A Contrasting Look at Self-Organization in the Internet and Next-Generation Communication Networks

David Alderson, California Institute of Technology
Walter Willinger, AT&T Labs—Research

ABSTRACT

This article examines contrasting notions of self-organization in the Internet and next-generation communication networks, by reviewing in some detail recent evidence regarding several of the more popular attempts to explain prominent features of Internet structure and behavior as "emergent phenomena." In these examples, what might appear to the nonexpert as "emergent self-organization" in the Internet actually results from well conceived (albeit perhaps ad hoc) design, with explanations that are mathematically rigorous, in agreement with engineering reality, and fully consistent with network measurements. These examples serve as concrete starting points from which networking researchers can assess whether or not explanations involving self-organization are relevant or appropriate in the context of next-generation communication networks, while also highlighting the main differences between approaches to self-organization that are rooted in engineering design vs. those inspired by statistical physics.

INTRODUCTION

The rapid development of new networking technologies, ranging from advances in the physical/link layer (e.g., ad hoc wireless networks) to innovations in the application layer (e.g., peer-to-peer), have created a need for more creative methods to analyze, model, and understand large-scale network behavior. In particular, as the engineering complexities and practical importance of these networks continue to grow, there is an increasing need for new methods to assess, control, and manage or mitigate the associated risks. For example, advances in high technology manufacturing have made it much easier to mass produce modular components and interconnect them than to understand how they will behave when so connected. The complexity inherent in today's networks creates significant challenges for their design, control, and operations, and, as we are all too often reminded, this complexity can contribute to fragilities (including the potential for cascading failure events) in these networked systems whose designs are otherwise believed to be extremely robust.

Due to the enormous complexity and potential risks associated with the operation and management of such systems, it is becoming increasingly important that these networks have certain *self-organization* properties — ranging from self-configuration in their startup, to self-adaptation to changes in the operating environment, to self-healing in the presence of component failures or losses — that will minimize the need for human intervention. Such design objectives are fundamental, for example, to the deployment of wireless networks whose operations are based on ad hoc discovery and routing between network nodes. Self-organization is also an increasingly important feature for the wired Internet, particularly in the context of peer-to-peer and other emerging applications, as well as for understanding the spatiotemporal dynamics of existing TCP-type transport and IP-type routing protocols.

At the same time, a striking feature of modern communication networks is the apparent ubiquity of certain "emergent phenomena" — empirical discoveries of the network's large-scale structure or behavior that were not an explicit part of their design and come instead as a complete surprise, defy conventional wisdom, and cannot be explained or predicted within the framework of the traditionally considered mathematical models. For example, the discovery of apparent power law distributions in network connectivity, ranging from the router-level to the autonomous system (AS) level to the Web, cannot be explained by traditional models of random graphs. Similarly, empirically observed self-similar scaling properties in network traffic conflict with the long-held belief that such traffic is Poisson in nature. Such phenomena have attracted attention from researchers across disciplines, and there is now a growing body of complex systems literature that attempts to explain many of these Internet-related "emergent" properties in an elegant and appealingly simple manner, often relying on concepts of statistical

0163-6804/05/$20.00 © 2005 IEEE

physics where notions of emergence and self-organization are closely related.

The purpose of this article is to examine contrasting notions of self-organization in the Internet and next-generation communication networks. To do so, we review in some detail recent evidence regarding several of the more popular claims about the role of self-organization to explain prominent features of Internet structure and behavior. Our results indicate for these examples that what might appear to the non-expert as "emergent self-organization" in the Internet actually results from well conceived (albeit perhaps ad hoc) design, with explanations that are mathematically rigorous, in agreement with engineering reality, and fully consistent with network measurements. Unfortunately, these explanations typically require substantial knowledge about the architecture and operation of the Internet, the design principles underlying its main protocols, and theories of highly regulated, nonlinear, large-scale, far-from-equilibrium dynamic systems that are often unfamiliar to researchers outside the domain. So while the appeal of approaches that attempt to avoid such complexities is understandable, this article sends a cautionary note about the potential success of approaches that ignore them entirely. In particular, we emphasize the distinction between modeling perspectives based on assumptions of *randomness vs. design*, and discuss their implications for Internet modeling and analysis in general. The hope is that a clearer understanding of these distinctions helps to provide a concrete foundation on which members of the networking community can assess whether or not such approaches to self-organization are appropriate to the study of particular network domains.

