HYPGRAPHS: An Approach for Analysis and Assessment of Graph-Based and Sequential Hypotheses

Martin Atzmueller[1] and Andreas Schmidt[1] and Benjamin Kloepper[2] and David Arnu[3]

[1] University of Kassel, Research Center for Information System Design
{atzmueller, schmidt}@cs.uni-kassel.de
[2] ABB Corporate Research Center Germany
benjamin.kloepper@de.abb.com
[3] RapidMiner GmbH
darnu@rapidminer.com

Abstract. The analysis of sequential patterns is a prominent research topic. In this paper, we provide a formalization of a graph-based approach, such that a directed weighted graph/network can be extended using a sequential state transformation function, that "interprets" the network in order to model state transition matrices. We exemplify the approach for deriving such interpretations, in order to assess these and according hypotheses in an industrial application context. Specifically, we present and discuss results of applying the proposed approach for topology and anomaly analytics in a large-scale real-world sensor-network.

1 Introduction

The analysis of sequential patterns, e. g., as a sequence of states, is a prominent research topic with broad applicability, ranging from exploring mobility patterns [7] to technical applications [9]. The DASHTrails approach [7] provides a comprehensive modeling approach for comparing hypotheses with such sequences (trails), in order to identify those hypotheses that show the largest evidence concerning the observed data.

This paper presents the HYPGRAPHS analysis approach (extending DASHTrails) for analyzing and assessing sequential hypotheses in the form of transition matrices given a directed weighted network. The application context is given by (abstracted) alarm sequences in industrial production plants in an Industry 4.0 context. Specifically, we consider the analysis of the plant topology and anomaly detection in alarm logs. The assessment of the static structure can help in identifying problems in the setup of the production plant, while dynamic relations can be applied for the analysis of unexpected (critical) situations. Our contribution is summarized as follows:

1. We outline a flexible analytics approach for assessing graph-based and sequential hypotheses in a *graph interpretation* of weight-attributed directed networks.
2. For that, we show how to embed the recent DASHTrails [7] approach for distribution-adapted *modeling and analysis* of sequential hypotheses and trails. Furthermore, we motivate and show the advantages of this state-of-the-art Bayesian approach compared to a typically applied frequentist approach for testing network associations.
3. Furthermore, we outline the application of the proposed approach in an industrial context, for the analysis of plant *structures* in industrial production contexts, as well as for detecting *anomaly indicators* concerning disrupting dynamic processes.

In: Postproceedings of the ECML-PKDD Workshop on New Frontiers in Mining Complex Patterns (NFMCP 2016), LNCS/LNAI, Springer Verlag, Berlin/Heidelberg, Germany, 2017 (In Press)

The remainder of the paper is structured as follows: Section 2 discusses related work. After that, Section 3 outlines the proposed method. Next, Section 4 presents results of a case study of HYPGRAPHS in the industrial context. Finally, Section 5 concludes with a discussion and outlines interesting directions for future work.

2 Related Work

The investigation of sequential patterns and sequential trails are interesting and challenging tasks in data mining and network science, in particular in graph mining and social network analysis. A general view on modeling and mining of ubiquitous and social multi-relational data is given in [5] focusing on social interaction networks. Orman et al. [18] defines a sequence-based representation of networks. Then the sequential patterns are used to characterize communities. For comparing hypotheses and sequential trails, the HypTrails [20] algorithm has been applied to sequential (human) navigational trails derived from web data. In [7] we have presented the DASHTrails approach that incorporates probability distributions for deriving transitions. Extending the latter, the proposed HYPGRAPHS framework provides a more general modeling approach. Using general weight-attributed network representations, we can infer transition matrices as *graph interpretations*, while HYPGRAPHS consequently also relies on Markov chain modeling [15, 21] and Bayesian inference [21, 22].

Sequential pattern analysis has also been performed in the context of alarm management systems, where sequences are represented by the order of alarm notifications. Folmer et al. [11] proposed an algorithm for discovering temporal alarm dependencies based on conditional probabilities in an adjustable time window. To reduce the number of alarms in alarm floods, Abele et al. [3] performed root cause analysis with a Bayesian network approach and compared different methods for learning the network probabilities. Vogel-Heuser et al. [23] proposed a pattern-based algorithm for identifying causal dependencies in the alarm logs, which can be used to aggregate alarm information and therefore reduce the load of information for the operator. In contrast to those approaches, the proposed approach is not only about detecting sequential patterns. We provide a systematic approach for the analysis of (derived) sequential transition matrices and their comparison relative to a set of hypotheses. Thus, similar to evidence networks in the context of social networks, e. g., [17], we model transitions assuming a certain interpretation of the data towards a sequential representation.

The detection and analysis of anomalies and outliers [12] in network-structured data is a novel research area, e. g., for identifying new and/or emerging behavior, or for identifying detrimental or malicious activities. The former can be used for deriving new information and knowledge from the data, for identifying events in time or space, or for identifying interesting, important or exceptional groups [4, 19]. In contrast to approaches for anomaly detection that only provide a classification of anomalous and normal events, we can assess different anomaly hypotheses: applying the proposed approach, we can then generate an anomaly indicator – as a potential kind of second opinion method, e. g., for assessing the state of a production plant that can help for indicating explanations and traces of unusual alarm sequences. Then, using the network representation, we can analyze anomalous episodes relative to structural (plant topology) as well as dynamic (alarm sequence) episodes.

