KEYWORDS › Clustering › Cluster Configuration › Solution Space Exploration

CLUSTERING TECHNIQUE FOR DSMs

Florian G. H. Behncke[1]
behncke@pe.mw.tum.de

Doris Maurer[1]
doris.maurer@tum.de

Lukas Schrenk[1]
lukas.schrenk@gmx.net

Danilo M. Schmidt[1]
schmidt@pe.mw.tum.de

Udo Lindemann[1]
lindemann@pe.mw.tum.de

[1]Institute of Product Development, Technische Universität München, Germany

ABSTRACT

This paper provides a clustering technique (CT) for Design-Structure-Matrices (DSMs) that explore the entire solution space of cluster configurations (CCs) for a given system. Therefore the paper gives an overview of established CTs for exclusive clusters as basis for the development of an alternative CT that generates all possible CCs. These configurations are assessed against a set of performance metrics, which are selected by the decision makers to determine the quality of the cluster. Through a representation of the CCs in a portfolio according to the values of the performance metrics, decision makers are provided with a ranking of CCs as s support for their decision. Thereby, it is observed that established CTs do not capture the entire solution space of CCs and therefore miss comparable configurations. As a result, decision makers are not necessarily equipped with ideal CCs by established CTs.

1. Introduction

Solving complex problems is common practice for system engineers, which counter this challenge by decomposing and integrating the object of investigation – complex systems *(Pimmler & Eppinger, 1994)*. Products, processes or organizations are vital examples of these systems in industry *(Browning, 2001)*. The decomposition or integration of technical products for instance applies at a system-to-component-level or a product-to-system-level *(Pimmler & Eppinger, 1994)*, depending on the level of concretion required by the system engineer. The DSM supports the process of decomposition and integration as established tool for representing and analyzing complex systems *(Browning, 2001)*. Clustering and sequencing are essential techniques that are applied for the analysis of certain types of DSMs, which can be divided into static- and time-based representatives *(Browning, 2001)*. Thereby, clustering is "finding subsets of DSM elements *(i.e. clusters or modules – sometimes also termed chunks)* that are mutually exclusive or minimally interacting" *(Sharman & Yassine, 2004)* and therefore focus on static DSMs. Literature provides numerous techniques for clustering that optimize an initial DSM by reordering the rows and columns to achieve a blocked matrix *(Pimmler & Eppinger, 1994)*. This paper focuses on the development of a CT for DSMs, which considers the challenges that the authors of this paper observed, while performing clustering in both industrial case studies at manufacturing firms as well as in case studies for educational purposes. The observations on established CTs are described in detail in the following and therefore draw the research agenda for the paper at hand. Commercial and research tools for DSMs featuring CTs mostly present a singular CC to decision makers for a given system represented in a DSM. This specific configuration is considered ideal according the objective function of the underlying optimization algorithm of the CT. However, a subsequent manual clustering oftentimes reveals comparable or even better CCs for a given system, which can have two reasons. One reason is that certain CTs miss configurations within the solution space due to the path dependency of algorithms reported by Sharman & Yassine *(2004)* or the fact that these techniques do not scan the entire solution space for CCs. **Figure 1** illustrates the path dependency that is obvious for perfectly equal clusters, "where it is purely a matter of chance how any clustering algorithm would present an answer" to this problem *(Sharman & Yassine, 2004)*.

Another reason for the identification of comparable or even better CCs for a given system is that decision makers consider further implications of the system *(i.e. constraints)* for the manual clustering. These implications influence the decision on clusters significantly through a reduction of the solution space without being explicitly mentioned in the given DSM (**Figure 1**). As a result, decision makers favor configurations that may not ideally fulfil the objective function of the CT, but better achieve the overall objectives of the system. Doubtless, this is rather no shortcoming of CT than a matter of modeling a system for a certain purpose. However, this is a major challenge of decision makers that may not have the overview of multitude of effects on the system as well as the ability to capture all effects in a DSM. The latter refers to optimize multiple domains, which focuses on a global optimum rather than local optima *(Yassine, 2010)*. Another observation especially in industrial case studies is that clustering is more about finding the best compromise without having an overlook of all effects on the system. Providing decision makers with a singular CC that best fulfils a generic objective function of a CT is not supporting the required decision support. Making more CCs available to the decision makers ensures a more comprehensive consideration of the solution space and equip decision makers with more transparency on alternative CCs that better achieve the overall objective of the system. Thereby, implications on the system can be considered even though they are not explicitly mentioned or modelled in the given DSM. As a result, the objective of the paper is to develop a CT which explores the entire solution space for CCs. Furthermore, this paper aspires a CT that provides multiple CCs with comparable cluster quality.

