Topic modeling for anomaly detection in telecommunication networks

H. Joe Steinhauer¹ · Tove Helldin¹ · Gunnar Mathiason¹ · Alexander Karlsson¹

Received: 31 January 2019 / Accepted: 18 June 2019
© The Author(s) 2019

Abstract

To ensure reliable network performance, anomaly detection is an important part of the telecommunication operators' work. This includes that operators need to timely intervene with the network, should they encounter indications of network performance degradation. In this paper, we describe the results of an initial experiment for anomaly detection with regard to network performance, using topic modeling on base station run-time variable data collected from live Radio Access Networks (RANs). The results show that topic modeling clusters semantically related data in the same way as human experts would and that the anomalies in our test cases could be identified in latent Dirichlet allocation (LDA) topic models. Our experiment further reveals which information provided by the topic model is particularly usable to support human anomaly detection in this application domain.

Keywords Telecommunication anomaly detection · Topic modeling · Decision-making

1 Introduction

To maintain a high quality of service with regard to reliability, robustness and speed is of utmost importance for today's wireless telecommunication networks. This importance is anticipated to dramatically increase in the future, where also emergency services like ambulance, health service and fire department will rely almost entirely on wireless communication.

In wireless telecommunication systems that to some extent are self-organizing (Imran and Zoha 2014; Aliu et al. 2013), as well as in many other highly complex systems, detection and identification of abnormal behavior is part of the operators' monitoring and maintenance work. A good and stable network performance needs to be ensured through timely and appropriate decisions about possible interventions in the network.

The related analysis of the events in the network is currently work-intensive, even though it is assisted by a multitude of automated monitoring systems that aid the human operators to better understand complex network behavior. Due to the anticipated increasing requirements of the

networks (Pirinen 2014), the challenges telecommunication operators face are constantly increasing. Hence, additional automated processes are needed for monitoring, analyzing, and interpreting the network's behavior.

To this end, telecommunication companies are investigating how artificial intelligence and exploratory data analysis can aid the human operator to establish and maintain a good situation awareness (Endsley and Kiris 1995) with regard to the network's behaviors and causality of network events. In the work presented here, we suggest the application of topic modeling (Blei et al. 2003) as a means for anomaly detection within network behavior.

This builds on our previous work (Helldin et al. 2018) where the assumption that topic modeling can be used to appropriately capture what is going on in the network has been validated through expert evaluation. We also identified what significant information can be extracted from the topic model that would aid an operator while monitoring network performance. As a result, we suggested an interactive dashboard where different aspects of the topic model are presented in a cognitively comprehensible way that fulfills the operators' information needs.

In the present work we go one step further and apply topic modeling not only to identify what is going on in the network, but also to detect if the current network behavior is different (anomalous) to what is normal for the network. Once anomalous network behavior is detected, the topic

✉ H. Joe Steinhauer
joe.steinhauer@his.se

¹ School of informatics, University of Skövde, Skövde, Sweden

model can further be used to identify possible root-causes for the anomaly. Note, that anomalous behavior within a wireless telecommunication system can be highly context sensitive and look very different for different parts of the network, different base stations and different situations.

Furthermore, anomalous network behavior does not necessarily manifest itself in decreased network performance. In the same way as an observable decreased network performance in parts of the network not necessarily constitutes an anomaly. These are two reasons why the development of an automated anomaly detection system, needs to be build step by step, including expert knowledge already in the early stages of the design process.

As a first step we focus on detecting and analyzing anomalies in cases where network performance degradation has been observed. Here, we would like to see if we can detect the anomaly, which means that we (a) will see a significant difference between the normal and the incident model and (b) that the root-cause for the performance degradation can be inferred from the comparison of normal and incident model. Note, that the anomaly, that was the reason for the network performance degradation might have occurred a significant amount of time before the degradation was noticeable.

Our work follows the design framework proposed by Koh et al. (2011) and has been described in more detail in Helldin et al. (2018). In this approach the focus is put on iterative and early prototyping together with the telecommunication experts. This allows us, as stated by Koh et al. (2011), to incorporate the experts early in the design and development process and increases the experts' awareness about which information can be potentially provided by the methods used. An approach like this is particularly important for our work since only experienced telecommunication experts are able to understand and interpret the data which are input into the topic model. Furthermore, only the telecommunication expert can validate if the topic models produced are meaningful with regard to telecommunication needs. These emphasizes as well that it is vital to involve the expert throughout the whole development process.

This paper is structured as follows: in Sect. 2 we briefly describe the background of our work, including anomaly detection, topic modeling and its visualization as well as the risks faced when topic models are interpreted by humans. In Sect. 3 we describe our approach for anomaly detection using topic modeling in the telecommunication domain and in Sect. 4 we present and validate our results, which are further discussed in Sect. 5. We conclude the paper in Sect. 6 and outline ideas for future work.

2 Background

2.1 Anomaly detection in telecommunication networks

Anomaly detection considers the task of identifying patterns in data that do not conform well to normal behavior, (Chandola et al. 2009). These patterns are often called anomalies or outliers. A common approach towards anomaly detection is to define a region of normal behavior and label any observation in the data that does not belong to this region as an anomaly, (Chandola et al. 2009). However, this process is often associated with several challenges. For example, to define what constitutes as normal behavior is not easy since there might be a fine line between what is to be considered "normal" and "anomalous". Moreover, what constitutes as normal behavior can be largely sensitive to context and time, as well as the fact that what is to be considered "normal" can be constantly evolving.

