PROCESS MINING APPROACHES TO DETECT ORGANIZATIONAL PROPERTIES IN CYBER-PHYSICAL SYSTEMS

Complete Research

Matzner, Martin, University of Muenster, ERCIS, Münster, Germany, martin.matzner@ercis.uni-muenster.de
Scholta, Hendrik, University of Muenster, ERCIS, Münster, Germany, hendrik.scholta@ercis.uni-muenster.de

Abstract

Cyber-physical systems (CPS) are service systems that connect physical and cyber elements through global networks. CPS put upon sensors and actuators as well as omnipresent status data of smart products in order to facilitate the design of innovative service offerings. CPS typically require the cooperation of several actors such as manufacturers of smart products and service providers. Organizational mining uses event log data produced by information systems to explore organizational structures and to analyze social networks representing communication structures. Hence, organizational mining is a promising approach for changing the organization of several actors in a CPS for the better and for improving the delivery of innovative CPS service offerings. However, approaches of this kind so far have not been discussed with regard to CPS. From a review of the literature on organizational mining this article therefore identifies 18 different approaches, and it discusses their requirements and possible challenges and obstacles of using them in a CPS. The main results from the analysis include that organizational mining may generally be well applicable to CPS while some serious challenges related to CPS characteristics such as distribution in space, different levels of granularity, and time issues require for further research.

Keywords: Organizational Mining, Cyber-Physical Systems, Process Mining, Role Mining.

1 Process Mining and Cyber-Physical Systems

Cyber-physical systems (CPS) are an emerging research field that will have great impact on the life of all human beings. A CPS is a system consisting of physical and cyber elements connected via a global network (Geisberger et al., 2011). Application areas of a CPS comprise advanced automotive systems, traffic control, smart grids, process control, medical systems, manufacturing and many others (Shi et al., 2011). The importance of CPS is stressed by the fact that the research field of CPS is part of the Presidential Innovation Fellows program of the U.S. president (White House, n.d.). The term process mining covers approaches which aim to extract process-related information out of event log data (van der Aalst, 2011), for instance an organizational model of resources involved in a process. Today, many systems keep an event log such as ERP systems, workflow management systems and also embedded systems (Song and van der Aalst, 2008) which are an important part of CPS (Geisberger et al., 2011).

However, the feasibility and benefits of mining organizational properties from process logs in a CPS have so far not been investigated. Consequently, this paper will give answers to the following research
questions: (1) What are mining techniques which derive information regarding the organizational perspective to get insights of the underlying organizational structure of a CPS? (2) What data has to be logged in a CPS so that these techniques can be applied to a CPS? (3) What are possible challenges which prevent using mining techniques in the context of a CPS?

The main contribution of this paper is to provide an overview of algorithms to mine organizational properties and analyze their suitability for CPS. The remainder of this paper is structured as follows: Section 2 provides research background on process mining and directs the focus of this paper. Section 3 explains the research process and gives an overview on organizational model mining approaches. Especially, extant work is analyzed regarding its data requirements to find out which information needs to be present so that the approaches work. Section 4 synthesizes this information with characteristics of CPS scenarios in an attempt to explain potential obstacles of using mining techniques in a CPS and to explore which approaches are most suitable for a CPS. Section 5 concludes with a summary and discusses directions for future research. Appendix A gives a synopsis of the review results.

2 Organizational Mining and Related Research Fields

Process mining may be employed for different purposes: Discovery, conformance, and enhancement (van der Aalst, 2011). Discovery aims at creating models from log data such as process models visualizing the observed behavior or organizational models. Conformance comprises the comparison of a given model to observations derived from the log. Enhancement aims for improving the way an organization operates. For any of these purposes, process mining may analyze event logs from different perspectives: The control-flow, time, case, and organizational perspectives (van der Aalst, 2011). The control-flow perspective deals with the sequences of tasks and their representation as process models. The time perspective investigates the frequency and timing of events, e.g. to identify bottlenecks and to control the workload of resources. The case perspective analyses specific process instances, their paths through the process and the associated elements such as resources and data objects. The organizational perspective investigates the resources involved in a process and their relationships. It uses log data in order to, firstly, explore the organizational structure including organizational units and roles as well as, secondly, to present and analyze the social network indicating the communication structure. Both analysis goals “organizational model mining” and “social network mining” are subsumed by the term “organizational mining” which is the main focus of this paper (Song and van der Aalst, 2008).

Organizational model mining can be further differentiated into task-based organizational model mining and case-based organizational model mining (Song and van der Aalst, 2008). Task-based organizational model mining identifies teams or organizational units as it searches for resources performing similar tasks. Case-based organizational model mining searches for people working on the same cases. Social network mining can construct and investigate either a social network of originators or a social network of organizational entities which consist of several originators (Song and van der Aalst, 2008). Consequently, organizational model mining identifies which elements exist in an organization and groups them either according to similar tasks or similar cases. Social network mining derives how the elements of an organization work together, i.e. how information/cases flow through the organization.

