PHYSICAL REVIEW E **69**, 016113, 2004

The Topology of Large-Scale Engineering Problem-Solving Networks

by

Dan Braha[1, 2] and Yaneer Bar-Yam[2, 3]

[1] Faculty of Engineering Sciences
Ben-Gurion University, P.O.Box 653
Beer-Sheva 84105, Israel
E-mail: brahad@bgumail.bgu.ac.il

[2] New England Complex Systems Institute
Cambridge, Massachusetts 02138, USA
E-mail: yaneer@necsi.org

[3] Department of Molecular and Cellular Biology,
Harvard University, Cambridge, MA 02138, U.S.A.

ABSTRACT

The last few years have led to a series of discoveries that uncovered statistical properties, which are common to a variety of diverse real-world social, information, biological and technological networks. The goal of the present paper is to investigate, for the first time, the statistical properties of networks of people engaged in distributed problem solving and discuss their significance. We show that problem-solving networks have properties (sparseness, small world, scaling regimes) that are like those displayed by information, biological and technological networks. More importantly, we demonstrate a previously unreported difference between the distribution of incoming and outgoing links of directed networks. Specifically, the incoming link distributions have sharp cutoffs that are substantially lower than those of the outgoing link distributions (sometimes the outgoing cutoffs are not even present). This asymmetry can be explained by considering the dynamical interactions that take place in distributed problem solving, and may be related to differences between the actor's capacity to process information provided by others and the actor's capacity to transmit information over the network. We conjecture that the asymmetric link distribution is likely to hold for other human or non-human directed networks as well when nodes represent information processing/using elements.

I. INTRODUCTION

Distributed problem solving, which often involves an intricate network of interconnected tasks carried out by hundreds of designers, is fundamental to the creation of complex manmade systems [1]. The interdependence between the various tasks makes the system development (referred to as Product development, PD) fundamentally iterative [2]. This process is driven by the repetition (rework) of tasks due to the availability of new information (generated by other tasks) such as changes in input, updates of shared assumptions or the discovery of errors. In such an intricate network of interactions, iterations occur when some development tasks are attempted even though the complete predecessor information is not available or known with certainty [3]. As this missing or uncertain information becomes available, the tasks are repeated to either verify an initial estimate/guess or to come closer to the design specifications. This iterative process proceeds until convergence occurs [3-5].

Design iterations, which are the result of the PD network structure, might slow down the PD convergence or have a destabilizing effect on the system's behavior. This will delay the time required for product development, and thus compromise the effectiveness and efficiency of the PD process. For example, it is estimated that iteration costs about one-third of the whole PD time [6] while lost profits result when new products are delayed in development and shipped late [7]. Characterizing the *real-world* structure, and eventually the dynamics of complex PD networks, may lead to the development of guidelines for coping with complexity. It would also suggest ways for improving the decision making process, and the search for innovative design solutions.

The last few years have witnessed substantial and dramatic new advances in understanding the large-scale structural properties of many real-world complex networks [8-10]. The availability of large-scale empirical data on one hand and the advance in computing power have led to a series of discoveries that uncovered statistical properties, which are common to a variety of diverse real-world social, information, biological and technological networks including the world-wide web [11], the internet [12], power grids [13], metabolic and protein networks [14, 15], food webs [16], scientific collaboration networks [17-20], citation networks [21], electronic circuits [22], and software architecture [23]. These studies have shown that many complex networks exhibit the "small-world" property of short average path lengths between any two nodes despite being highly clustered. The second property states that complex networks are characterized by an inhomogeneous distribution of nodal degrees (the number of nodes a particular node is connected to) following a power law distribution (termed "scale free" networks in [29]). Scale-free networks have been shown to be robust to random failures of nodes, but vulnerable to unexpected failure of the highly connected nodes [24]. A variety of network growth processes that might occur on real networks, and that lead to scale-free and small-world networks have been proposed [9, 10].

Planning techniques and analytical models that conceive the PD process as a network of interacting components have been proposed before [3, 25, 26]. However, others have not yet addressed the large-scale statistical properties of real-world PD task networks. In the research we report here, we study such networks. We show that task networks have properties (sparseness, small world, scaling regimes) that are like those of other biological, social and technological networks. We also demonstrate a previously unreported observation involving an asymmetry between the distribution of incoming links and the probability of outgoing links.

