PhD-FSTM-2021-046
The Faculty of Sciences, Technology and Medicine

DISSERTATION

Defence held on 08/07/2021 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN Sciences de l'Ingénieur

by

Yamila Mariel OMAR

Born on 27 July 1988 in Mar del Plata, Buenos Aires (Argentina)

COMPLEX NETWORK IN MANUFACTURING SUITABILITY AND INTERPRETATION

Dissertation defence committee

Dr Peter Plapper, dissertation supervisor
Professor, Université du Luxembourg

Dr Stephan Leyer, Chairman
Professor, Université du Luxembourg

Dr Thomas Engel, Vice Chairman
Professor, Université du Luxembourg

Dr Martin Atzmüller
Professor, Universität Osnabrück

Dr Till Becker
Professor, Hochschule Emden-Leer

Affidavit

I hereby confirm that the PhD thesis entitled "Complex Networks in Manufacturing - Suitability and Interpretation" has been written independently and without any other sources than cited.

Luxembourg, 7th July 2021

Yamila Mariel Omar

"If your dreams do not scare you, they are not big enough."

Ellen Johnson Sirleaf
First elected female head of state in Africa

University of Luxembourg

Abstract

Faculty of Science, Technology and Medicine
Department of Engineering

Complex Networks in Manufacturing – Suitability and Interpretation

by Yamila Mariel OMAR

The fourth industrial revolution, and the associated digitization of the manufacturing industry, has resulted in increased data generation. Industry leaders aim to leverage this data to enhance productivity, boost innovation and generate new manners of competition. In this work, out of the many domains within the manufacturing sector, production will be explored. To this end, the mathematical tools of network science are utilized to characterize and evaluate production networks in terms of complex networks.

In a manufacturing complex network, nodes represent workstations, and directed edges abstract the material flow that occurs among pairs of workstations. These types of complex networks are known as "material flow networks" and are used to study issues associated with manufacturing systems in the domain of production at the intra-enterprise level. While some research on the subject exists, this work will demonstrate that the use of complex networks to describe and evaluate manufacturing systems constitutes a nascent research field. In fact, the limited existing literature tackles a vast number of issues raising more questions than providing answers.

This work aims to answer a number of those open questions. Firstly, which complex network metrics are suitable in the context of manufacturing networks will be determined. As a consequence, unsuitable metrics will be identified as well. To accomplish this, the flow underlying assumptions of popular complex network metrics is studied and compared to those of manufacturing networks. Furthermore, other existing complex network metrics with more appropriate underlying assumptions, but not yet explored in the context of manufacturing, are proposed and evaluated. Then, the appropriate interpretation of suitable complex network metrics in terms of Operations Research is provided. Finally, shortcomings of these metrics are highlighted to caution practitioners regarding their use in industrial settings.

Acknowledgements

I would like to start by thanking Prof. Dr. Peter Plapper for offering me the opportunity to pursue my PhD in the University of Luxembourg. At the time, I had international research experience in Materials Science & Engineering. Many doubted my ability to pursue research in Data Science, yet Prof. Plapper did not show any hesitation. I will always be grateful for his vote of confidence.

I also wish to thank Prof. Dr. Stephan Leyer and Prof. Dr. Thomas Engel. Their input as CET members was instrumental to guide my research and shaped the work presented hereafter.

I'd like to extend my gratitude to the external jury members, Prof. Dr. Till Becker and Prof. Dr. Martin Atzmüller for reviewing and evaluating my work. It is an honor to share my findings with you.

I'm extremely grateful to the Luxembourg National Research Fund for funding my PhD through an Industrial Fellowship since 2017. *Villmools Merci!* Yet, an Industrial Fellowship would have never been possible without Husky Injection Molding Systems, my partnering company. To Mr. J. C. Witz, Mr. Christophe Keller and the entire team, thank you!

Special thanks to all my colleagues, past and present. Sharing this experience with you made the long way less lonely. Thank you for the scientific discussions, the moms-can-conquer-it-all elevator speeches, and the endless words of kindness. I cannot imagine the past five years without you.

I cannot leave the University of Luxembourg without mentioning my Bachelor and Master students. Teaching you, whether in the classroom or during your Thesis or study projects, proved more instructive to me that you can ever imagine. You might have learned the curriculum, but you taught me how to teach better and, in this process, strengthened my empathy, compassion and patience.

In closing, all this work would not have been possible without the support of my husband, Carlo, and that of our son, Elia. Their encouragement kept me going, even when this quest seemed impossible.

Contents

List of Figures

List of Tables

List of Abbreviations

APICS	American Production and Inventory Control Society
BA	Barabási-Albert model
BDN	Binary directed network
BUN	Binary undirected network
CIRP	*College International pour la Recherche en Productique*
DC-SBM	Degree corrected stochastic block model
EDA	Exploratory Data Analysis
EU-27	European Union - 27 member countries
EUR	Euro (currency)
FAHP	Fuzzy Analytic Hierarchy Process
OR	Operations Research
PR	PageRank
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
SBM	Stochastic block model
SCC	Strongly connected component
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
TT	Total manufacturing time
VAT	Value adding time
VSM	Value stream mapping
WDN	Weighted directed network
WIP	Work in progress
WUN	Weighted undirected network

List of Symbols

$\mathcal{A}$	Adjacency matrix of elements a_{ij}
a_{ij}	Element of the adjacency matrix $\mathcal{A}$
B	Base set for product family identification
C	Family count in product family identification
C_i	Capacity of node v_i
$C_B(v_i)$	Betweenness centrality of node v_i
$C_C(G)$	Average clustering coefficient for graph G (for BUN)
$C_C(v_i)$	Clustering coefficient of node v_i (for BUN)
$C_C^D(G)$	Average clustering coefficient for graph G (for BDN)
$C_C^D(v_i)$	Clustering coefficient of node v_i (for BDN)
$\tilde{C}_C^D(v_i)$	Clustering coefficient of node v_i (for WDN)
$\tilde{C}_C^D(G)$	Average clustering coefficient for graph G (for WDN)
$C_H(v_i)$	Path-transfer flow entropy centrality of node v_i
$c(v_i, v_j)$	Capacity of the edge (v_i, v_j)
$c_f(v_i, v_j)$	Residual capacity of the edge (v_i, v_j)
$D(v_i)$	Downstream degree of node v_i
E	Set of edges
$\mathbf{e}$	Vector indicating nodes where random walkers can jump to in PageRank
F	Graph corresponding to product families. Typically, F_1, F_2 and F_3
F_s	Family set
$f(v_i, v_j)$	Flow in the edge (v_i, v_j)
G	Graph, also denoted $G = (V, E)$ or $G(V, E)$
H	Shannon information content or entropy
i	Node v_i, $i \equiv v_i$
J	Jaccard similarity of sets
J_{th}	Threshold for the Jaccard similarity of sets
$K(i, j)$	Number of paths from node v_i to node v_j
k_i	Degree of node v_i
k_i^{in}	In-degree of node v_i
k_i^{out}	Out-degree of node v_i
$k_i^{\leftrightarrow}$	Number of bilateral edges between v_i and its neighbors
M	Transition matrix of elements m_{ij}
m_{ij}	Element of the transition matrix M
N	Number of nodes in G. $N = \|V\|$.
P	Path
P_s	Path set
p_{ij}	Probability that flow starting on v_i end in v_j
p_{x_i}	Probability of an outcome in R
R	Set of possible outcomes (random variable)
S	Subgraph in G
s	Source node
s_i	Strength of node v_i

s_i^{in}	In-strength of node v_i
s_i^{out}	Out-strength of node v_i
t	Sink node
T_i	Total number of triangles that v_i could form (undirected graph)
T_i^D	Total number of triangles that v_i could form (directed graph)
t_i	Number of triangles formed by v_i (undirected graph)
t_i^D	Number of triangles formed by v_i (directed graph)
t_m	Total number of manufactured items
U	List of non-family path sets
V	Set of nodes, $V \neq \emptyset$ and $V = \{v_0, v_1, ..., v_N\}$
$\mathbf{v}$	Probability distribution vector in PageRank formulation
v_i	Node i, $v_i \equiv i$
$\mathcal{W}$	Weight matrix of elements w_{ij}
w_{ij}	Element of the weight matrix $\mathcal{W}$
X	Path set. X or X_i are used interchangeably.
β	Probability that a random walker follows an outgoing link in PageRank
$\Gamma(v_i)$	Set of neighboring nodes of node v_i
$\sigma(s,t)$	Number of geodesic paths between nodes s and t
$\sigma(s,t\|v_i)$	Number of geodesic paths between nodes s and t passing through v_i
$\sigma_k(v_i)$	Probability that the flow stops on node v_i
$\tau_k(v_i)$	Probability that the flow is transferred from node v_i

1 Introduction

In 2017, almost 1 in 10 enterprises (8.8%) in the EU-27's non-financial business economy were classified to "manufacturing" according to Eurostat [1]. As shown in Fig. 1.1, this represents almost 2 million enterprises corresponding to a vast range of activities. It includes small-scale businesses where manufacturing follows traditional production techniques, as well as very large enterprises that require parts and components from multiple suppliers to collectively manufacture a complex product. Despite this heterogeneity, this sector is the second largest employment contributor to the non-financial business economy, employing 28.5 million people in 2017. In addition, it is the largest contributor to the value added generated by the non-financial business economy, which amounted to EUR 1820 billion in 2017 only.

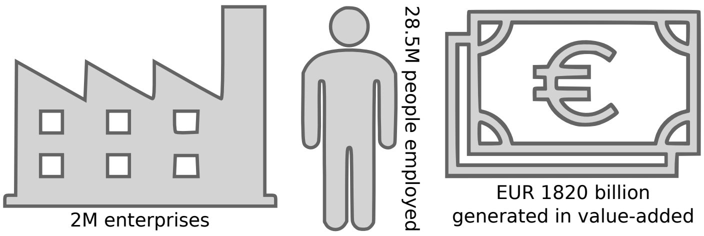


FIGURE 1.1: Infographic on the 2017 Eurostat figures for the manufacturing sector.

In contrast with these optimistic figures, the gross operating rate of the manufacturing sector was 9.8% in 2017, below the non-financial business economy average. This means that this sector had the second lowest level of profitability within the non-financial business economy for the EU-27 [1]. In fact, this underwhelming performance has been reported previously by Omar *et. al.* [2] citing Eurostat figures from 2013, indicating that the industry has been under pressure to reduce costs and increase margins while competing with developing economies for a long time. Thus, with the objective to increase profitability, manufacturing enterprises are evolving in a process termed the "Fourth Industrial Revolution" and popularized as "Industry 4.0" [3–9].

1.1 Industry 4.0

"*Industrie 4.0*" or "Industry 4.0" was a 2011 German initiative [3–9]. In Luxembourg, it was termed "The Third Industrial Revolution" according to Jeremy Rifkin's homonymous book [10], which later informed policy and strategy for the Luxembourgish Ministry of Economy [11]. In the Anglo-Saxon world similar concepts are used, such as as the "Industrial Internet" [12–15] and "Smart Manufacturing" [16–18]. All these different labels are characterized by partially overlapping concepts and terminology.

Although there is a growing body of research in the subject, the definition of "Industry 4.0" has not been agreed upon and thus, remains somehow dynamic. This is justified, partly, because the phenomenon is in its infancy and subject to a rapidly evolving technological landscape. Yet, Culot *et. al.* [9] explained that there is agreement in a number of aspects:

- Industry 4.0 is enabled by key technologies associated to increased digitalization and connectivity such as physical-digital interfaces, networks and data-processing technologies. Cyber-physical systems, the Internet of Things and Cloud Computing are often listed as Industry 4.0 enablers.
- Industry 4.0 has a number of characteristics consistent with the aforementioned key enabling technologies. These are virtualization, real-time information sharing and autonomy. Typically, intra- and inter-enterprise process integration is expected as well, making interoperability a distinctive characteristic and cyber-security an absolute need.
- Industry 4.0 is expected to increased profitability as a result of higher productivity and flexibility. Mass customization and new business models arise as a result.

Both, the enabling technologies and the characteristics of Industry 4.0, point to a differentiating factor with respect to previous industrial revolutions. Industry 4.0 enterprises are expected to generate massive amounts of data and utilize it to fulfill their expectations. As a result, scientific output increasingly focuses on data analytics algorithms and methods for the manufacturing sector.

1.2 Data in Manufacturing

From all the aspects surrounding Industry 4.0, the massive amount of generated data in manufacturing is a key ingredient for this work. Omar *et. al.* [2] explained that the use of manufacturing data to derive insights can

> ... enhance productivity [19] and competitiveness, boost innovation [20] and growth as well as generate new manners of competition [21] and value capture [22] across organizations. [It] contributes to an organization's agility by providing timely and accurate information [23]. In addition, the prevalent use of data ensures transparency, aids the discovery of market needs [24], uncovers process or service variability, improves performance [19] and assists in the adoption of more sustainable practices [25].

Furthermore, Omar *et. al.* identified five manufacturing domains in which data analytics can be a differentiating factor [2]. From these five domains, listed in Fig. 1.2, this work centers on the manufacturing data generated by ubiquitous cyber-physical systems in "production". This should be of particular interest for industry practitioners for several reasons. Firstly, they must leverage the data to guarantee the necessary return on investments to justify the acquisition of these new technologies. Secondly, the data generated by production lines through real-time monitoring of operations has the potential to maximize yield, reduce waste [26], cut maintenance and operational costs, optimize schedules [27] and support lean manufacturing endeavors [28].

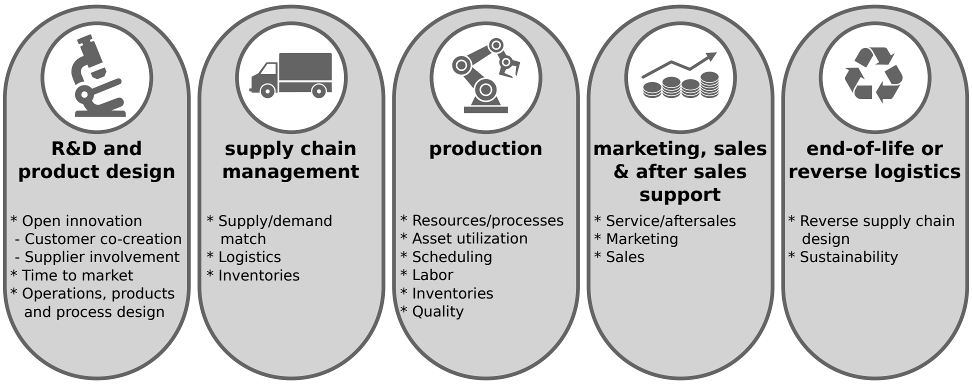


FIGURE 1.2: Data domains in manufacturing according to the work of Omar *et. al.* [2].

1.3 Potential of Complex Networks for Manufacturing

Manufacturing data analytics can take many different forms, such as machine learning [29–31] and big data analytics [32–34]. One promising yet rarely explored approach is "network science". The National Research Council defines "network science" as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena" [35]. In simple terms, network science is an academic field that studies complex networks where actors are represented as nodes and the relationship between them as edges. This field is truly multidisciplinary: its theories and methods are rooted in other fields such as mathematics (graph theory), physics (statistical mechanics), computer science (data mining and data visualization), statistics (inferential modeling) and sociology (social structure).

Advocacy for the use of "network science" in Operations Research dates back to 2008 [36, 37]. And while Alderson [36] promoted complex networks as a way to study network topology and associated statistical aspects, Nagurney [37] directed attention to the importance of incorporating flow related aspects. Yet, the existing literature within the domain of "production" in manufacturing is limited, as will be shown in Chapter 3. One possible explanation for this lack of research, may be the tendency of network science to abstract away domain-specific functions that typically drive engineers and practitioners [36]. As Alderson noted, one must determine which aspects of the problem are essential and which can be safely ignored [36]. Analogously, one must determine how to appropriately model the system under study. Furthermore, he pointed out that special attention should be payed to the underlying assumptions of the problem formulation and solution when utilizing methods from network science.

Yet some of these apparent shortcomings may be used to the practitioner's advantage. Analysis of manufacturing complex networks can aid production line design when the available data is limited. It can also be used to rapidly evaluate different alternatives, reducing the solution space before using computationally expensive simulations. After all, the majority of the algorithms and methods used to characterize complex networks are readily available and far less computationally expensive than traditional simulations. Thus, evaluating the applicability of existing complex networks metrics for Operations Research and appropriately interpreting results are open (and seldomly transited) research avenues worth exploring.

2 Preliminaries

2.1 Complex networks

A complex network, also called graph G, is composed of a finite set of nodes V and a set of edges E. It is thus generally indicated as $G(V, E)$. The set of nodes $V = \{v_0, v_1, ..., v_N\}$ must not only be finite, but also cannot be empty, i.e. $V \neq \emptyset$. The set of edges E is composed of pairs of nodes (v_i, v_j). These pairs indicate some kind of relationship between nodes.

Two nodes joined by an edge are typically called *adjacent* or *neighboring*. When the edges are unordered the graph is called *undirected* and, when the edges are ordered the graph is termed *directed*. Thus, in an undirected graph the edges (v_i, v_j) and (v_j, v_i) are equivalent while in a directed graph they are not.

A graph G can be completely described using an adjacency matrix $\mathcal{A}$. The elements of $\mathcal{A}$ have only two possible values: $a_{ij} = 1$ if there exists an edge between nodes v_i and v_j and 0 otherwise. A graph that is fully described by an adjacency matrix is known as *binary* or *unweighted*. Sometimes the edges carry a numerical value abstracting a property of the edge. This property can be a distance, the strength of a relationship, the number of transactions between pairs of nodes, etc. In this case, graph G is better described by means of a weight matrix $\mathcal{W}$ of elements $w_{ij} > 0$. These graphs are called *weighted*.

A graph G is called *connected* if, for every pair of distinct nodes v_i and v_j, there is a path from v_i to v_j. When this requirement is not held, the graph is said to be *disconnected*.

In directed graphs, a strongly connected component (SCC) can be defined. Thus, a SCC in a directed $G = G(V, E)$ is a maximal set of vertices $C \subseteq V$ such that for every pair of vertices u and v in C, there exists both a path $u \rightsquigarrow v$ and a path $v \rightsquigarrow u$. Thus, u and v are reachable from each other.

2.2 Notation

Throughout this work, a notation convention has been chosen. Node notations v_i and i are considered equivalent, $v_i \equiv i$, and used interchangeably. This is generally the case to avoid multiple subscripts where necessary.

2.3 Typology of centrality metrics

Much of this work deals with the calculation and interpretation of centrality metrics for manufacturing networks. It is thus important to understand the underlying assumptions that each metric entails. Some centrality metrics are only applicable to undirected graphs, others to directed ones. Formulation modifications may be necessary when dealing with a weighted graph instead of a binary one. Some metrics require connectedness, others do not. Yet one of the most relevant assumptions has to do with the type of network flow.

Borgatti [38] classified centrality metrics based on the way that the traffic flows through the network. There are thus, two aspects of flow worth studying: the type of route followed and the method in which the traffic propagates.

Types of routes in complex networks: According to Borgatti's classification [38], there are four possible routes that the traffic may follow. These are:

- **Paths:** sequence of linked nodes in which neither nodes nor edges are repeated. A path P of length n from node v_i to node v_j is an ordered sequence of distinct nodes $P = \{v_0, v_1, ..., v_n\}$ with $v_0 = v_i$, $v_n = v_j$ and $(v_t, v_{t+1}) \in E$ for $t = 0, 1, ..., n-1$.

- **Geodesics:** shortest paths, i.e. the shortest sequence of linked nodes in which neither nodes nor edges are repeated. The notation is analogous to paths, with the caveat that P is the shortest path from v_i to v_j.

- **Trails:** sequence of linked nodes where the nodes can be repeated, however each edge (v_t, v_{t+1}) can appear only once.

- **Walks:** sequence of linked nodes where both nodes and edges can be repeated.

Traffic propagation methods: Following Borgatti's classification [38], there are three possible ways in which the traffic propagates along the chosen route. These are:

- **Parallel duplication:** the traffic flows from a node v_i to all its neighbors $v_j \in \Gamma(v_i)$ simultaneously. A simple example is an e-mail message with multiple recipients. The message is sent to all of them simultaneously.

- **Serial duplication:** the traffic flows from one node to another one at a time, yet possibly occurring in multiple places. An example of serial duplication is gossip.

- **Transfer process:** type of traffic flow where an item can only be one place at a time and moves from node to node in the network. Money or any tangible good flow through a network following a transfer process.

TABLE 2.1: Centrality measures according to Borgatti's typology.

	Parallel Duplication	Serial Duplication	Transfer
Paths	Freeman Closeness [39] Freeman Degree [39]		Tutzauer Entropy [40]
Geodesics		Freeman Closeness [39]	Freeman Closeness [39] Freeman Betweenness [41]
Trails	Freeman Closeness [39] Freeman Degree [39]		
Walks	Freeman Closeness [39] Freeman Degree [39] Bonacich Eigenvector [42] PageRank [43]		

Each one of the centrality metrics discussed in this work makes assumptions with respect to the flow, i.e. the route followed and the traffic type. A list of centrality measures classified according to Borgatti's typology is presented in Table 2.1. It should be noted that in manufacturing networks, the flow is best described as following *paths* and propagating via a *transfer process*.

3 Literature Review

3.1 Introduction

The use of complex networks in manufacturing has been an enduring, albeit slow moving, trend. This is easily observed in the systematic literature review presented by Li *et. al.* in 2017 [44]. In their work, the authors investigated the application of complex networks in advanced manufacturing systems. They classified the existing literature in two main categories: manufacturing complex networks applied inside the enterprise and those describing inter-enterprise relationships. In turn, several sub-categories and specific issues were identified to further subdivide the literature, as shown in Table 3.1.

TABLE 3.1: Areas of application of complex networks in advanced manufacturing systems according to Li *et. al.* [44]

Category	Sub-category	Specific Issues
Intra-enterprise	Product design	Part & component Product module Product family & platform Product design & development
	Production	Manufacturing process Production line Manufacturing system
Inter-enterprises	Collaboration	Collaborative design & innovation Collaborative manufacturing Knowledge management
	Services	Manufacturing grid Service oriented manufacturing Cloud manufacturing
	Supply chain	
	Logistics	
	Organizational structure	Enterprise network Industry cluster

However, this classification criteria is not unique. Funke and Becker [45] analyzed manufacturing complex networks pertaining to material flows and identified three "perspectives on material flow": global production networks, supply chain networks and material flow networks. For example, on a global production network, companies are abstracted as nodes and contracts between them as edges. This perspective shares elements of the "collaboration" and "organizational structure" sub-categories presented in [44]. In the supply chain networks of Funke and Becker

[45], nodes represent plants and edges, deliveries much like in the “supply chain” sub-category in [44]. And in a material flow network, nodes are the abstract representation of machines and edges indicate material flows, similarly to the special issue of “manufacturing systems” in the “production” sub-category in [44].

In this work, “intra-enterprise” manufacturing networks constructed using “production” data are studied. Nodes represent workstations and edges, material flows among pairs of nodes. Thus, this work is concerned with the “manufacturing systems” special issue identified by Li *et. al.* [44], or equivalently, the “material flow networks” described by Funke and Becker [45].

In the remainder of this Chapter, a comprehensive literature review of manufacturing complex networks is presented. The focus is on “intra-enterprise” networks within the domain of “production”. More specifically, the literature pertaining to complex networks for “manufacturing systems” and/or “material flow networks” is discussed. The methodology followed for this systematic literature review is described in Section 3.2, followed by some general observations and a narrative description of the state of the art. Finally, analysis and classification of the research field makes it possible to determine the areas for further research explored in this work.

3.2 Methodology

3.2.1 Sources

The sources consulted for this literature review include

- Web of Science (primary source),
- ScienceDirect,
- Google Scholar, and
- Others, mostly consisting of records referenced in those identified by the sources indicated above.

3.2.2 Search criteria

The search criteria for Web of Science is reported in Table 3.2. Note that in all cases, the indexes used are SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI and the time-span chosen was “All years”.

A search in “TOPIC” looks for the search terms in the title, abstract, author keywords, and Keywords Plus. A search in “TITLE” looks for the search terms in the article title. The title refers to the title of a journal article, proceedings paper, book, or book chapter.

The same terminology in Table 3.2 was used to search for records in the other sources consulted. In ScienceDirect the search was conducted on title, abstract and

TABLE 3.2: Search criteria utilized in Web of Science for the literature review regarding complex networks in manufacturing.

Where	Search terms	# records found
TOPIC	(manufacturing AND "degree centrality")	15
TOPIC	(manufacturing AND "strength centrality")	0
TOPIC	(manufacturing AND "betweenness centrality")	18
TOPIC	(manufacturing AND "clustering coefficient")	8
TOPIC	(manufacturing AND pagerank)	12
TOPIC	(manufacturing AND "entropy centralilty")	0
TOPIC	(manufacturing AND "closeness centralilty")	0
TOPIC	(manufacturing AND "eigenvector centralilty")	0
TOPIC	(manufacturing AND "material flow network*")	19
TITLE	(manufacturing AND "complex network*")	13
TITLE	(manufacturing AND "social network analysis")	7

keywords for all years. In Google Scholar, the search was conducted on the title only for records in the year range 2010 to 2021.

3.2.3 Selection criteria

The systematic procedure followed to select relevant records is summarized in Fig. 3.1. In short, once all records were located, duplicates were discarded. The remainder records were filtered by title and then by abstract. In total 30 records were considered for full text analysis. For the fully reviewed records, the selection criteria required the following:

> Manufacturing complex networks described in the record must correspond to the "intra-enterprise" category, "production" sub-category and "manufacturing system" special issue in Li *et. al.* classification [44] **or, equivalently** they must correspond to the "material flow networks" perspective of material flow from Funke and Becker [45]. In these types of manufacturing networks, nodes represent workstations and edges are the abstract representation of material flows.

As a result, 18 records have been included in the literature review presented in this work. These are listed in Table 3.3 in chronological order.

It should be noted that the manufacturing networks under study share structural and flow similarities with "supply chain networks". However, these will not be discussed here. Readers interested in supply chain networks were nodes represent plants and edges describe deliveries should refer to [64–68] for practical examples.

3.3 General observations

The 18 records included in this literature review correspond to a time span of one decade, from 2012 to 2021. However, there is no clear trend over said period of time as shown in Fig. 3.2a. The research field is currently dominated by Becker's research

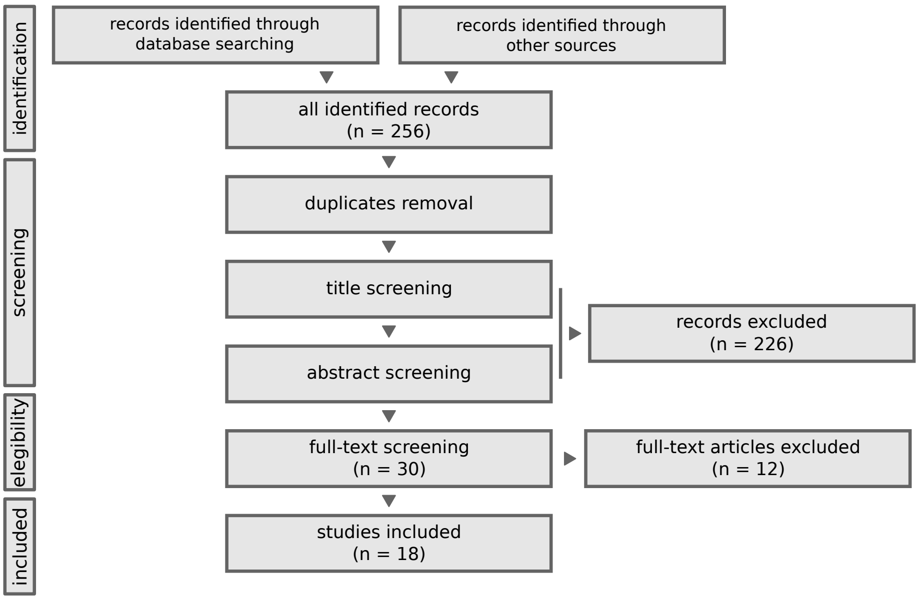
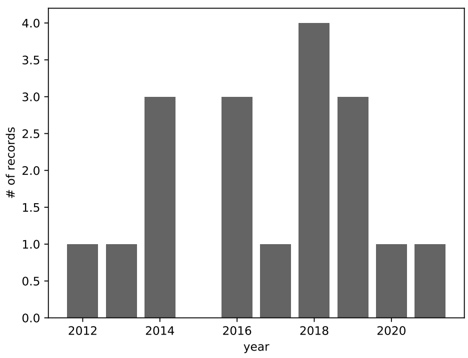


FIGURE 3.1: Methodology followed to select relevant records for the literature review. This methodology shares similarities with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) method [46].

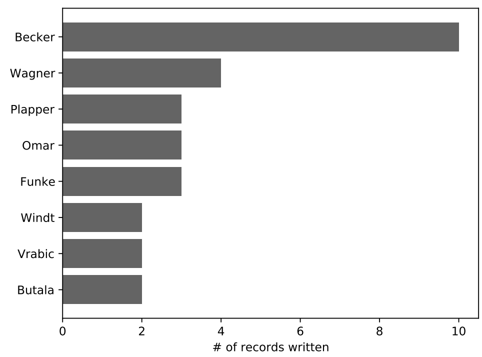
TABLE 3.3: Records included in the literature review.

Ref.	Year	Use case	Metrics
[47]	2012	autonomous work system detection	clique percolation method
[48]	2013	material flow anomaly detection	other
[49]	2014	topological characteristics and dynamical behavior	other
[50]	2014	production characterization	in-degree, out-degree, degree, betweenness, clustering
[51]	2014	bottleneck detection	degree, betweenness, pagerank
[52]	2016	quality assessment	pagerank
[53]	2016	key machine identification	in-degree, out-degree, PageRank, TOPSIS
[54]	2016	clusters in production networks	other
[55]	2017	generation of material flow networks	random walks
[56]	2018	synchronizability of the manufacturing network	ratio of the largest and first non-zero eigenvalues of the laplacian matrix
[57]	2018	modeling dynamic manufacturing networks	stochastic block model
[58]	2018	production characterization	in-degree, out-degree, in-strength, out-strength, betweenness, clustering, pagerank
[59]	2018	generation of material flow networks	random walks
[60]	2019	link prediction	stochastic block model
[61]	2019	production characterization, bottleneck detection	flow network
[62]	2019	bottleneck detection	combined
[45]	2020	material flow network prediction	stochastic block model
[63]	2021	routing flexibility	path-transfer entropy

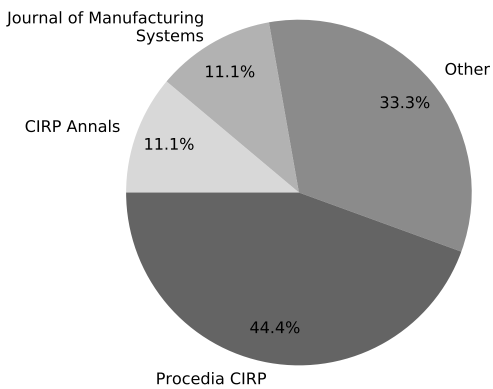
team as shown in Fig. 3.2b, yet other authors have made noteworthy contributions. Vrabič and Butala pioneered the field back in 2012 while Omar and Plapper's work has expanded it in recent years. Finally, it is noteworthy that most of the relevant literature appeared in CIRP associated journals, as shown in Fig. 3.2c, indicating industrial interest.



(A)



(B)



(C)

FIGURE 3.2: General observations drawn from the records included in the literature review. (A) Number of records by year of publication. (B) Number of records by author last name. (C) Distribution of records by journal.

3.4 State of the art

While complex networks had been used to describe a wide range of natural, social and engineered phenomena, it was Vrabič *et. al.* who first demonstrated the potential of this mathematical construct for manufacturing [47]. In their work, the authors identified “work systems” as nodes and explained that “work pieces” travel between them while being transformed into products. They explained that complex networks offer an objective view of the manufacturing system ignoring layout or functional similarities between nodes. They also pointed out that the data required could be easily extracted from the enterprise manufacturing execution system. In addition, this work provided a number of characteristics typical of a manufacturing complex network. The authors described the graph both as directed, indicating the direction of the material flow, as well as weighted, i.e. accounting for the number of "work pieces" transferred between "work systems". In fact, when furthering their research [48], Vrabič *et. al.* provided the mathematical formalism $G = (V, E, w)$ for manufacturing networks. In this formalism, V were the network nodes (i.e. the "work systems" or workstations), E were the connections among nodes created by the flow of materials and w was the weight of said connections, i.e. the number of transactions between "work systems".