Self-Organization in Statistical Physics and Engineering

We begin by noting that there is no single mathematically rigorous definition of "self-organization." However, there are distinctly different uses of this term, and in the following we discuss two areas (i.e., statistical physics and engineering) where the differences are particularly striking, with far-reaching implications for the study of large-scale complex network structure and behavior.

Power Laws and Scaling in Complex Systems

Within the physics community, most attempts at formally defining self-organization involve the concept of *self-organized criticality* (SOC) introduced by Bak [1] or its close relative, the *edge-of-chaos* (EOC) approach by Kauffman [2]. These concepts leverage the tools of statistical mechanics — a very successful and highly influential physical theory of the last century — and are rooted in an assumption that the large-scale structure and behavior of complex systems can be understood in terms of random ensembles and their statistical properties. System design and functionality play a secondary role in such formulations, except perhaps in the form of constraints on macroscopic statistics, and the emphasis is on "likely" configurations that are otherwise governed by randomness. The focal point then becomes explaining "emergent" behavior in the complex structure of the underlying system. In particular, SOC and EOC have their origins in attempts to explain the widespread appearance of emergent phenomena that reveal themselves in the form of scaling phenomena and power-law-like statistics of characteristic events observed in numerous geophysical, astrophysical, biological, engineered, economic, and cultural systems.

The intuition behind SOC is that large-scale highly interactive dissipative systems will drive themselves (i.e., "self-organize") to a critical state where characteristic events follow a power law distribution, and once in this state, the interactions between the individual component parts of the system induce a kind of global organization. Since this emergence of scale-free structures is generally viewed as a hallmark of systems at a critical point of a continuous phase transition, SOC is reminiscent of scale invariance in statistical mechanics models at criticality, except that there is no need for tuning some parameter to a unique value to achieve criticality; rather, order "emerges" from chaos. Similar in spirit, the main idea behind EOC is that there is typically a single parameter, or density, that describes the otherwise generic and random underlying system, and the point of greatest complexity and thus interest lies near a bifurcation point in this parameter.

Following in the footsteps of SOC and EOC, *scale-free networks* (SFNs) [3, 4] have been proposed as a framework for understanding, interpreting, and explaining power laws in network connectivity. The argument is that *"for many complex systems, we have to first understand the topology that describes how the diverse constituents interact with each other. This is fundamentally a physics problem, since it involves randomness and self-organization living side by side, and that is best addressed by the tools of statistical mechanics."*[1] SFN models typically use random, evolutionary network growth via preferential attachment (i.e., newly arriving nodes are more likely to connect with existing nodes that already have many nodes) in order to replicate the statistical power law connectivity signatures observed in many prominent technological, biological, or social networks. Because of the apparent ubiquity of power law degree distributions throughout natural and manmade networks, and the inability of traditional random graph models to explain them, SFNs have been promoted as a universal approach to understanding the self-organizing nature of complex networks, including the Internet [3, 4]. SOC, EOC, and most recently SFNs have dominated much of the scientific literature concerning "complex systems," and what makes these concepts so attractive to physicists is the fact that they deal with systems that tend spontaneously to reach a critical state, often characterized by a power law, and with minimal or no fine tuning of some control parameter. When applied to engineered systems, an alluring feature of these approaches is that they appear to be simple and generic; that is, randomness is their

Following in the footsteps of SOC and EOC, scale-free networks have been proposed as a framework for understanding, interpreting, and explaining power laws in network connectivity.

[1] *Quote by A.-L. Barabási, as reported in* PhysicsWeb*, 26 July 2000; http://physicsweb.org/articles/news/4/7/10/1.*

Most engineering-based approaches to the study of self-organization assume that the system of interest has underlying degrees of freedom that can be used to achieve a desired level of performance or automation.

main driver, not system-specific details or domain knowledge. However, their main appeal to physicists and engineers alike is that it combines two fascinating concepts — self-organization and critical behavior — in an attempt to explain a third, equally fascinating notion: complexity. While it is certainly fair to say that the Internet is teeming with complexity and emergence in the sense of exhibiting unintended power law distributions, self-similarity, and fractal-type behavior, a more pressing question here is whether or not this statistical physics perspective of complexity is capable of providing a sound and verifiable foundation for a theory of next-generation networks that allows for a systematic and integrated treatment of both its engineered structure and observed behavior.

