PROCESS MINING AND DATA AGGREGATION STRATEGIES FOR BUSINESS COLLABORATION NETWORKS

Kerstin Gerke[1], Gerrit Tamm[2]
[1]SAP Research, CEC Dresden, Germany
[2]Institute of Information Systems, Humboldt University, Germany

Abstract
Business collaboration and information sharing becomes more and more important for companies, mainly driven by economics of scale and the strategic value of network knowledge. In this paper we will describe how process mining results will increase the willingness of cross company data aggregation and the usage of shared business collaboration infrastructures. The results of our contributions are based on process and data analyses in the customer relation management of a leading German passenger airline. As information sharing is considered a central challenge among supply chain partners, we describe the role of data aggregation as a requirement for process mining in collaborative infrastructures. Aggregating data on the one hand can preserve companies from compromising their local data. On the other hand valuable network information may get lost. We use process mining to derive collaborative business process models. Therefore, we evaluate the impact of data aggregation on the significance of process mining results.

1. Introduction

Today, many companies seek for collaborative contributions from supply chain partners to raise customer satisfaction and to keep costs low. Although collaboration promises significant potentials, companies will be more dependent on the flow of data outside their own boundaries due to the disrupted flow of data.

In the traditional supply chain there is a sequential data and material flow among companies. Data is typically captured and stored at multiple places. Individual planning cycles are often unsynchronized and based on assumptions of the customers' expected behavior. Resulting problems are commonly described as the Bullwhip Effect [4]. The sequential supply chain is designed to manage internal operations. However, today's companies face a strong economic pressure to increase competitiveness, and to operate on a market becoming more dynamic and progressively more global. They are continuously challenged to engage in alliances and business networks to increase the velocity and transparency of their supply chain. Consequently, the traditional supply chain is moving towards complex multi company networks. A business collaboration network constitutes a community of customer focused companies being aligned by collective objectives to intelligently adjust to changing economic environments. To efficiently coordinate and control cross organizational operation, companies need well designed, integrated and cross boarder managed business processes [1]. However, an actual description and formalization of the business processes is missing in many companies. This may be explained with the fact that the manual process

modeling is a time-consuming, fault-prone and particularly expensive activity. Therefore, most business process models are seldom updated and do not correspond to the actual execution of business processes. Although the execution is typically recorded by modern information systems (IS) the valuable information is rarely used. Process mining technology is able to automatically derive business process models from process data. To extract cross organizational business processes with process mining methods raises the challenge that companies need to share their data. However, although business collaboration network partners realize that combining their data has some mutual benefit; few are willing to reveal their raw data. They fear to compromise their internal data and to become vulnerable among network partners. An integrated process analysis based on aggregated data instead of raw data may help to incorporate knowledge into business processes while protecting the autonomy of companies.

With our approach we first extract a business process model from raw data with process mining techniques. Second, we classify business processes in homogeneous groups according to the underlying process point of view. For this purpose, we define various variables so that the business process complexity can be expressed quantitatively. The variables will be used further in clustering. Third, we derive a business model from the aggregated database. To discuss these issues in detail the remainder of this paper is organized as follows: The following section introduces the claim settlement process of a passenger airline and the sensitivity of the enclosed data. Section 3 and 4 present established approaches for data aggregation and process mining technologies. Our concept to develop aggregated models is presented in section 5. In section 6 we discuss the results of data aggregation and process mining in business collaborative networks, and we evaluate whether data aggregation can be leveraged to support cross company modeling. Finally, in section 7 we formulate conclusions based on our current findings.

