Abstract

This thesis describes a technique for lossy compression of an array of similar simulation results. The developed techniques use similarities between simulation results to achieve a more efficient compression than if each simulation model would be treated separately. We focus on crash test simulation results, which are generated in large numbers in vehicle development. For our compression approach, we decompose the vehicle model into parts that are defined in the simulation results. A crash test simulation file consists to a large extent of time-dependent node and element variables. By decomposing the model into components and knowing which components occur in which simulation results, the time-dependent variables of a part can be extracted for all simulation results. A data matrix is created per variable and per component, in which each column corresponds to a time step of a simulation result. The Predictive Principal Component Analysis (PPCA) method is applied to these matrices. The PPCA method first executes a PCA on a data matrix. A specialized optimization process is used to determine the number of principal components to be used to reconstruct the matrix. We use the reconstruction of the dimensionality reduction as a prediction of the data matrix. The PPCA methods can be applied both as offline method directly to a whole set of simulation results and as a online method, which allows to add simulation results afterwards. In both the online and offline variants, several freely selectable parameters occur which are crucial for the compression quality and speed. These parameters are investigated both theoretically and empirically on the basis of 3 test data sets. The results are compared with the state-of-the-art compression tool FEMZIP. In addition, the learning factor was introduced, which can be used to classify if a procedure benefits from the fact that a large amount of simulations are compressed rather than just one. The result of the PPCA prediction is a residual matrix that has the same dimension as the original data matrix. The induced Markov chain (iMc) encoder was used to compress the residual matrix. The iMc encoder is the first encoder that uses the topology of the grid as side information and - at the same time - can be applied in cases of a big alphabet. Besides the description of the practical implementation of the iMc encoder, the induced Markov chains are derived as the underlying data model. On the one hand, the entropy of the data set can be determined on the basis of the data model, which is advantageous for the runtime of the optimization step of the PPCA methods. On the other hand, the quality of the iMc coding can be investigated theoretically. It is shown that the stationary distribution of the underlying Markov chain for all practical applications differs only slightly from the distribution that is induced by the relative frequencies. In addition, the deviation can be estimated only on the basis of the topology of the grid. The evaluation of the iMc encoder is carried out in comparison to the encoders Rice and zlib on the residual matrices, which arise from an application of the PPCA Online and Offline methods on our benchmark data sets. The combination of the PPCA Offline method with the iMc encoder achieved the best results for all data sets. For the PPCA Online method, the results are not clear, as there are cases where the iMc encoder performs best as well as cases where the Rice encoder performs best. Since the Rice encoder is faster than the iMc encoder, we recommend to use the Rice encoder for the PPCA Online method.

Chapter 1

Introduction

Today, big data is a buzzword for a collection of data sets that is too large and too complex to process using traditional data processing algorithms. Since large automotive companies simulate several petabytes of data, the numerical simulations themselves represent a major data problem.
The focus of the application field examined in this thesis is the simulation of car crashes which dates back to the early 1980s [58]. This research area is of central importance for the development and quality management of automobile manufacturers, but at the same time leads to a constantly increasing demand for memory and bandwidth. Safety regulations and quality management require that part of the simulations be stored. The duration can range from a few months to several years. On the one hand, this leads to a large number of stored data records. On the other hand, the simulation results increase because the engineers want to improve the accuracy of the simulation results, e.g. by refining the model. The growing computing power makes this refinements possible.
A simulation data management system (SDMS) is usually applied to organize and handle a large number of simulation results. An SDMS enables the use of compression methods, that exploit redundancies between similar simulation results. An SDMS allows the definition of parts by connectivity and initial coordinates, which can only be stored once for a series of simulations instead of every time they occur in a simulation [118]. In addition, the time-dependent data of similar simulations, which make up the largest part of the data in the car crash result files, show high correlations among each other. It is, however, due to variations in the geometry, not trivial to exploit these correlations.
We have decided to duplicate nodes that occur in several components and to accept the overhead produced by this. Then, for each part individually, we apply a dimensionality reduction method to the information of all time steps for all available simulations that contain this part. Due to the decompression speed requirement, we further propose to perform a principal component analysis because its decompression can be accomplished by fast matrix operations and has the properties of random time step access and a progressive transmission. Since a simulation of a car crash takes several hours to a day, the time consumption of the compression is not crucial and an asymmetric compression method, in which the compression consumes more computational power, time, and memory, can be applied.
A disadvantage of the principal component analysis (PCA) - as with all dimensionality reduction methods - is that the bound of reconstruction error is not sharp [79]. But limiting