The research of complex networks in manufacturing was soon extended by other scientists. Beber and Becker [49] focused on the analysis and characterization of manufacturing systems modeled as binary directed graphs $G = (V, E)$. They explained that the binary representation of the network was sufficient to describe the topology of the system. Among the advantages of complex networks in manufacturing, they cited the easy automation of the model generation and facilitated modeling process. However, they also pointed out that modeling manufacturing systems as complex networks suffers from a high level of abstraction.

Becker *et. al.* [50] were the first to apply traditional complex network topology metrics with the aim to characterize a manufacturing system in terms of Operations Research. The objective of their research was clear: they aimed "to investigate the relation between structure and performance". Thus, they studied binary, directed manufacturing networks using topological metrics such as degree, betweenness centrality and clustering coefficient as well as using traditional performance figures such as throughput time, lateness and work in progress. The authors found that manufacturing networks displayed distinct topological values when compared to equivalent random networks. In addition, they determined that the relation between topological metrics and performance indicators is not linear. The authors explained that their work could be used in place of classical methods such as computationally expensive simulation studies.

Blunck *et. al.* [51] proposed the use of complex network topological metrics as a mean to identify workstations that correspond to production bottlenecks. They compared the results of metrics such as degree, betweenness centrality and PageRank to the existing bottleneck identification heuristic approach, considering two throughput enhancement scenarios: capacity increase and batching. Under capacity increase alleviation, the bottleneck heuristics approach outperformed complex network topological metrics. Under batching alleviation, the bottleneck heuristic produces similar results to degree centrality. While the results of this work were underwhelming, a very important conclusion was drawn by the authors: complex network topological metrics cannot be universally applied. In fact, attention should be payed to the underlying flow type of the network as explained by Borgatti [38]. Notably, this is the first record in the research field to use Borgatti's argument as an explanation for poor results when compared to classical domain methods. Previously, Becker *et. al.* [50] had only indicated that the shortest paths followed by betweenness centrality may not accurately describe a manufacturing network.

Ai-ming *et. al.* [52] proposed a node ranking method to evaluate quality control networks. Their objective was to guide quality management activities which they claimed are often carried out blindly. In their quality control network $G = (V, E, w)$, the nodes V are workstations, the directed edges E represent material flows of unqualified batches and, the edges weight w denotes the number of unqualified batches traversing said edge. They then developed a node ranking algorithm based on PageRank and used it in combination with hierarchical clustering. This methodology allowed them to determine the workstations that should be prioritized in terms of quality control and improvement efforts.

Becker and Wagner [53] hypothesized that, in the quest to equip manufacturing systems with the necessary technology to be converted into cyber-physical production systems, some machines should be prioritized given their position in the production system. Thus, the authors aimed at identifying these key workstations from the complex network representation of the manufacturing system. They used several traditional topological metrics (degree, betweenness, closeness and eigenvector centrality) in addition to structural hole metrics (effective size, efficiency, constraint and hierarchy) and combined them using fuzzy analytic hierarchy process (FAHP) and TOPSIS. Their results showed that PageRank does a better job than the proposed combined metric. Just like Blunck *et. al.* [51], Becker and Wagner claimed that this could be explained by Borgatti's assertion [38], meaning that their proposed metric does not accurately represent the underlying flow of manufacturing networks.

Wagner and Becker [54] is the first recorded instance where a manufacturing system representation as a complex network is deemed a "material flow network". The authors specifically describe the components of these networks: nodes represent workstations and directed, weighted edges are the material flow among nodes. In their work, Wagner and Becker [54] aimed to identify clusters of workstations under the assumption that nodes within a cluster have similar characteristics or functions, similarly to [47]. For this purpose, they constructed manufacturing networks based on data corresponding to different time slots varying between 1 week and 3 months. They found that the statistical properties of the cluster structure in a production network depends on the selected time slot. The authors concluded that short time slots should be used when the objective is to identify autonomous acting clusters.

Wagner and Becker [55] noticed that the necessary data to build material flow networks and conduct complex network analysis is limited, thus hindering the advancement of this research field. They proposed a random walk based method to generate material flow networks that share similar structure and properties with original manufacturing networks. The procedure involves the generation of an undirected, unweighted network using either the Barabási-Albert (BA) model [69] or the Directed Configuration Model [70]. Then random walks are taken on these networks where the path length is chosen between 1 and $|V|$ with uniform probability and the start and end nodes of the walk are chosen at random. Wagner and Becker [55] evaluated the generated networks considering three variables: path length, number of random walks, and network topology. They concluded that shorter paths lengths better match the properties of the original manufacturing network. They also observed that the number of random walks has a greater effect on short path lengths, and thus should be chosen large enough. Finally, their results indicated that BA was better suited than Configuration Model for their objective. Later on, the authors furthered their work in order to allow for the generation of artificial material flow networks with specified cluster structures [59].

Chankov *et. al.* [56] studied the use of synchronizability of networks, a well studied field, in material flow networks. This metric is calculated as the ratio between the largest and the first non-zero eigenvalues of the Laplacian matrix, where the Laplacian matrix is obtained by subtracting the adjacency matrix from the degree matrix. The authors found that synchronizability is dependent on connectivity and the occurrence of linear elements (i.e. portions of production sequences with no alternative paths). It should be noted that this research differs from others in this literature review. This is because the synchronizability metric is well studied for undirected graphs and its use in directed networks is still unclear. Thus, the directed material flow network is transformed into an undirected graph in Chankov *et. al.*'s work [56]. It is unclear from the study what are the effects, if any, of this choice of representation.

Funke and Becker [57] evaluated the use of stochastic block model (SBM) and degree-corrected stochastic block model (DC-SBM) in manufacturing networks created with production data sliced weekly and monthly. The objectives were 1) to evaluate if the inferred internal structure of the network is consistent over time and, 2) whether it is possible to distinguish a real material flow network between a number of randomly generated networks with similar properties. The internal structure is found to be consistent over time. Furthermore, longer time scales have less fluctuations and thus, the internal structure is more easily captured. Finally, the authors found that DC-SBM can definitely identify true material flow networks, outperforming the simple SBM.

Omar *et. al.* [58] studied complex network topological metrics in the context of manufacturing aiming to provide the appropriate interpretation in terms of Operations Research. In a way, Omar *et. al.* furthered the work of Becker *et. al.* [50]. The authors interpreted the in-degree as the number of incoming links directly connected to a specific workstation. They then concluded that it is simply the number of upstream suppliers. The out-degree was interpreted as the number of outgoing links, and thus measures the number of downstream processes of a specific workstation. In this regard, they disagreed with Kim *et. al.* [64] and their analysis of supply chain networks where the in- and out-degrees were defined as a measure of "difficulty" in managing in-coming flows and customer needs respectively. Instead, Omar *et. al.* [58] proposed the use of the in- and out-strengths as a measure of the supply and demand loads on a workstation [58]. Furthermore, they explained that betweenness centrality determines structurally central workstations assuming these stand on the geodesic path connecting pairs of workstations. Yet they pointed out that shortest paths do not necessarily correspond to manufacturing paths, in agreement with Becker *et. al.* [50]. They stressed that, for this reason, the underlying assumptions of betweenness centrality may not hold for manufacturing networks. The authors also evaluated the interpretation of the clustering coefficient and agreed with previous research where the clustering coefficient was interpreted as a metric of interconnectedness [50, 58]. Finally, Omar *et. al.* analyzed the PageRank algorithm [58]. They explained that this algorithm is a probabilistic method that ranks workstations by importance based on effective processing paths. However, it should be noted that, in reality, the PageRank algorithm follows walks and not paths. The authors claimed that “the node importance [obtained from PageRank] measures the workload build-up of a node while accounting for inter-dependencies among pairs of nodes”. While they suggested it could be used to determine bottlenecks, no indications as to how were provided.

Funke and Becker [60] proposed a link prediction method based on DC-SBM for manufacturing networks. The objective was to advice production planning and production control in situations where existing methods fail due to lack of data. The proposed link predictor based on DC-SBM is used in ensemble learning to score new links based on the material flow networks generated for the past two weeks of data. The authors demonstrated that their method outperforms other existing link prediction methods.

Omar and Plapper [61] proposed the use of the mathematical formalism of “flow networks” for manufacturing systems, where “flow networks” should not be mistaken for the “material flow networks” defined by Funke and Becker [45]. The former are used to solve the optimization problem known as “maximum flow” under two assumptions: steady state flow of materials and zero intermediate inventory levels. The authors then obtain the theoretical maximum production rate and identify the bottleneck workstation based on capacity constraints.

Zhu *et. al.* [62] introduced a metric called “bottleneck degree index” to determine the production bottlenecks. It should be noted that the complex network constructed for this purpose has process operations as nodes and the relationship between nodes as edges. It is unclear what this relationship is. The authors stated that associations of processes, materials, machines and tools between nodes account as a relationship. It can then be process sequence, the impact of an operation on upstream and/or downstream processes or, other factors. The proposed “bottleneck degree index” is calculated, among other things, from a combination of the degree, closeness and betweenness centralities and, the clustering coefficient.

Funke and Becker [45] used the degree-corrected stochastic block model (SBM) to analyze “material flow networks” and predict changes in material flow. Their research focused on determining the appropriate data temporal aggregation to accurately represent a production system using SBM. They stressed that the "material flow networks" under study correspond to "the material flow of a single or a selection of multiple products or components within plants". This is noteworthy because similar assertions were made by Omar and Plapper [61] where they stressed that certain methods should be "constrained to a single product family of a specific product family mix and [results] are likely to differ from the values obtained for a different product family or mix".

Finally, Omar and Plapper [63] proposed the use of an entropy based metric called path-transfer entropy centrality to determine routing flexibility in manufacturing networks. Specifically, the authors used the probability distribution needed for the path-transfer entropy metric to evaluate manufacturing networks in terms of routing flexibility. The authors explain that the probability distribution of the binary representation of the network informs about routing flexibility potential while that of the weighted network is related to how much said flexibility is exercised in practice. In fact, the use of the binary, directed representation of manufacturing networks for structural analysis is shared by other authors [49, 50].

3.5 Discussion and outlook

The research field of manufacturing systems modeled as complex networks is relatively new, spanning only a decade. The number of works is limited mostly due to the difficulty in securing the necessary proprietary data, as noted by [55], but also due to the limited number of research groups involved. Nonetheless, it has fueled the interest of scientists and practitioners to tackle specific issues related to material flow networks, as shown in Table. 3.4.

TABLE 3.4: Classification of the literature regarding manufacturing systems modeled as complex networks.

Category	Sub-category	Specific Issues
Production	Characterization	Topological characterization Key machine identification Quality assessment Routing flexibility Clustering & autonomous work systems
	Scheduling & forecasting	Synchronizability Link prediction
	Troubleshooting	Abnormal flow detection Bottleneck identification
Artificial material flow networks		

The majority of the works are focused on the characterization of production networks. Yet, the specific issues tackled range from topological characterization [49, 50, 58] to identification of clusters [54] or key machines [53], quality assessment [52] or routing flexibility [63]. Regardless of how disparate these works are, they share unique concerns:

- how to best model manufacturing systems as complex networks?
- which metrics are appropriate considering the underlying flow network?
- what is the correct interpretation of these metrics is terms of Operations Research?

Some works explored aspects of production scheduling and forecasting, and thus focused on network synchronizability [56] and link prediction [60]. Yet more research is needed to consolidate their findings. Furthermore, it would be interesting to evaluate how characteristics of the production line, such as routing flexibility [63], could inform scheduling and forecasting.

Another set of records has been dedicated to production troubleshooting. In particular, bottleneck identification [51, 61, 62] has raised interest because of the potential to simplify the task by avoiding computationally expensive simulations. Yet the usefulness of the many proposed metrics have yet to be confirmed by other researchers.

Finally, some research focused on the issue precluding the field from advancing: the lack of data. In this sense, mechanisms were proposed in order to generate artificial material flow networks sharing similar structure and properties to real ones [55, 59]. Yet this field is nascent and in need for further development. Furthermore, it suffers from the very same problem it is trying to solve: real material flow networks are necessary to evaluate the performance of algorithms designed to generate artificial ones.

4 Aims and Scope

4.1 Problem definition

The use of complex networks to describe and evaluate manufacturing systems constitutes a nascent research field. As shown by the systematic literature review in Chapter 3, there are at this time a limited number of works tackling a vast number of issues. Yet, some prevalent areas of open research have interested multiple scientists. These are presented hereafter.

Mismatch between underlying network flow and centrality metrics: Several authors recognized that blindly utilizing complex network metrics for manufacturing networks does not produce the desired results [50, 51, 53, 58, 63]. In fact, when Alderson advocated the use of complex networks in the field of Operations Research [36], he warned the scientific community to consider the underlying assumptions of the problem formulation and solution. It thus remains that clarification regarding the circumstances in which each metric is applicable is needed.

Appropriate interpretation of complex network metrics in the context of manufacturing: Following the problem described above, the results obtained from utilizing complex network metrics for manufacturing networks requires the appropriate interpretation. This interpretation is domain specific, and thus cannot be borrowed from other more developed research fields such as social network analysis. Taking into account the research up to date and that of related fields such as supply chain networks, the interpretation of complex network metrics in the context of manufacturing flow networks requires further research.

4.2 Objectives

Following the problem definition introduced in the previous Section, this work aims to satisfy the following objectives:

1. Clarify the areas of applicability of popular existing complex network metrics in the context of manufacturing networks.
2. Propose and exemplify the use of other existing complex network metrics whose underlying assumptions better represent manufacturing networks.

3. Consolidate the appropriate interpretation of these metrics within Operations Research, and specifically in the domain of complex networks representing manufacturing systems.

4.3 Structure of this work

The structure of this thesis is summarized in Fig. 4.1. After a general introduction to the European manufacturing sector and Industry 4.0 in Chapter 1, preliminary complex networks concepts are introduced in Chapter 2. A systematic literature review of the state-of-the-art of complex networks in manufacturing follows in Chapter 3. Then the aims and scope of this work are presented here, in Chapter 4. The data used in this work and the methodology to construct the appropriate complex network is detailed in Chapter 5, followed by five self-contained Chapters.

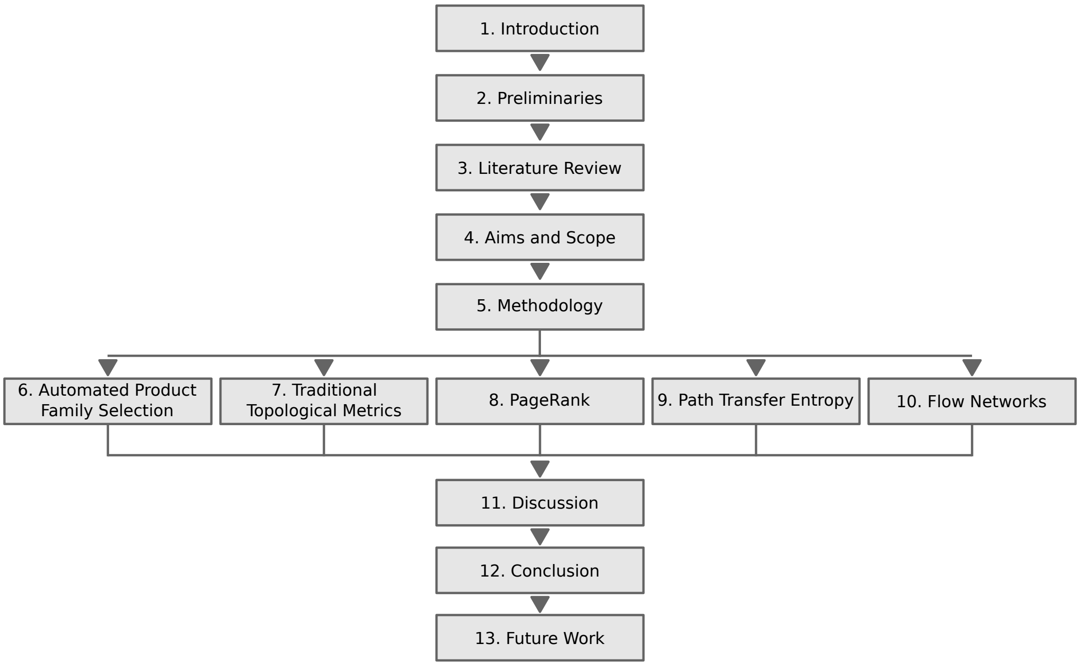


FIGURE 4.1: Structure of the thesis.

In Chapter 6, the procedure to obtain product families automatically is detailed. Chapter 7 describes the use of traditional complex network topological metrics in the context of manufacturing and provides the appropriate interpretation in terms of Operations Research. Chapter 8 explores the application of the PageRank algorithm in manufacturing, highlighting usefulness and shortcomings. Chapter 9 introduces the use of path-transfer entropy in order to determine the potential for routing flexibility in manufacturing production networks, as well as the extent to which said flexibility is exercised. Chapter 10 discusses the use of flow networks to determine the theoretical maximum production rate that the manufacturing production network can attain under the assumptions of steady state flow and zero intermediate inventory levels.

Chapters 6 to 10 are followed by three closing Chapters. A general discussion on the suitability, interpretation and shortcomings of complex network metrics in manufacturing systems is presented in Chapter 11. Conclusions are listed in Chapter 12. Finally, avenues for future research are presented in Chapter 13.

5 Methodology

5.1 The dataset

For this work, a dataset corresponding to a Bosch production line is used. The data is hosted by Kaggle and was made available for the "Bosch Production Line Perfomance Competition"[71]. This dataset was deliberately chosen as it is freely available and thus enables reproducibility and further development by the scientific community. Furthermore, it is an excellent example of the data produced by Industry 4.0 production lines.

The available data is anonymized. However, some general information regarding the dataset is available. As indicated in Fig. 5.1, the data corresponds to all manufactured parts during an unspecified period of time. The manufacturing facility consists of 4 manufacturing lines and 52 work stations. Furthermore, the data is stored in separate files corresponding to their type:

- time-stamp data recorded by multiple sensors on each work station,
- numeric data measured by each of those sensors , and
- network communication failure data.

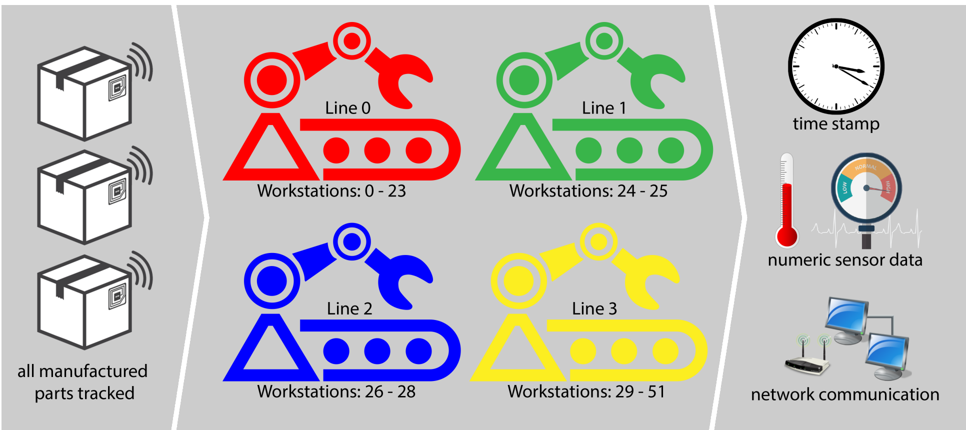


FIGURE 5.1: The Bosch manufacturing dataset is composed of time-stamp data corresponding to manufactured items in 4 manufacturing lines composed of 52 workstations for an unspecified period of time. In addition, anonymized numeric sensor data and network communication failure data is available.

5.2 From raw data to a complex network

In this Section, the step-wise procedure to generate a complex network from time-stamp raw data is presented. A summary of said procedure is shown in Fig. 5.2.

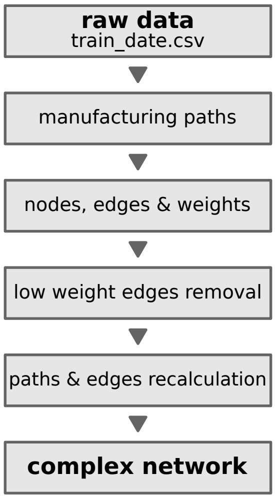


FIGURE 5.2: Step-wise methodology to obtain sets of nodes, edges and weights from raw time-stamp data

5.2.1 Raw data

For this work only the data contained in the "train_date.csv" file is relevant. An extract of said file is shown below:

```
Id,L0_S0_D1,L0_S0_D3,L0_S0_D5,L0_S0_D7,L0_S0_D9,L0_S0_D11,L0_S0...
4,82.24,82.24,82.24,82.24,82.24,82.24,82.24,82.24,82.24,82.24,8...
6, , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , ...
7,1618.7,1618.7,1618.7,1618.7,1618.7,1618.7,1618.7,1618.7,1618....
9,1149.2,1149.2,1149.2,1149.2,1149.2,1149.2,1149.2,1149.2,1149....
11,602.64,602.64,602.64,602.64,602.64,602.64,602.64,602.64,602....
13,1331.66,1331.66,1331.66,1331.66,1331.66,1331.66,1331.66,1331...
14, , , , , , , , , , , , , , , , , , , , , , , , , , , , , ,...
...
```

The first line contains column headers, where `Id` represent the manufactured item ID and `Lx_Sy_Dz` represents a time-stamp measurement registered in line `x`, workstation `y`, and feature `z` associated with a sensor measurement saved in the numerical features data file. The subsequent lines contain observations for individual manufactured items. Missing values indicate that the item was not processed at the specific workstation. In total, the file contains 1156 distinctive time features organized in columns, and close to 1.2 million observations (organized in rows) corresponding to manufactured items.

A full exploratory analysis was conducted on the raw data to gain further insights. The interested reader can find those results reported in Appendix A.

5.2.2 Manufacturing paths

A manufacturing path is understood as a sequence of visited workstations in chronological order. To obtain the manufacturing paths from the raw data shown in the previous Section, the visited workstations y were identified for each manufactured item. These visited workstations were ordered chronologically according to the value of the anonymized time-stamp. Furthermore, given that multiple items follow the same path, only unique paths and their frequency were stored. As an example, a small subset of manufacturing paths are reproduced along with their frequency in Table 5.1.

TABLE 5.1: A small subset of manufacturing paths and their frequency.

Frequency	Path
28137	$\{v_{24}, v_{26}, v_{29}, v_{30}, v_{33}, v_{34}, v_{36}, v_{37}\}$
8322	$\{v_{24}, v_{26}, v_{29}, v_{30}, v_{34}, v_{33}, v_{36}, v_{37}\}$
2370	$\{v_0, v_1, v_2, v_4, v_6, v_8, v_9, v_{29}, v_{30}, v_{33}, v_{34}, v_{36}, v_{37}\}$
2287	$\{v_0, v_1, v_2, v_5, v_7, v_8, v_9, v_{29}, v_{30}, v_{33}, v_{34}, v_{36}, v_{37}\}$
1005	$\{v_{12}, v_{13}, v_{15}, v_{17}, v_{18}, v_{20}, v_{22}, v_{29}, v_{30}, v_{33}, v_{34}, v_{36}, v_{37}\}$

5.2.3 Nodes, edges & weights

From the manufacturing paths obtained in the previous Section, one can trivially calculate the corresponding set of edges E, set of nodes V and the edges weights w_{ij}:

- The set of nodes V corresponds to the workstations in the manufacturing network. The union of all the path sets P gives rise to V.
- Each path $P = \{v_0, v_1, ..., v_i, ..., v_n\}$ allows the identification of a set of edges $E_P = \{(v_0, v_1), ..., (v_{i-1}, v_i), (v_i, v_{i+1}), ..., (v_{n-1}, v_n)\}$. The union of all the E_P gives rise to the set of edges E.
- Finally, for each edge (v_i, v_j), the edges weight w_{ij} is the frequency with which said edge appeared in the list of manufacturing paths.

This procedure identifies 52 nodes and 635 edges.

5.2.4 Removal of noisy data points

The removal of noisy data, corresponds to two steps in Fig. 5.2. The first step entails the analysis of the set E. This procedure identifies a number of edges with very low weight. Low weight edges are likely the result of noisy data recording. An arbitrary weight threshold of 0.1% of the total manufactured items was chosen to filter out "noisy" edges.

Once the clean set of edges was identified, the next step required the recalculation of both manufacturing paths and edges weights. The manufacturing paths were cleaned by removing those containing edges that were filtered out. The resulting clean paths subset contains 95.77% of the original manufacturing paths. The weight of the edges was recalculated from the clean paths dataset in the same fashion described earlier. The result is a set of 307 edges which visit only 50 out of 52 original nodes.

5.2.5 Complex network

The step-wise procedure described in the previous Sections produces a complex network comprised of 50 nodes, 307 edges and their corresponding weights. The full manufacturing network is plotted in Fig. 5.3. The color of the nodes follows the manufacturing line colors in Fig. 5.1.

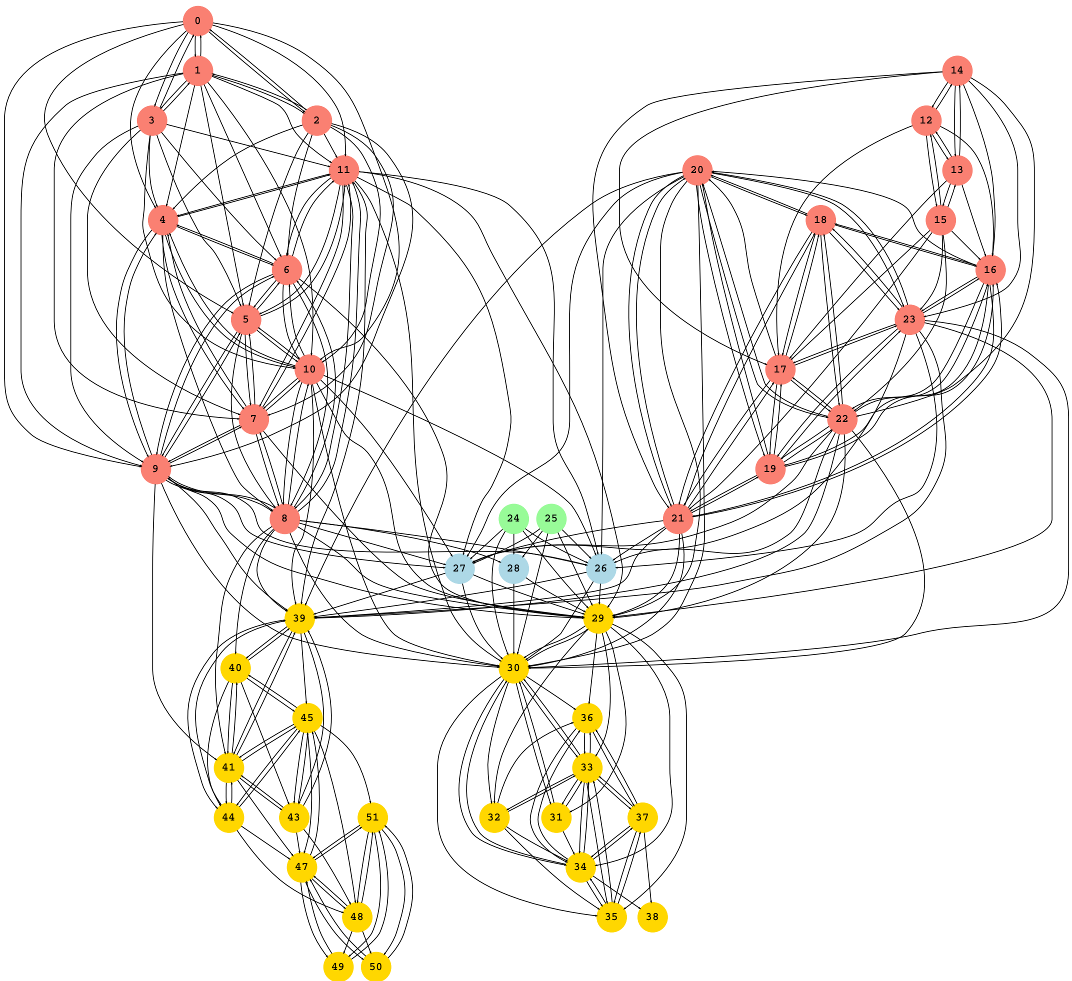


FIGURE 5.3: Complex network produced from the manufacturing raw time-stamp data. The color of the nodes follows the manufacturing line colors in Fig. 5.1.

5.3 Code

All the code used in this work was developed in Python 3 and is available in several online repositories on GitHub. Readers are encouraged to download and re-utilize it as needed, as long as the code is appropriately cited.

- The code for this Chapter can be found in [72]. Access request is necessary at this time. To reproduce the complex network presented in this Chapter, the reader should run the `from_timestamp_to_paths.py` script followed by `path_data_cleaning.py`.
- The code for Chapter 6 is available in [73].
- To reproduce the results presented in Chapter 7, the reader is encouraged to run the script `traditional_topological_metrics.py` in [72].
- The results presented in Chapter 8 can be reproduced by running the script `main_pagerank.py` in [72].
- The code used in Chapter 9 is available in [74].
- The code necessary to reproduce the results in Chapter 10 is available in the `flow_networks` directory in [72].

6 Automated Product Family Selection

6.1 Introduction

In value stream mapping (VSM), product families are determined using the principle of similarity. According to this principle, products with similar process sequences are grouped into the same family, i.e.

> "a product family [...] includes all products which are produced in similar production steps and on mostly identical machines" [75].

The procedure typically entails the visual identification of items that belong to a product family by determining the production steps they undergo and the machines in which they are processed. In the era of Industry 4.0, which is characterized by at scale production in automated factories and massive data generation, this is not ideal. In fact, this procedure is prone to human error. Thus, in this Chapter, an algorithm that conducts product family identification automatically without human intervention is developed.

6.2 High-level procedure for product family selection

In order to automate the product family selection process, the steps involved need to be identified. The starting data corresponds to the clean manufacturing paths obtained in Chapter 5. Note that each manufacturing path provides a set of workstations in which the item was processed in chronological order. Following the definition in [75], the high-level procedure is as follows:

1. Since the definition of product family does not constrain the manufacturing process to always follow the same chronological order, "path sets" are considered instead of "paths". A "path set" is simply the workstations visited ignoring the chronological order.

2. In order to find the biggest product families, the path sets should be ordered by decreasing number of frequency. Note that the frequency counts the number of items that have this exact path set.

3. Determine *family sets* or *bases*, i.e. a set of workstations that are common for a specific family.

4. Allocate path sets to family sets using a similarity metric.

6.3 Jaccard similarity of sets

Step 4 in the previous Section requires the allocation of each path set P_s to a family set F_s. Given the definition of product family, P_s must be a subset of or equal to F_s, i.e. $P_s \subseteq F_s$. In such scenario, the *Jaccard similarity of sets* is an appropriate similarity metric.

Definition: Given two sets S and T, the **Jaccard similarity** is $J(S,T) = |S \cap T|/|S \cup T|$, i.e. the ratio of the size of the intersection of S and T to the size of their union.

As a result, in order to allocate path sets to family sets, an arbitrary similarity threshold must be decided upon. In Section 6.6, a heuristic approach is proposed.

6.4 Automating Product Family Selection

The entire automation procedure is summarized in Fig. 6.1. Steps 1 and 2 in Section 6.2 are easily automated. The issue arises when a family set or base has to be determined (Step 3).