The paper is organized as follows: In Sec. II, we present evidence suggesting that PD task networks have the small-world property. We also demonstrate the distinct roles of incoming and outgoing information flows in distributed PD processes by analyzing the corresponding in-degree and out-degree link distributions. In Sec. III we provide our conclusions.

II. RESULTS

A. Small world properties

We analyzed distributed product development data of different large-scale organizations in the United States and England involved in vehicle development, operating software development, pharmaceutical facility development, and a sixteen story hospital facility development. A PD distributed network can be considered as a directed graph with N nodes and L arcs, where there is an arc from task v_i to task v_j if task v_i *feeds information to* task v_j. The information flow forming the directed links between the tasks has been based on structured interviews with experienced engineers, and design documentation data (design process models). In all cases, the repeated nature of the product development projects and the knowledgeable people involved in eliciting the information flow dependencies reduce the risk of error in the construction of the product development networks. More specifically, the vehicle development network was identified by *directly* questioning at least one engineer from each task “where do the inputs for the task come from (another task)?” and “where do the outputs generated by the task go to (another task)?” The answers to these questions are used to construct the network of information flows [33]. The three other larger networks [34] have been constructed based on data flow and design-process model diagrams (see [35] for a detailed description). An example of a diagram from the sixteen-story hospital facility process model is shown in Figure 1.

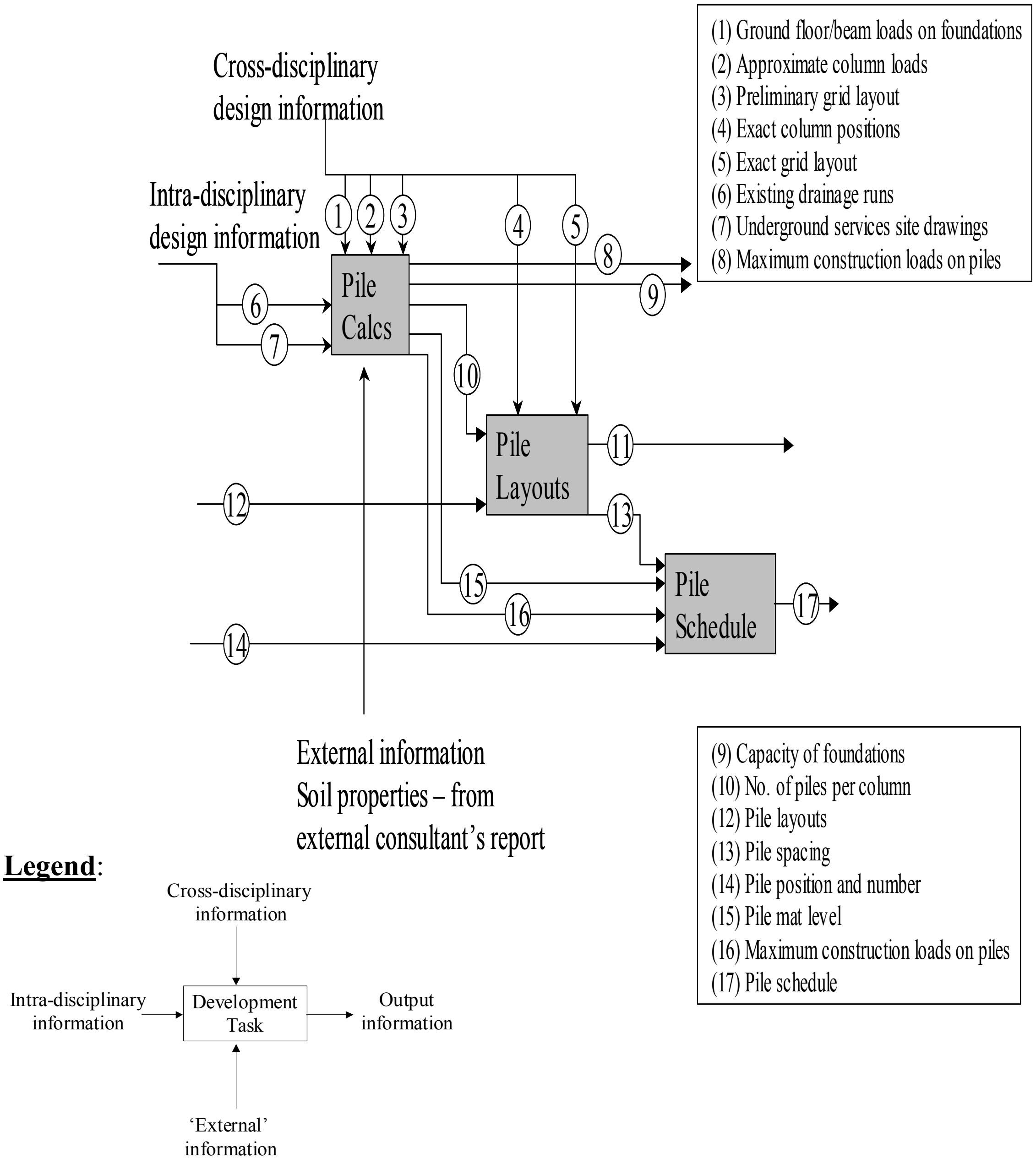
Cross-disciplinary design information
Intra-disciplinary design information
Pile Calcs
Pile Layouts
Pile Schedule
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
(1) Ground floor/beam loads on foundations
(2) Approximate column loads
(3) Preliminary grid layout
(4) Exact column positions
(5) Exact grid layout
(6) Existing drainage runs
(7) Underground services site drawings
(8) Maximum construction loads on piles
External information
Soil properties – from external consultant's report
(9) Capacity of foundations
(10) No. of piles per column
(12) Pile layouts
(13) Pile spacing
(14) Pile position and number
(15) Pile mat level
(16) Maximum construction loads on piles
(17) Pile schedule
Legend:
Cross-disciplinary information
Intra-disciplinary information
Development Task
Output information
'External' information