A related research area is role mining – a concept for role engineering – that aims at deriving artifacts for role-based access control (RBAC) systems (Coyne, 1995). RBAC systems do not grant permissions directly to performers but associate permissions to roles and assign performers to appropriate roles (Ferraiolo et al., 1995). There exist top-down and bottom-up approaches to identify roles and permissions (Jafari et al., 2009). The top-down approaches analyze processes from a high-level and may use process models and job descriptions to derive roles (Jafari et al., 2009). The bottom-up approaches – commonly known as role mining approaches – take actual performer-permission assignments from system configuration data (Molloy et al., 2008). A permission can allow a performer to access one or several objects or give the necessary privileges to perform specific actions in the system (Sandhu et al., 1999). Role mining algorithms can either use current performer-permission assign-
ments configured in a system or actually practiced assignments as observed in a log (Jafari et al., 2009). If an event log records the permissions which have been used to perform a task in some way, then role mining algorithms can also be used to derive organizational properties out of an event log.

The scope of this paper is visualized in Figure 1. The relevant areas are gray-shaded. The paper focuses the organizational perspective of process mining. Role mining approaches based on log data are considered as relevant while top-down role engineering approaches and role mining approaches which rely on system configuration data are not. Notably, there exist further areas which deal with the organizational perspective that were excluded such as staff assignment mining and performance analysis. Staff assignment mining aims at deriving rules which define the profile a person needs to have in order to be able and allowed to perform a task (Ly et al., 2005). However, this topic does not include information about the organizational relationship between people. Performance analysis is excluded as it affects all perspectives (Nakatumba and van der Aalst, 2009) and does not reveal information about an organizational structure. Moreover, approaches which are not based on an event log are ignored such as attempts to construct a social network based on e-mail interactions (e.g., Farnham et al., 2004). Obviously, CPS is focused in this paper but has not been linked to the other areas yet.

![Figure 1. Organizational Mining and Related Research Fields: Scope of this Paper](image)

### 3 Approaches for Mining Organizational Properties

#### 3.1 Research Method

In order to discover suitable approaches to mine organizational properties in a CPS, a literature review has been performed based on the works of Levy and Ellis (2006) and Webster and Watson (2002). A keyword search has been done by one researcher on these sources: Google Scholar, SpringerLink, CiteSeerX, and ScienceDirect. Combinations of these keywords have been used: Organization, Organizational, Model, Role, Social Network, Mining, Process, Workflow, Event, Log, Audit Trail, History Data, CPS, Cyber-Physical System. Additionally, forward and backward searches have been conducted. After identifying relevant papers according to headline and abstract, these papers have been read and the criteria in Table 1 have been developed to categorize and classify the approaches. Finally, eleven papers have been rated as relevant sources. Appendix A summarizes the categorization of the approaches and should be consulted regularly while reading the following subsections.

The criterion **Discipline** reflects the explanations in Section 2. In Section 4 **Required Log Information** is used to analyze the approaches’ applicability in a CPS while **Data Input** indicates if other data input than a log is needed. **Output** describes the desired analysis goal. **Domain** indicates a restrictive domain of an approach which may hinder its adoption in a CPS. A **performer** is a resource object of a company such as a human being or a machine. The **event type** describes the kind of an event such as start or delegate. Some algorithms need to know the temporal order of events, whereas others need an exact timestamp. The former is indicated by **time**, the latter by **timestamp**. The approaches have been evalu-
ated according to the criteria as well as structured and ordered systematically. These results and deeper descriptions of the approaches and their characteristics are presented in the following.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description and/or Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline</td>
<td>Task-based organizational model mining, case-based organizational model mining, social network of originators mining, social network of organizational entities mining or role mining</td>
</tr>
<tr>
<td>No.</td>
<td>A unique number to identify the approach in this paper</td>
</tr>
<tr>
<td>Source</td>
<td>The sources which present the approach</td>
</tr>
<tr>
<td>Required Log</td>
<td>The information the log has to offer regarding an event so that the approach can be applied including case, task, performer, time, timestamp, event type, permission, session</td>
</tr>
<tr>
<td>Information</td>
<td>The data basis the approach has to be supplied with</td>
</tr>
<tr>
<td>Data Input</td>
<td>Output, the result of the approach</td>
</tr>
<tr>
<td>Domain</td>
<td>Indicates whether the approach is restricted to a dedicated application area</td>
</tr>
</tbody>
</table>

Table 1. Criteria for Characterizing the Detected Approaches

3.2 Approaches for Organizational Model Mining

Approach 1 (Name in source: “Default mining”) is the simplest method for organizational model mining. Like the other approaches 2 to 7 for task-based organizational model mining, it requires a log containing information about the performer and task of an event as input and returns a task-based organizational model. For every task, it creates one organizational unit and assigns all performers to it who have executed this task at least once (Song and van der Aalst, 2008 p.306). A disadvantage is the dependence on the quantity of tasks. Hence, if there are a lot of tasks there will also be many organizational units – one for each task.

