Classification of Systems from Component Characteristics

A. S. Shaja
Doctoral Student, Department of Aerospace Engineering,
IIT Bombay, Mumbai, India - 400076,
Email: shaja@aero.iitb.ac.in

K. Sudhakar
Professor, Department of Aerospace Engineering, IIT
Bombay, Mumbai, India - 400076,
Email: sudhakar@aero.iitb.ac.in

Abstract

In recent times there is a growing interest in systems architectures in domains like biological systems and social networks leading to useful insights and generalizations. These developments have opened up possibility of investigating architectures of complex engineering systems on similar lines. Architecture of a system can be abstracted as a graph, wherein the nodes/vertices correspond to components and edges correspond to interconnections between them. Graphs representing system architecture have revealed motifs or patterns. Motifs are recurring patterns of 3-noded (or 4, 5 etc.) sub-graphs of the graph. Over-represented motifs have offered insights into the basic functionality of systems in some cases. Concept of motif significance profiles (i.e., proportions of various motifs present in a system) has also given rise to interesting insights. These profiles are seen to be highly correlated across systems of the same family (i.e., very similar proportions of motifs are present in systems of same type). Recently these profiles are proposed as classifiers for system architecture. We now show that the same classification of systems can be arrived at by merely looking at characteristics of components/nodes from which systems are synthesized. In other words, we argue that the motif significance profile of a system is due to the properties of the individual components that form the system. We have shown this by considering a vast variety of systems (38 systems arbitrarily chosen) ranging from – biological systems, languages, electronic circuits, software systems and mechanical engineering systems.

Keywords

Systems architecture, Motifs, Over-represented patterns, Pattern signatures, Component characteristics, System architecture families.

1. Introduction

1.1. Complex Systems Architectures

Recent years have witnessed a growing interest in the study of complex systems architectures, in domains like biological systems and social networks [Duncan J Watts]. Unifying principles have emerged [Boccaletti S et al]. Literature has commented on the hesitation of researchers in complex engineering systems, to look at their problems, in the light of emerging ideas in complex systems in general. "Engineering should be at the centre of these developments, and contribute to the development of new theory and tools" [Ottino J M]; "Engineers seem a little bit indifferent as if engineering is at the edge of the science of complexity" [Zhi-Qiang Jiang et al].

Architecture is the fundamental structure of components of a system - the roles they play, and how they are related to each other and to their environment [ANSI IEEE Standard 1471]. The dictionary definition of complexity refers to interconnected/interwoven components. Complexity of a system scales with the number of components, number of interactions, complexities of the components and complexities of interactions [Edward Crawley et al]. Complex engineering systems are synthesized from a large number of components coupled to each other, giving them a physical architecture. They evolve through a design process that is best represented by a large number of connected tasks, giving them a technical architecture. Also, they are created by collaborating groups of people giving them an organizational architecture. These architectural views pose interesting possibilities in respect of searching new understanding in complex engineering systems [Tyson R Browning]. Architecture of a system (say for example from engineering, biology, sociology) can be abstracted as a network/graph, where the nodes/vertices correspond to components in the system and edges correspond to interconnection between them.

1.2. Background

Previous literature on complex systems research in biology defines motifs *(also referred to as over-represented patterns)* as recurring sub-graphs of the graph. In biology, the analysis of network motifs has led to interesting insights in the areas of protein-protein interaction prediction [Albert L et al] and analysis of temporal gene expression patterns [Ronen M et al]. For instance, incase of sensory transcription networks of biology the discovered motifs have been theoretically and experimentally shown to perform signal-processing tasks. It is also argued in the literature that motifs can represent the simple building blocks of complex networks whose selection may possibly be one way to understand the basic functionality of a system. However, in other domains how important these ideas related to motifs are for supporting the global architecture is an area of recent research. Very similar ideas related to over-represented patterns/motifs namely chains [Villas Boas et al], pairwise disconnectivity index [Goemann B et al] have also been proposed in literature. Milo et al [Milo R et al] have proposed an approach to study similarity in the structure of networks, based on the significance profile (SP) of motifs in the network. They have discovered super families based on 3 node motifs from networks from biology, languages etc. Shaja and Sudhakar (authors) have recently discovered super families based on 3 node motifs for engineering systems ranging from softwares, mechanical systems and electrical circuits [Shaja A S, Sudhakar K] Recently motif signatures have also been proposed as a classifier for networks [Ahnert S.E., Fink T.M.A].

1.3. New Insights based on Component Characteristics

In this manuscript, we show that almost the same super families' classification which was previously achieved by [Milo R et al, Shaja A S, Sudhakar K] can be achieved by merely looking at the individual components/nodes characteristics. In other words, we argue that the classification of the system architectures across families using motif significance profile is due to the properties of the individual components that form the system. We have shown this by considering a vast variety of systems (38 systems arbitrarily chosen) ranging from – biology systems, languages, electronic circuits, software systems and mechanical engineering systems.

