Revealing cause-effect relations in comorbidities analysis using process mining and tensor network decomposition

Joan T. Matamalas
Dept. Eng. Informàtica i Matemàtiques
Universitat Rovira i Virgili
Tarragona, Spain
joantomas.matamalas@urv.cat

Alex Arenas
Dept. Eng. Informàtica i Matemàtiques
Universitat Rovira i Virgili
Tarragona, Spain
alexandre.arenas@urv.cat

Antoni Martínez-Ballesté
Dept. Eng. Informàtica i Matemàtiques
Universitat Rovira i Virgili
Tarragona, Spain
antoni.martinez@urv.cat

Agusti Solanas
Dept. Eng. Informàtica i Matemàtiques
Universitat Rovira i Virgili
Tarragona, Spain
agusti.solanas@urv.cat

Carlos Alonso-Villaverde
Unidad de enfermedades crónicas
Xarxa Sanitària i Social Santa Tecla
Tarragona, Spain
calonvi@gmail.com

Sergio Gómez
Dept. Eng. Informàtica i Matemàtiques
Universitat Rovira i Virgili
Tarragona, Spain
sergio.gomez@urv.cat

***Abstract*—The existence of certain comorbidities, the co-occurrence of different diseases in the same individual, is well-known in the medical community. However, finding temporal cause-effect relations between those diseases constitutes a big challenge. Clinical records can be used to extract the required information, but their analysis is elusive due to the vast and heterogeneous amount of data. We propose a new methodology for time-preserving tensor networks decomposition to be applied in the analysis of big data problems where the temporal dimension of the key factual fields must not be modified. This methodology will also allow the creation of a new process mining modeling which can capture the cause-effect relations as low-order tensors associated to the transitions of the mined processes, and whose structure takes the form of a multilayer complex network. All these theoretical and methodological advances will allow their application to real biomedical data to analyze comorbidities.**

***Index Terms*—tensor decomposition, process mining, comorbidities, network science**

I. Introduction

The wide adoption of information systems in companies, research institutions and administrations has opened the door to the collection and storage of massive amounts of data. In particular, many problems in biology, medicine and health care, generate huge amounts of high dimensional data to be analyzed. Institutions and companies (*e.g.*, hospitals, banks, ISPs or accounting firms) are progressively adopting events-logging policies that lead to the generation of data in the form of log files containing detailed lists of events/actions performed on their information systems (*e.g.*, patient-related events in hospitals, access control events in critical infrastructures, resources use in chemical companies, etc.). In the health care domain, this trend is specially relevant and it is fostered by the adoption of the concept of Smart Health [1].

In general, event-driven data are stored in (semi-)structured files such as plain texts, CSV files, and relational and non-relational databases. By analyzing these data the correctness and suitability of processes and company's policies could be determined, and the identification of hidden dependencies and correlations is also possible. However, the analysis of big event-driven data requires novel methodologies to efficiently process them in tolerable computational times while maintaining accuracy and precision.

A commonly accepted, compact, mathematical representation for this massive multidimensional data are tensors. Tensors are organized multidimensional arrays with multiple indices. The order of a tensor is the dimension of the array needed to represent it, or equivalently, the number of indices needed to label a component of that mathematical object. A challenge in big data analysis consists in using linear algebra machinery to reduce the dimensionality of tensors and, thus, of the data to investigate. Mathematically well-grounded lower-order approximations are then a subject of intense research in the area [2]–[5]. The mathematical approximations used to reduce tensors have started to attract attention in the context of big data analysis [6]. In particular, *tensor networks* (TN), a countable collection of tensors connected by contractions, promise to be a very useful approach to big data analysis in distributed computing. The first breakthroughs on the analysis of tensors decomposition were provided by Hitchcock [7] and Tucker [8], with the Canonical Polyadic and Tucker decompositions, respectively. However, only recently a few scalable tensor decomposition strategies have been proposed,

This work has been supported by the Generalitat de Catalunya project 2017-SGR-896, Spanish MINECO project FIS2015-71582-C2-1, and Universitat Rovira i Virgili project 2017PFR-URV-B2-41. AA acknowledges financial support from the ICREA Academia and the James S. McDonnell Foundation.