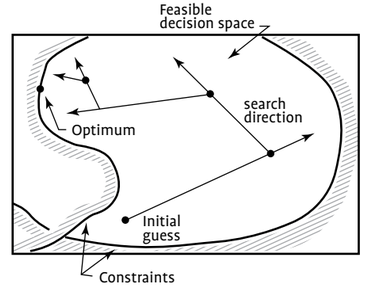


FIGURE 1. *Solution space for CCs (Deb, 2001)*

2. Research Methodology

This paper presents a CT that explores the entire solution space of CCs. Therefore, this paper provides an initial literature review on CTs and an overview on performance metrics for applied objective functions of CTs *(section 3)*. The review covers conference publications of the DSM conference and international journals that are related to clustering in DSMs. The inclusion of approaches for the review in section 3 depends on the presence of the appliance of a certain CT in a DSM. With a content analysis of the derived publications, section 3 provides a classification of CTs, which are the basis for the development of the CT *(4)* as key contribution of the paper at hand. In order to evaluate the internal validity and operability of the CT, section 5 provides an academic case study. The case study features a system with five elements to accent the comprehension of the solution space exploration of CCs *(Appendix)* in favor for a system that is as close as possible to an industrial appliance.

3. Clustering Techniques and Performance Metrics for Clusters

This paragraph summarizes relevant CTs and performance metrics for the assessment of the quality of clusters. There are two major classes of CTs according to the affiliation of objects to several clusters *(non-exclusive)* or one singular cluster *(exclusive)*. The latter is focused by the paper at hand. This class is divided into two groups *(extrinsic and intrinsic)* that differ in terms of labeling objects before the clustering. Within the intrinsic CTs, there are two representatives; hierarchical that generate a sequence of nested partitions and partitional, which generate single partitions *(Jain & Dubes, 1988)*. As CTs for DSMs do not require for a labeling of objects before the clustering, intrinsic techniques including hierarchical and partitional ones are considered by the paper at hand. Moreover, the techniques are limited to the ones using raw data *(Everitt, Landau, Leese, & Stahl, 2011)*. This leaves the Centroid-, Median-, and Ward's method as hierarchical as well as k-means as partitional technique *(Jain & Dubes, 1988) (Lance & Williams, 1967)*. The procedure of these methods are described in the following:

Hierarchical techniques first compute the proximity matrix containing the distance between each pair of patterns, where each pattern is treated as a cluster *(1)*. The second step *(2)* is to find the most similar pair of clusters using the proximity matrix. These are merged from two clusters into one cluster. This is performed as long as all clusters are in one cluster *(3)*. Thereby, the different techniques; Centroid-, Median- and Ward's method differ in terms of the measure for the distance *(2)*. The Centroid distance calculates the squared euclidian between mean vectors *(centroids)*, while the Median distance considers weighted centroids *(Everitt et al., 2011)*. Ward's method defines the distance as "Increase in sum of squares within clusters, after fusion, summed over all variables" *(Everitt et al., 2011)*.

The procedure of partitional techniques is described according to Jain et al. *(1999)* as: "*(1)* Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set. *(2)* Assign each pattern to the closest cluster center. *(3)* Recompute the cluster centers using the current cluster memberships. *(4)* If a convergence criterion is not met, go to *(2)*. Typical convergence criteria are: no *(or minimal)* reassignment of patterns to new cluster centers, or minimal decrease in squared error".

All presented techniques have in common that they generate a cluster centroid, to which objects are allocated. The resulting clusters depend on the given DSM, which means the results differ depending on the format of the given DSM. As a result, the techniques use different objective functions to assess the quality of the clusters. **Table 1** gives an overview of metrics that are provided by literature. Furthermore, it provides a description on the objective of the different performance metrics. The list does not claim to be complete, however it indicates the diversity of metrics.