2. Customer Relationship Management Process of a Passenger Airline

Deep knowledge about business processes is seen as the key prerequisite for modeling them efficiently. We look at a claim settlement process of a German passenger airline. Interactions among business partners are handled with the application “Interaction Center” (IAC) of a CRM system. The IAC processes claims involving multiple processing steps or multiple processors in- and outside the passenger airline. Therefore it supports the processing and communication flow of cross organizational business. Interactions among business partners are activities that are classified into contacts and tasks that need to be accomplished within a defined period of time. Activities handle services, business transactions, and process incoming and outgoing e-mails, faxes and phone-calls. Activities succeeding one another are created as subsequent contacts or tasks. Processing starts with a customer’s inquiry or claim. If the customer has no record, the complainant is created as a new business partner. In case the claim can be assigned to a pre-defined claim settlement procedure an activity “Communication Operation” is created. The activity triggers a written reply to the business partner and settles the case. If the claim needs further consideration, the activity “Customer Relations” is typically created. It handles all interactions among customers. The first activity “Customer Relations” serves as a starting point. Depending on the circumstances of the case the claim is passed on the responsible group which can be located within the passenger airline or in an affiliate. A claim for a lost suitcase, for instance, will be transferred to the group “Lost & Found”. The transfer takes place by generating a subsequent activity “Lost & Found”. The activity “Customer Payments” clarifies claims being associated to payments. The status indicates how much of the activity has been completed. Activities are stamped with the name of the employee who created the activity and the date and time of creation. The resolution of customer issues implies sensitive information that can affect the business reputation of the passenger airline

in the public. The circumstances which have been considered for the settlement or non settlement, information about payments, as well as the rejection rates of refunds may reveal whether a company is responsive to customers' needs.

3. Process Mining for Business Collaborative Infrastructures

Process-aware information systems like Customer Relationship Management (CRM) systems and Enterprise Resource Planning (ERP) systems record activities during the execution of business processes and store them in event logs. These logs are the starting point for process mining, and incorporate knowledge about what is actually happening in an enterprise.

The process mining technology can be seen as a special data mining method. The goal of data mining is to discover knowledge out of huge data volume. Process mining aims at the automatic extraction of a specific kind of knowledge: The process knowledge [11]. This knowledge is extracted with the help of process mining algorithm providing a business process model [7].

Process mining analyses the process knowledge from different perspectives [12]. The control flow describes the sequence in which the different activities are executed and allows answering the following questions: How are the processes actually being executed? How is the distribution of all process instances over the different paths through the process? The organizational aspect takes care of the executor's behavior and supports the answering of the questions: What are the business rules in the process model? Are the rules indeed being obeyed? What is the communication structure and dependencies among people? The instance level deals with the questions: How compliant are the process instances with the deployed process models? What is the throughput time of instances? What is the most frequent or critical path for every process model?

Various algorithms have been developed and implemented in ProM to discover different types of models, e.g. Petri nets or Event Process Chains (EPC) describing the behavior observed in a log. ProM is the process mining workbench consisting of the open source products ProM und ProMImport [12]. It is developed by the technical university of Eindhoven. In addition to the process mining algorithm ProM offers a multiplicity of mining tools including export and conversion functionalities, and various analyses for monitoring and verifying process models. A lot of work has been done in the area of process mining, for instance [9, 10]. In [9] the authors introduce an approach for mining precise models by clustering a log. The partition of the log takes place on the basis of the frequencies of tasks. We focus on specific process characteristics revealing details about the process complexity.

One major benefit of process mining techniques is that information is objectively exploited according to an event log. Modeling becomes independent from what people believe that is happening in the company [3]. The capability to formally describe business processes and to weave them into a comprehensive context is even more important for modeling cross company business process models. Furthermore, the automation improves sustainably the cost and time factor of modeling. Resulting patterns, regularities or anomalies in the supply network provide important support to continually learn from exceptions and allow proactive action. The integration of process mining technologies in the business process control contributes to a higher degree of transparency of the network knowledge like customer demands, inventory, and capacities. Thus, typical conflicts between goals of minimizing inventory and an optimal utilization of the transport capacities e.g. can be solved collaboratively. Time and goods buffer can be diminished.

4. Process of Data Aggregation and Process Mining

Figure 1 shows how a business collaboration network is managed by using data aggregation strategies and process mining technologies.

