Control and Cybernetics

vol. **47** (2018) No. 2

Temporal alarm pattern discovery in mobile telecommunication networks based on binary series analysis*

by

Artur Maździarz

Ph.D. student, Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw Poland
artur.mazdziarz@gmail.com

Abstract: Highly-advanced systems, such as mobile telecommunication networks, characterized by increased complexity, make maintenance routines difficult. Amount of data to be analyzed in a short time during fault diagnosis of the mobile telecommunication networks strongly justifies the need to automate alarm correlation and root cause analysis. A major challenge in the establishment of alarm correlation is to determine how to reflect the alarm flow inertia. Thus, adequate temporal alarm pattern discovery methods should be used in fault diagnosis for correlation-related purposes. Automatic temporal alarm pattern discovery allows fast generation of root cause analysis hypotheses and supports effective troubleshooting of network problems. The process for fault propagation throughout the network is manifested by the time lag between the root-cause alarm and potentially linked symptoms, as well as weakening correlation strength with time. The paper presents a novel method for alarm correlation analysis in mobile telecommunication networks, based on binary series analysis. The method allows for discovery of causal relationship between alarms with dynamic alarm correlation window size estimation.

Keywords: mobile telecommunication networks, fault diagnosis, root cause analysis, Dice similarity coefficients, Hamming distance, temporal pattern mining

1. Introduction

The major challenges during the troubleshooting of faults in a system as complex as a telecommunication network are the amount of data and limitations to analysis time. The volume of troubleshooting data during fault propagation for large networks can easily exceed several dozens of alarms per second. In

*Submitted: September 2018; Accepted: December 2018

the case of faults that make an impact on network usability for a large number of end users, problem resolution time is crucial and very important for service providers. There are numerous benefits of automating fault diagnosis routines. By automating the troubleshooting process, the time needed for identifying a potential source of the problem may be considerably shortened, what influences downtime and QoS (Quality of Service) figures for the network. Quick troubleshooting facilitates satisfying the Customers' SLAs (Service Level Agreement). In addition, less skilled personnel can be involved in network operation routines, thus reducing network maintenance costs (Samba, 2006).

In this paper we focus on the correlation between fault management events (alarms), which is a part of fault localization phase of fault diagnosis. In order to holistically cope with automation in fault diagnosis field, work can be pursued in two separate areas, data format on the one hand and efficient data correlation methods on the other. From the input data format perspective, the best option for fast processing is the binary format. In our approach, we convert raw data, collected by the Network Management System into binary format, where the appearance of a certain alarm type at a given moment of time is represented with ones. The data correlation method should be characterized by fast processing as well as easy interpretation of the results. For identifying the relationship between alarm events we use the *Dice* coefficient with two derivative coefficients, *Dice*1 and *Dice*2, which can be interpreted as empirical estimates of a conditional probability. They indicate how many occurrences of a given alarm trigger other alarms to occur at the same time. The proportion of positive matches in relation to all the occurrences of a given alarm is converted into estimates of conditional probability, represented by Dice1 or Dice2 coefficients. As we can analyse this aspect pairwise from the perspective of the first alarm in an analysed pair or of the second alarm in the pair, we also derive the possibility to discover the direction of the relation. In order to reflect the alarm flow inertia, which is manifested by the time lag between the root-cause and the associated effect-alarms, we apply a discrete, bidirectional binary shift to the binary series and calculate the respective coefficients for these modified input sequences. The time correlation window is estimated based on the *Hamming* distance between the analyzed binary series. In addition, we use a graph-type structure to represent the causal relations of alarm events, with the level of uncertainty expressed by the estimates of conditional probabilities as the fault propagation model. The graph has been selected as fault propagation model due to its intuitive and easy to analyse interpretation. The selected model makes it possible to uncover alarm sequences, which are central to the goal of correlating alarm events (Steinder and Sethi, 2004; Samba, 2006; Boulotas et al., 1994).

The paper is organized as follows. Section 2 briefly describes the Mobile Telecommunication Network architecture and summarizes the role of the Root Cause Analysis process in daily network maintenance. Section 3 introduces similarity coefficients for binary series. In Section 4 we present alarm correlation methodology, based on the calculation of *Dice*, *Dice*1, *Dice*2 coefficients and *Hamming* distance for binary alarm data representation. Section 5 describes

the experiments and the results achieved. Finally, concluding remarks are given in Section 6.

2. Preliminaries and problem statement

A Mobile Telecommunication Network consists of two major functional subsystems: Radio Access Network (RAN) and the Core Network (CN).

The RAN is responsible for managing radio resources, involving strategies and algorithms for controlling power, channel allocation and data rate. It allows the user terminal to access network services. The Core Network is mainly responsible for high level traffic aggregation, routing, call control/switching, user authentication and charging. The entire network is managed by the OSS (Operations Support System), sometimes also called NMS (Network Management System) (Datta and Niharika, 2013).

The fault management domain of the network is characterized by several definitions and notations that are central to this paper (Steinder and Sethi, 2004; Bouillard et al., 2013).

- **Event** is an exceptional condition occurring in the operation of hardware or software in a managed network; an *instantaneous occurrence* at a time.
- **Alarm** is a notification about an event.
- **Event correlation** is the process of establishing relationships between network events.
- **Alarm correlation** is the process of grouping alarms, which refer to the same problem in order to highlight those, which indicate the possible root cause.
- **Root causes** are events, that can cause other events, but are not caused by other events; they are associated with an *abnormal state* of network infrastructure.
- **Error**, **Fault** or **Failure** is a discrepancy between the observed or computed value or condition and a true value or condition that is assumed to be correct.
- **Symptoms** are external manifestations of failures (errors), which are observed as alarms.

Fault diagnosis typically involves three processes: fault detection, fault localization (also called fault isolation or root cause analysis) and fault identification (testing of the possible hypotheses) (Steinder and Sethi, 2004).

As presented in Fig. 1, in the first stage of the entire process, fault detection has to take place, i.e. the network element has to detect a malfunction and send a notification (alarm) to the Network Management System or the Network Management System itself should obtain the faulty status of the Network Element. Fault detection is the process of collecting information, related to the malfunction of the network components (network elements) in the form of alarms (Steinder and Sethi, 2004). In the next step, the alarms are analysed and potential fault hypotheses are isolated (Root Cause Analysis, Fault Localization, Fault Isolation). Fault localization or Root Cause Analysis (RCA)

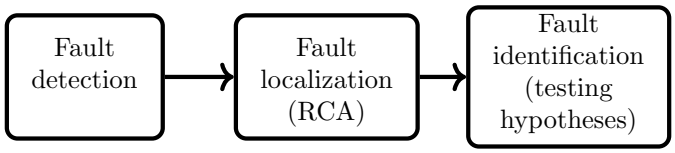

Figure 1. Visualization of fault diagnosis process (after Steinder and Sethi, 2004)

is the process of identifying the origins of the faults. It involves several steps of correlating events (including alarms), which occurred over a certain period of time, together with technical knowledge about the system under analysis (Steinder and Sethi, 2004; Bhaumik, 2010). Finally, the proposed root cause analysis hypotheses are tested and validated (fault identification). Furthermore repairing actions can be taken (Steinder and Sethi, 2004; Bouloutas et al., 1994; Bhaumik, 2010; Raghavan, 2015).

