Are technological and social networks really different?

Daniel E. Whitney

Engineering Systems Division
Massachusetts Institute of Technology
Cambridge, MA 02139
dwhitney@mit.edu

David Alderson

Control and Dynamical Systems
California Institute of Technology
Pasadena, CA 91125
alderd@cds.caltech.edu

The use of the Pearson coefficient (denoted r) to characterize graph assortativity has been applied to networks from a variety of domains. Often, the graphs being compared are vastly different, as measured by their size (i.e., number of nodes and arcs) as well as their aggregate connectivity (i.e., degree sequence D). Although the hypothetical range for the Pearson coefficient is $[-1, +1]$, we show by systematically rewiring 38 example networks while preserving simplicity and connectedness that the actual lower limit may be far from -1 and also that when restricting attention to graphs that are connected and simple, the upper limit is often very far from $+1$. As a result, when interpreting the r-values of two different graphs it is important to consider not just their direct comparison but their values relative to the possible ranges for each respectively. Furthermore, network domain (”social” or ”technological”) is not a reliable predictor of the sign of r. Collectively, we find that in many cases of practical interest, an observed value of r may be explained simply by the constraints imposed by its D, and empirically such constraints are often the case for observed $r < 0$. In other cases, most often for $r > 0$, other explanations must be sought.

1 Introduction

Newman [1] observed that the Pearson degree correlation coefficient r for some kinds of networks is consistently positive while for other kinds it is negative. Several explanations have been offered [2, 3]. In this paper we offer a different explanation based on embedding each subject network in the set of all networks sharing the subject network's degree sequence (denoted here as D).

Our primary contribution is to show with 38 example networks from many domains that the degree sequence dictates in large part the values of r that are possible. More precisely, we show that, although D does not necessarily determine the observed value of r, it conclusively determines the maximum and minimum values of r that each subject network could possibly have, found by rewiring it while preserving its D, its connectedness, and its simpleness. Approaching the problem this way reveals interesting properties of D that affect the range of possible values of r. In particular, we observe for the networks studied here that those exhibiting $r < 0$ are considerably more constrained in their allowable r-values than those exhibiting $r > 0$. After studying these properties and their underlying mathematics, we ask if the alternate wirings are semantically feasible, in an effort to see how the domain of each network might additionally constrain r.[1][2]

2 Observed data and mathematical analysis

Table 1 lists the networks studied and their properties of interest. The values of r_{max} and r_{min} were obtained by systematically rewiring each subject network while preserving connectivity and degree sequence. By "systematically," we mean that pairs of nodes are selected at random for degree-preserving rewiring, but the rewiring is accepted only if the network remains connected, self-loops and multiple edges between nodes are forbidden, and the value of r increases (or decreases). This type of rewiring procedure was used previously by Maslov et al. [4], who argued that graph properties such as assortativity only make sense when the graph of interest is compared to its "randomized" counterpart. The message of this paper is similar in spirit, but focuses on empirical evidence across

[1] No causality is implied. The domain may well provide the constraints that shape D . The present paper does not attempt to assign a causal hierarchy to the constraints.

[2] The networks analyzed here are all simple and connected. The reason for restricting the analysis to connected graphs is that in some cases of disconnected graphs, such as the physics coauthors and company directors analyzed by Newman and Park [3], including or excluding the smaller disconnected components from the calculation of r can have a huge numerical effect. The physics coauthor network consists of one large component with 145 nodes plus 28 tightly clustered isolated components having 2 to 5 nodes each. The value of r for the entire network of 29 components is 0.1515, whereas for the large component alone it is 0.0159 and for the other 28 as a group the value is 0.6795. The small clusters easily exhibit unusually large values of r inasmuch as the calculation obscures the small number of nodes in each one and it is easy to get a large positive r from 5 or 6 nodes with many mutual links. In order to avoid any confusion that could be caused by such disparity of r values in one network, this paper studies only connected networks.

a variety of domains.

The networks in Table 1 are listed in ascending order of r. It should be clear from this table that one find networks of various types, such as "social," "biological," or "technological," having positive or negative values of r. This indicates that networks do not "naturally" have negative r, nor should one require a special explanation why social networks have positive r. All empirical conclusions drawn from observations are subject to change as more observations are obtained, but this is the conclusion we draw based on our data.

