Received 28 December 2020
Revised 24 December 2021
Accepted 30 December 2021

Data-driven approach to predict the sequence of component failures: a framework and a case study on a process industry

Sara Antomarioni, Filippo Emanuele Ciarapica and Maurizio Bevilacqua
Department of Industrial Engineering and Mathematical Science, Università Politecnica delle Marche, Ancona, Italy

Abstract

Purpose – The research approach is based on the concept that a failure event is rarely random and is often generated by a chain of previous events connected by a sort of domino effect. Thus, the purpose of this study is the optimal selection of the components to predictively maintain on the basis of their failure probability, under budget and time constraints.

Design/methodology/approach – Assets maintenance is a major challenge for any process industry. Thanks to the development of Big Data Analytics techniques and tools, data produced by such systems can be analyzed in order to predict their behavior. Considering the asset as a social system composed of several interacting components, in this work, a framework is developed to identify the relationships between component failures and to avoid them through the predictive replacement of critical ones: such relationships are identified through the Association Rule Mining (ARM), while their interaction is studied through the Social Network Analysis (SNA).

Findings – A case example of a process industry is presented to explain and test the proposed model and to discuss its applicability. The proposed framework provides an approach to expand upon previous work in the areas of prediction of fault events and monitoring strategy of critical components.

Originality/value – The novel combined adoption of ARM and SNA is proposed to identify the hidden interaction among events and to define the nature of such interactions and communities of nodes in order to analyze local and global paths and define the most influential entities.

Keywords Association rules, Social network analysis, Predictive analytics, Predictive maintenance, Decision making, Big data analytics

Paper type Research paper

1. Introduction

Operations and processes are continuously under the control of the decision-makers in order to improve organizational performance (Komljenovic *et al.*, 2016; Bhattacharjee *et al.*, 2020). Typically, the asset maintenance activities help to guide the physical performance of maintenance equipment and tasks efficiently, trying to maximize the Return on Investment (ROI) of the asset (Van Horenbeek and Pintelon, 2014). Asset managers have to select the most appropriate maintenance policy for company plants; moreover, every day, they need to decide when to maintain each asset, what tasks need to be done and which parts need to be replaced at each maintenance interval (Madu, 2000). The European Committee for Standardization, in BS EN 13306:2017, defined the following maintenance policy for an industrial environment:

International Journal of Quality & Reliability Management
Vol. 40 No. 3, 2023
pp. 752-776
Emerald Publishing Limited
0265-671X
DOI 10.1108/IJQRM-12-2020-0413

(1) Corrective maintenance: an intervention is carried out after the occurrence of a failure in order to restore the normal system functioning;

(2) Preventive maintenance: maintenance is carried out at predefined intervals or conditions;

(3) Predictive maintenance: maintenance is carried out according to a forecast of the significant parameters of a component based on a thorough analysis of known characteristics.

In literature, there are many models of data-driven decision support systems for predictive maintenance. Some of these models are used to implement condition-based maintenance (CBM) solutions (Benanne and Yacout, 2012; Lin and Tseng, 2005), others are used to implement modeled or simulated predictive maintenance (statistically predictive) (Gerum *et al.*, 2019; Antomarioni *et al.*, 2019). In the asset maintenance field, the main research focus is on predicting the occurrence of component failures to reduce unexpected events and the consequent interruption of the production processes (Chuang *et al.*, 2019). Less attention has been concentrated on developing a framework for the decision-making process to achieve satisfying levels of reliability and to avoid wasting resources using huge piles of unstructured data. Existing research is valuable, but quantitative methods to structurally analyze and predict relationships between component failures and avoid them through the predictive replacement of critical ones, to the best of the authors' knowledge, are not present in the literature.

For this reason, the asset maintenance framework proposed in this work aims to address this research gap through the introduction of an innovative decision-making tool based on a data-driven methodology, using data collected through sensors and management systems in order to look for ways to give support to maintenance managers in predictive and prescriptive analytics. In particular, the framework proposed in this work is based on the concept that a failure event is rarely random and is often generated by a chain of previous events connected by a sort of domino effect (Bubbico *et al.*, 2018). Hence, the data-driven relies on the joint adoption of Association Rule Mining (ARM) and Social Network Analysis (SNA) to define the hidden interactions between components that lead to a domino effect between failures. The conjunction of these two methodologies is helpful because the ARM will be used to identify the interaction among events and the SNA to define the nature of such interactions. In comparison to previous works, this framework allows researchers to identify communities of nodes in order to analyze local and global paths and define the most influential entities.

This framework will be explained and tested through a case study of a medium-sized oil refinery. Above all, in the process industry, the various components (pumps, valves, pipes, tanks, . . .) are physically connected to each other, and a fault or simply a maintenance or a revamping on a component can trigger a series of events in other components (malfunctions, maintenance or failures). It is, therefore, necessary to have a tool capable of connecting these events and these entities, avoiding, in this way, disruptions causing delays, bottlenecks or accidents (Ishizaka and Nemery, 2014), mainly focusing on the components having a high influence on the plant (Gupta *et al.*, 2013).

The rest of the paper is as follows. This introduction is followed by a literature review that analyses the methods and objectives of data-driven decision support systems (Section 2). Section 3 presents our framework and data-driven decision support system, while the case study and the data set are described in Section 4. The discussion about theoretical contribution and practical implications are described in Section 5. To conclude, Section 6 summarizes the paper and outlines future research directions.

2. Literature review

Different data-driven methods for predictive maintenance have been proposed in the literature. Most of these methods implement CBM solutions, while others are used to

implement modeled or simulated predictive maintenance (statistically predictive). Both these approaches aim to define critical assets for which a physical plant owner should allocate maintenance resources. The condition-based methods focus on time and/or condition monitoring data (often provided with sensors) and statistical trending (Yam *et al.*, 2001), while the latter is focused on a prediction or simulation based on an expected potential for failure (Kaparthi and Bumblauskas, 2020); for instance, in Bumblauskas *et al.* (2017), a predictive perspective is adopted to anticipate faults and an improve equipment reliability and availability through Markov models. Létourneau *et al.* (1999), instead, focus on defining a predictive maintenance policy through the implementation of several techniques (i.e. instance-based learning, Naive Bayesian classifier and decision trees).

Another important classification that can be made on data-driven models is related to the objectives that these models seek to achieve. For example, some models aimed at predicting the remaining useful life of an asset (e.g. Liu *et al.*, 2019; Wu *et al.*, 2019; Baptista *et al.*, 2016), even considering the lifespan of a part and the lifetime maintenance cost (Kim *et al.*, 2019), and others again attempt to predict failures or their causes (e.g. Baptista *et al.*, 2018; Kumar *et al.*, 2019; Tang *et al.*, 2019). Different techniques can be integrated to reach the proposed objectives. In this regard, Langone *et al.* (2015) integrate three different techniques to address the maintenance of industrial machines, namely, clustering, non-linear autoregressive model and least squares support vector machines (SVM). Similarly, Saravanan *et al.* (2010) and Kankar *et al.* (2011) use both the SVM and artificial neural networks for fault diagnosis in gearboxes, while Salahshoor *et al.* (2010) fuse SVM and adaptive neuro-fuzzy inference for detecting and diagnosing failures in industrial steam turbines.

According to this classification, Table 1 describes the main literature contributions proposed in the field of the maintenance policy regarding data-driven predictive maintenance techniques and the objectives of the papers. Six main objectives have been identified in analyzing such contributions: fault prediction, fault detection and diagnosis, optimal maintenance schedule definition, equipment reliability and availability, normal behavior modeling and, lastly, remaining useful life (RUL) estimation. Regarding the data-driven techniques, instead, 17 of them have been taken into account. As presented in Table 1, neural networks are widely applied in all the fields described by the six objectives. Indeed, for their versatility in modeling all kinds of processes, they can be applied for modeling several classes of problems. For instance, Gerum *et al.* (2019) apply a recurrent neural network to study rail and geometry defects in order to schedule maintenance interventions. The artificial neural network deployed by Bangalore and Tjernberg (2015), instead, serves as a fault detector, as well as the radial basis function neural network employed in Gharoun *et al.* (2019), that also compare its performance with the adaptive neuro-fuzzy inference. Kusiak and Verma (2012), though a neural network, address both the fault prediction and the normal behavior modeling of wind turbines' bearings. Izquierdo *et al.* (2019) focus on the adoption of an artificial neural network to monitor and improve the reliability of assets, aiming to integrate the operational context information collected from them, while both Li *et al.* (2013) and Mazhar *et al.* (2007) focus on the RUL prediction using the same construct.

Among the other techniques, the SVM and Markov models were widely used in several applications. More in detail, according to this literature review, authors mostly use SVM to address fault prediction, diagnosis and detection problems. For example, Datong *et al.* (2009) propose an online time-series fault prediction, while Purarjomandlangrudi *et al.* (2014) compare SVM and anomaly detection algorithms to diagnose failures in rolling element bearings. Baptista *et al.* (2016) applied SVM to predict the RUL in the aeronautic field, while Medjaher *et al.* (2012) pursued the same objective through Gaussian Hidden Markov models in studying bearings' useful life. Chen *et al.* (2019), instead, used Hidden Markov models for RUL estimation as well as to schedule maintenance interventions.

