blankenhorn C, Wanke S (2008) A cell outage detection algorithm using neighbor cell list reports. In: Hummel K, Sterbenz J (eds) Self-Organizing Systems, Lecture Notes in Computer Science, vol 5343, Springer Berlin Heidelberg, pp 218–229
51. 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
52. Networks NGM (2008) Recommendation on SON and O&M Requirements. Tech. rep., NGMN, URL "http://www.ngmn.org/"
53. Networks NGM (2008) Use Cases related to Self Organising Network, overall description. Tech. rep., NGMN, URL "http://www.ngmn.org/"
54. 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
55. 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
56. Rabin N (2010) Data mining dynamically evolving systems via diffusion methodologies. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel
57. 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
58. Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. SIGMOD Rec 29(2):427–438
59. Ramiro J, Hamied K (2012) Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE, 1st edn. Wiley Publishing
60. Scully N, et al (2008) D2.1: Use cases for self-organising networks. URL http://www.fp7-socrates.eu
61. Sesia S, Baker M, Toufik I (2011) LTE - The UMTS Long Term Evolution: From Theory to Practice. John Wiley & Sons
62. 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
63. Turkka J, Ristaniemi T, David G, Averbuch A (2011) Anomaly detection framework for tracing problems in radio networks. In: The 10th International Conference on Networks, ICN 2011
64. Turkka J, Chernogorov F, Brigatti K, Ristaniemi T, Lempiäinen J (2012) An approach for network outage detection from drive-testing databases. Journal of Computer Networks and Communications
65. Yilmaz ONC, Hämälainen J, Hämälainen S (2011) Self-optimization of random access channel in 3rd generation partnership project long term evolution. Wirel Commun Mob Comput 11(12):1507–1517

PII

DATA MINING FRAMEWORK FOR RANDOM ACCESS FAILURE DETECTION IN LTE NETWORKS

by

Sergey Chernov, Fedor Chernogorov, Dmitry Petrov, Tapani Ristaniemi

Proc. 25th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2014

http://dx.doi.org/10.1109/PIMRC.2014.7136373

PIII

N-GRAM ANALYSIS FOR SLEEPING CELL DETECTION IN LTE NETWORKS

by

Fedor Chernogorov, Tapani Ristaniemi, Kimmo Brigatti, Sergey Chernov

Proc. 39^{th} IEEE International Conference on Acoustics, Speech and Signal Processing, 2013

http://dx.doi.org/10.1109/ICASSP.2013.6638499

PIV

DETECTION OF SLEEPING CELLS IN LTE NETWORKS USING DIFFUSION MAPS

by

Fedor Chernogorov, Jussi Turkka, Tapani Ristaniemi, Amir Averbuch

Proc. 73rd IEEE Vehicular Technology Conference (VTC Spring), 2011

http://dx.doi.org/10.1109/VETECS.2011.5956626

PV

AN APPROACH FOR NETWORK OUTAGE DETECTION FROM DRIVE-TESTING DATABASES

by

Jussi Turkka, Fedor Chernogorov, Kimmo Brigatti, Tapani Ristaniemi, and Jukka Lempiäinen

Journal of Computer Networks and Communications, Volume 2012 (2012), Article ID 163184
http://dx.doi.org/10.1155/2012/163184

PVI

COGNITIVE SELF-HEALING FOR FUTURE MOBILE NETWORKS

by

Fedor Chernogorov, Ilmari Repo, Vilho Räisänen, Timo Nihtilä, Janne Kurjenniemi

Proc. 11th IEEE International Wireless Communications & Mobile Computing Conference (IWCMC), 2015