In Table 1, the kinds of networks, briefly, are as follows: social networks are coauthor affiliations or clubs; mechanical assemblies comprise parts as nodes and joints between parts as edges; rail lines comprise transfer stations or rail junctions as nodes and tracks as edges; food webs comprise species as nodes and predator-prey relationships as arcs; software call graphs comprise subroutines as nodes and call-dependence relationships as arcs; Design Structure Matrices (DSMs) [9] comprise engineering tasks or decisions as nodes and dependence relationships as arcs; voice/data-com systems comprise switches, routers and central offices as nodes and physical connections (e.g., wire or fiber) as arcs; electric circuits comprise circuit elements as nodes and wires as arcs; and air routes comprise airports or navigational aids as nodes and flight routes as arcs.

In Table 1, we introduce the notion of *elasticity*, defined here as $e = |r_{\max} - r_{\min}|/2$, which reflects the possible range of r relative to the maximum range $[-1, 1]$ obtained for all networks having the same degree sequence. We call a degree sequence with large e *elastic*, while a degree sequence with small e is called *rigid*. The vastly different observed ranges for possible values of r can be explained by a closer look at the respective degree sequences for each network and the way in which they constrain graph features as a whole. In the remainder of this paper, when refering to the degree sequence D for a graph, we mean a sequence $\{d_1, d_2, \ldots, d_n\}$, always assumed to be ordered $d_1 \geq d_2 \geq \ldots \geq d_n$ without loss of generality. The average degree of the network is simply $\langle d \rangle = n^{-1} \sum_{i=1}^{n} d_i$.

For the purposes of this paper, we define the Pearson coefficient (known more generally as the correlation coefficient [5]) as

$$r = \frac{\sum_{(i,j)} (d_i - \bar{d}_i)(d_j - \bar{d}_j)}{\sqrt{\sum_{(i,j)} (d_i - \bar{d}_i)^2 (d_j - \bar{d}_j)^2}}, \tag{1}$$

where, for a network having m links, we define $\bar{d}_i \equiv m^{-1} \sum_{(i,j)} d_i = (2m)^{-1} \sum_k (d_k)^2$. Here, $\bar{d}_i$ is the average degree of a node seen at the end of a randomly selected link. It is easy to see that $\bar{d}_i = \bar{d}_j$ when averaging over links (i, j), so in what follows we will simply refer to $\bar{d}$. Observe that $\bar{d} \neq \langle d \rangle$. In fact, $\bar{d} = (n^{-1} \sum_i d_i^2)/(n^{-1} \sum_j d_j) = \langle d^2 \rangle / \langle d \rangle$, so $\bar{d}$ is a measure of the amount of variation in D.

Figure 1 shows that the most rigid D are characterized by a few dominant nodes of relatively high degree, with the remaining vast majority of nodes having relatively low degree, equivalent to a small supply of available edges and implying