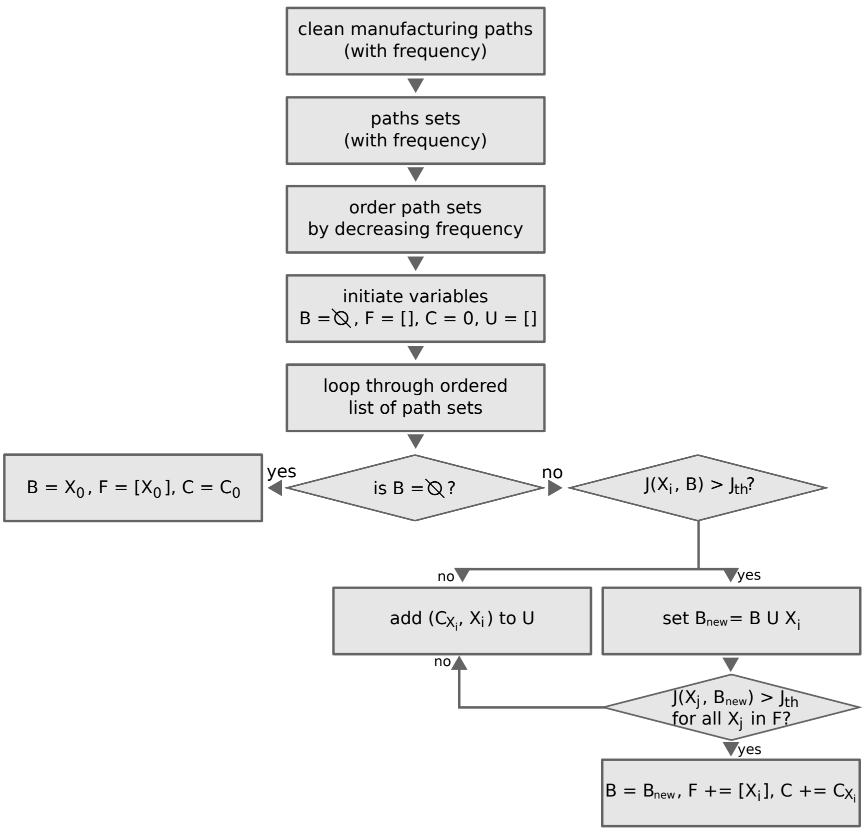


FIGURE 6.1: Procedure to obtained one product family in an automated fashion.

In traditional VSM, the family set was determined as the union of the visited workstations of family items. Items were grouped into families though a visual, manual, error prone and labor intensive process. However, a key component was the engineer's *a priori* knowledge regarding the production line and the manufactured items. The developed process does not have this *a priori* knowledge. Thus, to automate the product family selection, a family set is initialized as an empty set. Workstations are added to it as each path set is evaluated according to the similarity criteria described in the previous Section. The procedure, summarized in Fig. 6.1, is described as follows.

Starting with an ordered list L of path sets X_i in decreasing number of counts C_{X_i}, where $L = [(C_{X_i}, X_i)]$:

4. Initialize the following variables:
   - an empty base set $B = \emptyset$,
   - an empty list of family members $F = []$,
   - a family count of zero $C = 0$,
   - a list to save non-family path sets $U = []$.
5. Loop through the ordered list of path sets X_i (ordered in decreasing value of counts):
   (a) If $B = \emptyset$,
   - set $B = B \cup X_0 = X_0$, i.e. add the workstations of the first path set X_0 (the one with the highest count) to the base set B;
   - set $F \leftarrow X_0$, i.e. add the X_0 path set to the list of family members F.
   - set $C = C + C_{X_0}$, i.e. add its count to the family count.

   (b) If $B \neq \emptyset$,
   i. verify if $J(X_i, B) \geq J_{th}$. If the answer is yes,
   ii. verify if, given a new family base $B_{new} = B \cup X_i$, the following inequality $J(X_j, B_{new}) \geq J_{th}$ holds for all $X_j \in F$. If the answer is yes,
   iii. set $B_{new} = B \cup X_i$, add X_i to F and C_{X_i} to C.

   However, if either condition (i) or (ii) is not met, Step (iii) is ignored. In that case, one must add the pair (C_{X_i}, X_i) to U.

The pseudo-code for this procedure is shown in Algorithm 1. The corresponding Python 3 code is available in GitHub [73].

Algorithm 1 Find product family from path sets list

```
function ProductFamily(L, J_th)
    Sort(L, reverse = True)                          ▷ Sort L in reverse order
    for (C_{X_i}, X_i) in L do
        if B = ∅ then
            B ← B ∪ X_i
            F = Append(X_i)
            C = C + C_{X_i}
        else
            J = |B ∩ X_i|/|B ∪ X_i|                  ▷ J = J(B, X_i) is Jaccard Similarity
            if J ≥ J_th then
                B_new = B ∪ X_i
                if J(B_new, f_i) ≥ J_th ∀f_i ∈ F then
                    B ← B_new
                    F = Append(X_i)
                    C = C + C_{X_i}
                else
                    U = Append((C_{X_i}, X_i))
                end if
            else
                U = Append((C_{X_i}, X_i))
            end if
        end if
    end for
    return B, F, C, U
end function
```

6.5 Resulting product families

The iterative application of the procedure described in Algorithm 1, produces the results presented in Table 6.1 and Fig 6.2. In this case, a Jaccard similarity threshold $J_{th} = 0.63$ was used. This number was determined using a heuristic approach described in Section 6.6.

TABLE 6.1: Product families identified in the clean manufacturing paths from the Bosch Production Line data using the automated family search procedure described in Algorithm 1 with a threshold of $J_{th} = 0.63$. In this table, the base set elements are indicated using i for clarity. Note that $i \equiv v_i$.

Family	Base set	Base size	Count	%
F_1	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 35, 36, 37	20	528657	46.66
F_2	24, 25, 26, 27, 29, 30, 33, 34, 35, 36, 37	11	245225	21.64
F_3	12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 29, 30, 31, 33, 34, 35, 36, 37	20	184986	16.33
	Total items accounted for		958868	84.62

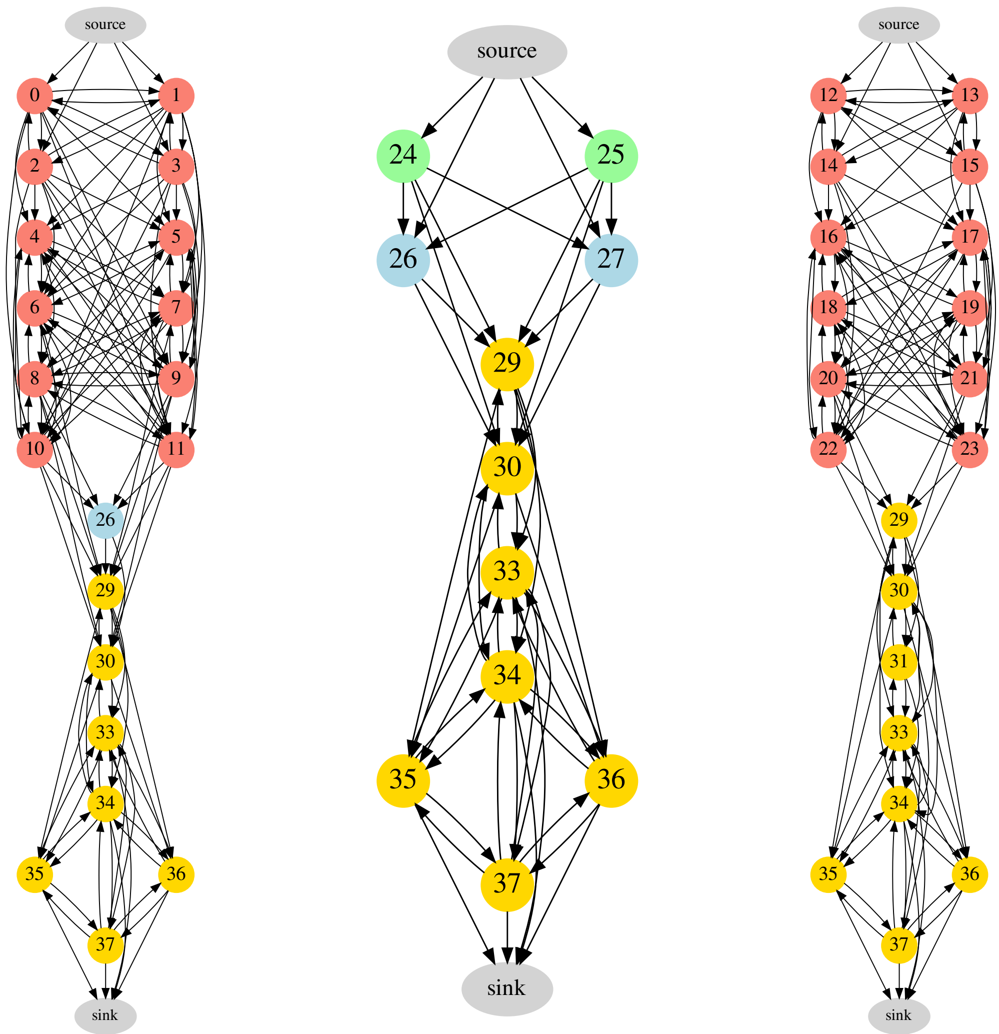


FIGURE 6.2: Graphs corresponding to the product families found automatically. The color of the nodes corresponds to the production line colors in Fig. 5.1. A *source* and a *sink* nodes have been added to facilitate visualization.

6.6 Heuristic study to determine the best J_{th}

In order to determine the optimal Jaccard similarity threshold value of $J_{th} = 0.63$ reported earlier, a brute force method was applied.

Firstly, an arbitrary Jaccard threshold $J_{th} = 0.7$ was used to determine the three biggest product families. For each family, the path set family member with the highest count was found. These are reported in Table 6.2.

Then, these high count path sets were used as a starting base (instead of the empty set) for family searches using J_{th} values ranging from 0.5 to 0.9. For each J_{th} considered, the fraction of manufactured items contained in each family was recorded. The optimal J_{th} was chosen graphically, coinciding with the cumulative curve $(F_1 + F_2 + F_3)$ inflection point. The fraction of items vs. threshold value curves

TABLE 6.2: Path sets with highest count for each product family. In this table, the path set elements are indicated using i for clarity. Note that $i \equiv v_i$.

F	Path set with highest count	Count
F_1	{0, 1, 2, 5, 6, 8, 11, 29, 30, 33, 34, 36, 37}	10569
F_2	{24, 26, 29, 30, 33, 34, 36, 37}	58375
F_3	{12, 13, 15, 17, 18, 20, 23, 29, 30, 33, 34, 36, 37}	3866

for each family as well as the cumulative curve are shown in Figure 6.3.

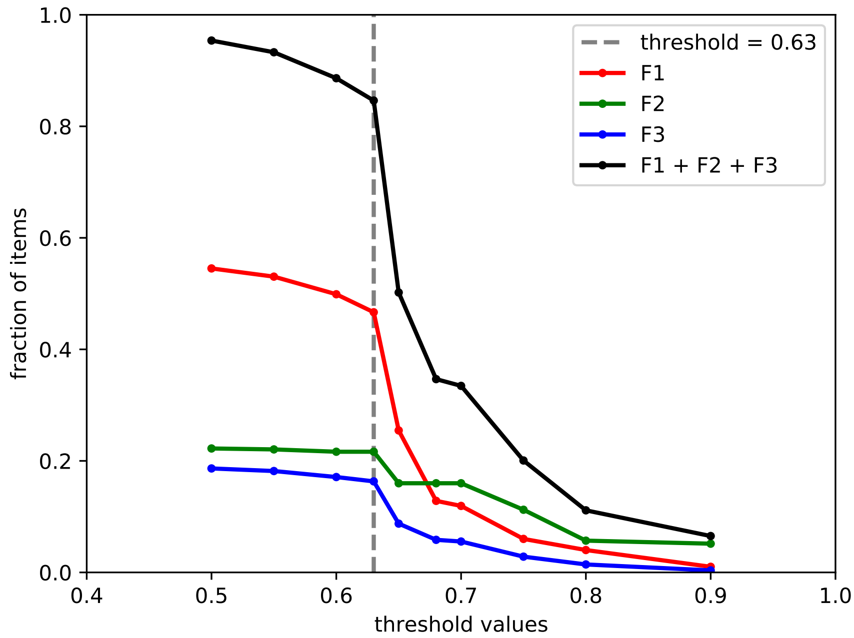


FIGURE 6.3: Results from the heuristic approach to Jaccard threshold determination for product family selection.

In order to interpret the results in Figure 6.3, the accompanying Tables 6.3, 6.4 and 6.5 are of help. It can be observed that at very high threshold values, $J_{th} = 0.9$, the family set is the starting base. As the threshold is decreased, i.e. the requirement of similarity is relaxed, the family set adds new workstations and the family count increases. From Figure 6.3, it can be concluded that as the similarity requirement is relaxed and we approach a threshold of $J_{th} = 0.63$, the total number of items accounted for (black line) increases steadily. At $J_{th} = 0.63$ there is an inflection point, and further relaxation of the threshold produces families that add new workstations but do not incorporate an important fraction of items. This means that at low threshold values, the obtained families deviate from the definition of product family in Value Stream Mapping. Thus, further relaxation of the threshold below $J_{th} = 0.63$ should be avoided.

TABLE 6.3: Jaccard similarity threshold analysis for family F_1. In this table, the base set elements are indicated using i for clarity. Note that $i \equiv v_i$.

J_{th}	Base set	Fraction
0.5	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38}	0.55
0.55	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38}	0.53
0.6	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 27, 29, 30, 33, 34, 35, 36, 37}	0.5
0.63	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 35, 36, 37}	0.47
0.65	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 35, 36, 37}	0.25
0.68	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 36, 37}	0.13
0.7	{0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 36, 37}	0.12
0.75	{0, 1, 2, 5, 6, 7, 8, 9, 10, 11, 26, 29, 30, 33, 34, 36, 37}	0.06
0.8	{0, 1, 2, 5, 6, 7, 8, 9, 11, 26, 29, 30, 33, 34, 36, 37}	0.04
0.9	{0, 1, 2, 5, 6, 8, 11, 26, 29, 30, 33, 34, 36, 37}	0.01

TABLE 6.4: Jaccard similarity threshold analysis for family F_2. In this table, the base set elements are indicated using i for clarity. Note that $i \equiv v_i$.

J_{th}	Base set	Fraction
0.5	{33, 34, 35, 36, 37, 24, 25, 26, 27, 28, 29, 30}	0.22
0.55	{24, 25, 26, 27, 28, 29, 30, 33, 34, 35, 36, 37}	0.22
0.6	{33, 34, 35, 36, 37, 24, 25, 26, 27, 29, 30}	0.22
0.63	{33, 34, 35, 36, 37, 24, 25, 26, 27, 29, 30}	0.22
0.65	{33, 34, 35, 36, 37, 24, 25, 26, 29, 30}	0.16
0.68	{33, 34, 35, 36, 37, 24, 25, 26, 29, 30}	0.16
0.7	{33, 34, 35, 36, 37, 24, 25, 26, 29, 30}	0.16
0.75	{33, 34, 35, 36, 37, 24, 26, 29, 30}	0.11
0.8	{33, 34, 36, 37, 24, 26, 29, 30}	0.06
0.9	{33, 34, 36, 37, 24, 26, 29, 30}	0.05

TABLE 6.5: Jaccard similarity threshold analysis for family F_3. In this table, the base set elements are indicated using i for clarity. Note that $i \equiv v_i$.

J_{th}	Base set	Fraction
0.5	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38}	0.19
0.55	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38}	0.18
0.6	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 29, 30, 31, 33, 34, 35, 36, 37}	0.17
0.63	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 29, 30, 31, 33, 34, 35, 36, 37}	0.16
0.65	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 29, 30, 31, 33, 34, 35, 36, 37}	0.09
0.68	{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 29, 30, 31, 33, 34, 35, 36, 37}	0.06
0.7	{12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 29, 30, 31, 33, 34, 35, 36, 37}	0.06
0.75	{12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 29, 30, 31, 33, 34, 36, 37}	0.03
0.8	{12, 13, 15, 17, 18, 19, 20, 21, 23, 29, 30, 31, 33, 34, 36, 37}	0.01
0.9	{12, 13, 15, 17, 18, 20, 23, 29, 30, 31, 33, 34, 36, 37}	0.004

6.7 Summary

In this Chapter,

- a simple algorithm to automatically determine product families was presented.
- the Jaccard similarity of sets was proposed to identify product family members.
- a heuristic approach was proposed for the definition of the similarity threshold.
- the families obtained respect the definition of product family widely accepted in VSM that says that "a product family [...] includes all products which are produced in similar production steps and on mostly identical machines" [75].

7 Traditional Topological Metrics

7.1 Introduction

The study of complex networks entails the analysis of the network structure, the characteristics of the flow, the identification of clusters or group of nodes of special significance, and much more. For these purposes, many metrics have been developed. While these metrics are used across scientific domains, their interpretation tends to be domain specific. That is, the analysis of the metric underlying assumptions have to be evaluated against the system being abstracted as a complex network. The result of this evaluation should determine the applicability of the metric in the context under study. If indeed the metric is applicable, its interpretation will be domain specific.

In this Chapter, a number of traditional structural metrics for complex networks are discussed. Their application is demonstrated on the Bosch manufacturing network produced following the methodology discussed in Chapter 5. The results are then compared to other works available in literature. Furthermore, the metrics are discussed with regards to the Borgatti typology [38] to which they belong to (see Section 2.3). Thus, the goal of this Chapter is to determine the applicability of traditional topological metrics in manufacturing networks, as well as their interpretation in terms of Operations Research.

7.2 Node degree

The node degree k_i accounts for the number of direct neighbors of a node. It is defined for undirected as well as directed graphs.

For an undirected graph, k_i is simply the number of nodes j to which i is adjacent. The degree k_i can be calculated directly from the adjacency matrix:

$$k_i = \sum_j a_{ij} = \sum_j a_{ji}. \tag{7.1}$$

For directed graphs, the in-degree k_i^{in} and out-degree k_i^{out} are defined. k_i^{in} is the number of in-going links while k_i^{out} indicates the number of outgoing links. They can also be calculated using the adjacency matrix:

$$k_i^{\text{in}} = \sum_j a_{ji} \tag{7.2}$$

$$k_i^{\text{out}} = \sum_j a_{ij} \tag{7.3}$$

Even though it is seldomnly used, the degree k_i of a node in a directed graph can be calculated as the sum of these two components [76]:

$$k_i = k_i^{\text{in}} + k_i^{\text{out}} = \sum_j a_{ji} + \sum_j a_{ij}. \tag{7.4}$$

7.2.1 Application example

The in- and out-degree of all 50 nodes in the Bosch manufacturing network (see Fig. 5.3) were calculated and are reported in Table 7.1. A number of observations can be made:

- k_i^{in} values fluctuate between 0 and 17.
- The nodes with the highest k_i^{in} values are v_{30}, v_{29} and v_{39} with $k_i^{\text{in}} = 17$, 16 and 14 respectively.
- Nodes v_{24} and v_{25} have an in-degree of 0. This is an artifact of the calculation since these two nodes are the starting workstations of product family F_2 (see Fig. 6.2). Thus, their incoming flow consists of raw materials. The value of 0 then indicates that they are not dependent on the work of any upstream workstation.
- The starting nodes of product families F_1 and F_3 do not have an in-degree value of zero because they form triangles among themselves (see Section 7.6).
- k_i^{out} values fluctuate between 0 and 13.
- The nodes with the highest k_i^{out} are v_8 and v_9 with $k_i^{\text{out}} = 13$ and 11 respectively. The third place is shared by 7 different nodes, with a value of 10.
- Node v_{38} has $k_i^{\text{out}} = 0$. This is because node v_{38} is an end node (manufacturing finalizes there). Other end nodes exist and they are evidenced when analyzing the strength (see Section 7.3). End nodes are not evident from a degree perspective on this network because they form triangles among themselves (see Section 7.6).

7.2.2 Interpretation

As explained in the literature review presented in Chapter 3, a number of authors have demonstrated the use of the degree in manufacturing networks [50, 51, 53, 58, 62]. For example, Becker *et. al.* [50] explained that in a manufacturing network abstracted as a directed graph, the in-degree indicates the number of upstream workstations that v_i is directly connected to. Equivalently, the out-degree specifies the

TABLE 7.1: Results obtained from calculating the in- and out-degree, the in- and out-strength, the betweenness centrality and the clustering coefficient of the Bosch manufacturing network reported in Fig. 5.3

v_i	k_i^{in}	k_i^{out}	s_i^{in}/t_m	s_i^{out}/t_m	$C_B(v_i)$	$C_C^D(v_i)$	$\tilde{C}_C^D(v_i)$
0	3	8	0.18	0.57	1.0	0.28	0.0162
1	3	10	0.41	0.57	1.25	0.29	0.0123
2	2	9	0.28	0.29	0.25	0.3	0.0118
3	2	9	0.28	0.28	0.25	0.3	0.0116
4	9	6	0.28	0.28	3.42	0.28	0.0085
5	9	6	0.29	0.29	3.42	0.28	0.0086
6	9	7	0.29	0.29	8.66	0.27	0.0074
7	9	7	0.28	0.28	8.66	0.27	0.0072
8	7	13	0.57	0.57	46.27	0.19	0.0078
9	9	11	0.19	0.19	66.69	0.16	0.0039
10	9	10	0.19	0.19	33.18	0.17	0.0051
11	9	10	0.19	0.19	33.18	0.17	0.0051
12	3	5	0.05	0.2	1.0	0.28	0.0094
13	3	5	0.15	0.2	1.0	0.28	0.0101
14	2	7	0.1	0.1	11.36	0.21	0.0046
15	2	7	0.1	0.1	11.36	0.21	0.0047
16	9	6	0.1	0.1	20.38	0.23	0.0019
17	9	6	0.1	0.1	20.38	0.23	0.0019
18	6	6	0.1	0.1	2.0	0.33	0.0028
19	6	6	0.1	0.1	2.0	0.33	0.0028
20	7	10	0.2	0.2	31.1	0.24	0.003
21	7	10	0.07	0.07	53.81	0.16	0.0016
22	7	10	0.07	0.07	53.81	0.16	0.0016
23	7	10	0.06	0.06	53.81	0.16	0.0016
24	0	5	0.0	0.16	0	0.2	0.005
25	0	5	0.0	0.07	0	0.2	0.0035
26	10	3	0.19	0.19	11.0	0.13	0.0038
27	10	3	0.1	0.1	11.0	0.13	0.0028
28	3	2	0.01	0.01	0	0.2	0.0015
29	16	7	0.96	0.96	129.0	0.11	0.0065
30	17	7	0.96	0.96	112.57	0.11	0.0065
31	3	3	0.03	0.03	0	0.44	0.0128
32	3	4	0.02	0.02	0	0.39	0.0064
33	8	7	0.96	0.93	22.33	0.25	0.0228
34	8	6	0.96	0.89	48.33	0.26	0.0259
35	6	3	0.47	0.44	8.0	0.35	0.0297
36	6	3	0.49	0.45	8.0	0.35	0.0281
37	4	5	0.96	0.18	2.17	0.33	0.0371
38	2	0	0.02	0.0	0	0.5	0.0083
39	14	5	0.04	0.04	208.68	0.12	0.0006
40	4	5	0.04	0.04	5.73	0.36	0.0024
41	7	6	0.04	0.04	61.31	0.26	0.0016
43	4	5	0.02	0.02	26.24	0.32	0.0016
44	4	5	0.02	0.02	26.24	0.32	0.0015
45	6	7	0.04	0.04	84.91	0.25	0.0014
47	8	5	0.04	0.04	75.48	0.21	0.0011
48	5	4	0.04	0.04	18.17	0.28	0.0018
49	3	2	0.02	0.02	0	0.5	0.0028
50	3	2	0.02	0.02	0	0.5	0.0029
51	5	4	0.04	0.01	7.61	0.3	0.0016

number of downstream workstations that v_i is directly linked to [50]. Blunck *et. al.* [51] used the degree centrality as a means to identify bottleneck workstations.

Their research indicated that under batching alleviation, the traditional bottleneck heuristic produces similar results to degree centrality. Zhu *et. al.* [62] also aimed at determining bottlenecks using a combined metric containing the degree centrality. Becker and Wagner [53] used degree centrality in combination with other metrics and TOPSIS with the objective to identify key machines that should be prioritized when equipping a manufacturing network with cyber-physical systems. However, they indicated that PageRank was better suited for said task. Finally, Omar *et. al.* focused on appropriately interpreting degree centrality in terms of Operations Research, agreeing with Becker *et. al.* previous interpretation [50].

While it was not discussed in the literature review, the degree centrality has also interested authors studying supply chain networks, which are structurally related to manufacturing networks. For example, Kim *et. al.* [64] linked the in-degree to the degree of difficulty faced by a node when managing incoming flows. The authors explained that the in-degree could be regarded as a metric of each node's operational load coming from upstream suppliers. The out-degree, on the other hand, was said to specify not only the size of the adjacent downstream tier, but also the difficulty faced by each node in managing the needs of customer nodes [64]. However, it should be pointed out that the node degree does not take actual material flow between neighboring nodes into consideration. This is because the node degree is calculated from the adjacency matrix where all existing edges are weighted equally. Thus, using the node degree as a metric of operational load is misleading [58].

According to Borgatti's typology of centrality metrics described in Section 2.3 and Table 2.1, the degree centrality corresponds to the "parallel duplication" traffic propagation method. In terms of routes, the degree is associated with "paths", "trails" and "walks". Practitioners should thus exercise caution since the underlying flow of manufacturing networks corresponds to the traffic type "transfer process" following "paths". In this work, it is proposed to limit the use of degree centrality to the following:

> *In a manufacturing network where workstations are abstracted as nodes and directed edges represent material flows, the in-degree is best interpreted as the number of direct upstream suppliers. Equivalently, the out-degrees is interpreted as the number of direct downstream customers. The degree is* ***not*** *a metric of operational load.*

7.3 Node strength

The node strength is the natural generalization of the node degree for the case of weighted graphs [58, 76]. In this case, it is calculated from the weight matrix. Similarly to the node degree, it is defined for both undirected and directed graphs. For undirected graphs it can be calculated as

$$s_i = \sum_j w_{ij} = \sum_j w_{ji}. \tag{7.5}$$

For directed graphs, the in- and out-strength are defined as follows

$$s_i^{\text{in}} = \sum_j w_{ji} \tag{7.6}$$

$$s_i^{\text{out}} = \sum_j w_{ij}. \tag{7.7}$$

7.3.1 Application example

The in- and out-strength of all 50 nodes in the Bosch manufacturing network shown in Fig. 5.3 were calculated. In Table 7.1 the normalized values of the in- and out-strengths are reported. Normalization is done by dividing the value of the in- or out-strength by the total number of manufactured items t_m, i.e. s_i^{in}/t_m and s_i^{out}/t_m. Thus, a value close to 1 indicates that all manufactured items have been processed by node v_i.

A number of observations can be made when observing the results reported in Table 7.1:

- As a general rule, $s_i^{\text{in}} = s_i^{\text{out}}$, i.e. material flow conservation laws apply. Thus the material flowing into a workstation, leaves it after processing. However, when $s_i^{\text{in}} - s_i^{\text{out}} < 0$, node v_i acts as a source. In this case, node v_i corresponds to a "starting node" and thus receives raw materials from outside the network. When $s_i^{\text{in}} - s_i^{\text{out}} > 0$, node v_i acts as a sink. Thus, it corresponds to a "finishing node", manufacturing can conclude there and finished products are delivered outside the network.

- The highest normalized in-strength value of 0.96 is shared by nodes v_{29}, v_{30}, v_{33}, v_{34} and v_{37}. The next highest normalized in-strength value is 0.57 for node v_8 followed by 0.49 for node v_{36}.

- The highest normalized out-strength value is of 0.96 for nodes v_{29} and v_{30}, followed by 0.93 (node v_{33}) and 0.89 (node v_{34}).

- From the results reported in the previous two bullet points, it is concluded that nodes v_{29} and v_{30} are intermediate nodes since $s_i^{\text{in}} = s_i^{\text{out}}$. On the other hand, nodes v_{33}, v_{34} and v_{37} are finishing nodes, since $s_i^{\text{in}} - s_i^{\text{out}} > 0$.

7.3.2 Interpretation

The node strength is a rarely explored metric for manufacturing networks as demonstrated by the literature review in Chapter 3. Only Omar *et. al.* proposed its use [58]. In their work, the authors explained that the strength is a measure of the workload that a node is subjected to [58]. Differently from the degree, the strength is based on the weight matrix instead of the adjacency matrix. Thus material flows are accurately accounted. As a result, this metric is a better measure of the workload to which a workstation is subjected to than the degree. This holds specially true for networks with highly heterogeneous material flows between sets of nodes, as is the case of the Bosch manufacturing network.

As the natural generalization of the degree, the strength shares the same underlying flow according to Borgatti's typology (see Section 2.3). Then the traffic propagation type is "parallel duplication" and the route can be that of "paths", "walks" or "trails". Practitioners are advised to use this metric with caution given that its underlying assumptions do not appropriately match manufacturing networks. The following interpretation is proposed:

> *In a manufacturing network where workstations are abstracted as nodes and directed edges represent material flows, the in-strength is a measure of the incoming workload from suppliers. The out-strength is a measure of the difficulty faced by the node to satisfy downstream customer needs.*
>
> *In general, $s_i^{in} = s_i^{out}$. However, when the source of raw materials and the sink for finished goods are not abstracted as nodes, differences between the in- and out-strength can be used to identify starting and ending manufacturing nodes. When $s_i^{in} - s_i^{out} < 0$, node v_i acts as a source, and thus corresponds to a starting node. When $s_i^{in} - s_i^{out} > 0$, node v_i acts as a sink, and thus manufacturing can conclude there.*

7.4 Betweenness centrality

The betweenness centrality C_B, typically attributed to Freeman [41], measures the fraction of times in which a node v_i appears on the geodesic (a.k.a. shortest) path σ between any two other nodes s and t. It can be determined as follows:

$$C_B(v_i) = \sum_{s,t \in V} \frac{\sigma(s,t|v_i)}{\sigma(s,t)} \tag{7.8}$$

Its calculation is far from trivial. A matrix based calculation is described in [77]. Yet, a faster algorithm was developed by Brandes [78] and later extended to weighted networks [79]. The reader should note that "Algorithm 11" presented in [79] for weighted networks contains an error. The accumulation part is missing. A factor of $w(v,w)$ should be applied to $\sigma[v]/\sigma[w]$. The erratum is available in [80].

7.4.1 Application example

The (not-normalized) betweenness centrality for all 50 nodes of the weighted, directed manufacturing network in Fig. 5.3 was calculated. Results are reported in Table 7.1. A few observations can be drawn:

- The value of $C_B(v_i)$ varies between 0 and 208.68.
- The highest values of $C_B(v_i)$ correspond to 208.68, 129.0 and 112.57 for nodes v_{39}, v_{29} and v_{30} respectively. Note that, nodes v_{29}, v_{30} and v_{39} also have the highest k_i^{in}, as shown earlier in Section 7.2.
- From the graph in Fig. 5.3, it can be observed that these nodes receive incoming flow from multiple upstream nodes and, after processing, they conduct the flow to downstream nodes. This is specially evident for node v_{39}, which with few exceptions, controls the downstream flow. On the other hand, nodes v_{29} and v_{30} share this gatekeeper quality with respect to their downstream tier.

7.4.2 Interpretation

Betweenness centrality was originally introduced to quantify the importance of an individual in a communication network. In this case, importance is understood as an individual's potential to control information flows between pairs of other individuals [41]. Nodes with high betweenness centrality are typically termed "gatekeepers", since they can prevent the flow of information.

Following Borgatti's typology of centrality metrics described in Section 2.3, the underlying network flow for betweenness centrality is assumed to follow "geodesics". The traffic type is that of a "transfer process". In this sense, it can be understood why several authors have evaluated this metric for manufacturing networks [50, 51, 53, 58, 62]. This metric assumes the same traffic propagation method that manufacturing networks follow. However, as noted by several authors [50, 51, 58] "geodesics" do not accurately describe the route followed by the flow in manufacturing networks. "Paths" are considered more appropriate [58, 63].

The literature review presented in Chapter 3 demonstrated that available research differs in the use and interpretation of betweenness centrality. Becker *et. al.* described this metric as a measure of a node's potential to impede or facilitate material flows [50]. They explained that nodes deemed structurally central stand between others exerting a high degree of control on the flow of materials. According to [50], these nodes could easily starve their downstream tier. Yet Blunck *et. al.* determined that betweenness centrality performed poorly as a metric to identify bottlenecks [51], contradicting [50]. Yet, Zhu *et. al.* used a combined metric for bottleneck identification that is based, among other metrics, on betwenness centrality with positive results [62]. Also in a combined metric but with the objective to identify key machines, Becker and Wagner used betweenness centrality as well but reported it underperformed compared to PageRank [53].