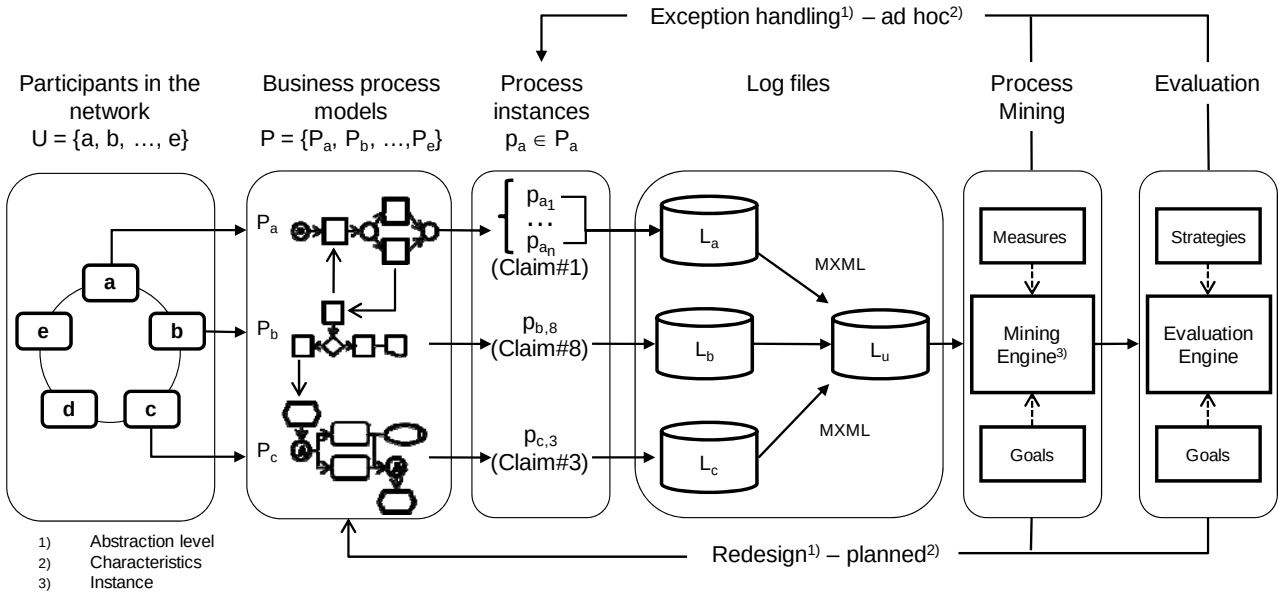

Figure 1: Manage a business collaboration network with process mining technologies

Participants in a network use information systems (IS) to execute collaborative business processes with respect to an a priori business process model. While initiating a process, a process instance is created that can consist of different activities like to receive a claim or an inquiry. The IS typically support logging capabilities that register what has been executed in the organization and store them into log files. These logs are the starting point for process mining. To meet the requirement for process mining in collaborative infrastructures the log files need to be formalized and consolidated. The Mining Extensible Markup Language (MXML) e.g. allows formalizing the log files. MXML is a generic XML-based format that the process mining framework ProM is able to read.

As environments in business collaborative networks are inherently distributed and decentralized, heterogeneous data from a great number of partners and sources has to be consolidated and aggregated. The data has to be reassessed by incomplete and obvious dissonant data. In the transforming process cleared data is enriched by business logic such as aggregated variables revealing the complexity of a business process. Goal of the data aggregation is to (i) maximize the significance of process mining result, and to (ii) preserve companies from compromising their sensitive data. Clustering techniques like K-Means are used to group data into homogeneous groups that are not known before. However, in the context of process mining we are looking for homogeneity of the underlying processes according to the sequence, timing and execution of activities. Distinguishing homogeneous groups appears to be important, because it allows cutting off those groups with sensitive data the company does not want to share.

The process mining engine extracts the process knowledge of the network in consideration of its measures, goals, and strategies. Depending of the abstraction level of the discovered deviations from the current business process model a redesign of the business process model or an exception handling at run time is necessary. The evaluation engine interprets the mining results and prepares

individualized recommendations for action. It is responsible for the exploitation and return of the mining results. Thus, the concerned participants respectively the applications are to be identified.

5. Implementation

In this section we give an architectural overview of the used components. We describe the data selection, the extraction of the process models, and the results of clustering.