the maximal absolute error is a key demand for the engineers to apply lossy data compression to simulation results. Due to strong restrictions of precision, it is almost impossible to apply a dimensionality reduction directly as a black box method and achieve a good compression rate. Therefore, we use the reconstruction of the PCA as an approximation for the initial data and calculate the residual between these two data sets. Moreover, we apply a lossy data compression on the residual and keep the low dimensional representation as side information, see [80]. We call this strategy Predictive Principal Component Analysis (PPCA). PPCA can be categorized as a prediction method, since we compress a residual instead of the original data set, see Section 3.3.2. Moreover, we want to quantize the principal components and the coefficients for both a better compression rate and to assure that the decompression on different machines generates identical results. Therefore, in the decompression phase, we use only integer arithmetic for matrix calculations. The proposed dimensionality reduction method only exploits directly the redundancies between simulations and time steps. But the data is usually defined on a finite element grid. Therefore, we investigate how to encode integers assigned to the nodes of a finite graph. The topological relation of neighbored nodes will be transformed into a value-based relation by calculating so-called transition probabilities. Combined with the relative frequencies of the node values, these probabilities form a Markov chain as the initial distribution. Therefore, we call our strategy induced Markov chains (iMc). Since each node is identified with a random variable, the iMc approach can be categorized as a special case of Bayesian networks. We will investigate the properties of the iMc especially its data compression properties. We prove that the iMc of connected parts is mean ergodic, and, therefore, the entropy as a measure of the amount of information is well defined. The encoding of Markov chains and their trajectories is a well known field in the information theory [121]. For a practical implementation, we determine a minimum span tree for the graph. For this tree the iMc statistics are determined, which are used as secondary information for our arithmetic encoder. We call this approach iMc encoder. Furthermore, we investigate alternative approaches to generate statistics for our arithmetic encoder.

In summary, when simulating a car crash we are often confronted with a high redundancy of the data, e.g. between time steps, neighbors and models. For a good compression rate it is crucial to eliminate these redundancies. This task can be tackled in two ways. The first is to predict the data so that the resulting variables are independent or at least nearly independent and encode the residual. The second is to find and exploit the remaining dependencies in the data. A prediction usually modifies the distribution of a data set. We will investigate a combination of these two approaches. For the prediction part, we will use the PPCA dimensionality reduction approach. Regarding the remaining dependencies, we propose the iMc as a two-pass universal encoder, which is based on coding and sending empirical data measurements rather than a coding method based on an a priori probability model.

1.1 Overview of proposed lossy compression method

In this section, we give a short overview on all steps of our compression method that is presented in this dissertation. We focus on the compression of time dependent coordinates

and variables since the topology of the grid only needs to be stored once. Although, the PPCA can be combined with other encoders and the iMc encoder can be combined with different prediction methods, we see the combination of PPCA and iMc in several application fields as beneficial and propose it as a combined method.
We handle two cases. First, we consider cases where we have access to all simulation results we want to compress and second a case where simulations can be added after a first compression with all available simulation results. Therefore, the first situation is the initial stage of the second one. Furthermore, we assume that a user provides absolute precision for coordinates and variables. The reconstruction error of our compression approach must not exceeded the provided absolute precisions. A rough overview on the PPCA Offline method can be found in Figure 1.1.
In the situation where simulation results become available one after the other, the PPCA Online method can be applied, which is shown in Figure 1.2. For further information on the PPCA method, we refer to Chapter 4.
Since we do not exploit the topological relation in this method it is meaningful to apply a specified encoding scheme. This can be done by the induced Markov chain encoder. A short overview of its workflow can be found in Figure 1.3. In Chapter 5, the induced Markov chains are introduced in detail.
Finally, we combine the PPCA and iMc encoding, see Figure 1.4. The advantage of this combination is that during the optimization process the coding of graph-based data does not have to be carried out completely, but the expected size can be determined using entropy. This saves run time unlike when compression is fully executed.