FIG. 1. Example of a diagram from a design process model used to construct the sixteen-story hospital facility development (adapted from [35]).

An example of one of these distributed PD networks (operating software development) is shown in Figure 2. Here we consider the undirected version of the network, where there is an edge between two tasks if they exchange information between them (not necessarily reciprocal). We see that this network is sparse ($2L/N(N-1)=0.0114911$) with the average total degree of each node only 4.116, which is small compared to the number of possible edges $N-1=465$. A clear deviation from a purely random graph is observed. We see that most of the nodes have low degree while a few nodes have a very large degree. This is in contrast to the nodal degree homogeneity of purely random graphs, where most of the nodal degrees are concentrated around the mean. The software development network also illustrates the 'small-world' property, which can be detected by measuring two basic statistical characteristics. The first characteristic is the average distance (geodesic) between two nodes, where the distance $d(i,j)$ between nodes v_i and v_j is defined as the number of edges along the shortest path connecting them. The characteristic path length L is the average distance between any two vertices:

$$L=\frac{1}{N(N-1)}\sum_{i\neq j}d_{ij} \tag{1}$$

The second characteristic measures the tendency of vertices to cluster in densely interconnected modules. The clustering coefficient C_i of a vertex v_i is defined as follows. Let vertex v_i be connected to k_i neighbors. The total number of edges between these neighbors is at most $k_i(k_i-1)/2$. If the actual number of edges between these k_i neighbors is n_i, then the clustering coefficient C_i of the vertex v_i is the ratio

$$C_i=\frac{2n_i}{k_i(k_i-1)} \tag{2}$$

The clustering coefficient of the graph, which is a measure of the network's potential modularity, is the average over all vertices,

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \qquad (3)$$

Small-world networks are a class of graphs that are highly clustered like regular graphs ($C_{\text{real}} \gg C_{\text{random}}$), but with small characteristic path length like a random graph ($\ell_{\text{real}} \approx \ell_{\text{random}}$). For the software development network, the network is highly clustered as measured by the clustering coefficient of the graph ($C_{\text{software}} = 0.327$) compared to a random graph with the same number of nodes and edges ($C_{\text{random}} = 0.021$) but with small characteristic path length like a random graph ($\ell_{\text{software}} = 3.700 \approx \ell_{\text{random}} = 3.448$).