The application of other techniques like, for instance, ARM is more limited in the maintenance field. Indeed, applications can be found for predicting component failures

Table 1.
Summary of literature contributions classified by their objectives and data-driven techniques applied

Data-driven techniques	Objectives					
	Fault prediction	Fault detection and diagnosis	Optimal maintenance schedule	Equipment reliability and availability	Normal behavior modeling	RUL estimation
Support Vector Machine	Datong *et al.* (2009), Baptista *et al.* (2018), Langone *et al.* (2015)	Saravanan *et al.* (2010), Salahshoor *et al.* (2010), Kankar *et al.* (2011), Purarjomandlangrudi *et al.* (2014)			Crespo Márquez *et al.* (2019)	Baptista *et al.* (2016), Liu *et al.* (2019)
K-nearest neighbors	Baptista *et al.* (2018)	Wang (2016)				
Regression	Baptista *et al.* (2018), Langone *et al.* (2015)	Schlechtingen and Santos (2011)				
Neural Networks	Baptista *et al.* (2018), Kusiak and Verma (2012), Bangalore and Tjernberg (2015), Gharoun *et al.* (2019)	Saravanan *et al.* (2010), Kankar *et al.* (2011), Schlechtingen and Santos (2011)	Lin and Tseng (2005), Yam *et al.* (2001), Gerum *et al.* (2019)	Lin and Tseng (2005), Izquierdo *et al.* (2019), Yam *et al.* (2001)	Kusiak and Verma (2012), Crespo Márquez *et al.* (2019)	Mazhar *et al.* (2007), Wu *et al.* (2019), Li *et al.* (2013)
Random Forest	Baptista *et al.* (2018)		Gerum *et al.* (2019)		Crespo Márquez *et al.* (2019)	
Instance-based learning			Létourneau *et al.* (1999)			
Naive Bayesian classifier			Létourneau *et al.* (1999)			
Decision trees	Romanowski and Nagi (2001)		Létourneau *et al.* (1999), Romanowski and Nagi (2001)	Romanowski and Nagi (2001)		
Logical Analysis of Data			Bennane and Yacout (2012)			

(*continued*)

Table 1.

Data-driven techniques	Objectives					
	Fault prediction	Fault detection and diagnosis	Optimal maintenance schedule	Equipment reliability and availability	Normal behavior modeling	RUL estimation
Adaptive neuro-fuzzy inference	Gharoun *et al.* (2019)	Salahshoor *et al.* (2010)				
Association Rules	Antomarioni *et al.* (2019), *This work*	Cunha *et al.* (2006)			Crespo Márquez *et al.* (2019)	
Case-based reasoning	Bahga and Madisetti (2011)					
Anomaly detection algorithm		Purarjomandlangrudi *et al.* (2014)				
Markov models	Bumblauskas *et al.* (2017), Gerum *et al.* (2019)	Kumar *et al.* (2019)	Sharma *et al.* (2018), Chen *et al.* (2019)	Bumblauskas *et al.* (2017)		Medjaher *et al.* (2012), Chen *et al.* (2019)
Clustering	Langone *et al.* (2015)	Kumar *et al.* (2019)				
Bayesian Networks		Tang *et al.* (2019), Sattari *et al.* (2021)	Abbassi *et al.* (2016)			
Social Network Analysis	*This work*		Kim *et al.* (2019), *This work*			

(Antomarioni *et al.*, 2019) or for diagnosis (Cunha *et al.*, 2006). Other authors, like Crespo Márquez *et al.* (2019), use ARM to explain the results obtained through an artificial neural network and compare the performances with a random forest and an SVM.

According to this literature review in the asset maintenance research field, the research focus is mainly on predicting the occurrence of component failures in order to reduce unexpected events and the consequent stoppage of the production processes. Thus, awareness in the decision-making process is mandatory for achieving satisfying levels of reliability and avoiding the waste of resources (Sattari *et al.*, 2021). The SNA is even rarer in this field of research: to the best of the authors' knowledge, only Kim *et al.* (2019) apply such technique to define the optimal maintenance schedule in a cost reduction perspective, even though they do not employ the ARM, neither optimize the component selection taking into account time, budget and personnel constraints. Integrating these techniques can provide a valuable methodology for having a deeper understanding of the relations among events through graphic representation. Indeed, as shown in other application fields (e.g. human factor risk management and environmental risk management), they resulted in being successful (Ciarapica *et al.*, 2019; Bevilacqua and Ciarapica, 2018). In this context, the proposed framework, combining different Big Data Analytics techniques, provides an approach to expand upon previous work in the areas of prediction of fault events and monitoring strategy of critical components. Among the identified contributions, the work by Romanoski and Nagi (2001), by applying the decision tree algorithm to identify the critical subsystem and define the optimal maintenance schedule, addresses the same objectives as the proposed work. Remarkably, both methodologies ensure a graphical representation of the solution, even though the techniques employed are different. However, from the author's perspective, the main advantage in the proposed work is the ability to take into account operative constraints (e.g. time and budget), instead of only highlighting the critical subsystems. Also, the graphical representation provided in this work supports immediately noticing the most critical components in the network, while in the work by Romanoski and Nagi (2001) – even though the logical path can be followed – the criticality of the subsystems analyzed is not immediately identifiable.

3. Asset maintenance framework

The emerging techniques in the Big Data Analytics field can provide valid support for the decision-making process since they allow the simultaneous analysis of several data sources and a wide amount of data. In this context, this paper proposes an asset maintenance framework based on a three-layer model. Each layer is devoted to a specific activity, namely, data collection, data management and, lastly, a data-driven decision support system. The framework, reported in Figure 1, will be described in the following paragraphs. It is created considering data belonging to the process industry, but it can also be generalized and applied to different industrial fields by adapting the data sources.

3.1 Data collection layer

The data collection layer represents the first layer of the model, being the aim of the framework – the development of data-driven support for the decisional process. In this sense, the quality of the entire framework relies on appropriate data collection. In this approach, three different macro-categories of data sources are considered.

On-field reports: monitoring the operations of an industrial plant is fundamental to controlling it; maintenance department supervisors have to check the sub-plants in order to notice and register any possible malfunctioning affecting its performance. During each inspection, the supervisor has to record all the relevant information and create a report; in case of abnormal events, there are procedures to follow and, possibly, immediate corrective interventions to perform, annotating these details too. Usually, such reports follow unstructured or semi-structured paths since they are often characterized by free-text

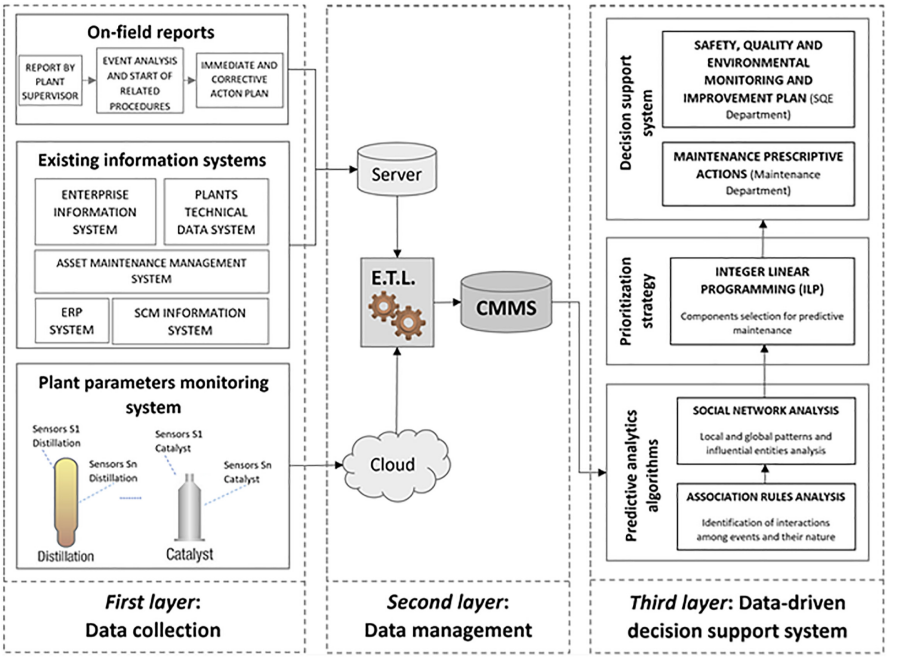

Figure 1.
Asset maintenance framework

annotations, making their computerization non-univocal or, at least, complex. However, these data contain useful information due to supervisors' broad knowledge of the process.

Existing information systems: data coming from the on-field reports are integrated with the information systems of the company. Information on both the normal operating conditions and on adverse events are stored in such systems, like:

(1) EIS: it collects the administrative data, work orders type (e.g. specific replacement of a component, lubrication . . .) and the related costs, purchasing orders, corrective interventions, their details and costs;

(2) ERP (Enterprise Resource Planning): it stores information regarding resource and inventory management;

(3) Plant technical data system: data regarding product and process characteristics in terms of design and functioning;

(4) Asset maintenance management system: data stored in this information system regards all the maintenance activities carried out in the plant (corrective, preventive or predictive), highlighting the date, the kind of intervention, the broken component (or set of components), the team in charge of the intervention, the duration of the intervention, etc.

(5) Supply chain management system: it records the data from suppliers and customers – in both cases regarding their general data, order data, real-time status and quality rate.