5.1. Architectural Overview

For our method we combined and developed several tools. The (i) IAC (Interaction Center) is an application of the CRM (SAP CRM 5.0). We utilize data from the IAC for process mining and for data aggregation. During a test phase employees of the passenger airline tested the functionality of the IAC following a pre-defined test catalogue. Therefore, all process sequences necessary for the productive operation are supposed to be in the system. The (ii) CB (Case Builder) is an application we developed for the extraction of data and the building of cases. Afterwards, we load the extracted data to the (iii) SAP Business Intelligence system (SAP BI 7.10). We enrich the data with business logic and convert it into the MXML format. Furthermore, SAP BI provides the method for clustering. ProM (iiii) reads MXML files and allocates the algorithm for the actual process mining.

5.2. Data Selection and Preparation

The IAC does not support logging capabilities. Consequently, we analyze whether a file can be extracted from a transactional system that fits the requirements of process mining. First of all, we identify the data belonging to one case. We consider a case as a collection of activities, partners, and dependencies between activities. Activities correspond to individual steps in the claim settlement process of a business partner. Business partners are responsible for the enactment of activities. Dependencies determine the execution sequence of the activities. To retrieve the cases we extract single activities, determine the dependencies from data flow among them, and generate corresponding events. To assure that each event can be properly associated to a case the generation of events adheres to the following assumptions [8]. Each event refers to an activity and a case. Events are totally ordered. In the log file events are sequentially recorded, even though activities may be executed in parallel. An activity performs a different category, e.g. “Lost & Found” or “Customer Payments”. It is important to be able to distinguish what category an activity belongs to. We only consider claim settlement cases that are completed being expressed by the status “Complete” and “Rejected”. The selection of the data results in a database with 590 claims. Once the cases are retrieved, the transformation process unifies parameter values being used synonymously and handles missing values. Finally, we formalize the events into the MXML format.

5.3. Creating Process Models with Process Mining

We use the “Genetic Algorithm plugin” available in ProM to extract the business model. The “Genetic Algorithm plugin” uses genetic algorithms. These algorithms start with an initial possible process model. A fitness measure indicating the quality of the model is assigned to the process model. Populations evolve by selecting the fittest process models and generating new models using genetic operators such as crossover and mutation. Every possible process model in the population is represented as a causal matrix showing the dependencies between tasks. Finally, a set of process models decreasingly ordered by the fitness value is provided [3]. We add an artificial start and end event named “Claim Settlement Process”, because there are various start and end events.

