31st Annual INCOSE international symposium
virtual event
July 17 - 22, 2021

Developing a Topic Network of Published Systems Engineering Research

Dr Rudolph Oosthuizen
Graduate School of Technology Management, University of Pretoria and CSIR
South Africa
+27827336355
rudolph.oosthuizen@up.ac.za

Abstract. This paper investigates past systems engineering research structure by developing a network of research topics from an extensive corpus of publications. The bibliometric data analysis from the published research articles provides valuable information on past progress and future scientific discipline opportunities. Topic modelling, a form of unsupervised machine-learning-based natural language processing, is applied to extract the main topics from the titles and abstracts of a wide range of papers published about systems engineering. The co-occurrence of research topics in these papers provides the data for generating network diagrams. A visual and quantitative network analysis of these topics revealed several cliques and clusters of research topics in systems engineering. Systems engineering researchers should consider these relationships between these research topics to plan a systems engineering research project.

Introduction

Since its formalization after the Second World War, systems engineering has grown into the multidisciplinary discipline and profession that it is today (Valerdi & Davidz 2009; Brill 1998). However, relative to other engineering fields, systems engineering is still relatively young and growing. Systems engineering requires focused research to continue developing theories, supported by processes and tools, to cope with engineering projects' increasing complexity (Sage 1998). Progress in a scientific field requires focused research to ensure long-term intellectual establishment. Evaluation of the research topic landscape structure and publication trends in a research field may detect meaningful research opportunities and requirements (Eker et al. 2019; Antons, Kleer & Salge 2016).

One way of investigating the relationships between research topics in a scientific field is to construct a network diagram. Network analysis of its structure may reveal connections between research topics to identify possible opportunities and research focus areas (Borgatti et al. 2009). This paper aims to extract and analyze the research topics in a wide range of articles from research published about systems engineering. The co-occurrence statistics of these topics provide the data for constructing topic network diagrams.

This paper will first provide background about the bibliometric analysis method to trace a research field's development and evolution. The research in this paper implements machine-learning-based natural language processing (NLP) due to difficulties experienced by bibliometric analysis of an extensive set of documents. This paper's research implemented an automatic software-based process to capture bibliometric data from published research on systems engineering to extract topics through topic modelling. The dominant research topics were then processed to generate network diagrams for deeper analysis. This research's output identifies the systems engineering research structure based on the most prominent relationships between these topics.

Bibliometrics and Natural Language Processing

The term bibliometrics is etymologically composed of "Biblio" and "Metrics". "Biblio" is the Greek word for books, and the term "Metrics" refers to measurement. Therefore, bibliometrics refers to the quantitative and systematic statistical analysis of words in books or documents to visualize patterns and trends. The bibliometric analysis relies on the assumption that scientific research produces knowledge that is published in the scientific literature (Jia et al. 2018; Jiang, Qiang & Lin 2016). Scientists codify research outputs in publications as the building blocks of science. The publication of scientific research is the primary source for transfer knowledge. Peer reviews provide validation for published papers (Keathley et al. 2015). Bibliometric analysis is suited to most scientific, engineering, and behavioral research fields and is increasing in popularity for evaluating and mapping scientific outputs (Kalantari et al. 2017) (Chen & Xie 2020).

Extracting research topics from the published text in scientific papers is one form of bibliometric analysis. Historically, topics from papers were manually assigned to a predetermined list, based on the researchers' subjective judgment or subject-matter experts (Eker et al. 2019; Keathley et al. 2015; Kalantari et al. 2017; Jia et al. 2018). Over the past few decades, the unprecedented growth and accessibility of scientific publications create an information overload of researchers' textual data. In this case, a manual process may suffer from bias and may miss the underlying latent topics. Due to systems engineering's interdisciplinary nature, papers may also cover multiple topics (Gretarsson et al. 2012).

NLP with machine-learning algorithms enables the automated processing of text. NLP can extract keywords, relationships between words and text clusters to categorize, summarize or classify documents(Agrawal, Fu & Menzies 2018). Further processing may identify patterns to extract the data's underlying structures and relationships (Lee & Kang 2018) (Chen et al. 2019; Sohn et al. 2018). Topic modelling is a popular approach for bibliometric analysis as part of NLP. These algorithms were initially developed to improve internet search algorithms. Topic modelling is a quantitative statistical and unsupervised machine-learning technique to perform automated semantic latent topics extraction from an extensive set of documents (Jussila et al. 2017; Blei, Carin & Dunson 2010).

Latent Dirichlet Allocation (LDA) is a popular topic modelling algorithm for NLP to discover semantic latent topics and structures (Chen et al. 2019). The algorithm assumes that the mixture of words in a document contains a set of latent topics where each topic has its probability distribution over words, as intended by the author (Kim & Kang 2018; Jiang, Qiang & Lin 2016; Kunc, M. and Mortenson, M. and Vidgen 2012). The model is learned from the text documents by approximately inferring the posterior distributions of parameters from the word vectors (Gretarsson et al. 2012). Topic modelling requires no training data or labels. The algorithm identifies major thematic clusters from the text with a probability distribution over a collection of topics, with each topic having a cluster of words (Suominen, Arho and Toivanen, 2016). The algorithm then analyses the occurrence of hidden relationships between these words to define topics as a probability distribution over the selected terms (Kim & Kang 2018; Jiang, Qiang & Lin 2016; Kunc, M. and Mortenson, M. and Vidgen 2012).

Topic modelling is more comprehensive and faster than other manual methods. The output also enables researchers to search and filter the corpus of papers for focusing more in-depth analysis of critical elements during an exploratory literature review (Agrawal, Fu & Menzies 2018; Kunc, M. and Mortenson, M. and Vidgen 2012). This paper's research implements network analysis to structure and analyze the extracted research topics to investigate research in the systems engineering domain.

Network Analysis

Network and graph theory provides a visual and mathematical process for investigating a network structure based on nodes and edges (Alnajem, Mostafa & ElMelegy 2020; Borgatti et al. 2009). Network visualizations are helpful for the qualitative assessment of these networks. The graphical representation of the network reveals the structural relationships between the nodes (Hevey 2018). Variations in the visual representation may highlight different attributes, and the emerging patterns of interaction. (Alnajem, Mostafa & ElMelegy 2020; Borgatti et al. 2009; Hao et al. 2018).