Plant monitoring system: the functioning of the plant is monitored by a series of sensors measuring the production process data, like flow, pressure and density. Besides, some of the

components are equipped with embedded sensors so that their state is currently monitored, generating a large amount of data to analyze. Each of them has its own IP address and communicates with a cloud-based application. Hence it is fundamental that cloud resources are allocated efficiently (Ergu *et al.*, 2013). This information integrates the systems mentioned above, giving a complete overview of the process.

Data coming from the aforementioned sources have to be integrated in order to extract information and knowledge for making informed decisions. Thus, in the second layer, the management of the collected data is performed.

3.2 Data management layer

The data management layer that is the second layer of the framework aims at integrating data coming from different data sources into a unique one. More specifically, the information contained in the company's server and cloud-based applications have to be merged, cleaned and transformed, in order to create a unique source to perform the analysis in the last step.

In this process, all the possible problems affecting the data have to be solved in order to analyze only a consistent set of data. For example, errors in recording the measures (e.g. misreading, repetitions) have to be removed or replaced with valid ones, while heterogeneities generated by different terminologies used in each source have to be standardized. In addition, some data could be filtered, selecting only the attributes considered relevant for the study. The Extraction, Transformation and Loading (ETL) process is carried out to integrate data from the original sources to a data warehouse, the computerized maintenance management system (CMMS). Specifically, the plant monitoring and supply chain management data are extracted from a Cloud Application, while the other ones come from the company's server. The use of a CMMS ensures a global view of a company's operations since it allows the collection of clean data from all the sources, integrating them and providing an aggregation of historical and real-time conditions. This technology is particularly useful in the case of an elevated number of components to monitor and maintain (Marquez and Gupta, 2006). In this way, the predictive analysis can be performed relying on a reliable, integrated data warehouse.

3.3 Data-driven decision support system layer

The data-driven decision support system developed in this study aims at defining the optimal set of components to predictively be maintained in order to achieve high-reliability levels, avoiding the occurrence of failures. The analysis carried out in this layer is organized in three steps, capitalizing on two predictive analytics techniques, ARM and SNA and on the optimization of an Integer Linear Programming (ILP) model.

Specifically, through the ARM, the failure analysis is carried out: from the data stored in the CMMS, information on the failures which occurred on the analyzed asset is extracted to identify the sets of components frequently failing together and the corresponding failure probability. Indeed, ARM aims at individuating the attribute-value pairs that frequently occur together (Buddhakulsomsiri *et al.*, 2006). In this way, the knowledge of the asset behavior is deepened by extracting previously unknown patterns from the data. The SNA is then used for relating the components frequently failing together and identifying the possible failure propagations among the related components. In this context, the use of the graph theory underlying the SNA facilitates the understanding of the association among component failures. It provides a global view of the interactions among the components frequently failing together. Finally, considering the Out Degree (OD) and the Betweenness Centrality (BC) metrics extracted from the SNA, two different strategies are defined: first, the components deserving particular monitoring are identified through the definition of a minimum OD threshold; then, when a failure on a component occurs, an ILP model is solved to

define the optimal set of components to be predictively replaced. In this case, the BC is used in the objective function. By optimizing the ILP model, the decision-maker is completely guided by the information extracted from the data analytics.

In the following sub-paragraphs, the description of the three phases of the analysis is developed, together with a formal definition of the techniques applied and the ILP model formulation.

3.3.1 Preliminary failure analysis. Considering the data collected, prepared and stored in the CMMS, a study of the failures occurring on the asset object of the analysis is required. This step is important for the identification of both critical components and their relationships: that is, it is important to identify the sets of components frequently failing together (within a given time interval). In this sense, the methodology selected to study such relationships is the ARM.

Before mining the Association Rules (ARs), two aspects have to be taken into account:

(1) The components included in the study: depending on the characteristics of the asset, it is important to define whether all the components are relevant for the analysis or only some of them. Indeed, the study aims to create an interrelation among the critical components: the components whose replacement does not impact the working conditions of the industrial plant might be excluded from the analysis in order not to lose the focus on the critical ones.

(2) The limit for the time interval: since it is stated that the aim of the framework is to identify the relations among components frequently failing together (i.e. within a given time interval), the temporal dimension has to be limited in order to provide interesting results. Indeed, considering a time interval that is too short, does not provide any significant connection, while having an overly long interval does not provide any connection in the opposite sense, which presents false relations among failures. Again, even in this case, the expertise of the decision-makers is crucial.

3.3.2 Association rule. The aim of ARM is the identification of hidden and previously unknown relations in a wide amount of data, supporting the decision-makers in their processes. In the following, a formal definition of the ARs and the procedure to mine them is explained.

Let $K = \{k1, k2, \ldots, kn\}$ be a set of n binary attributes named items and $T = \{t1, t2, \ldots, tm\}$ be a set of m transactions. Each transaction t_i is unique and contains a subset of the items (itemsets) selected from K. In our framework, an item is a component of the analyzed asset, while a transaction is a set of components failing within a defined time interval. As defined by Agrawal *et al.* (1993), an AR is an implication $\alpha \rightarrow \beta$, such that α and β are itemsets ($\alpha, \beta \subseteq K$) having no common items ($\alpha \cap \beta = \varnothing$). In other words, given a time interval of one week, the rule $\alpha \rightarrow \beta$ is defined if and only if component β fails within one week from the failure of component α. The strength of the rule can be defined through several metrics, among which, we recall:

(1) $\text{supp}(\alpha, \beta) = \frac{count\{\alpha \ U \ \beta\}}{m}$; the support of the rule, that is defined as the set of transactions containing both α and β. Remarkably, this measure represents the joint probability of having α and β in a transaction ($P(\alpha, \beta)$);

(2) $\text{conf}(\alpha \rightarrow \beta) = \frac{\text{sup}p\{\alpha,\beta\}}{\text{sup}p\{\alpha\}}$; the confidence of the rule, instead, is the set of transactions containing α, which also contain β. In this sense, the confidence can be seen as the conditional probability $P(\beta \mid \alpha)$, so it provides a measure of the rule's strength.

The ARM is performed according to the following roadmap:

(1) Define the frequent itemsets, namely, the itemsets appearing in T more frequently than user-specified minimum support; the algorithm selected in this study is the FP-growth (Han *et al.*, 2007);

(2) Considering each itemset IS defined in the previous step, all the ARs A → B are generated such that A U B = IS.

According to the aim of the study, we are interested in creating the relations among components frequently failing together so that the Social Network (SN) describing such relationships can be created and analyzed.

3.3.3 Social network analysis. As defined by Otte and Rousseau (2002), an SN is the representation of a social structure. It can be described by an ordered pair of vertices (or nodes) and connected by edges (E), G = (V, E). The classical application of SNA regards the analysis of social structures and of the interactions among a set of actors: the actors are the nodes of the network, while the interactions are the edges. An SN is generated considering the associations among components extracted in the previous step of the analysis. In the current approach, the SNA is used to represent and analyze the relations among components frequently failing together, which is for representing the ARs mined.

Specifically, in the proposed framework, the actors, thus the nodes, are the components, while the interactions (arcs) are the concurrent failures: that is, if node i is directly connected to node j ($i \rightarrow j$), it means that the rule $i \rightarrow j$ is mined in the previous step, indicating that when component i fails, usually components j fails as well. The probability of such a conditional event is given by the confidence of the rule. The confidence value of the rule represents the weight of the arc.

In order to define which nodes might be more critical in terms of failure probability, two SN metrics are applied for the analysis:

(1) OD: is calculated as the weighted sum of the arcs outgoing from a node (Knoke and Yang, 2008). Specifically, OD represents a measure of how much a node is connected to another: the higher the OD, the higher the probability that one of the following components fails.

(2) BC: is determined as follows: the shortest weighted paths between all couple of nodes are determined; the BC value equals the sum of the shortest weighted paths on which the node appears (Brandes, 2001). In other words, the BC measures how much a node is influent across the network (Scott and Carrington, 2011) since a node having a high BC value can be considered as a bridge among separate portions of the network. Thus, if a component fails, the right candidate for predictive maintenance would be a component characterized by a high BC value.

In the current framework, the OD (1) is considered as an indicator of the risk of failure of subsequent components: indeed, the OD is calculated for each node, sorting them in descending order. The components at the top of the list should be carefully monitored since they represent the most critical ones. The BC (2), instead, is considered when a failure on a component occurs: the failed component is definitely replaced but also a predictive intervention on the consequent ones could be performed, with the aim of avoiding the propagation of a failure chain across the network. In defining the best set of components to be predictively replaced, the decision-maker is supported by an ILP model, whose formulation can be interpreted as follows:

$$\max \sum_j BC_j x_j \tag{1}$$

$$\sum_j c_j x_j \leq \boldsymbol{B^{max}} \tag{2}$$

$$\sum_j t_j x_j \leq \boldsymbol{T^{max}} \tag{3}$$

$$\sum_j R_j x_j \leq \boldsymbol{R^{max}} \tag{4}$$

$$x_j \in \{0, 1\} \ \forall \ j \tag{5}$$

where the decision variable x_j represents the j components of the assets. It can assume a value of 1 if component j is selected for the predictive maintenance, or 0, otherwise, as expressed in constraint (5). The objective function (1), to maximize, assures that the components having the highest BC are selected. Constraint (2) requires that the selection is performed according to a predefined maximum budget ($\boldsymbol{B^{max}}$), considering the cost of each component (c_j). Constraints (3) and (4), similarly, require the selected components to respect a maximum amount of time ($\boldsymbol{T^{max}}$) and resources ($\boldsymbol{R^{max}}$) to perform the intervention. Indeed, among the data, the time required to replace a component (t_j) and the number of operators necessary for replacing the component (r_j) is known.