A decision support tool for anomaly detection often needs to be developed for a particular domain and set of tasks, providing alerts to the operator when data points fall outside the general, contextual normal model. However, an alert is often associated with a set of events that are correlated. Then, as posed by Livnat et al. (2005), the severity of an event must be carefully analyzed in the context of its appearance, including when the event happened and which other events that have occurred. Furthermore, an investigation of how the events are evolving needs to take place.

One way to overcome the problems stated above is to include a domain expert who knows what constitutes as normal given a specific time and place, sorting through the alerts generated by the anomaly detection tool. However, yet another challenge is to develop an anomaly detection alert system that is highly sensitive when it comes to detecting the outliers. At the same time it should be specific enough to not raise too many false alerts which would possibly lead to severe usability issues.

In telecommunication networks, it is important to keep track of the network's performance and to quickly identify the cause of a performance decrease. Today, key performance indicators (KPIs) are used for performance monitoring. This means, that the network's run-time performance is monitored through run-time variables, i.e. performance measurement counters and gauges, collected at telecom base stations. The values of these variables at the time of reading provide input to the respective KPI measures, which can be the run-time variable itself or a relative measure including several variables, e.g. connection requests received in relation to connections established.

Each KPI is then monitored to ensure that it remains within acceptable performance limits and a performance

alert is raised each time the respective measurement lies outside the acceptable value range. Consequently, performance alerts can be associated with run-time variables represented in the KPIs. The respective KPIs need then to be interpreted by experts to gain insight about where to look for the root-cause of the problem. A single telecom base station is home to more than a thousand different run-time variable counters and only a subset of them are currently used as input for the KPIs. We argue, that if a data analysis tool would make use of more run-time variables, the root-causes of the anomalies could be identified more efficiently.

Another problem is the aforementioned avoidance of false positives of KPI alerts. Since activities in wireless telecommunication networks are subject to many factors (time of day, weather conditions, maintenance work, events that group people together at places e.g. festivals) normal network traffic can look very different in different circumstances. Therefore, KPI's must be configured robustly enough not to trigger alerts when there is no reason for it.

For example, in case of a festival where many people participate, also many cell phones might be trying to connect to the nearest base station, stretching its capacity to the limit. This might result in an elevated number of connection request failures. However, it constitutes expected network behavior and should not be regarded as an anomaly and hence no alert should be issued.

On the downside, this includes that the same robustness is expressed towards differences in network behavior that are indicating problems. For example, if a near by base station is out of order and cell phones that should connect to that base station are now trying to connect to our base station, this will also lead to an elevated number of connection request failures. In this case, we are dealing with an anomaly that needs intervention. A simple KPI that triggers alerts based on a threshold value that relates to the elevated number of connection request failures is, however, not able to distinguish between these two scenarios. What is needed, is a system that is able to identify the type of anomaly on a more detailed level and that also allows for checking against the context of the situation.

Another problem that needs to be addressed is that today, faults that occur in one place in the network often lead to a cascade of alerts that not necessarily starts from the root-cause location itself. Because of one cause, many subsequent parts in the network might experience problems, generating a stream of alerts for the operator to review. Due to the complexity and flexibility of the wireless telecommunication system, it is often too difficult to identify the root-cause of such an alarm flood based solely on the provided alerts.

As such, telecommunication operators are in need of automated interactive monitoring support tools, that make it possible to identify the type and characteristics of a current

problem. Once such an aid is in place it can be used in combination with context information to decide if the situation constitutes a network anomaly that needs intervention, or rather a false positive (as in case of the festival).

With the help of exploratory data analysis (Tukey 1962) for big data, and its means for exposing statistical relationships within the data, telecommunication companies are now engaging in building decision support systems for additional performance monitoring that make use of the vast amount of available run-time variable data. It is expected that the exposed statistical relationships in the data also refer to semantically meaningful co-occurring events and can help the operator to better understand the network's behavior. The variables can be used to build normal models of wireless telecommunication network behavior and these models can then be utilized for interactive anomaly detection. This means that they provide better and faster means to identify network degradation which consequently leads to possibly earlier interventions with the network in form of counter-measures to the identified problems. This, in turn, leads to an increased quality of service.

While designing such a tool, an important issue to consider is to keep the operators' workload on a reasonable level as well as to design the anomaly detection support system in a way that aids the operators to increase their situation awareness (see for instance Endsley 2012). One step towards this goal is to identify the type of network behavior automatically and to base this identification not only on the KPI's that are responsible for triggering an alert (in the previous example an elevated connection request failure for one base station), but also on all other available run-time variables (specifically interesting for this example are the run-time variables from the surrounding base stations), which will lead to a more holistic picture of the network's behavior.

To build such a system, domain experts are needed to evaluate if potential approaches for this automation produce results that are conform with the human interpretation and analysis of the given network behavior. Therefore, we apply a design process where domain experts and operators are included early and continuously, as proposed by, for example, Koh et al. (2011).