The network map visualizes the degree of interconnectedness amongst the semantic research topics. The connecting edges' relative strength highlights the different densities in the network (Smiraglia, 2015). Network analysis also includes a mathematical analysis of the nodes and their relationships to quantify network parameters using graph theory concepts. Typical metrics include the centrality, distance between nodes, number of interactions, and the number of bridged interactions between clusters of nodes. These metrics quantify relationships, the ranking of nodes, and relationships (Borgatti et al. 2009).

Method

The research process applied in this paper, as seen in Figure 1, evaluated and analyzed the titles and abstracts from articles published in the field of systems engineering. Python-based algorithms implement and execute the steps in the process. The authors also successfully implemented this process to assess the research trends for technology management and systems engineering research (Oosthuizen and Pretorius, 2020).

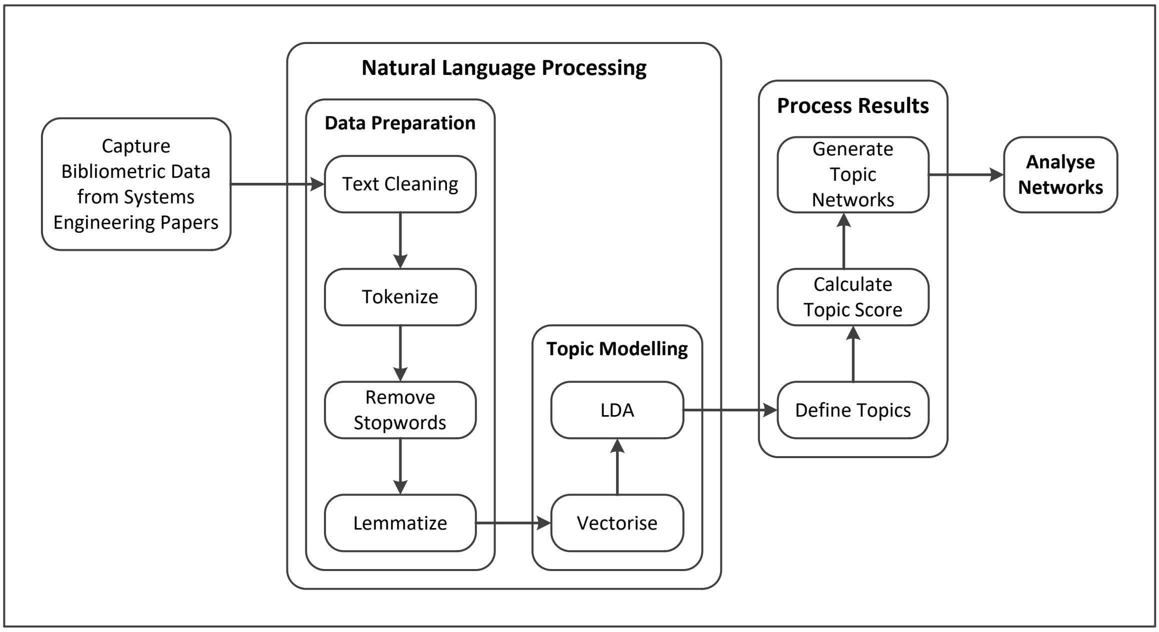

Figure 1: Bibliometric Analysis Process

Data Capture

Automated topic modelling for analysis of publications in a research field requires bibliometric data in the correct text (digital) format. Therefore, a prerequisite for effective topic modelling is access to bibliometric data (e.g., authors, year, affiliation, title, abstract, keywords) of publications in the research field in focus. The Pyblometrics library in Python, from Rose and Kitchin (Rose & Kitchin 2019), provides seamless interfacing to the Scopus library to access bibliometric data from paper retrieved using the crude and basic search terms of "systems engineering". When publishing this paper, the authors could not access a similar tool to retrieve data from other academic publication repositories (i.e. Web of Science). However, the number of papers retrieved should provide an adequate

The application retrieved 68 572 documents for processing. After removing the documents not containing the required information and limiting the range from 1990 onwards, a total of 62 136 papers remained for processing. The publication dates of these papers ranged from 1938 to 2021. Since Isoaho, Gritsenko and Mäkelä (2019) found that effective topic modelling requires at least 1000–2000 documents of about 100–200 words in length, this corpus should be adequate for topic modelling.

Data Preparation

Since abstracts are relatively short, combining the titles and abstracts also increase the size of the text corpus for improved processing and topic extraction. The title contains the most representative words for the article. Abstracts summarize the problem and the research results (Yeo & Jeong 2020; Agrawal, Fu & Menzies 2018). The captured bibliometric text from the papers on "systems engineering" requires cleaning, preparation and structuring to be suitable for topic modelling. The abstracts often contain publisher and copyright information and special and spurious characters due to the text retrieval application in Python. Text preparation also includes the following NLP steps (Agrawal, Fu & Menzies 2018; Ferris 2009) (Lin et al. 2016):

- Convert essential systems engineering terms or phrases into their abbreviated forms to retain their semantic meaning (e.g. "system of systems" to "SoS" and "model-based systems engineering" to "MBSE").
- Normalize spelling differences of keywords between the United Kingdom and the United States English (e.g., 'modelling' vs 'modeling', 'behaviour' vs 'behavior').
- Convert all the text to lowercase and remove punctuation, letters, and numbers as they may affect processing the text.
- Tokenize the text using spaces to extract the linguistic units from sentences or paragraphs.
- Remove stop words, which are common and frequent in a language and not adding real meaning to the text (e.g. "and", "the", "if", "a", etc.). These stop words tend to have a high frequency, adding noise and increasing the dimensionality during text vectorization.
- Lemmatize the text to reduce its dimensionality to combine similar terms. Lemmatization applies a vocabulary with a morphological analysis to transform the base word.
- Remove words of an academic nature that appear commonly in an abstract (e.g. "address", "represent", "aim", "paper", "review", "article", "author", "discuss", etc.) that will not contribute to extracting research topics in the field of systems engineering.

Topic Modelling

The Scikit-learn library in Python provides the LDA algorithm to perform topic modelling in this research. The first step is to vectorize the text into a document term matrix. Vectorisation may also extract common phrases if certain words frequently co-occur, which may have a meaning independent of the individual words. The LDA algorithm processes the document term matrix to produce the following outputs:

- The required number of topics with their describing terms (words and phrases).
- The importance of the terms to each topic.

- The probability of each topic to be present in each document.
- The perplexity of the topic model.

The primary input parameter to be defined for the LDA algorithm is the required number of topics for extraction. The algorithm then associates each of the topics to each of the documents in the text corpus. Expert and domain knowledge is still required to name and describe each of these topics (Eker et al. 2019; Agrawal, Fu & Menzies 2018).