It follows that researchers have used betweenness centrality with varied degrees of success. As explained earlier, following Borgatti's typology for centrality metrics [38], betweenness centrality is applicable to networks whose flow follows "geodesics" (a.k.a shortest paths) and where the propagation corresponds to a "transfer process". Since this underlying flow differs from that in manufacturing networks, Omar *et. al.* considers its applicability questionable [58]. Practitioners are advised against using this metric given that its underlying assumptions do not appropriately match manufacturing networks. The following conclusions are drawn:

> *Betweenness centrality can be used to identify nodes that may act as gatekeepers when the flow follows shortest paths. However, in modern manufacturing networks, it is possible that those nodes may be circumvented. Thus, at this time, betweenness centrality is considered a poor choice for bottleneck and/or key machine identification.*
>
> *Note, however, that used in conjunction with a network graph as exemplified in the previous Section, betweenness centrality may aid the identification of nodes that may disconnect upstream suppliers from their downstream tier. More research is this regard is encouraged.*

7.5 Clustering coefficient

The clustering coefficient $C_C(v_i)$ was firstly introduced by Watts and Strogatz [81] in 1998. It measures the likelihood that two neighbors of a node v_i are adjacent. Equivalently, one can calculate the ratio between the number of triangles t_i in which v_i is one vertex and the number of all possible triangles that v_i could form, T_i, as follows:

$$C_C(v_i) = \frac{t_i}{T_i} = \frac{2t_i}{k_i(k_i-1)}. \tag{7.9}$$

This formulation [81] is applicable to the case of binary undirected networks (BUN). Multiple generalizations have been introduced to extend the application of this metric to other types of graphs. The extension of the clustering coefficient to weighted undirected networks was developed by Saramaki *et. al.* [82]. The appropriate formulation for directed graphs, both binary (BDN) and weighted (WDN) was introduced by Fagiolo [83].

The clustering coefficient of a BDN, $C_C^D(v_i)$, is defined as the ratio between all *directed* triangles that node v_i forms in the network (t_i^D) and the number of all possible triangles that v_i could form (T_i^D) [83]. Its formulation is:

$$C_C^D(v_i) = \frac{t_i^D}{T_i^D} = \frac{\sum_j \sum_h (a_{ij}+a_{ji})(a_{ih}+a_{hi})(a_{jh}+a_{hj})}{2[k_i(k_i-1)-2k_i^{\leftrightarrow}]} \tag{7.10}$$

where k_i is the degree of node v_i. As described in Section 7.2, the total degree in a directed graph can be calculated as $k_i = k_i^{\text{in}} + k_i^{\text{out}}$. $k_i^{\leftrightarrow}$ is the number of bilateral edges between v_i and its neighbors (i.e. the number of nodes v_j for which both edges, $v_i \rightarrow v_j$ and $v_j \rightarrow v_i$, exist). This can be calculated as $k_i^{\leftrightarrow} = \sum_{j \neq i} a_{ij}a_{ji}$.

Fagiolo [83] presented the straightforward extension of the clustering coefficient of BDN to the case of WDN. The latter are characterized by a $N \times N$ weight matrix $W = \{w_{ij}\}$. Note that it is required that all weights are in the range of 0 to

1, $w_{ij} \in [0,1]$. Thus, if some $w_{ij} > 1$, all weights in the network must be divided by $max_{i,j}\{w_{ij}\}$ in order to fulfill this requirement. Then, the clustering coefficient $\tilde{C}_C^D(v_i)$ for a WDN is calculated as:

$$\tilde{C}_C^D(v_i) = \frac{\tilde{t}_i^D}{T_i^D} = \frac{\sum_j \sum_h (w_{ij}^{1/3} + w_{ji}^{1/3})(w_{ih}^{1/3} + w_{hi}^{1/3})(w_{jh}^{1/3} + w_{hj}^{1/3})}{2[k_i(k_i - 1) - 2k_i^{\leftrightarrow}]} \tag{7.11}$$

It is noteworthy that $C_C(v_i)$, $C_C^D(v_i)$ and $\tilde{C}_C^D(v_i)$ give the local clustering coefficient for node v_i. In many applications however, a network clustering coefficient $C_C(G)$ is of interest. Watts and Strogatz [81] proposed calculating $C_C(G)$ as the average clustering coefficient for all networks. It thus follows that $C_C = N^{-1} \sum_i C_C(v_i)$ for BUN, $C_C^D = N^{-1} \sum_i C_C^D(v_i)$ for BDN, and $\tilde{C}_C^D = N^{-1} \sum_i \tilde{C}_C^D(v_i)$ for WDN.

Note that other averaging methods exist. For example, Borgatti *et. al.* [84] proposed a weighted average approach, where the weights are the number of pairs of nodes in each node's ego network, $k_i(k_i - 1)/2$. They called this the "weighted overall clustering coefficient". However, in this work, that of Watts and Strogatz will be utilized.

7.5.1 Application example

The local clustering coefficient for all 50 nodes in the directed manufacturing network in Fig. 5.3 were calculated. For this, the network was considered first as binary and then as weighted. This is because there is agreement in literature [63, 85] that, for certain graph metrics, the binary version provides information regarding the structure of the network while the weighted version informs about the practical implementation. The results are reported in Table 7.1, columns $C_C^D(v_i)$ and $\tilde{C}_C^D(v_i)$. The following observations are drawn:

- The average clustering coefficient for the BDN is $C_C^D = N^{-1} \sum_i C_C^D(v_i) = 0.263$, indicative of modest structural connectedness.
- Local clustering coefficients $C_C^D(v_i)$ for the BDN climb up to 0.5 for nodes v_{38}, v_{49} and v_{50}. Yet it can be as low as 0.11 for nodes v_{29} and v_{30}.
- The average clustering coefficient for the WDN is $\tilde{C}_C^D = N^{-1} \sum_i \tilde{C}_C^D(v_i) = 0.007 \ll 1$, pointing to very low connectedness in practice.
- The local clustering coefficients for the WDN is also $\tilde{C}_C^D(v_i) \ll 1$ for all 50 nodes. The highest values recorded are $\tilde{C}_C^D(v_{37}) = 0.0371$, $\tilde{C}_C^D(v_{35}) = 0.0297$ and $\tilde{C}_C^D(v_{36}) = 0.0281$. Yet, it can be as low as 0.0006 for node v_{39}.

7.5.2 Interpretation

The clustering coefficient metric developed by Watts and Strogatz [81] had the objective to capture the extent to which a network had areas of high and low density.

In fact, $C_C(v_i)$ measures the density of ties for node v_i ego network, i.e. the density of connections among nodes connected to v_i [84]. The individual clustering coefficient is averaged among all nodes in the network to obtain the graph clustering coefficient. This value is typically interpreted as a measure of connectedness, and it ranges from 0 for a network composed of isolated nodes to 1 for fully connected network.

The use of the clustering coefficient in the literature corresponding to manufacturing networks is limited, as shown in Chapter 3. Becker *et. al.* used the binary undirected formulation of the clustering coefficient [50]. They claimed that $C_C(G)$ describes the network under study in terms of connectedness. They regarded high values as indicative of highly interconnected nodes as would be the case of workstations in cellular manufacturing. On the other hand, low clustering coefficient values were associated with serial manufacturing plants. Omar *et. al.* proposed the use of the weighted directed formulation of the clustering coefficient instead [58]. Yet the authors agreed with the previous interpretation proposed by Becker *et. al.* [50].

Following the agreed interpretation [50, 58], the Bosch manufacturing network studied here is indicative of mostly serial manufacturing. However, observation of Fig. 5.3 shows groups of nodes that seem highly interconnected. Further analysis indicates that these groups of nodes generally belong to the same strongly connected component (see Appendix B). In order to determine the level of connectedness in these strongly connected components, the sub-graph clustering coefficient is calculated.

Subgraph clustering coefficient

The SCC corresponding to the manufacturing network depicted in Fig. 5.3 were calculated and are reported in Appendix B. The subgraph S corresponding to each SCC can be easily identified. S contains all nodes and edges within a specific SCC. Edges connecting nodes in the SCC with other network nodes are not considered. An example corresponding to the SCC containing nodes v_0, v_1, v_2 and v_3 is shown in Fig. 7.1. Note that edge weights have been scaled as required by the clustering coefficient definition presented earlier.

The subgraph clustering coefficient is calculated considering the directed subgraph first as binary and then as weighted. Results are reported in Table 7.2. It can be observed that the binary subgraph clustering coefficient $C_C^D(S)$ is in line with values reported in literature [50] showing moderate structural connectedness. However, in practical terms, even within the strongly connected components, manufacturing is mostly serial as evidenced by the weighted subgraph clustering coefficient $\tilde{C}_C^D(S)$.

This work, thus proposes the use of the average clustering coefficient of SCC instead of that of the full network for the case of manufacturing networks. Manufacturing networks, regardless of their level of connectedness, fulfill the purpose of transforming raw materials into finished goods. While certain flexibility can be granted to groups of production steps, there is a generally prescribed order that

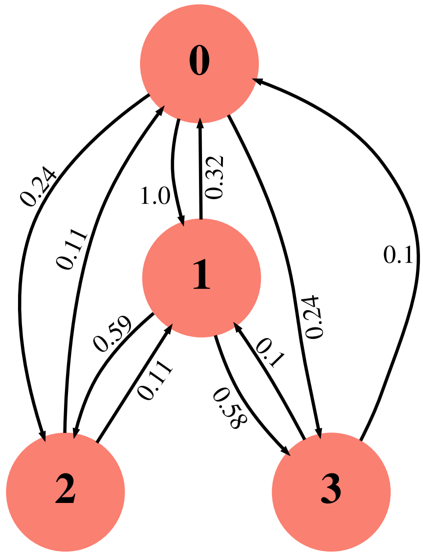


FIGURE 7.1: Subgraph corresonding to the strongly connected component composed of nodes v_0, v_1, v_2 and v_3.

TABLE 7.2: Subgraph clustering coefficient. In this table, the elements in each SCC are indicated using i for clarity. Note that $i \equiv v_i$.

Strongly connected component	$C_C^D(S)$	$\tilde{C}_C^D(S)$
$[0, 1, 2, 3]$	0.42	0.13
$[4, 5, 6, 7, 8, 9, 10, 11]$	0.34	0.04
$[12, 13, 14, 15]$	0.42	0.11
$[16, 17, 18, 19, 20, 21, 22, 23]$	0.34	0.02
$[24]$		
$[25]$		
$[26]$		
$[27]$		
$[28]$		
$[29, 30, 31, 32, 33, 34, 35, 36, 37]$	0.35	0.02
$[38]$		
$[39, 40, 41, 43, 44, 45, 47, 48, 49, 50, 51]$	0.35	0.05

must be followed. As a result, the average clustering coefficient of the full network is expected to be low. But, in cases of cellular manufacturing, groups of production steps will be exchangeable and thus produce highly interconnected subgraphs. Thus, the subgraph average clustering coefficient is a better metric of connectedness for manufacturing networks.

It can then be concluded that,

In a manufacturing network where workstations are abstracted as nodes and directed edges represent material flows, the clustering coefficient is a measure of connectedness. The average clustering coefficient should be calculated for the subgraphs corresponding to each strongly connected component.

The average subgraph clustering coefficient obtained when modeling the subgraph as a binary directed graph provides information regarding the potential *structural connectedness of the network. High values are indicative of high potential flexibility of the production steps involved in the specific SCC. Alternatively, low values are indicative of low potential flexibility.*

The average subgraph clustering coefficient obtained when modeling the subgraph as a weighted directed graph is indicative of the level of connectedness exercised in practice. Once again, high values represent high levels of flexibility in production steps in practice; and low values represent the opposite.

7.6 Triangles in directed graphs

The calculation of the clustering coefficient entails the analysis of the triangles in a directed graph. This analysis has not been carried in literature yet, as evidenced in the literature review presented in Chapter 3. For the sake of completeness, the analysis of the directed triangles in the Bosch manufacturing network of Fig. 5.3 is presented hereafter.

In directed graphs, four patterns of directed triangles can be identified from node v_i's perspective (see Figure 7.2). It should be noted that these triangles have edges pointing in different directions and thus conduce to a completely different interpretation in terms of the resulting flow pattern [83]. The four possible patterns are:

- *cycle*: a cyclical relationship among node v_i and any two neighbors is observed. There are two possible such triangles: $v_i \to v_j \to v_h \to v_i$ or $v_i \to v_h \to v_j \to v_i$.
- *middleman*: one of v_i's neighbors, for example v_j can reach a third node v_h directly ($v_j \to v_h$) or using v_i as intermediary ($v_j \to v_i \to v_h$). There exist two possible such triangles, one in which v_j uses v_i to reach v_h and the other in which v_h uses v_i to reach v_j.
- *in*: node v_i has two incoming edges. There are two possible *in* triangles: one containing the edge $v_j \to v_h$ and the other containing edge $v_h \to v_j$.
- *out*: node v_i has two outgoing edges. There are two possible *out* triangles: one containing the edge $v_j \to v_h$ and the other containing edge $v_h \to v_j$.

7.6.1 Application example

The fraction of each type of triangle present in the network was calculated. Note that the difference between the BDN and WDN is negligible. Thus, only the fractions for the WDN are reported in Table 7.3. It can be observed that the *out* triangle type is the most predominant, while the *cycle* type is encountered the least. This is expected in a manufacturing network where items flow from source to sink. In addition, it is a good sign that the fraction of *middleman* triangles is high. This indicates that there are alternative routes that the flow can follow to surpass blocking nodes.

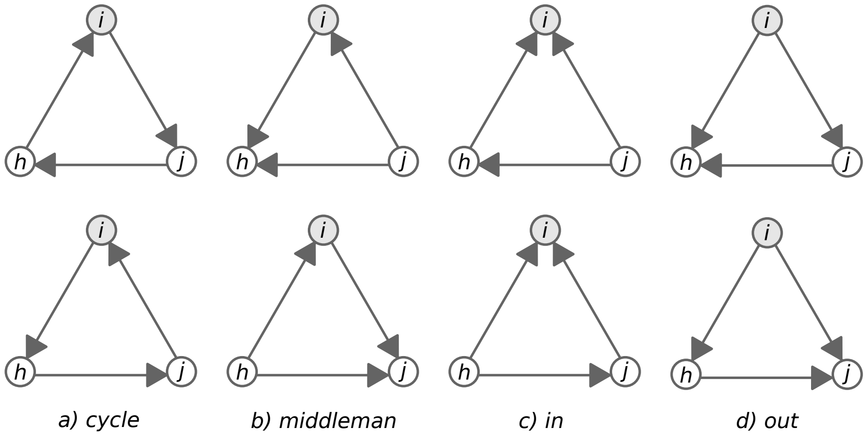


FIGURE 7.2: Types of triangles in directed networks as defined in [83].

TABLE 7.3: Fraction of *cycle*, *middleman*, *in* and *out* triangles present in the Bosch manufacturing network of Fig. 5.3.

Network	cycles	middleman	in	out
WDN	0.151	0.273	0.268	0.308

7.7 Summary

In this Chapter, a number of well known topological metrics for networks have been presented. Their formulation is summarized in Table 7.4, considering variations for different types of networks. The use of these topological metrics for manufacturing networks was also explored. In particular, the metrics interpretation in literature was contrasted with the flow underlying assumptions they are based on, leading to a clear understanding of the information they provide in the context under study. The following conclusions were reached:

- The in-degree is interpreted as the number of direct upstream suppliers. It does not provide information about workloads.
- The out-degree is interpreted as the number of direct downstream customers. It does not provide information about operational load.
- The in-strength is a measure of the incoming workload from suppliers.
- The out-strength is a measure of the difficulty faced by the node to satisfy downstream customer needs.
- The betwenness centrality is considered a poor choice for characterizing manufacturing networks. Its flow underlying assumptions do not match those of manufacturing networks and thus, this metric should be avoided. It does not appropriately identify bottlenecks and/or key machines.
- The clustering coefficient provides a measure of connectedness. For manufacturing networks the average clustering coefficient should be calculated for

the subgraphs corresponding to each strongly connected component. While applying this metric to a binary directed subgraph provides information regarding the potential flexibility of production steps within a SCC, using the weighted directed version measures how much this flexibility is exercised in practice.

TABLE 7.4: Summary of topological metrics

metric	formulation	applicability
degree	$k_i = \sum_j a_{ij} = \sum_j a_{ji}$	BUN, WUN
in-degree	$k_i^{\text{in}} = \sum_j a_{ji}$	BDN, WDN
out-degree	$k_i^{\text{out}} = \sum_j a_{ij}$	BDN, WDN
strength	$s_i = \sum_j w_{ij} = \sum_j w_{ji}$	WUN
in-strength	$s_i^{\text{in}} = \sum_j w_{ji}$	WDN
out-strength	$s_i^{\text{out}} = \sum_j w_{ij}$	WDN
betweenness	$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t\|v)}{\sigma(s,t)}$	(*)[1]
clustering coef.	$C_C(v_i) = \frac{t_i}{T_i} = \frac{2t_i}{k_i(k_i-1)}$	BUN
clustering coef.	$C_C^D(v_i) = \frac{t_i^D}{T_i^D} = \frac{\sum_j \sum_h (a_{ij}+a_{ji})(a_{ih}+a_{hi})(a_{jh}+a_{hj})}{2[k_i(k_i-1)-2k_i^{\leftrightarrow}]}$	BDN
clustering coef.	$\tilde{C}_C^D(v_i) = \frac{\tilde{t}_i^D}{T_i^D} = \frac{\sum_j \sum_h (w_{ij}^{1/3}+w_{ji}^{1/3})(w_{ih}^{1/3}+w_{hi}^{1/3})(w_{jh}^{1/3}+w_{hj}^{1/3})}{2[k_i(k_i-1)-2k_i^{\leftrightarrow}]}$	WDN

[1](*) Different formulations and algorithms exist for different types of networks.

8 PageRank

8.1 Introduction

One of the big inventions of the 20th century was efficient and accurate Web search, namely the "PageRank" algorithm that powers Google's search engine [43, 86]. Before Google, many search engines existed. They worked by listing terms, i.e. words or strings of characters, found in each crawled website in an inverted index (a data structure that makes it easy, given a term, to find all the places where such term occurs). When a search query was issued, the pages that contained the query terms were presented to the user in decreasing order of relevancy and frequency of term appearance. These deprecated search engines had a drawback: they could be easily fooled by *spammers* by simply adding terms unrelated to the actual content of the site, rendering them useless.

To counter "term spam", Google introduced two innovations. Firstly, PageRank simulates the behavior of common Web surfers by following links chosen at random, and the importance of a website is determined by the number of surfers that visit it. Secondly, the content of a page is not judged only by the terms appearing on it but by the terms used in or in the vicinity of links pointing to that page, making it harder for spammers to add unrelated terms. In other words, Google trusts what other websites say about a specif site rather than the site itself.

Although PageRank was created for web ranking, it has been applied in multiple disciplines, including vulnerability assessment of water supply networks [87], ranking scientific publications [88] and evaluating logistics networks [89, 90]. In this Chapter, the PageRank algorithm is adapted to determine workstation importance in a manufacturing line. The objective is to determine this metric suitability and its appropriate interpretation in terms of Operations Research.

8.2 Basic algorithm

The PageRank algorithm is a function that assigns a real number to each node. A node can be a website if we study the World Wide Web, a company when analyzing supply chain networks, or a workstation while dealing with a manufacturing network. The objective is to determine the importance of a node by assigning higher PageRank values to more significant nodes.

PageRank is defined for directed graphs whose transition matrix M describes what happens to a random walker after one step. M has N rows and columns corresponding to the N nodes in the graph. Each element m_{ij} in row i and column j has a value of $1/k_j^{\text{out}}$ if node j has k outgoing edges, and one of them points to node i. Otherwise, $m_{ij} = 0$. Thus, the transition matrix is an stochastic matrix, i.e. the sum of each column equals 1 [91].

The PageRank is calculated iteratively. Initially, a random walker has an homogeneous probability distribution $\mathbf{v}_0$ where $v_{0,i} = 1/N$. After one step, the probability distribution is $\mathbf{v}_1 = M\mathbf{v}_0$. After two steps, $\mathbf{v}_2 = M(M\mathbf{v}_0) = M\mathbf{v}_1$, and so on. Thus, the probability distribution at the next step $\mathbf{v}_{t+1} = M\mathbf{v}_t$ since the probability that a random walker will be at node i in the next step is $v_{t+1,i} = \sum_j m_{ij}\mathbf{v}_{t,j}$. As described before, m_{ij} is the probability that a random walker at node j will move to node i in the next step, and $v_{t,j}$ is the probability that the surfer was at node j in the previous step t.

This process converges to $\mathbf{v}' = M\mathbf{v}$, i.e. further multiplying the probability distribution $\mathbf{v}$ by the transition matrix M does not change the result $\mathbf{v}'$, provided that two conditions are met. Firstly, the graph is strongly connected, meaning that it is possible to get from any node to any other node. Secondly, there are no dead ends, i.e. there are no nodes with zero outgoing edges. In practice however, these strong assumptions are seldomly met.

To avoid such occurrences, the basic PageRank algorithm is modified to account for *taxation*, where it is assumed that a random walker has a finite probability of leaving the network at any step and new walkers are started at each page. The PageRank algorithm considering taxation is as follows

$$\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\frac{\mathbf{e}}{N} \tag{8.1}$$

where $\mathbf{v}'$ and $\mathbf{v}$ are the probability distribution vectors at the new and previous steps, M is a transition matrix, β is a chosen constant (usually in the range between 0.8 and 0.9) that accounts for the random walker's finite probability of leaving the network, $\mathbf{e}$ is a vector of all 1s and N is the number of nodes. The first term of the equation $\beta M\mathbf{v}$ represents the probability β that the walker follows an outgoing edge from the current node, while the second term $(1-\beta)\mathbf{e}/N$ represents the finite probability $(1-\beta)$ of a random walker to start at any given node.

8.3 Proposed modification for manufacturing networks

In order to calculate the PageRank of a manufacturing network, a few modifications to the algorithm are proposed. Firstly, the transition matrix M should no longer be based on degree values, but on material flows. Thus, each element m_{ij} represents the fraction of items leaving node v_i for node v_j, and they are obtained by dividing the item count of edge (v_i, v_j) by the total number of items leaving v_i to guarantee stochasticity as required.

Secondly, taxation is modified to better represent manufacturing networks. While semi-finished goods may leave the network at any step of processing, as is the case with faulty items, raw materials do not start their journey at any random workstation. For that reason, the finite probability of a random walker starting at any given node must be modified as follows:

$$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\frac{\mathbf{e}_S}{t_m} \tag{8.2}$$

where $\mathbf{e}_S$ is a vector that contains item counts in the workstations where manufacturing can commence and 0s in the rest. t_m is total number of manufactured items.

8.4 Results

Three different formulations of the PageRank algorithm were considered for the manufacturing network of Fig. 5.3. These formulations, summarized in Table 8.1, are:

- $\mathbf{PR_1}$: The original PageRank algorithm with taxation. No modifications considered.
- $\mathbf{PR_2}$: The PageRank algorithm with taxation where edge weights are considered, i.e. the transition matrix is based on material flows. However, the taxation portion $(1-\beta)\mathbf{e}/N$ is not modified and thus, a random walker can jump to any other node in the network.
- $\mathbf{PR_3}$: The PageRank algorithm with taxation considering all modifications listed in the previous Section to better resemble the manufacturing process. Thus, not only edge weights are considered, also the vector $\mathbf{e}_S$ has values different from zero only on nodes corresponding to starting points in the manufacturing network. Thus, when a random walker "jumps", new walkers can only be initiated on nodes where manufacturing starts.

TABLE 8.1: Summary of PageRank formulations evaluated.

case	formulation	notes
PR_1	$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\frac{\mathbf{e}}{N}$	M is based on the out-degree. New walkers can be initialized on all nodes.
PR_2	$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\frac{\mathbf{e}}{N}$	M is based on the out-strength. New walkers can be initialized on all nodes.
PR_3	$\mathbf{v}' = \beta M \mathbf{v} + (1-\beta)\frac{\mathbf{e}_S}{t_m}$	M is based on the out-strength. New walkers can only be initialized on manufacturing start nodes.

The results obtained from calculating the PageRank on the Bosch manufacturing network of Fig. 5.3 are reported in Table 8.2 and Fig. 8.1. In all cases, the value of β was set to 0.85.

TABLE 8.2: Results obtained from calculating the PageRank for three different cases of the Bosch manufacturing network reported in Fig. 5.3

v_i	PR_1	PR_2	PR_3
0	0.004	0.005	0.077
1	0.004	0.007	0.07
2	0.004	0.006	0.031
3	0.004	0.006	0.031
4	0.009	0.008	0.026
5	0.009	0.008	0.026
6	0.01	0.011	0.02
7	0.01	0.011	0.02
8	0.011	0.021	0.035
9	0.01	0.009	0.011
10	0.01	0.008	0.013
11	0.01	0.008	0.013
12	0.005	0.005	0.027
13	0.005	0.007	0.024
14	0.005	0.006	0.011
15	0.005	0.006	0.011
16	0.012	0.007	0.009
17	0.012	0.008	0.009
18	0.01	0.01	0.007
19	0.01	0.01	0.007
20	0.012	0.02	0.012
21	0.011	0.008	0.004
22	0.011	0.008	0.004
23	0.011	0.008	0.004
24	0.003	0.003	0.024
25	0.003	0.003	0.011
26	0.011	0.008	0.021
27	0.011	0.006	0.01
28	0.005	0.003	0.001
29	0.027	0.051	0.06
30	0.042	0.049	0.053
31	0.018	0.004	0.001
32	0.018	0.004	0.001
33	0.055	0.079	0.065
34	0.054	0.086	0.067
35	0.036	0.047	0.034
36	0.036	0.051	0.037
37	0.038	0.09	0.064
38	0.017	0.011	0.005
39	0.028	0.006	0.002
40	0.015	0.009	0.002
41	0.021	0.011	0.001
43	0.017	0.007	0.001
44	0.017	0.007	0.001
45	0.029	0.015	0.001
47	0.058	0.031	0.002
48	0.032	0.041	0.002
49	0.03	0.025	0.001
50	0.03	0.028	0.001
51	0.049	0.052	0.002

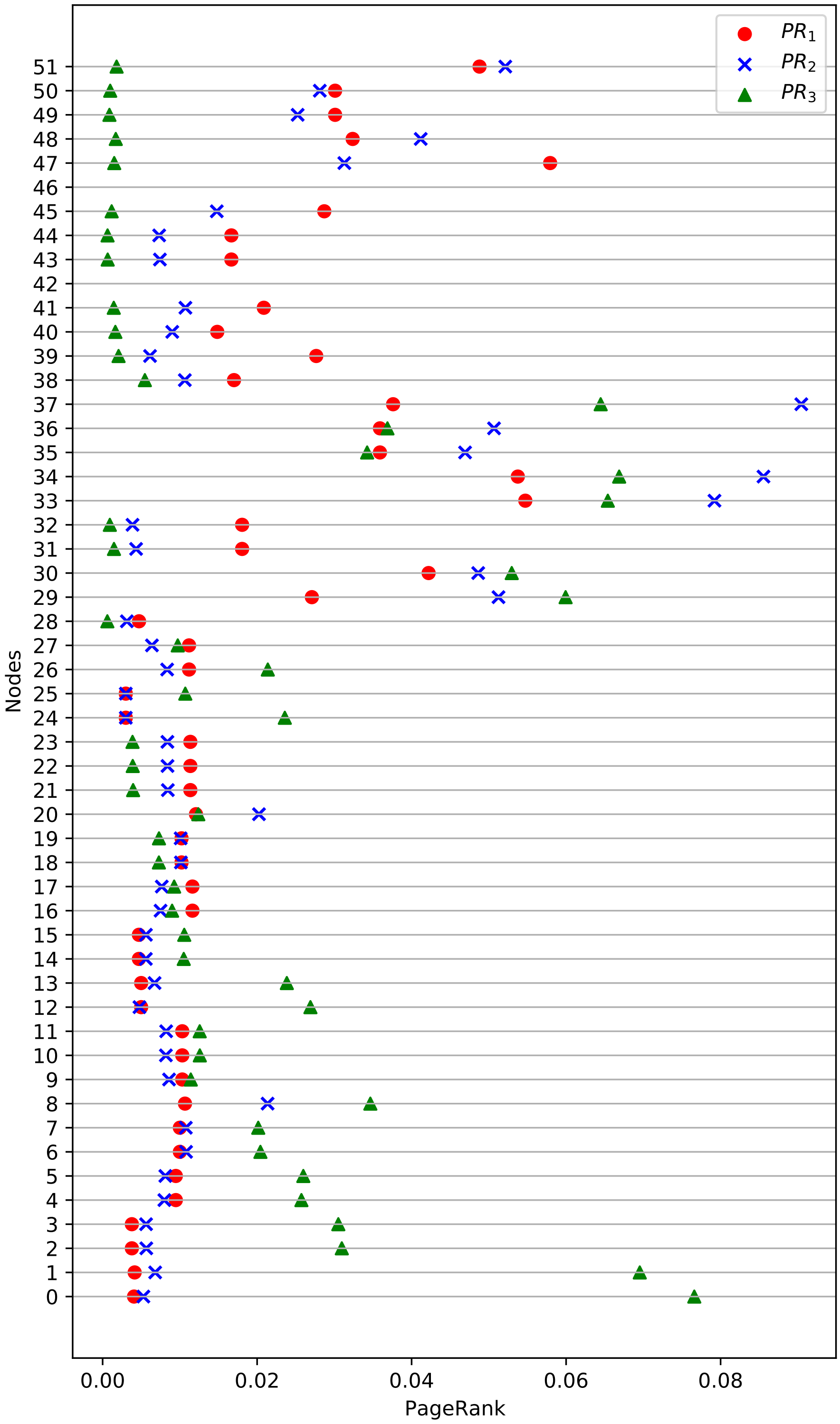


FIGURE 8.1: PageRank value for each node in the Bosch manufacturing network. Note that nodes v_{42} and v_{46} do not belong to the network. See Chapter 5 for details.

8.4.1 Analysis by case

$\mathbf{PR_1}$: In Fig. 8.1, the original PageRank formulation is denoted as PR_1 and indicated by big red dots. It can be observed that groups of upstream nodes have roughly the same value of PR_1. These groups corresponds to upstream strongly connected components (see Appendix B). For example, note the upstream SCC containing nodes $[v_0, v_1, v_2, v_3]$, where all nodes have $PR_1 = 0.004$. It follows that the PR_1 formulation assigns roughly the same PageRank value to the members of each upstream strongly connected component.

This does not hold true, however, for downstream nodes which have much more variable PageRank values. Note, for example, the SCC containing nodes $[v_{29}, v_{30}, v_{31}, v_{32}, v_{33}, v_{34}, v_{35}, v_{36}, v_{37}]$, where the value of PR_1 is highly variable within the range of 0.018 to 0.055. In general, this formulation tends to boost ending nodes centrality since their k_i^{in} tends to be higher than their k_i^{out} (see Table 7.1). It is worth remembering that this was indeed the objective of the PageRank algorithm: to identify nodes to which many other nodes point to and declare them important.

$\mathbf{PR_2}$: The first modification proposed to the original algorithm involves the use of the out-strength (i.e. the actual material flows) in the transition matrix instead of the out-degree. Then the transition matrix is normalized by column to fulfill the stochasticity requirement. This formulation was denoted as PR_2 and is indicated by blue crosses in Fig. 8.1.

In general, this formulation does not seem to have a great effect on upstream nodes. Note for example the nodes within the upstream SCC $[v_0, v_1, v_2, v_3]$, where the value of PR_2 ranges from 0.005 to 0.007, compared to 0.004 for PR_1. There are however, the notable exceptions of nodes v_8 and v_{20}. These exceptions can be explained by the fact that these nodes process a higher proportion of goods than the other nodes within the same SCC (see s_i^{in}/t_m or s_i^{out}/t_m in Table 7.1).

The situation is different for downstream nodes. The PR_2 formulation boosts the PageRank of highly transited downstream nodes and reduces that of less frequent ones. Note for example the nodes contained in the SCC $[v_{29}, v_{30}, v_{31}, v_{32}, v_{33}, v_{34}, v_{35}, v_{36}, v_{37}]$. Nodes v_{31} and v_{32} have a low workload (see s_i^{in}/t_m or s_i^{out}/t_m in Table 7.1) and thus $PR_2 < PR_1$ for them. On the other hand, the rest of the nodes in this SCC have a high workload (see s_i^{in}/t_m in Table 7.1)[1]and thus $PR_2 > PR_1$ for them.

In conclusion, while this formulation makes corrections accounting for actual material flows, once again the highest centrality values is assigned to ending nodes.

[1]In this case, the reader should assess the workload looking into s_i^{in}/t_m in Table 7.1 given that nodes v_{33}, v_{34}, v_{35}, v_{36} and v_{37} are ending nodes. As a result, $s_i^{\text{in}}/t_m > s_i^{\text{out}}/t_m$ for this nodes, and thus s_i^{in}/t_m is a better metric of workload than s_i^{out}/t_m.