2.2 Topic modeling

Topic modeling (Blei et al. 2003; Blei 2012) refers to a family of statistical models that aims to capture cluster patterns in count data within groups. These models have mainly been applied for text analysis where they are used to detect patterns in terms of word count frequencies in documents. The idea is that certain proportions of words are related to a certain *topic*. Hence, cluster patterns are essentially captured by the relative frequency of counts. These are then used in a probabilistic model in order to identify topics, i.e.,

probability distributions over count tokens. As an example, if a counter always has a specific proportion to another counter across several groups, then that proportion is likely to be captured as a topic. A topic model is a hierarchical model consisting of two layers; (1) a top layer of *shared cluster patterns* and (2) a *group layer* where the shared cluster patterns are used in order to model the data in the group. More formally, by assuming that we have groups $g \in \mathcal{G}$ of count data $n_1^g, \dots, n_m^g$, where m is the number of different counters and k denotes a count token indexed by $\{1, \dots, m\}$, one can describe a topic model by its generative structure:

1. Draw l topics (probability distributions over tokens) $p(k|t) \sim \text{Dirichlet}(\alpha)$, where $t \in \{1, \dots, l\}$
2. Draw topic proportions $\theta_g \sim \text{Dirichlet}(\beta)$, $g \in \mathcal{G}$
3. For each group $g \in \mathcal{G}$
 - (a) Draw a topic $t \sim \text{Multinomial}(\theta_g)$
 - (b) Draw a token $k \sim p(k|t)$
 - (c) For the size of g , repeat from (a)

Topic modeling, or rather hierarchical probabilistic models, has previously been used for anomaly detection (Xiong et al. 2011) by utilizing formal rules for determining whether or not an anomaly is evident. In contrast, we here need to utilize a more exploratory approach including human expertise to identify if the somewhat anomalous behavior of the network in fact constitutes an anomaly of the kind that needs human intervention.

2.3 Topic visualization

The output of the topic modeling algorithm, the constructed topic model, is usually given as a set of probability distributions over the variables. These variables are originally, in topic modeling for text analysis, the words within the document corpus. Each topic that could be found by the algorithm is hence a probability distribution of words. A rather mathematical description like that is often not easily comprehended by human users and needs considerable insight and time to be understood. As stated by Smith et al. (2017) “existing algorithms for understanding large collections of documents often produce output that is nearly as difficult and time consuming to interpret as reading each of the documents themselves”. This is also highlighted by Chang et al. (2009), who argue that topic models do not automatically provide meaning to their analysts, but need to be manually interpreted and evaluated together with domain experts.

One way of alleviating the human user’s work is to visualize the output from the topic model. Note, that in different application areas different aspects of the topic model might be of interest and hence there might not be one single visualization that fits every task. Several ways of visualizing

topic models have been suggested, where the most common representations are either graph based, as e.g. presented in Gretarsson et al. (2012), Dou et al. (2011) and Smith et al. (2014), or matrix or text based as described in Chuang et al. (2012), Chaney and Blei (2012) and Gardner et al. (2010). Other approaches focus on, for example, time aspects [see for instance Havre et al. (2002), Liu et al. (2012) or Cui et al. (2011)]. On a more detailed level, word clouds (Gardner et al. 2010; Klein et al. 2015) and network graphs (Smith et al. 2014; Gretarsson et al. 2012) are used but the majority of topic visualization tools uses word lists (Chaney and Blei 2012; Gardner et al. 2010) and word lists with bars (Sievert and Shirley 2014; Gardner et al. 2010) as their main visualization techniques.

In our work we will use topic modeling on run-time counter variables used in telecom base stations. For the expert who is analyzing the topics it will be of interest what variables are included in each topic and to what degree. It is also interesting to see what variables are the most significant for the topic and which ones are the most frequent. An ordered variable list with bars can provide this information, hence we have chosen LDAvis (Sievert and Shirley 2014) for topic visualization.

LDAvis shows each topic, i.e., $p(k|t)$, by presenting each counter’s relevance to that topic by a user-specified convex combination of the probability of a certain count within the topic together with this probability in relation to the probability of the count based on all groups. The result is presented on a logarithmic scale. For further details see Sievert and Shirley (2014, 2015). The size of the topics, estimated by the topic proportions θ_g and number of counts in each group $g \in \mathcal{G}$, is visualized as well as the distances between topics.

2.4 Evaluating topic models and human topic interpretation

Topic modeling is well-established as a text analysis approach for understanding the topics of a corpora. However, it is also necessary to evaluate human comprehension of a generated topic model in order to understand the accuracy of the topics in a topic model, in contrast to the held-out likelihood such as perplexity of test data and unseen documents.

Letting humans interpret topics for the purpose of topic evaluation, in an application area already known by the human, poses the risk of confirmation bias. This means there is a risk that topics are interpreted overly positively and regarded to be more meaningful than they really are. This may result in the belief that generated topics are able to explain more than they actually can. Thus, there is a need to evaluate the quality of human topic interpretation, in particular when applying topic models in new domains.

Topic evaluation measures, such as topic coherence (Mimno et al. 2011), are used to indicate how close a topic model represents the actual human comprehension of the topics generated by an algorithm. Human evaluation of topics is regarded as a gold standard, however costly due to the labor required. Multiple approaches have been developed for automatic evaluation, as well as measures of how well algorithms for topic models generate topics suited for human comprehension (Wallach et al. 2009; Röder et al. 2015).