1.2 Outline

In Chapter 2, we provide information about the content of crash test simulation results, how they are calculated and how they are managed by simulation data management systems (SDMS). Moreover, focus on the contents and properties of crash test simulation results.

In Chapter 3, we start with a short introduction to information theory and the resulting limitations of data compression. In addition, we depict the state-of-the-art compression methods in the context of simulation results. We distinguish between lossy and lossless compression and methods that exploit certain dependencies in time, space and as a new strategy similarities between sets of simulations. Moreover, we list the components of a state-of-the-art compression method in the context of crash test simulation results.

In Chapter 4, we introduce predictive principal component analysis (PPCA) that uses a linear dimensionality reduction method, namely PCA for the prediction of time-dependent data of simulation results. We distinguish between an offline method that is applied to all available data sets and an online method that can be used if the data sets are not available at the time of an initial compression. Since the PCA is a linear method and a crash test is a non-linear problem, we investigate how the residual of the prediction can be compressed efficiently.

Since we do not exploit the dependencies based on the topology of the finite element

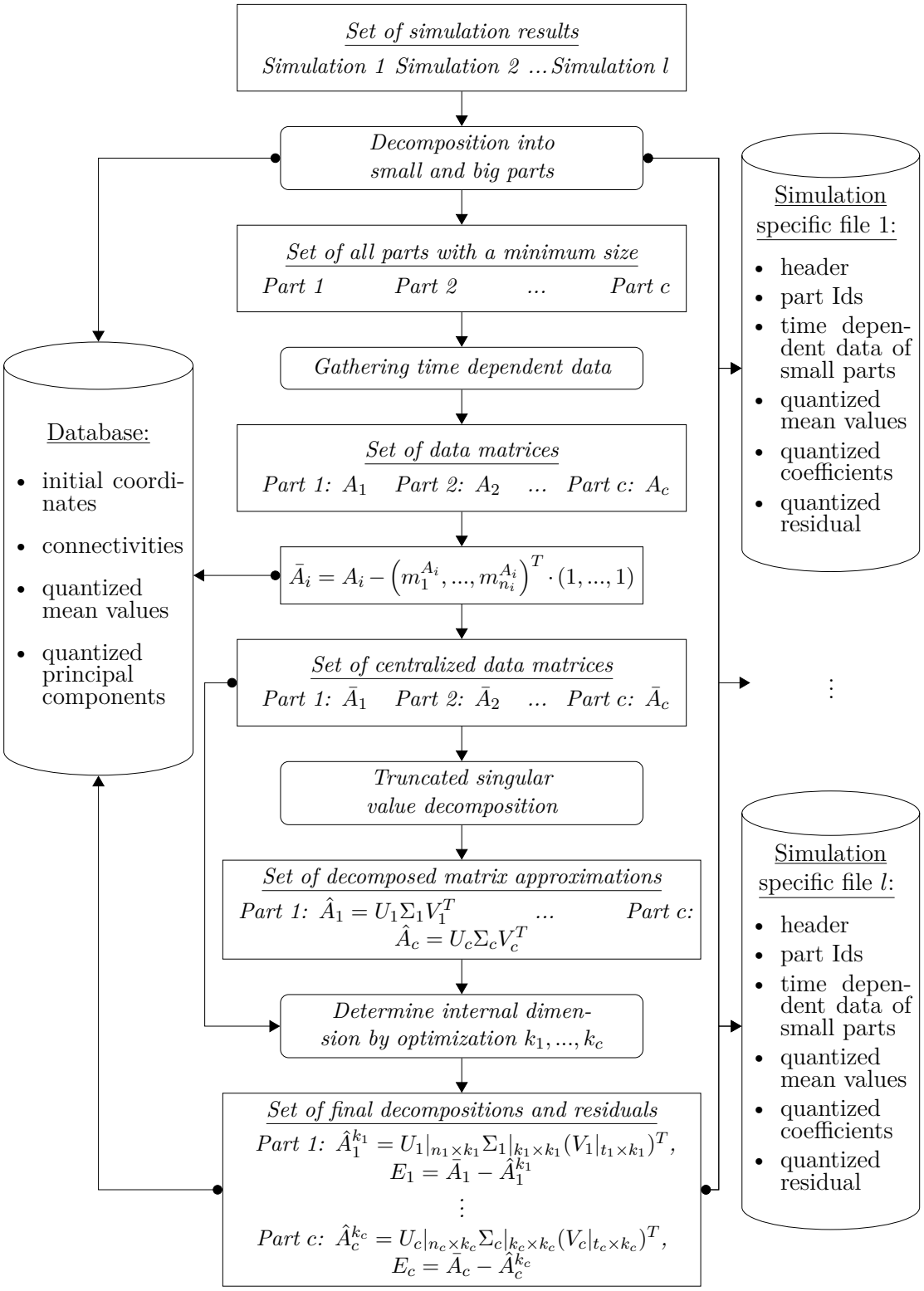

Figure 1.1: Workflow of the proposed compression method including the PPCA Offline method. We assume that two parts are identical if they are identical in the first time step.