2. Theoretical Background

2.1. Motifs

"Motifs are recurring sub-graphs of interactions from which the networks are built" [Milo R et al]. Some patterns of interconnections occur in real networks in numbers that are considered significant. Motifs can be of any size from n=2 to N-1, where N is the total number of nodes in the network. Let us consider a directed network with N nodes and look for motifs of size n=3. There are $^NC_3$ different combinations of triplets of nodes in an N-noded network. Some triplets out of $^NC_3$ need not form a connected graph, and are not sub-graphs (an example is when out of 3 nodes 2 nodes are connected to each other and the third does not have an edge with the first two). A connected triplet is a 3-noded sub-graph. For a 3-noded sub-graph there are 13 patterns possible as shown in Fig 3.1.

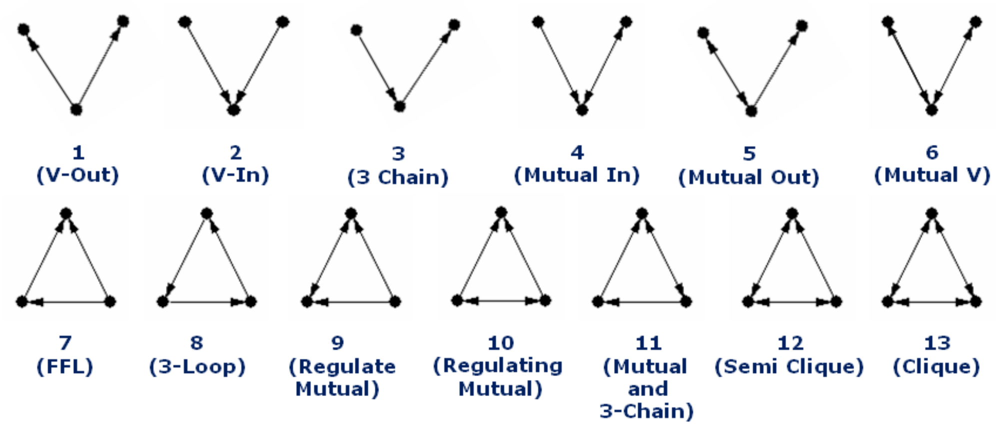


Fig 3.1 All 13 patterns for connected triplets

Each of the $^{N}C_3$ triplets, if it is a sub-graph, will assume one of the 13 patterns. One can count the occurrence of each pattern for all $^{N}C_3$ triplets and define a vector, $\mathbf{P}_{real}$, of size 13. In a network, the count for a particular pattern may be high, which by itself is not considered important. It is possible that such high count for that pattern is unavoidable for a network synthesized using the N nodes that preserve the degree distribution of the real network. To investigate this, randomized networks are created [Milo R et al] using same N nodes, i.e., the number of nodes and their degree distribution is preserved. Each randomized network defines a pattern count vector, $\mathbf{P}_{rand-i}$. A large number of randomized networks (i=1 to m) will define a vector of means, $\boldsymbol{\mu}_{rand}$ and a vector of standard deviations, $\boldsymbol{\sigma}_{rand}$, of 13 patterns. For the real network we can check the significance of j^{th} pattern by,

$$S_j = (P_{real-j} - \mu_{rand-j})/\sigma_{rand-j} \text{ for } j=1 \text{ to } 13$$

For a normally distributed random number, $-3 \leq S_j \leq 3$ implies a rare occurrence (3σ limit). Any pattern with its $S_j > 2$ is considered a motif and is an over-represented pattern [Milo R et al],. Any pattern with its $S_j < -2$ is an anti-motif, and is an under-represented pattern.

2.2. Motif Significance Profile (MSP)

S is a vector of size 13 that defines significance of 13 patterns in the real network. Milo R et al argue that **S** is influenced by the size of the network and propose normalization of **S** to make it largely independent of network size. Thus, **significance profile** vector, **Z** is defined as $Z_j = S_j / |\mathbf{S}|$. This makes comparison of networks of varying sizes possible.

2.3. Correlation of Motif Significance Profiles

Standard correlation coefficients (Pearson correlation coefficient) between **Z** vectors of two systems is used as a measure of **similarity between significance profiles** of two systems. The correlation coefficient can vary from -1 to +1. A value of +1 implies that the 13 patterns are present to the same extent in both systems i.e., if a particular pattern is over-represented (under-represented) in one system it will be over-represented (under-represented) in the other system to the same extent.

