HYPGRAPHS: An Approach for Modeling and Comparing 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 first 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 compare these and according hypotheses in the context of an industrial application. Specifically, we present first results of applying the proposed approach for topology analysis and anomaly analytics in a large-scale 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 [9], to technical applications [10]. The DASHTrails approach [9] 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 modeling and analysis approach (extending DASHTrails) for analyzing and comparing 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 the HYPGRAPHS approach for comparing graph-based and sequential hypotheses in a *graph interpretation* of weighted, directed & attributed networks.
2. For that, we show how to embed the recent DASHTrails [9] approach for distribution-adapted *modeling and analysis* of sequential hypotheses and trails: we motivate and discuss the advantages of this state-of-the-art Bayesian approach compared to a typically applied frequentist approach for testing network associations.
3. Furthermore, we exemplify 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.

2 Related Work

The investigation of sequential patterns (as sequences of events) and sequential trails are interesting and challenging tasks in data mining and network science, in particular in graph mining and social network analysis, e. g., [4, 14]. A general view on modeling and mining of ubiquitous and social multi-relational data is given in [5] focusing on social interaction networks, captured during certain events, e. g., during conferences or in workgroup environments [7]. Navigational patterns, as sequential (link) patterns in online systems, have been analyzed and modeled, e. g., in [18, 20]. In contrast to that, our approach focuses on the modeling and comparing sequential patterns as hypotheses in a graph-based network representation. For comparing hypotheses and sequential trails, the HypTrails [19] algorithm has been proposed. In [9] we have presented the DASHTrails approach that incorporates probability distributions for deriving transitions utilizing HypTrails. Extending the latter, the proposed HYPGRAPHS framework models transition matrices as *graph interpretations*, while HYPGRAPHS consequently also relies on Markov chain modeling [16, 20] and Bayesian inference [20, 21].

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. [12] 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. [22] 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 its 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. Then, we can identify important influence factors.

3 Method

We first provide an overview on the proposed approach. After that, we first describe our industrial application context, before we outline the proposed approach in more 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. Transferring core principles of the DASHTrails approach for modeling and analysis of distribution-adapted sequential hypotheses and trails that we have presented in [9] 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. We perform the three following steps, that we outline below in more detail:

1. Modeling: Determine a transition model given the respective weighted network using a *transition modeling function* $\tau : \Omega \times \Omega \rightarrow \mathbb{R}$. 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 construct according sequential transition matrices.
2. Estimation: Apply HypTrails, cf., [19] 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.

3.2 Industrial 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: Development of assistance systems to support plant operators in critical situations"[4] (short FEE) [8] 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 about how to handle this particular situation. In this paper, we apply the HypGraphs approach in this application context, both for static and dynamic analysis.

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 anomalies captured by according anomaly hypotheses. Doing that, we assume that the sequence of alarms indicates 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. Doing that, we can identify specific exceptional alarm sequences (as indicators) both for anomaly detection and diagnostics. Here, for example, we can target the (modeled) plant behavior (regarding documentation and process configuration), as well as process diagnosis and introspection.

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

3.3 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* $\tau : \Omega \times \Omega \rightarrow \mathbb{R}$. For that, we can utilize, for example, weights of a network corresponding to observed alarm frequencies. The transition modeling function τ then captures a certain interpretation of these weights. In the case of hypotheses, these are derived from 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.

For assessing a set of hypotheses that consider different transition probabilities between the respective states, we apply the core Bayesian estimation step of HypTrails [19] 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, i. e., frequencies of transitions between the respective states, according to the (observed) data. In addition, we utilize a set of specific hypotheses given by (row-normalized) stochastic matrices. 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.

Following [19], 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 k the higher 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.

The quadratic assignment procedure [15] (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 [15] 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, resulting in a statistical significance level.

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 [15]: we do 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.

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; hypotheses are based on the (normalized) transition probabilities, i. e., utilizing the weighted network and the defined transition modeling function. 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. Also, more advanced distribution-adapted modeling options, e. g., window-based or uniform smoothing are possible. We refer to [9, 19] for more details on modeling and inference.

4 Case Study

In Industry 4.0 environments like complex industrial production plants, intelligent data analysis is a key technique. Below, we first outline the collected datasets before we describe results of a case study of HypGraphs in the industrial context in detail.