3.3.4 Predictive maintenance strategy. The following bullet list aims at summarizing the main stages of the predictive maintenance strategy explained in the previous sections, providing a useful roadmap to be followed by the maintenance department.

During the normal operating conditions of the asset, the procedure proposed in this framework is the following:

(1) Monitor the components having high values of OD, specifically all the components j such that $OD_j > OD_{max}$ in order to detect failures early;

(2) When a failure on component i is detected:

  (2.1) Perform a corrective intervention on i;

  (2.2) Extract the set of consequent components using the ARM (all components j such that $i \rightarrow j$);

  (2.3) Create the SN graph and calculate the BC_j for all j components;

  (2.4) Solve the ILP model (1)–(5);

  (2.5) Perform a predictive maintenance intervention on the optimal set of components identified in 2.4.

(3) Return to 1.

In the next paragraph, an application of the presented framework is proposed, in order to clarify its explanation.

4. Research approach application

In the following, an application of the research approach proposed is deployed. The data belong to a medium-sized Italian oil refinery plant. The refinery was established in 1950 and has an extension of about 700,000 m^2. It currently has a processing capacity of 3,900,000 tons/year (equivalent to 85,000 barrels/day). It is equipped with a land shipment system for a potential of about 12,000 tons/day and a sea reception system through marine terminals for tankers up to 400,000 tons. The direct employees are about 500, while the induced personnel is about 2,000 people, representing mainly electromechanical, engineering, instrumentation and software, transport companies. Its production process is certified for what concerning quality procedures (ISO 9001), environmental management (ISO 14001) and safety (OHSAS, 18001).

The focus of the study is on the topping sub-plant. The time interval of reference for the analysis regards a period of three years, during which operational data (e.g. flows, density

and pressures) have been monitored: in the case of missing data, the mean value of the previous working day has been used to replace them, as well as in the case of anomalous measurement reported (e.g. out of scale values). In addition, the work orders and maintenance activities required for the components of the plants have been considered, compared and integrated with the notes taken by the supervisors of the plant, in order to define whether all the activities performed on the plant have been inserted into the information system and to ensure consistency among the two information sources. As required by the first layer of the decision support system, the integrated data are taken from the CMMS of the refinery.

Considering the data of the CMMS, the preliminary failure analysis is carried out: in all, 82 components are monitored in the sub-plant. Statistically, 46 of them have been considered for the analysis since they caused 615 failures over the 767 failures that occurred in three years, which is more than 80%. In order to define the time interval to consider the failures "concurrent," the maintenance department members have been interviewed to understand, based on their experience, which interval could be suitable for searching related failures. According to the interviewees, the maximum interval is set to two weeks: this means that the relations searched in the data concern component failures taking place at a distance of maximum two weeks. Then, the association rules describing such relations are mined. An excerpt from the rules extracted is reported in Table 2. The rules can be interpreted as follows: a failure of component C15 is followed by the failure of C2 within a two-week time interval with the confidence of 0.866, hence in 88.6% of the cases. Remarkably, when C2 fails, also C15 fails as well in the following two weeks.

In Appendix, the list of the components' ID and their related name is reported.

Taking into account all the ARs mined, the graph is built (Figure 2) and the SNA is performed. The 46 components represent the nodes of the SN, while their relations are the AR identified in the previous step. In all, 724 arcs connect the 46 nodes. As noted before, the weights assigned to the arcs are the respective confidence values of the corresponding rule. The thickness of each arc is proportional to the confidence of the relationship represented. For example, according to the representation, the confidence of the rule C41 → C15 (confidence = 1.000) is higher than the one of C41 → C25 (confidence = 0.375). For the sake of clarity, the weights are not reported in Figure 2.

The size of the nodes, instead, is proportional to the OD of the node itself; even its color is furtherly indicative of the OD: in particular, pink nodes are characterized by a high level of OD, while green ones by a lower level and the more intense the corresponding color, the higher the OD.

At this stage of the analysis, the calculation of SNA is required, and hence, for each node, the OD is determined and reported in Table 3. Then the OD_{max} threshold has to be defined in

$\alpha \rightarrow$ B	Confidence
C15 → C2	0.866
C2 → C15	1
C15 → C40	0.657
C15 → C13	0.657
C40 → C15	0.92
C13 → C15	0.958
C2 → C40	0.677
C2 → C13	0.677
C40 → C2	0.84
C13 → C2	0.875
.....	

Table 2.
Excerpt of the association rules mined

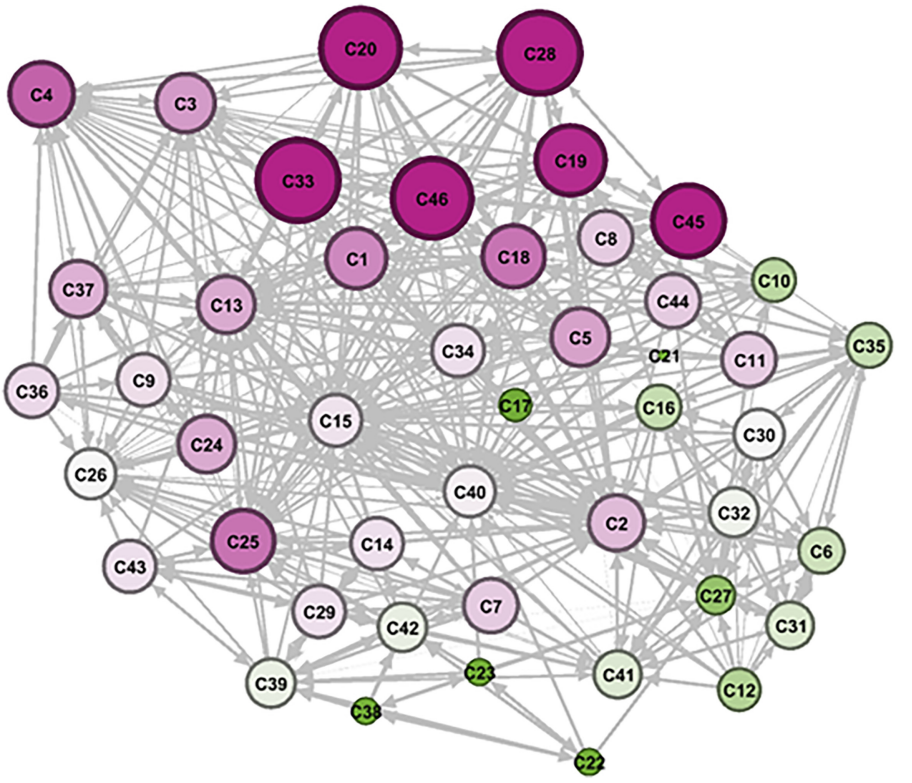

Figure 2.
Representation of the Social Network characterized by 46 nodes and 724 arcs

order to identify the components that need to be carefully monitored by the operators: as explained before, the higher the OD, the higher the probability of failure of one of the consequent components.

Undoubtedly, the selection of the OD_{max} threshold has an impact on the operations of the maintenance department members: the higher the threshold value, the lower the number of components to be monitored. At the same time, it corresponds to a higher risk of failure. On the contrary, if the threshold is too low, there would be a high number of components to be carefully monitored, and the effort may not be repaid by the benefit. Considering the values reported in Table 3, an $OD_{max} = 14.00$ has been identified by company maintenance managers as a good compromise since it would require the careful monitoring of five components (C33, C28, C46, C20 and C45). Lowering the threshold, for instance, to 12 would imply the double of the components (C33, C28, C46, C20, C45, C19, C4, C25, C18 and C1) to be monitored, making this activity more onerous in terms of person-hours.

When a failure on a component occurs, on the other hand, it is necessary to decide whether to perform a predictive intervention on the consequent ones. For this purpose, the ILP model presented in the previous section is used. For instance, let us consider the failure and the replacement of component C15 (this component presents the highest value of BC). According to the association rules mined and the consequent SN created, 40 components (the ones highlighted in yellow in Figure 3) are related to C15. Hence, it would not be realistic to replace all of them in advance.

Component	OD	Component	OD
C33	15.83	C43	10.67
C28	15.83	C29	10.67
C46	15.43	C14	10.67
C20	15.43	C15	10.61
C45	14.13	C40	10.44
C19	13.67	C30	10.29
C4	12.71	C26	10.28
C25	12.44	C32	10.13
C18	12.42	C42	10.00
C1	12.07	C39	10.00
C3	11.79	C41	9.75
C5	11.67	C31	9.67
C13	11.50	C6	9.40
C24	11.50	C35	9.25
C37	11.40	C16	9.22
C2	11.26	C10	9.00
C7	11.00	C12	8.80
C44	11.00	C27	8.08
C11	11.00	C17	7.00
C7	11.00	C38	6.00
C36	10.80	C23	6.00
C9	10.71	C22	6.00
C34	10.69	C21	3.00

Table 3.
Out degree values of the social network's nodes

Therefore, the BC value for all the components is calculated and given as the objective function of the ILP model, as well as the other data reported in Table 4, ranking them in descending BC. Such ranking allows us to visualize the most influential among the network, i.e. the ones that have a higher criticality, at the top of the table; while, as we descend along the table, the remaining components will gradually become less influential and, therefore, less troublesome.