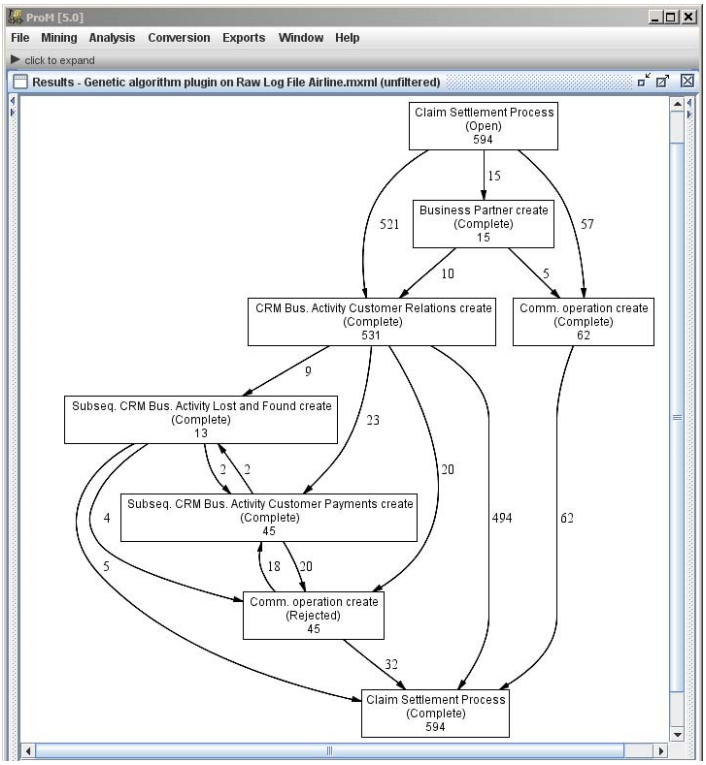

Figure 2: Mined claim settlement process

The dependency/frequency graph visualizing the claim settlement process of the passenger airline is shown in Figure 1. The graph is limited to an illustratable process model only considering the activities "Customer Relations", "Customer Payments", and "Communication operation". According to the employees of the passenger airline the complete model reflects the execution of the claim settlement which is described in section 2. Each node of the mined processes represents one activity. The first line of the node depicts the description of the activity, e.g. "Business Partner create". The second line reflects the status, e.g. "Complete". The third line shows the frequency of the activity in the log file. Nodes are connected with directed edges representing a mined dependency relation between two activities.

5.4. Data Aggregation Model, Enrichment and Quality

Since we know that it is possible to extract a relevant process model from raw data, we will search for complaint cases that are homogeneous related to the complexity of the claim processing. Process complexity is the degree to which a process is difficult to analyze, understand or explain characterized by the number and intricacy of activity, activity categories, interfaces, and other process characteristics [2]. In collaboration with the employees and business partners of the passenger airline we agree on five variables that may allow us to differentiate the claim processing concerning its complexity [5]. We calculated the average completion time of the considered cases as 3.5 days. Hence, we define "days" as the basic time unit. Every process consists of a sequence of activities. Let A be the activity $A = \{a_1, a_2, \dots, a_n\}, n \in \mathrm{N}$ and let S be the sequence of activities $S = \{a_1 \to a_2,\ a_2 \to a_3, \dots, a_{n-1} \to a_n\}, n \in \mathrm{N}$.

1. $A_{diff} = \#\{a_i : a_i;\ a_j \in A, a_i \neq a_j\}$ counts the total number of different activities being executed within the case. An activity typically refers to an organizational unit in or outside the passenger airline. We assume that the more different activities are involved, the more

complex the case is likely to be. A case containing the sequence of A-B-A-C-D implies e.g. $A_{diff} = 4$.

2. $A_{time} = \frac{A_i}{t}, A_i = \#\{a_1, a_2, \ldots, a_n, t\}, n \in N$ represents the number of activities within the case per time unit. We suppose that the more activities per time unit, the more complex the case is expected to be. If customer C_1 contacts the passenger airline four times within six days, whereas customer C_2 interacts only two times, the case of customer C_1 appears to be more complex. We calculate $A_{time(C1)} = \frac{4}{6} = 0.7$ and $A_{time(C2)} = \frac{2}{6} = 0.3$.

3. S_{diff} states the total number of shifts between activities counted by their total number within the case. We presume the more a case is being passed on, the more complex it is likely to be. Assume customer C_1 has a complaint claim involving the sequence of A-B-A. Consider now that customer C_2 has a claim expressed by the sequence of A-A-C-A-A. Obviously the case of C_1 is more complex although both are transferred two times. We have $S_{diff\ (C1)} = \frac{2}{3} = 0.7$ and $S_{diff\ (C2)} = \frac{2}{5} = 0.4$.

4. $S_{time} = \frac{S_{diff}}{t}$ records the number of shifts within the claim counted by the total number of activities per time unit. We assume the more shifts per time unit, the more complex the case is supposed to be. For example, consider that customer C1 has a case involving the sequence of A-A-C-A-D-A within four days, whereas the claim case of customer C_2 results in the sequence A-A-C-A-A-D-A-A within eight days. Case C_1 *is more complex than case C2* because *we have* $S_{time(C1)} = \frac{0.7}{4} = 0.25$, and $S_{time(C2)} = \frac{0.5}{4} = 0.06$.

5. S_{mean} calculates the mean of number of shifts counted by the total number of activities per existing activity category. Within the case, for each activity category S_{diff} is calculated, next the mean is computed. The higher the mean, the higher the complexity of the case is expected to be. If $S_{mean\ (C1)} = 0.3$ and $S_{mean\ (C2)} = 0.2$, case C_1 is more complex than C_2.