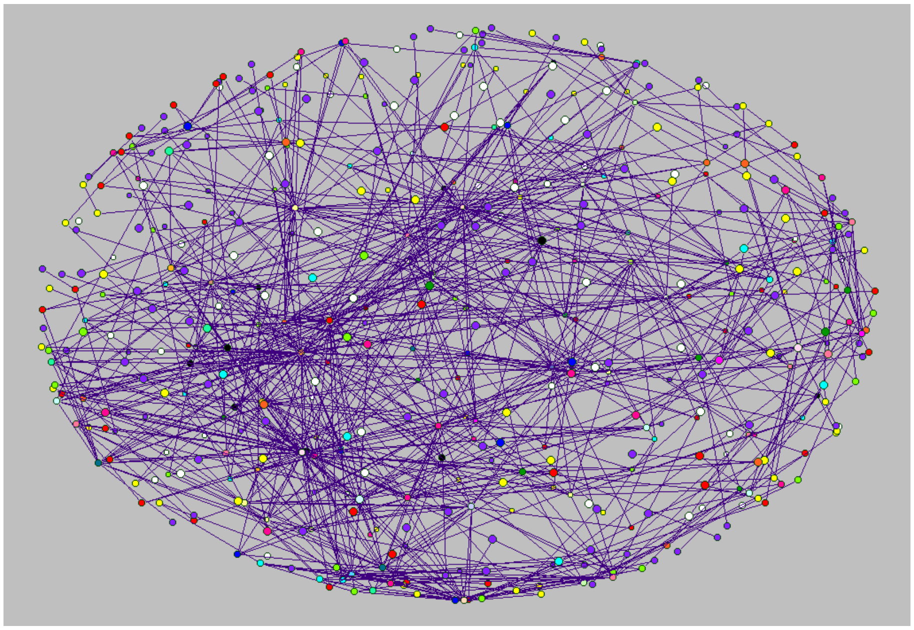

FIG. 2. Network of information flows between tasks of an operating system development process. This PD task network consists of 1245 directed information flows between 466 development tasks. Each task is assigned to one or more actors ("design teams" or "engineers") who are responsible for it. Nodes with the same degree are colored the same.

In Table 1, we present the characteristic path length and clustering coefficient for the four distributed PD networks examined in this paper, and compare their values with random graphs with the same number of nodes and edges. In all cases, the empirical results display the small-world property ($C_{\text{real}} \gg C_{\text{random}}$ and $\ell_{\text{real}} \approx \ell_{\text{random}}$).

TABLE 1 Empirical Statistics of the four large-scale PD Networks

Network	N	L	C	ℓ	C_{random}	ℓ_{random}
Vehicle	120	417	0.205	2.878	0.070	2.698
Operating Software[*]	466	1245	0.327	3.700	0.021	3.448
Pharmaceutical Facility	582	4123	0.449	2.628	0.023	2.771
Sixteen story Hospital Facility[*]	889	8178	0.274	3.118	0.024	2.583

[*] We restrict attention to the largest connected component of the graph, which includes ~82% of all tasks for the Operating Software network, and ~92% of all tasks for the Sixteen story Hospital Facility network.

Shorter development times, improved product quality, and lower development costs are the key factors for successful complex PD processes. The existence of cycles in the PD networks points to the seemingly undeniable truth that there is an inherent, iterative nature to the design process [2]. Each iteration results in changes that must propagate through the PD network requiring the rework of other reachable tasks. Consequently, late feedback and excessive rework should be minimized if shorter development time is required.

The functional significance of the small-world property can be attributed to the fast information transfer throughout the network, which results in immediate response to the rework created by other tasks in the network. The high clustering coefficient of PD networks suggests an inherently modular organization of PD processes; i.e., the organization of the PD process in clusters that contain most, if not all, of the interactions internally and the interactions or links between separate clusters is eliminated or minimized [1-3]. The dynamic model developed in [5] shows that a speed up of the PD convergence to the design solution is obtained by reducing or 'ignoring' some of the task dependencies (i.e., eliminating some of the arcs in the corresponding PD network). A modular architecture of the PD process is aligned with this strategy.