3. Architectures of diverse Systems

3.1. Brief Details of considered Systems

In this paper we consider 38 arbitrarily chosen **systems** from literature and study their architectures. Systems considered range from aircraft engine [Manuel E Sosa et al], softwares [Software graph data], electronic circuits [ISCAS High level models, ISCAS'89 benchmark data], robot [Amro M. Farid et al], refrigerator [Thomas U. Pimmler and Steven D. Eppinger], bacteria e-coli [Bacteria e-coil, yeast S], yeast S. cerevisiae [Bacteria e-coil, yeast S], language networks [Kaufman L, Rousseeuw P J.]. These 38 systems are of vastly different sizes (ranging from minimum 16 components to maximum 23843 components). Table 3.1 briefly identifies each of the 38 systems along with their sizes. In electronic circuits, nodes represent component gates and edges represent the interconnection between gates. In case of software systems, nodes represent a software class and edges represent reference between classes. In mechanical systems, nodes represent physical components and edges represent exchange of energy, material or signal between components. In case of biological systems, nodes represent genes and edges represent direct transcription interactions. In case of languages, each node represents a word and an edge occurs when one directly follows the other in the text.

3.2. Similarities in MSP across Systems

Shaja and Sudhakar in [Shaja A S, Sudhakar K, INCOSE 2009] have studied the MSP of the considered engineering systems (aircraft engine, softwares, electronic circuits, robot and refrigerator). Uri Alon et al have studied the MSP of the above considered languages and biology systems. They investigated the similarities in significance profiles across the considered systems by computing correlation coefficient between each pair. This combined information can be presented as a square matrix of size 38. Diagonal elements of the matrix represent similarity of significance profile of a system with itself and are always +1. Off diagonal elements can take values in the range

-1 to +1. In Figure 3.2 the full range of values (-1 to +1) is grouped into 3 regions and indicated by 3 different colors for visual impact.

- +0.65 ≤ Correlation coefficient ≤ +1.00, Positively correlated, red color
- \- 0.65 < Correlation coefficient < +0.65, Weak or no correlation, light green color
- \- 1.0 ≤ Correlation coefficient ≤ -0.65, Negatively correlated, blue color

Shaja and Sudhakar in [Shaja A S, Sudhakar K, INCOSE 2009] performed clustering using a standard clustering algorithm (partitioning around medoids algorithm [Kaufman L et al]), where edge weights are clustering coefficients. The square matrix after regrouping of systems based on clustering is shown in Fig 3.2. There are 5 distinct groups as revealed by 5 red colored blocks along the diagonal. These 5 blocks contain systems whose significance profiles are all positively correlated with respect to each other. It is interesting to note that the above grouping coincides with standard grouping as biology, languages, software, electrical and mechanical.

System no	System Name	Nodes	System no	System Name	Nodes	System no	System Name	Nodes
S1	Digital Fractional Multiplier (s208)	122	S14	ECAT (c1355)	1355	S27	ALU (c7552)	3718
S2	Digital Fractional Multiplier (s420)	252	S15	Forward logic chips (s9234)	5844	S28	PLD (s641)	433
S3	Digital Fractional Multiplier (s838)	512	S16	Forward logic chips (s13207)	8651	S29	ECAT (c1908)	913
S4	Ecoli	423	S17	Forward logic chips (s15850)	10383	S30	ALU (c3540)	1719
S5	Yeast	688	S18	Forward logic chips (s38417)	23843	S31	Traffic control system (s562)	217
S6	Apword	1096	S19	Forward logic chips (s38584)	20717	S32	Aircraft Engine	54
S7	Linux	5420	S20	Traffic control system (s400)	186	S33	Refrigerator	16
S8	Mysql	1501	S21	PLD (s820)	312	S34	Robot	28
S9	Vtk	778	S22	Traffic control system (s382)	182	S35	English	7724
S10	Xmms	1097	S23	ALU (74181)	87	S36	French	9424
S11	Traffic control system (s444)	205	S24	PLD (s832)	310	S37	Japanese	3177
S12	PLD (s713)	447	S25	ECAT (c499)	243	S38	Spanish	12642
S13	ALU (c2670)	1350	S26	ALU (c880)	443			