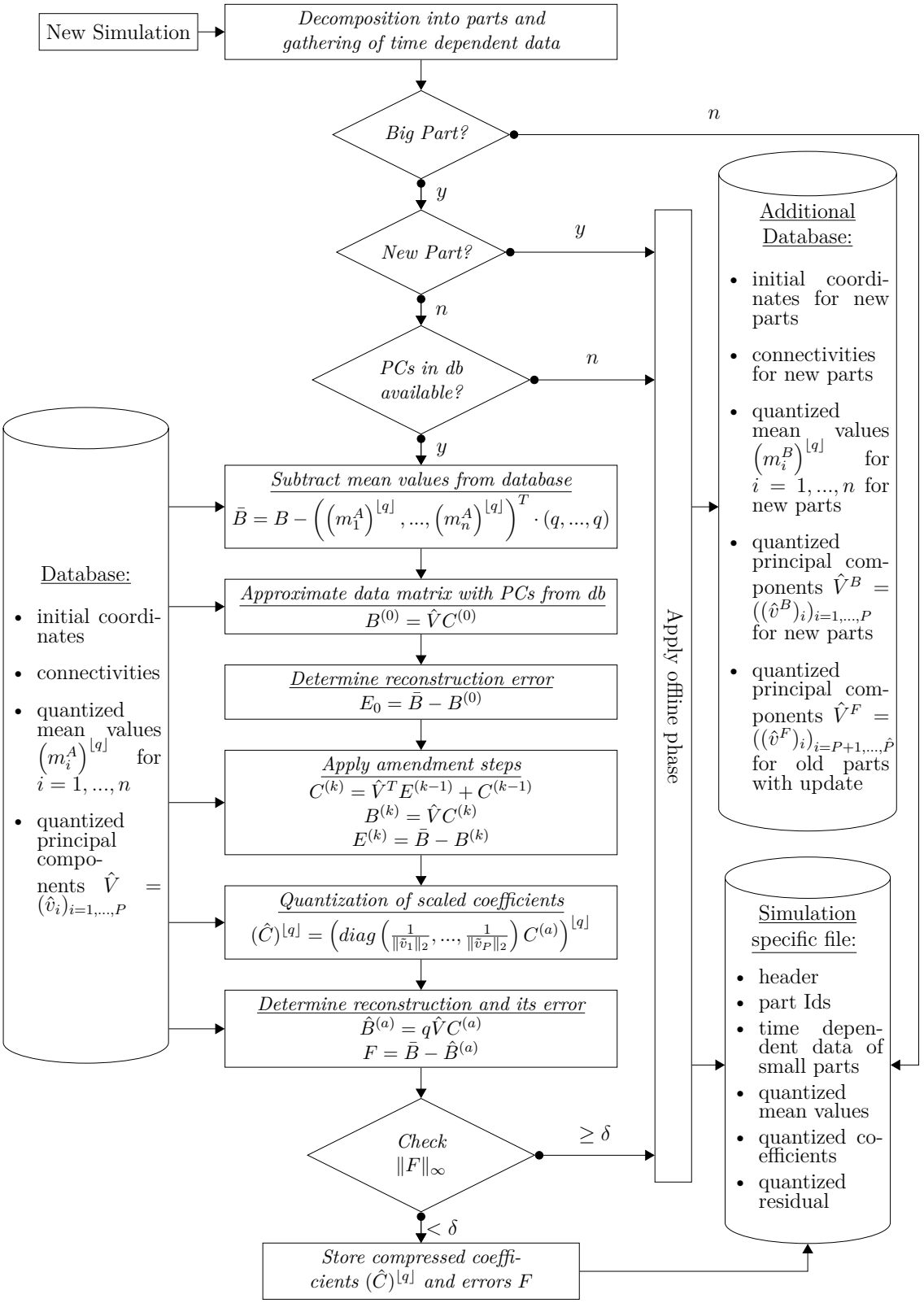

Figure 1.2: Workflow of the proposed compression method including the PPCA Online method. In this example, we add one simulation result. In our example we have P number of principal components (PCs), the number of nodes n, and a amendment steps.

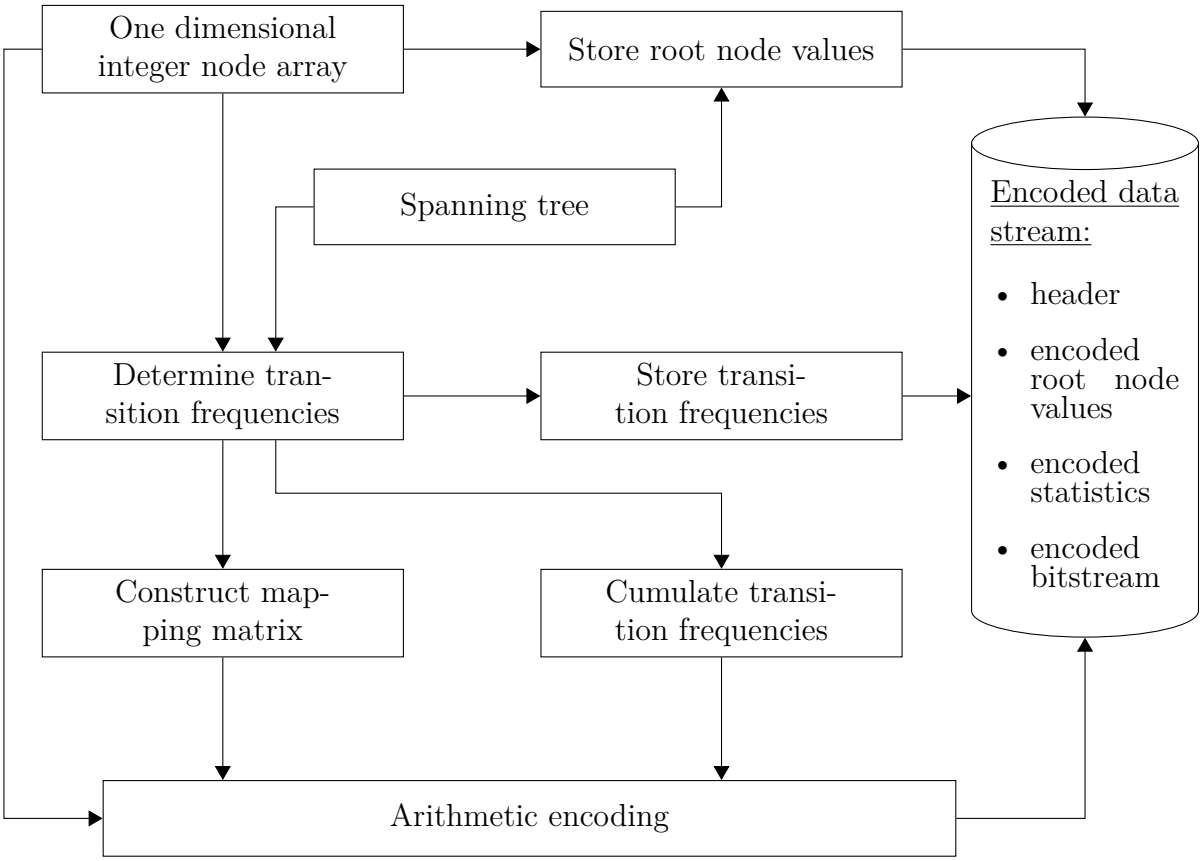

Figure 1.3: Diagram of the induced Markov chain workflow for one part.

mesh, we introduce the induced Markov chains (iMc) in Chapter 5. We begin with the application of the new induced Markov chains within an entropy encoder. Thereafter, we define the iMcs as a statistical model. We can combine the encoding scheme with different prediction methods that additionally reduce the dependencies in the investigated data sets. Moreover, we categorize and delimit the iMc from graphical models like Bayesian networks and Markov Random Fields (MRF).