B. In-degree and out-degree distributions

We compared the cumulative probability distributions $P_{in}(k)$ and $P_{out}(k)$ that a task has more than k incoming and outgoing links, respectively (see Figure 2)[30]. For all four networks, we find that the in-degree and out-degree distributions can be described by power-laws with cutoffs introduced at some characteristic scale k^*; $k^{-\gamma} f(k/k^*)$ (typically the function f corresponds to exponential or Gaussian distributions). More specifically, we find scaling regimes (i.e., straight-line regimes) for both $P_{in}(k)$ and $P_{out}(k)$; however, the cutoff k^* occurs lower (by more than a factor of two) for $P_{in}(k)$ than for $P_{out}(k)$.

The presence of cutoffs in the in-degree and out-degree distributions is consistent with a conjecture by Amaral et al. [17] that physical costs of adding links and limited capacity of a node should lead to a power-law regime followed by a sharp cutoff (this conjecture has been tested for undirected networks). Our empirical results are also consistent with Mossa et al. [31] who suggest that making decisions on new Internet links, based on filtered information, leads to an exponential cutoff of the in-degree distribution for networks growing under conditions of preferential attachment. Both Amaral et al [17] and Mossa et al. [31] comment that, in the context of network growth, the presence of costly connections, limited capacity of a node, or limited information-processing capability of a node are not unlike the so-called “bounded rationality” concept of Simon [28]. Our findings suggest that although the cutoff may be attributed to constraints on the information-processing capacities of the actors carrying out the development process (in accordance with the “bounded rationality” concept) , there is an *asymmetry* between the distributions of incoming and outgoing information flows. The narrower power law regime for $P_{in}(k)$ suggests that the costs of adding incoming links and limited in-degree capacity of a task are higher than their counterpart out-degree links. We note that this is consistent with the realization that bounded rationality applies to incoming information, and to outgoing information only when it is different for each recipient, not when it is duplicated. This naturally leads to a weaker restriction on the out-degree distribution.

An additional functional significance of the asymmetric topology can be attributed to the distinct roles of incoming and outgoing links in distributed PD processes. The narrow scaling regime governing the information flowing into a task implies that tasks with large incoming connectivity are practically absent. This indicates that distributed PD networks strive to limit conflicts by reducing the multiplicity of interactions that affect a single task, as reflected in the incoming links. This characteristic reduces the amount and range of potential revisions that occur in the dynamic PD process, and thus increases the likelihood of converging to a successful solution. This empirical observation is found to be consistent with the dynamic PD model (using *linear systems theory*) developed in [5]. There it was shown that additional rework might slow down the PD convergence or have a destabilizing effect on the system’s behavior. As a general rule, the rate of problem solving has to be measured and controlled such that the total number of design problems being created is smaller than the total number of design problems being solved.

The scale-free nature of the outgoing communication links means that some tasks communicate their outcomes to many more tasks than others do, and may play the role of coordinators (or product integrators see [5]). Unlike the case of large numbers of incoming links, this may improve the integration and consistency of the problem solving process; thus reducing the number of potential conflicts. Product integrators put the separate development tasks together to ensure fit and functionality. Since late changes in product design are highly expensive, product integrators continuously check unfinished component designs and provide feedback to a large number of tasks accordingly.