4.1 Prioritization strategy and what-if scenarios

The parameters of the work are set in collaboration with the maintenance department of the topping sub-plant, considering that a participatory approach allows a larger view of the entire contest (Marinakis *et al.*, 2017). In addition, this decision enables the decision-makers to be consistent with their actual policies. Specifically, the maximum budget allowed for predictive maintenance of this plant ($\boldsymbol{B}^{max}$) is set to 3,000 € per day, while the maximum time ($\boldsymbol{T}^{max}$) is 350 min. In addition, a maximum of five operators ($\boldsymbol{R}^{max}$) can take part in predictive maintenance activities. Considering these parameters and the data provided in Table 4, after the failure of C15, the results obtained recommend the replacement of components C2, C25 and C18, obtaining a total BC value of 252.22 (see Experiment 1 in Table 5). As presented in Figure 4, the items C15, C2, C25 and C18 are closely connected to each other and are characterized by a wide number of ingoing and outgoing edges, making them critical in terms of influence among the network. The available time is saturated, as well as the number of operators employed in the operations. The budget needed to satisfy the requirement of such a solution, on the other hand, is lower than the $\boldsymbol{B}^{max}$ (2,109 € out of 3,000 €).

Hence, a sensitivity analysis is performed to understand whether, adjusting the parameters, a relevant improvement could be obtained. Indeed, budget, time and human resources allocation is a critical activity for decision-makers, especially in large organizations (Li *et al.*, 2019). Thus it is important to verify the impact of their decision and, possibly, adjust them.

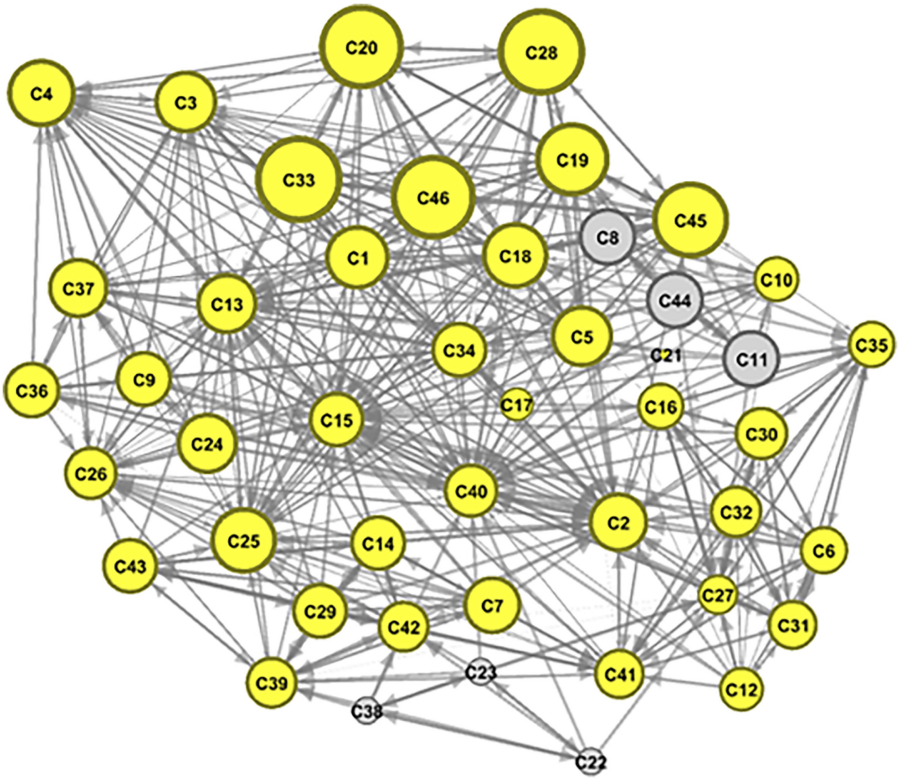

Figure 3.
Social Network representation highlighting C15's consequent components

Increasing the $\boldsymbol{R}^{max}$, without modifying the other parameters, has no impact on the solution found, as presented in Experiment 2 of Table 5; while, increasing the $\boldsymbol{T}^{max}$ by 25% – hence, extending it to 437.5 min – allows an improvement of the selected components: C2, C25, and C39 are selected for predictive maintenance (Experiment 3 in Table 5). Even in this case, the constraint on $\boldsymbol{R}^{max}$ is saturated, while budget and time ones are not. Leaving $\boldsymbol{T}^{max}$ unchanged while increasing $\boldsymbol{R}^{max}$ provides the same solution; hence, it is decided to furtherly increase both the $\boldsymbol{T}^{max}$ up to 525 min and the $\boldsymbol{R}^{max}$ up to 10. The solution provided by this scenario recommends the replacement of four components (C2, C25, C27 and C39) as reported in Table 5 (Experiment 4). Introducing a further increment on the budget, hence increasing the $\boldsymbol{B}^{max}$ by 25 and 50%, assures the selection of 5 (C2, C25, C4, C18 and C39) and 6 (C2, C25, C4, C18, C32 and C39) components, respectively. As shown in lines with Experiments 5 and 6 of Table 5, there is no relevant increment in terms of objective function: this is justified by the fact that the further selected component (C32) has a low value in BC (0.09). Hence, in this case, it may not be convenient to increase the budget so much.

To summarize the results reported in Table 5, it can be said that the current parameter setting allows the execution of some predictive replacements. Hence, it is acceptable. According to Experiments 1 and 2, the scarce resource is time: there is no need to increase budget availability and human resources if no adjustment is made in terms of time. This should be the first modification to be made should the company decide to make more investment in the asset maintenance perspective.

Component	c_j [€]	t_j [min]	R_j	BC
C40	5,931	600	2	192.79
C2	1,184	250	1	188
C13	2,300	750	1	104.53
C42	1,311	286	1	74.66
C39	1,274	175	1	74.66
C25	80	10	3	55.79
C41	235	299	1	54.17
C26	289	300	1	51.66
C5	2,881	223	1	32.58
C27	190	60	2	30.22
C3	4,094	255	1	20.66
C34	650	300	1	17.95
C4	1,009	120	2	13.88
C37	2,100	150	3	13.88
C35	1,627	495	1	9.1
C18	845	90	1	8.43
C19	2,103	66	1	7.81
C45	735	300	1	7.43
C1	3,074	146	1	6.88
C9	1,281	423	1	5.65
C36	3,288	248	1	5.13
C16	2,500	800	1	4.29
C6	1,010	206	1	1.75
C12	2,118	357	1	1.66
C43	2,950	386	1	0.18
C29	577	529	1	0.18
C14	207	299	1	0.18
C31	1,233	607	1	0.09
C32	402	68	1	0.09
C30	4,063	333	1	0.09
C46	2,930	329	1	0
C20	5,041	122	1	0
C33	2061	212	1	0
C28	2,302	340	2	0

Table 4. List of C15's consequent components and their associated repair cost (c_j), time (t_j), number of operators required (R_j) and Betweenness Centrality (BC_j)

	Selected components	BC	T^{max}	$\sum_j t_j x_j$	B^{max}	$\sum_j c_j x_j$	R^{max}	$\sum_j r_j x_j$
Experiment 1	C2, C25, C18	252.22	350	350	3,000	2,109	5	5
Experiment 2	C2, C25, C18	252.22	350	350	3,000	2,109	10	5
Experiment 3	C2, C25, C39	318.45	437.5	435	3,000	2,538	5	5
Experiment 4	C2, C25, C27, C39	348.67	525	495	3,000	2,728	10	7
Experiment 5	C2, C25, C4, C18, C39	362.55	700	615	3,750	3,737	10	9
Experiment 6	C2, C25, C4, C18, C32, C39	362.65	700	683	4,650	4,139	10	10

Table 5. Summary of the what-if scenarios

4.2 Results analysis

According to the ARs mined, several relationships among component failures have been identified. Such failures might not be the ones expected by the technicians of the plant, even though they have a large experience in the field. Indeed, one of the main theoretical