In Chapter 6, we state our results for the PPCA for three benchmark data sets, see Section 6.2 and compare them with those of the state-of-the-art tool FEMZIP™ [124]. Furthermore, we combine the PPCA and the iMc and compare it with the combination of PPCA with zlib encoder [111] and the Rice encoder [110, 145].

In Chapter 7, we conclude with a short summary of our results and an outlook on future work.

1.3 Contributions

The most important part of this thesis is the development and elaboration of the mathematical background of the two procedures Predictive Principal Component Analysis (PPCA), see Section 4.3, and induced Markov chain (iMc) encoder, see Section 5.1, as

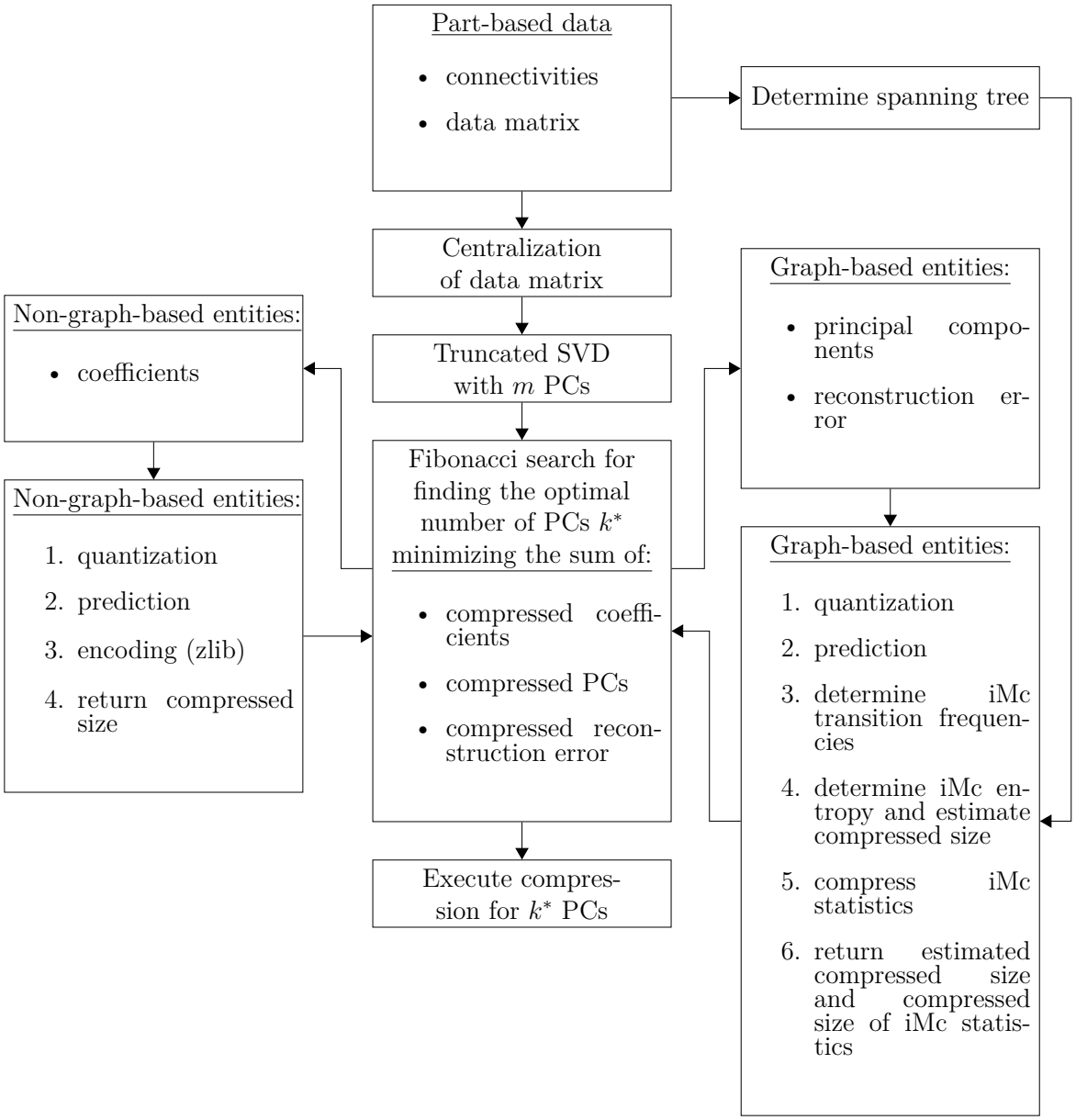

Figure 1.4: Combination of PPCA and iMc encoding. As part of the optimization process for graph based data, there is no need to actual encode the data, since iMc is an entropy encoder and the entropy can be determined applying the transition frequencies.

well as their combination, see Chapter 6. Both PPCA and the iMc encoder have already been published by the author, see [90, 96]. The PPCA is a machine learning technique that allows the reconstruction of compressed results with a given accuracy. This distinguishes the PPCA from other machine learning methods, which update a database on which decisions are made rather than extend it like the PPCA. An important contribution is made by the amendment steps, which make it possible to reuse the principal components meaningfully despite the lossy compression. A detailed differentiation of the PPCA from other procedures can be found in Section 4.1.

The iMc encoder is the first graph-based encoder that can also be used with a large al-

phabet. An important feature is that the overhead caused by the underlying model of the encoder can be estimated upwards. The estimation depends only on the topology of the graph, not on the alphabet. A differentiation of the iMc encoder from other methods can be found in Section 5.8.

The combination of PPCA offline method and iMc encoder proves to be the method that achieves the best compression rates. Due to the entropy encoder's property and the a priori determination of transition probabilities, the compressed size can be determined without having to perform the encoding completely. This is an advantage in the time-consuming optimization process determining the intrinsic dimension.