Each alarm event possesses five major attributes: alarm number, alarm description, alarm type, alarm severity, and name of the alarming object (the network element). The alarm number is a unique number, which identifies a fault. Typically, *alarm numbers* are divided into ranges (classes) representing a specific subsystem, network element type and alarm type. The *alarm description* inside an alarm frame is a very short, compact description of the fault that usually contains a few words. The *alarm type* can be specified as communication, or, for example, equipment type. The *alarm severity* specifies the importance of the fault and describes the class of an alarm. It can take one of the following nominal values: critical, major, medium, minor or warning. The name of the object is an object identification label, which clearly identifies the network element, which sent the alarm event signal. An important note is that in practice we are not dealing with signaling of single alarm events, but rather with a class of alarms, which represent a certain category of problems, related to a certain network element type.

The graph models, often referred to as fault propagation graphs, have been found to be very useful in developing efficient diagnostic algorithms (Padalkar et al., 1991). Let $G = (V, E)$ denote a fault propagation model graph, where V denotes the set of nodes (vertices), alarms in our case, and an edge $(u, v) \in E$ (arc) denotes the fact that an alarm represented by node u is linked with the alarm represented by node v. If u and v form an ordered pair, u is the tail of the arc and v is the head, then the arc is directed from u to v and is represented by an arrowhead in v ($u \to v$). Specifically, a directed arc represents a causal relation between the alarm events. If (u, v) is unordered, the respective arcs are referred to as undirected arcs and represented with a simple line. The characteristics of arcs define the graphs themselves as either directed or undirected. It is also possible for a graph to include both directed and undirected arcs, and in such a case the graph is called partially directed or mixed (Bang-Jensen and

Gutin, 2009; Diestel, 2005). In our case, the existence of the arc (edge) and its configuration are defined by Dice, Dice1 and Dice2 coefficients, calculated on the basis of a binary representation of an alarm occurrence.

In this paper, we consider several aspects of alarm correlation. The literature of the subject proposes several approaches, dealing with alarm correlation. It is clear that the telecommunications and IT domains share similar approaches to coping with the alarm correlation task holistically. It is also clear that there are developments in many domains with respect to the alarm (alert) correlation area, such as Network Management Systems, Supervisory Control And Data Acquisition field (SCADA), IT Security (Salah et al., 2013) as well as Software Engineering, aimed at pinpointing the root cause of software failures (Abreu et al., 2009). A special challenge in the telecommunication field is to select a technique, which will cope efficiently with multi-layer complex mobile networks that generate big amounts of symptoms (alarms). There is no well-established taxonomy for alarm correlation techniques in the literature. We analyzed two taxonomic approaches. In the classification proposed by Kim et al. (2011), we distinguish four major categories of correlation methodologies. These are: Rule-Based Alarm Correlation, Codebook-Based Alarm Correlation, Case-Based Alarm Correlation and Mining-Based Alarm Correlation (Kim et al., 2011). Rule-Based Alarm Correlation is a manual methodology of creating correlation rules for identifying the root causes from symptoms collected from the network, based on expert knowledge and experience (Banerjee et al., 2009). The Codebook-Based approach is also based on expert knowledge, but it defines a binary matrix format to establish the relation between the problems and symptoms. The appearance of a particular symptom for a given problem is denoted by 1, while lack of relation between the symptom and the problem is coded by 0. The symptoms can be later analysed in a causality graph, created on the basis of the matrix (EMC White Paper, 2009; Jian and Ming, 2008). In Case-Based Alarm Correlation we use past experience to solve current problems. The methodology is based on good documentation and quick access to knowledge database (Lewis, 1995). The Mining-Based Alarm Correlation methodology is based on data mining algorithms, which are able to isolate correlated alarms (alarm clusters) and propose root cause analysis hypotheses automatically. The main challenge in this approach is the processing time for big alarm data sets (Jukic and Kunstic, 2009; Vaarandi, 2003, 2005). Among the remaining challenges in this methodology, the main problem is the selection of the appropriate lag between events to establish the correlation between symptoms and the associated problems. There are already several publications, proposing probabilistic solutions to this specific problem (Expectation Maximization Based - likelihood maximization, for example). However, discovering dependencies among multiple events and the accuracy of the relations discovered remain an open topic for future work (Zeng et al., 2014). Another classification of correlation techniques is provided by Salah et al. (2013) where the authors propose a comprehensive taxonomy that takes into account the number of data sources, the type of the application (NMS, IT Security, SCADA), the correlation method and the data

distribution architecture type (centralized, distributed, hierarchical). According to this proposal, correlation analysis methods are divided into three categories: Similarity-Based methods, Sequential-Based methods and Case-Based methods. The Similarity-Based methods cluster and aggregate alerts based on the similarity of their attributes. In the Sequential-Based category, the methods correlate the alerts by discovering causal relations among them. This group of methods includes Graph-Based methods, Codebook-Based methods, Markov models, as well as Bayesian networks and neural networks. According to this taxonomy, Case-Based methods include all the methods that rely on the existence of a knowledge-based system which stores past experiences, previously observed scenarios and solutions. Following the taxonomy proposals presented above, our alarm correlation methodology based on *Dice*, *Dice*1, *Dice*2 coefficients and *Hamming* distance can be classified as Mining-Based/Sequential-Based method.

3. Similarity coefficients for binary series. Characteristics of the Dice, Dice1, Dice2 coefficients and Hamming distance

Binary sequences have proven their usefulness in many fields. Binary representation of the data makes it possible to calculate similarity or dissimilarity (distance) using relatively simple metrics. Thanks to the binary representation of the data all operations on respective series are very effective and fast from the processing perspective. The metrics of similarity for binary series are used in many areas like ecology (Jaccard, Forbs), biology (Jaccard, Dice-Sorensen, Kulczynski, Driver-Kroeber-Ochiai), ethnology, taxonomy, geology, image recovery, chemistry, biometric patterns identification and many others (Choi et al., 2015; Hubalek, 1982; Warrens and Joost, 2008; Jousselme and Maupin, 2012; Wijaya et al., 2016). An important question to be answered in a number of implementations is how similar binary sequences are. This explains why so many different similarity or dissimilarity coefficients have been proposed and analysed, see for example Choi et al.(2015) or Hubalek (1982). In this paper the *Dice*, *Dice*1 and *Dice*2 similarity coefficients are used.

To be more precise, let us denote by E a finite set of all alarms and assume a discrete time scale with the unit of time set to 1 second. Next, for any alarm $\hat{e} \in E$ define a binary sequence $e = (e_1, e_2, ...)$ in such a way that $e_k = 1$ if the alarm $\hat{e}$ occurred at time k and $e_k = 0$ otherwise. Now, to evaluate the conditional probability $P(\hat{e}|\hat{f})$ concerning two alarms $\hat{e}$ and $\hat{f}$, a suitably selected similarity coefficient for binary sequences may be used.