Chang et al. (2009) suggested that the semantic meaning of topic models can be measured and they present a user study where their evaluation of the interpretability of the topics consisted of two components, *word intrusion* and *topic intrusion*. The study first evaluated how well a topic algorithm created cohesive interpretable topics, while the latter evaluated how well the association between the documents and the topics made sense to humans. To evaluate word intrusion, six randomly ordered words were presented, and the human task was to detect and classify an *intruder* word, selected from a pool of words with low resemblance with the topics of the corpus. For poor-coherence topics, the users would choose the intruder word more randomly than otherwise. Topic intrusions were evaluated to see how well the topic decomposition of a document fitted the topic model. In the work of Chang et al. (2009), the topic models evaluated were generated by the latent Dirichlet allocation (LDA), probabilistic Latent Semantic Indexing (pLSI), and Correlated Topic Models (CTM) algorithms. It was shown that LDA had the highest precision on human evaluations for word intrusions, while LDA and pLSI performed best for topic intrusions.

3 Topic modeling for telecommunication anomaly detection

The long term and overall purpose of our project is to automate wireless telecommunication network monitoring, which includes automated anomaly detection. The work presented in this paper builds on our previous work (Hell-din et al. 2018) where we use topic modeling as a tool for exploratory data analysis for wireless telecommunication network data. This concluded a first step towards our goal as it identified that topic modeling can be used to group events within the network in a manner that is conform with the telecommunication experts' interpretation.

The focus of the work presented here is to demonstrate that the topic models, build on incident free network data, can be used as normal models within anomaly detection. Here a topic model over a time period wherein an incident occurred is compared to the normal model. We present that in our test cases, the normal and incident model show significant and relevant differences. Furthermore, we state that

these differences provide information about possible root-causes for the anomaly.

Based on these findings, our future work entails to further automate the root-cause localization process. That means, however, that telecom operators need to be given significant time to manually perform anomaly detection with the help of the topic model. From this, they will acquire new expert knowledge about how to correctly interpret the information from the topic models. This expert knowledge then needs to be elicited and combined with information provided by other monitoring systems as well as context knowledge, and implemented in an automated tool that will point out identified anomalies together with their possible root-causes.

The data on that our experiments are based contains telecommunication run-time variable readings collected from base stations of the type eNodeB within a Fourth Generation (4G) Radio Access Network (RAN). In terms of topic modeling, we consider each run-time variable as a word. The numerical reading of that variable which counts the frequency with that a specific event occurs within a given time interval, is regarded to be the counterpart to what, in traditional topic modeling for text, would be representing the frequency with that a specific word occurs within a document. The collection of run-time variable readings, collected from one base station for the given time interval, then corresponds to one document. Hence, each base station produces one document per measured time interval. The collection of these documents then forms the document corpus for that time interval.

A document from a specific base station hence encodes which processes, such as cell phones establishing connection with the base station or hand-over of a call from the base station to another, the base station was engaged in. Each of these procedures are build out of a sequence of causally dependent events.

In the same way as a textual document might address several topics, the base stations typically engage in several different procedures during the measured time interval. A simple example of a behavior characterizing an identified topic is the establishing of a connection between a cell phone and a base station. This procedure is defined in terms of the comprised signals: first the base station registers a connection request from the cell phone, it then tries to build up a connection which includes a number of steps, each representing an event related to the underlying core network. Each of these events is also registered in its respective run-time variable counter. Finally, when the connection is established (or has failed to be established) the base station also registers this in the connection established or connection failed counter.

Another example is a hand-over of a connection between two base stations in case a cell phone is moving away from the geographical area of the base station it is currently

connected to and closer to another base station. The hand-over needs to be initiated by a request, the wireless core net infrastructure at the second base station needs to be build up to take over the cell phone, and then the cell phone needs to be taken over and, at last, the first base station can terminate the core infrastructure the cell phone was using prior to the hand-over.

As base stations are usually busy handling many different processes at the same time, (e.g. a multitude of cell phone connections and hand-overs, etc.) and many of these processes entail partially the same events (e.g. events that relate to the wireless core network) different procedures contribute to the incrementation of the run-time variable counters. Hence, just from the counter data itself it is difficult to infer what procedures have actually produced this data.

The topic model then re-establishes this information based on statistical features in the data set. As a result we can see that events belonging to one type of process (e.g. hand-overs) are presented as a topic. For our work we used LDA with Gibbs sampling as a topic model and the LDAvis package Sievert and Shirley (2015) to present the results. Based on this, a telecommunication expert could, through comparison of a normal model with an incident model, determine if an anomalous network behavior was present and what it consisted of. Figure 1 depicts an

overview of the anomaly detection process using topic modeling. In the following we describe this process together with one of the tested scenarios in more detail.

For each of our experiments, we created a topic model from a dataset representing one hour of normal network traffic, and a second topic model from a dataset of one hour of network traffic during which an incident occurred. The incidents, which have previously been identified by telecommunication specialists, consists, for example, of the unavailability of a mobile management entity (MME). The MME is the control plane node of the 4G mobile core network and is involved in e.g. mobile registration, security procedures and mobility procedures. It maintains information about locations of mobile devices and is responsible for selecting the closest available gateway.

Due to the good evaluation results for LDA with regard to human evaluation for word intrusion and topic intrusion as described by Chang et al. (2009), LDA was applied to build the topic models. The purpose of the experiments was to investigate, through expert validation, if topic modeling through LDA, could be used to assist the telecommunication operators to identify anomalies in the telecommunication network. Since interaction with the visualization needs a good amount of familiarization, one expert was

Fig. 1 The anomaly detection process using topic modeling. **a** Run-time data is collected from the network during normal and typical network traffic. **b** A topic model (the normal model) is constructed. **c** Run-time data is collected whenever the network is in operation. **d** When there is reason to suspect an incident in the net a topic model

of the current run-time data is created. **e** An expert telecom operator compares the current model with the normal model in order to identify anomalies and their potential root-causes. The operator needs also to take context information into account that might explain the different network behavior

given substantial time and support to gain a good understanding and familiarity with the visualization.