Because the algorithm implements a random number generator to initiate training the LDA, different runs with the same text input may result in slightly different output topic sets. Determining an optimal parameter set to stabilize the LDA model is not easy with such an unsupervised, data-driven algorithm. There is not yet a standard approach to evaluate these models. Perplexity is an important measure to determine the statistical goodness of fit, or quality, of the topic model (Blei, Carin & Dunson 2010). A lower perplexity indicates a better and perhaps more appropriate LDA model (Agrawal, Fu & Menzies 2018; Kunc, M. and Mortenson, M. and Vidgen 2012).

Selecting a high topic number may generate noise in the form of spurious and similar topics, while too few topics result in complex and overlapping sets (Shin et al. 2018). Due to the large size and wide range of possible topics, the number of topics required was set to 25 for this research. The discussion of the results will cover the identification and processing of the extracted topics.

Post Processing

The topics were analyzed and identified through inspection and compared to a previous detailed topic modelling research publication of the author on INCOSE's primary publication platforms, namely the Systems Engineering Journal and international symposia proceedings (Oosthuizen and Pretorius, 2020). In that paper, a narrow and controlled sample of 622 systems engineering journal papers and 3 694 papers for the symposium proceedings provided a set of 20 systems engineering research topics. A team of systems engineering experts identified and validated the topics. These topics offer a baseline to group and assign the topics from the research in this paper. For this paper, the number of topics was increased to 25 due to the text corpus's increased size to extract possible additional topics.

The LDA output also includes the distribution probability of each topic per paper. Therefore, it is possible to count the total distribution per topic over all the text corpus documents. The extracted topics provide the nodes for the network diagram to be constructed. The co-occurrence probabilities of all topics per paper are summed to provide the relationship or edges between the topics for creating a network diagram. This approach is similar to the method applied by Zhang et al. (2017). The free software, Gephi, then visualizes the node and edge data in a network diagram for further analysis. Detailed network analysis was performed with Social Network Visualiser (SocNetV) 2.8.

Results and Discussion

Systems Engineering Research Topics

Table 1 presents the comparison between the initial (baseline) systems engineering research topics from INCOSE's journal and symposium proceedings (Oosthuizen and Pretorius, 2020) to the topics extracted from the larges text corpus in this paper. Inspecting the new topic terms from this research, the author aligned them with the previously validated (baseline) topics. The topics from the research for this paper compares well with the baseline topics. The only baseline topics not present in the current research topics list are Risk management and Systems Engineering Management. This may be attributed to the slightly different focus of the corpus of documents in this research.

Table 1: Comparison of baseline (Oosthuizen and Pretorius, 2020) to newly extracted topics

No	Baseline Topics	New Topics	New Topic Terms
1	Architecting	Architecting	architecture, component, reliability, system, base, application, approach, implementation, analyze,
2	Modelling	Modelling	modeling, base, system, approach, analyze, development, methodology, application, use, tool
3	Risk management		
4	Integration	Integration	environment, strategy, effectiveness, impact, base, analyze, integration, development, performance,
5	Systems Engineering Methodology	Systems Engineering Methodology	methodology, evaluation, measurement, analyze, base, system, effectiveness, application, technique,
6	System Operation	System Operation	management, operational, system, integration, implementation, development, application, analyze, information, base
7	Systems Engineering Capability	Systems Engineering Capability	engineering, system, system engineering, development, approach, application, process, industry, concept, complex
8	Project management	Project management	project, planning, development, implementation, management, industry, system, work, approach, base
9	Systems Engineering Practice	Systems Engineering Practice	problem, solution, methodology, base, application, effectiveness, system, approach, technique, complex
10	Tools and Cost	System Cost	cost, activity, effectiveness, impact, life, system, development, decision, process, performance
11	Systems Engineering Management		
12	Systems Engineering Processes	Systems Engineering Processes	process, business, modeling, approach, activity, base, analyze, implementation, application,
13	Decision Support and Frameworks	Decision Support and Frameworks	information, technology, decision, development, system, application, base, use, analyze, integration
14	System Life Cycle	System Life Cycle	base, methodology, maintenance, accuracy, life, modeling, effectiveness, analyze, application,
15	Software Engineering	Software Engineering	software, tool, development, application, system, technique, use, implementation, engineering,
16	Complexity	Complexity	complex, knowledge, analyze, relationship, system, base, information, concept, approach, domain
17	Requirements	Requirements	requirement, system, development, integration, capability, process, industry, engineering, technology
18	Product development	Product development	product, development, industry, process, life, analyze, approach, methodology, base, management

No	Baseline Topics	New Topics	New Topic Terms
19	Design	Design	design, system, development, implementation, approach, integration, engineering, process, base
20	Education and Program Management	Education	learning, task, domain, impact, base, approach, application, work, performance, use
21		Quality	quality, efficiency, work, effectiveness, system, use, performance, analyze, application, base
22		Failure Analysis	system, failure, analyze, reliability, effectiveness, operational, base, methodology, application
23		Measures of Effectiveness	performance, dynamic, base, analyze, effectiveness, application, technique, evaluation, methodology,
24		Human Factors	approach, user, structure, base, application, technique, use, analyze, implementation, requirement
25		Services	service, industry, application, base, approach, architecture, implementation, business, development,
26		System Analysis	Function, value, base, analyze, effectiveness, methodology, approach, modeling, use, relationship
27		Simulation	simulation, theory, methodology, base, effectiveness, analyze, application, modeling, development,

However, some of the new topics from this paper could not be confidently linked to the existing set of topics; these generated new topics to map systems engineering research. The discrepancy may be attributed to the more generic sources of systems engineering research processed in this paper. However, the comparison in Table 1 shows the usefulness of implementing an unsupervised machine-learning approach for topic modelling. Figure 2 shows the sum of the probabilities of allocating each topic to each paper. This indicates the popularity of the topics over the whole text corpus and will serve as the node strength in the network analysis later in this paper.

The topic allocation of systems engineering capability is so high because the papers published about systems engineering's ultimate aim are to grow the field and establish a systems engineering capability. This may be seen as the overarching or high-level topic being built up by the other topics. The second-tier topics, methodology, modelling, and failure analysis, may be seen as critical elements to establish a systems engineering capability. The remainder of the topics can be characterized as the building blocks of a systems engineering capability.

Generating a Topic Network

Although the topics extracted from systems engineering literature seem interesting, their actual contribution will surface through additional processing to extract trends and patterns. However, this information will be processed to derive a structure of the systems engineering domain in this paper. This structure should assist analysis for deeper insight into the systems engineering research domain. Outputs from this analysis may identify research opportunities as well as developing educational curricula.