Let us consider two binary sequences $e = (e_1, \ldots, e_N)$ and $f = (f_1, \ldots, f_N)$, and define their scalar product in the standard way, i.e.

$$e \bullet f = \sum_{k=1}^{N} e_k f_k. \tag{1}$$

It is easy to notice that $e \bullet f$ is equal to the number of positions on which both sequences are set to 1. In the context of alarm correlation, this may be interpreted as the number of times when both alarms (alarm classes), $\hat{e}$ and $\hat{f}$, occurred simultaneously. Next, assume that $\bar{e}$ and $\bar{f}$ denote the sequences obtained from e and f in such a way that all ones are replaced by zeros and all zeros by ones. Then the following four variables summarize all information contained in the sequences e and f:

- $a = e \bullet f$ - the number of positions, on which elements of both sequences are equal to 1
- $b = \bar{e} \bullet f$ - the number of positions, on which elements of f are equal to 1 whereas elements of e are equal to 0
- $c = e \bullet \bar{f}$ - the number of positions, on which elements of e are equal to 1 whereas elements of f are equal to 0
- $d = \bar{e} \bullet \bar{f}$ - the number of positions, on which elements of both sequences are equal to 0

The above four variables can be now used to construct the so-called contingency table:

Table 1. Contingency table

		f	
		1	0
e	1	$a = e \bullet f$	$b = \bar{e} \bullet f$
	0	$c = e \bullet \bar{f}$	$d = \bar{e} \bullet \bar{f}$

The $Dice$, $Dice1$ and $Dice2$ similarity coefficients are defined as:

$$S_{Dice} = \frac{2 * a}{2 * a + b + c}\,. \tag{2}$$

$$S_{Dice1} = \frac{a}{a + b}\,. \tag{3}$$

$$S_{Dice2} = \frac{a}{a + c}\,. \tag{4}$$

The Dice coefficients may be interpreted as the empirical estimates conditional probability of an alarm e given the alarm f, or an alarm f given the alarm e (Warrens and Joost, 2008). This conditional probability characteristic $P(\hat{e}/\hat{f})$ or $P(\hat{f}/\hat{e})$ is the basis of our methodology. The Dice coefficient was proposed for binary variables by Gleason (1920), Dice (1945), Sorenson (1948), Nei and Li (1979), and popularized by Bray (1956), Bray and Curtis (1957) and Warrens and Joost (2008). The coefficient is a symmetrical, two-way similarity coefficient, and takes the values from the interval $[0, 1]$. It compares the number of coincident appearances of 1s ("1" "1" situation) in the analysed binary series

to the cumulative total number of appearances of 1s in both binary sequences. The $Dice1$ and $Dice2$ coefficients were proposed by Dice (1945), Wallace (1983), and Post and Snijders (1993) (see Warrens and Joost, 2008). Coefficients $Dice1$ and $Dice2$ are asymmetrical coefficients, which estimate conditional probability between a pair of identical length binary series. They take values from the interval $[0, 1]$. These coefficients compare the number of coincident occurrences of 1s in the analysed binary series to the total number of occurrences of 1s in analysed binary series. The $Dice1$ coefficient compares the coincident occurrences of 1s to the total number of the occurrences of 1s in the first binary series, while the $Dice2$ coefficient estimates the relation to the total number of occurrences of 1s in the second binary series. An additional advantage of Dice coefficients is the simplicity and the speed of computation, which is in line with our main goal (Hubalek, 1982; Warrens and Joost, 2008). The Dice coefficient-based methodology can be considered as methodology for correlating binary representations of the appearance of alarms and it can be used as discovery engine for probability relations in fault propagation models.

Apart from the group of normalized similarity coefficients there are a lot of unnormalized distance (dissimilarity) measures for binary series. One of the examples of very popular dissimilarity measure is the so-called $Hamming$ distance, which is widely used in information analysis. This distance metric is named after the American mathematician Richard Hamming (1915-1998). The $Hamming$ distance is defined as follows, taking into consideration the contingency table of Table 1:

$$D_{Hamming} = b + c\,. \tag{5}$$

It expresses the number of different values of bits in the same position in a pair of identical length binary series. The $Hamming$ distance measures the number of configurations of bits where "1" and "0" are placed in the same position in the analysed binary series ("0" "1" ; "1" "0" cases) (Niederreiter and Winterhof, 2015).

4. Alarm correlation identification. The methodology

The goal of the methodology is fast identification of alarms, which are correlated, among big alarm data sets. The method should be able to evaluate causal relation between correlated alarms, taking into consideration an inertia of the alarm flow. The relation should be represented by a strength measure, which takes values from the $[0, 1]$ interval. Finally, there should be a graphical representation of the results, shown in the form of a graph structure.

The methodology is applied to gathered data and generates on demand alarm dependencies from the analyzed samples. The RCA (root cause analysis) conclusions drawn can be treated as the snapshot of current FM (fault management) situation or can be used for the future as learned alarm patterns. This approach is similar to the pattern recognition concept, where we recognize patterns in the

analyzed data set and use predefined data subsets for further analysis and data classification.

For this purpose, a collection of data, describing incidents and alarms in a real telecommunication network is used. For all the alarms, observed during a fixed time interval, the moments of time when they occurred have been recorded. Information gathered in this way, after conversion to a binary format, will make it possible to estimate the unknown relation measure and to infer possible alarm correlations. Other information, regarding collected alarms i.e.: alarm severity, alarm type, name of the alarming object, including a topology reference, are used in the fault localization process for validating the RCA hypotheses.

The Dice coefficients are interpreted as empirical estimates of conditional probability of there being a relation between the alarms. Taking into account the individual characteristics of the coefficients, in the first stage we propose to use the Dice coefficient for preliminary identification of the possible correlation of the alarms. Based on the experiments we selected the value of 0.2 as Dice coefficient threshold for detecting a possible correlation between two alarms.

After isolating potential alarm candidates for correlation, $Dice$, $Dice1$ and $Dice2$ coefficients are used to evaluate the strength and the direction of the correlation between the binary representations of the occurrence of alarms. In other words, we use $Dice$ coefficient as a preliminary correlation detector, which is followed by the application of $Dice$, $Dice1$ and $Dice2$ coefficients in order to obtain more precise conclusions about correlations.

A very important aspect to note is that $Dice$, $Dice1$ and $Dice2$ coefficients can have the same value. This is the case of binary series with the same number of occurrences of 1s/0s. If $Dice$, $Dice1$ and $Dice2$ coefficients have the same value, the direction of the relation cannot be determined by the algorithm. In such a situation, the value of the coefficients only approximates the strength of the relation and is represented by undirected arc in the fault propagation model graph. In this specific case, the expert needs to evaluate this correlation hypothesis.

For binary series with a different number of 1s/0s (asymmetrical binary series), $Dice1$ and $Dice2$ coefficients are used for the evaluation of the strength and the direction of the relation.