Network	n	m	$\langle d \rangle$	fraction of nodes $d_i > \bar{d}$	$r_{\min}$	r	$r_{\max}$	"elasticity" $\frac{\lvert r_{\max} - r_{\min} \rvert}{2}$
Karate Club	34	78	4.5882	0.1471	-0.806	-0.4756	-0.0139	0.396
"Erdos Network" (Tirole)	93	149	3.204	0.0645	-0.6344	-0.4412	0.0197	0.3073
"Erdos Network" (Stiglitz)	68	85	2.50	0.0882	-0.6417	-0.4366	-0.0528	0.2945
Scheduled Air Routes, US	249	3389	27.22			-0.39		
Littlerock Lake* food web	92	997	10.837	0.337		-0.3264		
Grand Piano Action 1 key	71	92	2.59	0.197	-0.7262	-0.3208	0.8955	0.8108
Santa Fe coauthors	118	198	3.3559	0.0593	-0.5098	-0.2916	0.1412	0.325
V8 engine	243	367	3.01	0.0122	-0.2932	-0.269	-0.1385	0.07735
Grand Piano Action 3 keys	177	242	2.73	0.2034	-0.5375	-0.227	0.7461	0.6418
Abilene-inspired toynet (Internet)	886	896	2.023	0.0158	-0.2300	-0.2239	-0.0379	0.096
Bike	131	208	3.1756	0.0458	-0.435	-0.2018	0.18	0.3075
Six speed transmission	143	244	3.4126	0.1	-0.3701	-0.1833	0.3431	0.3565
"HOT"-inspired toynet (Internet)	1000	1049	2.098	0.0170	-0.1847	-0.1707	-0.009	0.08785
Car Door* DSM	649	2128	3.279			-0.1590		
Jet Engine* DSM	60	639	10.65			-0.1345		
TV Circuit*	329	1050	6.383	0.018		-0.109		
Tokyo Regional Rail	147	204	2.775	0.3401	-0.8779	-0.0911	0.6820	0.7799
FAA Nav Aids, Unscheduled	2669	7635	5.72			-0.0728		
Mozilla, 19980331* software	811	4077	5.0271	0.0259		-0.0499		
Canton food web*	102	697	6.833	0.157		-0.0694		
Mozilla, all components*	1187	4129	3.4785			-0.0393		
Munich Schnellbahn Rail	50	65	2.6	0.34	-0.8886	-0.0317	0.4870	0.6878
FAA Nav Aids, Scheduled	1787	4444	4.974			-0.0166		
St. Marks* food web	48	221	4.602	0.146		-0.0082		
Western Power Grid	4941	6594	2.6691	0.2022	-0.69	0.0035	0.9	0.795
Unscheduled Air Routes, US	900	5384	11.96			0.0045		
Apache software call list*	62	365	5.88			0.007		
Physics coauthors	145	346	4.7724	0.1517	-0.652	0.0159	0.553	0.6025
Tokyo Regional Rail + Subways	191	300	3.1414	0.4188	-0.8864	0.0425	0.8467	0.8665
Traffic Light controller* (circuit)	133	255	1.9173			0.0614		
Berlin U- & S-Bahn Rail	75	111	2.96	0.48	-0.778	0.0957	0.5051	0.6415
London Underground	92	139	3.02	0.413	-0.9257	0.0997	0.7589	0.8423
Regional Power Grid	1658	2589	3.117	0.1695	-0.5095	0.1108	0.8096	0.6596
Moscow Subways	51	82	3.216	0.1765	-0.9958	0.1846	0.7758	0.8613
Nano-bell (telephone)	104	121	2.327			0.2196		
Broom food web*	82	223	2.623			0.2301		
Company directors	6731	50775	15.09	0.1703	-0.65	0.2386	0.89	0.77
Moscow Subways + Regional Rail	129	204	3	0.4191	-0.8970	0.2601	0.7641	0.8305

Table 1: **Networks Studied and Some of Their Properties, Ordered by Increasing Pearson Degree Correlation r.** Each network is simple, connected, and undirected unless marked *. In the case of the physics coauthors, company directors, and software, only the largest connected component is analyzed. Table omissions correspond to cases where only summary statistics (and not the entire network) were available or where the network was directed (complicating the calculation and interpretation of $\bar{d}$). Social networks were obtained from published articles and data available directly from researchers. Their definitions of node and edge were used. The Santa Fe researchers data were taken from Figure 6 of [7]. Air route and navigational aid data were taken from FAA data bases. Mechanical assemblies were analyzed using drawings or exploded views of products. DSM data were obtained by interviewing participants in design of the respective products. Rail and subway lines were analyzed based on published network maps available in travel guides and web sites. Food web data represent condensation to trophic species. Software call list data were analyzed using standard software analysis tools. The traffic light control circuit is a standard benchmark ISCAS89 circuit. "Nano bell" is a modern competitive local exchange carrier operating in one state with a fiber optic loop network architecture. Its positive value for r reflects this architecture. The regional bell operating company (RBOC) that operates in the same state has a legacy copper wire network that reflects the tree-like architecture of the original AT&T monopoly, and in this state its network's r is -0.6458. This statistic is based on ignoring all links between central offices. Adding 10% more links at random between known central offices brings r up to zero. The RBOC would not divulge information on these links for competitive reasons.

a small value of $\langle d \rangle$. This gives D a rather "peaked" appearance. By comparison, the more elastic D have a more gradually declining degree profile.

The importance of $\bar{d}$ in determining r can be easily seen from Equation 1. Positive r is driven by having many nodes with $d_i > \bar{d}$ that can connect to one another. However, for networks with large $\bar{d}$, there are typically fewer such nodes, and thus many more connections in which $d_i > \bar{d}$ but $d_j < \bar{d}$. The implication is that for highly variable D in which there are only a few dominant high degree nodes larger than $\bar{d}$, then most connections in the network will be of this latter type, and r will likely be negative. This line of reasoning is suggestive but not conclusive, since only an evaluation of Equation 1 can determine the sign or magnitude of r. Yet as a heuristic, it succeeds in distinguishing the rigid D from the elastic D studied here.