Table 3.1 38 systems considered

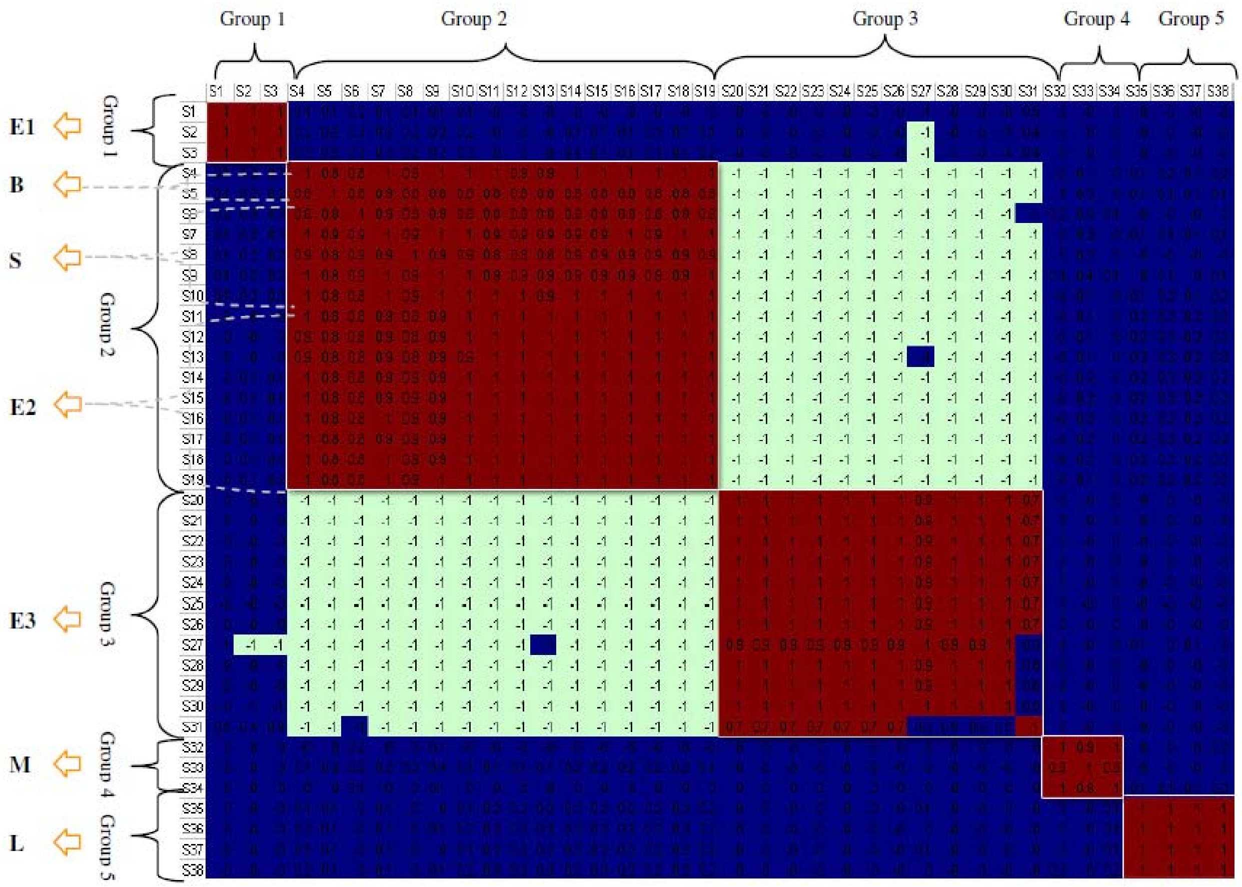


Fig 3.2 Similarities in significance profiles (MSP) across all 38 systems after clustering