No.	Objective of performance metric	Reference
M.1	Minimize the proportion of outer relations of the cluster; maximize the proportion of inner relations of the cluster	Adopted from (Lindemann, Maurer, & Braun, 2009)
M.2	Maximize the proportion of relations used in the cluster	Adopted from (Newman, 2003) and (Kreimeyer, 2009)
M.3	Minimize the squared euclidian distance between mean vectors (Centroid)	(Everitt et al., 2011)
M.4	Minimize the squared euclidian distance between weighted centroids (Median)	(Everitt et al., 2011).
M.5	Increase the sum of squares within clusters, after fusion, summed over all variables (Ward)	(Everitt et al., 2011)
M.6	Minimize the sum of discrepancies between a point and its centroid	(Berkhin, 2006)
M.7	Minimize the sum of internal and external relations of clusters	(Yassine, 2010)

Table 1. Performance metrics for the assessment of clusters

4. Clustering Technique for Exclusive Clusters

The CT consists of a sequence of four steps that cover the generation of the solution space of CCs *(CT.1)*, the acquisition of information on relations between elements of the given system *(CT.2)*, the selection of performance metrics for the CCs *(CT.3)*, and an export of a set of ideal CCs *(CT.4)*. Thereby, the technique focuses on exclusive clusters *(non-overlapping)*, while aspiring an extension for non-exclusive clusters in future work. The first step *(CT.1)* of the CT focuses on generating the entire solution space for CCs. Thereby, the CCs of a given system are calculated analytically by the Stirling number of the second kind *(S)*, where n is the number of elements within the system and k is the size of the cluster *(number of elements in the cluster) (Bronstein, Semendjajew, Musiol, & Mühlig, 1999)*. This gives the number of CCs *(S)* for a specific size of the cluster *(k)*:

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\cdot\binom{k}{i}\cdot i^{n}$$

In order to achieve the total number of CCs *(M)* – representing the entire solution space – the Stirling number of the second kind *(S)* is calculated for each size of clusters *(M)*, according to the following formula:

$$M(n,k) = \sum_{i=0}^{k} n - i$$

The deduction of the solution space of CCs is performed by a numerical calculation in MATLAB. This step *(CT.1)* is performed without information on the given system and just requires the number of elements of the system as input parameter. As a result, this step can be considered as pre-processing of the CT.

The next step *(CT.2)* focuses on the acquisition of information on the relations between elements of the given system. Therefore, the relations are recorded in template for a DSM that is supported by a form in the conducted MATLAB program.

Based on the numerous objective functions provided by literature, the third step *(CT.3)* provides decision makers with an overview of performance metrics for the assessment of the quality of the clusters. **Table 1** summarizes the metrics that are offered to decision makers through a form in the MATLAB program.

The last step *(CT.4)* of the CT merges the information provided by the previous steps, so that the information on the given system *(CT.2)* is entered to the generated CCs *(CT.1)*. These configurations are then assessed against the selected performance metrics *(CT.3)*. Based on this results, the MATLAB program displays the CCs in a portfolio using their values for the selected performance metrics. Furthermore, MATLAB provides the blocked DSMs of the top CCs to decision makers.

5. Academic Case Study

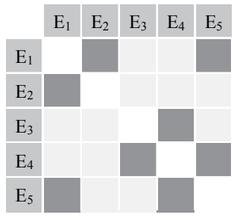


FIGURE 2. *DSM and flow chart*

The CT presented in section 4 is applied to an academic case study to prove the internal validity and operability. The case study features a system with five elements that are interrelated according to the given system (**Figure 2**). Thereby, elements are linked bi-directionally. E2 influences E1, E1 influences E5, E5 influences E4 and E4 influences E3 and vice versa. As a result, **Figure 2** illustrates the outcome of step CT.2.

Based on the number of elements MATALB is generating the entire solution space for CCs using the presented formulas *(section 4)* with n = 5 number of elements within the system (**Figure 2**). **Table 2** illustrates the total number of CCs for the different sizes of clusters k. The appendix gives an overview of the 52 CCs for the given system *(CT.1)*.

	k = 0	k = 1	k = 2	k = 3	k = 4	**k = 5**	∑
n = 5	0	1	15	25	10	1	52

TABLE 2. *Number of CCs*

The 52 CCs are evaluated against a set of performance metrics for clusters (**Table 1**) that are selected by the decision maker using the corresponding template in MATLAB *(CT.3)*. For the case study two different performance metrics are selected. The metric M.1 focuses on the assessment of the quality of clusters through the relations within the cluster as well as the relations of the cluster to other elements of the system. The other metric M.2 derives the number of relations of the system, which are captured by the cluster as an indicator of the quality of the clusters from a system perspective. Based on this results, the last step *(CT.4)* features the illustration of the CCs according to their values for the performance metrics M.1 and M.2 in a portfolio (**Figure 3**). Thereby, CCs with a high cluster quality are located in the bottom-left area close the origin of the co-ordinates. Configurations with a low cluster quality are positioned in the top-right area of the portfolio. CCs on the parabolic line have the same cluster quality according to the two performance metrics.