PR$_3$: In addition to using material flows for the transition matrix, a second modification to the original formulation was proposed. This entailed allowing random walkers to be started only on nodes where manufacturing can begin. This formulation, called PR_3, is indicated as green triangles in Fig. 8.1.

The resulting PageRank is a very interesting. Starting nodes have their centrality values boosted as a result of the newly formulated taxation term. However, not all of them are boosted equally since $\mathbf{e}_S/t_m$ is based on the actual number of manufactured items processed by each starting node. Its combination with a material flows based transition matrix slashes the centrality values of nodes within the biggest, yet least transited, downstream SCC $[v_{39}, v_{40}, v_{41}, v_{43}, v_{44}, v_{45}, v_{47}, v_{48}, v_{49}, v_{50}, v_{51}]$. As a result, the highest centrality values are held by a combination of frequent starting, middle and ending nodes.

8.5 Discussion

As shown in the literature review presented in Chapter 3, a number of authors have used the PageRank algorithm when studying manufacturing networks with varied objectives. Blunck *et. al.* [51] used PageRank for bottleneck identification under capacity increase and batching alleviation. However, the authors reported that traditional methods for bottleneck identification outperformed PageRank. They suggested that the underlying assumptions of PageRank may not properly abstract manufacturing networks.

Other authors reported more positive outcomes. Ai-ming *et. al.* [52] modified the PageRank algorithm to generate a ranking of workstations that should be prioritized in terms of quality control and improvement efforts. Becker and Wagner [53] showed that the PageRank algorithm could be used to create a ranking of machines in order to efficiently equip a production line with the appropriate technology to be converted into a cyber-physical production system. Finally, Omar *et. al.* [58] suggested that PageRank could rank workstations by importance based on effective processing paths. They went on to claim that "the node importance [obtained from PageRank] measures the workload build-up of a node while accounting for interdependencies among pairs of nodes" [58].

As explained in Section 2.3, centrality metrics can be classified using Borgatti's typology (see Table 2.1). This typology evaluates two characteristics of the flow: the routes followed and the propagation method. In the case of PageRank, it was clearly stated in its definition that random walkers are followed on each time step. Thus, the traffic flows through *walks*. In addition, the traffic propagates through *parallel duplication* [92]. It can be noted that PageRank shares similarities with eigenvector centrality. Not only, both PageRank and eigenvector centrality have the same Borgatti typology, but also they are both obtained from matrix operations providing an exact result. In fact, the PageRank creators, Brin and Page, explained that "PageRank [...] corresponds to the principal eigenvector of the normalized link matrix" [43].

As a consequence, PageRank is not very well suited to evaluate manufacturing networks since these correspond to the *paths* and *transfer process* cell in Borgatti's typology. Thus, practitioners are advised against using PageRank as a centrality metric for the characterization and analysis of manufacturing networks where nodes represent workstations and (weighted) directed edges, material flows. Consequently, an interpretation for this metric in terms of Operations Research is not provided.

8.6 Summary

In this Chapter, the use of the PageRank algorithm is exemplified and analyzed in the context of manufacturing networks. It is concluded that, while the metric has been applied in previous research in this domain, it is not appropriate. PageRank is a centrality metric whose underlying flow follows *walks* and whose traffic propagates by means of *parallel duplication*. As a result, PageRank occupies an entirely different cell in Borgatti's typology when compared to the underlying flow of manufacturing networks. It is thus considered inappropriate for characterization of these networks and consequently, no interpretation in terms of Operations Research is provided.

9 Path-Transfer Entropy

The concept of information theory originated in the need to quantify the fundamental limits of signal processing. Typically attributed to Shannon [93], information entropy quantifies the average number of bits needed to store or communicate a message. A message with N different symbols cannot be stored or communicated in less than $\log_2 N$ bits. Shannon's entropy determines a lower limit below which no message can be further compressed. In addition, Shannon's information theory has also been regarded as a measure to quantify uncertainty, or entropy, in a system [94]. It allows to quantify the uncertainty involved in predicting the value of a random variable, i.e. the amount of randomness or freedom of choice. It is defined as follows:

Definition 1. *For an ensemble $X(R, p_{x_i})$, where R is the set of possible outcomes (the random variable), $N = |R|$ and p_{x_i} is the probability of an outcome in R. The Shannon information content or entropy of X is given by*

$$H(X) = -\sum_{i=1}^{N} p_{x_i} \log_2 p_{x_i} \tag{9.1}$$

where calculating $H(X)$ requires the mass distribution probability of ensemble X.

9.1 Path-transfer entropy

Tutzauer [40] proposed an entropy centrality metric called "path-transfer entropy". Based on Shannon's entropy [93], Tutzauer's path-transfer entropy is valid under the following assumptions regarding the network flow:

- The flow follows *paths*, where a path is a sequence of linked nodes in which neither nodes nor edges are repeated.
- The flow progresses by means of a *transfer process*, that is an item can only be in one place at a time and it moves from one node to another in the network sequentially.

As a result, Tutzauer's centrality metric occupies a specific cell in Borgatti's typology (see Section 2.3).

9.1.1 Formula derivation

To calculate Tutzauer's path-transfer entropy, the mass probability distribution p_{x_i} in Eq. 9.1 has to be calculated first. The procedure to derive the appropriate probability distribution, called p_{ij} in Tutzauer's work, is described hereafter.

As noted in the previous Section, Tutzauer assumes the traffic progresses by means of a transfer process [40, 95]. Thus, a node receiving the transfer can pass the object to any adjacent node (that is to say, a neighboring node with which it shares an edge). It must be noted that the chosen adjacent node must not have yet appeared on the path up to this point, since paths do not allow for neither node nor edge repetitions. Using the convention that both v_t and $v(t)$ are valid and equivalent notations to represent a node, the following terms are defined:

Definition 2. *For a given path $P = \{v_0, v_1, ..., v_n\}$, a **downstream edge** of a vertex $v_t \in P$ is any edge $(v_t, w) \in E$ with $w \neq v_s$ for $s < t$.*

Definition 3. *A **downstream vertex** of v_t is any vertex w such that (v_t, w) is a downstream edge.*

Definition 4. *The **downstream degree** $D(v_t)$ of a vertex v_t is the number of downstream edges it has and can be calculated from the adjacency matrix as follows:*

$$D(v_t) = \sum_{j=1}^{N} a_{v(t),j} - \sum_{s<t} a_{v(t),v(s)}. \tag{9.2}$$

The term $\sum_{j=1}^{N} a_{v(t),j}$ counts the total number of edges incident with v_t, i.e. its degree. The term $\sum_{s<t} a_{v(t),v(s)}$ subtracts from that total the number of edges incident with vertices that have already appeared earlier in the path.

At any step of the process, the probability that a transfer is passed to a downstream node is equal to 1 divided by the downstream degree. Likewise, the probability that a node stops the flow is 1 divided by the downstream degree. Thus, the transfer and stopping probabilities are defined as follows:

Definition 5. *In general, there are $K(i,j)$ paths from i to j. Let P_k be one of such paths, assumed to be of length n_k. The **transfer probability** of a vertex $v_t \in P_k$ is:*

$$\tau_k(v_t) = \frac{1}{D(v_t)} \tag{9.3}$$

*and the **stopping probability** is*

$$\sigma_k(v_t) = \frac{1}{D(v_t)}. \tag{9.4}$$

To obtain the single path probability, i.e., the likelihood that a flow beginning at $i = v_0$ ends at $j = v(n_k)$ by traveling along the path $P_k = \{v_0, ..., v(n_k)\}$, one must calculate the product of the transfer probability of the first $n_k - 1$ nodes and multiply it by the stopping probability of the last vertex in the path. Then,

Definition 6. *The overall probability that a flow starting at i ends at j is given by the combined path probability:*

$$p_{ij} = \sum_{k=1}^{K(i,j)} \sigma_k(j) \prod_{t=0}^{n(k)-1} \tau_k(v_t). \tag{9.5}$$

Finally, the path-transfer centrality $C_H(i)$ of vertex i is then given by the entropy:

$$C_H(i) = -\sum_{j=1}^{N} p_{ij} \log_2 p_{ij}. \tag{9.6}$$

Furthermore, to put the centrality score on a zero-to-one scale, it must be divided by the maximum entropy which is well known to be $\log_2 N$. Hence the relative centrality of vertex i is given by:

$$C'_H(i) = \frac{C_H(i)}{\log_2 N}. \tag{9.7}$$

9.1.2 Generalizations

The formulae in the previous Section were derived for connected undirected, unweighted graphs. Here we present the generalizations to other types of networks.

Directed graphs. Non-symmetric relationships among nodes are represented as directed graphs. However, the formulae for undirected graphs defined above carries over to directed graphs with no changes.

Loopless vertices. In the original description, flow could be stopped in two ways: the lack of downstream vertices and the vertex choosing itself as downstream vertex. When nodes cannot elect to stop the flow, i.e. there are no (v_t, v_t) edges, vertices are said to be loopless. In such a case, the appropriate transfer and stopping probabilities are:

$$\tau_k(v_t) = \begin{cases} 0 & D(v_t) = 0 \\ \frac{a_{v(t),v(t+1)}}{D(v_t)} & D(v_t) \neq 0 \end{cases} \tag{9.8}$$

$$\sigma_k(v_t) = \begin{cases} 1 & D(v_t) = 0 \\ \frac{a_{v(t),v(t)}}{D(v_t)} & D(v_t) \neq 0 \end{cases} \tag{9.9}$$

Weighted graphs. Using the convention that the adjacency matrix $\mathcal{A} = \{a_{ij}\}$ which contains only zeroes and ones is replaced by the weight matrix $\mathcal{W} = \{w_{ij}\}$ which contains the weights of the edges instead, the equations presented carry through directly. It must be noted that the downstream degree $D(v_t)$ is now more of a measurement of downstream strength since it is the sum of the weights of downstream edges.

Disconnected graphs. The entropy centrality is calculated for all non-zero combined path probabilities, i.e. $p_{ij} \neq 0$. To extend the calculation to disconnected networks, the centrality equation is updated to

$$C_H(i) = -\sum_{\substack{j \in V \\ p_{ij} \neq 0}} p_{ij} \log_2 p_{ij}. \tag{9.10}$$

Summary of the most general entropy centrality equations

Downstream degree:

$$D(v_t) = \sum_{j=1}^{N} a_{v(t),j} - \sum_{s<t} a_{v(t),v(s)}$$

Transfer and stopping probabilities:

$$\tau_k(v_t) = \begin{cases} 0 & D(v_t) = 0 \\ \frac{a_{v(t),v(t+1)}}{D(v_t)} & D(v_t) \neq 0 \end{cases} \qquad \sigma_k(v_t) = \begin{cases} 1 & D(v_t) = 0 \\ \frac{a_{v(t),v(t)}}{D(v_t)} & D(v_t) \neq 0 \end{cases}$$

Probability of path p_{ij}:

$$p_{ij} = \sum_{k=1}^{K(i,j)} \sigma_k(j) \prod_{t=0}^{n(k)-1} \tau_k(v_t)$$

Path-transfer centrality:

$$C_H(i) = -\sum_{\substack{j \in V \\ p_{ij} \neq 0}} p_{ij} \log_2 p_{ij} \qquad C'_H(i) = \frac{C_H(i)}{\log_2 N}$$

9.2 Path-transfer entropy and manufacturing

As explained above, Tutzauer's path-transfer entropy occupies a specific cell of Borgatti's typology of centrality metrics [38] (see Section 2.3). It is then appropriate for networks where the flow follows *paths* and progresses by means of a *transfer process*. These underlying assumptions perfectly describe the flow in manufacturing networks.

The next pressing questions is "what does path-transfer entropy centrality tells us about a network?". The limited existing literature provides vague answers. Tutzauer claimed that "flow beginning at a highly central node will stop with nearly equal probability at all other nodes, and a flow beginning at a less central node will have a much more uneven distributions of probabilities" [40]. Similarly, Oggier *et. al.* [85, 96] described the spread of the flow as more even when originating from a highly central node. This metric interpretation is simply the re-wording of one very well know property held by the Shannon's information entropy formulation: the entropy increases as the probability values p_{x_i} in Eq. 9.1 become equal and reaches its maximum when all the p_{x_i} are exactly equal [93, 97]. Yet, these authors do not address a pressing fact. As the number of nodes and edges in graph G increases, the number of paths grow combinatorially [97]. Consequently, utilizing methods that search for all the paths in the network provides results with very little to no differenciation between nodes in terms of centrality [98].

Given that literature seems to pay special attention to the probability distribution p_{ij} instead of the node entropy centrality $C_H(i)$, this Chapter will focus on answering the following questions:

- What is the meaning of p_{ij} in terms of Operations Research?
- How does p_{ij} change for binary and weighted manufacturing networks?
- What information does path-transfer entropy centrality provides in the context of manufacturing networks?

9.3 Methodology

In this Chapter, the manufacturing complex network used was obtained following the methodology presented in Chapter 5. In addition, one further step is required. To account for the stopping probability on final manufacturing nodes, self-loops must be added. These self-loops exist, in principle, only for the nodes where manufacturing can conclude. Their weight is determined by the frequency with which a given node is the final node in the list of clean manufacturing paths. The resulting network, shown in Fig. 9.1, contains 50 nodes and 319 edges.

In order to answer the questions posed in the previous Section, the probability distribution p_{ij} will be calculated for two different representations of the graph:

Binary, directed graph: Firstly, a structural analysis of the network is conducted. For this, the binary, directed network is used. In order to capture the effect of the flow on the vertices, a mathematical artifact is needed. Thus, self loops are added to all nodes [85]. The graph contains 50 nodes and 357 edges. Then, the corresponding probability distribution p_{ij} is calculated and analyzed.

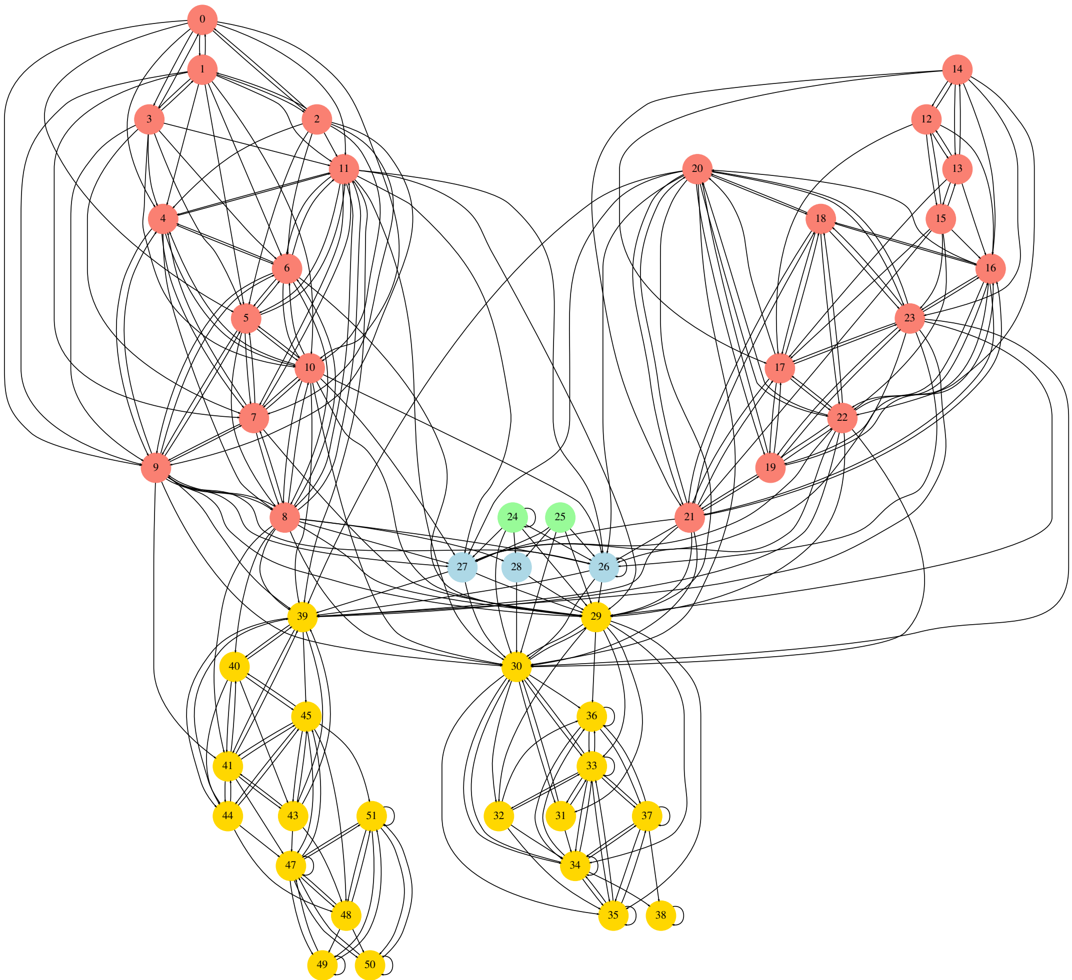


FIGURE 9.1: Bosch manufacturing network including self-loops on nodes where manufacturing can conclude only. The color of the nodes follows the manufacturing line colors in Fig. 5.1.

Weighted, directed graph: Secondly, analysis of the original weighted network with the added self-loops on ending nodes only is carried out. The graph, shown in Fig. 9.1, has 50 nodes and 319 edges.

In both cases, the computation of all the paths in the network and the paths probabilities was carried out using the UL High Performance Computer (HPC) Iris Cluster [99]. For this work, a single node containing 28 cores running at 2.4 GHz with 112 GB RAM was used. The code was implemented in parallel, i.e. each core was assigned the task of obtaining all the p_{ij} values for node v_i.

This work shares similarities with Omar *et. al.* [63]. The interested reader should refer to the cited article and its associated code available in [74].

9.4 Results

In this Section, the results obtained when calculating the probability distribution p_{ij} for the binary and weighted networks are presented and discussed.

9.4.1 Binary, directed graph with self-loops on all nodes

The probability distribution p_{ij} for all the nodes in the network under study are presented graphically in Fig. 9.2. In addition, the entropy centrality values for each node v_i are reported in Table 9.1. Hereafter, a number of observations are noted.

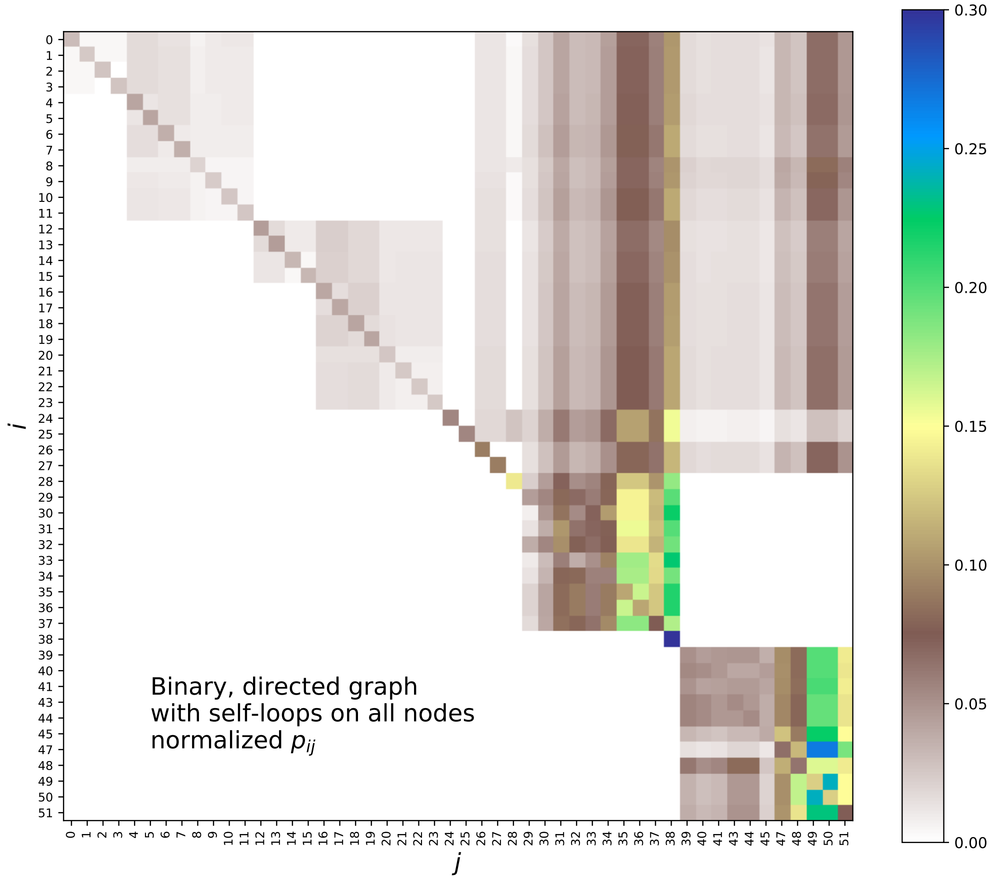


FIGURE 9.2: Heat-map denoting the probability distribution p_{ij} obtained for the binary directed network with self-loops on all nodes. *Note:* the colorbar has been chosen to facilitate visualization. It should be noted that $p_{38,38} = 1$.

Much as in the case of traditional serial manufacturing, upstream nodes can reach downstream ones, yet the reverse is not possible. To better visualize this, nodes where manufacturing can start were identified and are shown in the left graph

TABLE 9.1: Entropy centrality values for the binary, directed network with self-loops on all nodes.

<table>
<tr><th>v_i</th><th>$C_H(i)$</th><th># nodes in SCC</th><th>$\overline{C_H}$</th><th>σ_{C_H}</th></tr>
<tr><td>0</td><td>0.835</td><td rowspan="4">4</td><td rowspan="4">0.833</td><td rowspan="4">0.001</td></tr>
<tr><td>1</td><td>0.834</td></tr>
<tr><td>2</td><td>0.832</td></tr>
<tr><td>3</td><td>0.832</td></tr>
<tr><td>4</td><td>0.815</td><td rowspan="8">8</td><td rowspan="8">0.810</td><td rowspan="8">0.004</td></tr>
<tr><td>5</td><td>0.815</td></tr>
<tr><td>6</td><td>0.811</td></tr>
<tr><td>7</td><td>0.811</td></tr>
<tr><td>8</td><td>0.806</td></tr>
<tr><td>9</td><td>0.807</td></tr>
<tr><td>10</td><td>0.806</td></tr>
<tr><td>11</td><td>0.806</td></tr>
<tr><td>12</td><td>0.851</td><td rowspan="4">4</td><td rowspan="4">0.848</td><td rowspan="4">0.002</td></tr>
<tr><td>13</td><td>0.851</td></tr>
<tr><td>14</td><td>0.846</td></tr>
<tr><td>15</td><td>0.846</td></tr>
<tr><td>16</td><td>0.821</td><td rowspan="8">8</td><td rowspan="8">0.817</td><td rowspan="8">0.004</td></tr>
<tr><td>17</td><td>0.821</td></tr>
<tr><td>18</td><td>0.820</td></tr>
<tr><td>19</td><td>0.820</td></tr>
<tr><td>20</td><td>0.812</td></tr>
<tr><td>21</td><td>0.814</td></tr>
<tr><td>22</td><td>0.814</td></tr>
<tr><td>23</td><td>0.814</td></tr>
<tr><td>24</td><td>0.722</td><td>1</td><td></td><td></td></tr>
<tr><td>25</td><td>0.722</td><td>1</td><td></td><td></td></tr>
<tr><td>26</td><td>0.741</td><td>1</td><td></td><td></td></tr>
<tr><td>27</td><td>0.741</td><td>1</td><td></td><td></td></tr>
<tr><td>28</td><td>0.580</td><td>1</td><td></td><td></td></tr>
<tr><td>29</td><td>0.561</td><td rowspan="9">9</td><td rowspan="9">0.547</td><td rowspan="9">0.012</td></tr>
<tr><td>30</td><td>0.542</td></tr>
<tr><td>31</td><td>0.546</td></tr>
<tr><td>32</td><td>0.564</td></tr>
<tr><td>33</td><td>0.522</td></tr>
<tr><td>34</td><td>0.539</td></tr>
<tr><td>35</td><td>0.550</td></tr>
<tr><td>36</td><td>0.550</td></tr>
<tr><td>37</td><td>0.546</td></tr>
<tr><td>38</td><td>0</td><td>1</td><td></td><td></td></tr>
<tr><td>39</td><td>0.564</td><td rowspan="11">11</td><td rowspan="11">0.549</td><td rowspan="11">0.029</td></tr>
<tr><td>40</td><td>0.565</td></tr>
<tr><td>41</td><td>0.562</td></tr>
<tr><td>43</td><td>0.569</td></tr>
<tr><td>44</td><td>0.569</td></tr>
<tr><td>45</td><td>0.534</td></tr>
<tr><td>47</td><td>0.469</td></tr>
<tr><td>48</td><td>0.583</td></tr>
<tr><td>49</td><td>0.543</td></tr>
<tr><td>50</td><td>0.543</td></tr>
<tr><td>51</td><td>0.540</td></tr>
</table>

of Fig. 9.3. Take for example $v_i = v_0$, which corresponds to the most frequent starting node, and note from its p_{ij} in Fig. 9.2 that it can reach any node v_j except for j values between 12 and 25 inclusive. On the other hand, nodes that act predominantly as end nodes (see right graph in Fig. 9.3), can only reach a few nodes that are also downstream. For example, taking $v_i = v_{37}$ means that only nodes v_j with j between 29 and 37 inclusive can be reached.

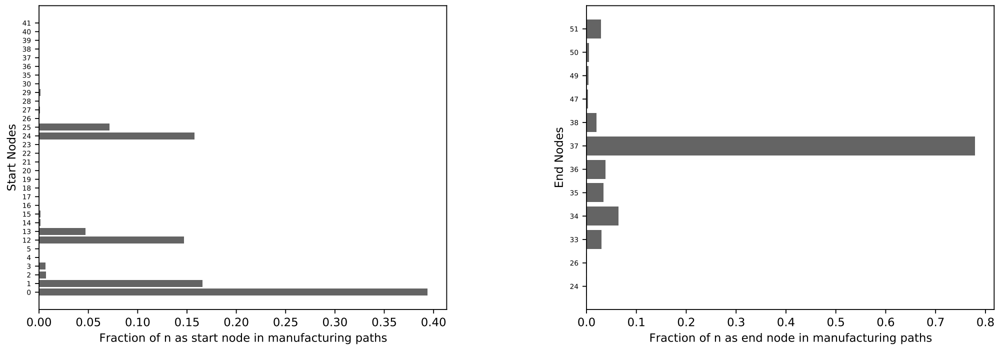


FIGURE 9.3: *Left:* Nodes that appear as start nodes in the clean manufacturing paths with their corresponding fraction. *Right:* Nodes that appear as end nodes in the clean manufacturing paths with their corresponding fraction.

It should also be noted that a great portion of $p_{ij} = 0$. These are depicted as white cells in the heat-map of Fig. 9.2. From Eq. 9.5, it is clear that p_{ij} is different from zero when the set of paths $K(i, j)$ between nodes v_i and v_j is non empty. Thus $p_{ij} \neq 0$ if and only if there is at least one path from v_i to v_j. Then, a great number of $p_{ij} = 0$ is expected in a manufacturing network. Even in a highly flexible manufacturing line, there is a source of raw materials and a sink of manufactured products. In addition, not all processing steps can be interchanged. There is always some required order of precedence [100].

Yet another aspect to notice in Fig. 9.2, is that groups of nodes that are reachable from each other form clusters. These clusters are the strongly connected components (SCC) described in Appendix B. Furthermore, all nodes within a SCC have similar values in terms of entropy centrality as reported in Table 9.1. This similarity in $C_H(i)$ has already been highlighted in literature and attributed to the fact that this entropy metric requires searching for all paths in the network [98]. Note that the number of paths in a network increases as the number of nodes and edges increases [97] leading to a loss in metric sensitivity. This, in turn, traslates to increased difficulty in differentiating nearby neighbors. As a result, nodes within a SCC have very similar $C_H(i)$ values.

The existence of SCC and the similarity in entropy centrality values provides valuable information regarding the structure of the manufacturing network in terms of its operation and routing flexibility. One could argue that the ordering of process steps within a SCC can be altered, yet the sequence in which the SCC are visited cannot. This points to a manufacturing network that shares properties of flexible manufacturing [100]: the system can potentially alter the order of a number of manufacturing operations, yet a required partial precedence structure exists.

Finally, the fact that the probability distribution p_{ij} for any node v_i in the network is uneven in j is mention worthy. This holds true within SCC as well. As a consequence, entropy is not maximized and there is some certainty as to what the final destination of the flow starting in node v_i is [63]. Following the previous example and from observation of Fig. 9.2, one can confidently say that flow starting on node $v_i = v_0$ has a high probability of stopping on node $v_j = v_{37}$. And flow departing from node $v_i = v_{37}$ will stop with high likelihood on node v_j with j either 34, 35 or 37. Once again, this points to some predefined or likely path in this manufacturing network. In fact, in literature [63], this uneven distribution of probability is said to be accentuated for serial production lines.

9.4.2 Weighted, directed graph with self-loops on end nodes only

In the previous Section, the binary graph was considered. In this one, the weighted network with self-loops on end nodes only will be utilized. While binary graphs are used to study the network structure, weighted networks provide information related to “the skew in strength of the relationships” [85]. Thus, if the binary graph pointed towards the level of manufacturing flexibility attainable by this network, the weighted graph will indicate the degree to which said flexibility is exercised.

The probability distribution p_{ij} for all the nodes in the network under study are presented graphically in Fig. 9.4. Contrary to the structural analysis in the previous Section, the level of exercised flexibility is low. In fact, the following can be concluded:

- For any node v_i where $i \leq 37$, manufacturing will likely conclude in node $v_j = v_{37}$.
- For node $v_i = v_{38}$, manufacturing concludes with absolute certainty in node $v_j = v_{38}$ since $p_{38,38} = 1$.
- For any node v_i where $i > 38$, manufacturing will likely conclude in node $v_j = v_{51}$.

9.5 Note on computational issues

The calculation of Shannon’s entropy entails a choice regarding the level of granularity of the analysis [94]. This follows from it being a metric requiring the counting of discrete elements or events. In other words, even though the formulae necessary to calculate the entropy centrality value of a node is quite straight-forward, the computation requires to find all paths emanating from the node under study [40]. In

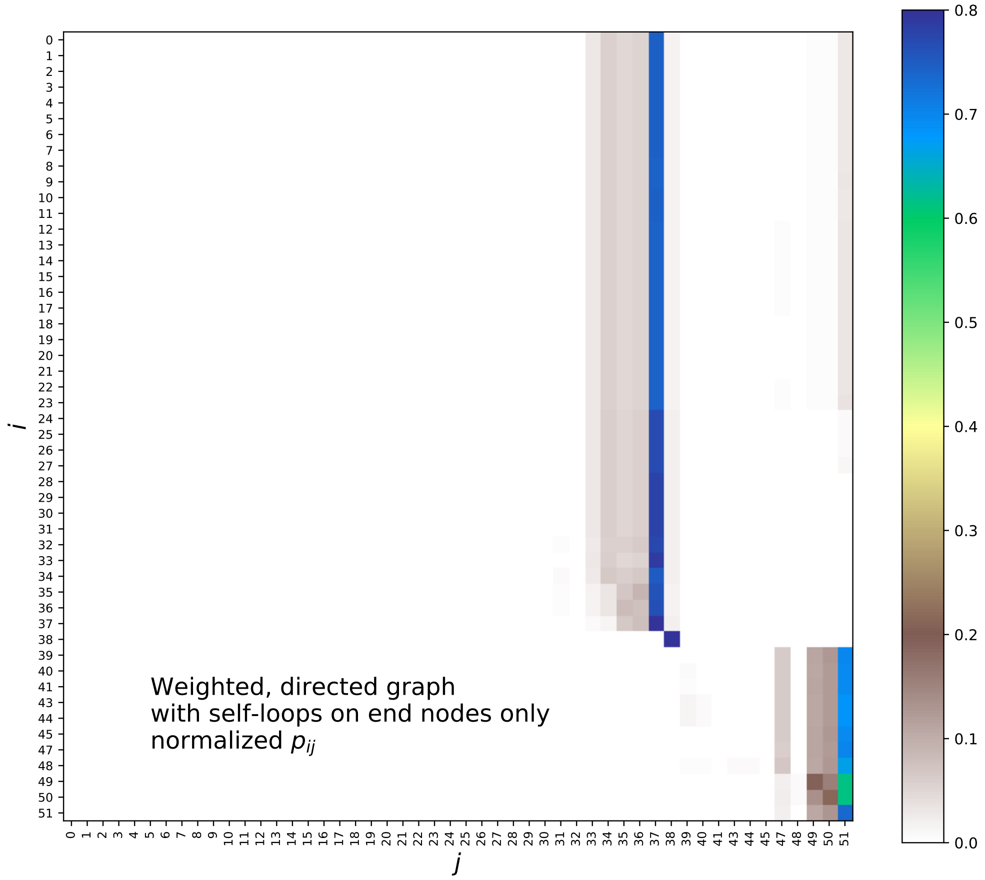


FIGURE 9.4: Heat-map denoting the probability distribution p_{ij} obtained for the weighted directed network with self-loops on end nodes only. *Note:* the colorbar has been chosen to facilitate visualization. It should be noted that $p_{37,37} = 0.8103$ and $p_{38,38} = 1$.

order to rank nodes by means of their centrality measure, the entropy centrality of each node must be calculated and thus, every path in graph G must be found. For large and dense graphs (where the number of edges is close to N^2), the number of paths is exceedingly large and computationally prohibitive.