All the next approaches use a performer-task matrix. Given a log storing the performer and task of each event, this matrix shows how often each performer has conducted each task (Song and van der Aalst, 2008). The example in Table 2 (left) assumes the following events were recorded: (Benjamin, task2); (Daniel, task3); (Moritz, task1); (Benjamin, task1); (Daniel, task3); (Daniel, task3). The first string indicates the performer, the second string represents the task of an event. Approach 2 then again returns a task-based organizational model and requires a log including performer and task of events as input. The performer-task matrix is processed as follows (Alves de Medeiros et al., 2008): All performers which have conducted the same tasks form one organizational unit. Then, a hierarchy is formed in which all performers of a higher organizational unit can execute a subset of the tasks, which a lower level organizational unit can execute. In this setting, the algorithm only regards if or not a task has been executed by a performer but neglects frequencies of execution. Alternatively, minimal execution numbers can be set as thresholds. Notably, similar approaches exist in the RBAC context such as that of BAUMGRASS (2011; 2012). However, they are not classified as role mining as they do not operate on permissions but create an organizational model consisting of roles based on executed tasks.

Table 2. Performer-Task Matrix 1 (left) and Performer-Task Matrix 2 (right)

Approach 3 ("Metrics based on joint activities"), SONG AND VAN DER AALST (2008) calculate the similarities of the performers employing the Minkowski distance, Hamming distance or Pearson’s correlation coefficient based on the performer-task matrix. A low distance between performers indicates a high similarity. Afterwards, a graph can be constructed with performers as nodes and similarities between the performers as weights of the edges. A threshold for minimal similarity is defined to remove edges between unsimilar performers. Each remaining connected sub graph forms one organi-
zational unit. Let \( T \) be the set of tasks, \( t \in T \) a task and \( t_{pi} \) the number of times performer \( p_i \) executed \( t \). Then the Minkowski distance between \( p_i \) and \( p_j \) is defined as (van der Aalst et al., 2005 p.568):

\[
d_{M1}(p_i, p_j) = \left( \sum_{t \in T} |t_{pi} - t_{pj}|^\theta \right)^{\frac{1}{\theta}}
\]

If \( \theta = 2 \), it is the Euclidean distance. A disadvantage of the Minkowski distance is that it hardly recognizes similarities if two performers have a different volume of work such as full-time and part-time workers (van der Aalst et al., 2005). In addition to the performers in Table 2, assume another performer who has executed task 3 15 times. While this performer and Daniel seem to be similar, their distance is 12. A solution is to use a logarithmic scale by applying the \( \log_k(x + 1) \)-function to the performer-task matrix. Alternatively, the Hamming distance can be used which disregards absolute frequencies and is more robust. The Hamming distance between two performers is calculated in the following way with \( |T| \) denoting the quantity of tasks (van der Aalst et al., 2005 p.568):

\[
d_{HA}(p_i, p_j) = \sum \delta\left(t_{pi}, t_{pj}\right)
\]

where \( \delta(x, y) = \begin{cases} 0 & \text{if } (x > 0 \land y > 0) \lor (x = y = 0) \\ 1 & \text{otherwise} \end{cases} \)

The Hamming distance returns the relative number of tasks only one of the two performers has executed while the other has not, i.e. a logical XOR operation. However, different ratios of executed tasks are not taken into account (Jin et al., 2007). If the performer-task matrix 2 in Table 2 (right) is assumed, the Hamming distance between the two performers is 0 although they have different profiles.

Pearson’s correlation coefficient divides the covariance by the product of both standard deviations. Thus, it evaluates a linear relationship and its values range from \(-1\) to \(+1\) where \(-1\) indicates a small similarity between the performers and \(+1\) a high similarity (van der Aalst et al., 2005 p.568):

\[
\rho(p_i, p_j) = \frac{\sum_{t \in T} (t_{pi} - \frac{\sum_{t \in T} t_{pi}}{|T|})(t_{pj} - \frac{\sum_{t \in T} t_{pj}}{|T|})}{\sqrt{\sum_{t \in T} (t_{pi} - \frac{\sum_{t \in T} t_{pi}}{|T|})^2 \sum_{t \in T} (t_{pj} - \frac{\sum_{t \in T} t_{pj}}{|T|})^2}}
\]

Approach 4 (“Organization mining based on the similarity of the performer”) is very similar to approach 3 because it also suggests three similarity measures and builds an organizational structure of performers who executed similar tasks based on a graph and similarity threshold. The Minkowski distance, spatial similarity and Pearson’s correlation coefficient are used here. In opposite to the other metrics, the spatial similarity takes weights of tasks into account in order to emphasize core tasks. With \( \omega_t \) as weight of task \( t \), it is defined as follows (Ni et al., 2011 p.223):

\[
sim_{spa}(p_i, p_j) = \sum_{t \in T} \frac{1}{1 + \omega_t |t_{pi} - t_{pj}|}
\]

A drawback of this metric is that the weights are derived from the properties of tasks so that these properties have to be known in advance (Jin et al., 2007).