As some CCs have the same quality according to the performance metrics M.1 and M.2, **Figure 3** just illustrate one representative for each configuration *(diamond-shaped data points)*. For instance CC 2.1 is a representative for CC 2.5 and CC 2.7. Furthermore, triangle-shaped data points in-

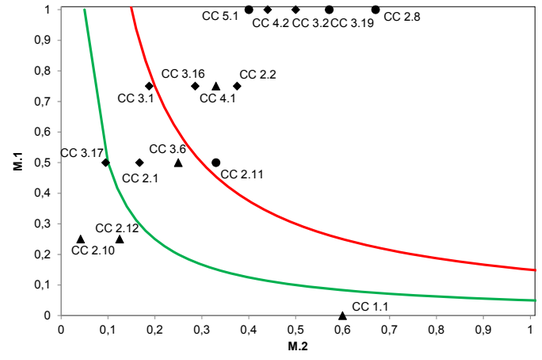


FIGURE 3. *Quality assessment of the solution space for CCs*

dicate that these CCs are derived by the other CTs presented in section 3. For example CC 2.3 is identified by the k-means method and is represented by CC 2.10 in the portfolio. **Table 3** gives an overview of the representative CCs that are not explicitly illustrated in **Figure 3**.

According to the values of the CCs for the performance metrics M.1 and M.2 the established CTs are able to find sound configurations for different sizes of clusters. Hence, they are not able to find all CCs which provide good values for the performance metrics *(e.g. CC 3.17 or 2.1)*. As a result, the exploration of the entire solution space reveal further CCs that are not captured by established techniques. This provides decision makers with more alternatives including a configuration that may better achieve the objectives of the system. Thereby, decision makers may consider implications on the system even though they are not explicitly mentioned in the given DSM.

Besides the mentioned advantages of the conducted CT, there are some limitations. This paper focuses on exclusive clusters, which is not covering overlapping CCs. Those need to be considered in a subsequent version of the MATLAB program, as overlapping clusters better capture the challenges of industrial appliances. With an extension to non-exclusive clusters the already high number of configurations increases significantly, which requires for corresponding techniques to pre-assess the resulting CCs to limit the solution space to promising configurations. Furthermore, the assessment of CCs strongly depends on the selected performance metrics. Addressing this limitation is partly considered by the MATLAB program through giving decision makers a choice upon the performance metrics. Moreover, the MATLAB program paves the way for a multi-criteria evaluation of CCs as the relevant performance metrics are already in place. However, there is still significant room to improve the support of decision makers at finding ideal performance metrics for their specific applications.

CC 2.1	CC 2.5 and CC 2.7
CC 2.2	CC 2.4, CC 2.6, and CC 2.9
CC 2.10	CC 2.3
CC 2.11	CC 2.14 and CC 2.15
CC 2.12	CC 2.13
CC 3.1	CC 3.3, CC 3.4, CC 3.9, CC 3.10, and CC 3.12
CC 3.2	CC 3.5, CC 3.8, CC 3.11, CC 3.14, and CC 3.15
CC 3.6	CC 3.7 and CC 3.13
CC 3.16	CC 3.18; CC 3.20; CC 3.22; CC 3.23; CC 3.25
CC 3.17	CC 3.17, CC 3.21, and CC 3.24
CC 4.1	CC 4.4, CC 4.8, and CC 4.10
CC 4.2	CC 4.3, CC 4.5, CC 4.6, CC 4.7, and CC 4.9
CC 2.1	CC 2.5 and CC 2.7