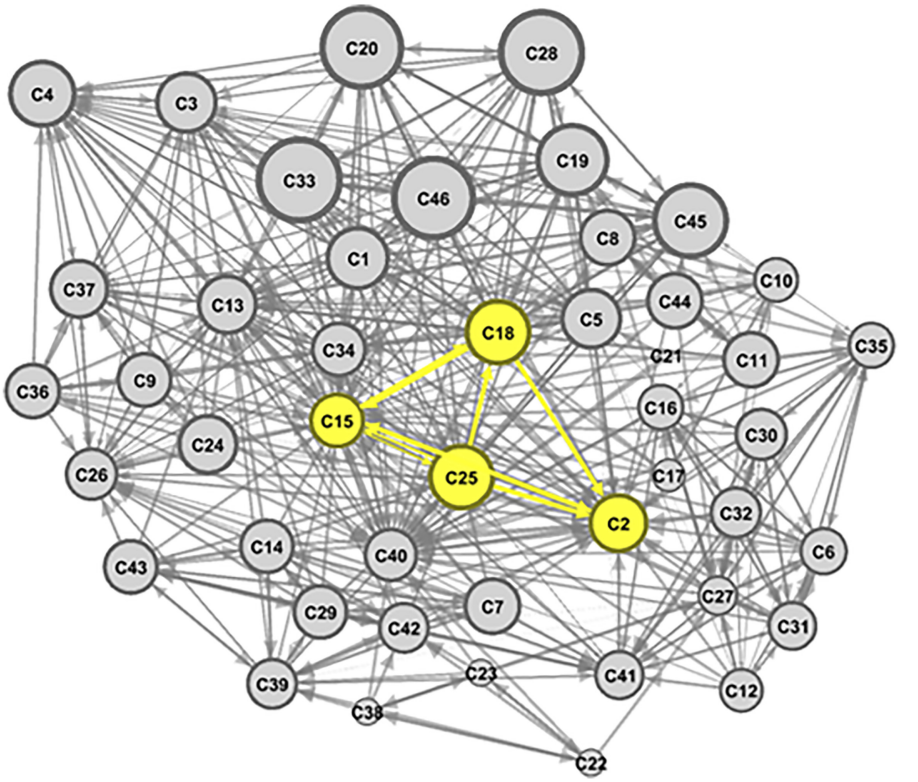

Figure 4.
Social Network representation highlighting the components selected in Experiment 1

contributions proposed in this work is that it is data-driven. Hence, the driver followed to define the components to be replaced is the information extracted by the data rather than the technical and physical structure of the process. For instance, rule C15 → C2 indicates that when a failure occurs on component C15, that is, a controller, the coupling C2 also requires a replacement within two weeks, quite likely since the confidence is 0.886. This means that the controller C15 and the coupling C2 are likely to be replaced concurrently in a two-week time interval. Similarly, the ARs C15 → C40 (confidence = 0.657) indicates that in more than 65% of cases after the failure of the controller C15, even the sealing device C40 has to be replaced. The explication of such relationships is evident from the data since the application of the proposed approach relies on a solid dataset and a relevant amount of data, which is fundamental to deploying a data-driven framework.

Considering the prioritization of the components to be replaced, results show that the rationale is similar to the mining of the association rules. Indeed, the objective function of the ILP model takes into account the influence of each component across the SNA since it aims at selecting those having the highest BC, respecting the constraints. Recalling the example proposed in Section 4, we can say that when the failure of the controller C15 occurs, the most critical successors according to their BC value would be C40, C2, C13, C42, C39 and C25, that are, respectively, the sealing device, the coupling, the insulation, the transmitting device, the measurement instrumentation and the lighting. In all the what-if scenarios tested, the coupling and the lighting systems are selected, the instrumentation is selected in Experiments 3–4–5–6, while the other ones are excluded to

respect the constraints imposed by the company policies in favor of components characterized by lower BC, as well as lower resource requirement. For instance, in the "case-base," i.e. Experiment 1, the drainer (C18) is replaced or maintained together with the controller (that is the one effectively experiencing the failure), the coupling (C2) and the lighting (C25). As previously stated, data provide the support for the execution of such interventions, even though there might not seem to be any actual relations, furtherly highlighting the benefits driven by the implementation of the approach. Indeed, the reliability of the plant is ensured through the adoption of the proposed framework since the domino effect among failures frequently occurring together is limited by anticipating the maintenance of critical components.

4.3 Discussion

As presented in previous sections, the aim of this work is to extract information from data and apply them to identify the relationships between component failures so that they can be avoided through predictive replacements. In this way, their impact on the process can be eliminated or, at least, limited. A data-driven approach is selected in this work that is based on a process industry case study. Indeed, in this sector, it is important to have complete control and monitoring policies due to the hazard related to the operations (Ciarapica *et al.*, 2019). The proposed asset maintenance framework does not entirely modify the current procedure in place in the case study: they should be considered as an addition to the present one. Therefore, they have to be used both for online and off-line asset maintenance activities to ensure the resilience of the system, i.e. the ability of a system to absorb and resist adverse occurrences. For example, Failure Modes Effects and Criticality Analysis (FMECA) should also be carried out, in order to identify the possible failure modes and prioritize them by the risk priority number. It is specifically useful to build a baseline of the potential failure modes and effects but also has some criticalities that can be overcome by the introduction of specific predictive maintenance policies. Indeed, as pointed out by Ahmed *et al.* (2021), FMECA does not take into account coexisting failures and mutual relationships among them and it is also reasonably vulnerable to human error.

The implementation of quality management systems (e.g. ISO 9001) represents an opportunity to achieve benefits in terms of business process optimization and advantages like rationalization and cost reduction. In relation to the maintenance aspects, thanks to quality management systems, it is possible to reduce internal errors, resulting in less waste and more production efficiency. Considering the refinery sector in which the proposed case study is situated, fewer errors are related to lower risk and higher safety levels. It is worth noticing that component failures are somewhat intrinsic in the process, so they cannot be totally eliminated by the introduction of data-driven maintenance policies. However, they can be sensibly reduced and, thanks to the joint adoption of quality management systems and maintenance policies, they can be addressed and efficiently treated in a formal manner (Maletič *et al.*, 2014). As previously mentioned, the case study is proposed basing on an oil refinery plant. However, the same problem can be addressed in the process industry and also in massive production lines. Indeed, the specific parameters can be adjusted accordingly, and total monitoring of the plant can be performed.

5. Contribution and implications

The proposed framework aims at helping maintenance managers come to better, more informed decisions in the day-to-day business practices in order to maximize availability, minimize failures and optimize costs of asset maintenance. From this framework, both theoretical contributions and managerial implications can be extracted.

5.1 Theoretical contribution

The theoretical contribution provided in this work is essential of a twofold nature: in the problem addressed and in the methodology used.

The problem addressed regards a research gap in the literature: the prediction of the domino effect between component failures. It is important to underline the importance of using the proposed framework in all companies where there is this domino effect between failures or malfunctions of components. The results obtained have shown that this phenomenon often occurs in the analyzed plant. It is easy to predict that this behavior is present in many process industries, where the various components (pumps, valves, pipes, tanks, . . .) are physically connected to each other.

In addition, this work adopts a data-driven perspective. Hence, the decision-maker implementing such a framework on the process industry relies on the information extracted by the data rather than on the technical and physical structure of the system – that is, instead, the rationale followed by the model-driven paradigm. This vision expands the body of knowledge of the plant technicians by integrating it with the insights derived from the data analytics.

From a methodological point of view, different techniques have been combined in the proposed framework providing complementary contributions from a theoretical point of view. In particular, the ARM method provided researchers with tools to solve the problems related to the use of statistical analysis like the elevated number of variables, the independence assumptions and the distribution of collected data. The intrinsic organization and complexity of the data collected might jeopardize the use of traditional tools for analysis since the variables showed some critical features. The method based on ARs offers many readable patterns (rules) explaining the interactions between two or more variables. In addition, it eliminates the need to formulate a research hypothesis for each failure event before doing a formal evaluation that may become practically infeasible even for a moderately sized set of variables (Antomarioni *et al.*, 2021).

The key contribution of the SNA concerns the possibility of identifying nodes communities in the network created through the ARs and defining the nature of such connections. Furthermore, it allows asset managers to verify the existence of missing or false links in the network, eliminating errors in the data collection process that would have been unnoticed. In existing literary contributions, as shown in Section 2, only Kim *et al.* (2019) proposed the implementation of an SNA for the synchronous replacement of components. The main difference with their work resides both in the application area and in the definition of the relations among components: indeed, the application area is the construction industry, while the relationships among components to be replaced are model-driven. The approach proposed in this work, on the other hand, is data-driven since the relationships among failures are derived from the records of previous breakages. Considering other existing contributions, as shown in the literature review, one can mention the study by Antomarioni *et al.* (2019), where the optimal selection of the components to be repaired in an oil refinery is proposed. The present application goes beyond the approach proposed in the other work in which we consider not only the probability of failure of individual components but also the connections between different components.

5.2 Implications for practitioners

The proposed framework – developing tools for monitoring critical components and predicting fault events – can help different refinery departments, as well as other process industries. The proposed data-driven decision support system enables asset managers to turn predictive analytics insight into prescriptive analytics action by converting information on what is likely to happen in maintenance activities, transforming the raw data into useful

and applicable knowledge. In particular, the framework aims to be useful for the maintenance planner, who needs to decide when to maintain each asset, what tasks need to be done and which parts need to be replaced at each maintenance interval in order to meet reliability targets at an optimal cost. The combined use of ARM and SNA highlights the domino effect among events, with both a visual perspective of the network and the relations existing among the components being determined through a data-driven technique.

Moreover, the integration of ILP helps the maintenance planner to schedule maintenance activities. It is valid to support the definition of the components to be prioritized for maintenance, taking into account the resource constraints (e.g. time, budget and number of employees) actually existing in the company.

These tools are also important for the parts planner, who needs to decide how many of each of the spare parts are needed in which locations and when so that they can maximize first-time fix rates and reduce spare parts acquisition and holding costs.