4. Similarities through Component Characteristics

4.1. Component Characteristics

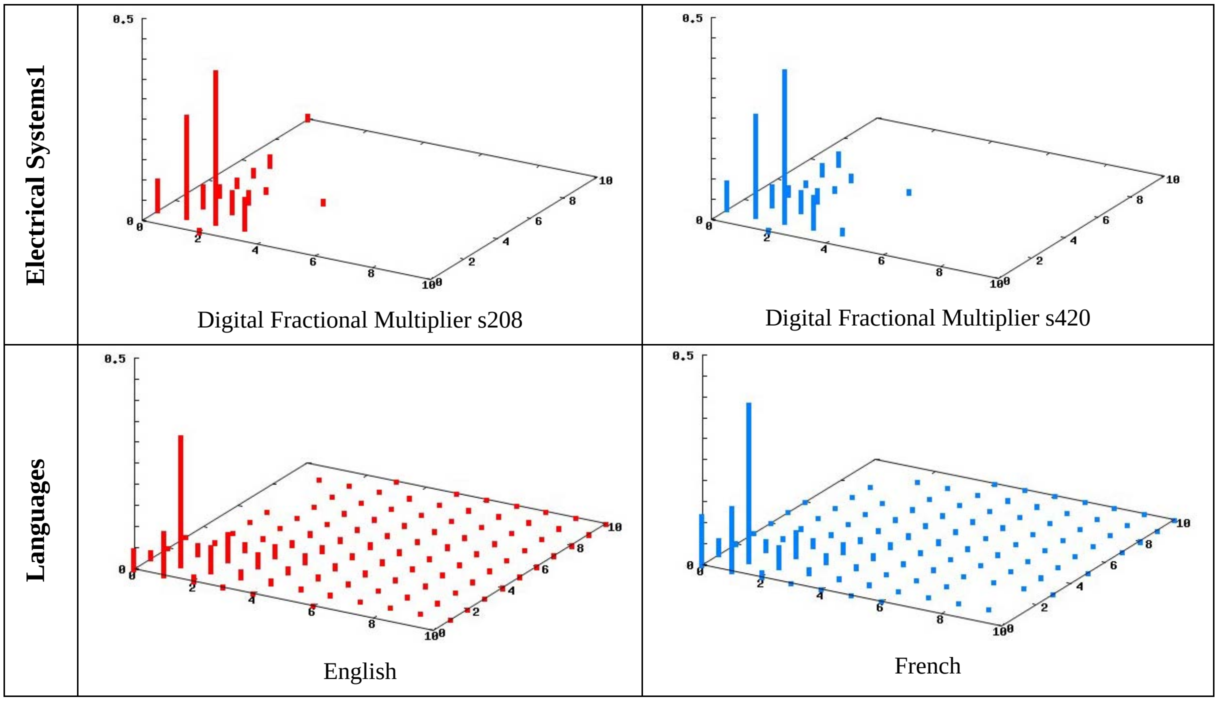
We define the component characteristics in terms of the in-degree and out-degree of a component/node. Each system can be characterized by distribution (counts) of nodes of different in/out-degrees that constitute the system. The correlation between distributions of pairs systems can be estimated. We find that the correlation coefficient so defined can classify systems in respective superfamilies.

4.2. Component Characteristic Profiles (CCP)

Let us now consider an integer plot – considering in-degree of component on X-axis, out-degree of component on Y-axis and count of components having the corresponding in-degree, out-degree pairs on Z-axis. One can visualize this as a histogram – i.e., each unit-square on the plot shows how many components have that in/out-degree. The in-degree (X-axis) and out-degree (Y-axis) are considered till only 10 unit counts. Adequacy of the length of X and Y axis is confirmed by the fact that very few components across all systems considered have more than 10 in-degree or out-degree. In order to compare systems of various sizes (number of nodes), we normalize the plot. This is done by dividing the number in each square (Z-axis) by the total components of that system. Let us call this as the Component Characteristic Profile of a System.

4.3. Similarities in Component Characteristic Profiles across Systems

We create component characteristic profiles for all the considered systems. We find striking resemblances of the CCPs of systems within the same family/group. Fig 3.4 shows CCPs of 5 system families, with two systems from each family to highlight how similar they look within each family and how different they look across families.