If the first binary alarm representation in an analysed pair is the root-cause, the $Dice1$ coefficient will have higher value than $Dice$ and $Dice2$ coefficients ($Dice1 > Dice \wedge Dice1 > Dice2$). The value of $Dice1$ coefficient will indicate the direction and the strength of the relation. Analogously, if the second binary alarm representation in an analysed pair is the root-cause, $Dice2$ with its value will be used for the identification of the direction and the estimation of relation strength ($Dice2 > Dice \wedge Dice2 > Dice1$).

The alarms linked to a given incident usually come with a certain delay, which is related to the reaction time of interconnected network elements for a problem. A visualization of the alarm correlation is presented in Fig. 2. An important aspect to be noted is that an alarm event in our approach does not have any duration. The alarm representation takes the value of "1" only at the

time instant when it is generated.

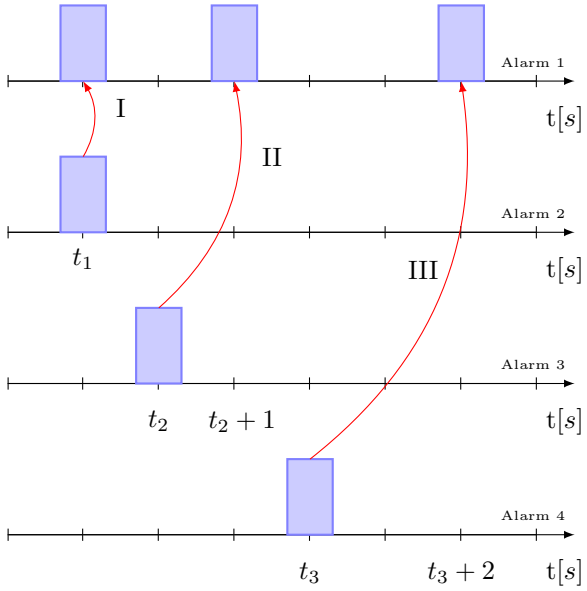

Figure 2. Visualization of alarm correlation, backward (positive) shift example: I) Alarm 1 and Alarm 2 correlated without time shifting. II) Alarm 1 and Alarm 3 correlated with backward (positive) time shift by 1 second. III) Alarm 1 and Alarm 4 correlated with backward (positive) time shift by 2 second

In order to reflect the inertia of the alarm flow, which is manifested by the time lag between the root-cause alarm and the potential linked effect alarms, we model the correlation with the *Hamming* distance (Hd) between binary series, representing the alarm occurrence time. We use the *Hamming* distance as the correlation level control function, which helps our correlation engine, based on rolling $\widetilde{Dice}$ coefficients to evaluate how a temporal shift (τ) influences the correlation. In order to detect the occurrence of a correlation in the time function and simulate the decreasing value of the relation strength with the increasing applied delay (τ), we calculate a weighted average of $\widetilde{Dice}(\tau)$, $\widetilde{Dice1}(\tau)$, $\widetilde{Dice2}(\tau)$ coefficients for the given time lag interval. We implement the calculation of $\widetilde{Dice}(\tau)$ (similarly $\widetilde{Dice1}(\tau)$, $\widetilde{Dice2}(\tau)$) coefficients for binary series, shifted in time by τ (temporal shift implementation).

The model expressions of $\widetilde{Dice}$, $\widetilde{Dice1}$, $\widetilde{Dice2}$ coefficients are presented below:

$$\widetilde{Dice} \simeq \frac{1}{N}\sum_{\tau}(\widetilde{Dice}(\tau) * (1.5)^{-|\tau|}) \tag{6}$$

$$\widetilde{Dice}1 \simeq \frac{1}{N_1}\sum_{\tau}(\widetilde{Dice1}(\tau) * (1.5)^{-|\tau|}) \tag{7}$$

$$\widetilde{Dice}2 \simeq \frac{1}{N_2}\sum_{\tau}(\widetilde{Dice2}(\tau) * (1.5)^{-|\tau|}) \tag{8}$$

$$\forall(\widetilde{Dice}(\tau) > 0 \wedge \widetilde{Hd}(\tau) \leq \widetilde{Hd}(0)) \tag{9}$$

$$\forall\widetilde{Dice}(\tau) : \frac{\widetilde{Dice}(\tau) * (1.5)^{-|\tau|}}{\sum\limits_{\tau}\widetilde{Dice}(\tau) * (1.5)^{-|\tau|}} > 0.1 \tag{10}$$

$$\forall\widetilde{Dice1}(\tau) : \frac{\widetilde{Dice1}(\tau) * (1.5)^{-|\tau|}}{\sum\limits_{\tau}\widetilde{Dice1}(\tau) * (1.5)^{-|\tau|}} > 0.1 \tag{11}$$

$$\forall\widetilde{Dice2}(\tau) : \frac{\widetilde{Dice2}(\tau) * (1.5)^{-|\tau|}}{\sum\limits_{\tau}\widetilde{Dice2}(\tau) * (1.5)^{-|\tau|}} > 0.1 \tag{12}$$

where N, N_1, N_2 are the numbers of respective coefficients fulfilling conditions (9) through (12).

An important remark is that we take into consideration only weighted Dice coefficients, which represent at least 0.1 of the overall weighted average value.

In the method we allow only for the shift, which does not increase the $Hamming$ distance, what is represented by the expression (9). On the other hand, we can also interpret the $Hamming$ distance as a sort of ability for correlation after we apply the lag τ. The ability is represented by the number of different bits in the same position in the analysed binary series. The motivation behind this approach is that the difference of bits in the same position enables the conditions to achieve the correlation after applying a temporal shift between the binary series. In addition, the weakening correlation effect between the alarms in time function is represented by the rolling exponentially-weighted average.

We selected the exponential function of the form

$$f(\tau) = coefficient(\tau) * (1,5)^{-|\tau|}$$

for our model as we aim to achieve the weakening effect on the level of 0.2 for the lag applied equal $\tau = 4$ in relation to the full correlation without the lag applied ($coefficient(4) = 1 \Rightarrow f(4) = 0.1975$).

The algorithmic scheme for $\widetilde{Dice}(\tau)$, $\widetilde{Dice1}(\tau)$, $\widetilde{Dice2}(\tau)$ coefficients calculation is presented in Fig. 3.