The expert was at first included in the initial experiments and asked to investigate whether the LDA topic model represented meaningful topics. Later the expert was asked if he, while comparing a topic model of one hour of normal network behavior and a topic model of one hour of network behavior during which an anomaly occurred, could identify that there (a) was an anomaly present in the second topic model and (b) if the nature of the anomaly could be identified from the topic model.

To validate if the LDA topic modeling captured the telecommunication topics and also the occurring anomalies correctly, the expert needed to be aware of the ground truth of both, the normal traffic data and the type of the anomaly. Hence, the expert himself had chosen these particular incidental cases as appropriate test cases. This however, also implies the risk of the validity threat of confirmation bias, as described in Sect. 2.4. However, the expert was aware of the threat and is, provided his role at the telecommunication company, not interested in promoting topic modeling or any other anomaly detection method unless it has been proven reliable.

Furthermore, during the experiments the expert identified which information that was valuable to him and why. Based on this information we could establish different and more detailed visualizations, concentrating on the information needs of the telecommunication operators. These, cognitively easier comprehensible abstractions of relevant information will, in the continuation of our work, be used in a larger scale user study.

4 Results and validation

To investigate whether topic modeling can be used for anomaly detection in the telecommunication domain, we firstly needed to analyze if the topics found in both models (normal and incident) for our test cases describe procedures, i.e. consecutive causal events, that are in accordance with how telecommunication experts and operators would cluster the same events. Since our intention was to use LDA to aid a human operator to interactively identify anomalies, a good match between the output of LDA and the operators' own interpretation of what constitutes a topic is necessary. In previous analyses the expert aiding us in our study has confirmed that topics build by LDA are conform with how human experts would expect the events to be grouped. In the closer investigated cases of the present study, both the normal models and the incident models also satisfied this need.

Since the number (k) of topics that LDA will generate from the provided data needs to be given beforehand, the question arises how a suitable number of topics can be

identified. It also raises the question if the number of topics is dynamic and could be different for the normal model and the incident model. These questions have been thoroughly investigated by the telecommunication expert. Topic models with different numbers of topics have been generated for each data set and the expert has analyzed the content of the topics for each number of topics. As a result, the expert could identify a suitable k for each dataset.

As the purpose of our work is anomaly detection, we only need to identify the appropriate number of topics once in order to build the normal model. When this model is in place, the differences that are of interest are how the content within the given topics changes when incidents occur.

As an example consider the establishment of a connection between a cell phone and a base station. The run-time variables involved in this procedure (topic) are, among others, connection request, connection established and connection failed. When there is no problem in the network connections requested usually are successfully established and hence the variable counters for connection request and connection established are more often incremented than the connection failed counter. However, even if the counter for failed connections is not increased that often, it is still part of the topic. When network problems arise and more connection failures occur the counter is incremented more often and the relationships between these three counters changes, but the topic itself stays the same.

Our experiments show, that when the number of topics given is higher than the number of topics that the human expert would identify as individual topics, LDA will generate several topics that are so similar to each other that the expert can easily identify them as instances of the same topic. In the LDavis visualization these topics are situated almost at the exact same position in the diagram and hence are also here easily identifiable as belonging together. As expected, when the number of topics is too small, the topics generated are not separating the procedures that the human expert wants to see as distinguished groups.

In the example cases analyzed, four topics seemed to be enough to capture the network traffic correctly when provided with the dataset containing this specific set of run-time variables. However, from the visualization we recognize, that in the anomaly case, when a larger number is used, the topics that are placed at almost the same spot are now slightly separated from each other. The expert's manual analyses of this phenomenon did not reveal that these slightly different topics showcase differences that pinpoint to certain effects of the anomaly and are still instances of the same topic. However, we should not dismiss the possibility that there might be some merit in the differences that are not yet comprehensible for a human expert. Here lies the potential that exploratory data analysis might reveal new knowledge about the anomaly that

human experts so far did not take into consideration. It is hence part of our future work to investigate these differences further. In the present paper we focus instead on the differences within the four topics that match the expert's understanding of the network behavior. We will consider topic 5, which represents the group of topics 1–5 in the eight topic representation, topic 6, topic 7 and topic 8.

For these topics we needed to investigate whether the topics generated for the incident hour showcase relevant differences compared to the normal network traffic hour. According to the expert's analysis, this is the case. Several systematical changes are noticeable in the incident models already at the top level of the LDAvis representation. We will look at one particular case of anomaly for demonstration, where during the incident hour the MME that serves the observed RAN area becomes unavailable. In order to omit proprietary information, we need to present an anonymized representation of the LDAvis output in Fig. 2

and we describe the procedures in the network on a conceptual level.

In the diagram in Fig. 2 topic 6 is highlighted, which means that the bar charts presented on the right side of the figure present the counters for events (run-time variables) that are most relevant for topic 6. The counters are ordered with regard to their relevance for the topic (as opposed to being ordered with regard to how frequently the event is occurring).

In the bar chart we can see that for normal and incident model different variables are considered to be most relevant. Besides counters 5 and 6 that are at the top of the list for both models, the list has changed significantly. It can also be observed that counters 5 and 6 have higher values in the incident model and that the respective events counted have also occurred within other topics beside topic 6 (observable through the blue proportion of the respective bar in the bar chart), which was not the case in the normal model.