Other approaches follow iterative ways to group performers as organizational units. One option is to use standard clustering procedures such as agglomerative hierarchical clustering as suggested in approach 5 (“Hierarchical organizational mining”) (Song and van der Aalst, 2008). At the beginning each performer forms a single cluster. Then the algorithm finds the two nearest clusters and merges them to a single cluster. This is repeated until all performers are in a single cluster or a given number \( k \) of clusters is reached. The profiles of the performer-task matrix and the metrics of approach 3 are used to determine the distance between clusters. Results of hierarchical clustering can be visualized in a dendrogram which allows deriving flat organizational models (Song and van der Aalst, 2008). If a flat model is created with a cut-off value of 0.3 for the example in Figure 2, the organizational units are \{John\}, \{Clare, Sue\}, and \{Jane, Mona, Pete, Robert, Fred, Mike\}. However, the benefit of hierarchical clustering is that it is able to create hierarchical models. In this case, one can start with merging Pete, Robert, Fred and Mike to one organizational unit named \( ou1 \). Then one can construct \( ou2 \) includ-
ing own and Jane and Mona, and so forth. Hence, one organizational unit may comprise performers and other organizational units. However, hierarchical relations (superior units are supervisors of inferior units) are not depicted. In addition to hierarchical clustering, VAN DER AALST (2011 p.226) suggests to employ k-means clustering.

Besides generic clustering procedures, there are those dedicated to organizational model mining. JIN ET AL. (2007) develop a two-step method (approach 6) based on the idea of the k-center problem. At the beginning, for each performer there is one potential unit and the performer becomes the center of the unit. All performers (including the center), who are more similar to the center than a given threshold, are associated to this potential unit. In the second step, intersections between the association sets of the potential units are eliminated so that each performer is finally assigned to exactly one unit. For this purpose, the potential unit with the highest average similarity between its members is taken as actual unit and its members are deleted from all remaining potential units. This is done iteratively until each performer is assigned to one organizational unit. Due to the drawbacks of the Minkowski and Hamming distances mentioned above, JIN ET AL. choose the cosine metric as similarity measure. Geometrically, $sim_{cos}$ is the cosine of the angle of two vectors and returns values ranging from 0 to 1. A high similarity is indicated by a value of 1. According to JIN ET AL. (2007) it outstands Minkowski and Hamming distances since it takes into account both times and kinds of executed tasks. It is defined as follows (Jin et al., 2007 p.672):

$$sim_{cos}(p_i, p_j) = \frac{\sum_{t \in T} (t_{pi} \cdot t_{pj})}{\sqrt{\sum_{t \in T} t_{pi}^2} \cdot \sqrt{\sum_{t \in T} t_{pj}^2}}$$

Approach 7 (“Organization mining based on grid clustering”) by NI ET AL. (2011) is a dedicated algorithm based on grid clustering. A grid space is constructed with one dimension for every task. Each performer is mapped into this space in a way that the coordinate for each dimension is the number of executions of the according task by the performer as to the performer-task matrix. Each dimension is divided into about $\sqrt{|P|}$ parts so that a grid consisting of $\sqrt{|P|}|T|$ rectangular units results where $|P|$ denotes the number of performers. The density of a rectangular unit is the relative number of performers mapped to points in the unit in relation to the number of all performers. The algorithm starts with constructing the first cluster as the rectangular unit with the highest density. Then it successively adds rectangular units to this cluster if the geometric distance and density difference between a rectangular unit and the elements in the cluster are sufficiently low. The other clusters are constructed accordingly until all performers are in a cluster.

Besides the first, approach 8 (“Mining activity set taken by each role”) is the only approach from this class which can assign a performer to more than one organizational unit directly. It requires the case and time in addition to the performer and task (Zhao et al., 2012; 2009). The additional information is used to explore sequence dependencies between tasks. A task $g$ directly depends on another task $f$ if there exists at least one case where $g$ directly follows $f$ and there are no cases where $f$ follows $g$ (Zhao et al., 2009 p.302). The algorithm forms groups of succeeding tasks each of which is executed by a role. All performers who executed at least one task of a task group are assigned to the task group and the role. The algorithm merges succeeding tasks until their degree of diversity is higher than a given threshold. The next following task is the beginning of a next group. The degree of diversity is calcu-
lated as a kind of dispersion measure for the relative number of times performers have conducted tasks of the task group. Only performers who executed at least one task of the task group are considered. Hence, the degree of diversity is low if each task of a task group has been executed by each relevant performer almost the same relative number of times which is calculated in comparison to the number of times a performer conducted any task.

Similarly to approach 15, ZHAO ET AL. (2009) create an approach to construct a social network of organizational entities. It connects two organizational units if there are two succeeding events regarding one case: The first event is executed by a performer of the first unit and the second event is executed by a performer of the second unit. In 2012, they pick up this idea to develop an interaction degree of diversity between units (Zhao et al., 2012). It measures the dispersion of the number of communications between members of a specific unit to all other units, i.e. their members. This degree is high if some members communicate to other units intensively and others do not. The authors develop a genetic algorithm whose fitness function combines the diversity degree of execution times of approach 8 and the diversity degree of interaction. This algorithm is approach 9 ("Role Identification Optimization Based on Genetic Algorithm"). It requires the same input as approach 8 and returns task-based teams but also considers communication structures. It is classified as an approach for social network of organizational entities mining and task-based organizational model mining since it combines elements of both areas.