Finally, the maintenance technicians of the refinery, who need to determine the root cause of failures, decide on the best fix and determine whether an asset should be repaired or replaced, in order to minimize turn time, reduce repair cost and eliminate rework. These decisions must be made for each asset, although each asset has a unique configuration, history, usage, environment, conditions and parameters, which begins with the commissioning and start-up steps. In this context, the importance of data analytics tools to determine the best decision option and action plan for each asset becomes evident. Indeed, the proposed framework aims at integrating the analysis of large amounts of data in everyday processes in order to support real-time decision-making. Decisions in real-time that drive efficient maintenance operations, increase equipment reliability, uptime, safety and reduce overall costs.

6. Conclusions

This article provides a data-driven system based on a combination of ARs and directed weighted SN to identify and analyze the relationships among component maintenance activities and predict the domino effect among component failures. Through the metrics provided by the SNA and ILP, the decision-making process for the selection of the component to predictively maintain is supported. The reliability of the asset is addressed in three ways: (1) having better control of the critical components, enabling more rapid interventions on faults; (2) anticipating the substitution of probable failing components to avoid further interruptions in the production flow in the future, with a positive impact on the availability of the plant and on the reduction of downtimes; and (3) the root causes of failure chains can be identified after the introduction of the proposed methodologies, due to possibility of visualizing the relationships between failures.

The main advantages resulting in a data analytics approach to asset maintenance can be summarized as follows: (1) the data management process is clearly defined through the layered framework presenting the roadmap from the collection of Big Data, ETL process, integration and analysis and, finally, decision-making; (2) remarkable cause–effect relationships in the refinery processes can be identified during the asset maintenance activities; and (3) the bias deriving from basing the technical management of the refinery only on the technicians' knowledge is overcome through the implementation of objective decision-making tools.

The idea of using ARM and SNA for asset maintenance is should not be seen as an attempt to replace the traditional maintenance procedure but as a complementary method to be integrated into these types of activities: indeed, the operations performance should benefit from the innovation performance (Hong *et al.*, 2019).

The main limitation of this work regards the data collection phase. Indeed, in general, there might be a lag between the occurrence of the event and the recording of its consequence, complicating the understanding of the relationship between the event and the maintenance activities. Furthermore, the refinery piping systems consist of kilometers of pipes operating at different conditions, mostly underground and therefore real-time monitoring of their condition is rather costly. For this reason, there is sometimes not a complete panorama of their data.

Further development of this work may regard the extension of the testing case to the whole refinery plant, aggregating all data, as well as the integration of the current approach with multi-objective optimization models or with multi-criteria decision-making approaches, taking into account other risk categories and decision criteria.

References

Abbassi, R., Bhandari, J., Khan, F., Garaniya, V. and Chai, S. (2016), "Developing a quantitative risk-based methodology for maintenance scheduling using Bayesian Network", *Chemical Engineering Transactions*, Vol. 48, pp. 235-240.

Agrawal, R., Imieliński, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", in *Acm Sigmod Record*, ACM, New York, NY, Vol. 22 No. 2, pp. 207-216.

Ahmed, U., Carpitella, S. and Certa, A. (2021), "An integrated methodological approach for optimising complex systems subjected to predictive maintenance", *Reliability Engineering and System Safety*, Vol. 216, 108022.

Antomarioni, S., Pisacane, O., Potena, D., Bevilacqua, M., Ciarapica, F.E. and Diamantini, C. (2019), "A predictive association rule-based maintenance policy to minimize the probability of breakages: application to an oil refinery", *The International Journal of Advanced Manufacturing Technology*, Vol. 105, pp. 3661-3675.

Antomarioni, S., Lucantoni, L., Ciarapica, F.E. and Bevilacqua, M. (2021), "Data-driven decision support system for managing item allocation in an ASRS: a framework development and a case study", *Expert Systems with Applications*, Vol. 185, 115622.

Bahga, A. and Madisetti, V.K. (2011), "Analyzing massive machine maintenance data in a computing cloud", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 23 No. 10, pp. 1831-1843.

Bangalore, P. and Tjernberg, L.B. (2015), "An artificial neural network approach for early fault detection of gearbox bearings", *IEEE Transactions on Smart Grid*, Vol. 6 No. 2, pp. 980-987, doi: 10.1109/TSG.2014.2386305.

Baptista, M., de Medeiros, I.P., Malere, J.P., Prendinger, H., Nascimento Jr, C.L. and Henriques, E. (2016), "Improved time-based maintenance in aeronautics with regressive support vector machines", *Annual Conference of the Prognostics and Health Management Society*.

Baptista, M., Sankararaman, S., de Medeiros, I.P., Nascimento Jr, C., Prendinger, H. and Henriques, E.M. (2018), "Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling", *Computers and Industrial Engineering*, Vol. 115, pp. 41-53.

Bennane, A. and Yacout, S. (2012), "LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance", *Journal of Intelligent Manufacturing*, Vol. 23 No. 2, pp. 265-275.

Bevilacqua, M. and Ciarapica, F.E. (2018), "Human factor risk management in the process industry: a case study", *Reliability Engineering and System Safety*, Vol. 169, pp. 149-159.

Bhattacharjee, P., Dey, V. and Mandal, U.K. (2020), "Risk assessment by failure mode and effects analysis (FMEA) using an interval number based logistic regression model", *Safety Science*, Vol. 132, 104967.

Brandes, U. (2001), "A faster algorithm for betweenness centrality", *Journal of Mathematical Sociology*, Vol. 25 No. 2, pp. 163-177.

BS EN: 13306 (2017), *Maintenance Terminology*, BSI Standards Publication.

Bubbico, R., Greco, V. and Menale, C. (2018), "Hazardous scenarios identification for Li-ion secondary batteries", *Safety Science*, Vol. 108, pp. 72-88.

Buddhakulsomsiri, J., Siradeghyan, Y., Zakarian, A. and Li, X. (2006), "Association rule-generation algorithm for mining automotive warranty data", *International Journal of Production Research*, Vol. 44 No. 14, pp. 2749-2770.

Bumblauskas, D., Gemmill, D., Igoua, A. and Anzengruber, J. (2017), "Smart Maintenance Decision Support Systems (SMDSS) based on corporate big data analytics", *Expert Systems with Applications*, Vol. 90 No. 2017, pp. 303-317.

Chen, Z., Li, Y., Xia, T. and Pan, E. (2019), "Hidden Markov model with auto-correlated observations for remaining useful life prediction and optimal maintenance policy", *Reliability Engineering and System Safety*, Vol. 184, pp. 123-136.

Chuang, S.Y., Sahoo, N., Lin, H.W. and Chang, Y.H. (2019), "Predictive maintenance with sensor data analytics on a Raspberry Pi-based experimental platform", *Sensors*, Vol. 19 No. 18, p. 3884.

Ciarapica, F., Bevilacqua, M. and Antomarioni, S. (2019), "An approach based on association rules and social network analysis for managing environmental risk: a case study from a process industry", *Process Safety and Environmental Protection*, Vol. 128, pp. 50-64.

Crespo Márquez, A., de la Fuente Carmona, A. and Antomarioni, S. (2019), "A process to implement an artificial neural network and association rules techniques to improve asset performance and energy efficiency", *Energies*, Vol. 12 No. 18, p. 3454.

Cunha, C.D., Agard, B. and Kusiak, A. (2006), "Data mining for improvement of product quality", *International Journal of Production Research*, Vol. 44 Nos 18-19, pp. 4027-4041, doi: 10.1080/00207540600678904.

Datong, L., Yu, P. and Xiyuan, P. (2009), "Fault prediction based on time series with online combined kernel SVR methods", *2009 IEEE Instrumentation and Measurement Technology Conference*, IEEE, Singapore, pp. 1163-1166.

Ergu, D., Kou, G., Peng, Y., Shi, Y. and Shi, Y. (2013), "The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment", *The Journal of Supercomputing*, Vol. 64 No. 3, pp. 835-848.

Gerum, P.C.L., Altay, A. and Baykal-Gürsoy, M. (2019), "Data-driven predictive maintenance scheduling policies for railways", *Transportation Research Part C: Emerging Technologies*, Vol. 107, pp. 137-154.

Gharoun, H., Keramati, A., Nasiri, M.M. and Azadeh, A. (2019), "An integrated approach for aircraft turbofan engine fault detection based on data mining techniques", *Expert Systems*, Vol. 36 No. 2, e12370.

Gupta, S., Bhattacharya, J., Barabady, J. and Kumar, U. (2013), "Cost-effective importance measure: a new approach for resource prioritization in a production plant", *International Journal of Quality and Reliability Management*, Vol. 30 No. 4, pp. 379-386, doi: 10.1108/02656711311308376.

Han, J., Cheng, H., Xin, D. and Yan, X. (2007), "Frequent pattern mining: current status and future directions", *Data Mining and Knowledge Discovery*, Vol. 15 No. 1, pp. 55-86.

Hong, J., Liao, Y., Zhang, Y. and Yu, Z. (2019), "The effect of supply chain quality management practices and capabilities on operational and innovation performance: evidence from Chinese manufacturers", *International Journal of Production Economics*, Vol. 212, pp. 227-235.

Ishizaka, A. and Nemery, P. (2014), "Assigning machines to incomparable maintenance strategies with ELECTRE-SORT", *Omega*, Vol. 47, pp. 45-59.