Engineering Design and Performance in Complex Systems

Most engineering-based approaches to the study of self-organization assume that the system of interest has underlying degrees of freedom that can be used to achieve a desired level of performance or automation. Thus, in contrast to the physics-based approach outlined above, where complexity (in the form of power laws, scaling, or fractals) is explained by simple underlying forces (e.g., randomness, preferential attachment), the engineering approach seeks to create simplicity in function or performance through the use of (often hidden) underlying complexity. This approach requires incorporating knowledge of the system's functional objectives, the details of its component parts, and the specifics of its operating environment to yield descriptions that explain the observed structure or behavior but are also fully consistent with engineering reality and available measurements.

In the context of the Internet, most attempts at understanding large-scale structure and behavior have primarily focused on the interaction of specialized, decentralized, asynchronous components instead of searching for features that "emerge out of randomness." That is, while random models are common in engineering-based approaches, they are not required, and incorporating randomness into a model often serves the sole purpose of accounting for uncertainty (e.g., environmental or operational) that needs to be managed. The resulting uncertainty models are typically mixed with hard system-specific constraints, and represent either an ensemble of network designs or a single robust design, depending on the underlying design objective, but all designs are highly constrained and "hand-crafted" (i.e., extremely rare from a traditional random network model perspective) to achieve a certain level of performance. Thus, while randomness is a prominent feature of many engineering models, it is typically less important than system-specific aspects such as performance, functionality, resource constraints, or design trade-offs inherent in the engineering process.

One approach to exploring the structure and behavior of such highly-engineered, complex systems is based on *highly or heuristically optimized tolerance or trade-offs (HOT)*, a conceptual framework in support of an engineering mindset that attempts to uncover which type of (possibly implicit) design process is responsible for the prevalence of such "emergent" phenomena [5]. The HOT approach has shown that the introduction of even minimally realistic trade-offs yields models and outcomes that are dramatically different from those based on assumptions of randomness. As discussed below, simple toy models using the HOT framework have demonstrated how engineering design easily generates highly variable (power law) event sizes once functional performance and robustness trade-offs are considered [5, 6]. When modeling Internet structure or behavior, the overriding concern of a HOT-based approach is ultimately not to generate or reproduce power laws or self-similar scaling per se, but to understand the functional objectives implicit in Internet design and, in the context of observed high variability in Internet connectivity and traffic, to identify the main mechanisms underlying their prevalence. In fact, given the many ways that power law relationships naturally arise [7], it becomes clear that an ability to generate them does not "explain" anything and should not be counted as evidence for a proposed model or theory. Additional support for this perspective comes from arguments originally due to Mandelbrot, who observed that since power law distributions enjoy certain invariance properties (e.g., to marginalization, mixtures, maximization), once high variability appears in real data, power law relationships become a natural outcome of the processes that measure them (see [8] for additional details). These arguments serve as a cautious reminder that to understand power laws or scaling behavior, there is no inherent need for special explanations or models like SOC/EOC/SFNs, and HOT becomes just one of many possible frameworks that try to capture this difference in perspective.

The Internet as a Case Study

The Internet is a particularly attractive case study in self-organization, since a detailed understanding of the underlying technology together with the ability to do detailed measurements means that any conjectures about "emergent" properties can be unambiguously resolved. Here, we consider several examples where the popular view of self-organization from statistical physics leads to direct contradictions with the engineering perspective on network structure and behavior.