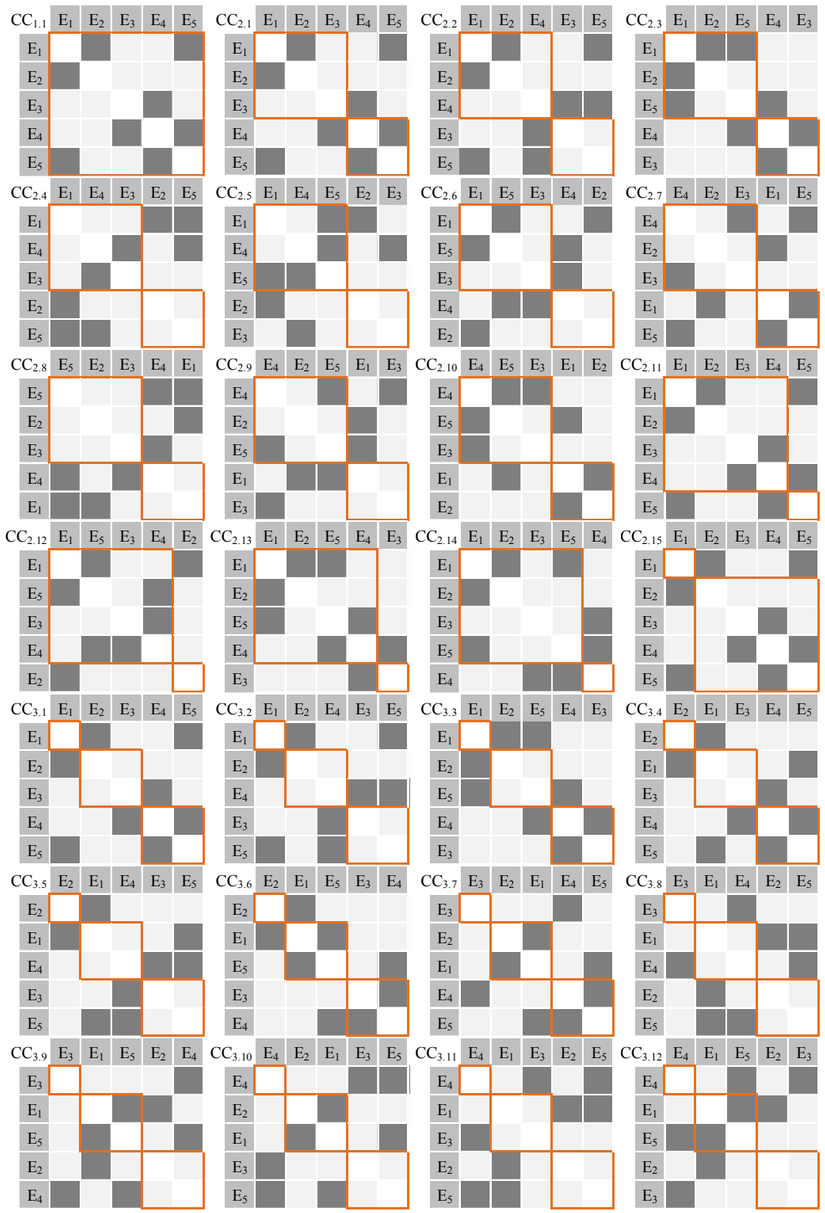
TABLE 3. *Representative CCs*

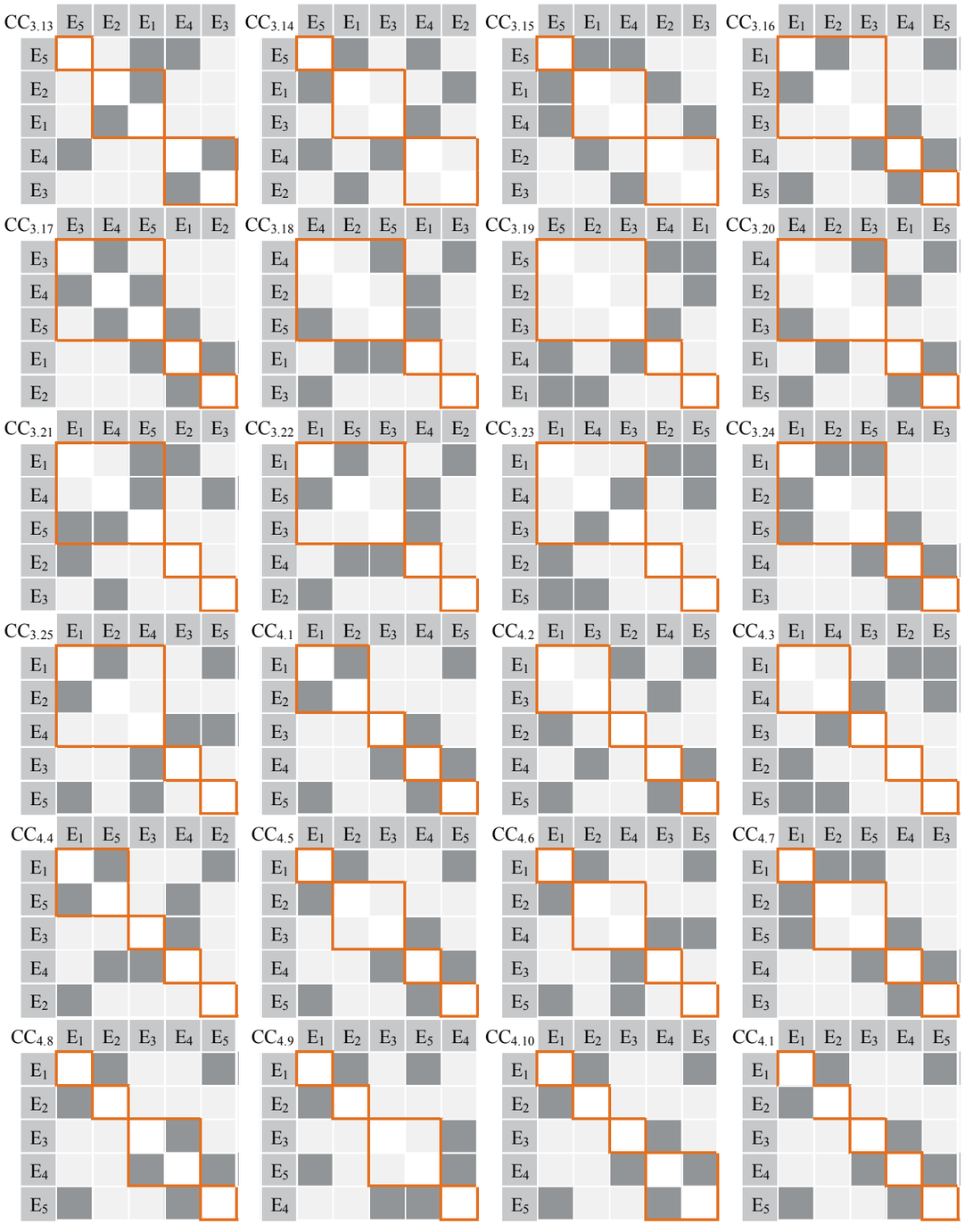
6. Summary and Outlook

This paper presents a CT that searches the entire solution space of CCs as established techniques miss comparable good configurations. This challenges decision makers that may not be able to capture all implications of the system in a DSM. The conducted CT provides several CCs to decision makers. This allows the consideration of implications through the selection of a certain configuration. The implementation of the CT in MATLAB allows the selection of different performance metrics for the assessment of the cluster quality. A first step of future work is to extend the CT to non-exclusive clusters. As the solution space for non-exclusive clusters provides even more CCs a further step is the development of a pre-assessment of clusters in terms of their quality using performance metrics *(table 1)*. This procedure of iterative assessment allows to limit the solution space to promising CCs and therefore reduces the required capacity for the assessment and export of ideal CCs. The pre-assessment combined with the advancement of the MATLAB program allows the validation of the CT at an industrial application with a larger systems. Furthermore, this advancement of the MATLAB program facilitates an empirical analysis on the correlation of performance metrics to case study specific problems to identify whether certain performance metrics better predict the performance of a specific problem in an industrial application than others.

APPENDIX

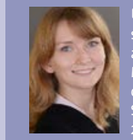
The following matrices (DSMs) illustrates the entire solution space of CCs for the system of an academic case study (section 5). The index in the top-left cell of the DSMs defines the CC through the size of the cluster (first number of the index) and the sequential number of CCs with a specific cluster size (second number of the index).





authors

F. G. H. Behncke is a scientific assistant at the Institute of Product Development (Faculty of Mechanical Engineering) at the Technische Universität München, Germany (www.pe.mw.tum.de). He graduated in mechanical engineering in 2010 with a focus on vehicle and production technology. His research is centers on Systems Engineering, Complexity Management, Product Design as well as Supply Chain Management. He dedicates his research career to the synchronization between the product and supply chain design, which manifests in his PhD topic on "Design for Procurement – Matching between product architecture and supply network design in early phases of product development".