978-1-5386-8161-9/18/$31.00 ©2018 IEEE

e.g., the Memory-Efficient Tucker in [9].

The main problem to face in tensor networks decomposition is that of interpreting the resulting low-rank projections. For many applications, this is not a drawback given that the main objective is to deal with a dimensionality reduction of the multi-dimensional data with no restrictions. However, in some cases, the data is time-stamped and proper analysis requires to preserve this particular dimension (time) in the data. This is the case of *process mining* analyses, a young research discipline (initial studies date back to the late 1990s) aiming at discovering, monitoring and improving real processes by extracting knowledge from event log files. Process mining promises to be particularly important in highly dynamic environments such as health care [10].

In health care, time-stamped/time-dependant data records convey longitudinal meaning on the evolution of patients and the set of diseases that they have suffered throughout their lives. The analysis of these data can reveal the presence of one or more concurrent diseases, the so-called *comorbidities* [11], and might help to explain the reasons for a given patient to develop them, which is the objective of this work.

This paper is organized as follows. In section II, we describe the background methodology that we need to tackle the causal-effect relations in comorbidities, namely, tensor decomposition and tensor networks, process mining, and comorbidities. Then, in section III we detail our proposed methodology. Finally, in section IV we discuss the main breakthroughs of this new approach and future perspectives.

II. Background

A. Tensors, tensor decomposition and tensor networks

In the last few years tensors have been found to be useful in a completely different field: data analysis. With the possibility to collect huge amounts of information, beyond the traditional relational databases, new approaches have emerged to try to understand the structure and relations between data. For example, suppose we are interested in the relationships between people (or companies, products, diseases, etc.). We could use a matrix $\boldsymbol{A}$ to represent these relations, *e.g.* with component A_{ij} accounting for the number of interactions between individuals i and j. If the interactions can be of different types, a matrix is not enough: either you start using sets of independent matrices (one for each kind of interaction), or you proceed to replace them with a 3*rd*-order tensor $\underline{\mathbf{T}}$ with components $T_{i,j,k}$, where the new index k is used to identify the kind of interaction. If the persons can also be classified in different categories, then we could end up with a 5*th*-order tensor, $T_{i_1,c_1,i_2,c_2,k}$, where c_1 is the category of individual i_1, and the same for person 2. We realize from this example that the restriction of using just two indices (*i.e.*, matrices) to describe data is rather arbitrary, and that there are situations in which using tensors instead of matrices is more convenient, since it has more expressive power, and the relation between the mathematical object (the tensor) and the reality it describes is much closer. This approach has been used in [12] to establish a tensorial mathematical formulation of multilayer networks, which has become the natural framework for the study of pairwise interactions of diverse types between elements. See Figure 1 for a visual representation of ordinary linear objects, tensors, and some standard operations on them.

From all the theory of tensor algebra, the part which has received more attention in relation to its application to data analysis is that of tensor decomposition. The objective of tensor decomposition consists in finding a set of (usually lower order) tensors which, when combined using standard tensor operations, recover the original one. A couple of examples for the case of matrices (2*nd*-order tensors) are the well-known Principal Components Analysis (PCA) and Singular Value Decomposition (SVD). When the rank of the matrix is not maximum, both PCA and SVD allow for loseless dimensionality reduction. Otherwise, the dimensionality reduction is accomplished with a controlled information loss. In the same way, generalizations of SVD exist for higher order tensors, such as the Canonical Polyadic [7] and Tucker [8] decompositions. More recently, new decomposition schemes have appeared, *e.g.*, the Hierarchical Tensor decompositions [13], [14] and the Tensor Trains [15], [16], which can be considered as belonging to the emerging class of Tensor Networks. The main idea of Tensor Networks is to have distributed collections of tensors, each one of low order (between 1 and 6), which are combined in pairs using index contractions, thus enabling the reconstruction of tensors of arbitrarily high order.