Internet Traffic and Self-Similarity

The empirical discovery that measured traffic rates on links in the Internet exhibit *self-similar* features [9] has continued to fascinate researchers across different disciplines. Self-similar Internet traffic constitutes a prototypical "emergent" phenomenon because its ubiquity is both unintended and surprising, thus inviting a spectrum of potential explanations. These range from being a consequence of dynamic instabilities or bifurcations, where the details of the system's design, architecture, and components are

largely irrelevant, all the way to arising naturally within the confines of the Internet's hourglass protocol architecture, where it can be understood in terms of underlying architectural guidelines, design principles, and applications. While the former is straightforward statistical physics and implies that the observed self-similar nature of Internet traffic is a signature of chaotic dynamics or criticality (e.g., [10, 11]), the latter engineering-based explanation identifies application-layer traffic properties (i.e., heavy-tailed characteristic of file or Web document sizes) as largely responsible for the self-similar scaling behavior of aggregate traffic at the IP layer over sufficiently large timescales. In fact, Mandelbrot's renewal reward processes (or their close relatives, Cox's immigration-birth models) and their limiting regimes can be viewed as highly simplified formulations of traffic self-similarity arising within an appropriate HOT framework, where the sizes of transferred files represent the main source of environmental uncertainty, and the limit regimes capture first-order effects due to constraints imposed by TCP/IP and the overall network structure (for details, see [9, references therein]).

As detailed in [12], under scrutiny with measured traffic and by exploiting their rich semantic context to associate individual packets with meaningful higher-layer networking quantities such as IP flows, TCP connections, or even sessions where possible, empirical evidence and networking reality have consistently favored the engineering explanation of self-similarity over its statistical physics counterparts. An important lesson from these opposite explanations is that because the Internet, like many advanced technologies, hides great internal complexity in exchange for ease of use, it invites appealingly simple explanations that — while they may appear to capture its "essence" without the burden of details concerning its protocols or designed nature — collapse quickly under scrutiny with real data and are easily refuted by applying varying amounts of domain knowledge. While one can certainly imagine that the sort of simple (but inefficient) systems advocated by a statistical physics view of networking [10, 11] could be built and made to operate in a chaotic, bifurcating, or critical state, it is crucial to realize that the actual Internet, with its routers, hosts, protocols, and user behavior, is far more than such a simple construct.

Internet Topology and Power Law Distributions

Recent attempts to characterize the structure of the Internet as well as other complex networks have focused on cataloging certain statistical properties and then investigating mechanisms that might yield them. One feature of Internet topology that has received considerable attention is the distribution of node degree (i.e., connectivity), whether or not it follows a power law, and what if anything power law node degree distributions imply about the "robust yet fragile" nature of the Internet. As noted above, the prevalence of power law statistics in the connectivity of the Internet has made SFNs a popular theory — one that suggests that simple underlying mechanisms in the evolution of networks cause them to self-organize into structures exhibiting high variability in connectivity. In addition to their signature power laws in node degree distribution[2] and evolutionary growth via preferential attachment, scale-free graphs are commonly associated with the presence of highly connected, central *hub* nodes that are critical to the overall connectivity of the network. As noted in [4], *"Networks containing such important nodes, or hubs, tend to be what we call 'scale-free,' in the sense that some hubs have a seemingly unlimited number of links and no node is typical of the others. These networks also behave in certain predictable ways; for example, they are remarkably resistant to accidental failures but extremely vulnerable to coordinated attacks."* That is, the network is claimed to have a few crucial hub routers, through which most traffic must pass and which hold the network together — giving it "error tolerance" to random node failures since most nodes have low connectivity (i.e. are non-hubs), but also "attack vulnerability" to efforts that target these hubs. Thus, when applied in the context of the Internet's router-level topology, claims of a scale-free structure paint a picture of a previously overlooked Achilles' heel and suggest a defensive strategy whereby resources should be devoted to protecting the few hubs. In fact, the hub-like structure of scale-free graphs is such that the epidemic threshold is zero for contagion phenomena [14], thus suggesting that the natural way to stop computer viruses/worms is also to protect these hubs.