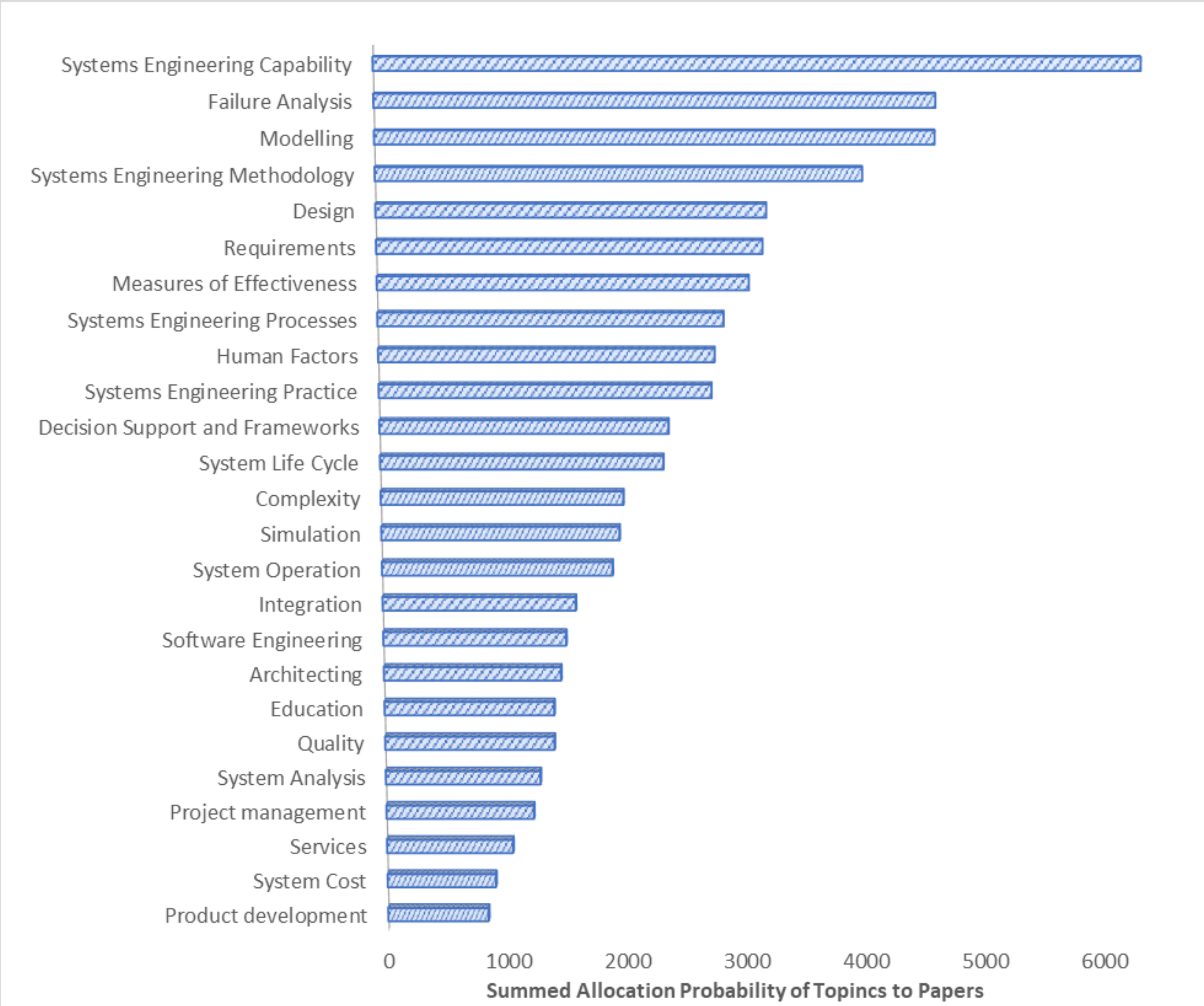

Figure 2: Topic Allocation to Papers

As discussed before, a network consists of nodes and edges. The topics extracted in this paper provide the nodes for the network diagram. The edges for the network are calculated by summing the co-occurrence probabilities of topics per paper. The output of this process is an N2 matrix consisting of the nodes and their edges. Figure 2 shows the basic network diagram of all the nodes and their edges. Every node is connected to all the other nodes in this network, resulting in a fully connected or mesh network topology. Because the LDA maps all papers to each topic, even a minimal probability eventually contributes to a significant value for every edge when summing over 62 136 papers. A substantial contributor to this structure may be the interdisciplinary nature of systems engineering. The topic co-occurrence values indicate the edges' relative strengths. The nodes' level of degree centrality determines their sizes. The prominent topics (nodes) include methodology, design, human factors, processes and modelling.