Because the binary shift for the binary series under analysis can be imposed symmetrically (bidirectionally), we apply the shift in both directions. The maximum value of the shift (τ) for the binary series is determined by the $Hamming$ distance $\widetilde{Hd}(\tau)$ and the value of $\widetilde{Dice}(\tau)$ coefficient. We use the values of τ

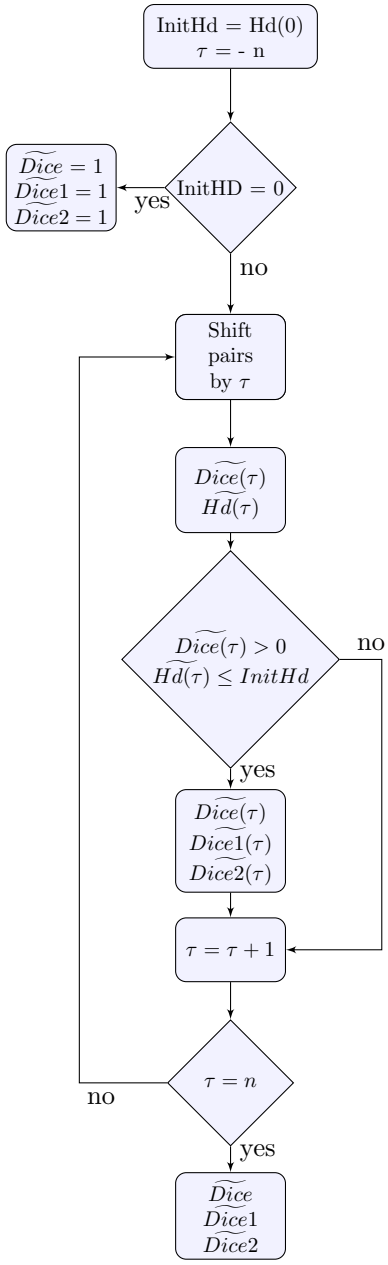

Figure 3. Calculation of $\widetilde{Dice}(\tau), \widetilde{Dice}1(\tau), \widetilde{Dice}2(\tau)$ coefficients with Hamming distance $\widetilde{Hd}(\tau)$ control, the algorithmic scheme

from the interval $[-n; n]$, such that do not increase the initial value of the $Hamming$ distance $InitHd = \widetilde{Hd}(\tau = 0)$ for positive values of $\widetilde{Dice}(\tau)$ coefficient ($\widetilde{Dice}(\tau) > 0$). The absolute, maximum value of τ fulfilling the criteria, specified by the expression (13) below, determines the maximum value of the shift applied and defines the correlation window (14):

$$\{\tau : \tau \in [-n; n] \cap Z, (\widetilde{Hd}(0) \geq \widetilde{Hd}(\tau), \widetilde{Dice}(\tau) > 0) \tag{13}$$

$$\tau_{max} = max(\mid \tau \mid) \tag{14}$$

where n represents the length of the analysed binary series and Z is the set of integer values.

In other words, we use weighted average in order to represent the mean value of the Dice coefficients, calculated in specified correlation-window based on the $Hamming$ distance. By using the rolling exponentially weighted average, we scale the contribution of $\widetilde{Dice}(\tau)$, $\widetilde{Dice1}(\tau)$, $\widetilde{Dice2}(\tau)$ coefficients to the total mean depending on the time lag applied to analysed binary series. In addition, for RCA hypotheses generation we select only the most significant components for the rolling $\widetilde{Dice}$, $\widetilde{Dice}1$, $\widetilde{Dice}2$ coefficients calculation, as expressed by (10), (11), (12).

Thanks to the $Hamming$ distance and the $Dice$ coefficients characteristics it is sufficient to calculate the $Dice$ coefficients only for binary series with no zero value of the $Hamming$ distance. For the binary series with the initial $Hamming$ distance of zero (without applying binary shift) we have full correlation expressed only by the 1s in the same position in the analysed binary series. In this case there is no need for further analysis and the alarms can be classified as correlated.

The entire RCA methodology can finally be summarized in Fig. 4.

5. Experiments and results

The proposed approach has been applied to real data obtained from one of the main telecommunication operators. The network under analysis consists of a variety of different 2G, 3G and 4G network elements. The overall data set contains information about 1 440 813 alarm events, gathered from July 2014 to May 2015. For each alarm, five attributes are stored, the time of occurrence and a numerical ID which depends, in a unique way, on the source of the alarm, its priority, severity and a brief description of what has happened. All the attributes are used in the final RCA hypotheses validation.

Numerical experiments were carried out on a PC with Intel(R) Core(TM) i7-4600U 2.1 GHz processor, 16 GB main memory and 64-bits MS Windows operating system. We used the R package environment version 3.3.1. The performance of the Dice-based alarm correlation methodology has been evaluated by executing tests on independent sets of collected alarms. We tested the speed of the algorithm by measuring correlation processing time for alarm subsets extracted from selected samples. The subsets contained 10, 100, 1 000, 10 000

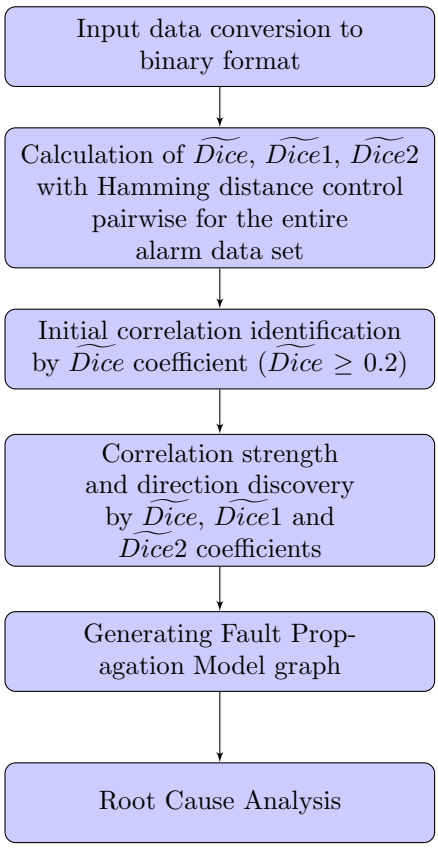

Figure 4. Root Cause Analysis based on $\widetilde{Dice}$, $\widetilde{Dice}1$, $\widetilde{Dice}2$ coefficients

and 20 000 alarms. As shown in Table 2 and in Fig. 5, for four samples the correlation processing time for all the samples with up to 1 000 alarm events is around 10 seconds. In the worst case we achieved 1176 seconds correlation processing time for the sample with 20 000 alarms.

Table 2. Dice correlation methodology performance

Alarm events number	**Sample1 correlation time [s]**	**Sample2 correlation time [s]**	**Sample3 correlation time [s]**	**Sample4 correlation time [s]**
10	**1**	**2**	**1**	**2**
100	**2**	**2**	**2**	**3**
1000	**6**	**12**	**7**	**9**
10000	**28**	**446**	**60**	**89**
20000	**374**	**1176**	**121**	**300**

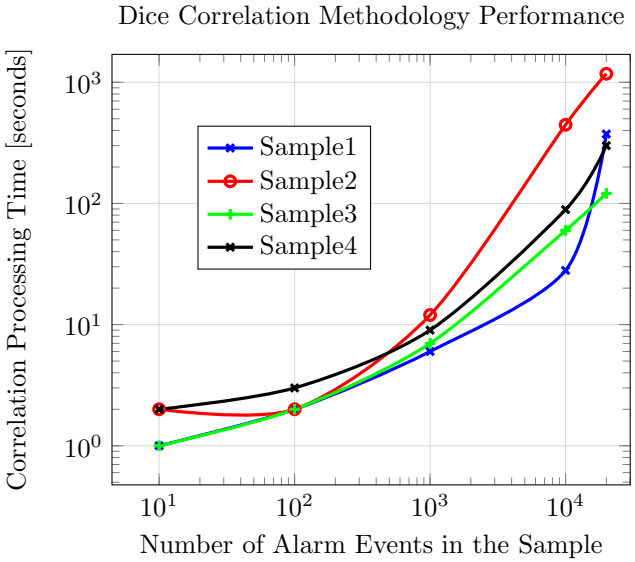

Figure 5. Performance of Dice-coefficient-based alarm event correlation methodology

The proposed Dice-based algorithm was even able to construct a large FPM consisting of several thousands of alarms, e.g. up to 64 000 alarms in a single sample. For clarity reasons, we analyze a part of the generated FPM containing seven alarm events and link them with arcs, as this is presented in Fig. 6.