Under scrutiny with connectivity data from actual networks and relying on available information about existing router technologies, recent work applying the HOT framework to the wired Internet [6] has shown that considerable insight into router-level topology can be achieved from thinking about designs that allow the network to effectively carry a projected overall traffic demand, subject to trade-offs that have to be made between what is technologically feasible vs. economically sensible. That is, the HOT formulation for topology design involves optimizing functional objectives (e.g., performance), subject to constraints on their components (e.g., routers), with an explicit source of uncertainty (e.g., traffic demands) against which solutions must be robust. More specifically, assume that two-way network traffic is exchanged between all pairs (i, j) of end nodes i and j, the flow W_{ij} of traffic between i and j is given by $W_{ij} = \rho w_i w_j$, where ρ is some global constant, every end node i has a total bandwidth demand w_i, and flows are otherwise uncorrelated from one another. Under this "gravity model," the performance measure for a given network is then its maximum throughput, computed as

$$\begin{aligned} \max_{\rho} \quad & \sum_{ij} W_{ij} \\ \text{subject to} \quad & QW \le b, \end{aligned} \tag{1}$$

where Q is the routing matrix obtained using standard shortest path routing. $Q = [Q_{kl}]$, with $Q_{kl} = 1$ if flow l passes through router k, and Q_{kl}

The prevalence of power law statistics in the connectivity of the Internet has made scale-free networks a popular theory — one that suggests that simple underlying mechanisms in the evolution of networks cause them to self-organize into structures exhibiting high variability in connectivity.

[2] *The definition of a scale-free graph has never been made mathematically precise (see [13] for details). Although it can easily be shown to result in contradictions, the presence of power law node degrees is often adopted in the literature as a de facto definition of the term* scale-free.

The simplistic view to network modeling conflicts directly with the Internet's legendary robustness to router failure and completely misses the most obvious attack vulnerability: that the very same robustness can be essentially hijacked by malicious end users.

= 0 otherwise. W is the vector of all flows W_{ij}, indexed to match the routing matrix Q, and b is a vector consisting of all router bandwidth capacities (see [6] for details). Such optimization-based formulations suggest that a reasonably "good" design for router-level Internet connectivity is one in which the core network is constructed as a sparse mesh of high-speed low-connectivity routers that carry heavily aggregated traffic over high-bandwidth links. Accordingly, this mesh-like core is supported by a hierarchical tree-like structure at the edges whose purpose is to aggregate traffic. Moreover, this result suggests that any sensible network design process with minimally realistic assumptions will produce something qualitatively similar. On a practical level, the hub-like structure inherent in scale-free models of the Internet's router-level topology violates what can reasonably be constructed from existing router technologies and, if built, would result in unacceptable performance and/or exorbitant cost. Indeed, because router technology is fundamentally limited in the number of data packets that can be processed in any unit of time, and because there is a need for high bandwidth connections in backbone routers, the number of connections for such routers must be relatively low. Figure 1 highlights the extreme differences between a scale-free type of network design based on preferential attachment and a designed network resulting from the HOT framework. Like the real Internet, the HOT network exhibits high connectivity only at the network edge for purposes of traffic aggregation, and economic drivers of customer willingness to pay for bandwidth in combination with highly variable population density become the clear drivers of power-law-type connectivity at the network edge. However, the loss of such hub nodes at the network's periphery results in only local disruptions to connectivity and does not result in the type of Achilles' heel inherent in SFN models. Perhaps more important, the simplistic view of network modeling in which connectivity alone is sufficient to describe the "robust yet fragile" features of the Internet creates a picture that is, at best, misleading. It conflicts directly with the Internet's legendary robustness to router failure [15] and completely misses out on the most obvious attack vulnerability: that the very same robustness can be essentially hijacked by malicious end users to launch distributed denial of service attacks or network worms, directed against other end users or possibly against the physical infrastructure as a whole.

The Internet's Large-Scale Behavior at the TCP/IP Layer

To a first approximation, the Internet's router-level topology represents the interface where the traffic demand generated at the application layer meets the raw bandwidth provided by the physical network infrastructure. IP routing and TCP ensure that, subject to a number of technological and economic-related constraints, most of this raw bandwidth is delivered in an effective fashion to the application layer. This picture sharpens by expanding on the HOT theme to consider a formulation that offers a rich conceptual framework for addressing the Internet's ability to self-organize in the engineering sense, achieving efficiency by solving implicitly given global utility maximization problems across the Internet in a completely decentralized and fully distributed fashion.