The observed values of r, $r_{\max}$, and $r_{\min}$ from Table 1 are plotted in Figure 2. The range $[r_{\min},\ r_{\max}]$ provides the background against which the observed r should be compared, not $[-1, +1]$. When the observed $r < 0$, the whole range is wholly or mostly < 0. When the observed $r > 0$, the whole range approximates $[-1, +1]$. Networks of all types may be seen across the whole range of r in this figure.

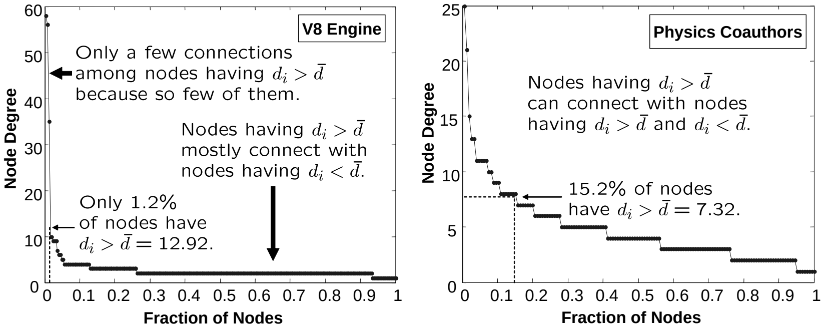


Figure 1: **Degree Profiles of Two Networks in Table 1.** A greater fraction of nodes have $d_i > \bar{d}$ in the physics coauthors (right) than in the V8 engine (left), consistent with increasing elasticity. Abscissa: fraction of all nodes. Ordinate: degree of each node.

3 Domain analysis

The preceeding data and indicators lead us to a striking conclusion: *in some cases whether a network has $r < 0$ or $r > 0$ may be simply a function of network's degree sequence D itself.* For example, if the entire range of allowable r is negative, then no domain-specific "explanation" is required to justify why the network has $r < 0$. Networks with rigid D are obviously more constrained than those with elastic D, and why an individual network gives rise to a particular r-value when the mathematically feasible range is largely unconstrained by its D remains an important question. In such cases, it then makes sense to ask whether the domain-specific features of the system necessarily constrain the network to

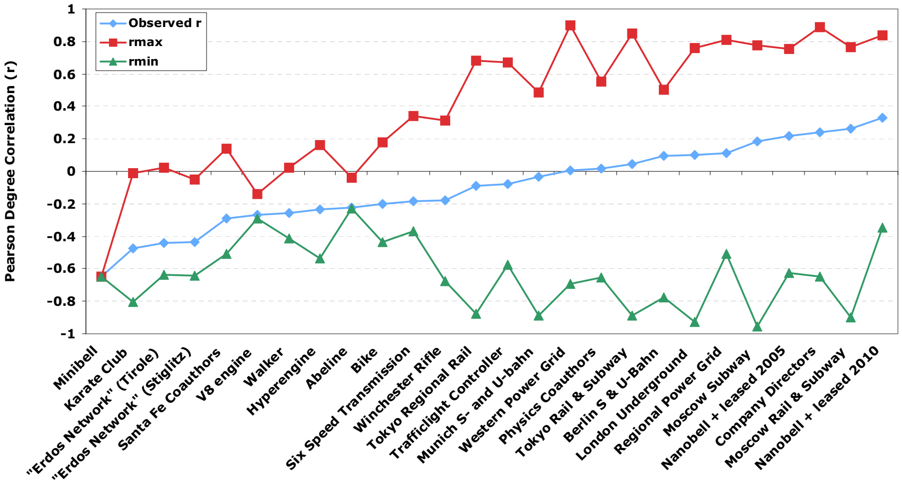


Figure 2: Relationship between r and its range for selected networks.

the observed (or nearby) r-values? Or conversely, do all r-values within the mathematically possible range plausibly correspond to real systems?