This is a 2G technology alarm event correlation example. The alarm events attributed to this case are presented in Table 3. For all the event pairs, which fulfill the correlation criteria $\widetilde{Dice} \geq 0.2$, the $\widetilde{Dice}$, $\widetilde{Dice}1$, $\widetilde{Dice}2$ coefficients,

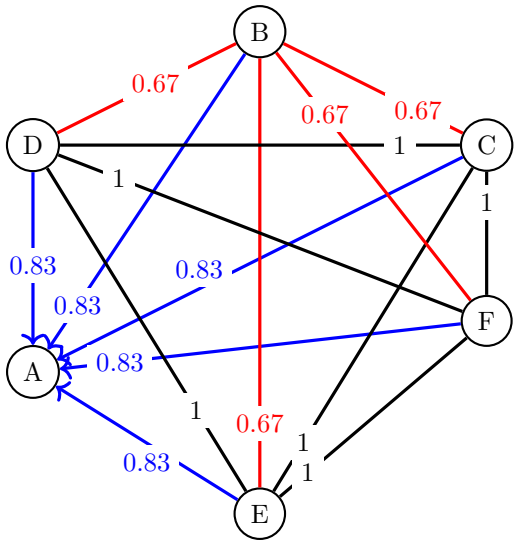

Figure 6. 2G Technology FPM example

the *Hamming* distance ($InitHD$) and the applied maximal temporal shift τ_{max} values are presented in Table 4.

Table 3. 2G technology alarm events correlation example - alarm events attributes

Alias	**Alarm Event.Network Element**	**Occurrence time**
A	TRX FAULTY.PLMN-PLMN/BSC1/BCF1	46747 46748
B	TRX FAULTY.PLMN-PLMN/BSC1/BCF1/BTS4/TRX-26	46747
C	TRX OP. DEGRADED.PLMN-PLMN/BSC1/BCF1/BTS1/TRX-2	46748
D	BTS OP. DEGRADED.PLMN-PLMN/BSC1/BCF1/BTS2	46748
E	BTS OP. DEGRADED.PLMN-PLMN/BSC1/BCF1/BTS1	46748
F	TRX OP. DEGRADED.PLMN-PLMN/BSC1/BCF1/BTS2/TRX-8	46748

In this example we present the correlation related to TRXes (transceivers). The alarm, which is reported by the Network Element of the type BCF (BCF1) "TRX FAULTY", is caused by alarms "TRX OP.DEGRADED" on two TRXes connected to separate BTSes (sectors). The "BTS OP.DEGRADED" alarms are also the consequence of the TRX problem.

In this case, the Dice-based methodology generated a partially directed FPM graph. We have a set of directed arcs, for which $\widetilde{Dice} \neq \widetilde{Dice}1 \neq \widetilde{Dice}2$ and a set of undirected relations, for which $\widetilde{Dice} = \widetilde{Dice}1 = \widetilde{Dice}2$. The entire set of alarm events should be isolated as one correlation hypothesis. The example shows the accuracy and reliability of the methodology. The alarm event A is the only node in the FPM graph to which the directed arcs converge for all other

Table 4. $\widetilde{Dice}, \widetilde{Dice}1, \widetilde{Dice}2$ coefficients, initial *Hamming* distance and τ_{max} values for 2G technology example

Alarm Event.1	**Alarm Event.2**	$\widetilde{Dice}$	$\widetilde{Dice}1$	$\widetilde{Dice}2$	$InitHD$	τ_{max}
A	E	0.55	0.42	0.83	1	1
A	C	0.55	0.42	0.83	1	1
A	D	0.55	0.42	0.83	1	1
A	G	0.55	0.42	0.83	1	1
A	B	0.55	0.42	0.83	1	1
E	C	1	1	1	0	0
E	D	1	1	1	0	0
E	F	1	1	1	0	0
E	B	0.67	0.67	0.67	2	1
C	B	0.67	0.67	0.67	2	1
C	D	1	1	1	0	0
C	G	1	1	1	0	0
D	B	0.67	0.67	0.67	2	1
D	G	1	1	1	0	0
F	B	0.67	0.67	0.67	2	1

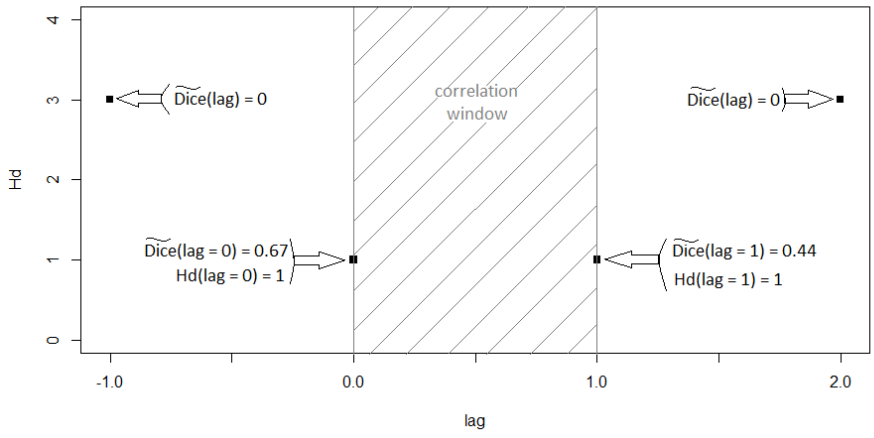

Figure 7. The correlation method visualization for alarm events A and B from the 2G technology example, the *Hamming* distance (Hd) and $\widetilde{Dice}(\tau)$ coefficient in the function of lag τ

alarm events in the sample. The alarm event A is caused by alarm events B, C, D, E, F with the probability of 0.83. From the event time occurrence data, presented in Table 3, it is clear that the relation between the events is established on the basis of the coincidence of the occurrence of events A, B, C, D, E, F. The

alarm event A occurred twice and it has been classified as the effect of the events B, C, D, E, F. In this example, we can see the initial *Hamming* distance value on the level of 1 and also the same value of the time lag τ_{max} applied. In Fig. 7 we present the values of $\widetilde{Dice}$ coefficient together with *Hamming* distance in the function of temporal lag applied for the correlation of alarm events A and B from the 2G technology example. We can see that two values of the $\widetilde{Dice}$ coefficient are taken into consideration for estimating the correlation. It is the $\widetilde{Dice}$ coefficient for the lag value of 0 and the $\widetilde{Dice}$ coefficient for the time lag of 1.