As we use the SAP Business Intelligence system (SAP NetWeaver BI 7.0) for the transformation process we choose the clustering method K-means that is available in the system. K-means is described in more detail in [6]. We defined the maximum number of clusters to be formed by the K-means method as three. We identify the clusters "Simple" (93%), "Moderate complex" (2%), and "Complex" (5%) from the enriched data. The time depending variables do not contribute to a meaningful clustering of the raw data. The most important reason for this is probably due to the fact, that a test person undertakes a complete case. Thus, wait and transfer times are omitted.

Table 1: Value distribution of the clustering output

Cluster	A_{diff}		S_{diff}		S_{mean}	
	Range	Number	Range	Number	Range	Number
Simple	0.8-2.0	552	0.0-0.7	552	0	552
Moderate Complex	2.8-3.2 3.6-5.0	11 1	0.4-0.9 0.9-2.3	10 2	0.2-0.3	12
Complex	1.6-2.0 2.8-5.0	6 23	0.4-0.9 0.9-1.8	24 5	0.4-0.7	29

That is why we concentrate on the residual three variables A_{diff}, S_{diff}, and S_{mean}. Table 1 shows the value distribution of the clustering output. In cluster "Moderate Complex" e.g. the number of different activities of 11 cases ranges between 2.8 and 3.2. The values can be interpreted as

described below: The first cluster is the largest one showing high correlation with the variables A_{diff} and S_{diff}, and very small correlation with the rest. Accordingly, cluster one represents cases with few different activities and none or few shifts between them. A hypothesis can be that cluster one contains only easy and simple claims. The second and third cluster both correlate with all variables. They represent cases with various different activities and shifts among them. The key differentiation between them is given by S_{mean}. The average shift per activity category of cluster three is more distinctive. Therefore, we define cluster two as "Moderate complex" and cluster three as "Complex". The differentiation of S_{mean} can be explained by the claim settlement process. The "Lost & Found" processing is associated with plenty of interactions. Typically, further inquiries are necessary asking for detailed information about the content of a suitcase, insurances, or sales slips. Passengers canceling their flights have a legal claim to refund taxes. In this case there will be no further inquiries. Thus, these cases are probably linked with the simple cluster. The limited number of the two small clusters can be explained with the fact that the more complex and coherent cases are tested less. We conclude that it is worthwhile to distinguish our log file based on raw data and to split it up according to the clusters. Applying the process mining method "Genetic Algorithm plugin" on the resulting three log files we receive the process models shown in Figures 3 to 5.

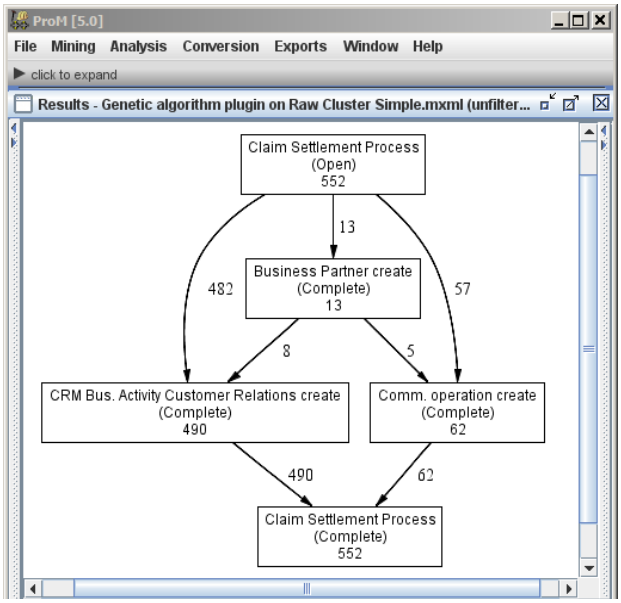

Figure 3: Cluster "Simple"

In Figure 3 we show the mined business process model using only claim cases from cluster "Simple". In Figure 4 we only consider cases from cluster "Moderate complex". The discovered model uses only cases from cluster "Complex" is shown in Figure 5. It is possible to extract sub-models from the common model based on the clustered information. All sub models are meaningful and can exist on their own.