Among the many benefits of using tensor network decomposition for large-scale data analysis we should highlight the following: efficient compressed formats for large multidimensional data; distributed data representation; numerical stability and robustness to noise of lower-order tensor approximations; unified framework for all data operations; natural multidimensional extensions of commonly used data analysis algorithms; availability of graphical representations of tensor networks, which simplify tensor manipulation; existence of efficient software libraries for full and sparse tensor representations and their operations.

B. Process mining

The continuous monitoring of the alignment between event data generated by business and functioning processes and their intended design was unusual and was mainly performed manually by human experts since very recently. From these initial experiences, it has been proven that monitoring and analyzing process events is highly beneficial both economically and organizationally. However, the cost and difficulty of manual analysis prevents many organizations from putting these techniques in place; specially when these analyses require huge amounts of multifaceted data.

With the aim to ease the analysis of events, the *process mining* discipline emerged, combining machine learning and data mining techniques on the one hand, and process modeling and analysis on the other. Its main goal is to *"discover, monitor and improve real processes by extracting knowledge from event logs readily available in today's information systems"* [17]. The processes obtained from process mining techniques are

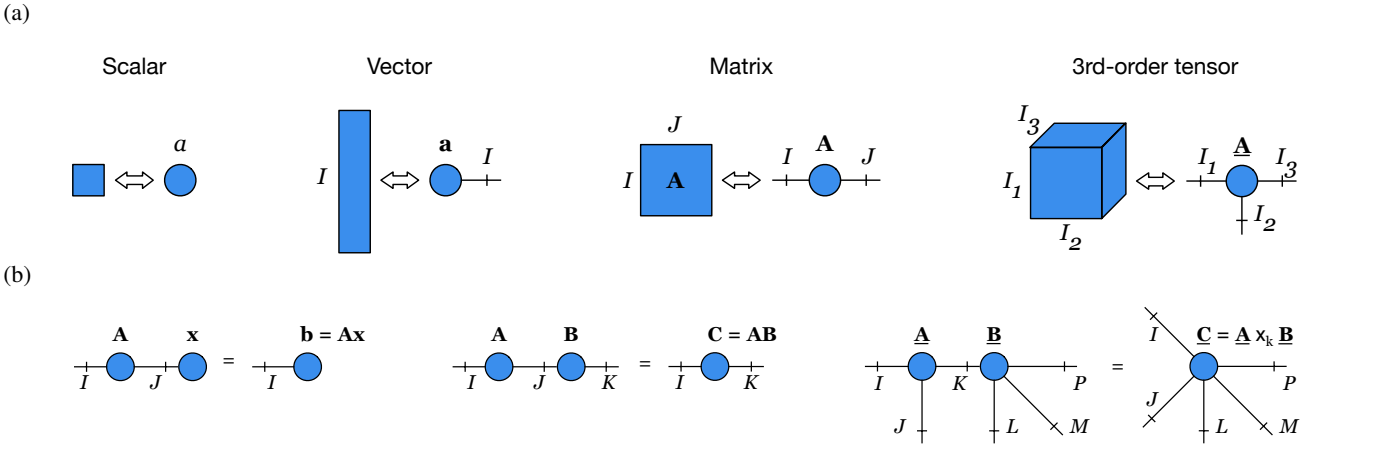

Fig. 1: **Tensor network diagrams notation**. (a) The basic buildings blocks for tensor network diagrams, from left to right: scalar, vector, matrix, and tensor representations. (b) Algebraic operations using tensor diagrams: matrix-vector multiplication, matrix by matrix multiplication, and contraction of two tensors.

represented as *process models* in some common notation, such as BPMN, UML or Petri nets. According to such notations, processes are represented as a group of activities (*i.e.* each activity is a defined step in the process), whose ordering/relationship is described as causal dependencies.