In contrast to the previous, approach 10 ("Metrics based on joint cases") returns a case-based organizational model (van der Aalst et al., 2005; Song and van der Aalst, 2008). It requires a log containing at least information about the performer and case. A graph is constructed with the performers as nodes. The weight of an edge between two performers expresses how often the performers worked at the same case. Assume a performer Kai who worked at the cases A, B, C and a performer Florian who dealt with cases C, D, E, F. The edge from Kai to Florian has the weight 1/3 because Kai works at three cases and in one of them Florian participates. The weight of the edge from Florian to Kai is 1/4. The first option to obtain a case-based organizational model is to use a threshold as in the approaches 3 and 4. The second option is to disconnect performers with a high centrality from all other performers and then construct each organizational unit as a remaining connected sub graph. The second option may be necessary if there is a performer who works as a connector between several teams so that no meaningful connected sub graphs exist without removing his edges (Song and van der Aalst, 2008).

3.3 Approaches for Social Network Mining

The goal of social network mining approaches is to investigate the flow of work and information through an organization to reveal insights on the underlying structure. The approaches 11 to 14 are used to derive and analyze a social network of originators (cf. Appendix A). Approach 11 ("Metrics based on (possible) causality") requires a log recording the case and performer of an event (van der Aalst et al., 2005; Song and van der Aalst, 2008) where events need to be temporarily ordered. A process model and the tasks of the events can be provided for a causality analysis. Approach 11 introduces two kinds of metrics for (potential) causality which can be used to construct a social network. The metrics determine the weights of the relations between two nodes. Relations with weights higher than a defined threshold become edges in the social network. The first kind of metric considers the handover of work. A handover of work occurs between two performers if there are two subsequently executed tasks at the same case, one conducted by the first performer and the other one executed by the second performer (van der Aalst et al., 2005). The second kind of metrics is subcontracting. Subcontracting occurs within one case if there is a task conducted by the second performer between two tasks which are completed by the first performer. For each kind of metric, there are three binary configuration parameters so that in total $2^3 = 8$ metrics of each kind are introduced by VAN DER AALST ET AL. (2005). Roughly speaking, these metrics count the relative number of times, handover of work or subcontracting take place between two performers depending on the selection of the parameters. The first
parameter determines whether direct or indirect succession is considered. Regarding handover of work, indirect succession means that there may be other tasks in-between the two relevant tasks. Regarding subcontracting, indirect succession means that there may be several tasks in-between two tasks conducted by one performer. The second parameter defines whether multiple transfers within one case should be considered. If multiple transfers are excluded, it is only relevant if there exists a handover in a case, but not if there are several. The last parameter determines whether every handover should be considered or only those whose causality is proven by a process model, which then also requires providing the model. A process model is beneficial to for instance exclude concurrent tasks from being counted for the handover of work metrics (van der Aalst, 2011).

Approach 12 (“Metrics based on special event types”) is an extension and specialization of approach 11 which requires the event type and the task as additional log information. The event type is used to provide metrics for constructing a social network with hierarchical relations (van der Aalst et al., 2005 pp.569–570). Thus, VAN DER AALST ET AL. only consider events of a specific type when constructing the social network. They define two metrics for the event type reassign which can also be applied to any other event type. A reassignment of a task between two performers is denoted by two successive events regarding the same case and task. The first event is conducted by the first performer and has the event type reassign. The subsequent event is executed by the second performer and has an arbitrary event type. Similar to the metrics in approach 11, both reassignment metrics count the relative number of times a reassignment took place from one performer to another one. One metric considers multiple transfers, the other one ignores them. In order to analyze social networks constructed with one of the approaches mentioned above, different ratios such as centrality notions can be calculated. Centrality notions can be used to for instance identify possible bottlenecks (Song and van der Aalst, 2008). VAN DER AALST ET AL. (2005) provide an overview of some applicable figures and indicators to analyze these social networks and refer to relevant literature.

While the approaches 11 and 12 follow the rationale that relations between two performers such as handovers and reassignments are indicated by a certain sequence of events, the approaches 13 and 14 use logs which record the event type and the two relevant performers as sender and receiver in one single event. In approach 13 (“Hierarchy mining algorithm”), HANACHI ET AL. (2012) aim to extract hierarchies from a social network by defining a hierarchical structure as a directed tree so that each performer has one direct supervisor at most. The social network is constructed as graph with labelled edges such that there is an edge from performer A to performer B for each event type which serves as event type of at least one event of the log with performer A as sender and performer B as receiver. Approach 13 only considers edges labelled with the event type delegate since it extracts hierarchies. After constructing the social network, it is checked whether the social network is a directed tree, i.e. a hierarchy. If not, it is checked for every node without a superior, whether there is a directed sub tree with this node as root. If there are hierarchical structures, these are outputted finally. Approach 14 (“Federation mining algorithm”) is related to approach 13 since it checks whether a social network has a federative structure (Hanachi et al., 2012). A social network has to fulfill several characteristics to be a federation. For example, it has to be connected and must not be a hierarchy. The algorithm proposed by HANACHI ET AL. (2012) checks whether a social network fulfills the requirements and returns an according boolean value.

The social network mining approaches presented so far create and analyze social networks of originators. Complementarily, approach 15 focuses on social networks of organizational entities, and approach 16 uses elements of both kinds of social network mining. Approach 15 (“Information flows between organizational entities”) consists of rules defining how to aggregate performers to organizational entities in a social network (Song and van der Aalst, 2008). For this purpose, it requires an organizational model and a social network of originators. Although it does not operate on a log directly, the input elements can be derived by using approaches described in this paper. The approach proceeds as follows (Song and van der Aalst, 2008 p.309): The organizational units specified in the organizational model become the nodes of the social network. There is an edge between two of these nodes if
there is at least one edge between a performer of the first organizational unit and a performer of the second organizational unit in the provided social network. The weight of such an edge is the absolute or relative sum of the weights of all edges between a performer of the first organizational unit and a performer of the second organizational unit.