The methodology was able to discover multiple undirected relations having probability of 0.67 between the events with a single occurrence within the 2-second time window. In this case, the initial *Hamming* distance takes the value of 2 and the applied temporal lag for this case is 1. Finally, the alarm events, which occurred at the same time, share the same correlation probability at the level of 1 without an established direction of the relation. In these circumstances, the initial *Hamming* distance takes the value of 0 and there is no temporal lag applied by the algorithm.

In the second example, we present a more complex 3G technology case, where two alarm events occurred in the data set many times. Alarm events occurrence times are presented in Tables 5 and 6. It is clear that manual correlation of these events would be very challenging, due to the amount of data (occurrence time references) to be correlated. From the technical point of view, we can see that 3G WCEL1 reports CELL OPERATION DEGRADED, most probably due to WBTS1 problem with the license, required for a WBTS operation. In this case, we obtained an undirected FPM graph with the relation strength 0.22. All three coefficients $\widetilde{Dice}, \widetilde{Dice}1, \widetilde{Dice}2$ have the value of 0.22 in this case.

In this example, the initial *Hamming* distance takes the value of 112 due to number of different time occurrences of the alarms. In this case, the maximum time lag applied is 2 seconds. In Fig. 8 we can see that only three values of $\widetilde{Dice}$ coefficient are taken into consideration for the alarm correlation in this case, the value for $\tau = 0$, the value for $\tau = 1$ and the value for $\tau = 2$, as only these values fulfill the requirement of not increasing the value of the *Hamming* distance with a positive, significant value of $\widetilde{Dice}$ coefficient.

6. Conclusions

The alarm correlation methodology, based on the exponentially weighted average of *Dice*, *Dice*1 and *Dice*2 coefficients, shows satisfactory accuracy, speed and reliability of generation of correlation hypotheses. In the here presented methodology, the *Hamming* distance is used to control the maximal value of the temporal shift τ to be applied in order to capture the correlation between the alarm events. The methodology generates the reasonable Fault Propagation Models for the Mobile Telecommunication Network. It is very effective from the computational point of view, and it is possible to run the algorithm on a PC. The proposed approach of using the *Dice*, *Dice*1, *Dice*2 coefficients with the

Table 5. 3G technology alarm events correlation example - attributes of the alarm event 1

CELL OPERATION DEGRADED.RNC1/WBTS1/WCEL1
320 450 782 834 964 1265 1519 1677 1924 2030 2197 2347
2434 2523 2671 2795 2849 2961 3073 3287 3413 3636 4412
4568 5408 5907 6064 7812 8665 8746 16401 17537 18104
19180 19339 1948819590 19827 19970 20070 20294 20457
20637 20814 20942 21290 21767 21856 22068 22344 22609
22791 22864 23009 23200 23286 23376 23440 23516 23769
23839 23972 24039 24305 24574 24728 25957 85821 86346

Table 6. 3G technology alarm events correlation example - attributes of the alarm event 2

BTS RESET NEEDED TO ACTIVATE A LICENSE.RNC1/WBTS1
320 452 784 836 965 1267 1520 1678 1924 2031 2198 2349 2434 2524
2672 2796 2850 2962 3074 3288 3414 3637 4413 4569 5409 5908 6065
7813 8666 8747 16402 17538 18104 19182 19340 1949019591 19827
19972 20071 20295 20458 20637 20814 20943 21291 21769 21856
22068 22345 2261122793 22864 23011 23202 23287 23379 23442
23517 23771 23841 23972 24041 24305 24575 24728 25958 85822
86347

Hamming distance as temporal lag estimation for generating Fault Propagation Models works very efficiently for models with several thousand symptoms (alarms).

The values of conditional probability estimates allow us to filter the most probable symptoms for network problems with the right priorities. The binary temporal shift introduced into the algorithm at the level, which does not increase the *Hamming* distance, provides a good model of the time correlation window in mobile telecommunication networks and makes it possible to correlate alarms more accurately. The exponentially weighted average of *Dice*, *Dice*1 and *Dice*2 coefficients simulates reasonably well the impact of alarm propagation time on the value of correlation strength.

The methodology is universal and works regardless of the mobile technology, which is used in the network (2G,3G,4G). It has been established that the methodology provides also a good base for constructing alarm correlation patterns. The patterns obtained could be used as predefined alarm correlation rules for reducing the alarm correlation effort in the future for alarm data sets for a given technology.

Comparing the methodology proposed here with the state-of-the-art approaches is a separte challenge, due to the lack of standard strategies for such comparison. Evaluating the performance of alert correlation techniques, as well

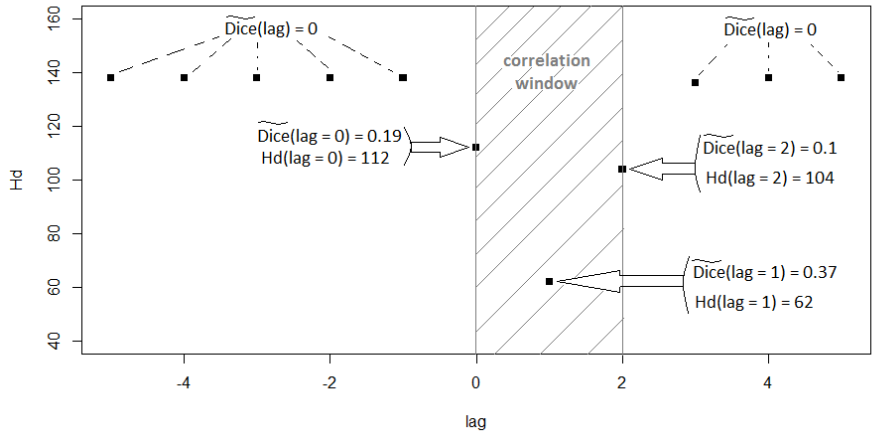

Figure 8. 3G Technology FPM example, $\widetilde{Dice}$ coefficient, *Hamming* distance (Hd) in the function of temporal lag τ

as any other aspect of the correlation, would require well established benchmarking. This point was also made by Abren et al. (2009). We propose six alarm correlation method attributes to be taken into consideration for comparison purposes: alarm correlation accuracy, which is also related to the level of alarm reduction (clustering), execution time scalability, correlation threshold calibration requirement, possibility of applying correlation time-window, and, finally, the ability to discover causal relationship between the alarm events. We compare Dice-based methodology with Similarity-based family of methods, referring to the Temporal-based methods and to the Sequential-based methods, where we use Bayesian Networks (BN) as the category representation.