3 Method

In the following section, we first provide an overview on the proposed approach. After that, we discuss the modeling and analysis steps in detail.

3.1 Overview

We start with a set of directed weighted networks. Then, we interpret these weights for constructing transitions between states (denoted by the nodes of the network) and compare this *data* to *hypotheses* that can also be constructed using the network-based formalizations. Adapting the modeling principles of the DASHTrails approach that we have presented in [7] to our network formalism, we model transition matrices given a probability distribution of certain states. We assume a discrete set of such states Ω corresponding to the nodes of the network (without loss of generality $\Omega = \{1, \ldots, n\}$, $n \in \mathbb{N}, |\Omega| = n$). Then, assuming a certain *network interpretation* of the weights of the edges, we construct transitions between states. As shown in Figure 1, we perform the three following steps, that we discuss below in more detail:

1. Modeling: Determine a transition model given the respective weighted network using a *transition modeling function* $\tau : \Omega \times \Omega \to \mathbb{R}$. Transitions between sequential states $i, j \in \Omega$ are captured by the elements m_{ij} of the transition matrix M, i. e., $m_{ij} = \tau(i, j)$. Then, we collect sequential transition matrices for the given network (data) and hypotheses.
2. Estimation: Apply HypTrails, cf. [20] on the given data transition matrix and the respective hypotheses, and return the resulting evidence.
3. Analysis: Present the results for semi-automatic introspection and analysis, e. g., by visualizing the network as a heatmap or characteristic sequence of nodes.

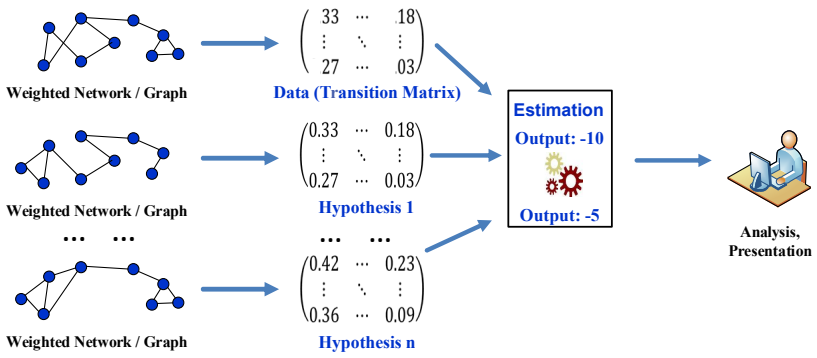

Fig. 1. Overview on the HYPGRAPHS modeling and analysis process.

3.2 Modeling and Comparing Graph-Based Network Interpretations

As outlined above, we derive the transition matrices (modeling transitions between states) using a certain *transition modeling function* $\tau : \Omega \times \Omega \rightarrow \mathbb{R}$, as described below. The transition modeling function τ captures a certain interpretation of these weights. In the case of hypotheses, these correspond to link traversal probabilities from one state to another state, represented by the respective individual nodes. Equivalently, we can represent the obtained directed and weighted graph in the form of an adjacency matrix, where the individual values of an entry (i, j) correspond to the weight of the link between nodes i and j; as an hypothesis this can be interpreted as a transition probability between two states i and j.

Modeling For modeling, we consider a sequential interpretation (according to the first order Markov property) of the original data with respect to the obtained transition probabilities (Markov chain). Thus, using τ, we can model (derived) transition matrices corresponding to the *observed data*, e. g., given frequencies of alarms on measurement points, as well as hypotheses on sequences of alarms. For data transition matrices, we need to map the transitions into derived counts in relation to the data; for hypotheses we provide the (normalized) transition probabilities. That is, for hypothesis testing, we can directly convert the weighted network using the defined transition modeling function (i.e., we convert the obtained values to probabilities by row-normalization).

For observed sequences, we can simply construct transition matrices counting the transitions between the individual states, e. g., corresponding to the set of alarms. Then, $\tau(i, j) = |suc(i, j)|$, where $suc(i, j)$ denotes the successive sequences from state i to state j contained in the sequence. For deriving transition matrices from a probability distribution over certain events, for example, we need to apply a more complex modeling approach. We refer to [7, 20] for more details on modeling and inference, respectively.

Assessment For assessing a set of hypotheses that consider different transition probabilities between the respective states, we apply the core Bayesian estimation step of HypTrails [20] for comparing a set of hypotheses representing beliefs about transitions between states. In summary, we utilize Bayesian inference on a first-order Markov chain model. As an input, we provide a (data) matrix, containing the transitional information (frequencies) of transition between the respective states, according to the (observed) data. In addition, we utilize a set of hypotheses given by (row-normalized) stochastic matrices, modeling the given hypotheses. The estimation method outputs a set of evidence values, for the set of hypotheses, that can be used for ranking these. Also, using the evidence values, we can compare the hypotheses in terms of their significance.

Specifically, hypotheses are expressed in terms of belief in Markov transitions, such that we distinguish between common and uncommon transitions between the respective states. Then, for each hypothesis, we construct the belief matrix for subsequent inference. Given the data (matrix), we elicit a conjugate Dirichlet prior and finally obtain the evidence using marginal likelihood estimation. Here, the evidence denotes the probability of the data given a specific hypothesis. Thus, this can also be interpreted as the relative plausibility of a hypothesis. Then, the hypotheses can be ranked in terms of their evidence.