In contrast to the approaches from above, approach 16 is restricted to the domain of knowledge maintenance (Li et al., 2011) which deals with the management of a knowledge repository. It requires a log with information regarding the case and performer of events which are temporarily ordered. The goal is to derive a social network which reflects the knowledge maintenance organization where people are grouped according to their knowledge level and area. A performer A belongs to a superior level in comparison to performer B if there is a handover from B to A. A superior unit covers all knowledge areas of its inferior units. Hence, performers who handover work to the same superiors deal with the same knowledge area which is indicated by the superior unit. The first step of the approach constructs a social network of originators taking the longest handover chain between two performers as the path between them in the network in order to derive all hierarchical relations. In the second step, a social network of organizational entities is derived by grouping performers together who reside at the same knowledge level and area. They have to handover work to the same people (area) and have to receive handovers from the same people (level), i.e. their successors and predecessors in the network have to be the same. Consequently, this approach constructs a specific network of the information flow in the knowledge maintenance process and reveals information of the underlying knowledge organization.

3.4 Approaches for Role Mining

This subsection now discusses event-log-based role mining approaches (see again Appendix A) which both can assign a performer to several roles. Approach 17 operates on a log of performer, permission and exact timestamp of an event. The basic idea is that permissions, which occur temporarily together very often, form one role (Jafari et al., 2009). Hence, if the distance of the timestamps of several events is low, then these permissions belong to one role. The algorithm (Jafari et al., 2009 p.260) starts with calculating all possible combinations of all recorded permissions, i.e. the power set of all permissions. Each combination of permissions becomes a potential role which is assigned to all the performers of these permissions according to the log. A score is assigned to every potential role which is computed based on the number of times the permissions of the potential role occur together for each performer and their temporal distance in the occurrence. The temporal distance is indicated by the timestamps of events. Afterwards, the potential roles are sorted according to their score. Starting with the potential role with the highest score, the potential roles become actual roles. This is done until each permission is at least covered by one role and each performer is assigned to at least one role. Since the complexity of this algorithm is exponential, JAFARI ET AL. (2009) improve it by considering coinciding permissions during the creation of potential roles.

Approach 18 is dedicated to web applications and uses a log which provides the session, performer and permission of an event. It assumes that the permissions are grouped to logical sessions before inputted to the algorithm (Gal-Oz et al., 2011). A logical session is a single logical operation such as a course registration within a login session of a performer. To perform this operation, several database objects have to be accessed and the according permissions are needed. The algorithm proceeds in three steps (Gal-Oz et al., 2011 pp.128–134). First, a hierarchy of sessions is created so that the set of permissions of a higher session is a proper subset of permissions of a lower session. The performers who had this session and their quantity of times they had this session (named support in the following) are assigned to each session node which represents a potential role. The number of roles is decreased in the subsequent steps. Second, the support of a potential role at a higher level is moved to an inferior potential role if the according performer is assigned to both potential roles. This can lead to roles with a total support of 0 so that they can be deleted. Third, the performers of a potential role are moved to one of its inferior potential roles if its number of performers or support are less than predefined
thresholds. This step can be useful since not every performer may have used all his necessary permissions and this step extends his set of permissions. Further, it can avoid an excessive amount of roles.

4 Challenges of Mining Organizational Properties in CPS

4.1 General Challenges

One main characteristic of a CPS is its high degree of distribution in space. Machines, GPS satellites or ERP systems are just some examples which may need to be integrated when establishing a CPS. In order to make organizational mining happen, data from all these elements needs to be brought together. In 2004, VAN DER AALST AND WEIJTERS (2004) claim that collecting all relevant information needed in an enterprise for starting with mining is far from being trivial since the information is scattered over many components and applications. Due to the integration of different kinds of systems and components from the physical and cyber space and the higher degree of distribution, it will even be more challenging in a CPS to obtain all that relevant information required for a mining approach.

Another general problem are different levels of granularity. VAN DER AALST (2011) criticizes that the level of abstraction of events provided by many applications differs from the level of granularity requested by end users being responsible for process or organizational management. In a CPS, the level of granularity of events does not only have to match between application and end user. Instead, the levels of abstraction between elements of the cyber space, the end user as well as physical components have to correspond. Whereas cyber components such as ERP systems may rely on events produced by human beings such as orders or invoices, events in the physical space are recorded by sensors, implying higher frequency and the record of events that might not be of interest for any human being. Techniques for complex event processing (CEP) may help overcoming this issue. CEP aims to combine atomic events to complex events in order to derive more valuable information (Eckert and Bry, 2009).