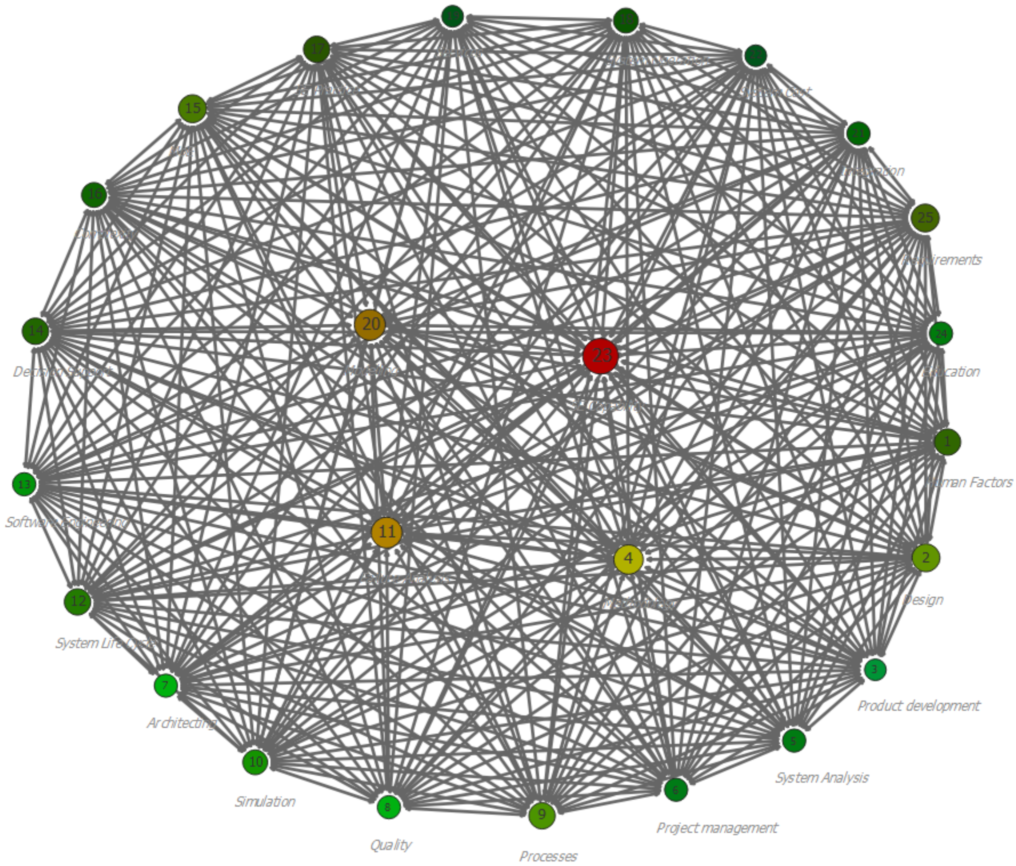

Figure 3: Complete Topic Network Graph

The network in Figure 2 is too dense for easy analysis because all the nodes are connected. Some of the thicker lines (edges) indicate some stronger connections. One solution is to filter out the less significant links. The network diagram of Figure 3 shows only the links with a value of 50% and higher of the maximum link value (5000). This network already indicates some discernible patterns. The visual inspection suggests that the critical nodes within processed system engineering papers are slightly different from Figure 2 as many of the lesser nodes lost most of their connections.

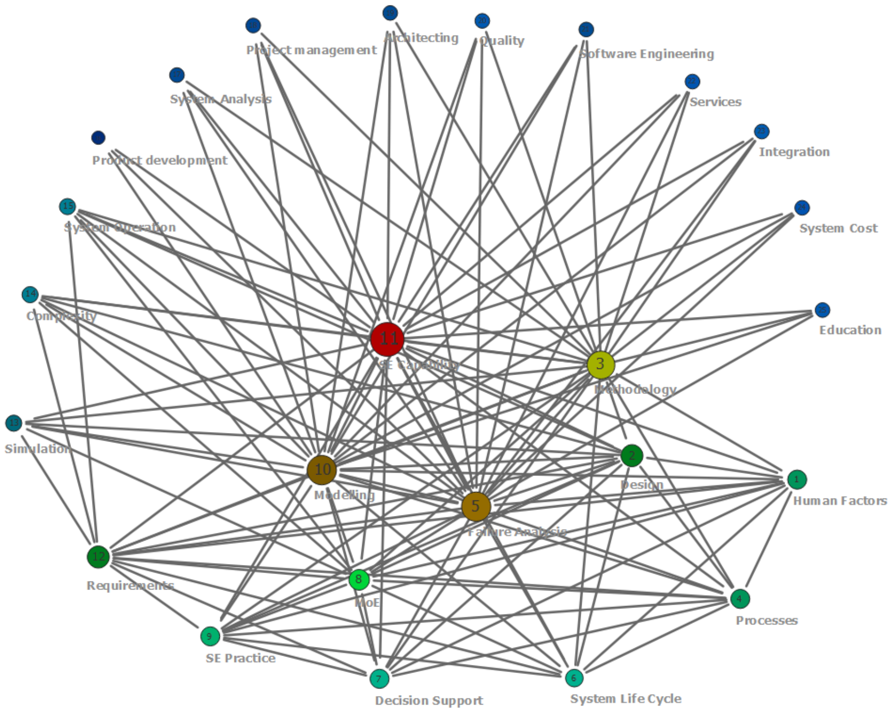

Figure 4: Topic Network Graph with 50% Connection of Maximum Edge

Note that the node numbers between the two graphs have changed, but the nodes' names and relative importance has remained the same. The prominent topics (nodes) include methodology, Failure Analysis, design, processes and human factors. This network also shows more clearly how the secondary and supporting nodes link to the central nodes. These links indicate the underlying building blocks of the research field.

Topic Network Analysis

Analyzing the structure of a research domain depends not only on the interpretation of the placement of nodes in the visual representation but also on examining the network statistics. Network analysis includes several mathematically processed parameters. The centrality value of a node is one of the key parameters of a network. Centrality is determined by the number of connections of a node to the other nodes in the network. Different centrality indices need to be considered in judging centrality, such as (Hevey 2018):

Node strength. The sum of all connections' weighted number and strength indicates a node's strength relative to other nodes. A node with many weak links may not be as important as a node with fewer but stronger relationships.

Degree Centrality (DC). In undirected networks, the Degree Centrality of a node is the sum of the edges' weights attached to that node.

In this paper, the focus is only on the Degree Centrality and the clustering of nodes to analyze the contribution of the various topics in systems engineering to the field's development. Table 2 shows the output on the SocNetV applications' analysis of the network diagrams' different centrality indexes. The analysis showed differences between the degree centrality values for the two versions of the network diagram from Figures 3 and 4. Figure 5 provides a bar chart that compares values for the topic node strength and the degree centrality values from Table 2. The values were adjusted by dividing each set of values per column by the maximum parameter value for plotting on the same axis.

Table 2: Comparison of Centrality Measures between the Total and Reduced Networks for each Node

Node Name	Node Strength	No. in Fig. 3	Degree Centrality Full Diagram	No. in Fig. 4	Degree Centrality 50% Diagram
SE Capability	6167	23	205514	11	205515
Failure Analysis	4520	20	167411	10	167412
Modelling	4517	11	167732	5	167732
Methodology	3935	4	153938	3	149002
Design	3161	2	136591	2	90461
Requirements	3131	25	135570	12	89884
MoE	3019	15	133047	8	88458
Processes	2814	9	129068	4	71368
Human Factors	2740	17	126572	9	70283
SE Practice	2715	1	126681	1	70330
Decision Support	2363	14	118743	7	62038

Node Name	Node Strength	No. in Fig. 3	Degree Centrality Full Diagram	No. in Fig. 4	Degree Centrality 50% Diagram
System Life Cycle	2320	12	116877	6	61226
Complexity	1990	16	109520	14	43329
Simulation	1956	10	109340	13	43274
System Operation	1897	18	108086	15	42892
Integration	1591	21	100950	23	26144
Software Engineering	1510	13	99065	21	25816
Architecting	1465	7	98419	19	25704
Education	1405	24	96768	25	25417
Quality	1405	8	96767	20	25417
System Analysis	1285	5	94332	17	24993
Project management	1227	6	92946	18	24752
Services	1050	19	88940	22	24056
System Cost	904	22	85725	24	23497
Product development	839	3	83867	16	18237

The first point of interest is the degree centrality compares well to the popularity or allocation probability per topic per paper. This pattern reflects the nature of the fully connected mesh network. Another observation from Figure 5 is that degree centrality values from the two graphs (full and reduced) follows the same pattern. Only at the nodes with the least strength values are the difference the greatest. However, it could still be helpful to support the visual interpretation of the analysis results.