Furthermore, a central aspect of the method is an additional parameter (k) indicating the *belief* in a given hypothesis: the higher the value ofis k the higher is the belief in the respective hypothesis matrix, i. e., its parameter configuration. Given a lower value of k, the hypothesis is assigned more tolerance, such that other (but similar) parameter configurations become more probable. Then, for assessing a hypothesis, we monitor its performance with increasing k, typically relative to the data itself (as a kind of upper bound), the uniform hypothesis (as a random baseline) and competing hypotheses.

In contrast, the quadratic assignment procedure [14] (QAP) is a frequentist approach for comparing network structures. For comparing two graphs G_1 and G_2, it estimates the correlation of the respective adjacency matrices [14] and tests a given graph level statistic, e. g., the graph covariance, against a QAP null hypothesis. QAP compares the observed graph correlation of (G_1, G_2) to the distribution of the respective resulting correlation scores obtained on repeated random row and column permutations of the adjacency matrix of G_2. As a result, we obtain a correlation value and a statistical significance level according to the randomized distribution scores.

As we will show in our experiments below, the applied Bayesian inference technique has significant advantages compared to the typically applied frequentist approach for comparing networks based on graph correlation using the QAP test [14]: we not only know whether a hypothesis is significantly correlated with the data, but we can also compare hypotheses (and their significance) relative to each other (given Bayes factor analysis, cf. [13]). In particular, this also holds for those hypotheses that are not correlated with the data, obtaining a total ranking for likely and unlikely hypotheses. Furthermore, we can express our *belief* in the hypothesis relative to the data, and analyze the impact of that on the evidence concerning the likelihood estimate.

4 Case Study

Below, we first outline our application context and discuss the instantiation of the proposed approach. After that, we discuss the collected datasets before we describe results of a case study of HYPGRAPHS in the industrial context in detail.

4.1 Application Context

In many industrial areas, production facilities have reached a high level of automation nowadays. Here, knowledge about the production process is crucial, targeting both static relations like the topological structure of a plant and the modeling of operator notifications (alarms), and dynamic relations like unexpected (critical) situations. Assessment of the static structure can help in identifying problems in the setup of the production plant. The dynamic relations involve analytics for supporting the operators, e. g., for diagnosis of a certain problem. The objective of the BMBF funded research project "Early detection and decision support for critical situations in production environments"[4] (short FEE) is to detect critical situations in production environments as early as possible and to support the facility operator with a warning or even a recommendation on how to handle this particular situation. The consortium of the FEE project

[4] http://www.fee-projekt.de

consists of several partners including also application partners from the chemical industry. These partners provide use cases for the project and background knowledge about the production process, which is important for designing analytical methods. In this paper, we utilize HYPGRAPHS in this application context, both for static (topology) and dynamic (alarm log) analysis.

4.2 HYPGRAPHS Instantiation

In an industrial production plant, alarms for certain measurement points occur if the value of the measurement is not within a specified value range. Therefore, by intuition, an alarm sequence (for a given point in time, or interval) represents an abstracted state of the production plant. Then, we can utilize the "normal" long running state of the plant as the "normal behavior", excluding known anomalous episodes.

We perform two kinds of analyses. First, we compare the normal behavior to the overall topology of the plant, i. e., corresponding to transitions between different functional units of the plant. Second, we compare the normal behavior to our anomaly hypotheses, which are defined by the captured anomalies. Doing that, we assume that the sequence of alarms indicates a certain normal or abnormal (process) behavior. We can then compare the (historic) long running state of the plant to the current state for obtaining indicators about possible normal or abnormal situations.

4.3 Dataset

In our experiments, we used a dataset from the FEE project that was collected in a petrochemical plant; it includes a variety of data from different sources such as sensor data, alarm logs, engineering- and asset data, data from the process information management system as well as unstructured data extracted from operation journals and instructions.

We used alarm logs for a period of two months as well as Piping and Instrumentation Diagrams (P&IDs) [10] which represent the topological structure of the facility, i. e., capturing the piping of the considered petro-chemical process along with installed equipment (pumps, valves, heat-exchangers, etc.) and instrumentation used to control the process. P&IDs are usually composed of several sub-diagrams with disjoint system elements. Connections between elements on different P&IDs are captured in textual form at the corresponding pipe or other connecting elements. Commonly, the structuring of P&IDs follows in some way the structure of the captured process and plant capturing different areas. In our dataset, the titles of P&IDs suggest such a structuring of the P&IDs around major equipment like tanks, reactors, processing columns, etc. (e.g. 'Input vessel - desorption plant', 'Preheater - desorption plant', 'Desorber - desorption plant', 'Steam/condensate - auxiliary materials'). We also used text data from the operation journals to verify anomalous events. The characteristics of the applied real-world dataset are shown in Tables 1-2. According to standards [1, 2] P&IDs are used to identify the measurements (temperatures, flows, level, pressures, etc.) in the process, using identifiers of the respective measurement points with up to 5 letters. The alarms in the alarms logs are defined based on measurements captured in the P&ID diagrams, usually as a threshold value on the corresponding measurements; the entries in the alarm log reference the measurements in the P&IDs by a matching identifier.