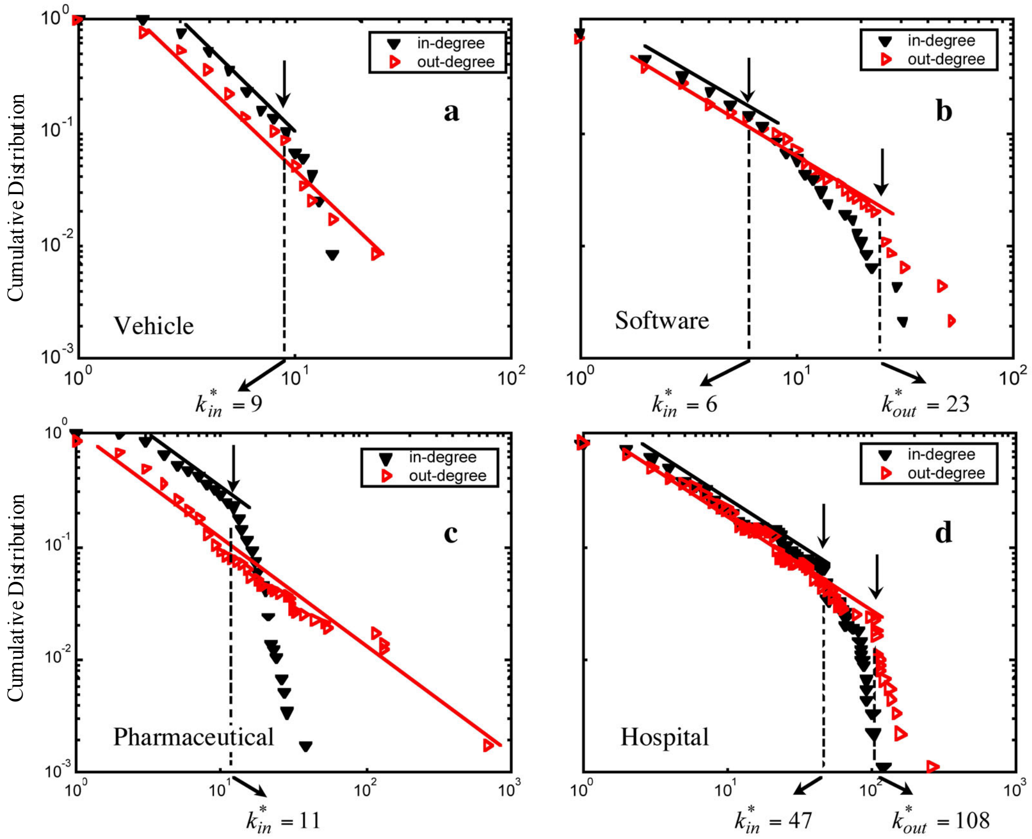

Figure 2 Degree distributions for four distributed problem solving networks. The log-log plots of the cumulative distributions of incoming and outgoing links show a power law regime (Pearson coefficient $R > 0.98$, $p < 0.001$) with or without a fast decaying tail in all cases. The in-degree distribution has a lower best-fit cutoff k^*_{in} in each case. **a**, Vehicle development with 120 tasks and 417 arcs. The exponents of the cumulative distributions are $\gamma^{in}_{vehicle} - 1$ and $\gamma^{out}_{vehicle} - 1$, where $\gamma^{in}_{vehicle} \approx 2.91$ and $\gamma^{out}_{vehicle} = 2.97$ denote the exponents of the associated probability density functions. **b**, Software development with 466 tasks and 1245 arcs, where $\gamma^{in}_{software} \approx 1.97$ and $\gamma^{out}_{software} \approx 2.17$. **c**, Pharmaceutical facility development with 582 tasks and 4123 arcs, where $\gamma^{in}_{pharmaceutical} \approx 1.8$ and $\gamma^{out}_{pharmaceutical} \approx 1.96$. **d**, Hospital facility development with 889 tasks and 8178 arcs, where $\gamma^{in}_{hospital} \approx 1.76$ and $\gamma^{out}_{hospital} \approx 1.89$.

III. CONCLUSIONS

The study of complex network topologies across many fields of science and technology has become a rapidly advancing area of research in the last few years [8-10]. One of the key areas of research is understanding the network properties that are optimized by specific network architectures [17, 23, 27, 31, 32]. Here we analyzed the statistical properties of real-world networks of people engaged in product development activities. We show that complex PD networks display similar statistical patterns to other real-world networks of different origins. In the context of product development, what is the meaning of these patterns? How do they come to be what they are? We propose several explanations for these patterns.