To illustrate, consider a physical network infrastructure modeled as a set of L links with finite capacities $c = (c_l, l \in L)$. They are shared by a set of N sources. Each source i uses a set $L_i \subseteq L$ of links. The sets L_i define an $L \times N$ routing matrix $R_{li} = 1$ if $l \in L_i$, and 0 otherwise. Each source i transmits at rate $x_i(t)$. These transmission rates determine the aggregate flow $y_l(t)$ at each link, $y_l(t) = \Sigma_i R_{li} x_i(t - \tau_{li}^f)$, where τ_{li}^f denotes the forward transmission delays from sources to links. Each link l maintains a *congestion measure* $p_l(t)$, termed *price*, that has different interpretations in different versions of TCP. A source i has access only to the *aggregate* price $q_i(t)$ in its route, $q_i(t) = \Sigma_l R_{li} p_l(t - \tau_{li}^b)$, where τ_{li}^b denotes the backward delays in the feedback path. Decentralization requires that source rates $x_i(t)$ be adjusted based only on aggregate prices $q_i(t)$, and prices $p_l(t)$ be adjusted based only on aggregate rates $y_l(t)$. This can be represented as $\dot{x}_i = F_i(x_i(t), q_i(t))$ and $\dot{p}_l = G_l(p_l(t), y_l(t))$, where different TCP protocols are modeled as different F_i, and different active queue management (AQM) schemes are modeled as different G_l. To understand the equilibrium of the network modeled this way, it is useful to associate a utility function $U_i(x_i)$ to each source i and consider the problem of maximizing aggregate utility subject to capacity constraints:

$$\max_{x_i \geq 0} \sum_i U_i(x_i), \qquad \text{subject to} \quad Rx \leq c \tag{2}$$

and its dual:

$$\min_{p \geq 0} \max_{x_i \geq 0} \sum (U_i(x_i) o - x_i \sum_l R_{li} p_l) + \sum_l p_l c_l. \tag{3}$$

The key idea in this duality model (see [16, references therein]) is to *interpret source rates as primal variables, prices as dual variables, and congestion control as a distributed primal-dual algorithm over the Internet to solve Eqs. 2 and 3. Different TCP/AQM protocols all solve the same prototypical constrained nonlinear program, but they use different utility functions and implement different iterative rules* (F_i, G_l) *to optimize them.* This approach combines with tools from control theory to provide rigorous proofs of the global efficiency, dynamics, and stability of these networking protocols, as well as their robustness to arbitrary network connectivity and transport delays (see [17, references therein]). Thus, what might appear to the nonexpert as "emergent self-organization" — a large-scale network running close to criticality where it achieves maximum information transfer and efficiency — actually results from the provable robustness of a highly organized design of decentralized and asynchronous protocols. Again, the substantial amount of domain-specific knowledge required for such explanations often makes them inaccessible or unattractive to nonspecialists, thus increasing the appeal of approaches that claim to avoid such complexities.