A more complex challenge which cannot be solved with additional logic is the temporal ordering of events and consistent recording of timestamps. Several techniques exist to achieve synchronized clocks between different parts of a CPS (Kopetz, 2012; Broman et al., 2013). In a highly distributed system, a global navigation satellite system such as GPS is an exemplary global time source that however suffers from vulnerabilities against several attacks and failures such as a lack of sufficient satellite signals for sensors in buildings (Broman et al., 2013). Network protocols such as IEEE 1588 may also be used to synchronize clocks. While time transfer within networks is accurate enough for most but not all application cases, issues remain with increasing accuracy resulting from asymmetry that results from the difference in transferring times of synchronizing messages from the source to the receiver and the other way around (Broman et al., 2013). Regarding integrating the physical and cyber space, KOPETZ (2012) holds the view that it is difficult to map a model of physical events to the cyber space, mainly due to the different time granularities between both spaces. Concluding from these issues, temporal problems in a CPS cannot be solved with additional logic. Instead, several technical challenges remain that make a correct time recording in all CPS difficult.

A related problem is the identification of cases, i.e. to determine which events form one process instance. The literature uses the term event correlation to describe this problem which means that events have to be related to each other (van der Aalst, 2011). If event data is distributed over several systems, it is difficult to establish event relations. VAN DER AALST (2011) argues that, however, it is quite easy to tackle this obstacle if logging functionality is designed from scratch. Unfortunately, this is hardly possible in a CPS due to its high variety of elements and the incorporation of existing systems such as ERP systems. Thus, the event log has to be enriched with information that enables an identification of a business object for every event, e.g. a unique identifier of an object, the location or time interval an event took place in. However, it has to be decided for each scenario individually whether a sufficient set of attributes can be obtained. Hence, the obstacles of event correlation can be solved with addi-
tional logic if information is supplied to distinguish different business objects. Consequently, the event correlation problem may be solved easier than the technical temporal obstacles.

Additional topics which should be focused include privacy and noise (van der Aalst et al., 2005). Mining approaches analyze data that may refer to individual human beings. The approaches allow assessing performance indicators such as the amount of work handed over. Hence, privacy should always be kept in mind. In addition to temporal contaminations, noise in general is a problem which may distort the results of mining approaches. In a CPS, noise can result from for instance corrupted machines which do not record the true attribute values of an event so that wrong values are logged, and it could be addressed with statistical approaches (Breuker and Matzner, 2013).

### 4.2 Individual Approach Evaluation

The discussed approaches differ in their data requirements. Table 3 recaps these characteristics. The approaches are framed according to their different disciplines. An ‘x’ indicates that an approach requires the respective information. If the ‘x’ is written in brackets, it is optional. A ‘2’ indicates that the approach requires two objects of this element to work.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>(x)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performer</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>2</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Type</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timestamp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permission</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Data Requirements of the Approaches

Many authors such as SONG AND VAN DER AALST (2008 p.306) and JIN ET AL. (2007 p.671) define an event as combination of a task and a performer. Hence, it can be concluded that the most rudimentary process log provides this information and that it should be available in every event-based CPS. An event type is recorded in an event log depending on the log’s comprehensiveness. There are no major obstacles which prevent elements in a CPS to log types. However, logic has to be implemented which recognizes which of the available event types has to be assigned. Permissions and sessions are very dedicated role mining artifacts and occur in the cyber part of a CPS. In physical parts, they will not occur because there do not exist permissions needed to get access to data objects. According to the explanations above, it is difficult to register time precisely in every CPS. These problems cannot be solved with additional logic or a more comprehensive log but require fundamental technical solutions. However, the case can be derived if additional logic is implemented. Again it depends on the comprehensiveness of the log whether two performers as sender and recipient of an event can be captured.

Most of the task-based organizational model mining approaches require only the task and the performer of an event. Hence, the approaches 1 to 7 are easily applicable in a CPS (Table 4). Since the approaches 8 and 9 require a temporary order of the events, it is difficult to apply them in every CPS. Approach 10 requires the case and performer so it can be implemented with additional logic. The approaches 11 and 12 are not suitable for every CPS at the moment as they rely on a temporal order of events. By contrast, the approaches 13 and 14 may be applicable because the second performer may be provided depending on the comprehensiveness of the log. The suitability of approach 15 depends on whether its inputs can be delivered. An organizational model can be provided easily due to the applicability of the approaches 1 to 7. Hence, the applicability of approach 15 depends on whether a social network of originators is available and this tends to be rather difficult. Approach 16 is not applicable since it relys on a temporal order of events and due to its domain. The suitability of the
The applicability of the approaches with regard to data requirements is divided into three categories: *Easy*, *Easy/Difficult* and *Difficult*. *Easy* approaches only require the task and performer. Approaches in *Easy/Difficult* need additional logic or more comprehensive logs. *Difficult* approaches require role mining artifacts, have a too restrictive domain or need technical innovations to be applicable in all CPS.