Tutzauer [40, 95] as well as several other authors [63, 85, 96–98] agree on the fact that searching for all paths in a graph is computationally prohibitive, particularly for moderate and large graphs. A commonly proposed solution is to calculate the path probability as the path is found. Paths should then be pruned when their probability falls below a user defined threshold, typically set to 10^{-6}. The implementation in this work does not make use of path prunning. Readers interested in path pruning should request a working implementation from Oggier *et. al.* [96].

The implementation developed for this work, searches for all paths emanating from node v_i using a generator function [101] in Python 3 to guarantee speed and low memory usage. Furthermore, the probabilities p_{ij} are calculated in parallel for each v_i using the High Performance Computer (HPC) Cluster in the University of Luxembourg [99]. Finally, the PyPy compiler [102] is chosen instead of CPython for speed. Code profiling[1] indicates that, in its current stage, the most time consuming routine in the code is not the path search. It is the calculation of the downstream degree as shown in Fig. 9.5.

9.6 Discussion

Tutzauer's path-transfer entropy is the only metric thus far discussed whose underlying flow assumptions match the flow of manufacturing networks. In fact, Tutzauer explicitly pointed out that his metric aimed at filling an empty cell in Borgatti's typology table [38, 40]. Yet, while the path-transfer entropy formulation was derived by Tutzauer in 2007 [40], its use to characterize manufacturing networks is quite novel. In fact, in the literature review presented in Chapter 3, there is only one work discussing it.

Omar and Plapper [63] studied the probability distribution necessary for Tutzauer's path-transfer entropy as a metric of manufacturing flexibility. Similar to this work, the authors concluded that analyzing the binary, directed graph provides information regarding the potential routing flexibility that the network could display. In addition, they agreed with Oggier *et. al.* [85] with respect to the use of the weighted, directed graph to evaluate the skew in the strength of the relationship among nodes.

It should be noted, however, that this metric suffers from loss of sensitivity for moderate to large graphs. As explained in Section 9.4, as the number of nodes and edges grows, so does the number of paths in the network. Consequently, there is little differentiation among the values of p_{ij} leading to many nodes having the same (or quite similar) value of path-transfer entropy. As the results in Table 9.1 demonstrated, this is specially true within strongly connected components.

It is also noteworthy that the path-transfer entropy formulation weighs all paths in G equally. However, very short and very long paths may not be representative of the flow in manufacturing networks. Note, for example, Wagner's study of walks on manufacturing graphs [55] which indicated that shorter walks were conducive to graphs that better matched the properties of the original manufacturing network. While currently outside the scope of this work, future research should consider weighing paths according to their length.

Thus, while a conclusive interpretation of path-transfer entropy in terms of Operations Research is elusive at this time, a number of aspects regarding this metric are agreed upon:

[1] Note that code profiling was conducted on the weighted, directed manufacturing graph corresponding to product family F_2 when the code is run in series.

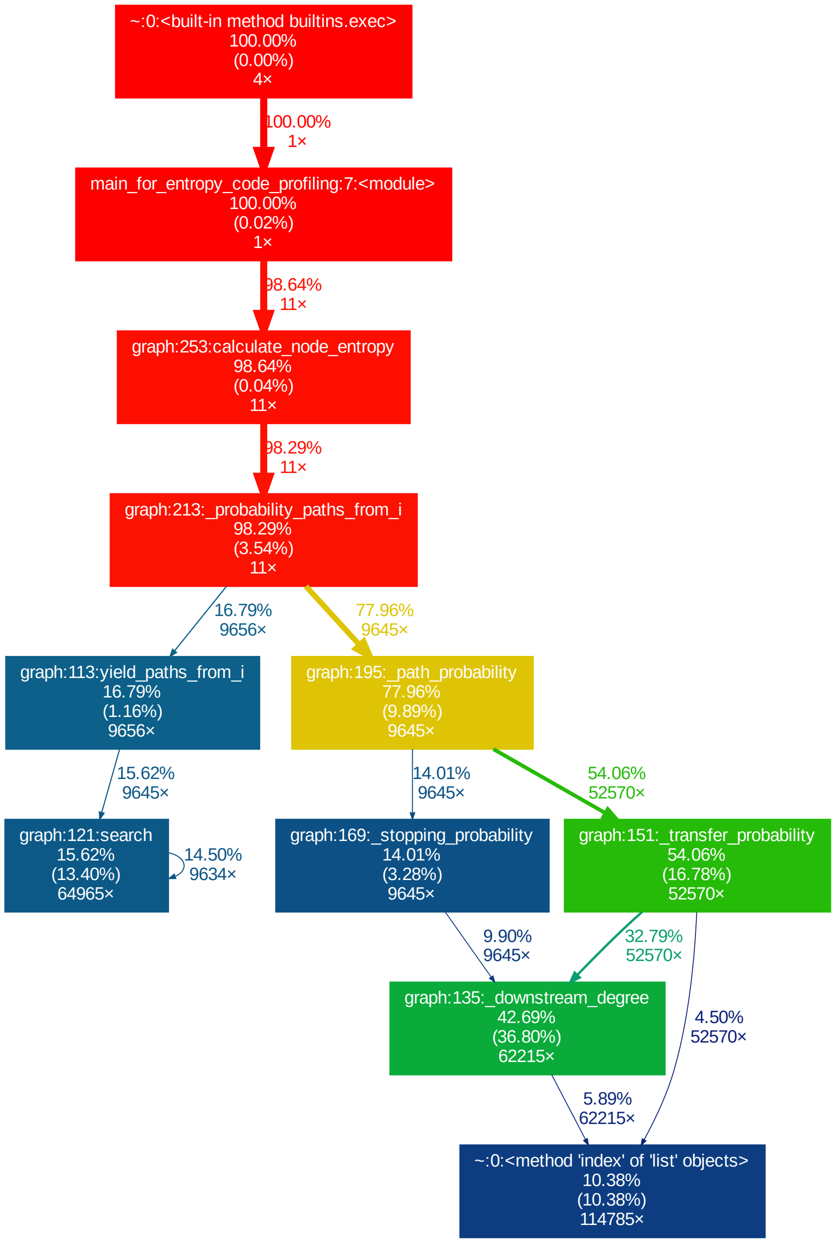


FIGURE 9.5: Path-transfer entropy centrality code profiling. *How to read this graph?* Each rectangle represents a function which name is indicated in the first line. The second line contains the percentage of the total CPU time spent on that function and its children functions. The third line contains the percentage of self time, that is to say not accounting for the time spent on children functions. The fourth line indicates the total number of calls to the function. The arrows point to children functions and indicate the percentage of total time transferred to the child function (first line) as well as the number of calls (second line). Finally, the color of the rectangles and arrows is associated with the percentage of time spent on them with red indicating a high percentage and blue indicating low[103].

In a manufacturing network where nodes represent workstations and directed edges, material flows, the underlying flow follows paths and proceeds through a transfer process. Tutzauer's entropy centrality is the only centrality metric whose underlying assumptions match those of manufacturing networks.

The probability distribution associated to the path-transfer entropy can be used to evaluate routing flexibility. Calculating the probability distribution of the binary, directed network indicates the potential level of flexibility that the manufacturing network can attain. On the other hand, the probability distribution of the weighted, directed graph speaks to the extent to which said flexibility is exercised in practice.

Practitioners should consider a number of aspects before deploying this metric. Firstly, this is a computationally expensive calculation, particularly for moderate, large and dense graphs, because it requires the identification all paths in G. Secondly, because all paths in G must be identified, the metric suffers from a loss of sensitivity that has yet to be addressed in literature. Finally, the formulation weighs all paths equally disregarding their length.

9.7 Summary

This Chapter illustrated the use of Tutzauer's path-transfer flow entropy centrality metric. Tutzauer explained that "flow beginning at a highly central node will stop with nearly equal probability at all other nodes, and a flow beginning at a less central node with have a much more uneven distribution of probabilities" [40]. Therefore, this metric was originally intended to rank nodes by importance. However, this work demonstrated that there is potential for more informative uses. For example, calculating the path probability distribution instead of the path-transfer entropy of a binary representation of a directed graph provides valuable insights regarding the network structure. Furthermore, the path probability distribution of the weighted representation informs on the effect of the skew in the strength of the relationship among nodes.

In the context of manufacturing networks, the path probability distribution of the network binary representation measures the potential for manufacturing flexibility. On the other hand, the path probability distribution of the network weighted representation determines the level to which said potential flexibility is exercised. This information is valuable in Operations Research and helps bridge the gap between potential and practical flexibility. Furthermore, it can aid production line design from the onset in order to ensure increased manufacturing flexibility [63].

10 Flow Networks

In all work presented up until now, the Bosch manufacturing network under study was abstracted as a directed graph with workstations as nodes and material flows as edges. In this Chapter, the concept of ”flow networks” is explored in order to answer questions about the material flow.

In a manufacturing flow network, goods with different levels of processing move through the system. They depart from a source, such as the raw materials storage room, and arrive to a sink, like a warehouse for finished goods. Two assumptions are made regarding the flow. Firstly, the flow is at steady state, i.e. raw materials leave the source at some steady rate, and finished goods arrive to the sink at the same rate. The second assumption is that of zero intermediate inventory levels, i.e. while directed edges act as conduits for material with a stated capacity, vertices act as junctions where material flows without accumulating.

In this Chapter, the appropriate procedure to determine the theoretical maximum production capacity and the location of bottlenecks will be discussed. For this purpose, the content will be organized as follows. Some key concepts of flow networks and flow need to be introduced first, followed by some operational excellence vocabulary. Then, the methodology followed to construct the flow network from the Bosch manufacturing data is presented. Note that this methodology has some important differences with that presented in Chapter 5. Finally, the optimization problem under study is stated and results are presented, followed by a number of conclusions.

This work shares similarities with that of Omar and Plapper [61].

10.1 Definitions

10.1.1 Flow networks

A flow network $G = (V, E)$ is a directed graph in which each edge $(v_i, v_j) \in E$ has a non-negative capacity $c(v_i, v_j) \geq 0$ [104]. The requirements of flow networks are as follows:

- A source s and a sink t vertex are present. If such nodes are not present or there existed more than one source or sink, a supersource and a supersink can be added. In this case, the capacity values are $c(s, v_i) = \infty$ and $c(v_i, t) = \infty$.

- If E contains an edge (v_i, v_j), then there is no antiparallel edge. In other words, the edge (v_j, v_i) is not present. In order to model a flow problem with antiparallel edges (v_i, v_j) and (v_j, v_i), the network must be transformed to an equivalent one containing no antiparallel edges. This is achieved by removing the edge (v_j, v_i) and adding a new vertex v' and two edges (v_j, v') and (v', v_i) with capacities equal to that of the original edge.
- Graph G is connected, i.e. for each vertex v_i there is a path $s \rightsquigarrow v_i \rightsquigarrow t$. Therefore, $|E| \geq |V| - 1$.

10.1.2 Flow

Given a flow network $G = (V, E)$ with a capacity function c, a source vertex s and a sink vertex t; the flow in G is a real-valued function $f : V \times V \to \mathbb{R}$ that satisfies the following properties:

- The flow from one vertex to another must be non-negative and must not exceed the capacity of the edge. Formally, $\forall v \in V$, the flow f follows the capacity constraint $0 \leq f(v_i, v_j) \leq c(v_i, v_j)$.
- The total flow into a vertex other than the source or sink must equal the total flow out of that vertex. Formally, $\forall v_j \in V - \{s, t\}$, it is required that

$$\sum_{v_i \in V} f(v_i, v_j) = \sum_{v_i \in V} f(v_j, v_i). \tag{10.1}$$

10.1.3 OR vocabulary

The Operations Research vocabulary refered to in this Chapter, is defined here. These definitions were obtained from the APICS Dictionary [105].

Bottleneck: A facility, function, department, or resource whose capacity is less than the demand placed upon it. For example, a bottleneck machine or work center exists where jobs are processed at a slower rate than they are demanded. *Synonym:* bottleneck operation.

Capacity: The capability of a worker, machine, work center, plant, or organization to produce output per time period. In simpler words, the number of units that can be filled per unit time.

Capacity utilization: Goods produced, or customers served, divided by total output capacity. It is a measure of how much of the capacity of an operation is being used.

Work in progress (WIP): A good or goods in various stages of completion throughout the plant, including all material from raw material that has been released for initial processing up to completely processed material awaiting final inspection and acceptance as finished goods inventory. *Synonym:* in-process inventory.

10.2 Complex networks vs. flow networks in manufacturing

As noted earlier, the type of directed graph representing a flow network differs from the complex network representation used in previous Chapters. In this Section, this difference is clarified.

From an Operations Research perspective, a manufacturing network is made of a group of workstations that add value to goods. The objective is to transform raw materials into a finished product. This can be abstracted mathematically as discussed in previous Chapters. Each workstation is represented by a node and, the material flow between pairs of workstations, by a directed edge. This type of manufacturing network was exemplified in Fig. 5.3. However, nodes and edges can also have properties summarizing other information related to the manufacturing network in question. For example, nodes can have capacity values (i.e. the manufacturing capacity of the workstation they represent). Edges, as seen before, have weight values indicative of the fraction of the total manufactured products that flow between pairs of workstations.

In this Chapter, however, manufacturing networks are represented in terms of flow networks (defined in Section 10.1.1). Each directed edge in a flow network acts as a conduit for material, an thus its capacity is known. In addition, each node represents a conduit junction and, with the exception of the source and sink, material must flow through nodes without accumulating in them. The reader should note, however, that the capacity of the edges in a manufacturing flow network may not necessarily have a one-to-one link to a property of the physical system [61]. If the capacity of the transport between nodes is assumed infinite (or is much greater than the capacity of all the workstations), then the effective capacity of the edges is constrained by that of the nodes. In this case, the appropriate capacity distribution method should be agreed upon in order to most accurately represent the physical world (see Section 10.3).

Flow networks are used to solve an optimization problem known as "maximum-flow". The objective is to compute the maximum rate at which materials can travel from the source to the sink without violating capacity constraints. In manufacturing, this translated to determining the network theoretical maximum production rate. As a reminder, this is done under two assumptions: steady state flow and zero intermediate inventory levels.

10.3 Capacity distribution

In a manufacturing network, the capacity of the workstations is known. In addition, the capacity of the transports between workstations is also known. When the transports are the limiting factor, their capacity should be considered for the edges. However, in cases where the capacity of the transport is much higher that that of the workstations, a different approach needs to be taken. In this case the capacity of the nodes is used. Yet, the definition of flow networks requires capacity values for the edges, not the nodes. Thus, one must decide how to distribute the capacity of a node to its out-going edges to best abstract the real manufacturing network as a flow network. A number of possible alternatives are introduced hereafter and summarized in Table 10.1.

TABLE 10.1: Summary of capacity distribution formulations.

based on	formulation	comments
out-degree	$c(v_i, v_j) = \dfrac{C_i}{k_i^{\text{out}}}$	Suitable when all customer nodes v_j are identical and have the same capacity C_j.
material flow	$c(v_i, v_j) = \dfrac{w_{ij}}{\sum_j w_{ij}} C_i$	This method should be reserved for cases where the only information available is the real material flows through the network.
customer node	$c(v_i, v_j) = \dfrac{C_j}{\sum_j C_j} C_i$	Appropriate when all customer nodes v_j perform the same task but have different capacity C_j.

10.3.1 According to the out-degree

One possibility is to make use of the out-degree of a node k_i^{out}, defined as $k_i^{\text{out}} = \sum_j a_{ij}$. This can be calculated from the adjacency matrix $\mathcal{A}$ as explained in Chapter 7, Section 7.2. Then, the capacity of each out-going edge from node v_i could be calculated as

$$c(v_i, v_j) = C_i / k_i^{\text{out}}. \tag{10.2}$$

Omar and Plapper [61] explained that this method is suitable when all customer nodes v_j are identical and thus have the same capacity C_j.

10.3.2 According to the actual material flows

A second option is to consider the actual material flow traversing edge (v_i, v_j). This information is available in the weight matrix $\mathcal{W}$. Then the capacity of a node would be distributed as follows

$$c(v_i, v_j) = \frac{w_{ij}}{\sum_j w_{ij}} C_i. \tag{10.3}$$

Omar and Plapper [61] suggested that this method should be reserved for cases where the only information available is the real material flows through the network.

The authors warn that this method should be used with caution. They explain that edge capacities determined using the actual material flows must be constrained to a single product family. In addition, these values are likely to differ from those obtained for a different product family. Thus, the flow network obtained cannot be utilized to determine the theoretical maximum production rate for a different product family. The authors also recommend to generate the weight matrix utilizing data from an extended period of time. This is done to circumvent bias due to day to day variations in manufacturing outputs. In this way, one would average out daily fluctuations.

10.3.3 Considering the capacity of customer nodes

A third option is to take the capacity of customer nodes into consideration. In this case, the capacity of workstation v_i is distributed as follows

$$c(v_i, v_j) = \frac{C_j}{\sum_j C_j} C_i. \tag{10.4}$$

Omar and Plapper [61] explained that this method should be used when all customer nodes v_j perform the same task but have different capacity C_j. This occurs, for example, when there are two different models of the same piece of equipment.

10.4 Methodology

In order to solve the optimization problem known as "maximum-flow", the Bosch manufacturing network under study must be converted into a flow network. The appropriate procedure, summarized in Fig. 10.1 is described hereafter. Note that, in this Chapter, the complex network corresponding to product family F_2 is used instead of the full Bosch manufacturing network. The reader should refer to Chapter 6 for details on how product families were identified.

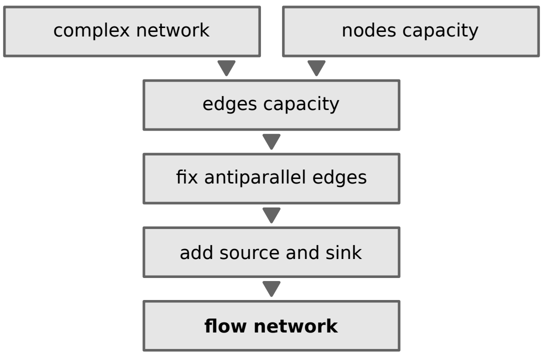


FIGURE 10.1: Step-wise methodology to obtain a flow network from a complex network when node capacity values are known.

10.4.1 Complex network

Following the step-wise procedure described in Fig. 10.1, the starting point to obtain the appropriate "flow network" for product family F_2 is its "complex network". The graph for F_2 is characterized by the weight matrix in Eq. 10.5. Note that two artificial nodes, a supersource s and a supersink t, have been added for clarity. These provide information regarding the nodes that act as starting points for manufacturing (v_{24}, v_{25}, v_{26} and v_{27}), and those where production can finalize (v_{33}, v_{34}, v_{35}, v_{36} and v_{37}). In addition, the weight listed for the appropriate edges (s, v_i) and (v_i, t) indicate the frequency with which manufacturing starts or finalizes in node v_i respectively.

$$
W = \begin{array}{c} \\ s \\ v_{24} \\ v_{25} \\ v_{26} \\ v_{27} \\ v_{29} \\ v_{30} \\ v_{33} \\ v_{34} \\ v_{35} \\ v_{36} \\ v_{37} \\ t \end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc} s & v_{24} & v_{25} & v_{26} & v_{27} & v_{29} & v_{30} & v_{33} & v_{34} & v_{35} & v_{36} & v_{37} & t \end{array} \\
\begin{pmatrix}
0 & 0.690 & 0.308 & 0.0002 & 0.002 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.472 & 0.171 & 0.037 & 0.010 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.189 & 0.088 & 0.024 & 0.007 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.557 & 0.105 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.217 & 0.045 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.813 & 0.139 & 0.035 & 0.005 & 0.003 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.162 & 0 & 0.607 & 0.198 & 0.018 & 0.015 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.018 & 0 & 0.642 & 0.150 & 0.143 & 0.015 & 0.031 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.002 & 0.150 & 0 & 0.292 & 0.313 & 0.170 & 0.074 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.017 & 0.042 & 0 & 0 & 0.392 & 0.042 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.010 & 0.030 & 0 & 0 & 0.422 & 0.046 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.077 & 0.053 & 0.027 & 0.035 & 0 & 0.807 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\end{array}
\tag{10.5}
$$

10.4.2 Nodes capacity estimation

As stated, the objective is to obtain the appropriate flow network, where nodes v_i are the abstract representation of workstations and edges have a capacity value $c(v_i, v_j)$, or simply c_{ij}, in the appropriate units (for example, units/hour). Following the steps shown in Fig. 10.1, in addition to the "complex network" corresponding to product family F_2, the "nodes capacity" is necessary. Note that the raw data from the Bosch production line used for this work does not indicate the capacity of each workstation. Thus, these capacity values must be estimated from the raw data contained in the "train_date.csv" file.

The original time-stamp data contained in the "train_date.csv" file was shown in Chapter 5. In summary, it is a comma-separated-value file with headers in the first line and observations in subsequent lines. Each observation provides a manufacturing item anonymized ID followed by a number of anonymized time-stamps associated to the workstations in which the item was processed. Missing values indicate that the item was not processed at the specific workstation. An extensive exploratory data analysis was conducted and is available to the interested reader in Appendix A.

In this work, the capacity of each workstation is estimated from the time-stamp data. It is assumed to be equal to the highest throughput of each workstation in units processed per unit of time. The reader should note that the assumption that such time period corresponds to maximum (or nearly maximum) capacity may not hold true for all networks. However, it is a good compromise to estimate workstation capacity.

The procedure to obtain the estimated workstation capacity is summarized in Fig. 10.2. The necessary steps are briefly described hereafter. The code necessary to reproduce these results is available in the `flow_networks` folder in [72].

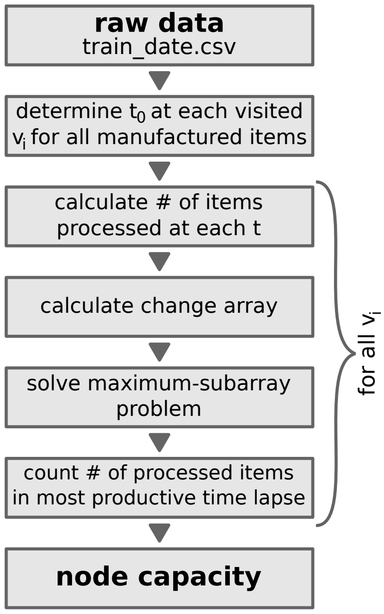


FIGURE 10.2: Step-wise methodology to estimate the capacity of each node.

1. The initial time-stamp of each manufactured item at each visited workstation is determined.

2. The number of items processed at each time-stamp value (from 0.0 to 1718.48 with granularity of 0.01) is calculated for each v_i. For example, the number of units processed in node v_0 for the time-lapse between 0.0 and 1.0 is shown in the left graph of Figure 10.3.

The objective is the to find the time of maximum productivity for each workstation. This is known as the "maximum sub-array problem". To solve it, a transformation is required. Instead of looking at the number of items processed at each time-stamp value, the change in items processed from one time-stamp to the next is more useful. The change at time-stamp t_i can be calculated as the difference between the number of items processed at time-stamp t_i and those processed at time-stamp t_{i-1}. Thus,

3. Calculate the "change array" for all v_i. As an example, the "change array" of node v_0 for the time-lapse between 0.0 and 1.0 is shown in the right graph of Figure 10.3.

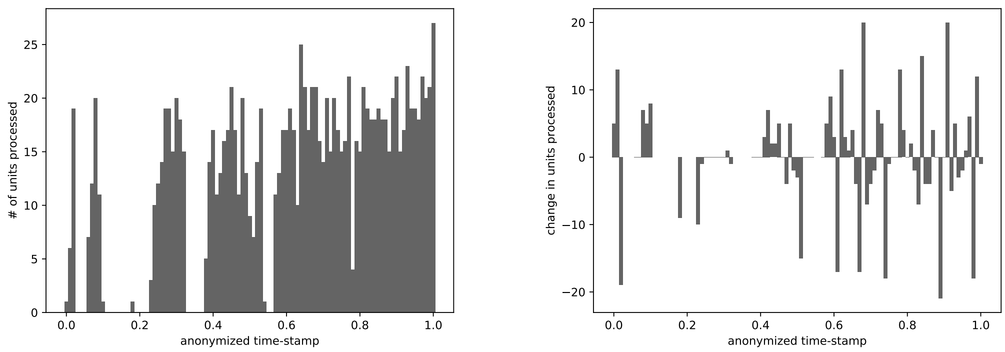


FIGURE 10.3: *Left:* Number of units processed vs. time (in anonymized units) for node v_0. *Right:* Change array vs. time (in anonymized units) for node v_0. In both graphs, only the fraction of data corresponding to time-stamps values between 0.0 and 1.0 is plotted for clarity.

4. Find the nonempty, contiguous sub-array whose values have the largest sum. The maximum sub-array problem has been widely studied and can be solved in $O(n)$ time using Kadane's algorithm [106]. The result obtained is the time lapse where the highest number of items were processed.

5. Estimate the capacity of v_i by counting the number of items processed in the most productive time lapse for each v_i and normalizing by time. The estimated capacity of each workstation is shown in Table 10.2.

TABLE 10.2: Capacity estimation (in units/day) for each workstation based on their most productive time.

v_i	C	v_i	C	v_i	C	v_i	C
v_0	880	v_{13}	364	v_{26}	167	v_{39}	42
v_1	1058	v_{14}	288	v_{27}	210	v_{40}	21
v_2	535	v_{15}	165	v_{28}	12	v_{41}	45
v_3	438	v_{16}	177	v_{29}	1498	v_{42}	4
v_4	457	v_{17}	185	v_{30}	1633	v_{43}	42
v_5	353	v_{18}	166	v_{31}	272	v_{44}	28
v_6	532	v_{19}	182	v_{32}	38	v_{45}	45
v_7	527	v_{20}	348	v_{33}	1608	v_{46}	0
v_8	880	v_{21}	122	v_{34}	1608	v_{47}	28
v_9	295	v_{22}	119	v_{35}	757	v_{48}	45
v_{10}	298	v_{23}	108	v_{36}	759	v_{49}	33
v_{11}	327	v_{24}	311	v_{37}	1499	v_{50}	26
v_{12}	364	v_{25}	132	v_{38}	41	v_{51}	51

10.4.3 Edge capacity estimation

In the previous Section, the "nodes capacity" was estimated from the original time-stamp data. As noted in Fig. 10.1, the "nodes capacity" in conjunction with the "complex network" for product family F_2 are necessary ingredients to produce a "flow network". The next step forward requires the calculation of the "edges capacity".

As discussed previously, for the production line under study, the capacity of the transports can be assumed to be infinite. Thus, the factor limiting the theoretical maximum production rate of the network is the capacity of the workstations. Consequently, node capacities need to be distributed among the out-going edges of each node v_i to obtain the necessary "edge capacities". In this work, node capacity is distributed according to the material flow values available in the weight matrix, as described in Section 10.3.2. This method is chosen given that no other information regarding this production line is available.

10.4.4 Dealing with anti-parallel edges

As described in Section 10.1.1, a flow network must not contain anti-parallel edges. If both edges (v_i, v_j) and (v_j, v_i) exist, the network must be transformed to an equivalent one containing no anti-parallel edges. This is achieved by removing the edge (v_j, v_i) and adding a new vertex v' and two edges (v_j, v') and (v', v_i) with capacities equal to that of the original edge. This procedure constitutes the "fix antiparallel edges" step in Fig. 10.1.

10.4.5 Adding source and sink nodes

From its definition in Section 10.1.1, it is known that a flow network requires two special nodes: a source node s and and a sink node t. However, in the Bosch manufacturing network, production can start is several different nodes and end in a few other nodes, as previously shown in Fig. 9.3. In fact, for product family F_2, manufacturing can start in nodes v_{24}, v_{25}, v_{26} or v_{27}, and finish in either node v_{33}, v_{34}, v_{35}, v_{36} or v_{37} as shown in Eq. 10.5. Thus, a supersource and a supersink nodes are added (step "add source and sink" in Fig. 10.1). The supersource contains out-going edges to all possible starting nodes. Similarly, the supersink node contains only in-going edges coming from nodes where manufacturing can conclude. The capacity of edges associated to either the supersource or the supersink is assumed infinite.

10.4.6 Final flow network

Following the steps described in the preceding Sections and summarized in Fig. 10.1, is a necessary precondition to obtain the "flow network" that will be used to solve the maximum-flow problem. In Fig. 10.4, the original complex network for product family F_2 (*left* graph) can be compared with the corresponding flow network (graph on the *right*). Note that, just like in Eq. 10.5, a source s and a sink t have been added to the complex network graph to facilitate comparison with its corresponding flow network.

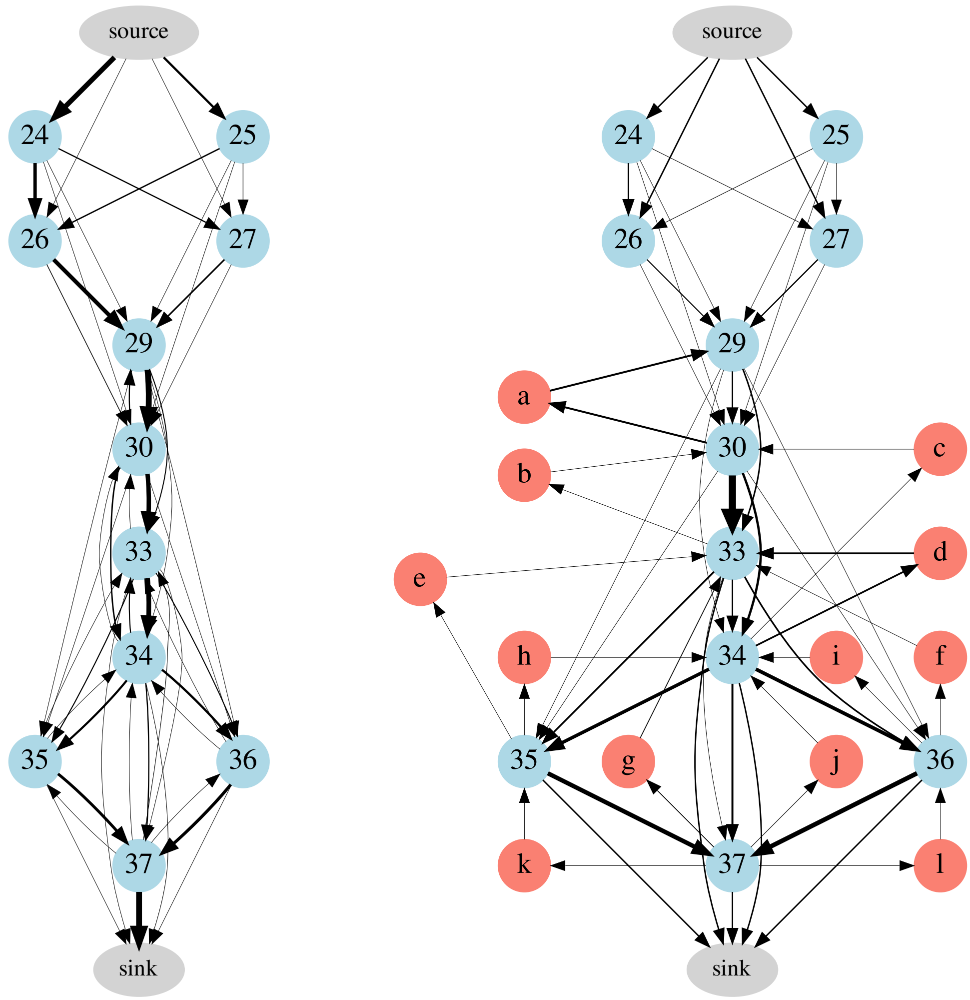


FIGURE 10.4: *Left:* Manufacturing network corresponding to product family F_2 where a source node s and a sink node t have been added to facilitate visualization. The thickness of the edges is related to the w_{ij} values in the weight matrix $\mathcal{W}$. *Right:* Flow network corresponding to product family F_2. The thickness of the edges represents their capacity c_{ij} and is thus, related to the capacity of the nodes.