Table 1. Characteristics of the real-world dataset (petrochemical plant) for a period of two months

	Count
Anomalies	4
P&IDs	63
P&IDs referenced in alarm log	55
Alarms referencing measurement points in P&IDs	59.623
Distinct alarms referencing P&IDs	327
P&ID transitions (between distinct P&IDs)	384
Topological connections (between distinct P&IDs)	299

4.4 Matrix Construction

Before constructing the transition matrices, we first identified anomalous events by looking at the operation journals. We used this background knowledge to divide the dataset into nine disjoint time slots with five normal and four abnormal episodes. For abnormal episodes, we empirically determined a time window of one hour spanning the anomalous event starting half an hour before the event and ending half an hour after the event. In practice, the length of this time window is a parameter that needs to be determined according to application requirements. All nine time slots together covered the whole time (two months). Note that we only used the alarms that could be mapped to a P&ID. The distribution of alarms and P&IDs for the different time slots is shown in Table 2.

Table 2. Overview on normal/abnormal episodes for the real-world dataset (petrochemical plant)

#	Episode	#Alarms	#Distinct alarms	#Distinct P&IDs
1	Normal1	10503	66	34
2	Abnormal1	86	12	9
3	Normal2	8382	91	31
4	Abnormal2	212	14	5
5	Normal3	6130	74	31
6	Abnormal3	220	17	7
7	Normal4	6318	89	29
8	Abnormal4	1516	127	30
9	Normal5	26256	278	44

For each time slot, we constructed a transition matrix M by counting the consecutive transitions in the sequence of the alarm log. Formally, let $A =< a_1, a_2, ..., a_n >$ be a sequence of alarms which represents the alarm log. We created a function, which maps alarms to P&IDs $\mathrm{map}(a_t)$ and retrieved the P&IDs contained in the alarm log $P = \{\mathrm{map}(a_t)|a_t \in A\}$. Then, the weights m_{ij} for the $|P| \times |P|$ transition matrix M are given by the number of transitions from p_i to p_j with $(p_i, p_j) \in P \times P$:

$$m_{ij} = |\{(a_t, a_{t+1}), a_t, a_{t+1} \in A, \mathrm{map}(a_t) = p_i, \mathrm{map}(a_{t+1}) = p_j\}|$$

For the data matrix corresponding to the alarm data, we can then just utilize the obtained count data, denoting the number of transitions between the states. For creating hypotheses, we normalize the data by row in order to obtain a stochastic matrix.

We also extracted data from the P&IDs corresponding to the plant organization in terms of functional units. As described above each P&ID corresponds to such a functional unit, containing several sensors that can then trigger respective alarms if the corresponding measurements are not within a specified value range. A P&ID shows the process and instrumentation structure and also links to other P&IDs with respect to certain flows (material, energy, information) that connects the process structure. Given the P&IDs in PDF format, we converted the data to XML and extracted the necessary information for modeling all possible (directed) links between the individual P&IDs in a network-based representation of the overall plant modeling.

4.5 Results and Discussion

According to our hypotheses, we expect that the functional units of the plant also model functional dependencies as observed by alarm sequences. Furthermore, we expect, that normal episodes (sequences) should be "close" to the normal (long running) behavior. Accordingly, abnormal sequences should be "away" from the normal (reference) behavior – in terms of evidence. As we will see below, we can confirm these hypotheses using Bayes factor analysis [13]. As a baseline, we furthermore apply the presented QAP method. Since a (data) transition matrix should be explained best by its corresponding hypothesis, we constructed a respective row-normalized data transition matrix. In addition, we constructed a uniform hypotheses (square matrix, all entries being 1) as a random baseline. A good hypothesis explaining the normal behavior should be between both, however, relatively close to the data.

Topological Analysis As previously discussed, the document structure of P&IDs capture to a certain extent the structure of the process plant they describe. Simply put, the designer of the P&IDs decided to put elements on the same diagram because they are closely related (although, sometimes graph layout consideration might override this rule of thumb). Consequently, the measurements captured on a P&ID are more closely related to measurements across different P&IDs. Since measurements are used to define alarm messages, it seems a valid assumption that consequently alarms in the alarm logs should reference measurements on the same P&ID with a higher probability than measurements from different P&IDs. Based on this assumption, we formulated our first hypothesis to test HYPGRAPHS on the industrial dataset: For topological analysis, we utilized the given P&ID graph containing directed links between the P&IDs. We checked whether the alarm sequences (normal behavior) can be explained by a uniform topology model, where we assume that transitions between all linked P&IDs are equally likely. The results are shown in Figure 2. We observe that the uniform topology hypothesis does not explain the data well since it is significantly away (larger k) compared to the data and close to the random baseline. In contrast, an "encapsulated topology" hypothesis fits the data relatively well, assuming that transitions in alarm sequences mainly occur local to the specific P&IDs. This confirms our expectations and indicates a good performance of plant and alarm management in general.

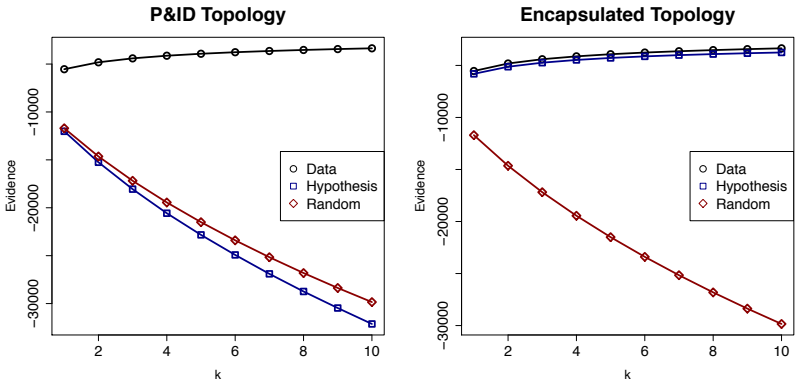

Fig. 2. Topological analysis: Uniform topology hypothesis and local topology hypothesis.