Most organizations typically associate the use of process mining techniques to the characterization of process models, but the scope of process mining is much broader. Although process mining is an emerging research field whose attention has increased within the research community in the last decade, its beginnings date back to the 1990s. Agrawal *et al.* [18] introduced the idea of modeling business processes from log data as activity graphs in the context of work flow management systems. Similarly, Datta [19] proposed a method to discover business process models using variants of finite state machines combined with probabilistic approaches in the context of work flow management and business process redesign. Cook and Wolf [20] proposed three methods for discovering software engineering processes from event data: (i) using neural networks, (ii) using finite state machines, and (iii) using Markov models. More concretely, the authors identified the last two methods as the most promising ones, whilst the neural-network-based was not sufficiently mature. Herbst [21] was one of the first in addressing the discovery of more complicated (and realistic) processes, which may contain duplicated tasks, in the context of work flow management by using inductive approaches.

C. Comorbidities analysis

Some chronic diseases (*e.g.,* diabetes, hypertension) can appear with different degrees of aggressiveness on the same subject and might evolve in a variety of ways. Each of these pathological processes do not follow a predefined course, independent from the other concurrent processes. These chronic pathologies can be exacerbated; or other acute disease may appear in an independent manner. In clinical practice, several indices are used to evaluate the immediate prognosis of the concomitance of chronic processes and the burden of diseases, such as: the Charlson index, the Comorbidity-polypharmacy score, Elixhauser comobidity measure, or the diagnosis-related group. These indexes help mainly to predict the risk of death. Depending on the country, between 30% and 60% of the population over 70 years old have comorbidities. However, knowledge about the combination of multiple complex chronic diseases is yet scarce.

Clinical practice reveals that the secondary prevention of a pathology can act as the inducer of a new comorbidity, or on the contrary the primary prevention of another. To understand this complexity, in the last decade, complex analyses have been carried out on a number of data sets resulting from different levels of studies of diseases (*e.g.,* data on genetics, transcriptomics, intracellular-signals, protein-protein interaction, metabolomics, epidemiological data, drugs design, cell cultures models, animal models, clinical expression, etc); and the term *network medicine* has been introduced. Due to the huge amount and diversity of the collected data, it has proven to be very difficult to obtain a global perspective on the problem and most of the times, narrower and more specific studies are performed.

Diseases can be represented as nodes that are interconnected by means of edges (like process mining techniques do), which represent transitions from one disease to another or, alternatively, the addition of one disease on the previous ones. As a result, extensive maps of interconnections between different pathologies can be obtained. Note that these nodes could represent other features such as syndromes, phenotypes or biomarkers; and the links between them represent their particular relationship; until now, extensive analyses of genomic associations have been carried out, but other crucial factors such as metabolic pathways, intracellular signaling pathways, transcriptomics and the interaction of various molecules in the cellular cytosol are progressively incorporated to the studies.