Successful PD processes in competitive environments are often characterized by short time-to-market, high product performance, and low development costs [7]. An important tradeoff exists in many high technology industries between minimizing time-to-market and development costs and maximizing the product performance. Considering the PD task network, accelerating the PD process can be achieved by "cutting out" some of the links between the tasks [5]. Although the elimination of some arcs should result in a speed up of the PD convergence, this might worsen the performance of the end system. Consequently, a tradeoff exists between the elimination of task dependencies (speeding up the process) and the desire to improve the system's performance through the incorporation of additional task dependencies. PD networks appear to be highly optimized when both PD completion time and product performance are accounted for. Recent studies have shown that an evolutionary algorithm involving *minimization of link density and average distance* between any pair of nodes can lead to non-trivial types of networks including truncated scale-free networks, i.e. $p(k) = k^{-\gamma} f(k/k^{*})$ [23, 27]. This might suggest that an evolutionary process that incorporates similar generic optimization mechanisms (e.g., minimizing a weighted sum of development time and product quality losses) might lead to the formation of a PD network structure with the small-world and truncated scale-free properties.

Another explanation for the characteristic patterns of PD networks might be related to the close interplay between the design structure (product architecture) and the related organization of tasks involved in the design process. It has been observed that in many technical systems design tasks are commonly organized around the architecture of the product [25]. Consequently, there is a strong association between the information flows underlying the PD task network and the design network composed of the physical (or logical) components of the product and the interfaces between them. If the task network is a "mirror image" of the related design network, it is reasonable that their large-scale statistical properties might be similar. Evidence for this can be found in recent empirical studies that show some design networks (electronic circuits [22] and software architectures [23]) exhibit small-world and scaling properties. The scale-free structure of design networks, in turn, might reflect the strategy adopted by many firms of reusing existing modules together with newly developed modules in future product architectures [2]. Thus, the highly connected nodes of the scale-free design network tend to be the most reusable modules. Reusing modules at the product architecture level has also a direct effect on the task level of product development; it allows firms to reduce the complexity and scope of the product development project by exploiting the knowledge embedded in reused modules, and thus significantly reduce the product development time.

We demonstrated a previously unreported difference between the distribution of incoming and outgoing links in a complex network. Specifically, we find that the distribution of outgoing communication links is scale-free (power law decay) with or without a cutoff. The distribution of incoming information flows always has a cutoff, and when both distributions have cutoffs the incoming distribution has a cutoff that is lower by more than a factor of two. From a product development viewpoint, the functional significance of this asymmetric topology has been explained by considering the dynamical interactions that take place in distributed problem solving. PD task networks are one example of directed social, communication or information networks composing a set of people or groups of people with some pattern of interactions between them [10]. Thus, the asymmetric link distribution is likely to hold for other directed networks as well when nodes represent information processing/using elements. This plausibility is based on a bounded-rationality argument originally put forward by Simon in the context of human interactions [28]. Accordingly, this asymmetry could be interpreted as indicating a limitation on the actor's capacity to process information provided by others rather than the ability to transmit information over the network. In the latter case, boundedness is less apparent since the capacity required to transmit information over a network is often less constrained, especially when it is replicated (e.g., many actors can cite a single article). In light of this observation, we expect a distinct cut-off distribution for in-degree as opposed to out-degree distributions when the network reflects communication of information between human beings as a natural and direct outcome of Simon's bounded rationality argument. It would be interesting to see whether this property can be found more generally in other directed human or non-human networks.

Acknowledgements

This work has been conducted while the first author was a Visiting Professor at the Massachusetts Institute of Technology, Center for Innovation of Product Development. The authors would like to thank Alberto Cividanes of the Massachusetts Institute of Technology for providing the vehicle development data at General Motors' Research & Development Center, Stephen Denker of The Business Process Architects for providing the operating software development data at a major telecommunication company, and to Simon Austin and Andy Newton of Loughborough University and ADePT Management Limited for providing the pharmaceutical facility development, and the sixteen story hospital facility development data.

REFERENCES

[1] C. Alexander, *Notes on the Synthesis of Form* (Harvard University Press, Cambridge, MA, 1964).

[2] D. Braha and O. Maimon, *A Mathematical Theory of Design: Foundations, Algorithms, and Applications* (Kluwer Academic Publishers, Boston, MA, 1998).

[3] A. Yassine, and D. Braha, Concurrent Engineering: Research and Applications **11** (3), (2003).