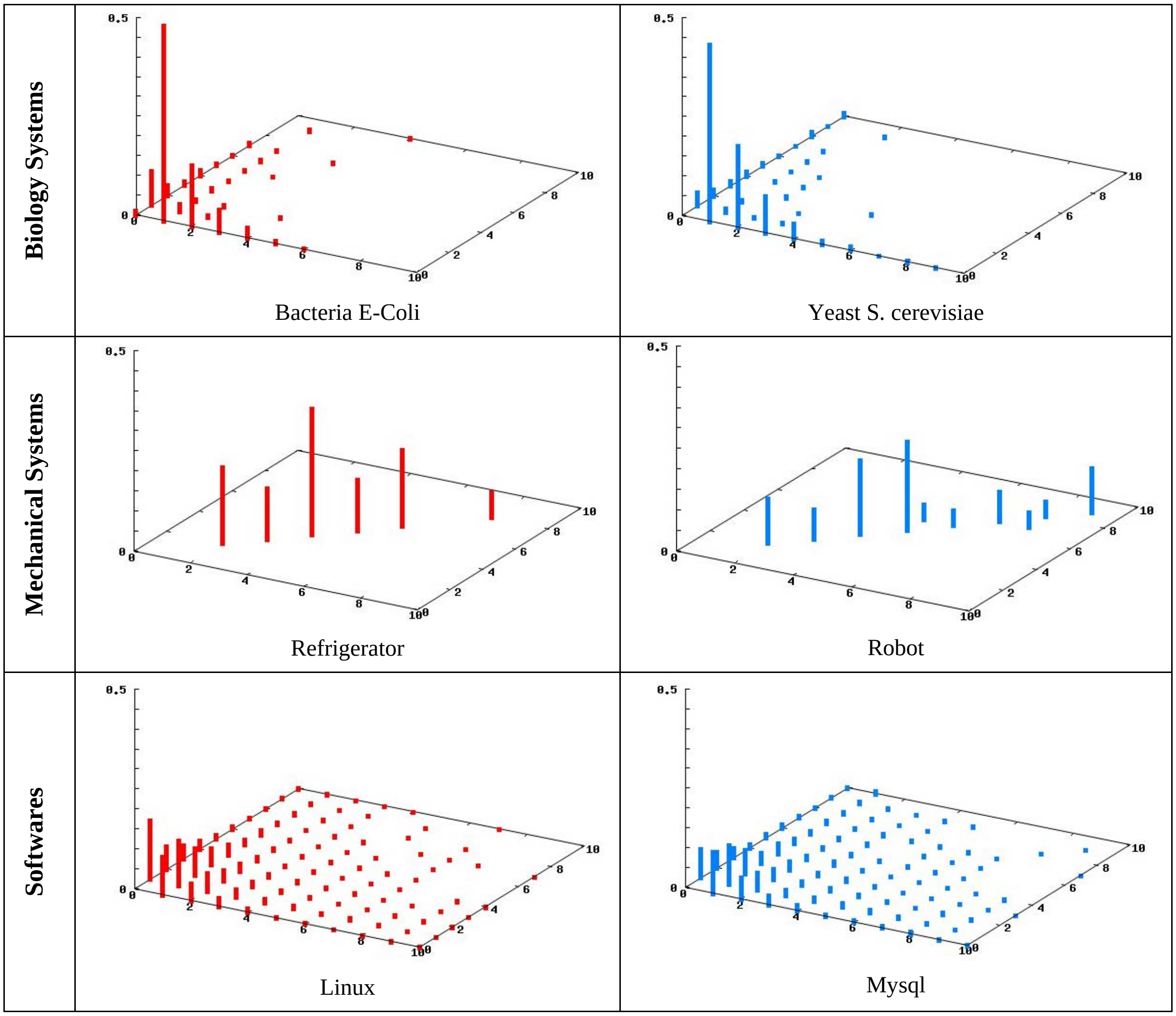


Digital Fractional Multiplier s208

Digital Fractional Multiplier s420

English

French



Bacteria E-Coli

Yeast S. cerevisiae

Refrigerator

Robot

Linux

Mysql

Fig 3.3 Similarities in Component Characteristic Profiles across Systems

We now propose a mathematical basis to capture the likeness or otherwise of a pair of CCP profiles. The Z is defined for each system as a real value (<1) for integer values of X & Y, each going from 0 to 11. ie. Z is a matrix of size 11 x 11 of real values. Z can also be visualized a vector size 11x11 =121. Thus CCP of each system is a vector of size 121. We can now define a correlation coefficient of a pair of systems as we did for MSP. This information is presented in Fig 3.4 as a square matrix (same terminology and legends as in Fig 3.2 are used) except the correlation in Fig 3.2 is based on MSPs and in Fig 3.4 it is based on CCPs. The corresponding systems falling inside the related superfamilies can be easily seen from the matrix.