Izquierdo, J., Márquez, A.C. and Uribetxebarria, J. (2019), "Dynamic artificial neural network-based reliability considering operational context of assets", *Reliability Engineering and System Safety*, Vol. 188, pp. 483-493.

Kankar, P.K., Sharma, S.C. and Harsha, S.P. (2011), "Fault diagnosis of ball bearings using machine learning methods", *Expert Systems with Applications*, Vol. 38 No. 3, pp. 1876-1886.

Kaparthi, S. and Bumblauskas, D. (2020), "Designing predictive maintenance systems using decision tree-based machine learning techniques", *International Journal of Quality and Reliability Management*, Vol. 37 No. 4, pp. 659-686, doi: 10.1108/IJQRM-04-2019-0131.

Kim, J., Han, S. and Hyun, C.-T. (2019), "Identification and reduction of synchronous replacements in life-cycle cost analysis of equipment", *Journal of Management in Engineering*, Vol. 35 No. 1, 04018058, doi: 10.1061/(ASCE)ME.1943-5479.0000673.

Knoke, D. and Yang, S. (2008), *Social Network Analysis*, SAGE, Thousand Oaks, CA.

Komljenovic, D., Gaha, M., Abdul-Nour, G., Langheit, C. and Bourgeois, M. (2016), "Risks of extreme and rare events in asset management", *Safety Science*, Vol. 88, pp. 129-145.

Kumar, A., Chinnam, R.B. and Tseng, F. (2019), "An HMM and polynomial regression based approach for remaining useful life and health state estimation of cutting tools", *Computers and Industrial Engineering*, Vol. 128, pp. 1008-1014.

Kusiak, A. and Verma, A. (2012), "Analyzing bearing faults in wind turbines: a data-mining approach", *Renewable Energy*, Vol. 48, pp. 110-116.

Langone, R., Alzate, C., De Ketelaere, B., Vlasselaer, J., Meert, W. and Suykens, J.A.K. (2015), "LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines", *Engineering Applications of Artificial Intelligence*, Vol. 37, pp. 268-278, doi: 10.1016/j.engappai.2014.09.008.

Létourneau, S., Famili, F. and Matwin, S. (1999), "Data mining to predict aircraft component replacement", *IEEE Intelligent Systems and Their Applications*, Vol. 14 No. 6, pp. 59-66.

Li, D., Wang, W. and Ismail, F. (2013), "Enhanced fuzzy-filtered neural networks for material fatigue prognosis", *Applied Soft Computing*, Vol. 13 No. 1, pp. 283-291.

Li, F., Zhu, Q. and Chen, Z. (2019), "Allocating a fixed cost across the decision making units with two-stage network structures", *Omega*, Vol. 83, pp. 139-154.

Lin, C.C. and Tseng, H.Y. (2005), "A neural network application for reliability modelling and condition-based predictive maintenance", *The International Journal of Advanced Manufacturing Technology*, Vol. 25 Nos 1-2, pp. 174-179.

Liu, Y., Hu, X. and Zhang, W. (2019), "Remaining useful life prediction based on health index similarity", *Reliability Engineering and System Safety*, Vol. 185, pp. 502-510.

Madu, C.N. (2000), "Competing through maintenance strategies", *International Journal of Quality and Reliability Management*, Vol. 17 No. 9, pp. 937-949, doi: 10.1108/02656710010378752.

Maletič, D., Maletič, M. and Gomišček, B. (2014), "The impact of quality management orientation on maintenance performance", *International Journal of Production Research*, Vol. 52 No. 6, pp. 1744-1754.

Marinakis, V., Doukas, H., Xidonas, P. and Zopounidis, C. (2017), "Multicriteria decision support in local energy planning: an evaluation of alternative scenarios for the sustainable energy action plan", *Omega*, Vol. 69, pp. 1-16.

Marquez, A.C. and Gupta, J.N. (2006), "Contemporary maintenance management: process, framework and supporting pillars", *Omega*, Vol. 34 No. 3, pp. 313-326.

Mazhar, M.I., Kara, S. and Kaebernick, H. (2007), "Remaining life estimation of used components in consumer products: life cycle data analysis by Weibull and artificial neural networks", *Journal of Operations Management*, Vol. 25 No. 6, pp. 1184-1193.

Medjaher, K., Tobon-Mejia, D.A. and Zerhouni, N. (2012), "Remaining useful life estimation of critical components with application to bearings", *IEEE Transactions on Reliability*, Vol. 61 No. 2, pp. 292-302.

Otte, E. and Rousseau, R. (2002), "Social network analysis: a powerful strategy, also for the information sciences", *Journal of Information Science*, Vol. 28 No. 6, pp. 441-453.

Purarjomandlangrudi, A., Ghapanchi, A.H. and Esmalifalak, M. (2014), "A data mining approach for fault diagnosis: an application of anomaly detection algorithm", *Measurement*, Vol. 55, pp. 343-352.

Romanowski, C.J. and Nagi, R. (2001), "Analyzing maintenance data using data mining methods", in *Data Mining for Design and Manufacturing*, Springer, Boston, MA, pp. 235-254.

Salahshoor, K., Kordestani, M. and Khoshro, M.S. (2010), "Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers", *Energy*, Vol. 35 No. 12, pp. 5472-5482.

Saravanan, N., Siddabattuni, V.K. and Ramachandran, K.I. (2010), "Fault diagnosis of spur bevel gear box using artificial neural network (ANN), and proximal support vector machine (PSVM)", *Applied Soft Computing*, Vol. 10 No. 1, pp. 344-360.

Sattari, F., Macciotta, R., Kurian, D. and Lefsrud, L. (2021), "Application of Bayesian network and artificial intelligence to reduce accident/incident rates in oil and gas companies", *Safety Science*, Vol. 133, 104981.

Schlechtingen, M. and Santos, I.F. (2011), "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection", *Mechanical Systems and Signal Processing*, Vol. 25 No. 5, pp. 1849-1875.

Scott, J. and Carrington, P.J. (2011), *The SAGE Handbook of Social Network Analysis*, SAGE, London.

Sharma, S., Cui, Y., He, Q., Mohammadi, R. and Li, Z. (2018), "Data-driven optimization of railway maintenance for track geometry", *Transportation Research Part C: Emerging Technologies*, Vol. 90, pp. 34-58.

Tang, K., Parsons, D.J. and Jude, S. (2019), "Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system", *Reliability Engineering and System Safety*, Vol. 186, pp. 24-36.

Van Horenbeek, A. and Pintelon, L. (2014), "Development of a maintenance performance measurement framework—using the analytic network process (ANP) for maintenance performance indicator selection", *Omega*, Vol. 42 No. 1, pp. 33-46.

Wang, D. (2016), "K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: revisited", *Mechanical Systems and Signal Processing*, Vol. 70, pp. 201-208.

Wu, J., Hu, K., Cheng, Y., Zhu, H., Shao, X. and Wang, Y. (2019), "Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term memory neural network", *ISA Transactions*.

Yam, R.C.M., Tse, P.W., Li, L. and Tu, P. (2001), "Intelligent predictive decision support system for condition-based maintenance", *The International Journal of Advanced Manufacturing Technology*, Vol. 17 No. 5, pp. 383-391.

Further reading

Bevilacqua, M., Ciarapica, F.E., Diamantini, C. and Potena, D. (2017), "Big data analytics methodologies applied at energy management in industrial sector: a case study", *International Journal of RF Technologies*, Vol. 8 No. 3, pp. 105-122.

Goebel, K., Saxena, A., Saha, S., Saha, B., Celaya, J., Srivastava, A.N. and Han, J. (2011), "Prognostic performance metrics", in *Machine Learning and Knowledge Discovery for Engineering Systems Health Management*, Vol. 147.

Hand, D.J. (1998), "Data mining: statistics and more?", *The American Statistician*, Vol. 52 No. 2, pp. 112-118.

Peng, Y., Kou, G., Shi, Y. and Chen, Z. (2008), "A descriptive framework for the field of data mining and knowledge discovery", *International Journal of Information Technology and Decision Making*, Vol. 7 No. 04, pp. 639-682.

Appendix

Table A1.
List of the components' ID and the corresponding name

ID	Component name	ID	Component name
C1	Undefined component	C24	Joint
C2	Coupling	C25	Lighting
C3	Alarm	C26	Indicator
C4	Ammeter	C27	Liquid-level indicator
C5	Area	C28	Level switch
C6	Auxiliary	C29	Lubrication
C7	Shovel	C30	Engine
C8	Keg	C31	Shovels
C9	Battery	C32	Oil Seal
C10	Burner	C33	Flooring
C11	Bypass	C34	Sampling valve
C12	Strap	C35	Button panel
C13	Insulation	C36	Refrigerant
C14	Condensation indicator	C37	Detector
C15	Controller	C38	Blower
C16	Bearing	C39	Instrumentation
C17	Caliber disc	C40	Sealing device
C18	Drainer	C41	Tracing
C19	Ecos	C42	Transmitting device
C20	Electrode	C43	Overfilling indicator
C21	Tube bundle	C44	Pipeline
C22	Filter	C45	Valve
C23	Fittings	C46	Shifter

Corresponding author
Sara Antomarioni can be contacted at: s.antomarioni@univpm.it