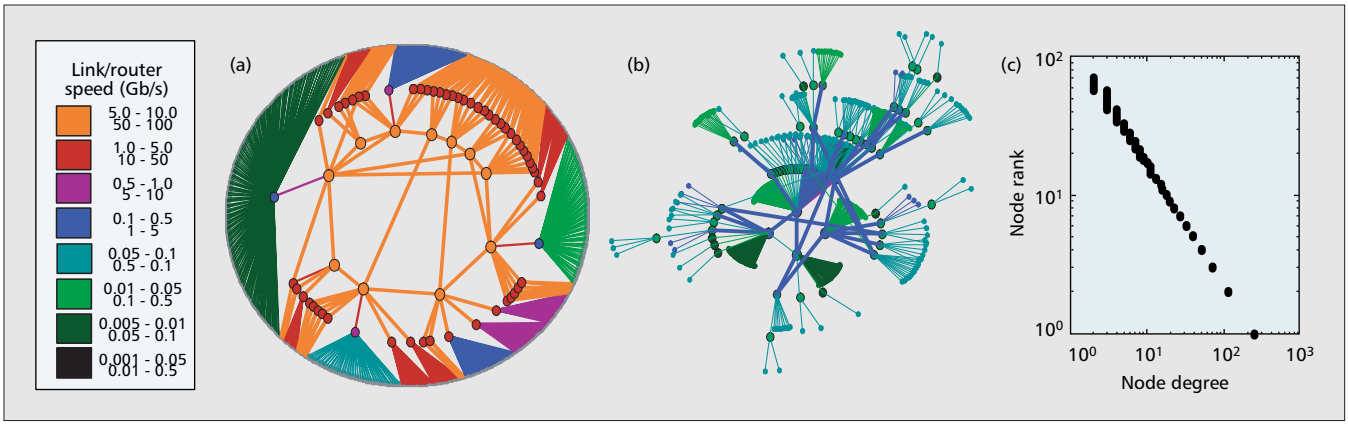

Figure 1. *The toy network in a) is the result of heuristic design based on the HOT framework, while the toy network in b) is the result of a preferential attachment construction consistent with scale-free networks. Both networks have the exact same node degree distribution, c). Links and nodes are color-coded to indicate their utilization when carrying the maximum flow defined by Eq. 2. The HOT network has a total throughput of 5.76×10^{11} b/s, while the SFN graph has a total throughput of 5.77×10^{9} b/s. The differences in network size notwithstanding, the examination of real networks and a need for high throughput suggest that the real router-level Internet is nothing like b) but is qualitatively more like a).*

Looking Ahead: A Cautionary Note

The differences in perspective described in this article emphasize the importance of examining the assumptions of a candidate modeling framework against the details of a particular application domain. In particular, is the premise of statistical mechanics compatible with the Internet? On one hand, the benefits of describing atoms or molecules in terms of random ensembles and analyzing them using techniques from statistical physics are undeniable. On the other hand, while the ever increasing number of network components makes modeling and analyzing large-scale networks such as the Internet a daunting task, under what circumstances is it reasonable to treat routers, hosts, or end users as generic "particles" that interact in a largely random fashion? It is certainly interesting and educational to know that specific Internet-related power law statistics and scaling phenomena could possibly arise as signatures of self-organized emergent phenomena, but many of the proposed explanations and models that are rooted in statistical physics and ignore engineering details have simply not held up under increasing scrutiny within the networking community. At the same time, alternate modeling frameworks like HOT have provided only a nascent understanding of the way in which trade-offs between performance and constraints lead to complex structure and behavior, and much work remains. And it is only appropriate to question similarly the presumption of alternative modeling approaches like HOT, namely that these complex (networking) systems are highly evolved — in the sense that these systems are fundamentally driven by an iterative process through which "good" designs are reused and improved on while "poor" designs are discarded — and that an optimization-based framework can appropriately capture them. Meanwhile, new paradigms arising in the areas of ad hoc wireless or sensor networks are challenging the existing framework for network design, and the increasing frequency with which power law and scaling phenomena are encountered in those areas, together with a natural desire for simple and universal explanations, support ongoing popularity of a statistical physics perspective that tends to associate such phenomena with critical phase transitions [14] or chaotic dynamics near a bifurcation point [18].

A critical question for network engineers and architects is: *What do these different modeling approaches have to say about the design, structure, and behavior of next-generation networks?* While there is an obvious appeal to the notion that one can ignore evolution, design, functionalities, and constraints — all the ingredients that make engineering different from physics — it remains to be seen how an appropriate use of emergence, self-organization, criticality, and statistical properties of random ensembles in the context of networking technology can produce systems that meet the performance, reliability, and predictability requirements of such systems. If recent experience with the wired Internet is an indication, network self-organization in the form of management simplicity will be a critical objective, but will likely be the result of deliberate and well-designed protocols rather than a feature that emerges out of randomness. In the meantime, the perspective advocated here is that tremendous insight in the modeling and analysis of self-organizing networks is available through a distinction between systems that are assumed to be inherently random from those that are explicitly or implicitly designed. And if sensor networks, where different applications will dictate different designs, constraints, and functionalities, are an indication, such a distinction will likely be highly application-specific, arguing for careful assessment of the suitability of as wide a range of models or theories as possible, with the SOC/EOC/SFN and HOT models considered here representing but two concrete (albeit contrasting) examples.

Acknowledgments

The authors are indebted to John Doyle for ongoing conversations regarding the HOT framework, to Lun Li for her insight and collab-

If recent experience with the wired Internet is an indication, network self-organization in the form of management simplicity will be a critical objective, but it will likely be the result of deliberate and well-designed protocols rather than a feature that emerges out of randomness.

oration on the modeling of Internet topology, and to Steven Low for helpful discussions about the primal-dual nature of TCP/AQM. Finally, we would also like to thank the reviewers for their constructive criticism and valuable feedback.