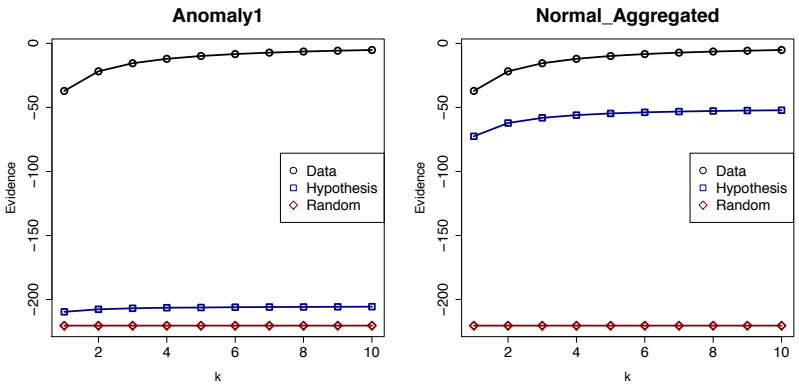

Fig. 3. Artificial local topology baseline: Example of an anomalous and a normal hypothesis.

Furthermore, we double-checked the data against an artificial baseline, assuming only transitions local to P&IDs (in that case, the transition matrix becomes a diagonal matrix). Results are shown in Figure 3. We observed strict differences between normal and abnormal episodes, two examples are shown in the figure. While abnormal situations are far away from the local hypothesis, normal situations are significantly closer, but these, cannot "explain" only local transitions, indicating that most transitions but not all conform to this artificial situation. We also checked the rankings of the normal and abnormal episodes comparing the respective hypotheses to the real data (normal behavior) and the artificial local topology baseline. Using Kendall's-Tau as a correlation measure (0.61), the ranking was not very consistent, indicating that the local topology assumption alone is too simple in order to be explainable by the observed data.

Overall, we observe that we can verify structural modeling assumptions using HYPGRAPHS (given in the P&ID structure) using the collected data from the alarm logs. We already observe distinct differences between abnormal and normal episodes.

Anomaly Analytics In the start phase of the FEE project, a series of workshops and interviews were executed for identifying potential Big Data and analytics applications. One of the identified analytics tasks was anomaly detection. The idea behind that application scenario is that retrospective analysis of disrupting events often uncovers that a situation could have been handled better, if the operators or process experts had been involved earlier and would have been pointed to the relevant data. Thus, we developed a description of the current and desired situation to identify the right analytics questions:

- **Current Situation**:
  - *Who*: Operator in the operating room, shift leader (in the operating room), process engineer, process manager (in the office).
  - *What*: Anomalies (e. g., uncommon oscillations) in a plant need to be recognized as early as possible. If such cases are not recognized by the operator, serious problems can occur (product is not usable, unplanned plant shutdown, etc.) and staff with higher expertise need to be informed.
  - Challenge: Anomalies are not easy to detect manually. New technologies like advanced controllers make anomalies even more difficult to detect. Furthermore, operators usually inform an expert when a problem has occurred and they are not able to handle it. In addition, diagnostics of an anomaly by process engineers and managers is usually time-consuming.
- **Desired Situation**:
  - System: informs the operator about a possible anomaly. The operator performs an analysis and diagnosis of the situation and informs the expert.
  - Expert: automated updates about possible anomalies; can track long term trends.
  - Users: pointed to relevant measurements for supporting diagnostic activities.

In the context of anomaly analytics, our results indicate the significance of the proposed HYPGRAPHS approach for specifically supporting analysis and diagnosis tasks.

In particular, for anomaly analysis of the alarm data, we used the partitioning of the dataset into normal and abnormal episodes. Then, we checked both abnormal and normal situations against the assumed "normal behavior" of the plant that is observed for the long running continuous process. In the analysis, we applied a typical estimation procedure using separate training and tests sets, such that the data and the tested hypotheses do not overlap in time. However, since we have only had data covering a two months period available we also tested the hypotheses against the aggregated normal behavior covering all normal episodes. It turned out, that the findings reported below are also consistent across these different evaluation periods; we observe the same (significant) trends, confirming the individual results even on larger scale.

Figure 4 shows the different anomaly hypotheses corresponding to the different anomaly episodes (cf. Table 2). We observe that the anomalies are well distinguishable (using Bayes factor analysis [13]). The anomalies are "well away" from data (more than factor 3 for higher k), indicating a significant deviation from the data. Furthermore, we observe distinct characteristics of the anomalies, observing the trends with increasing k. Anomalies 1-3 are of the same class and show similar characteristics, while Anomaly 4 conforms to another real-world class of a disrupting event, also showing different characteristics in terms of evidence. We also performed an analysis using the QAP procedure for the anomaly data, correlating the transition matrices corresponding to the normal behavior and the abnormal episodes. These results support the findings of