10.5 Optimization problem

The flow network illustrated in Fig. 10.4 was constructed with the objective to solve an optimization problem commonly known as the "maximum flow problem". The goal is then to "compute the greatest rate at which we can ship material from the source to the sink without violating any capacity constraints" [104]. This problem can be rephrased using manufacturing friendly vocabulary as follows [61]:

> *Given a manufacturing flow network of known edge capacities, determine the theoretical maximum production rate that the network can attain under the assumptions of steady state flow and zero intermediate inventory levels.*

The maximum flow problem can be solved using the Ford-Fulkerson method [104, 107], described in Section 10.6.

10.6 The Ford-Fulkerson algorithm

In order to succinctly explain the Ford-Fulkerson algorithm [107], a number of definitions are introduced hereafter.

Definition 7. *Given a flow network $G = (V, E)$ with a source s and a sink t, let f be a flow in G. Considering the pair of nodes $u, v \in V$, the **residual capacity** $c_f(u, v)$ can be defined as*

$$c_f(u,v) = \begin{cases} c(u,v) - f(u,v) & \textit{if } (u,v) \in E \\ f(v,u) & \textit{if } (v,u) \in E \\ 0 & \textit{otherwise.} \end{cases} \tag{10.6}$$

Following the definition of flow networks that does not allow the presence of anti-parallel edges, exactly one case in this equation applies to each ordered pair of vertices.

Definition 8. *Given a flow network $G = (V, E)$ and a flow f, the **residual network** of G induced by f is $G_f = (V, E_f)$, where $E_f = \{(u, v) \in V \times V : c_f(u, v) > 0\}$.*

Definition 9. *Given a flow network $G = (V, E)$ and a flow f, an **augmenting path** p is a simple path from source s to sink t in the residual network G_f.*

The algorithm. In the Ford-Fulkerson method [104, 107], edges are initialized with a flow value $f(v_i, v_j) = 0$. The flow is augmented iteratively. In each iteration of the method, an "augmenting path" p is found in the associated "residual network" G_f. As a result, the total flow from source s to sink t increases. However, the value of the flow of each individual edge $f(v_i, v_j)$ may increase or decrease as needed, as long the overall flow from source s to sink t is increased. The algorithm terminates when there is no "augmenting path" p available from source s to sink t. Upon termination, this process yields the maximum flow.

The appropriate pseudo-code is summarized in Algorithm 2. A detailed description of different algorithms and running times can be found in [104].

It should be noted that the augmenting path used in the Ford-Fulkerson method does not necessarily correspond to a manufacturing path. The method chooses paths following a breadth-first-search, which gives shortest paths from source to sink [108]. Despite this geodesic path search, the method correctly determines the maximum production capacity. In addition, reverse engineering of the maxed out edges allows the identification of limiting resources (bottleneck nodes).

Algorithm 2

```
function FORD-FULKERSON(G, s, t)
    for each edge (v_i, v_j) ∈ E do
        initialize flow f(v_i, v_j) to 0
    end for
    while there exists a path p from s to t in the residual network G_f do
        c_f(p) = min{c_f(u, v) : (u, v) is in p}
        for each edge (u, v) ∈ p do
            if (u, v) ∈ E then
                f(u, v) = f(u, v) + c_f(p)
            else
                f(v, u) = f(v, u) − c_f(p)
            end if
        end for
    end while
    return f
end function
```

10.7 Results

Application of the Ford-Fulkerson method produces the following results. The maximum production capacity for the manufacturing network under study, bearing the assumptions of steady state flow and zero intermediate inventory levels, is of 408 units/day. This can also be interpreted as 408 units should be scheduled for daily production in order to guarantee steady flow of materials from source to sink without accumulation and/or intermediate inventory.

Furthermore, the Ford-Fulkerson method returns the flow fraction graph shown in Fig. 10.5. This graph indicates the fraction of the edge capacity that has been consumed in order to reach the maximum production capacity. The flow fraction value is always in the range between 0 (i.e. 0%) and 1 (i.e. 100%). Comparing the flow fraction digraph in Fig. 10.5 against the original flow network in Fig. 10.4, allows to make observations about the flow. For example, it is clear that workstations v_{26} and v_{27} must be used at 100% capacity to reach the calculated steady state maximum production rate. Since the edges (v_{24}, v_{29}), (v_{24}, v_{30}), (v_{25}, v_{29}) and (v_{25}, v_{30}) are also used at 100% capacity, that translates to workstations v_{26} and v_{27} blocking the flow of any further WIP processed in workstations v_{24} and v_{25}. Thus, it can be concluded that workstations v_{26} and v_{27} are the bottleneck of this operation.

10.8 Discussion

The study of flow networks is hardly new. In fact, the work of Ford and Fulkerson dates to 1962 [107]. Furthermore, recognition of the potential of flow networks in Operations Research can be attributed to Alderson's 2008 paper in which he advocated for the use of network science in this field [36]. Yet, as indicated in Chapter 3, flow networks have only been used once in recent research to study manufacturing

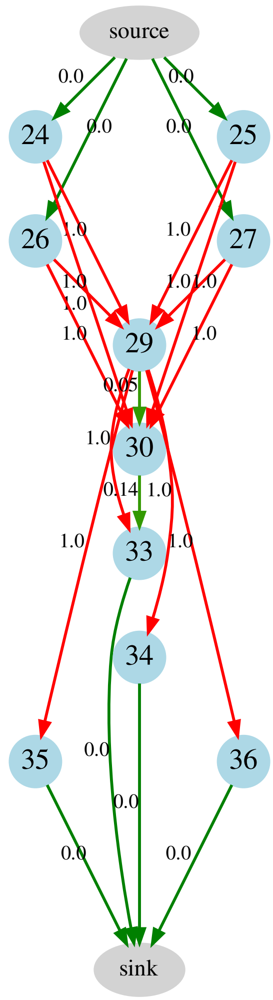


FIGURE 10.5: Flow fraction digraph produced by the Ford-Fulkerson method when applied to the flow network of product family F_2

networks. Omar and Plapper [61] proposed the use of flow networks to solve the maximum flow optimization problem in order to determine the maximum theoretical production rate. This is done under two assumptions: steady state flow and zero intermediate inventory levels. In addition, the authors explained how to identify limiting resources. The work in this Chapter is based on Omar and Plapper's [61].

In this Chapter, the maximum flow problem was solved by means of the Ford-Fulkerson method. The algorithm utilizes breadth-first-search to find the "augmenting paths". This search method finds geodesic paths between pairs of nodes. As a result, the Ford-Fulkerson method does not visit workstations following manufacturing paths. Instead shortest paths from source s to sink t are followed. Thus, not all workstations in the manufacturing network are "visited" in order to calculate the theoretical maximum production rate. However, this choice of path search guarantees computational efficiency, determining the maximum production rate of the whole system quickly. Yet, one must be cognizant that the flow fraction graph in Fig. 10.5 is not a representation of the manufacturing network under study. It is only intended to graphically identify the workstations that would be utilized at maximum capacity in the theoretical maximum production rate scenario.

These results are an exciting step towards the use of plant floor data for optimization purposes. Yet, caution should be exercised. The assumption of zero intermediate inventory levels can jeopardize throughput through blocking and starving in serial production lines [109]. And while the manufacturing network under study does not seem to be a serial production line at first sight, further analysis is required. Note, for example, that the findings in Chapter 9 indicated that, while there is potential for flexible routing of operations, manufacturing steps are still followed mostly in serial mode in practice. Thus, further research is necessary to determine the extent to which zero intermediate inventory levels may affect throughput [61].

At this time, the following interpretation is proposed in terms of Operations Research:

> *A manufacturing flow network is a graph $G = (V, E)$ where nodes represent workstations and directed edges indicate conduits for material flow with a stated capacity value $c(v_i, v_j)$. Flow networks are constructed with the objective of solving an optimization problem known as "maximum-flow". The solution provides the following information:*
>
> - *The theoretical maximum production rate, i.e. the maximum number of units that can be produced per unit of time under the assumptions of steady state flow and zero intermediate inventory levels.*
> - *The limiting resource(s). When the capacity of the edges is derived from the capacity of the workstations, the limiting resources are workstations that represent a bottleneck operation in the manufacturing network.*
>
> *Practitioners should exercise caution, since the assumption of zero intermediate inventory levels could potentially jeopardize throughput equivalently to serial production lines. More research is encouraged in this regard.*

10.9 Summary

This Chapter illustrated the use of flow networks as a tool to determine the maximum theoretical production rate that a manufacturing network can attain. This was done under two assumptions: the flow is at steady state and there is no intermediate inventory levels. In addition, the identification of limiting resources or bottlenecks was demonstrated from analysis of the flow graph produced when solving the maximum flow optimization problem. Finally, the appropriate metric interpretation was provided as well as a number of aspects to consider when applying it.

11 Discussion

Throughout this work, several well-known complex network metrics were introduced. Their use was illustrated using a manufacturing complex network constructed from time-stamp data that originated from a Bosch production line. The results obtained were discussed in the context of existing research and an appropriate interpretation was proposed for those metrics deemed suitable. Indeed, the objectives for this work, listed in Chapter 4, entailed

1. clarifying the areas where popular complex network metrics are applicable for manufacturing networks,
2. proposing the use of other existing complex network metrics whose underlying assumptions are more adequate for manufacturing networks, and
3. determining the appropriate interpretation of these metrics in terms of Operations Research, and specifically, for manufacturing networks.

To this end, Chapters 7, 8 and 9 focused on manufacturing networks in which nodes represent workstations and directed edges, material flows. Additionally, the number of transactions among pairs of workstations could be codified as an edge weight and used to differentiate structural from practical characteristics of the network. In Chapter 10, however, the manufacturing network under study was abstracted as a flow network. In a manufacturing flow network, nodes represent workstations but edges act as conduits for material with a stated capacity value.

In this Chapter, all the metrics introduced in this work are discussed in terms of three important aspects listed in Fig. 11.1. Firstly, their *suitability* is evaluated based on whether their underlying assumptions match those of manufacturing networks. Then, the metrics *interpretation* is re-instated, evaluated against said assumptions and framed in the context of Operations Research. Finally, the *shortcomings* of applicable metrics are discussed. A summary of the complex network metrics presented in this work along with their suitability and interpretation is shown in Table 11.1.

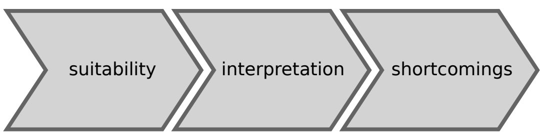


FIGURE 11.1: Aspects of the studied complex network metrics discussed in this Chapter.

TABLE 11.1: Summary of studied metrics listing their suitability and interpretation in the context of manufacturing networks.

Metric	Chapter	Suitable?	Interpretation
in-degree	7	Yes	Number of direct upstream suppliers
out-degree	7	Yes	Number of direct downstream customers.
in-strength	7	Yes	Measure of the incoming workload from suppliers.
out-strength	7	Yes	Measure of the difficulty faced by the node to satisfy downstream customer needs.
betweenness	7	No	N/A
clustering	7	Yes	Measure of connectedness.
PageRank	8	No	N/A
entropy	9	Yes	The associated probability distribution determines routing flexibility. From the binary, directed graph, the potentially attainable level of flexibility is determined. From the weighted, directed graph, the extent to which said flexibility is exercised in practice is measured.
flow network	10	Yes	Solving the maximum-flow problem provides the theoretical maximum production rate under the assumptions of steady state flow and zero intermediate inventory levels. Limiting resources can be identified using the flow fraction graph.

11.1 Metrics suitability

The main factor determining a metric suitability is its underlying assumptions. Specifically, centrality metrics make assumptions regarding the route that the flow follows on the network as well as the flow propagation method. As discussed in Section 2.3, these two factors create a two-dimensional typology widely known as "Borgatti's typology" [38]. A suitable metric must share the same typology as the network it is trying to describe.

The flow in manufacturing networks follows *paths* and progresses by means of a *transfer process*. A path is defined as a sequence of linked nodes in which neither nodes nor edges are repeated. In turn, a transfer process is a type of traffic flow where an item can only be one place at a time and moves from node to node in the network. Then, any metric utilized to characterize a manufacturing network should, in principle, belong to the same Borgatti's typology.

However, this rule could be defined as *sufficient* yet not always *necessary*. For example, the in- and out-degree were deemed suitable to describe manufacturing networks although their propagation method is *parallel duplication*, as indicated in Table 2.1. The same can be said of the in- and out-strength. Yet, these metrics application should be restricted to the domain specific interpretation provided, as discussed in the next Section.

Other metrics, like betweenness centrality and PageRank, were deemed unsuitable based on their underlying assumptions and the Borgatti's typology to which they belong. Betweenness centrality shares the propagation method of manufacturing networks, but the routes followed are *geodesics* instead of paths. Indeed, this metric was developed to identify *information gatekeepers* in communication networks. As a consequence, this metric performs poorly when the objective is the identification of gatekeepers in manufacturing networks [51] where the flow could overcome a gatekeeper following a different (lengthier) path. On the other hand, PageRank belongs to a completely different Borgatti's typology where the flow progresses following *walks* by means of *parallel duplication*. Consequently, the underlying assumptions are far removed from those of the networks under study.

Tutzauer's path-transfer flow entropy is the only centrality metric developed while explicitly stating that the flow follows paths by means of a transfer process. This metric, is thus, suitable for manufacturing networks given that it shares the same underlying assumptions. Yet, as discussed in Chapter 9 and in the following Section, its interpretation is anything but straight-forward.

It should also be noted that not all the metrics studied are centrality metrics. For example, the clustering coefficient and the maximum-flow problem in flow networks are not graph centrality metrics. They were deemed suitable based on their objectives and how those matched problems associated to manufacturing networks. The clustering coefficient, for example, is a metric of connectedness and thus, is useful to characterize a manufacturing network as a whole (or on a by-subgraph basis). The objective is to determine whether the production line is serial or not. The maximum-flow problem in flow networks was indeed developed with Operations Research objectives in mind [36, 107]. Its objective matches OR interests in determining maximum theoretical production rates. Furthermore, this work and Omar and Plapper's [61] identified the flow-fraction digraph usefulness for the identification of limiting resources.

In conclusion, practitioners should evaluate complex network metrics before blindly applying them to characterize manufacturing networks. Firstly, the underlying assumptions should be evaluated. In particular, a thorough analysis of similarities and differences between the problem the metric originally intended to solve and the current network under study must be made. Then, the characteristics of the flow for a specific (centrality) metric should be determined and compared with those of the network under study. While certain differences may coexist, they limit the metric interpretation. Thus, the use of metrics where the underlying assumptions match the flow characteristics of the network under study is encouraged.

11.2 Metrics interpretation

As indicated in Fig. 11.1, once a metric suitability has been determined, the appropriate interpretation must be agreed upon. The interpretations of all suitable metrics studied in this work are listed in Table 11.1. These were proposed following a thorough literature review, analysis of the metrics Borgatti's typology and underlying assumptions, and evaluation of the results when applied to manufacturing networks. The objective was to provide a useful and actionable interpretation in terms of Operations Research.

From Table 11.1, it can be observed that simple metrics whose typology does not perfectly match the underlying assumptions of manufacturing networks have a narrow interpretation. For example, the in- and out-degree are restricted to a mere accounting of suppliers and customer nodes respectively; and the in- and out-strength are circumscribed to the accounting of incoming and outgoing workloads. Practitioners are warned against using these metrics with more far-reaching interpretations. Specifically, users should be well aware of these metrics weaknesses as listed in the next Section.

When a metric underlying assumptions perfectly match those of manufacturing networks, their use and interpretation widens. For example, the probability distribution of the path-transfer flow entropy centrality was found to be useful in determining both potential and practical routing flexibility. In Chapter 9, it was shown that the path probability distribution of the binary, directed graph aided practitioners in determining the potentially attainable level of manufacturing flexibility. On the other hand, the path probability distribution of the weighted, directed graph measures the extent to which said flexibility is exercised in practice.

The interpretation of non-centrality metrics is typically more straight-forward. The clustering coefficient was developed as a measure of connectedness and that is how it is interpreted in the context of manufacturing network. While the traditional use of this metric was demonstrated in Chapter 7, some domain specific suggestions were made. In a manufacturing network, centrain groups of workstations are highly connected. This is because some flexibility is possible for certain production steps. However, there is a generally prescribed order that must be followed. In the words of Browne *et. al.* [100], "the system can potentially alter the order of a number of manufacturing operations, yet a required partial precedence structure exists". Thus, the strongly connected component subgraph clustering coefficient was proposed as a more appropriate metric of connectedness.

The maximum-flow problem solved with flow networks is another example of a non-centrality metric. In this case, the optimization problem was originally defined with Operations Research in mind. Thus, its interpretation is straight-forward. Solving the maximum-flow problem is conducive to the theoretical maximum production rate. It must be noted that two assumptions are made: the flow is on steady state and there are zero intermediate inventory levels. In addition, it was demonstrated in Chapter 10 that limiting resources can be identified using the flow fraction graph.

As expected, no interpretation is provided for metrics deemed unsuitable in Table 11.1. This is the case of betweenness centrality and PageRank. Note however, that in Chapter 7, more research is encouraged regarding the use of betweenness centrality in conjunction with the network graph to aid the identification of nodes that may disconnect upstream suppliers from their downstream tier. Yet, practitioners are warned against using betweenness centrality and PageRank for bottleneck and/or key machine identification.

11.3 Metrics shortcomings

Once a metric suitability is determined and their interpretation agreed upon, the only step left is to identify its shortcomings, as shown in Fig. 11.1. Indeed, understanding what a metric cannot do is as important to ensure it is appropriately used. Thus, the limitations and weaknesses of the metrics studied in this work are listed hereafter and summarized in Table 11.2.

TABLE 11.2: Summary of limitations suffered by the metrics presented in this work.

Limitations	Affect
Loss of seasonality Loss of historical changes	all metrics
Computational efficiency Loss of sensibility All paths are weighted equally	path-transfer flow entropy
Applicability due to far-reaching assumptions	flow networks

A first aspect to consider is that, in all cases, these metrics were utilized to characterize a network constructed using data from a two year long period of time. The justification given was "to average out data seasonality", which is typical of manufacturing. The drawback is the lack of a temporal component to this data. Take the in- and out-strength for example. These values are the number of processed items in a time lapse of two years. Thus, no information is available regarding peak and down times, for example. In addition, any historical changes are also lost. For example, the in- and out-degree may have changed over time. Some new connections may have been added and others discontinued. Yet this information is no longer available: all connection, regardless of when they occurred, are listed. The graph constructed is static, it aggregates two years of data and does not provide any information regarding variations in time. Some authors noticed this [45, 57] and proposed methods to evaluate if the inferred internal structure of the network is consistent over time. While this is outside the scope of this work, more research is encouraged in this regard.

Some metrics also have intrinsic limitations and weaknesses. For example, the path-transfer flow entropy discussed in Chapter 9 is computationally expensive. This metric, an thus the associated probability distribution used to evaluate routing flexibility, requires the search for all possible paths in G. Since the number of paths grows combinatorially as the number of nodes and edges increases [97], the time required to find them all and to then evaluate their probability increases. While some authors suggested path pruning as a possible solution [40, 85, 95, 96, 98], this work indicated that the path search may not be the limiting computational routine (see Section 9.5). In addition, the path-transfer flow entropy centrality suffers from loss of sensibility as well. This means that neighboring nodes tend to have very similar entropy values. As noted in Chapter 9, this still needs to be addressed in literature. Finally, all paths are weighted equally ignoring their length. Thus, more research is needed to understand how these aspects affect entropy centrality.

Yet other metrics may have assumptions that are too far removed from the realities of manufacturing. This is the case of the theoretical maximum production rate calculated from the maximum flow problem in flow networks. Note that this theoretical production rate is calculated based on two strong assumptions: steady state flow and zero intermediate inventory levels. As discussed in Chapter 10, the assumption of zero intermediate inventory levels can jeopardize throughput through blocking and starving in serial production lines [109]. Thus more research is encouraged to understand how zero intermediate inventory levels may affect throughput in manufacturing networks.

12 Conclusions

12.1 Summary

In this work, the mathematical tools of network science were deployed to characterize manufacturing networks. The literature review presented in Chapter 3 demonstrated that a limited number of records exists on the subject. These are concentrated on the past decade, from 2012 to the present, coinciding with the increased digitalization and data generation of the fourth industrial revolution. However, most existing literature fails to evaluate a number of important aspects regarding the suitability and adequate interpretation of the network science machinery deployed.

In order to fill this research gap, this work pursued objectives aimed at consolidating the use and understanding of network science in the context of manufacturing networks. Listed in Chapter 4, these objectives entailed:

- determining a complex network metric *suitability*, and thus identifying areas where popular existing complex network metrics can be applied in the context of manufacturing networks;
- proposing the use of less popular metrics whose underlying assumptions are better aligned with those of manufacturing networks, and thus make them more *suitable*; and
- providing an adequate and actionable *interpretation* for suitable metrics in terms of Operations Research.

While it was not explicitly listed as an objective, this work also stressed the *shortcoming* of suitable metrics. Indeed, a complete *interpretation* cannot be provided unless a metric limitations and weaknesses are fully understood.

In order to fulfill these objectives, this work used data corresponding to a Bosch production line to construct the appropriate manufacturing network. The necessary methodology was described in Chapter 5 and the full graph was shown in Fig. 5.3. In addition, product families were identified in Chapter 6, and their respective complex networks illustrated in Fig 6.2.

The rest of this work presented a number of self-contained Chapters where different complex network metrics were discussed. Chapter 7 was concerned with traditional topological metrics such as degree, strength, betweenness centrality and the clustering coefficient. Chapter 8 discussed the famous PageRank algorithm used behind Google search engine. Chapter 9 proposed the use of path-transfer flow entropy for manufacturing networks due to their matching underlying assumptions. And Chapter 10 promoted the use of flow networks and the solution to the maximum flow problem as an alternative for bottleneck identification.

These four Chapters have a number of commonalities. In them, each metric is defined and then exemplified on the Bosch production line network (or in a network corresponding to a product family). Then, a thorough analysis of each metric flow related assumptions was conducted. Special focus was drawn to whether the metric underlying assumptions match those of manufacturing networks. Thus, the first aspect considered is the *suitability* of each complex network metric for the problem at hand. In this regard, it was stressed that complex networks metrics should never be utilized blindly. Instead, the problem for which they were derived and the flow they describe should be analyzed and compared to those of the network under study. When proper agreement is found, the metric is considered suitable. Only suitable metrics should be used to characterize manufacturing networks.

For complex network metrics considered suitable in the context of manufacturing networks, an adequate and actionable *interpretation* was provided. The formulation of said interpretation stemmed from

- a thorough literature review where the metric use in manufacturing complex networks was evaluated;
- analysis of the results obtained when utilizing the metric on the Bosch manufacturing network; and
- the similarities between manufacturing networks and the problem for which the metric was originally derived.

In this work, it was stressed that the interpretation of complex network metrics has a strong domain-specific component. Caution was advised to practitioners. In fact, this work strongly suggests against "borrowing" interpretations from other (more developed) fields.

Along with each metric interpretation, the *shortcomings* of the suitable metrics were highlighted. In each Chapter, practitioners were advised on limitations and weaknesses that may have an effect on the applicability of the metrics in industrial settings. It was stressed that understanding what a metric cannot do is as important to ensure it is appropriately used.

Finally, these three aspects, suitability, interpretation and shortcomings, were discussed in Chapter 11 for all metrics. It was highlighted that, while matching underlying assumptions are sufficient to consider a metric suitable, exceptions do exist. However, when the underlying assumptions of a metric are removed from those of the complex network under study, the metric interpretation tends to be circumscribed. In addition, several limitations are easily identified indicating that their applicability is restricted. On the other hand, metrics whose underlying assumptions match those of manufacturing networks have a wider interpretation and open a number of research avenues.

12.2 Evaluation of objectives fulfillment

This work fulfilled the objectives listed in Chapter 4 and re-instated in the previous Section.

The suitability of popular existing complex network metrics in the context of manufacturing networks was discussed in Chapters 7 and 8. In fact, it was determined that in- and out-degree, in- and out-strength and the clustering coefficient were suitable while betweenness centrality and PageRank were not.

Other existing metrics whose underlying assumptions are better aligned with those of manufacturing networks were proposed and exemplified in Chapters 9 and 10. The probability distribution associated with the path-transfer flow entropy proved to be an interesting tool for evaluating routing flexibility in manufacturing. In addition, the flow fraction graph associated with solving the maximum flow problem of flow networks is adequate to identify limiting resources. In both cases, avenues for future research were identified and are discussed in detail in Chapter 13.

Finally, the appropriate Operations Research centered interpretation was provided for each suitable metric in their respective Chapter. In addition, each metric limitations and weaknesses were identified highlighting the shortcomings that practitioners may experience when deploying the metrics in industrial settings.

13 Future Work

This work has opened several avenues worthy of future research. These are listed hereafter.

Path-transfer flow entropy and clustering coefficient correlation: As noted in Chapter 7, the clustering coefficient is a measure of connectedness. It was proposed to use the subgraph clustering coefficient in order to measure the connectedness of strongly connected components, yet a local clustering coefficient for each node can also be determined. On the other hand, the path probability distribution associated with the path-transfer flow entropy centrality was used in Chapter 9 as a measure of routing flexibility. It can be argued that these two metrics have goals that share similarities. An interesting avenue for future research would entail the analysis of possible correlations between these metrics and clarification on the aspects in which they differ.

Potential of betweenness centrality to identify nodes that disconnect upstream suppliers from their downstream tier: As discussed in Chapter 7, in the Bosch manufacturing network example, the "important" nodes according to betweenness centrality were v_{29}, v_{30} and v_{39}. Observation of the manufacturing network graph in Fig. 5.3 led to the conclusion that these nodes can potentially disconnect upstream nodes from their downstream tier. Future research should determine whether this proposition holds true for all manufacturing networks.

Betweenness centrality based on paths instead of geodesics: In Chapter 7, it was explained that betweenness centrality measures the fraction of times in which a node v appears on the geodesic path σ between any two other nodes s and t. The metric was deemed unsuitable for manufacturing networks due to the fact that geodesics do not accurately represent the flow in manufacturing networks. Other authors have also recognized this shortcoming for other applications of the metric [110, 111]. Thus, future research should focus on the development, suitability and interpretation of a path-based betweenness centrality for manufacturing networks.

Path-transfer flow entropy - weighing paths considering their length: As noted in Chapter 9, all the paths in graph G are weighted equally when determining the path probability distribution p_{ij} between nodes v_i and v_j. However, this is unlikely to be a realistic representation of manufacturing networks were very short and very long paths are unlikely. In fact, in Appendix A it was shown that, for the Bosch manufacturing network, most manufactured items visited 8 or 13/14 workstations. Thus, further research should evaluate possible modifications to Tutzauer's formulation to properly weigh manufacturing paths according to their length.

A Exploratory Data Analysis on Raw Time-Stamp Data

As introduced in Chapter 5, the time-stamp data available for the Kaggle competition titled "Bosch Production Line Perfomance Competition"[71] was used to produce the manufacturing network studied. In this Appendix, an exploratory data analysis (EDA) on the raw data contained in the "train_date.csv" file is conducted. It provides information regarding:

- which workstations were visited by each manufactured item;
- the total number of workstations required to transform raw materials into finished goods;
- the time-stamp pattern and granularity;
- total manufacturing time for each recorded instance;
- span of time spent on each manufacturing line for each manufactured item.

In the following sections, this information is further analyzed to draw conclusions regarding the dataset studied.

A.1 Most visited workstations

Not all workstations are used at the same capacity. As observed in Figure A.1, while workstations v_{29}, v_{30}, v_{33}, v_{34} and v_{37} attend to approximately 95% of the manufactured items, workstations v_{42} and v_{46} process well below 1% of them.

A.2 Average number of visited workstations

In addition, the number of workstations required to process each manufactured good is not constant. As shown in Figure A.2, 71.2% of items require 13 or 14 workstations to be transformed from raw materials into finished goods, 19.5% pass through 8 workstations, and the remainder require any number between 0 (no data recorded) and 23 workstations.

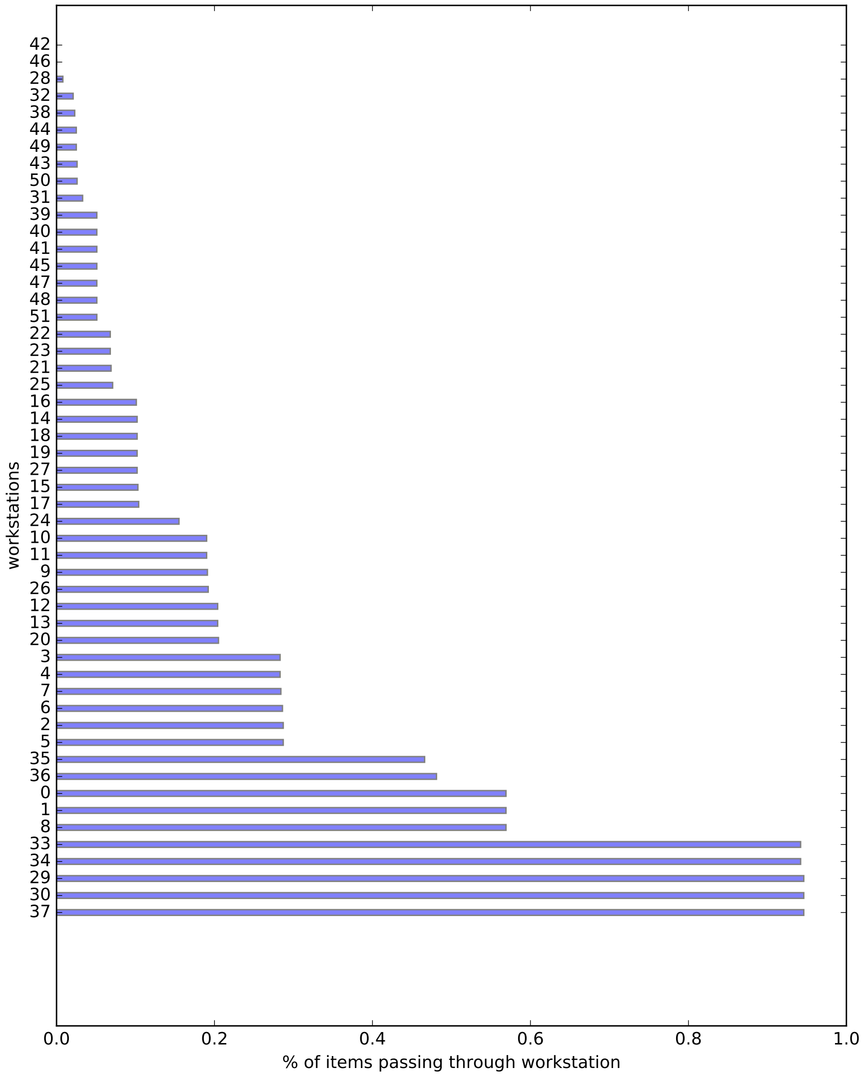


FIGURE A.1: Use frequency of each workstation in Bosch Production Line shows great heterogeneity in terms of usage.

A.3 Time-stamp

All time-stamp values in the dataset are anonymized. A plot showing the start time (`itime`) and the finish time (`ftime`) frequency is shown on Figure A.3. A few observations can be drawn:

- A periodic pattern seems likely.
- The time values vary between 0.0 and 1718.48 with granularity of 0.01.
- A gap can be observed in the mid-section of the plot.