Analyzing these maps or their equivalent process-like coun-

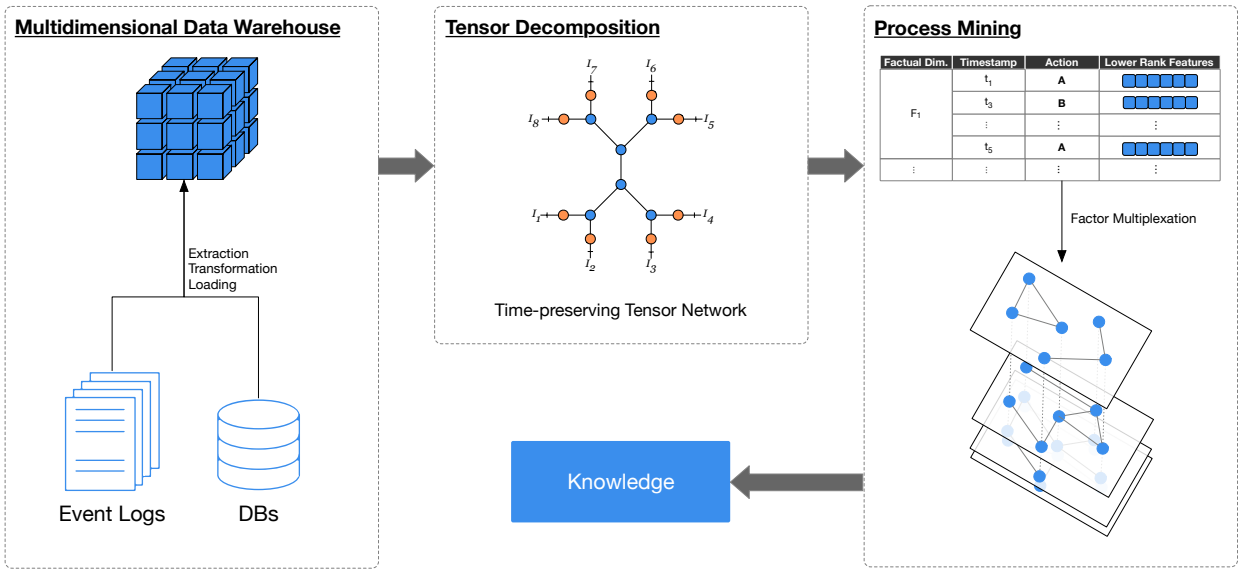

Fig. 2: **From data to knowledge**: Conceptual scheme of the methodological approach of the proposal.

terparts (also known as *diseasome maps*) of interactions is an important step towards the understanding of the association, relation and evolution of comorbidities and will give physicians greater predictive and preventive capabilities; However, current interaction maps are fairly limited and do not consider fundamental constraints related to time. Hence, there is a clear need for the development of interaction maps with time-constraints (*i.e.*, processes) that enable the analysis of comorbidities from a multidimensional perspective but, at the same time, considering the influence and importance of time (*i.e.*, concepts such as past, present, concurrency and future have to be fully considered).

III. Our proposal

The detection of cause-effect relations among diseases is the main objective of this proposal. However, the discovery process requires the analysis of huge volumes of data. Thanks to the use of tensor decompositions, transitions between nodes can be augmented with low-order tensors that help explain the cause-effect relations behind those transitions. Decomposed tensors and tensor networks can be fed to properly tuned process mining algorithms, and once transitions are identified, each of them can be associated with an aggregation of *lower rank features* that form a *low-order tensor*, able to represent the cause-effect relations hidden in the transitions/edges of the mined processes.

Current clinical practice considers diseasome maps with a static vision, which could be of great help when they are incorporated into medical knowledge, and are likely to provide new diagnostics. However, we approach comorbidities from a process mining perspective with time constraints that incorporates time-related concepts into diseasome maps, and allows the modeling and detection of inducers or protectors associated with the features, described by lower-order tensors in the links/edges representing transitions between nodes. By approaching the problem from this perspective, it is possible to show that *interactomes* can present distant associations with other nodes, the epigenetic effects of diseasomes, interactomes, and the heritability of the diseasome map itself. Nevertheless, process mining is not possible if the time dimension of the key factual fields is not respected. Unfortunately, the approximations involved in tensor network decomposition may break this temporal constraint. Thus, new tensor network decompositions must be developed, preserving time for the key fields and respecting also the controlled error and efficient computability requirements. Additionally, we take advantage of the connection between tensors and multilayer networks to select optimal unsupervised dimensionality reduction of tensor modes [22].

We show in Fig. 2 the scheme of the whole process. We create the proper theory to transform a multidimensional data warehouse into tensors (which is a natural representation for this kind of data), and we provide the theoretical means for the decomposition of those tensors into tensor networks that preserve important dimensions such as the time. Next, we apply novel process mining techniques on the aforementioned tensor networks to discover hidden processes within the data. In addition, we create low-order tensors to model the cause-effect relations that explain the transitions between nodes of the identified processes. By using this low-order tensors we are able to represent cause-effect relations with a multilayered structure inspired in complex-networks theory. Finally, we apply them to the analysis of comorbidities by using real biomedical data.