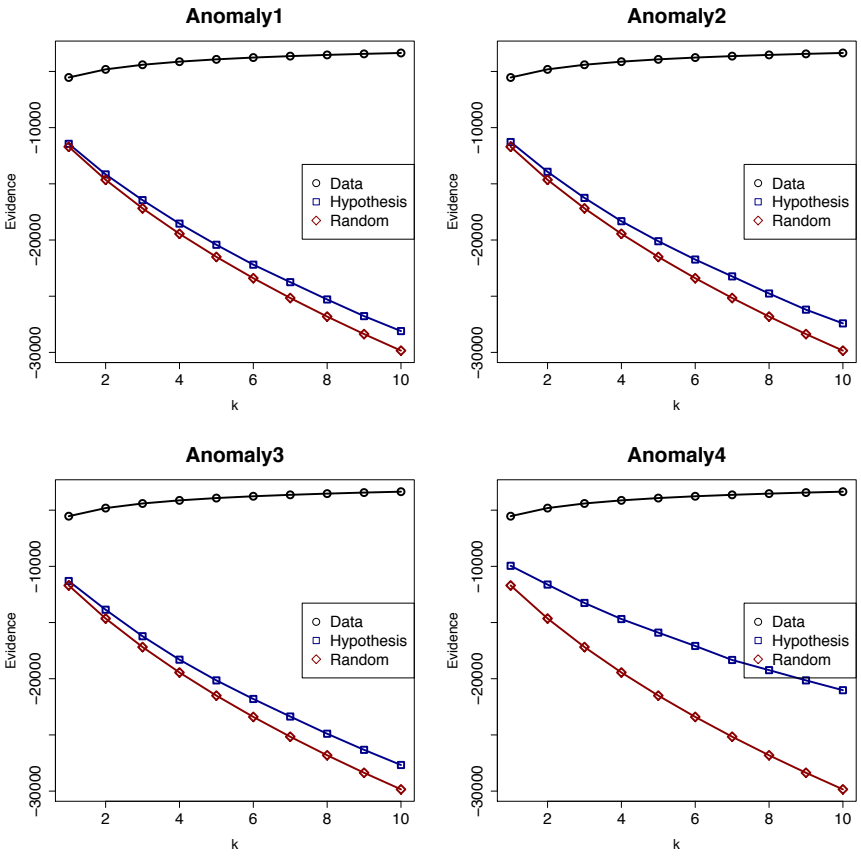

Fig. 4. Normal behavior (data) compared to different anomaly episodes (Anomaly1-4) and a random baseline (uniform hypothesis).

the Bayesian approach, showing a correlation close to zero that was not significant. However, while confirming the deviation, QAP does not allow to derive a (significance-based) ranking of the different hypotheses here, in contrast to our proposed approach.

Figure 5 shows results of comparing exemplary normal episodes (as hypotheses) with the normal behavior (data) – the results for the rest of the normal episodes show equivalent trends. We observe significant differences compared to the anomaly hypotheses. Using the Bayes factors technique, we also observe that the normal behavior is well detectable, the hypotheses are sufficiently "close" to the data hypotheses. In addition, we also compared shorter normal periods (using random samples of the normal behavior) in order to exclude control for the different sizes of the alarm distributions. The bottom right chart of Figure 5 shows an example - the findings confirm our results for the other episodes well.

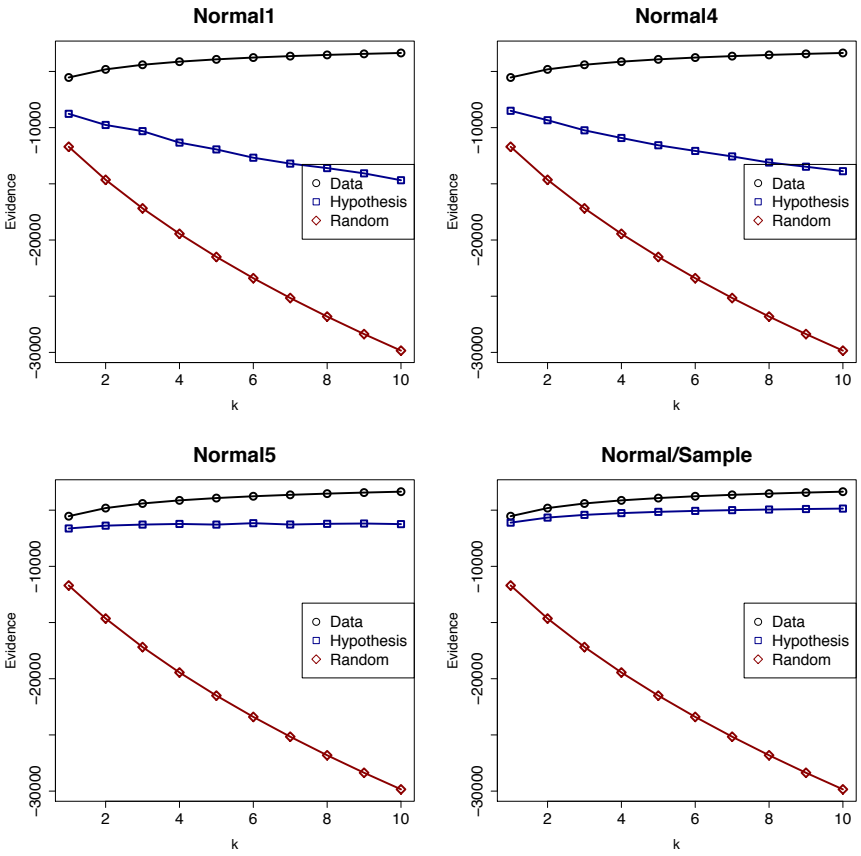

Fig. 5. Normal behavior (data) compared to different normal episodes and a random baseline (uniform hypothesis).