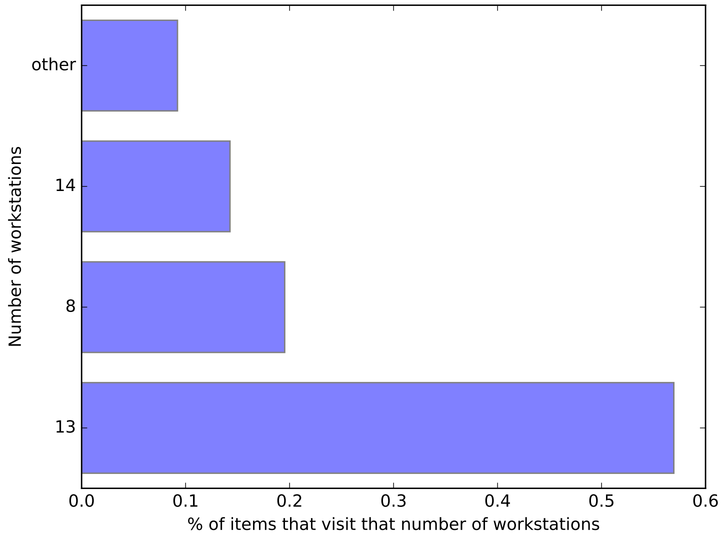


FIGURE A.2: Number of workstations required to transform raw materials into manufactured goods.

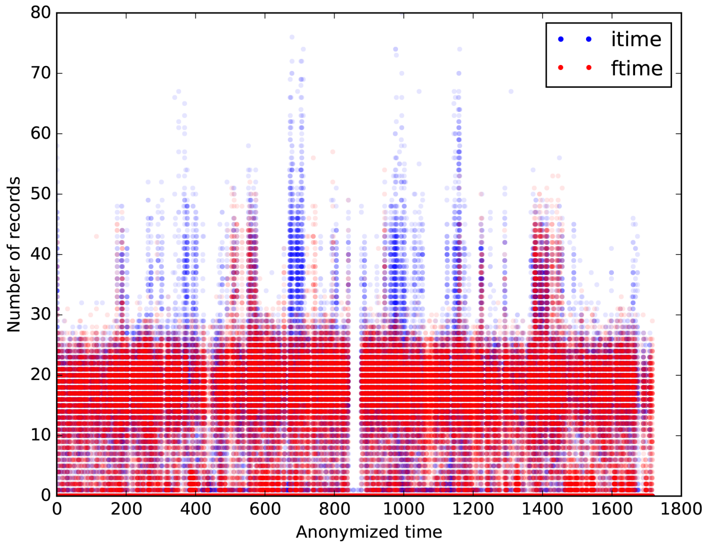


FIGURE A.3: Frequency graph for time-stamp values for both, `itime` which corresponds to the first time-stamp value recorded for each item (depicted in blue), and `ftime` which indicates the last time-stamp recorded for each item (colored in red).

The periodic pattern is further evidenced by calculating the auto-correlation coefficient [112] between frequency values for all possible time-stamp values, i.e. between 0.0 and 1718.48 with granularity of 0.01. As observed in Figure A.4, seven peaks can be counted between periodic maximums (indicated in red), whose period is ≈ 16.75. Assuming each of the inter-maximums peaks corresponds to one day of the week and periodic maximums account for weeks, the granularity of 0.01 would represent a time-frame of ≈ 6 minutes. Furthermore, under this assumption, the full dataset would account for 102.6 weeks ($\approx$ 2 years) of manufacturing data.

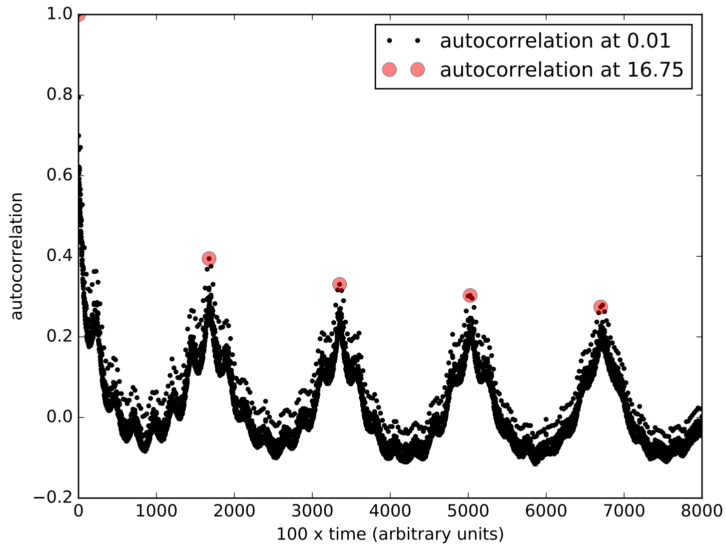


FIGURE A.4: Auto-correlation vs. time (in arbitrary units) showing periodic behavior. Given the existence of seven peaks between one periodic maximum and the next (indicated in red), it can be assumed that these periodic maximums account for weeks while intermediate peaks account for days. Since weeks (periodic maximums) are separated by a value of $\approx$ 16.75 (in anonymized units), it can be concluded that the granularity of the data (0.01) corresponds to $\approx$ 6 minutes.

A.4 Key performance figures

Some key performance figures are represented in Figure A.5 and Table A.1. The following observations can be extracted:

- For all represented figures, the values spread on multiple decades.
- While the required processing time in lines `L0`, `L1` and `L3` is roughly equivalent; it is considerably lower for line `L2`.
- The value adding time (VAT) is only a small portion of the total manufacturing time (TT). Using the mean values presented in Table A.1, the average lead-time is $10.72 \times 6 \approx 64$ minutes consisting of a value adding share of 5.4%.

[1] There are 357019 items processed in line `L2`. From these, only 301 were calculated to have spent a time of at least 0.01 in the line. The mean time in line `L2` is 0.002 ($\approx$ 1.2 minutes), i.e. the processing time is likely to be shorter than the data sampling time. Other possible explanation could be that the sensors' location is either at the beginning or end of the line and therefore, they do not appropriately record permanence time.

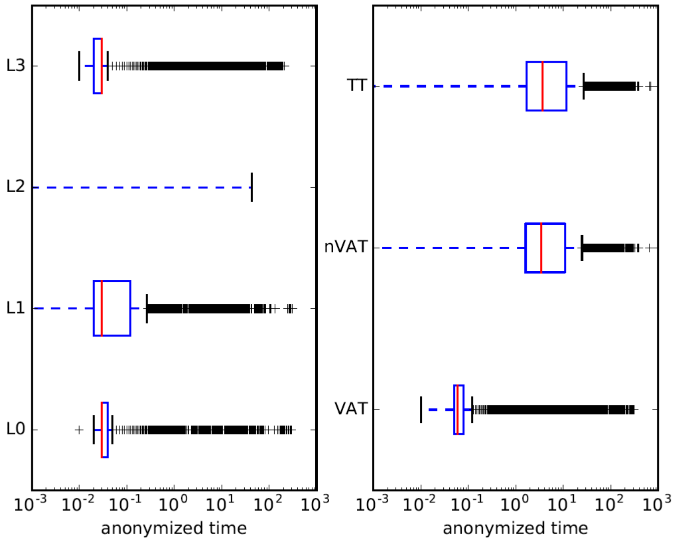


FIGURE A.5: *Left:* Boxplot representing manufacturing time in each processing line. *Right:* Boxplot representing total manufacturing time (TT), value adding time (VAT) and non-value adding time (nVAT)

TABLE A.1: Key performance figures for Bosch Manufacturing Line

	Min	1st quantile	Mean	3rd quantile	Max
L0	0.0	0.03	0.08	0.04	298.60
L1	0.0	0.02	0.77	0.12	321.38
L2[1]	0.0	0.00	0.00	0.00	42.80
L3	0.0	0.02	0.35	0.03	211.71
VAT	0.0	0.05	0.58	0.08	312.46
nVAT	0.0	1.61	10.14	10.93	662.27
TT	0.0	1.71	10.72	11.77	699.20

A.5 Conclusions

The EDA presented in this Appendix leads to the following conclusions:

- the time span of the dataset is of approximately 2 years;
- most items require either 8 or 13/14 workstations (out of 52 possible) to be fully processed;
- the workload of different workstations is highly heterogeneous; and
- the VAT is only a small fraction of the total processing time.

B Strongly Connected Components

B.1 Strongly connected components

As explained in Chapter 2, in directed graphs, a strongly connected component (SCC) can be defined as follows:

Definition 10. *A SCC in a directed graph $G = G(V, E)$ is a maximal set of vertices $C \subseteq V$ such that for every pair of vertices u and v in C, there exists both a path $u \rightsquigarrow v$ and a path $v \rightsquigarrow u$. Thus, u and v are reachable from each other.*

B.1.1 Application example

The SCC of the Bosch manufacturing network presented in Fig. 5.3 are listed in Table B.1.

TABLE B.1: Strongly connected components in the Bosch production line network in Fig. 5.3. In this table, the elements in each SCC are indicated using i for clarity. Note that $i \equiv v_i$.

Strongly connected components
$[0, 1, 2, 3]$
$[4, 5, 6, 7, 8, 9, 10, 11]$
$[12, 13, 14, 15]$
$[16, 17, 18, 19, 20, 21, 22, 23]$
$[24]$
$[25]$
$[26]$
$[27]$
$[28]$
$[29, 30, 31, 32, 33, 34, 35, 36, 37]$
$[38]$
$[39, 40, 41, 43, 44, 45, 47, 48, 49, 50, 51]$

Bibliography

[1] Eurostat, *Manufacturing statistics - NACE Rev. 2*, https://ec.europa.eu/eurostat/statistics-explained/index.php/Manufacturing_statistics_-_NACE_Rev._2, Online; accessed 2 March 2021, 2020.

[2] Y. M. Omar, M. Minoufekr, and P. Plapper, "Business analytics in manufacturing: Current trends, challenges and pathway to market leadership", *Operations Research Perspectives*, vol. 6, p. 100 127, 2019. DOI: 10.1016/j.orp.2019.100127.

[3] H. Kagermann, W.-D. Lukas, and W. Wahlster, "Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution", *VDI nachrichten*, vol. 13, p. 11, 2011.

[4] M. Hermann, T. Pentek, and B. Otto, "Design principles for industrie 4.0 scenarios", in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 3928–3937. DOI: 10.1109/HICSS.2016.488.

[5] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues", *Journal of Industrial Information Integration*, vol. 6, pp. 1–10, 2017. DOI: 10.1016/j.jii.2017.04.005.

[6] S. S. Kamble, A. Gunasekaran, and S. A. Gawankar, "Sustainable industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives", *Process Safety and Environmental Protection*, vol. 117, pp. 408–425, 2018. DOI: 10.1016/j.psep.2018.05.009.

[7] G. B. Pereira, A. d. P. Lacerda Santos, and M. G. Cleto, "Industry 4.0: Glitter or gold? a systematic review", *Brazilian Journal of Operations & Production Management*, vol. 15, no. 2, pp. 247–253, 2018. DOI: 10.14488/BJOPM.2018.v15.n2.a7.

[8] F. Galati and B. Bigliardi, "Industry 4.0: Emerging themes and future research avenues using a text mining approach", *Computers in Industry*, vol. 109, pp. 100–113, 2019. DOI: 10.1016/j.compind.2019.04.018.

[9] G. Culot, G. Nassimbeni, G. Orzes, and M. Sartor, "Behind the definition of industry 4.0: Analysis and open questions", *International Journal of Production Economics*, vol. 226, p. 107 617, 2020, ISSN: 0925-5273. DOI: 10.1016/j.ijpe.2020.107617.

[10] J. Rifkin, *The Third Industrial Revolution*. St. Martin's Press, 2011.

[11] The TIR Consulting Group LLC, *The Third Industrial Revolution Strategy Study*, https://goo.gl/VH73hU, Online; accessed 7 November 2017, 2016.

[12] P. C. Evans and M. Annunziata, *Industrial internet: Pushing the boundaries of minds and machines*, https://goo.gl/MR4eAP, General Electric, 2012.

[13] B. R. Reddy and A. Sujith, "A comprehensive literature review on data analytics in iiot (industrial internet of things)", *HELIX*, vol. 8, no. 1, pp. 2757–2764, 2018. DOI: 10.29042/2018-2757-2764.

[14] J. H. Nord, A. Koohang, and J. Paliszkiewicz, "The internet of things: Review and theoretical framework", *Expert Systems with Applications*, vol. 133, pp. 97–108, 2019. DOI: 10.1016/j.eswa.2019.05.014.

[15] H. Yang, S. Kumara, S. T. Bukkapatnam, and F. Tsung, "The internet of things for smart manufacturing: A review", *IISE Transactions*, pp. 1–27, 2019. DOI: 10.1080/24725854.2018.1555383.

[16] J. Davis, T. Edgar, J. Porter, J. Bernaden, and M. Sarli, "Smart manufacturing, manufacturing intelligence and demand-dynamic performance", *Computers & Chemical Engineering*, vol. 47, pp. 145–156, 2012. DOI: 10.1016/j.compchemeng.2012.06.037.

[17] H. S. Kang, J. Y. Lee, S. Choi, H. Kim, J. H. Park, J. Y. Son, B. H. Kim, and S. Do Noh, "Smart manufacturing: Past research, present findings, and future directions", *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 3, no. 1, pp. 111–128, 2016. DOI: 10.1007/s40684-016-0015-5.

[18] P. Zheng, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarok, S. Yu, X. Xu, *et al.*, "Smart manufacturing systems for industry 4.0: Conceptual framework, scenarios, and future perspectives", *Frontiers of Mechanical Engineering*, vol. 13, no. 2, pp. 137–150, 2018. DOI: 10.1007/s11465-018-0499-5.

[19] E. Brynjolfsson, L. M. Hitt, and H. H. K. Kim, "Strength in numbers: How does data-driven decisionmaking affect firm performance?", *Available at SSRN*, 2011. DOI: 10.2139/ssrn.1819486.

[20] T. May, *The new know: innovation powered by analytics*. John Wiley & Sons, 2009, ISBN: 978-0-470-46171-6.

[21] M. E. Porter and J. E. Heppelmann, "How smart, connected products are transforming competition", *Harvard Business Review*, vol. 92, no. 11, pp. 64–88, 2014.

[22] S. Williams, N. Williams, *et al.*, "The business value of business intelligence", *Business Intelligence Journal*, vol. 8, pp. 30–39, 2003.

[23] A. Gunasekaran, Y. Y. Yusuf, E. O. Adeleye, and T. Papadopoulos, "Agile manufacturing practices: The role of big data and business analytics with multiple case studies", *International Journal of Production Research*, vol. 56, no. 1-2, pp. 385–397, 2018. DOI: 10.1080/00207543.2017.1395488.

[24] B. Azvine, Z. Cui, D. D. Nauck, and B Majeed, "Real time business intelligence for the adaptive enterprise", in *E-Commerce Technology, 2006. The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, The 3rd IEEE International Conference on*, IEEE, 2006, pp. 29–29. DOI: 10.1109/CEC-EEE.2006.73.

[25] V. S. Yadavalli, J. D. Darbari, N. Bhayana, P. Jha, and V. Agarwal, "An integrated optimization model for selection of sustainable suppliers based on customers' expectations", *Operations Research Perspectives*, vol. 6, p. 100 113, 2019, ISSN: 2214-7160. DOI: 10.1016/j.orp.2019.100113.

[26] T. Gutowski, C. Murphy, D. Allen, D. Bauer, B. Bras, T. Piwonka, P. Sheng, J. Sutherland, D. Thurston, and E. Wolff, "Environmentally benign manufacturing: Observations from japan, europe and the united states", *Journal of Cleaner Production*, vol. 13, no. 1, pp. 1–17, 2005. DOI: 10.1016/j.jclepro.2003.10.004.

[27] J. Lee, B. Bagheri, and C. Jin, "Introduction to cyber manufacturing", *Manufacturing Letters*, vol. 8, pp. 11–15, 2016. DOI: 10.1016/j.mfglet.2016.05.002.

[28] S. Lizotte-Latendresse and Y. Beauregard, "Implementing self-service business analytics supporting lean manufacturing: A state-of-the-art review", *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1143–1148, 2018. DOI: 10.1016/j.ifacol.2018.08.436.

[29] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: Advantages, challenges, and applications", *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, 2016. DOI: 10.1080/21693277.2016.1192517.

[30] M. Sharp, R. Ak, and T. Hedberg, "A survey of the advancing use and development of machine learning in smart manufacturing", *Journal of Manufacturing Systems*, vol. 48, pp. 170–179, 2018, Special Issue on Smart Manufacturing, ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2018.02.004.

[31] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing", *Expert Systems with Applications*, vol. 166, p. 114 060, 2021, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2020.114060.

[32] A. Belhadi, K. Zkik, A. Cherrafi, S. M. Yusof, and S. El Fezazi, "Understanding Big Data Analytics for Manufacturing Processes: Insights from Literature Review and Multiple Case Studies", *COMPUTERS & INDUSTRIAL ENGINEERING*, vol. 137, 2019. DOI: {10.1016/j.cie.2019.106099}.

[33] S. Ren, Y. Zhang, Y. Liu, T. Sakao, D. Huisingh, and C. M.V. B. Almeida, "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions", *JOURNAL OF CLEANER PRODUCTION*, vol. 210, 1343–1365, 2019. DOI: {10.1016/j.jclepro.2018.11.025}.

[34] Y. Cui, S. Kara, and K. C. Chan, "Manufacturing big data ecosystem: A systematic literature review", *ROBOTICS AND COMPUTER-INTEGRATED MANUFACTURING*, vol. 62, 2020. DOI: {10.1016/j.rcim.2019.101861}.

[35] N. R. Council, *Network Science*. Washington, DC: The National Academies Press, 2005, ISBN: 978-0-309-10026-7. DOI: 10.17226/11516.

[36] D. L. Alderson, "Or forum—catching the "network science" bug: Insight and opportunity for the operations researcher", *Operations Research*, vol. 56, no. 5, pp. 1047–1065, 2008. DOI: 10.1287/opre.1080.0606.

[37] A. Nagurney, *Comment on "catching the network science bug" by david l. alderson*, 2008.

[38] S. P. Borgatti, "Centrality and network flow", *Social networks*, vol. 27, no. 1, pp. 55–71, 2005. DOI: 10.1016/j.socnet.2004.11.008.

[39] L. C. Freeman, "Centrality in social networks conceptual clarification", *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[40] F. Tutzauer, "Entropy as a measure of centrality in networks characterized by path-transfer flow", *Social networks*, vol. 29, no. 2, pp. 249–265, 2007. DOI: 10.1016/j.socnet.2006.10.001.

[41] L. C. Freeman, "A set of measures of centrality based on betweenness", *Sociometry*, pp. 35–41, 1977. DOI: 10.2307/3033543.

[42] P. Bonacich, "Power and centrality: A family of measures", *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[43] S Brin and L Page, "Anatomy of a large-scale hypertextual web search engine", in *Proc. 7th Intl. World-Wide-Web Conference*, 1998, pp. 107–117.

[44] Y. Li, F. Tao, Y. Cheng, X. Zhang, and A. Nee, "Complex networks in advanced manufacturing systems", *Journal of Manufacturing Systems*, vol. 43, pp. 409–421, 2017. DOI: 10.1016/j.jmsy.2016.12.001.

[45] T. Funke and T. Becker, "Complex networks of material flow in manufacturing and logistics: Modeling, analysis, and prediction using stochastic block models", *Journal of Manufacturing Systems*, vol. 56, pp. 296–311, 2020. DOI: 10.1016/j.jmsy.2020.06.015.

[46] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement", *PLoS med*, vol. 6, no. 7, e1000097, 2009. DOI: 10.1371/journal.pmed.1000097.

[47] R. Vrabič, D. Husejnagić, and P. Butala, "Discovering autonomous structures within complex networks of work systems", *CIRP Annals*, vol. 61, no. 1, pp. 423–426, 2012. DOI: 10.1016/j.cirp.2012.03.083.

[48] R. Vrabič, G. Škulj, and P. Butala, "Anomaly detection in shop floor material flow: A network theory approach", *CIRP Annals*, vol. 62, no. 1, pp. 487–490, 2013. DOI: 10.1016/j.cirp.2013.03.131.

[49] M. E. Beber and T. Becker, "Towards an understanding of the relation between topological characteristics and dynamic behavior in manufacturing networks", *Procedia CIRP*, vol. 19, pp. 21–26, 2014, 2nd CIRP Robust Manufacturing Conference (RoMac 2014). DOI: 10.1016/j.procir.2014.05.005.

[50] T. Becker, M. Meyer, and K. Windt, "A manufacturing systems network model for the evaluation of complex manufacturing systems", *International Journal of Productivity and Performance Management*, vol. 63, no. 3, pp. 324–340, 2014. DOI: 10.1108/IJPPM-03-2013-0047.

[51] H. Blunck, V. Vican, T. Becker, and K. Windt, "Improvement heuristics for manufacturing system design using complex network figures", *Procedia CIRP*, vol. 17, pp. 50–55, 2014. DOI: 10.1016/j.procir.2014.01.063.

[52] X. Ai-ming, G. Jian-min, and C. Kun, "Excavation of critical resource node for quality control of multi-variety mixed production shopfloor based on complex network property", *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 230, no. 1, pp. 169–177, 2016. DOI: 10.1177/0954405414558696.

[53] T. Becker and D. Wagner, "Identification of key machines in complex production networks", *Procedia CIRP*, vol. 41, pp. 69–74, 2016. DOI: 10.1016/j.procir.2015.12.006.

[54] D. Wagner and T. Becker, "Characterization of cluster structures in material flow networks: A network approach", *Procedia Technology*, vol. 26, pp. 277–283, 2016. DOI: 10.1016/j.protcy.2016.08.037.

[55] ——, "Evaluation of the applicability of random walks for generation of material flow network models", *Procedia CIRP*, vol. 63, pp. 488–492, 2017, Manufacturing Systems 4.0 – Proceedings of the 50th CIRP Conference on Manufacturing Systems. DOI: 10.1016/j.procir.2017.03.164.

[56] S. Chankov, M.-T. Hütt, and J. Bendul, "Understanding synchronizability of manufacturing networks: A multi-method study on structural network properties", *Journal of Manufacturing Systems*, vol. 46, pp. 127–136, 2018. DOI: 10.1016/j.jmsy.2017.11.007.

[57] T. Funke and T. Becker, "Stochastic block models as a modeling approach for dynamic material flow networks in manufacturing and logistics", *Procedia CIRP*, vol. 72, pp. 539–544, 2018, 51st CIRP Conference on Manufacturing Systems. DOI: 10.1016/j.procir.2018.03.209.

[58] Y. M. Omar, M. Minoufekr, and P. Plapper, "Lessons from social network analysis to industry 4.0", *Manufacturing Letters*, vol. 15, pp. 97 –100, 2018. DOI: https://doi.org/10.1016/j.mfglet.2017.12.006.

[59] D. Wagner and T. Becker, "Automated generation of material flow network models with cluster structures using random walks", *Procedia CIRP*, vol. 72, pp. 569–573, 2018, 51st CIRP Conference on Manufacturing Systems. DOI: https://doi.org/10.1016/j.procir.2018.03.249.

[60] T. Funke and T. Becker, "Forecasting changes in material flow networks with stochastic block models", *Procedia CIRP*, vol. 81, pp. 1183–1188, 2019, 52nd CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019. DOI: 10.1016/j.procir.2019.03.289.

[61] Y. Omar and P. Plapper, "Maximum flow of complex manufacturing networks", *Procedia CIRP*, vol. 86, pp. 245–250, 2019. DOI: 10.1016/j.procir.2020.01.005.

[62] F. Zhu, R. Wang, and C. Wang, "Intelligent workshop bottleneck prediction based on complex network", in *2019 IEEE International Conference on Mechatronics and Automation (ICMA)*, IEEE, 2019, pp. 1682–1686. DOI: 10.1109/ICMA.2019.8816432.

[63] Y. M. Omar and P. Plapper, "Entropy of complex manufacturing networks as a metric of flexibility", *Under Review*, 2021.

[64] Y. Kim, T. Y. Choi, T. Yan, and K. Dooley, "Structural investigation of supply networks: A social network analysis approach", *Journal of Operations Management*, vol. 29, no. 3, pp. 194–211, 2011. DOI: 10.1016/j.jom.2010.11.001.

[65] Y. Kim, Y.-S. Chen, and K. Linderman, "Supply network disruption and resilience: A network structural perspective", *Journal of operations Management*, vol. 33, pp. 43–59, 2015. DOI: 10.1016/j.jom.2014.10.006.

[66] V. S. Adami, J. R. Verschoore, and M. A. Sellitto, "Structure and complexity in six supply chains of the brazilian wind turbine industry", *The International Journal of Logistics Management*, vol. 32, no. 1, pp. 23–39, 2020. DOI: 10.1108/IJLM-01-2020-0039.

[67] A. K. Lau, Y. Kajikawa, and N. Sharif, "The roles of supply network centralities in firm performance and the moderating effects of reputation and export-orientation", *Production Planning & Control*, vol. 31, no. 13, pp. 1110–1127, 2020. DOI: 10.1080/09537287.2019.1700569.

[68] A. Potter and M. Wilhelm, "Exploring supplier–supplier innovations within the toyota supply network: A supply network perspective", *Journal of Operations Management*, vol. 66, no. 7-8, pp. 797–819, 2020. DOI: 10.1002/joom.1124.

[69] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, no. 5439, pp. 509–512, 1999. DOI: 10.1126/science.286.5439.509.

[70] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications", *Phys. Rev. E*, vol. 64, p. 026 118, 2 2001. DOI: 10.1103/PhysRevE.64.026118.

[71] Bosch and Kaggle, *Bosch production line performance competition*, https://www.kaggle.com/c/bosch-production-line-performance, 2016.

[72] Y. M. Omar, *Code for phd thesis*, https://github.com/YamiOmar88/Code_PhD_Thesis, 2021.

[73] ——, *Automated product family selection - Python 3 code*, https://github.com/YamiOmar88/product-family-selection, 2021.

[74] ——, *Manufacturing entropy - Python 3 code*, https://github.com/YamiOmar88/manufacturing_entropy_article, 2021.

[75] K. Erlach, *Value stream design - The way towards a Lean Factory*. Springer, 2013.

[76] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics", *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006.

[77] F. Harary, R. Z. Norman, and D. Cartwright, *Structural models: An introduction to the theory of directed graphs*. John Wiley & Sons Inc, 1965.

[78] U. Brandes, "A faster algorithm for betweenness centrality", *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001. DOI: 10.1080/0022250X.2001.9990249.

[79] ——, "On variants of shortest-path betweenness centrality and their generic computation", *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008. DOI: 10.1016/j.socnet.2007.11.001.

[80] ——, *Erratum - on variants of shortest-path betweenness centrality and their generic computation*, https://web.archive.org/web/20100518174803/http://www.inf.uni-konstanz.de/~brandes/publications/#journal.

[81] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. DOI: 10.1038/30918.

[82] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, "Generalizations of the clustering coefficient to weighted complex networks", *Physical Review E*, vol. 75, no. 2, p. 027 105, 2007. DOI: 10.1103/PhysRevE.75.027105.

[83] G. Fagiolo, "Clustering in complex directed networks", *Physical Review E*, vol. 76, no. 2, p. 026 107, 2007. DOI: 10.1103/PhysRevE.76.026107.

[84] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing social networks*. Sage, 2013.

[85] F. Oggier, S. Phetsouvanh, and A. Datta, "A split-and-transfer flow based entropic centrality", *PeerJ Computer Science*, vol. 5, e220, 2019. DOI: 10.7717/peerj-cs.220.

[86] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine", *Computer Networks*, vol. 56, no. 18, pp. 3825 –3833, 2012, ISSN: 1389-1286. DOI: http://dx.doi.org/10.1016/j.comnet.2012.10.007.

[87] J. A. Gutiérrez-Pérez, M. Herrera, R. Pérez-García, and E. Ramos-Martínez, "Application of graph-spectral methods in the vulnerability assessment of water supply networks", *Mathematical and Computer Modelling*, vol. 57, no. 7, pp. 1853–1859, 2013.

[88] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 06, P06010, 2007.

[89] B. Scholz-Reiter, F. Wirth, S. Dashkovskiy, T. Makuschewitz, M. Kosmykov, and M. Schönlein, "Application of the pagerank algorithm for ranking locations of a production network", in *Proceedings of 42nd CIRP Conference on Manufacturing Systems*, 2009.

[90] B. Scholz-Reiter, F. Wirth, S. Dashkovskiy, M. Kosmykov, T. Makuschewitz, and M. Schönlein, "An approach to model reduction of logistic networks based on ranking", in *Dynamics in Logistics*, Springer, 2011, pp. 91–103.

[91] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2014.

[92] W. To, "Centrality of an urban rail system", *Urban Rail Transit*, vol. 1, no. 4, pp. 249–256, 2015. DOI: 10.1007/s40864-016-0031-3.

[93] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[94] H. Zenil, N. Kiani, and J. Tegnér, "A review of graph and network complexity from an algorithmic information perspective", *Entropy*, vol. 20, no. 8, p. 551, 2018. DOI: 10.3390/e20080551.

[95] F. Tutzauer and B. Elbirt, "Entropy-based centralization and its sampling distribution in directed communication networks", *Communication Monographs*, vol. 76, no. 3, pp. 351–375, 2009. DOI: 10.1080/03637750903074727.

[96] F. Oggier, S. Phetsouvanh, and A. Datta, "Entropic centrality for non-atomic flow networks", in *2018 International Symposium on Information Theory and Its Applications (ISITA)*, IEEE, 2018, pp. 50–54. DOI: 10.23919/ISITA.2018.8664236.

[97] Y. M. Omar and P. Plapper, "A survey of information entropy metrics for complex networks", *Entropy*, vol. 22, no. 12, p. 1417, 2020. DOI: 10.3390/e22121417.

[98] A. G. Nikolaev, R. Razib, and A. Kucheriya, "On efficient use of entropy centrality for social network analysis and community detection", *Social Networks*, vol. 40, pp. 154–162, 2015. DOI: 10.1016/j.socnet.2014.10.002.

[99] S. Varrette, P. Bouvry, H. Cartiaux, and F. Georgatos, "Management of an academic HPC cluster: The UL experience", in *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, Bologna, Italy: IEEE, 2014, pp. 959–967. DOI: 10.1109/HPCSim.2014.6903792.

[100] J. Browne, D. Dubois, K. Rathmill, S. P. Sethi, K. E. Stecke, *et al.*, "Classification of flexible manufacturing systems", *The FMS magazine*, vol. 2, no. 2, pp. 114–117, 1984.

[101] A. M. Kuchling, *Functional programming howto*, https://docs.python.org/3/howto/functional.html, Accessed February 2021.

[102] *Pypy - a fast, compliant alternative implementation of Python*, https://www.pypy.org/.

[103] Fonseca, José, *gprof2dot Python script*, https://github.com/jrfonseca/gprof2dot, Online; accessed 19 November 2019.

[104] T. H. Cormen, C. E. Leiserson, R. R. L., and S. Clifford, *Introduction to algorithms*, Third Edition. MIT press, 2009.

[105] P. H. Pittman and J. B. Atwater, *APICS Dictionary: The Essential Supply Chain Reference, Fifteenth Edition*. APICS, Chicago, 2016.

[106] J. Bentley, "Programming pearls: Algorithm design techniques", *Communications of the ACM*, vol. 27, no. 9, pp. 865–873, 1984. DOI: 10.1145/358234.381162.

[107] L. R. Ford Jr and D. R. Fulkerson, *Flows in Networks*. Princeton University Press, 1962.

[108] E. F. Moore, "The shortest path through a maze", in *Proceedings of the International Symposium on the Theory of Switching*, Harvard University Press, 1959, pp. 285–292.

[109] R. Conway, W. Maxwell, J. O. McClain, and L. J. Thomas, "The role of work-in-process inventory in serial production lines", *Operations research*, vol. 36, no. 2, pp. 229–241, 1988. DOI: 10.1287/opre.36.2.229.

[110] M. J. Newman, "A measure of betweenness centrality based on random walks", *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005. DOI: 10.1016/j.socnet.2004.11.009.

[111] K. Jiang, D. Ediger, and D. A. Bader, "Generalizing k-betweenness centrality using short paths and a parallel multithreaded implementation", in *2009 International Conference on Parallel Processing*, 2009, pp. 542–549. DOI: 10.1109/ICPP.2009.76.

[112] A. Mangal and N. Kumar, "Using big data to enhance the bosch production line performance: A kaggle challenge", in *Big Data (Big Data), 2016 IEEE International Conference on*, IEEE, 2016, pp. 2029–2035. DOI: 10.1109/BigData.2016.7840826.