<table>
<thead>
<tr>
<th></th>
<th>Task-Based Org. Model Mining</th>
<th>Case-Based Org. Model Mining</th>
<th>Social Network of Originators Mining</th>
<th>Social Network of Org. Entities Mining</th>
<th>Role Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy/Difficult</td>
<td>10</td>
<td></td>
<td>13, 14</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Difficult</td>
<td>8, 9</td>
<td></td>
<td>11, 12</td>
<td>16</td>
<td>17, 18</td>
</tr>
</tbody>
</table>

Table 4.  Applicability of the Approaches

### 4.3 Usage Scenarios

Besides general information about roles, the organizational structure and the social network of a CPS, mining approaches can be used to get other valuable insights. We assume large-scale autonomous manufacturing including preventive maintenance services as a prototypical CPS service system scenario (Ray, 2013) based on event data analysis (Klein et al., 2013). A prerequisite is that the according approaches are applicable. A social network can be employed to identify possible single point of failures and bottlenecks which are nodes with a high centrality (Song and van der Aalst, 2008). Moreover, a node whose number of ingoing and outgoing edges decreases rapidly during one time period may indicate that the machine represented by the node is corrupted. By contrast, if the number of ingoing and outgoing edges exceeds a certain threshold, a machine may be overloaded and may get corrupted soon. Obviously, a low or high number of ingoing and outgoing edges does not always mean a damage or overloaded, but could be due to changes in the way of working to a more or less cooperating style.

In case of an overload or failure of a machine, a task-based organizational model can be used to identify machines which can support or substitute the affected machine. Using additional spatial information in this case can increase the accuracy by only considering machines which are located near to the affected machine. Furthermore, specialists can be identified which can be consulted to answer a question regarding a specific task (Alves de Medeiros et al., 2008). In addition, generalists can be identified. They can act as mediators if for instance several units or elements of a CPS work together.

### 5 Conclusion and Outlook

This paper aims to answer the following research questions or issues: The presentation and classification of mining approaches which can derive organizational properties, the derivation of their data requirements and analysis of possible problems for mining approaches in a CPS scenario. For this purpose, Section 2 sets the scope of this paper to organizational model mining, social network mining, and role mining. Section 3 contains the results of the literature review, the classification of the approaches and their data requirements. In total, 18 approaches have been discovered. Section 4 deals with the applicability of mining approaches in a CPS context. The applicability can be evaluated based on the data requirements of the approaches. Deriving a correct temporal order of events is hardly possible in all CPS due to technical limitations. The case of events may be recognized using additional logic. On the contrary, the task and performer of an event are obtained easily. Hence, especially most of the approaches aiming at task-based organizational model mining are most suitable for CPS.

However, all approaches can be improved since they do not use spatial information. According to TAN ET AL. (2009) spatial information is essential in a CPS and should be part of an event log. Hence, mining approaches may be extended so that they use location data in order to e.g. create an organizational model regarding spatial properties. Although role mining approaches cannot be applied in their intended ways, their ideas can be transferred to a CPS in future work. The simplest option is to substitute...
permission by task so the approaches operate on tasks. Whether this is semantically useful, has to be validated for each approach individually. Our study is not without limitations, which suggest opportunities for future research. Most notably, this work is restricted to the analysis of the approaches’ data requirements. This work can be extended by a case study which investigates the approaches empirically e.g. with regard to performance and accuracy, thus leading to a multi-dimensional quality assessment (de Weerdt et al., 2012). Approaches relying on a temporal order of events are hardly applicable in all CPS now. But CPS is an emerging research field so that many areas have a lot of potential for improvements. Consequently, approaches which require an event’s time will probably be well applicable in all CPS one day.

Appendix A: Classification of the Approaches

<table>
<thead>
<tr>
<th>No.</th>
<th>Source</th>
<th>Required Log Info.</th>
<th>Data Input</th>
<th>Output</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Song, v.d. Aalst, 2008</td>
<td>Task, performer</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Alves de Medeiros et al., 2008</td>
<td>Task, performer</td>
<td>Log</td>
<td>Hierarchical task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Song, v.d. Aalst, 2008; v.d. Aalst et al., 2005</td>
<td>Task, performer</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Ni et al., 2011</td>
<td>Task, performer</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Song, v.d. Aalst, 2008</td>
<td>Task, performer</td>
<td>Log</td>
<td>Hierarchical/flat task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Jin et al., 2007</td>
<td>Task, performer</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Ni et al., 2011</td>
<td>Task, performer</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Zhao et al., 2012; 2009</td>
<td>Case, task, performer, time</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>Zhao et al., 2012</td>
<td>Case, task, performer, time</td>
<td>Log</td>
<td>Task-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>Song, v.d. Aalst, 2008; v.d. Aalst et al., 2005</td>
<td>Case, performer</td>
<td>Log</td>
<td>Case-based org. model</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>Song, v.d. Aalst, 2008; v.d. Aalst et al., 2005</td>
<td>Case, performer, time (task for causality)</td>
<td>Log (proc. mod. for causality)</td>
<td>SN of originators</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>v.d. Aalst et al., 2005</td>
<td>Case, task, performer, time, event type</td>
<td>Log</td>
<td>SN of originators with hierarchical relations</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>Hanachi et al., 2012</td>
<td>Performer 1/2, event type</td>
<td>Log</td>
<td>Hierarchical structures in the SN</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>Hanachi et al., 2012</td>
<td>Performer 1/2, event type</td>
<td>Log</td>
<td>Boolean indicating if SN is federative or not</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>Song, v.d. Aalst, 2008</td>
<td>-</td>
<td>Log</td>
<td>SN of originators, org. mod.</td>
<td>SN of org. entities</td>
</tr>
</tbody>
</table>

Legend: RMining = Role Mining, SN = Social Network
References