IV. Discussion and Conclusions

The present proposal leads to important breakthroughs in the three different fields involved. With respect to comorbidities analysis, it contributes by providing new information on the relation/interaction of comorbidities in diseasome maps by using

process mining models built upon tensor decompositions with time preservation. This opens the door to the discovery of new paths for prevention, early detection and intervention. Also, it makes possible to show the epigenetic effects of diseasomes, and the heritability of the diseasome map itself. Additionally, the current state of knowledge on tensor decomposition and tensor networks requires the design of new mathematically grounded and computationally efficient tensor network decomposition algorithms, capable of preserving unaltered both the time dimension and the additional fields required for process mining. At the same time, process mining must be extended to be properly applied on tensor decompositions and tensor networks. Moreover, the transitions information, augmented with novel low-order-based tensor structure, allows to explain the cause-effect relations hidden within the transitions of the identified processes from multiple perspectives (*i.e.*, following a multilayered, complex-networks-inspired approach).

We expect to use this new approach with real multidimensional data provided by the Sant Pau i Santa Tecla Hospital in Tarragona, Spain, whose data warehouse contains more than 30 years of biomedical data from thousands of patients organized in more than 200 million heterogeneous records, totalling hundreds of terabytes of data. Thanks to the help of medical doctors and biologist we will be able to select the most relevant information related to comorbidities and create multidimensional data marts focused on the topic.

REFERENCES

[1] A. Solanas *et al.*, "Smart health: A context-aware health paradigm within smart cities," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 74–81, 2014.

[2] E. Stoudenmire and D. J. Schwab, "Supervised learning with tensor networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 4799–4807.

[3] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[5] F. Scholz, A. Mantzaflaris, and B. Jüttler, "Partial tensor decomposition for decoupling isogeometric galerkin discretizations," *Computer Methods in Applied Mechanics and Engineering*, vol. 336, pp. 485–506, 2018.

[6] A. Cichocki, "Tensor networks for dimensionality reduction, big data and deep learning," in *Advances in Data Analysis with Computational Intelligence Methods*. Springer, 2018, pp. 3–49.

[7] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[8] L. R. Tucker, "The extension of factor analysis to three-dimensional matrices," in *Contributions to mathematical psychology*, H. Gulliksen and N. Frederiksen, Eds. New York: Holt, Rinehart and Winston, 1964, pp. 110–127.

[9] T. G. Kolda and J. Sun, "Scalable tensor decompositions for multi-aspect data mining," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 363–372.

[10] A. Solanas, F. Casino, E. Batista, and R. Rallo, "Trends and challenges in smart healthcare research: A journey from data to wisdom," in *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, Sept 2017, pp. 1–6.

[11] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," *Medical care*, vol. 43, no. 11, pp. 1130–1139, 2005.

[12] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, "Mathematical formulation of multilayer networks," *Physical Review X*, vol. 3, no. 4, p. 041022, 2013.

[13] W. Hackbusch and S. Kühn, "A new scheme for the tensor representation," *Journal of Fourier analysis and applications*, vol. 15, no. 5, pp. 706–722, 2009.

[14] L. Grasedyck, "Hierarchical singular value decomposition of tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 2029–2054, 2010.

[15] I. V. Oseledets and E. E. Tyrtyshnikov, "Breaking the curse of dimensionality, or how to use svd in many dimensions," *SIAM Journal on Scientific Computing*, vol. 31, no. 5, pp. 3744–3759, 2009.

[16] I. V. Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[17] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[18] R. Agrawal, D. Gunopulos, and F. Leymann, "Mining Process Models from Workflow Logs," in *Proceedings of the International Conference on Extending Database Technology*. Springer, 1998, pp. 467–483.

[19] A. Datta, "Automating the Discovery of AS-IS Business Process Models: Probabilistic and Algorithmic Approaches," *Information Systems Research*, vol. 9, no. 3, pp. 275–301, 1998.

[20] J. E. Cook and A. L. Wolf, "Discovering Models of Software Processes from Event-Based Data," *ACM Transactions on Software Engineering and Methodology*, vol. 7, no. 3, pp. 215–249, 1998.

[21] J. Herbst, "A Machine Learning Approach to Workflow Management," in *Proceedings of the 11th European Conference on Machine Learning*, vol. 1810. Springer, 2000, pp. 183–194.

[22] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Structural reducibility of multilayer networks," *Nature communications*, vol. 6, p. 6864, 2015.