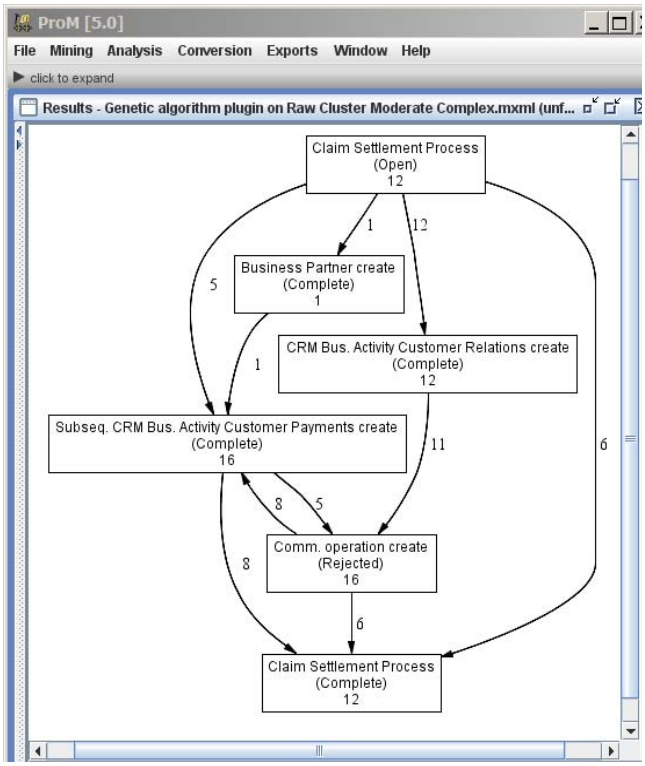

Figure 4: Cluster "Moderate complex"

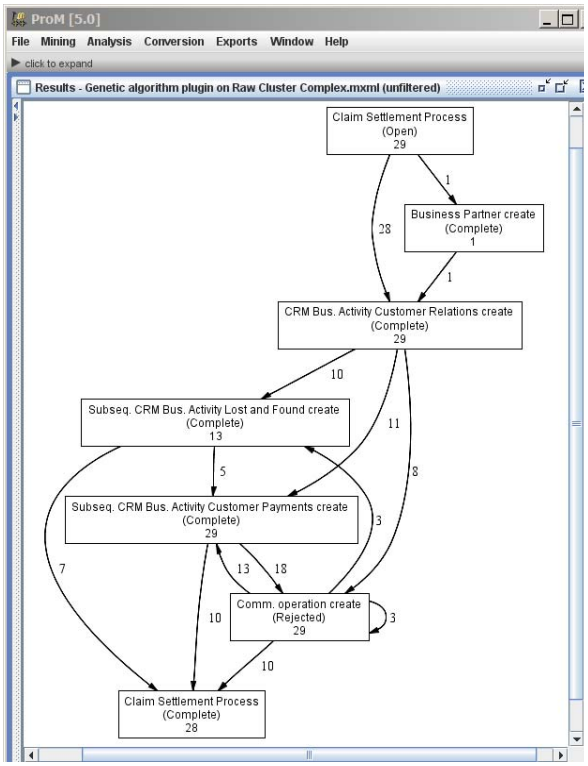

Figure 5: Cluster "Complex"

6. Evaluation

6.1. Extracting Business Process Models

It was possible to extract interpretable business process models. For the evaluation of the process models we use the fitness measurement guiding the "Genetic Algorithm plugin" and the analysis plugins "Behavioral Precision/Recall" and "Structural Precision/Recall" to measure how precise and complete the models are. Fitness is a function that evaluates how well the business process model is able to reproduce the behavior in the application [3]. The mined process model should be complete and precise from a behavioral perspective. The overall fitness of our model is 0.969. The fitness measures of the sub-models are 0.955, 0.978, and 0.967. Due to missing practical experiences there is no clear statement in the literature how to interpret these values. The closer the value tends to 1, the fitter the model is likely to be. Consequently, our mined models based on raw and aggregated data can be seen as complete and precise.