Fig. 2 LDAvis representation of the topics generated for **a** the normal network hour and **b** the incident hour. Topic 6 has been selected (marked with red) and its most significant counters are listed through the bar chart visualizations for the two cases. The blue bars depict the overall frequency of the counters in the corpora, while the red bars indicate the estimated counter frequency within the selected topic. In **b** the counters in group A represent events belonging to basic cell phone connection procedures whereas the counters in group B represent events belonging to subsequent signalling procedures (color figure online)

We can further identify that in the normal model counter values are approximately equal for many variables which indicates normal network behavior where typical telecommunication procedures take place without considerable problems. In the incident model, however, we can see a different structure, as counter values are considerably different, indicating that several of the events were more frequent than others. Hence, we observe that the ratio between events is different in the two models. Which events that appear more frequently and which that appear less frequently than in the normal model provides important information for the telecom operator.

The difference in frequency between events that represent basic cell phone connection procedures (marked as group A in the incident model) and the events representing subsequent signaling procedures (marked as group B) is here of particular interest to the operator. It indicates that cell phone connection requests are overly represented and are not met by the corresponding necessary signaling events. Note, that different values of the ratio between specific counters alone do not necessarily constitute an anomaly—in a different situation or a different part of the network the present ratio might be the norm. Hence, to identify this as an anomaly, either a normal model or a person with a very good understanding of this particular normal situation is needed.

The important aspects in the diagram to the left of Fig. 2 are the relationships between topics and their respective distances to each other, rather than their global positioning within the coordinate system. We can see on the left side of Fig. 2, that the distance between topic 6 and 7 as well as the distance between topic 5 and 6 has increased and that topics 5 and 7 are now stronger overlapping than in the normal model.

Once it is established that there is an anomaly present, information provided by other topics helps to identify its root-cause. Topics need now to be compared individually between the normal and incident model. The investigation, done by the telecom expert, revealed that the topics 5, 6 and 8 provided enough information to describe the incident. The expert's analysis of the topics revealed the following:

Topic 5 “Expected Signalling-5”, describes procedures typical for normal RAN operations which concern the exchange of information between the different nodes within the network. Example operations are the set up, control and termination of cell phone calls. Topic 5 groups the instances of these procedures that work smoothly and hence, the run-time variables that are dealing with these procedures are represented in equal proportions.

Topic 7 “Expected Signalling-7” also groups variables involved in typical signalling procedures on the RAN, and is hence much similar to topic 5. However, we can see that topic 7 also includes (in the normal model as well as in the incident model) a proportion of connection re-establishing

procedures, and signalling failure counters, which are both not significantly present in topic 5 in the normal model.

When we compare topic 5 in the normal model with topic 5 in the incident model we see that, even though the same run-time variables (describing the same type of events which represent typical network behavior) are represented in both models, they occur to a significantly lower degree in the incident model. Additionally, run-time variables that keep track of the failure of many of these events, here in particular events that are related to different types of signaling, are more prevalent in the incident model than in the normal model. This behavior indicates to the expert that something in the underlying network is not working as usual.

This analysis also explains why topic 5 and topic 7 are overlapping to a higher degree in the incident model than in the normal model. It is due to the higher amount of variables dealing with failing events in topic 5 in the incident model. These variables are not significantly present in topic 5 in the normal model, whereas they have been part of topic 7 in the normal model to approximately the same degree as in the incident model.

Topic 6 “Radio Connections” groups together run-time variables that are involved in handling wireless radio connections between different nodes in the network. It shows, among others, that the ratio between cell phones that request connections and base stations that accept connections is different in the normal and incident model. In the latter, the request connection rate is higher than the accept connection rate. This indicates that cell phones are trying and re-trying to connect to a base station but are frequently failing to establish a connection. With this over-representation of connection requests and connection failures in the incident case, the content of topic 6 has shifted towards dealing to a higher degree with these types of procedures.

Since these procedures are not equally present in topic 5, topic 6 has less in common with topic 5 and has hence moved further away from it in the graphical representation of the incident model. Additionally, as already discussed, the ratio between counters for basic cell phone connection procedures (group A) and subsequent signalling procedures (group B) is significantly different in the incident model.

Topic 8 “Hand-over” is characterized by run-time variables representing hand-over procedures. In this topic the hand-over failure rate is higher during the incident case. It can further be seen that the hand-over operations are registered in the eNodeBs but are failing significantly. A telecom expert knows, that a typical reason for this type of failure would potentially be an unavailable MME. Hence, topic 8 provides important information about where to look for the root-cause of the problem.

Taken together, what can be identified from the topic model is what the expert would have expected for this kind of problem. The model points out the symptoms of the

failure in topic 5, topic 6 and topic 8. The most likely root-cause can be inferred from expert knowledge once the symptoms are identified. Hence, we can verify through expert validation that a) wireless network procedures are identified correctly and clustered appropriately and meaningfully for a human expert and b) that a difference between the normal and the incident model manifest itself in the topics that represent the procedures that ought to be influenced by the anomaly. Note, that the anomaly only affects some topics, other topics that describe network procedures that are not affected by the unavailable MME are also not affected in the topic model.