For the mechanical assemblies, the answer is that not all values of r within the possible range correspond to functioning systems. The rewired bikes are not different bikes, but in fact meaningless snarls of spokes, pedals, wheels, brake cables, and so on. These networks are not only constrained by rigid D, they are functionally intolerant of the slightest rewiring. But the rewired coauthor networks, even at their extremes of positive and negative r, represent plausible coauthorship scenarios. A negative r scenario could arise in classic German university research institutes, where each institute is headed by a professor whose name is on every paper that the institute publishes. Some of the coauthors go on to head their own institutes and ultimately have many coauthors themselves, while the majority of the others go into industry and publish few papers after graduating. The result is a network with relatively few high-degree nodes connected to many low-degree nodes and only one, if any, connections to other high-degree nodes, leading to negative r. The opposite scenario could be observed at a large research institute devoted to biomedical research, where huge efforts by many investigators are needed to create each publication, and there are often 25 or 30 coauthors on each paper.[3] If such groups produce a series of papers, the result will be a coauthor network with positive r. The fact that coauthorship and other social networks have been found with both positive and negative r shows that such scenarios are plausible.

The same may be said of the Western Power Grid, where the observed connections are no more necessary than many other similar hookups, although the range of plausible connections is narrower than for coauthor networks and wider

[3] For all 55 reports published in Science in the summer and fall of 2005, the average number of authors is 6.9 with a standard deviation of 6.

than for v8 engines or bikes. In [8] it was shown that a communication network with a power law degree sequence could be rewired to have very different r-values and structure, and that the different structures could display very different total bandwidth capacity. While all these networks are plausible, engineering and economic criteria dictate one particular form. Interestingly, this form strongly resembles the planned form of the AT&T long distance network as of 1930 [6].

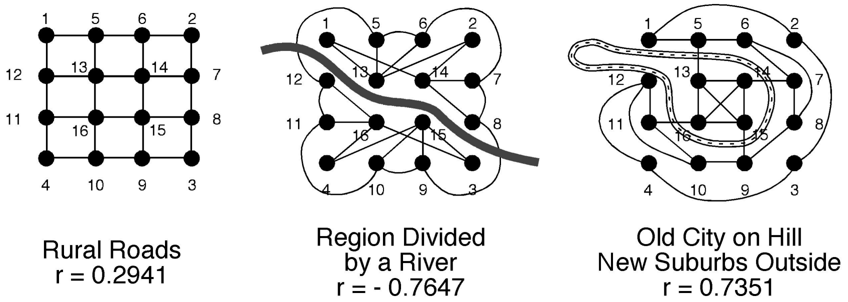
Food webs have been found that exhibit both positive and negative r. This fact reflects different predator-prey patterns in different food webs. Nevertheless, the findings in Table 1 must be viewed with caution since many food webs have been condensed by ecologists. Species that exist in huge variety but are eaten indiscriminately, such as plankton, are often condensed into a single node. "Trophic species" are created by condensing to a single node sets of species that have the same predators and the same prey. This results in different r values.

Large mechanical assemblies like the v8 and the walker have a few high degree nodes because those nodes support the large forces and torques that are typical of the operation of these devices. The six speed transmission and the bike similarly support large forces and torques but have a larger number of load-bearing parts and consequently fewer edges impinging on those parts. Assemblies with rigid parts are severely restricted in allowed magnitude of $\langle d \rangle$ by their need to avoid over-constraint in the kinematic sense. The average of $\langle d \rangle$ for rigid parts assemblies is around 3, and values exceeding 4.2 have not been observed in a set of over 50 assemblies. A mathematical derivation of this restriction appears in [10]. Elastic parts do not impose the severe mechanical constraint on their neighbors that rigid ones do, so the limit on $\langle d \rangle$ is not as severe. The entries in Table 1 bear this out. In the bike, the parts that create elasticity are the spokes while in the transmission they are thin clutch plates. Both kinds of parts appear in large numbers and connect to parts like wheel rims, hubs, and main foundation castings without imposing undue mechanical constraint. For these reasons the bike and the transmission have less peaked D, larger $\bar{d}$, and offer more options for rewiring, thus displaying a wider range of mathematically feasible values for r. Nonetheless, all of these rewirings are implausible and are not observed in practice.

Transportation networks may be tree-like or mesh-like, depending on the constraints and objectives under which they were designed or evolved, as the case may be. It is easy to show that regular trees have negative r while meshes have positive r. Planned urban rail and subway systems increasingly include circle lines surrounding a closely knit mesh, tending to push r toward positive values. If a simple grid is rewired to have respectively minimum and maximum r, we can easily imagine geographic constraints that make the rewired versions plausible, as shown in Figure 3.