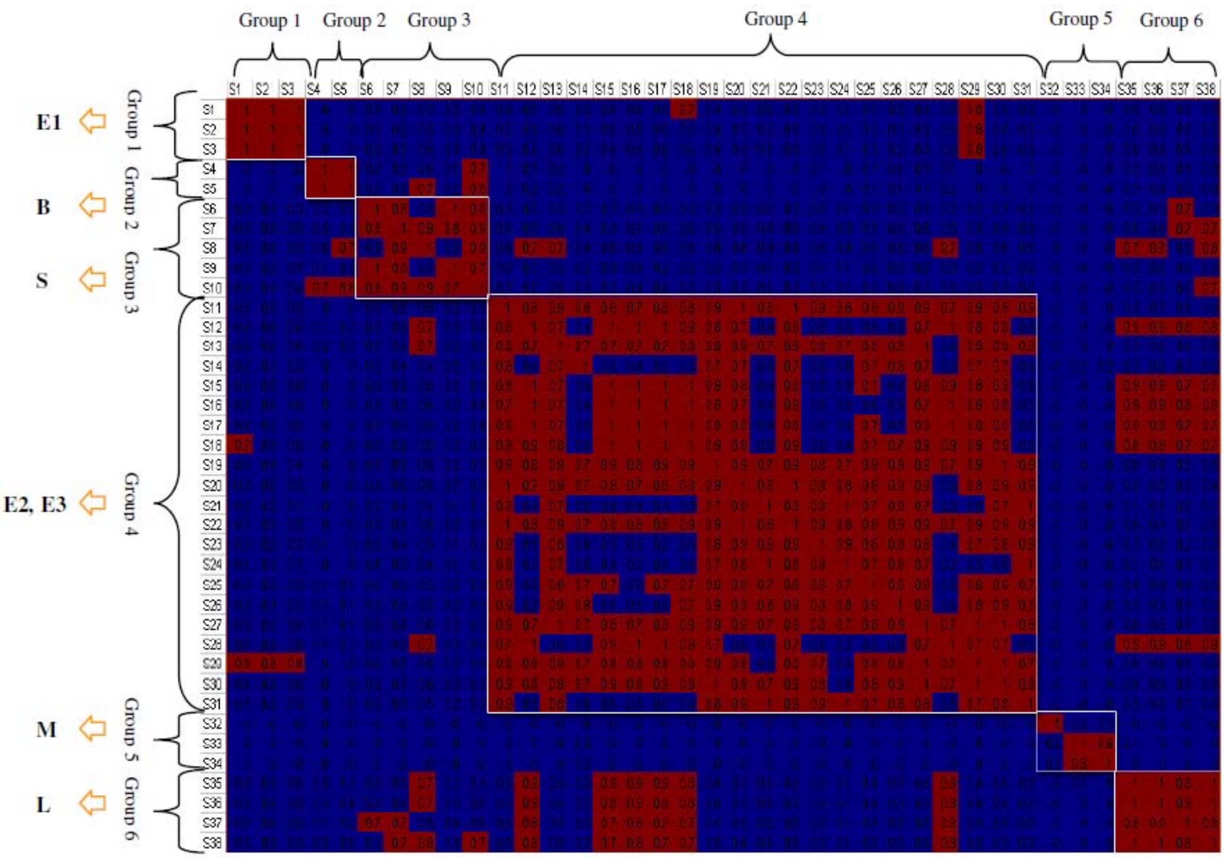


Fig 3.4 Similarities in component characteristic profiles (CCP) across all 38 systems after clustering

4.4. Discussion

Table 3.2 helps us to understand/compare how close are the correlations due to MSPs (Fig 3.2) and CSPs (Fig 3.4) of these families.

S.No	Family	Symbol in Figure 3.2, 3.4	System No	Average MSP Correlation	Average CCP Correlation	MSP Correlated?	CCP Correlated?
1	Electronic Circuits-I	E1	S1-S3	0.9	0.9	Yes	Yes
2	Biological Systems	B	S4-S5	0.9	0.9	Yes	Yes
3	Software Systems	S	S6-S10	0.9	0.9	Yes	Yes
4	Electronic Circuits-II	E2	S11-S19	0.9	0.8	Yes	Yes
5	Electronic Circuits-III	E3	S20-S31	0.9		Yes	Yes

6	Mechanical Engineering Systems	M	S32-S34	0.9	0.7	Yes	Yes
7	Languages	L	S35-S38	0.9	0.9	Yes	Yes

Table 3.2 Relationship between the correlation families due to MSPs and CCPs

We find that, all the systems of a family seems to be using similar distribution of in-degree and out-degree. This can be seen from Fig 3.3 and Fig 3.4 as well. Apart from the tabulated correlation of families, most of the super-families (marked as groups in Fig 3.2 and Fig 3.4) also have close similarities. For instance, the groups 1, 4, 5 formed by MSP and groups 1, 5, 6 formed by CCP respectively are similar. These groups stand out to be electronic circuits (digital fractional multipliers), mechanical systems and languages respectively. The superfamilies corresponding to biological systems and softwares stand as separate families with respect to CCPs, however they fall into the same superfamily with respect to MSP.

5. Conclusion & Directions

Ideas related to complex system architectures may give insight into previously complex and poorly understood phenomena in engineering domains. Albert Barabasi argues that, “The science of networks is experiencing a boom. But despite the necessary multi-disciplinary approach to tackle the theory of complexity, scientists remain largely compartmentalized in their separate disciplines”. The application of this complex system architectures theory is still in infancy and has very recently entered into study of engineering systems or their design. This paper has shown that the superfamiles classification can be arrived at by merely looking at components/nodes characteristics, i.e., the over-represented pattern signature classification of the system architectures across families is due to the properties of the individual components that form the system. This is shown by considering 38 arbitrarily chosen systems ranging from - biology systems, languages, electronic circuits, software systems and mechanical engineering systems. This study has thrown some new insights about Classification of Systems from Component Characteristics.

6. Acknowledgements

We thank Aeronautics Research and Development Board (ARDB), India for this project grant. We thank CASDE, IIT Bombay for an excellent working environment. We are also thankful to Mr R Mahesh for his support in plotting the graphs.

7. References

Duncan J Watts. 2004, Six Degrees: The Science of a Connected Age. Norton and Company, NewYork.