For the normal episodes, we also applied QAP analysis, using the graph correlation measure on transition matrices corresponding to the normal behavior and the respective normal episodes described above. Here, we observed significant ($p = 0.01$) correlation values between 0.42 and 0.72, with a ranking of the normal hypotheses that is consistent with the Bayesian approach. In essence, this suggests that our findings are rather robust against the selected statistical measure.

Retrospective as well as realtime analysis can be supported, for example, by according visualization approaches summarizing anomalous episodes in the form of heatmaps, or by directly tracing anomalous sequences on a detailed level of analysis. Figure 6 demonstrates an example of a heatmap visualization, showing data for the (aggregated) anomalies 1-3 compared to the long term behavior of the data.: Rows/columns of the matrix refer to the individual P&IDs. The aggregations refer to different types of real-world anomalies and we can observe distinct “fingerprints” of the transitional episodes.

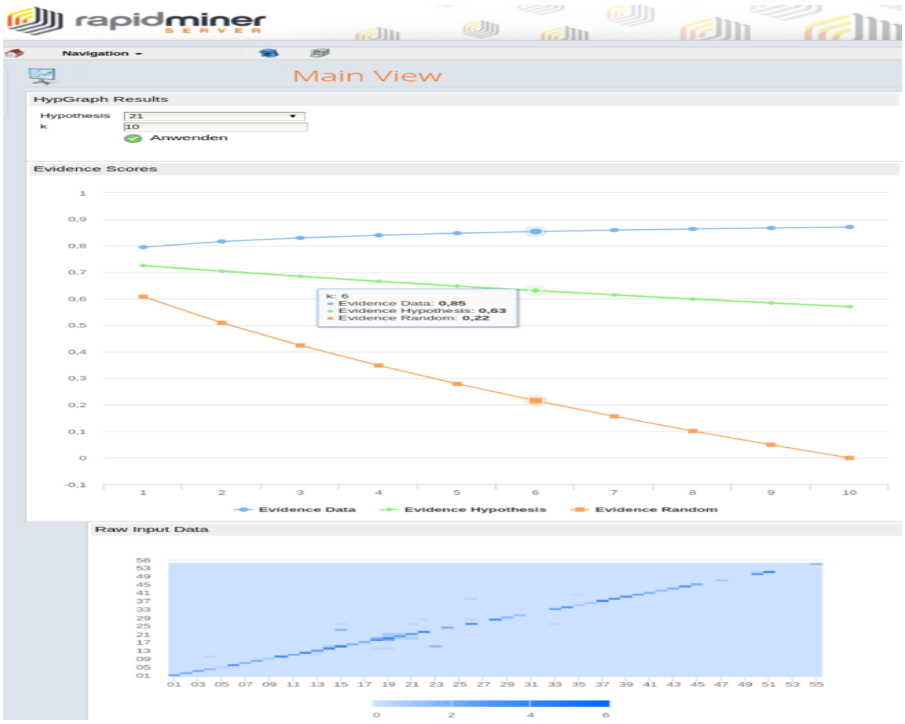

Fig. 6. Example of a dashboard with a heatmap visualization, showing "fingerprints" for the long term behavior and for an anomalous sequence. Rows/columns refer to the individual P&IDs.

Then, by inspecting the different cells (corresponding to transitions of alarms between a pair of P&IDs), the respective data points (sequences of alarms) can be assessed in detail, e. g., showing the corresponding alarm messages or sensor reading. Please note, that this visualization can be applied for static data, i. e., for retrospective analysis, as well as for dynamic analysis, e. g., utilizing a suitable time window for data aggregation on the current (alarm) log data stream.

In summary, these analysis results indicate the significance of the HYPGRAPHS approach for anomaly analytics, concerning detection, analysis and diagnosis tasks. Applying HYPGRAPHS we can compare different hypotheses to the "normal behavior" and identify normal and abnormal episodes in a data-driven way. In contrast to typical frequentist approaches like QAP, we can obtain a ranking of both the normal and abnormal episodes, enabling a comprehensive view on the data for anomaly analytics, complemented by suitable visualizations. Furthermore, there are several visualization options to be used for dashboards, e. g., the obtained evidence plots, using extended heatmaps, or a detailed view on sequences of nodes corresponding to individual alarms.

4.6 Big Data Aspects

With time periods longer than two months or with very detailed sensor readings, the amount of data can quickly get overwhelming for normal computation systems. In this case, a distributed storage and computation system can handle the requirements of evaluating several years of production data. The RapidMiner [16] platform, for example, can be integrated with Hadoop systems such that preprocessing and analytical processes built on a local machine can be transfered to the big data environment. In the context of the FEE project, we target a two layered architecture where long running and computationally expensive processes run in the Hadoop infrastructure and either the prepared data or the final models, in this case the transition matrix M, can be applied on a local machine. The computation can be executed, e. g., in a Spark/MapReduce [8] process and the orchestration and deployment can be handled with RapidMiner, for which the HYPGRAPHS approach is already implemented as an independent extension.[5]

5 Conclusions

This paper outlined the HYPGRAPHS approach for modeling and comparing graph-based and sequential hypotheses using first-order Markov chain models. Our application context is given by structural and anomaly analytics in Industry 4.0 contexts, i. e., of (abstracted) alarm sequences in industrial production plants. We applied a real-world dataset in an Industry 4.0 context, specifically in the scope of the FEE project.