4 Conclusions

This paper studied simple connected networks of various types and investigated the extent to which their degree sequences determined the observed value of



Rural Roads
r = 0.2941

Region Divided
by a River
r = - 0.7647

Old City on Hill
New Suburbs Outside
r = 0.7351

Figure 3: **Three Road Systems.** Left: a simple grid, typical of roads in Iowa or Nebraska. Center: the grid rewired to have minimum degree correlation, reflecting roads in a region divided by a large river or mountain range. Right: the grid rewired to have maximum degree correlation, reflecting an old European city as a citadel on high ground surrounded by new suburbs with a new geographically constrained road system.

r or the range of mathematically feasible values of r that they could exhibit. We found that certain characteristics of D, mainly a few dominant high degree nodes, small $\langle d\rangle$, and large $\bar{d}$ relative to $\langle d\rangle$ give rise to observed $r < 0$ and constrained r to a narrow range comprising mostly negative values. It is then of domain interest to understand why a particular network has a degree sequence D with these characteristics. For the rigid assembly networks, this can be traced to the fact that they must support large forces and torques and that they have a few high degree parts that perform this function while supporting the rest of the parts. For rigid social networks like the Karate Club and the Tirole and Stiglitz coauthorship networks, it can be traced to the presence of one or a few dominant individuals who control the relationships represented in the network.

For networks whose D does not have these restrictive characteristics, the observed value of r, while usually > 0 for the systems studied here, may not have direct meaning from either a mathematical point of view (because a wide range of r of both signs is mathematically feasible) or from a domain point of view (because other rewirings with very different r exist or are plausible). Thus, our findings contradict the claim made in [3], namely that "Left to their own devices, we conjecture, networks normally have negative values of r. In order to show a positive value of r, a network must have some specific additional structure that favors assortative mixing." The examples in this paper disaffirm such such generalizations and suggest instead that the observed r for any network should not be compared to $[-1, +1]$ but rather to the allowed range of r for that network, as dictated by its D. Characterizing in greater detail the implications of a system's r-value within that range will be an important topic of future research.

5 Acknowledgments

The authors thank Juyong Park, Jerry Davis, and Jennifer Dunne for sharing and explaining physics coauthor, company director, and food web data, respectively, and Sergei Maslov, Gergana Bounova and Mo-Han Hsieh for sharing and explaining valuable Matlab routines. The authors also thank Jehanzeb Noor, Katherine Steel, and Karen Tapia-Ahumada for data on the regional power grid, Phillipe Bonnefoy and Roland Weibel for data on the US air traffic system, Chintan Vaishnav, Jijun Lin and Daniel Livengood for several telephone networks, Nandan Sudarsanam for the traffic light controller, Carliss Baldwin and John Rusnak for Apache and Mozilla call list data, and Craig Rowles and Eric McGill for DSM data. Finally, the authors thank Chris Magee and Lun Li for valuable discussions.

Bibliography

[1] Newman, M., 2003, The structure and function of complex networks, *SIAM Review* **45**, 167.

[2] Maslov, S. and Sneppen, K., 2002, Specificity and Stability in Topology of Protein Networks, *Science* **296**, 910-913.

[3] Newman, M. E. J., and Park, J., 2003, Why Social Networks are Different from Other Types of Networks, *Physical Review E* **68**, 036122.

[4] Maslov, S., Sneppen, K., and Zalianyzk, A., 2004, Detection of topological patterns in complex networks: correlation profile of the internet, *Physica A* **333**, 529-540.

[5] Eric W. Weisstein. Correlation Coefficient. From *MathWorld–A Wolfram Web Resource.* http://mathworld.wolfram.com/CorrelationCoefficient.html

[6] Fagen, Ed., 1975, A History of Engineering and Science, in *The Bell System, The Early Years (1875-1925)*, Bell Telephone Laboratories.

[7] Girvan, M., and Newman, M. E. J., 2002, Community Structure in Social and Biological Networks, *PNAS* **99**, *12*, 7821-7826.

[8] Li, L., Alderson, D., Doyle, J. C., and Willinger, W., 2006, Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications, *Internet Mathematics* **2**, *4*, 431-523.

[9] Steward, D. V., 1981, *Systems Analysis and Management: Structure, Strategy, and Design*, PBI (New York).

[10] Whitney, D. E., 2004, *Mechanical Assemblies and their Role in Product Development*, Oxford University Press (New York).