4.1 Dataset

In our experiments, we used a dataset from the FEE project that was collected in a petrochemical plant and 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 operation instructions. We used alarm logs for a period of two months as well as Piping and Instrumentation Diagrams (P&IDs) [11] 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 data set, 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

For constructing the dataset, 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 (anomalous) 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&IDs 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). For creating hypotheses (below), we utilize the window-adjusted counts, representing the characteristic transitions, the uniform topology basis hypothesis, and topological transitions, respectively. We also extracted data from the P&IDs w.r.t. the plant organization in terms of functional units.

4.2 Results and Discussion

As a general hypothesis, 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]. Since a (data) transition matrix should be explained best by its according 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. Then, a good hypothesis explaining the normal behavior should be between both, however, relatively close to the data.

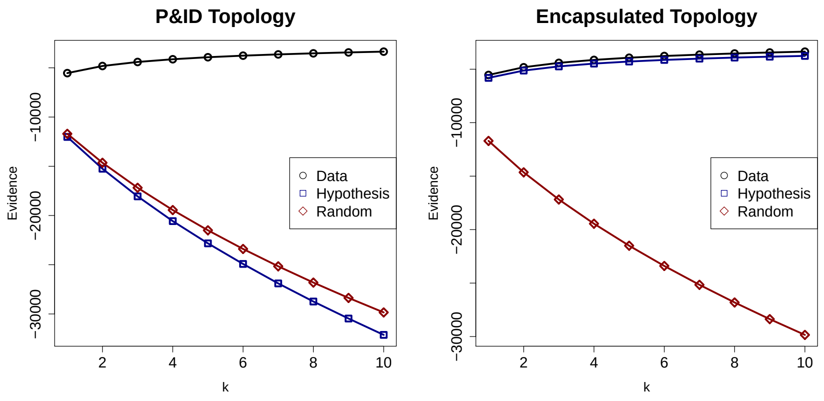

Fig. 1. Topological analysis: Uniform topology hypothesis, local topology hypothesis, artificial local baseline.

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 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 data set: We utilized the given P&ID graph containing directed links between the P&IDs. Then, 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 1. 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, also indicating a good performance of plant and alarm management in general, as observed in the data.

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), cf., Figure 2. Normal situations are significantly close, however, 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.

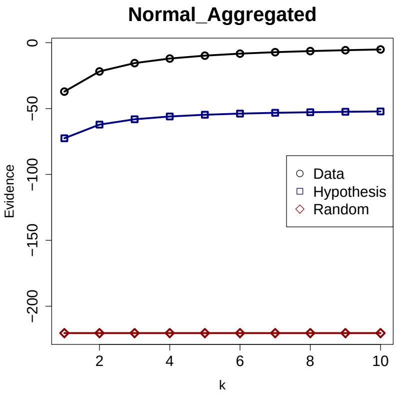

Fig. 2. Artificial local topology baseline: Example of a normal hypothesis.

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 an operator, serious problems can occur (product 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**: Early information about a possible (high probability) anomaly.
  - 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 supporting specifically analysis and diagnosis tasks, as described in the following.

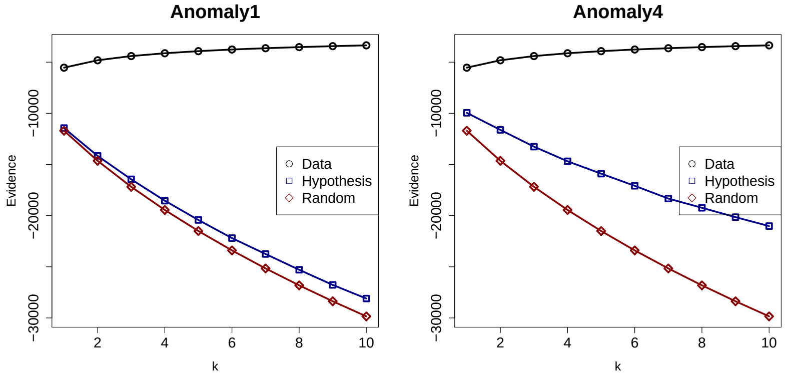

Fig. 3. Normal behavior (data) compared to different exemplary anomaly episodes formalized as hypotheses (Abnormal1 & Abnormal4) and a random baseline (uniform hypothesis). Other situations showed similar findings.

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, i. e., 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.