REFERENCES

[1] P. Bak, *How Nature Works: the Science of Self-Organized Criticality*, Copernicus 1996.
[2] S. Kauffman, *The Origins of Order*, Oxford Univ. Press, 1993.
[3] R, Albert and A-L, Barabási, "Statistical Mechanics of Complex Networks," *Rev. Modern Physics*, vol. 74, Jan. 2002.
[4] A.-L. Barabási and E. Bonabeau, "Scale-Free Networks," *Sci. Amer.*, vol. 288, 2003, pp. 60–69.
[5] J. M. Carlson and J. Doyle, "Complexity and Robustness," *Proc. Nat'l. Acad. Sci. USA*, vol. 99, suppl. 1, 2002, pp. 2539–45.
[6] L. Li *et al.*, "A First-Principles Approach to Understanding the Internet's Router-Level Topology," *Proc. ACM SIGCOMM 2004*, *Comp. Comm. Rev.*, vol. 34, 2004, pp. 3–14.
[7] M. E. J. Newman, "Power Laws, Pareto Distributions and Zipfs Law," *Contemporary Physics*, in press; available electronically: arXiv:cond-mat/0412004, Jan. 9, 2005.
[8] W. Willinger *et al.*, "A Pragmatic Approach to Dealing with High Variability in Network Measurements," *Proc. ACM SIGCOMM Internet Measurement Conf. '04*, 2004.
[9] K. Park and W. Willinger, Eds., "Self-Similar Network Traffic: An Overview," *Self-Similar Network Traffic and Performance Evaluation*, Wiley, 2000.
[10] R. Solé and S. Valverde, "Information Transfer and Phase Transitions in a Model of Internet Traffic," *Physica A*, vol. 289, 2001, pp. 595–695.
[11] A. Veres and M. Boda, "The Chaotic Nature of TCP Congestion Control," *Proc, IEEE INFOCOM 2000*, 2000.
[12] W. Willinger *et al.*, "Scaling Phenomena in the Internet: Critically Examining Criticality," *Proc. Nat. Acad. Sci. USA*, vol. 99, suppl. 1, 2002, pp, 2573–80.
[13] L. Li *et al.*, "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications," to appear, *Internet Mathematics*.
[14] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, Cambridge Univ. Press,2004.
[15] D. D. Clark, "The Design Philosophy of the DARPA Internet Protocols," *Proc. ACM SIGCOMM '88, ACM Comp. Commun. Rev.*, vol. 18, no. 4, 1988, pp. 106–14.
[16] S. H, Low, "A Duality Model of TCP and Queue Management Algorithms," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, Aug. 2003, pp. 525–36.
[17] A. Papachristodoulou, L. Li, and J. C. Doyle, "Methodological Frameworks for Largescale Network Analysis and Design," *Comp. Comm. Review*, vol. 34, no. 3, July 2004, pp. 7–20.
[18] L. Kocarev and G. Vattay, Eds., *Complex Dynamics in Communication Networks*, Springer-Verlag Series in Understanding Complex Systems.

BIOGRAPHIES

DAVID ALDERSON [M] (alderd@cds.caltech.edu) is currently a postdoctoral scholar in the Division of Engineering and Applied Science at the California Institute of Technology (Caltech), Pasadena. He received a B.S.E. in civil engineering and operations research from Princeton University, and M.S. and Ph.D. degrees from the Department of Management Science and Engineering at Stanford University. He is a member of INFORMS.

WALTER WILLINGER [F] (walter@research.att.com) is currently a member of the Information and Software Systems Research Center at AT&T Labs-Research, Florham Park, New Jersey. He received the Diplom (Dipl.Math.) from ETH Zurich, Switzerland, and M.S, and Ph.D. degrees from the School of ORIE, Cornell University, Ithaca, New York. He is co-recipient of the 1996 IEEE W. R. G. Baker Prize Award and the 1994 W. R. Bennett Prize Paper Award. He is a member of ACM, SIAM, and INFORMS.