Information on the clusters and cliques in the network will also provide valuable information about the topics' structure within the research field. A clique is a subset of a network where the nodes are more closely connected to one another than other network nodes. The nodes have all possible ties present among themselves (Borgatti et al. 2009). SocNetV applies the Bron–Kerbosch algorithm to produce a census of all maximal cliques in the network and reports. The Clique Sensus analysis in SocNetV produced 15 maximal cliques in Table 3 and the dendrogram in Figure 6. The node numbers refer to the Figure 4 numbers in Table 2.

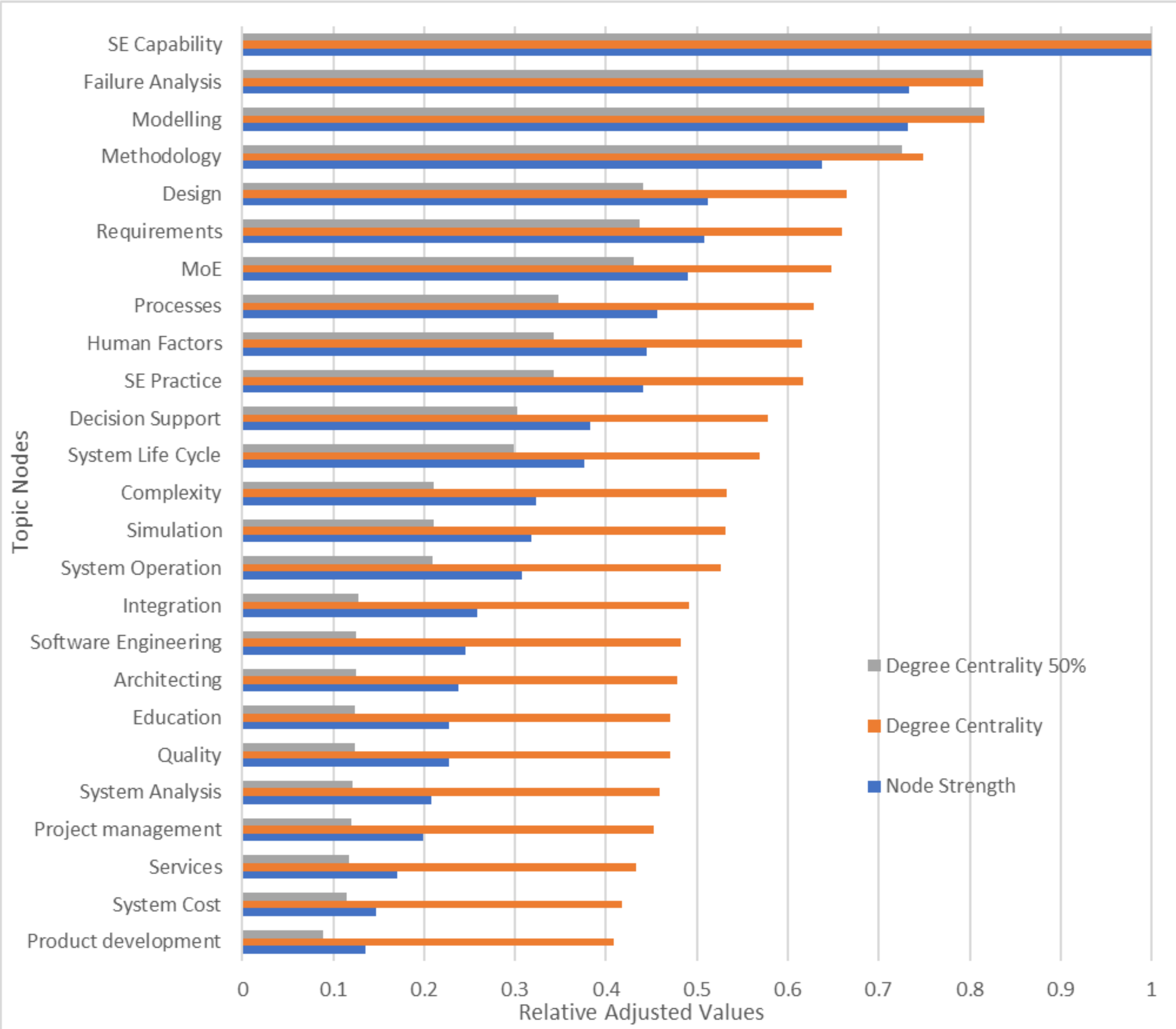

Figure 5: Network Analysis Values

The entire mesh network of Figure 3 presents one large clique, which is not helpful to achieve this paper's objective. Therefore, the reduced network of Figure 4 was also processed to identify cliques within the topic network. Since the comparison of the two graphs indicates a high level of similarity, a clique analysis of Figure 4 would provide the required insight into the field's structure.