In another incident case that we analyzed, a RAN reconfiguration caused the cells of the RAN to be restarted. It is interesting to note, that the consequences of this incident are similar to the previous case (unavailable MME). However, the root-cause is different and the impact on the network service is less severe. Also in this example the telecom expert could identify that an anomaly had occurred by comparing the normal model with the incident model. Additionally, the expert was again able to pinpoint the potential root-cause correctly.

Summarizing, the topic models captured combinations of run-time variables in the same way as the telecom expert would expect with regard to the procedures that are taking place in the network. Anomalies can be spotted when an incident model is compared to a normal model. The first indication is the relative positioning of topics in the visualization and the change in size of the topic circles.

To analyze what in particular has changed within the identified anomalous topics the expert needs to look at the specific topics individually and compare their respective normal and incident model. Here, the ratio between typical events within these topics together with changes in significance of run-time variables for the respective topics from normal to incident model are relevant.

LDavis provides all this information but, as previously described, requires some training for correct interpretation.

Hence, we suggest, that the information needed should be represented in a cognitively easier form. A simple suggestion to present the different ratios of significant events within topics, while at the same time providing an overview which topics include what run-time variables is presented in Fig. 3. A qualitative representation like this can now be used to design an experiment for a user study including additional experts and operators. This study will provide a more thorough evaluation of if and for what types of anomalies topic modeling can be used within telecommunication networks. It will also reveal information of how operators will use the information provided and hence provide the basis for further automation of the anomaly detection process.

5 Discussion

The purpose of the study presented here was to investigate whether topic modeling is a suitable tool for anomaly detection and identification of root-causes for anomalies within the telecommunication domain. The obtained results are affirmative to that question.

Testing a new approach for anomaly detection in an application area where only experts with considerable experience are able to decide if the approach is working, is dependent on a very close collaboration with the experts. Therefore, we have adopted the design approach suggested by Koh et al. (2011) where an iterative design process, including the domain expert early on, is promoted.

However, several limitations of our work exist. Firstly, only one expert was asked to analyze the test cases and the ground truth for the anomalies was known to him, hence interpretation bias might have influenced the results.

Secondly, so far, only a limited number of cases has been studied in great enough detail. The reason for these limitations are, as indicated, that before topic modeling can be applied on a larger scale on telecommunication network data, including more experts and operators, a good indication

Fig. 3 Matrix comparison of the generated topics and their counter distributions for **a** the normal network hour and **b** the incident hour. Topic 6 has been selected for inspection and its most significant counters for the normal/incident hour have been highlighted

needed to be established that the approach is indeed promising. This confirmation could only be provided by a domain expert who had to spend a considerably large amount of time to learn to interpret LDAvis. This was also needed to help us identify the features and values from the LDAvis representation that the expert uses for anomaly detection. As a result, we are now in a position where we can design a more elaborate experiment that provides the relevant information for anomaly detection and identification, but in a cognitively easier comprehensible way.

In our work, we generate topics for telecommunication data using LDA, which is somewhat different from topic models for text analysis. The difference between traditional topic model comprehension of text corpora, and topic modeling on telecommunication run-time variable readings needs to be further explored, but it can be expected, based on Chang et al. (2009), that LDA is the most accurate alternative out of the established topic models. Measuring word intrusions and topic intrusions in a user study for telecommunication run-time variables will allow for an objective evaluation of the topics generated, outlining our future work.

With some experience with the tool applied on the data, the telecom experts was able to identify a suitable number of topics for anomaly detection purposes. However, it cannot be excluded that depending on the task, the type of the anomaly, or the preferences of the operator, a different number of topics might be more suitable. It is also important to realize, that in order to effectively help the operator, the number of topics provided needs to match the operator's understanding of the network. This could entail that there might not be one correct number that fits every combination of anomaly and each operator's way of reasoning. Hence, to identify the number of topics automatically is, though interesting, of lesser importance in our work than to concentrate on the operators' needs and to build a tool where the number of topics can be changed dynamically on request and thereby allow the operator to "zoom-in" and "zoom-out" and see the analysis from different interpretation angles.

6 Conclusions and future work

Our results identify topic modeling as a possible tool for anomaly detection within wireless telecommunication network traffic data. It captures the procedures within the network in the same way that experts would and is able to show the relevant and, to the human expert, meaningful differences between a normal and an incident model in the investigated cases.

Following the expert's analysis work we were able to identify that the changes in relationships between run-time variable readings in the normal and the incident model are one of the most important information sources. This

information can be extracted and shown in an overview such as presented in Fig. 3.

The next step in our work is to validate our approach further, using the extracted representations, rather than the LDAvis visualization, by presenting experts and operators with cases of normal and abnormal network behavior topic models where the ground truth is available but not known to the expert/operator. This will eliminate the possible interpretation bias and also account for possibly different chains of human reasoning. Further user studies describing tests similar to measuring word intrusions and topic intrusions on telecommunication run-time variables will further allow for an objective evaluation of the topics generated.

As mentioned, a user study with telecom operators will reveal if the topic modelling approach will be advantageous to use, both in comparison with how anomaly detection is done today and in combination with other available tools for network monitoring. After operators have gotten used to the tool, their newly developed expert knowledge about anomaly detection using the model can be elicited for further automation.