The contents of Chapter 2 and Chapter 3 are mainly known facts. Some evidence not available in the literature but important for this work is given, e. g. Theorem 3.42 and Theorem 3.63. Finally, we investigate the term "ergodicity" in the context of Markov chains and information theory since they are used in different ways, see Section 3.1.3.

1.4 Danksagung

Ich möchte mich an dieser Stelle bei all jenen bedanken, die mich bei der Anfertigung dieser Doktorarbeit unterstützt haben.

Ich danke Frau Prof. Dr. Caren Tischendorf für die fortwährende Betreuung dieser Arbeit, die Motivation den mathematischen Kern meiner Anwendung zu finden und diesen sauber zu definieren. Bedanken möchte ich mich bei Dr. Martin Weiser, der als Mentor im Rahmen der Berlin Mathematical School stets als Ansprechpartner bereit stand und der mit seinem Feedback und seiner kritischen Expertise diese Arbeit vorangebracht hat. Bei Prof. Dr. Rudolph Lorentz bedanke ich für einen Forschungsaufenthalt an der Texas A&M University at Qatar und seine engagierte Betreuung.

Mein ausdrücklicher Dank geht an Clemens-August Thole, der mich bestärkt hat PCA Verfahren für die Kompression von Scharen von Simulationsergebnissen zu untersuchen, wie auch seine fortwährende Unterstützung. Er hatte immer ein offenes Ohr für meine Fragen und Probleme.

Auf dem Weg zu dieser Arbeit haben mich viele ehemalige und aktuelle Kollegen unterstützt, kritisch hinterfragt und motiviert. Insbesondere sind hier Dr. Matthias Rettenmeier, Dr. Lennart Jansen und Stefan Mertler zu nennen bei denen ich mich ausdrücklich bedanken möchte. Auch danke ich Frau Yvonne Havertz für das detaillierte Lektorat.

Auch Danke ich der Berlin Mathematical School (BMS), die die Reisekosten meines Konferenzbesuchs auf der IEEE Konferenz ISSPIT in Noida Indien übernommen hat, sowie eine Plattform für viele spannende und gewinnbringende Bekanntschaften bietet. Neben meinen BMS Mentor, möchte ich Dr. Benjamin Trendelkamp-Schroer herausheben.

Besonders möchte ich mich bei meinen Eltern bedanken ohne deren immer währende Unterstützung ich mein Studium in dieser Form nicht hätte durchführen können.