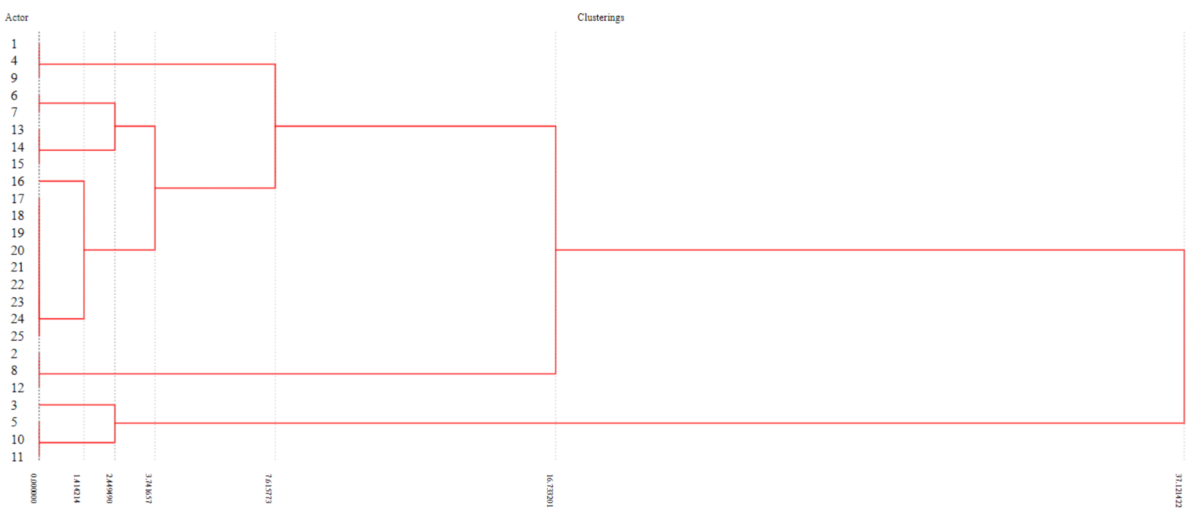

Figure 6: Clustering Dendrogram

Table 3: Cliques in the Reduced Network

Clique Number	Node Numbers	Clique Node Names
1	16 10 11 5	Product development, Modelling, SE Capability, Failure Analysis
2	17 10 11 5 3	System Analysis, Modelling, SE Capability, Failure Analysis, Methodology
3	19 10 11 5 3	Architecting, Modelling, SE Capability, Failure Analysis, Methodology
4	18 10 11 5 3	Project management, Modelling, SE Capability, Failure Analysis, Methodology
5	21 10 11 5 3	Software Engineering, Modelling, SE Capability, Failure Analysis, Methodology
6	20 10 11 5 3	Quality, Modelling, SE Capability, Failure Analysis, Methodology
7	23 10 11 5 3	Integration, Modelling, SE Capability, Failure Analysis, Methodology
8	22 10 11 5 3	Services, Modelling, SE Capability, Failure Analysis, Methodology
9	25 10 11 5 3	Education, Modelling, SE Capability, Failure Analysis, Methodology
10	24 10 11 5 3	System Cost, Modelling, SE Capability, Failure Analysis, Methodology
11	12 13 10 11 8 5 2 3	Requirements, Simulation, Modelling, SE Capability, MoE, Failure Analysis, Design, Methodology
12	15 12 10 11 8 5 2 3	System Operation, Requirements, Modelling, SE Capability, MoE, Failure Analysis, Design, Methodology
13	14 12 10 11 8 5 2 3	Complexity, Requirements, Modelling, SE Capability, MoE, Failure Analysis, Design, Methodology
14	12 10 11 8 9 7 4 5 2 3 1	Requirements, Modelling, SE Capability, MoE, SE Practice, Decision Support, Processes, Failure Analysis, Design, Methodology, Human Factors
15	12 10 11 8 9 6 4 5 2 3 1	Requirements, Modelling, SE Capability, MoE, SE Practice, System Life Cycle, Processes, Failure Analysis, Design, Methodology, Human Factors

The cliques in Table 3 provide some interesting relationships between sets of topic nodes from the reduced network diagram. It seems that the topics nodes with lower strengths are each grouped with a set of most prominent nodes. One way to interpret the table, using the first clique as an example, is that research into Product Development's topic depends on the topics Modelling, SE Capability, and Failure Analysis as a foundation. Also, the focus topics may have an application in the other higher-level topics that may provide requirements or constraints on the research. This goes for the other topics such Quality, Integration, Architecting, Project Management etc.

When planning research into specific topics within systems engineering, the researchers need to understand the research topic's relationships to the other topics within the field, especially those from the same clique.

Conclusion and Future Work

This research performed topic modelling on more than 60 000 papers published about systems engineering from the Scopus database. These papers' topics compare relatively well with previous research that focused on INCOSE's research publications (journal and international symposiums). The new topics from this broader and deeper analysis also added new topics to the existing set. By calculating the co-occurrence of these topics in papers, a topic network diagram was developed. Simplification of the graph also improved visual interpretation of the links between the topics (node) to provide a valuable overview of systems engineering research structure.

An in-depth quantitative and mathematical network analysis identified similarities in the centrality indices to support network graphs' deductions. This paper combined the visual qualitative and quantitative mathematical analysis to map the research topics into a three-tiered framework to understand the relationships better. The take away is that systems engineering researchers need to identify the topic building blocks and their interdependencies due to the multidisciplinary nature of systems engineering. They need to consider transdisciplinary research to ensure that systems engineering can grow and mature as a discipline.

Future steps in this research will include increasing the text sample size (article) from other databases (i.e. Web of Science) and refining the search terms to improve the text data's quality and relevance to process. This will support an in-depth validation of the research topics. The size of the topics list should also be increased for further visual and mathematical network analyses. Also, more profound insight into interpreting the network graphs and statistics are required to improve the value of this methodology's output.

Guidelines for generating the network of nodes and edges need to be established and refined. This includes methods for reducing the original fully connected mesh networks into a validated and verified reduced network, useful for analysis. Lastly, the prominence of the topic Failure Analysis in the output data warrants another look.

References

Agrawal, A, Fu, W & Menzies, T 2018, 'What is wrong with topic modeling? And how to fix it using search-based software engineering, *Information and Software Technology*, vol. 98, no. February, Elsevier, pp. 74–88.

Alnajem, M, Mostafa, MM & ElMelegy, AR 2020, 'Mapping the first decade of circular economy research: a bibliometric network analysis', *Journal of Industrial and Production Engineering*, Taylor & Francis, pp. 1–22.

Antons, D, Kleer, R & Salge, TO 2016, 'Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories', *Journal of Product Innovation Management*, vol. 33, no. 6, pp. 726–749.

Blei, D, Carin, L & Dunson, D 2010, 'Probabilistic Topic Models', *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65.

Borgatti, SP, Mehra, A, Brass, DJ & Labianca, G 2009, 'Network Analysis in the Social Sciences', *Science*, vol. 323, pp. 892–896.

Brill, J 1998, 'Systems engineering? A retrospective view', *Systems Engineering*, vol. 1, no. 4, pp. 258–266.

Chen, H, Wang, X, Pan, S & Xiong, F 2019, 'Identify Topic Relations in Scientific Literature Using Topic Modeling', *IEEE Transactions on Engineering Management*, IEEE, pp. 1–13.

Chen, X & Xie, H 2020, 'A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature', *Cognitive Computation*, Cognitive Computation, pp. 1–33.

Eker, S, Rovenskaya, E, Langan, S & Obersteiner, M 2019, 'Model validation: A bibliometric analysis of the literature', *Environmental Modelling and Software*, vol. 117, no. March, pp. 43–54.