6.2. Results of Clustering

The clustering method enables a distinction between simple and more difficult and complex cases. The specific clusters are consistent to the overall business process model. Thus, it is possible to apply process mining on conducted log files. Furthermore, it is feasible to identify specific process structures according to the cluster. The activity "Comm. operation", for instance, belongs to the simple claim settlement processing, whereas activity "Lost & Found" associates to the moderate complex processing. An interesting observation is that joining cluster "Moderate complex" and "Complex" allows the passenger airline to conceal the information about the rejected claims. This result indicates that it is possible to extract meaningful business process models based on aggregated data while protecting sensitive data of the enterprise. We conclude that data aggregation strategies can support cross company process modeling based on process mining. Of course one has

to be careful with respect to the outer validity of the result. However, we think that our method has high potential for a transfer to processes from a running CRM system in a wide variety of branches. Our approach has certain drawbacks with respect to statistical coverage for some aspects related to data aggregation. In return, we gained valuable insights into many aspects of process mining based on aggregated data in business collaborative networks. The clustered log files provide an information basis for a later reuse of the experience gained in the course of the process execution. The more difficult and complex claims can be reviewed regularly to identify any ongoing issues. A better understanding of the customers' needs contributes to improve service quality and to reduce time spent on the claim settlement.

7. Conclusion and Future Work

In this paper we have discussed the main challenges collaborative business networks face in sharing their data and the usage of business collaborative infrastructures e.g. shared process mining. We have described how we addressed them in our approach and method. Moreover, we have briefly described the claim settlement process of a passenger airline interacting among customers and business partners. Further research into addressing process mining, data aggregation, and data sharing with regard to cross organizational aspects has to be done. In the short run, the next steps are to replace the CRM test system and get logs from live systems instead. In the long run, we have to consider the granularity of the collaborative business processes and consider them on a common level of abstraction.

8. References

[1] J. Borzo. Business 2010 - Embracing the Challenge of Change. A Report from the Economist Intelligence Unit sponsored by SAP, The Economist Intelligence Unit, 2005.

[2] J. Cardoso. Process control-flow complexity metric: An empirical validation. IEEE International Conference of Services Computing, Chicago, USA, pages 167–173, 2006.

[3] A. K. A. de Medeiros, A. J. M. M. Weijters, and W. M. P. van der Aalst. Genetic Process Mining: An Experimental Evaluation. Data Min. Knowl. Discov., 14(2):245–304, 2007.

[4] J. W. Forrester. Industrial Dynamics. A Major Breakthrough for Decision Makers. Harvard Business Review, 36(4):37–66, 1958.

[5] L. Măruşter, T. Weijters, G. de Vries, A. van den Bosch, and W. Daelemans. Logistic-based Patient Grouping for Multidisciplinary Treatment. Artificial Intelligence in Medicine, 26(1-2):87–107.

[6] J. B. M. Queen. Some Methods for Classification and Analysis of Multivariate Observations. In L. M. L. Cam and J. Neyman, editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967.

[7] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Măruşter, G. Schimm, and A. Weijters. Workflow Mining: A Survey of Issues and Approaches. Data and Knowledge Engineering, 47(2):237–267, 2003.

[8] B. Weber, B. F. van Dongen, M. Pesic, C. Günther, and W. M. P. van der Aalst. Supporting Flexible Processes Through Recommendations Based on History. 2007.

[9] A. K. A. de Medeiros, A. Guzzo, G. Greco, W. M. P. van der Aalst, A. J. M. M. Weijters, B. F. van Dongen, and D. Saccà. Process Mining Based on Clustering: A Quest for Precision. In Business Process Management Workshops, pages 17–29, 2007.

[10] A. K. A. de Medeiros, A. J. M. M. Weijters, and W. M. P. van der Aalst. Genetic Process Mining: An Experimental Evaluation. Data Min. Knowl. Discov., 14(2):245–304, 2007.

[11] SCHIMM, G.: http://www.processmining.org, 2008

[12] DONGEN, B.F.; MEDEIROS, A. K. A.; VERBEEK, H.; WEIJTERS, A., The ProM Framework: A New Era in Process Mining Tool Support Lecture Notes in Computer Science, Applications and Theory of Petri Nets 2005, 26th International Conference, ICATPN 2005, Springer-Verlag Berlin Heidelberg, S. 444-454, 2005