Acknowledgements Open access funding provided by University of Skövde. This work was supported by the Swedish Knowledge Foundation under grant BISON—Big Data Fusion—in cooperation with Huawei Technologies Sweden AB. We would like to thank Anders Åhlén for sharing his knowledge throughout our work. The topic modeling was performed using the package topicmodels (Grün and Hornik 2011) in R (R Core Team 2017), and the LDAvis visualization was enabled by Sievert and Shirley (2014).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aliu OG, Imran A, Imran MA, Evans B (2013) A survey of self organisation in future cellular networks. *Commun Surv Tutor IEEE* 15(1):336–361
- Blei DM (2012) Probabilistic topic models. *Commun ACM* 55(4):77–84
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3(Jan):993–1022
- Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. *ACM Comput Surv (CSUR)* 41(3):15
- Chaney AJB, Blei DM (2012) Visualizing topic models. In: 6th Inter. AAAI conf. on weblogs and social media
- Chang J, Gerrish S, Wang C, Boyd-graber JL, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A (eds) *Advances in neural information processing systems* 22, Curran Associates, Inc., pp 288–296

- Chuang J, Manning CD, Heer J (2012) Termite: Visualization techniques for assessing textual topic models. In: Proc. of the Inter. working conf. on advanced visual interfaces, ACM, pp 74–77
- Cui W, Liu S, Tan L, Shi C, Song Y, Gao Z, Qu H, Tong X (2011) Textflow: towards better understanding of evolving topics in text. *IEEE Trans Vis Comput Graph* 17(12):2412–2421
- Dou W, Wang X, Chang R, Ribarsky W (2011) Paralleltopics: A probabilistic approach to exploring document collections. In: IEEE conf. on visual analytics science and technology (VAST), IEEE, pp 231–240
- Endsley M, Kiris E (1995) The out-of-the-loop performance problem and level of control in automation. *Hum Factors* 37:381–394
- Endsley MR, Jones DG (2012) Designing for situation awareness: an approach to user-centered design. CRC Press, Boca Raton
- Gardner MJ, Lutes J, Lund J, Hansen J, Walker D, Ringger E, Seppi K (2010) The topic browser: An interactive tool for browsing topic models. In: NIPS workshop on challenges of data visualization, vol 2
- Gretarsson B, O'donovan J, Bostandjiev S, Höllerer T, Asuncion A, Newman D, Smyth P (2012) Topicnets: visual analysis of large text corpora with topic modeling. *ACM Trans Intell Syst Technol (TIST)* 3(2):23
- Grün B, Hornik K (2011) topicmodels: an R package for fitting topic models. *J Stat Softw* 40(13):1–30
- Havre S, Hetzler E, Whitney P, Nowell L (2002) Themeriver: visualizing thematic changes in large document collections. *IEEE Trans Vis Comput Graph* 8(1):9–20
- Helldin T, Steinhauer HJ, Karlsson A, Mathiason G (2018) Situation awareness in telecommunication networks using topic modeling. In: 21st Inter. Conf. on Information Fusion
- Imran A, Zoha A (2014) Challenges in 5G: how to empower SON with big data for enabling 5G. *Netw IEEE* 28(6):27–33
- Klein LF, Eisenstein J, Sun I (2015) Exploratory thematic analysis for digitized archival collections. *Digital scholarship in the humanities*, pp 130–141
- Koh LC, Slingsby A, Dykes J, Kam TS (2011) Developing and applying a user-centered model for the design and implementation of information visualization tools. In: 15th Inter. conf. on information visualisation, IEEE, pp 90–95
- Liu S, Zhou MX, Pan S, Song Y, Qian W, Cai W, Lian X (2012) Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans Intell Syst Technol (TIST)* 3(2):25
- Livnat Y, Agutter J, Moon S, Erbacher RF, Foresti S (2005) A visualization paradigm for network intrusion detection. In: Proceedings from the sixth annual IEEE SMC information assurance workshop, pp 92–99
- Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing, association for computational linguistics, EMNLP '11, pp 262–272
- Pirinen P (2014) A brief overview of 5G research activities. In: 1st Inter. Conf. on 5G for ubiquitous connectivity (5GU), IEEE, pp 17–22
- R Core Team (2017) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL <https://www.R-project.org/>
- Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proc. of the 8th ACM inter. conf. on web search and data mining, ACM, WSDM '15, pp 399–408
- Sievert C, Shirley K (2015) LDAvis: interactive visualization of topic models. <https://CRAN.R-project.org/package=LDAvis>
- Sievert C, Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. In: Proc. of the workshop on interactive language learning, visualization, and interfaces, pp 63–70
- Smith A, Lee TY, Poursabzi-Sangdeh F, Boyd-Graber J, Elmqvist N, Findlater L (2017) Evaluating visual representations for topic understanding and their effects on manually generated labels. *Trans Assoc Comput Linguist* 5:1–15
- Smith AM, Hawes T, Myers M (2014) Hierarchy: Interactive visualization for hierarchical topic models. In: Proc. of the workshop on interactive language learning, visualization, and interfaces, pp 71–78
- Tukey JW (1962) The future of data analysis. *Ann Math Stat* 33(1):1–67
- Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009) Evaluation methods for topic models. In: Proc. of the 26th annual inter. conf. on machine learning, ACM, ICML '09, pp 1105–1112
- Xiong L, Poczos B, Schneider J, Connolly A, VanderPlas J (2011) Hierarchical probabilistic models for group anomaly detection. In: Gordon G, Dunson D, Dudík M (eds) Proc. of the 14th Inter. Conf. on artificial intelligence and statistics, PMLR, Fort Lauderdale, FL, USA, Proc. of machine learning research, vol 15, pp 789–797

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.