In summary, we considered the analysis of the plant topology and anomaly analytics in alarm logs, which was identified as one major application in the project. Our results indicate that the proposed HYPGRAPHS approach is well suited for analyzing and assessing the transition networks, respectively the corresponding alarm sequences. We could identify distinct differences between abnormal and normal episodes, e. g., in order to derive an anomaly indicator. We also verified the modeling of plant topology and alarm setup. The results can help for analysis and inspection of the corresponding alarm sequences, e. g., for detailed analysis and diagnosis of anomalies. This enabled directly the inspection, for example, of a deviating sequence through a drill-down into the data. Furthermore, results can be transparently visualized, e. g., in the form of heatmaps, and embedded into Big Data dashboards.

For future work, we aim to extend the analysis using high diversity data, i. e., with longer time periods, different event and anomaly settings. We are also investigating options for detecting descriptive anomaly patterns [6]. Furthermore, including more background knowledge on known relations on plant configuration and the extension to an unsupervised approach for anomaly detection is another interesting direction.

Acknowledgements

This work was funded by the BMBF project FEE under grant number 01IS14006. We wish to thank Leon Urbas (TU Dresden) and Florian Lemmerich (GESIS, Cologne) for helpful discussions, also concerning Florian's implementation of HypTrails[6] [20].

[5] `https://github.com/rapidminer/rapidminer-extension-hypgraphs`
[6] `https://bitbucket.org/florian_lemmerich/hyptrails4j`

References

1. Ansi/isa s51.1-1979 (r1993): Process instrumentation terminology
2. ISO 14617-6:2002 Graphical Symbols for Diagrams – Part 6: Measurement & Ctrl. Functions
3. Abele, L., Anic, M., Gutmann, T., Folmer, J., Kleinsteuber, M., Vogel-Heuser, B.: Combining Knowledge Modeling and Machine Learning for Alarm Root Cause Analysis. In: MIM. pp. 1843–1848. International Federation of Automatic Control (2013)
4. Akoglu, L., Tong, H., Koutra, D.: Graph Based Anomaly Detection and Description. Data Min Knowl Disc 29(3), 626–688 (May 2015)
5. Atzmueller, M.: Data Mining on Social Interaction Networks. JDMDH 1 (2014)
6. Atzmueller, M.: Detecting Community Patterns Capturing Exceptional Link Trails. In: Proc. IEEE/ACM ASONAM. IEEE Press, Boston, MA, USA (2016)
7. Atzmueller, M., Schmidt, A., Kibanov, M.: DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In: Proc. WWW 2016 (Companion). IW3C2 / ACM (2016)
8. Becker, M., Mewes, H., Hotho, A., Dimitrov, D., Lemmerich, F., Strohmaier, M.: SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails. In: Proc. WWW (Companion). ACM Press, New York, NY, USA (2016)
9. Buddhakulsomsiri, J., Zakarian, A.: Sequential Pattern Mining Algorithm for Automotive Warranty Data. Computers & Industrial Engineering 57(1), 137 – 147 (2009)
10. Cook, R.: Interpreting Piping and Instrumentation Diagrams. Blog-Entry (September 2010), `http://www.aiche.org/chenected/2010/09/interpreting-piping-and-instrumentation-diagrams`
11. Folmer, J., Schuricht, F., Vogel-Heuser, B.: Detection of Temporal Dependencies in Alarm Time Series of Industrial Plants. Proc. 19th IFAC World Congr pp. 24–29 (2014)
12. Hawkins, D.: Identification of Outliers. Chapman and Hall, London, UK (1980)
13. Kass, R.E., Raftery, A.E.: Bayes Factors. J Am Stat Assoc. 90(430), 773–795 (1995)
14. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Soc Networks 9, 171–186 (1987)
15. Lempel, R., Moran, S.: The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. Computer Networks 33(1), 387–401 (2000)
16. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: Rapid prototyping for complex data mining tasks. In: Proc. KDD. pp. 935–940. ACM, New York, NY, USA (2006)
17. Mitzlaff, F., Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Community Assessment using Evidence Networks. In: Analysis of Social Media and Ubiquitous Data. LNAI, vol. 6904. Springer, Heidelberg, Germany (2011)
18. Orman, G.K., Labatut, V., Plantevit, M., Boulicaut, J.F.: A Method for Characterizing Communities in Dynamic Attributed Complex Networks. In: Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 481–484 (2014)
19. Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., Samatova, N.F.: Anomaly Detection in Dynamic Networks: A Survey. WIREs: Comp. Statistics 7(3), 223–247 (2015)
20. Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Hyptrails: A Bayesian Approach for Comparing Hypotheses about Human Trails. In: Proc. WWW. ACM, New York, NY, USA (2015)
21. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Memory and Structure in Human Navigation Patterns. PLoS ONE 9(7) (2014)
22. Strelioff, C.C., Crutchfield, J.P., Hübler, A.W.: Inferring Markov Chains: Bayesian Estimation, Model Comparison, Entropy Rate, and Out-of-Class Modeling. Physical Review E 76(1), 011106 (2007)
23. Vogel-Heuser, B., Schütz, D., Folmer, J.: Criteria-based Alarm Flood Pattern Recognition Using Historical Data from Automated Production Systems (aPS). Mechatronics 31 (2015)