D. Maurer studied Aerospace between 2008 and 2012 at Technische Universität München and finished her degree with Bachelor of Science. Between 2012 and 2014 she continued with the master's programme Mechanical Engineering and majored in „Flow and Flight Physics" and in „Production Management". During her master's programme she started working on complexity management and cluster optimization. On the basis of two examples she develops two algorithms which support the thesis that information exchange at a minimum number of critical intersections leads to more robust processes. In the following she worked as scientific assistant at the institute of Product Development dealing with the topic of clustering techniques. She graduated at Technische Universität München with her master's degree in June 2014. Today she is working as a development engineer in Munich.

L. E.-M. Schrenk is an Aerospace Technology Student at the Technical University of Munich, currently working on his joined Master Thesis between the Technische Universität München, Germany and the Massachusetts Institute of Technology (MIT). He is a Student Research Assistant at the Institute of Product Development and works on Complexity Management and Structural Supply Chain Management. The topic of his Bachelor Thesis was the computational configuration of supply chain networks based on a product's architecture. Thereby a predecessor of the clustering method presented in the paper at hand has been applied to determine and analyze all possible product architecture and supply chain configurations.

D. M. Schmidt was born in Berlin (Germany) at July 1st 1988. He did his diploma degree in mechanical engineering at Technische Universität München and did his diploma thesis at Tel Aviv University about "Development of an Optimization Algorithm for Adaptable Product Architecture Design". He has graduated in 2012 and since then, he is a research assistant and PhD-student at the institute of Product Development at the Technische Universität München. His PhD's topic is "Increasing Customer Acceptance in PSS-Planning" and he is supervised by Prof. Udo Lindemann. The main research focus are Product-Service Systems, Customer acceptance, decision-making in product planning and decision processes. Other research deals with knowledge management, complexity management and analysis of industrial clusters.

U. Lindemann succeeded Professor Ehrlenspiel in 1995 as head of the Institute of Product Development at the Technical University of Munich. Within the time since 1995 until today he served as Dean for Study Affairs and as Dean of the Faculty Mechanical Engineering. Today he is a member of the Academic Senate of the Technical University Munich. He is co-publishers of the German journal „Konstruktion" and co-editor of several international journals. Since the initiation of the Design Society, he has been an active member, from 2007 to 2010 he served as its President. In addition he is an active member of a number of scientific societies and other organisations. 2008 he became a member of the German Academy of Science and Engineering.

references

Berkhin, P. *(2006).* A Survey of Clustering Data Mining Techniques. In Grouping multidimensional data *(pp. 25–71).* Berlin: Springer.

Bronstein, I. N., Semendjajew, K. A., Musiol, G., & Mühlig, H. *(1999).* Taschenbuch der Mathematik. Frankfurt am Main: Harri Deutsch.

Browning, T. R. *(2001).* Applying the Design Structure Matrix to System Decomposition and Integration Problems: A Review and New Directions. IEEE Transactions on Engineering Management, 48*(3),* 292–306. doi:10.1109/17.946528

Deb, K. *(2001).* Multi-objective optimization using evolutionary algorithms. New York: John Wiley & Sons.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. *(2011).* Cluster Analysis. Chichester: Wiley.

Jain, A. K., & Dubes, R. C. *(1988).* Algorithms for Clustering Data. New Jersey: Prentice Hall.

Kreimeyer, M. *(2009).* A Structural Measurement System for Engineering Design Processes. Lehrstuhl Für Produktentwicklung. München: Technische Universität München.

Lance, G. N., & Williams, W. T. *(1967).* A general theory of classificatory sorting strategies - I. Hierarchical system. Computer Journal, 9, 373–380.

Lindemann, U., Maurer, M., & Braun, T. *(2009).* Structural Complexity Management - An Approach for the Field of Product Design. Berlin: Springer.

Newman, M. E. J. *(2003).* The structure and function of complex networks. SIAM Review, 45*(2),* 167–256.

Pimmler, T. U., & Eppinger, S. D. *(1994).* Integration analysis of product decompositions. In American Society of Mechanical Engineers, Design Engineering Division *(Publication)* DE *(Vol. 68, pp. 343–351).*

Sharman, D. M., & Yassine, A. A. *(2004).* Characterizing Complex Product Architectures. Systems Engineering, 7*(1),* 35–60. doi:10.1002/sys.10056

Yassine, A. A. *(2010).* Multi-domain DSM: Simultaneou optimization of product, process & people DSMs. In 12th International Dependency and Structure Modelling Conference *(pp. 319–332).* Cambridge, UK.