Boccaletti S, Latora V, Moreno Y, Chavez M. and Hwang D. 2006, Complex networks: Structure and dynamics, *Physics Reports,* 424 (4): 175-308. doi:10.1016/j.physrep.2005.10.009

Ottino J M. 2004, Engineering Complex Systems, *Nature,* 427: 399. doi:10.1038/427399a

Zhi-Qiang Jiang, Wei-Xing Zhou, Bing Xu and Wei-Kang Yuan. 2007, Process flow diagram of an Ammonia Plant as a Complex Network, *AIChE Journal*, 53 (2): 423-428. doi:10.1002/aic.11071

ANSI IEEE Standard 1471. `http://www.iso-architecture.org/ieee-1471/`

Edward Crawley, Olivier de Weck, Steven Eppinger, Christopher Magee, Joel Moses, Warren Seering, Joel Schindall, David Wallace and Daniel Whitney. 2004, The Influence of Architecture in Engineering Systems. *MIT Engineering System Division Monograph*.

Tyson R Browning. 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

Albert L and Albert R. 2004, Conserved Network Motifs Allow Protein-Protein Interaction Prediction, *Bioinformatics* (20): 18, pp 3346-3352. doi:10.1093/bioinformatics/bth402

Ronen M, Rosenberg R, Shraiman B I and Alon U. 2002, Assigning Numbers to the Arrow: Parameterizing a Gene Regulation Network by Using Accurate Expression Kinetics, *Proceedings of the National Academy of Sciences* (99): 16, 10555-10560. doi:10.1073/pnas.152046799

Villas Boas, P.R. Rodrigues, F.A., Travieso, G., and Da Fontoura Costa, L., February 2008, Chain motifs: The tails and handles of complex networks, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 77(2), Article number 026106. doi: 10.1103/PhysRevE.77.026106

Goemann, B., Wingender, E., Potapov, A.P. May 2009, An approach to evaluate the topological significance of motifs and other patterns in regulatory networks, *BMC Systems Biology* 3 (53) doi:10.1186/1752-0509-3-53

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U. March 2004, Superfamilies of evolved and designed networks, *Science* 303 (5663), pp. 1538 - 1542 doi:10.1126/science.1089167

Shaja, A.S., Sudhakar, K., 20-23 July 2009, Overrepresented and Underrepresented Patterns in System Architectures across Diverse Engineering Systems, *19th Annual INCOSE International Symposium*, Singapore

Ahnert, S.E., Fink, T.M.A. 2008, Clustering signatures classify directed networks, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 78 (3), art. no. 036112, doi:10.1103/PhysRevE.78.036112

Milo R, Shen-Orr, Itzkovitz S, Kashtan N, Chklovskii D and Alon. 2002 , Network Motifs: Simple Building Blocks of Complex Networks. *Science*: 298, 824-827. doi:10.1126/science.298.5594.824

Manuel E Sosa, Steven D Eppinger and Craig M. Rowles. 2003, Identifying Modular and Integrative Systems and Their Impact on Design Team Interactions, *Journal of Mechanical Design*, 125 (2): 240-252. doi: 10.1115/1.1564074

Software graph data for specified software systems: `http://www.tc.cornell.edu/~myers/Data/SoftwareGraphs/index.htm`

ISCAS High level models: `http://www.eecs.umich.edu/~jhayes/iscas.restore/`

ISCAS'89 benchmark data: `http://www.pld.ttu.ee/~maksim/benchmarks/iscas89/bench/`

Amro M. Farid and Duncan C McFarlane. 2006, An Approach to the Application of the Design Structure Matrix for Assessing,Reconfigurability of Distributed Manufacturing Systems, *Proceedings of the IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications* , art. no. 1633429, pp. 121-126. doi: 10.1109/DIS.2006.10

Thomas U. Pimmler and Steven D. Eppinger. September 1994, Integration analysis of product decompositions, *ASME Design Theory and Methodology Conference Minneapolis.*

Datasets for Bacteria e-coil, yeast S. cerevisiae and language networks `http://www.weizmann.ac.il/mcb/UriAlon/groupNetworksData.html`

Kaufman L, Rousseeuw P J. 1990, Finding Groups in Data: An Introduction to cluster Analysis, New york, *John wiley & Sons Inc.*

8. Biography

A. S. Shaja is currently a doctoral student in Aerospace Engineering Department, Indian Institute of Technology Bombay. He previously did his M.Tech in Indian Institute of Technology Bombay. His recent research interests include Multidisciplinary Design Optimization, Systems Engineering, Networks etc.

K. Sudhakar is a Professor in Aerospace Engineering Department, Indian Institute of Technology Bombay, for the last 25 years. He was previously associated with Aerospace Industry as a Design Expert for 12 years. He completed his B.Tech, M.Tech and Doctorate from Indian Institute of Technology. Some of his recent research interests include optimization in design, Flight Mechanics, Systems Engineering, Mini Aerial Vehicle, Modeling and simulation etc.