Regarding the alarm correlation accuracy, we found that any Similarity-based method accuracy requires a priori selection of similarity threshold based on the experts' knowledge. In that case the accuracy depends on the experience. It is possible to apply optimization algorithm for the Similarity-based approach like cluster analysis in timeline, taking into consideration alarm occurrence time. In this approach, optimization can be seen as a variance minimization task across the entire data set for the selected number of alarm clusters. This type of method, though, has significant performance problems for big data sets and requires additional actions to reduce the input data set size (Maździarz, 2018). In the case of Bayesian Networks, the unsupervised discovery of the relation between alarms requires appropriate representation of alarm occurrence (high occurrence frequency). This method will not be able to discover correlation between seldom events. It is possible to incorporate expert knowledge in this method, as well, in the form of a priori conditional probability data (CPT

tables), but the accuracy in this case will also rely on the experts' knowledge. Regarding the method execution time scalability, based on the experiments we set up the target value of 10 seconds correlation time for 1000 alarms. This requirement is fulfilled by most of the similarity-based methods as they work on simple metric calculation for alarm attributes and compare the derived value of the metric with a certain threshold level. The Dice-based method performance depends strongly on the density of the data. For some of the samples, the threshold limit of 1000/10, as mentioned before, has been slightly exceeded. In the case of Bayesian Networks the performance of the method is the worst compared to other methods.

Practical implementation of BN for big data sets is difficult. Each of the alarm correlation methods requires a so-called method calibration in order to validate the correlation threshold. Regarding the correlation methods, where time is the major correlation attribute, the correlation time-window plays fundamental role in the correlation process. It is possible to apply the correlation time-window in any of the comparable methods, but optimizing the time-window size is a big challenge. In case of the Dice-based method we use the *Hamming* distance for estimating the time-window size. In other methods, this aspect requires significant optimization effort (Maździarz, 2018). Regarding the ability to discover the causal relation between the alarms, it is possible in the Dice-based method as well as in the Sequential-based methods. The Similarity-based methods do not provide the possibility to discover the causal relationship between the alarm events. A summary of comparison is shown in Table 7.

Table 7. Alarm correlation methods comparative summary

	Dice-based	**Similarity-based (Temporal-based)**	**Sequential-based (Bayesian Networks)**
Alarms reduction (clustering)	+	+	+
Accuracy	**high**	**medium**	**medium**
Execution time scalability [1000/10]	**medium**	**high**	**low**
Scores/correlation threshold	+	+	+
Correlation time-window	+	+	+
Causal relationship discovery - graph model	+	-	+

In the future, we will extend our model to consider also other alarm attributes, such as topology for alarm events correlation in mobile telecommuni-

cation networks. We will validate the usage of other binary similarity coefficients for correlating alarm events. Alarm event interrelations other than pairwise dependencies will also be in the scope of our research.

References

Abreu, R., Zoeteweij, P. and van Gemund, A.J. (2009) *Spectrum-based multiple fault localization. Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering.* IEEE Computer Society.

Banerjee, D., Madduri, V. and Srivatsa, M. (2009) *A Framework for Distributed Monitoring and Root Cause Analysis for Large IP Networks. 28th IEEE International Symposium on Reliable Distributed Systems September 2009.* IEEE Computer Society, 246-255.

Bang-Jensen, J. and Gutin, G. (2009) *Digraphs: Theory, Algorithms and Applications.* 2nd edn., Springer, Heidelberg.

Bhaumik, S.K. (2010) Root cause analysis in engineering failures. *Transactions of The Indian Institute of Metals*, **63**, (2-3), 297-299.

Bouillard, A., Junier, A., Ronot, B. (2013) *Alarms correlation in telecommunication networks.* [Research Report] RR-8321, INRIA.

Bouloutas, A., Calo, S., and Finkel, A. (1994) Alarm correlation and fault identification in communication networks. *IEEE Transactions on Communications*, **42**, 2/3/4.

Choi, S.S., Cha, S.H., and Tappert, C.C. (2015) *A survey of binary similarity and distance measures.* Department of Computer Science, Pace University New York, US.

Datta, R., Niharika, N. (2013) Comparative study between the generations of mobile communication 2G, 3G & 4G, *International Journal on Recent and Innovation Trends in Computing and Communication*, **1** (4).

EMCWhitePaper (2009) *Automating Root-Cause Analysis: EMC Ionix Codebook Correlation Technology vs. Rules-based Analysis.* November 2009, EMC White Paper Technology Concepts and Business Considerations.

Hubalek, Z. (1982) Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, **57**(4), 669-689.

Jian, W. and Ming, L.X. (2008) A novel algorithm for dynamic mining of association rules. *International Workshop on Knowledge Discovery and Data Mining Jan 2008.* IEEE, 94-99.

Jousselme, A.L. and Maupin, P. (2012) Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, **53**, 118-145.

Jukic, O. and Kunstic, M. (2009) Logical inventory database integration into network problems frequency detection process. *ConTEL 2009, June 2009.* IEEE, 361-365.

Kim, D.S., Shinbo, H., Yokota, H. (2011) *An Alarm Correlation Algorithm*

for Network Management Based on Root Cause Analysis. KDDI R&D Laboratories Inc. Fujimino Saitama Japan ICACT2011.

Lewis, L. (1995) *Managing Computer Networks - A Case-Based Reasoning Approach*, Telecommunications Library Artech House, Inc.

Maździarz, A. (2018) Alarm correlation in mobile telecommunication networks based on k-means cluster analysis method. *Journal of Telecommunications and Information Technology (JTIT)*, **2/2018**, 95-102.

Niederreiter, H. and Winterhof, A. (2015) *Applied Number Theory*. Springer, DOI: 10.1007/978-3-319-22321-6,102.

Padalkar, S., Karsai, G., Biegl, C., Sztipanovits, J., Okuda, K., and Miyasaka, N. (1991) Real-time fault diagnosis. *IEEE Expert*, 75-85.

Raghavan, A. (2015) *Root cause analysis. Management and Leadership - A Guide for Clinical Professionals*. Springer Verlag, 105-121.

Salah, S., Macia-Fernandez, G., and Diaz-Verdejo, J.E. (2013) A model-based survey of alert correlation techniques. *Computer Networks*, Elsevier, 1289-1317.

Samba, A. (2006) A Network Management Framework for Emerging Telecommunications Networks. *Modeling and Simulation Tools for Emerging Telecommunication Networks Needs, Trends, Challenges and Solutions*. Springer, Boston, MA.

Steinder, M., Sethi, A.S. (2004) A survey of fault localization techniques in computer networks. *Science of Computer Programming*, **53**, 165-194.

Vaarandi, R. (2003) A data clustering algorithm for mining patterns from event logs. *IP Operations and Management(IPOM 2003), October 2003*. IEEE, 119-126.

Vaarandi, R. (2005) Tools and techniques for event log analysis. PhD thesis, Tallinn University of Technology, Department of Computer Engineering , Estonia, June 2005.

Warrens, M.J. (2008) Similarity coefficients for binary data. Properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients. Doctoral dissertation, Leiden University, Netherlands.

Wijaya, S., Afendi, F.M., Latifah, I., Darusman, K., Altaf-Ul-Amin, M., and Kanaya, S. (2016) Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines. *BMC Bioinformatics* 17:520, DOI:10.1186/s 12859-016-1392-z.

Zeng, Ch., Tang, L., Li, T., Shwartz, L., Grabarnik, G.Y. (2014) Mining Temporal Lag from Fluctuating Events for Correlation and Root Cause Analysis, Network and Service Management (CNSM). *10th International Conference on Network and Service Management (CNSM) and Workshop*. IEEE, 19-27, ISBN: 978-3-901882-67-8.