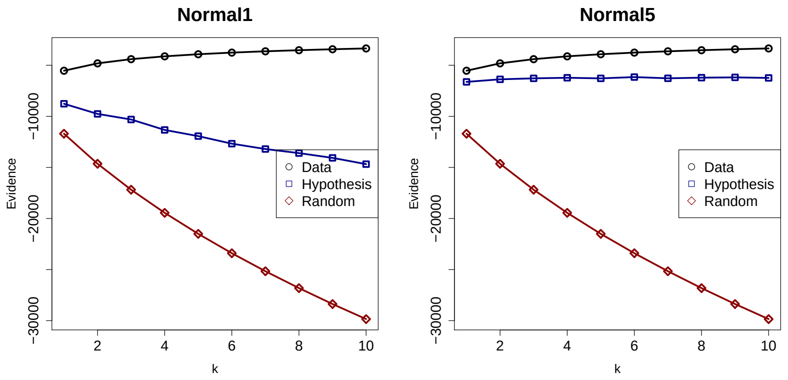

Fig. 4. Normal behavior (data) compared to different exemplary normal episodes and a random baseline (uniform hypothesis). Other situations showed similar findings.

Figure 3 shows the different anomaly hypotheses corresponding to the different anomaly episodes (cf., Table 2) for anomalies 1 & 4 (we observed similar results for the other situations). 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 as a baseline, correlating the transition matrices corresponding to the normal behavior and the abnormal episodes. These results support the findings of 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 4 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 4 shows an example - the findings confirm our results for the other episodes well. For the normal episodes, we also applied QAP analysis as a baseline, 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. 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. 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 results indicate the potential and significance of the HypGraphs approach for anomaly analytics, concerning detection, analysis and diagnosis tasks: 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.

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.

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, while 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. We can then directly inspect, for example, a deviating sequence since the approach allows for a direct 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 more (diverse) data, i. e., longer time periods and different event and anomaly settings, and investigate options for detecting descriptive anomaly patterns [6]. Furthermore, including more background knowledge on known relations on plant configuration is another interesting direction.

Acknowledgements

This work was funded by the BMBF project FEE under grant number 01IS14006E. We wish to thank Leon Urbas (TU Dresden) and Florian Lemmerich (GESIS, Cologne) for helpful discussions, also concerning Florian’s implementation of HypTrails[5] [19].

[5] `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. Atzmueller, M.: Analyzing and Grounding Social Interaction in Online and Offline Networks. In: Proc. ECML/PKDD. LNCS, vol. 8726, pp. 485–488. Springer, Heidelberg, Germany (2014)
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., Becker, M., Kibanov, M., Scholz, C., Doerfel, S., Hotho, A., Macek, B.E., Mitzlaff, F., Mueller, J., Stumme, G.: Ubicon and its Applications for Ubiquitous Social Computing. New Review of Hypermedia and Multimedia 20(1), 53–77 (2014)
8. Atzmueller, M., Kloepper, B., Mawla, H.A., Jäschke, B., Hollender, M., Graube, M., Arnu, D., Schmidt, A., Heinze, S., Schorer, L., Kroll, A., Stumme, G., Urbas, L.: Big Data Analytics for Proactive Industrial Decision Support: Approaches & First Experiences in the Context of the FEE Project. atp edition 58(9) (2016), (In Press)
9. 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)
10. Buddhakulsomsiri, J., Zakarian, A.: Sequential Pattern Mining Algorithm for Automotive Warranty Data. Computers Industrial Engineering 57(1), 137 – 147 (2009)
11. Cook, R.: Interpreting Piping and Instrumentation Diagrams. Blog-Entry (September 2010), `http://www.aiche.org/chenected/2010/09/interpreting-piping-and-instrumentation-diagrams`
12. 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)
13. Kass, R.E., Raftery, A.E.: Bayes Factors. J Am Stat Assoc. 90(430), 773–795 (1995)
14. Kibanov, M., Atzmueller, M., Scholz, C., Stumme, G.: Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions. Sci Chi Inf Sci 57 (2014)
15. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Soc Networks 9, 171–186 (1987)
16. Lempel, R., Moran, S.: The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. Computer Networks 33(1), 387–401 (2000)
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 (2011)
18. Pirolli, P.L., Pitkow, J.E.: Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations. World Wide Web 2(1-2) (1999)
19. 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)
20. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Memory and Structure in Human Navigation Patterns. PLoS ONE 9(7) (2014)
21. 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)
22. Vogel-Heuser, B., Schütz, D., Folmer, J.: Criteria-based Alarm Flood Pattern Recognition Using Historical Data from Automated Production Systems (aPS). Mechatronics 31 (2015)