Ferris, TLJ 2009, 'On the Methods of Research for Systems Engineering', *7th Annual Conference on Systems Engineering Research*.

Gretarsson, B, O'Donovan, J, Bostandjiev, S, Höllerer, T, Asuncion, A, Newman, D & Smyth, P 2012, 'Topicnets: Visual analysis of large text corpora with topic modeling', *ACM Transactions*

on Intelligent Systems and Technology, vol. 3, no. 2, pp. 1–26.
Hao, T, Chen, X, Li, G & Yan, J 2018, 'A bibliometric analysis of text mining in medical research', *Soft Computing*, vol. 22, no. 23, Springer Berlin Heidelberg, pp. 7875–7892.
Hevey, D., 2018 Network analysis: a brief overview and tutorial, Health Psychology and Behavioral Medicine, vol 6, no 1, pp. 301-328
Isoaho, K, Gritsenko, D & Mäkelä, E 2019, 'Topic modeling and text analysis for qualitative policy research', *Policy Studies Journal*.
Jia, Y, Wang, W, Liang, J, Liu, L, Chen, Z, Zhang, J, Chen, T & Lei, J 2018, 'Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences', *Computer Methods and Programs in Biomedicine*, vol. 166, Elsevier B.V., pp. 19–32.
Jiang, H, Qiang, M & Lin, P 2016, 'A topic modeling based bibliometric exploration of hydropower research', *Renewable and Sustainable Energy Reviews*, vol. 57, Elsevier, pp. 226–237.
Jussila, JJ, Mustafee, N, Aramo-Immonen, H, Menon, K, Hajikhani, A, Helander, N & ... 2017, *A Bibliometric Study on Authorship Trends and Research Themes in Knowledge Management Literature, 2th International Forum on Knowledge Asset Dynamics - Knowledge Management in the 21st Century: Resilience, Creativity and Co-creation*.
Kalantari, A, Kamsin, A, Kamaruddin, HS, Ale Ebrahim, N, Gani, A, Ebrahimi, A & Shamshirband, S 2017, 'A bibliometric approach to tracking big data research trends', *Journal of Big Data*, vol. 4, no. 1, Springer International Publishing, pp. 1–18.
Keathley, H, Bean, A, Chen, T, Vila, K, Ye, K & Gonzalez-Aleu, F 2015, 'Bibliometric analysis of author collaboration in engineering management research', *International Annual Conference of the American Society for Engineering Management 2015, ASEM 2015*, pp. 679–689.
Kim, J & Kang, P 2018, 'Analyzing international collaboration and identifying core topics for the "internet of things" based on network analysis and topic modeling', *International Journal of Industrial Engineering : Theory Applications and Practice*, vol. 25, no. 3, pp. 349–369.
Kuhn, TS 1962, *The Structure of Scientific Revolutions*, The University of Chicago Press.
Kunc, M. and Mortenson, M. and Vidgen, R 2012, 'A computational literature review of the field of System Dynamics from 1974 to 2017', *Journal of Simulation*, vol. 12, no. 2, pp. 115–127.
Lee, H & Kang, P 2018, 'Identifying core topics in technology and innovation management studies: a topic model approach', *Journal of Technology Transfer*, vol. 43, no. 5, Springer US, pp. 1291–1317.
Lin, JR, Hu, ZZ, Zhang, JP & Yu, FQ 2016, 'A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM', *Computer-Aided Civil and Infrastructure Engineering*, vol. 31, no. 1, pp. 18–33.
Oosthuizen, R & Pretorius, L 2020, 'Analysis of INCOSE Systems Engineering Journal and international symposium research topic trends', *Systems Engineering*.
Oosthuizen, R. & Pretorius, L 2020, 'Bibliometric analysis of technology management research topic trends', *Towards the Digital World and Industry X.0 - Proceedings of the 29th International Conference of the International Association for Management of Technology, IAMOT 2020*.
Rose, ME & Kitchin, JR 2019, 'pybliometrics: Scriptable bibliometrics using a Python interface to Scopus', *SoftwareX*, vol. 10, pp. 1–6.
Sage, AP 1998, 'Systems engineering: Purpose, function, and structure', *Systems Engineering*, vol. 1, no. 1, pp. 1–3.
Salminen, J & Wirtz, J 2021, 'Artificial Intelligence in Marketing : Topic Modeling, Scientometric Analysis , and Research Agenda', *Journal of Business Research*, vol. Forthcoming.
Shin, SH, Kwon, OK, Ruan, X, Chhetri, P, Lee, PTW & Shahparvari, S 2018, 'Analyzing sustainability literature in maritime studies with text mining', *Sustainability (Switzerland)*, vol. 10, no. 10, pp. 1–19.
Smiraglia, R.P., 2015. Domain analysis of domain analysis for knowledge organization: Observations on an emergent methodological cluster. Knowledge Organization, vol. 42, no. 8,

pp. 602-614.
Sohn, E, Noh, KR, Lee, B & Kwon, OJ 2018, 'Bibliometric network analysis and visualization of research and development trends in precision medicine', *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, IEEE, pp. 727–730.
Suominen, Arho;Toivanen, H 2016, 'Map of science with topic modelling: Comparison of unsupervised learning and human-assigned subject classification', *Journal of the Association for Information Science and Technology*, vol. 67, no. 10, pp. 2464–2476.
Valerdi, R & Davidz, HL 2009, 'Empirical research in systems engineering: Challenges and opportunities of a new frontier', *Systems Engineering*, vol. 12, no. 2, pp. 169–181.
Yeo, JS & Jeong, Y 2020, 'Pathway toward market entry of perovskite solar cells: A detailed study on the research trends and collaboration networks through bibliometrics', *Energy Reports*, vol. 6, Elsevier Ltd, pp. 2075–2085.
Zhang, Y, Chen, H, Lu, J & Zhang, G 2017, 'Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016', *Knowledge-Based Systems*, vol. 133, pp. 255–268.

Biography

Dr Rudolph Oosthuizen. Rudolph Oosthuizen joined South African Air Force in 1990 to obtain a B.Eng (Elec 1994), B.Eng (Hons) (Indus 1998) and MEM (2002). He performed systems engineering roles in electronic warfare and Command and Control. In 2008 he joined the CSIR and is a Senior System Engineer with more than 10 years experience in conceptual system modelling. He obtained a PhD in Engineering Management at UP in 2015. Of special interest is modelling the human element in sociotechnical systems and Artificial Intelligence for decision support and situation awareness. Since May 2020 he has been appointed as a senior lecturer at the GSTM. Rudolph is registered as a Professional Engineer with ECSA and a CSEP.