[4] M. Klein, H. Sayama, P. Faratin and Y. Bar-Yam, Concurrent Engineering: Research and Applications, **11** (3), (2003).

[5] A. Yassine, N. Joglekar, D. Braha, S. Eppinger and D. Whitney, *Research in Engineering Design* (to be published).

[6] S. M. Osborne, MSc. Thesis, Massachusetts Institute of Technology, 1993.

[7] K. B. Clark, *Management Science* **35**, 1247–1264 (1989).

[8] S. H. Strogatz, *Nature* **410**, 268-276 (2001).

[9] R. Albert and Barabási, A.-L., Reviews of Modern Physics 74, 47-97 (2002)

[10] M. E. J. Newman, SIAM Review **45**, 167-256 (2003).

[11] R. Albert, H. Jeong, and A.-L. Barabási, *Nature* **401**, 130 (1999).

[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *Comp. Comm. Rev.* **29**, 251-262 (1999).

[13] D. J. Watts, and S.H. Strogatz, *Nature* **393**, 440-442 (1998).

[14] H. Jeong, B. Tombor, R. Albert, Z. N. Oltavi, and A.-L. Barabási, *Nature* **407**, 651-654

(2000).

[15] H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai, *Nature* **411**, 41 (2001).

[16] J. M. Montoya and R. V. Solé, J. Theor. Bio. **214**, 405-412 (2002).

[17] L. A. N. Amaral, A. Scala, M. Barthélémy and H. E. Stanley, *Proc. Nat. Ac. Sci USA* **97**, 11149-11152 (2000).

[18] M. E. J. Newman, *Proc. Nat. Ac. Sci USA* **98**, 404 (2001).

[19] M. E. J. Newman, *Phys. Rev. E* **64**, 016131 (2001).

[20] M. E. J. Newman, *Phys. Rev. E* **64**, 016132 (2001).

[21] D. J. de S. Price, *Science* **149**, 510-515 (1965).

[22] R. F. Cancho, C. Janssen, and R. V. Solé, *Phys. Rev. E* **63** (2001).

[23] S. Valverde, R. F. Cancho, and R. V. Solé, Europhys. Lett. 60, 512-517 (2002).

[24] R. Albert, H. Jeong, and A.-L. Barabási, *Nature* **406**, 378-382 (2000).

[25] S.D. Eppinger, D.E. Whitney, R.P. Smith, and D.A. Gebala, Res. in Eng. Des. **6**, 1-13 (1994).

[26] D.V. Steward, IEEE Trans. on Eng. Man. **28**, 71-74 (1981).

[27] R. F. Cancho, and R. V. Solé, SFI Working Paper 01-11-068 (2001).

[28] H. A. Simon, *The Sciences of the Artificial* (MIT Press, Cambridge, MA, 1998).

[29] A.-L. Barabási, and R. Albert, *Science* **286**, 509-512 (1999).

[30] Note that a power-law distribution of the in-degree distribution (respectively, the out-degree distribution) $p_{in}(k) \sim k^{-\gamma_{in}}$ with exponent γ_{in} translates into a power-law distribution of the cumulative probability distribution $P_{in}(k) \sim \sum_{k'=k}^{\infty} k'^{-\gamma_{in}} \sim k^{-(\gamma_{in}-1)}$ with exponent $\gamma_{in} - 1$.

[31] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral, *Phys. Rev. Lett.* **88**, 138701 (2002).

[32] B. Shargel, H. Sayama, I. R. Epstein and Y. Bar-Yam, *Phys. Rev. Lett.* **90**, 068701 (2003).

[33] A complete description of the tasks, the list of interviewees, and the result of the survey are available at http://necsi.org/projects/braha/largescaleengineering.htmlFor further details regarding the data collection process at GM’s Research & Development Center see A. Cividanes, MSc. Thesis, Mechanical Engineering Department, Massachusetts Institute of Technology, 2002.

[34] Available at http://necsi.org/projects/braha/largescaleengineering.html

[35] S. Austin, A. Baldwin, B. Li and P. Waskett, *Design Studies* **20**, 279–296 (1999).