Editorial: The Future of Power Law Research

Michael Mitzenmacher*

Abstract

I argue that power law research must move from focusing on observation, interpretation, and modeling of power law behavior to instead consider the challenging problems of validation of models and control of systems.

1 The Problem with Power Law Research

To begin, I would like to recall a humorous insight from the paper of Fabrikant, Koutsoupias, and Papadimitriou [8], consisting of this quote and the following footnote.

"Power laws... have been termed 'the signature of human activity'...[1]"

"[1] They are certainly the product of one particular kind of human activity: looking for power laws..."

The study of power laws, especially in networks, has clearly exploded over the last decade, with seemingly innumerable papers and even popular books, such as Barabási's *Linked* [1] and Watts' *Six Degrees* [20]. Power laws are, indeed, everyhwere. Despite this remarkable success, I believe that research into power laws in computer networks (and networks more generally) suffers from glaring deficiencies that need to be addressed by the community. Coping with these deficiencies should lead to another great burst of exciting and compelling research.

To explain the problem, I would like to make an analogy to the area of string theory. String theory is incredibly rich and beautiful mathematically, with a simple and compelling basic starting assumption: the universe's building blocks do not really correspond to (zero-dimensional) points, but to small (one-dimensional) strings. Many different versions and variations of string theory exist. String theory is an incredibly popular area of physics, both with researchers and with the general populace. The mathematical richness, simple starting points, large number of variations, and popularity are all characteristics also shared by current research in power laws.

There is, however, an unpleasant problem with string theory: it has not been experimentally validated, and it is not currently clear what will be required to validate it. There has not been an actual experiment that shows that string theory successfully predicts something that is not also explained by other existing theories, and current suggestions for potential experiments to pass this hurdle apeear well outside the realm of what is possible in the foreseeable future. In an era where evolution and intelligent design have set the stage for the question of "What is science?", the issue that there is apparently no adequate way to verify or falsify string theory is rightly seen as a major flaw. This problem may be corrected in time, and I do not wish to suggest that string theory is not a valuable scientific contribution. The current state, however, creates a litany of challenging questions about the fruitfulness of this line of research.

I would argue that we are now facing a similar problem in research on power laws in computer networks, and this signals that we must actively change the research agenda. Specifically, while numerous models that yield power law behavior have been suggested, and in fact the number of such models continues to grow rapidly, no general mechanisms or approaches have been suggested that allow one to validate that a suggested model is appropriate. Like the string theorists, we have

*Harvard University, Div. of Engineering and Applied Sciences. Supported in part by NSF grant CCR-0121154. E-mail: michaelm@eecs.harvard.edu.

beautiful frameworks, theory, and models – indeed, we have perhaps far too many models – but we have been hesitant in moving to the next steps, which could transform this promising beginning into a truly rich new area of science.

To make my case, let me introduce a rough taxonomy of the types of results one could aim for in studying power laws.

(a) *Observe*: Gather data on the behavior of a system and demonstrate that a power law distribution appears to fit the relevant data.

(b) *Interpret*: Explain the significance of the power law behavior to the system.

(c) *Model:* Propose an underlying model that explains the power law behavior.

(d) *Validate:* Find data to validate, and if necessary specialize or modify, the model.

(e) *Control:* Use the understanding from the model to control, modify, and improve the sysytem behavior.

My argument is that most research on power laws has focused on observing, interpreting, and modeling, with a current emphasis on modeling. As a community, we have done almost nothing on validation and control, and we must actively move towards this kind of research. Unlike string theory, where validation and control are hampered by nature and constrained by what is physically possible, I believe that in networks we can, in most cases, successfully validate our models and use them to control systems.

As an example, consider a power law that has been the subject of significant research: the distribution of inlinks and outlinks on the directed graph corresponding to the pages of the World Wide Web (see, e.g., [3, 4, 14]). Early work on the web graph discovered these distributions followed a power law, and the significance for applications was clear. Indegree was used early on by search engines such as Altavista to rank web pages, and the power law distribution ensured that this approach was reasonably useful, as only a small number of pages were likely to have a significantly larger number of inlinks.

A significant amount of historical and recent theoretical work has led to models for why power law behavior occurs, as detailed in various surveys [18, 19]. For the web graph, the primary model of study has been variations of preferential attachment: the more links a web page has, the more links it is likely to obtain in the future. But even for this specific problem, there has been little or no systemic or theoretical work covering validation or control. Validation is important, because there are numerous models that yield power laws, and in many cases more than one of these models can reasonably be used to explain an observed power law behavior. For example, models other than preferential attachment have been used to describe link behavior, and some evidence has been given that preferential attachment is not a suitable explanation (see, e.g., [13]). On the issue of control, it now seems clear that the presence of search engines (especially Google) has affected how people link to web pages. There is less need for people to explicitly put in links in their web pages when the links are easily found on Google; however, if people reduce how often they link to pages, this might affect the performance of Google's algorithms. Could this impact and its effects have been modeled appropriately ahead of time? (A model for this situation was examined in [6].)

This path of observation, interpretation, and modeling has been repeated in many domains. I would argue, however, that without validating a model, it is not clear that one understands the underlying behavior, and therefore how the behavior might change over time. It is not enough to plot data and demonstrate a power law, allowing one to say things about current behavior; one wants to ensure that one can accurately predict *future* behavior appropriately, and that requires understanding the correct underlying model.

For example, barring major new changes, will the degree distribution of web pages still look the same two years from now? Recent work suggests the density of edges in various power law graphs, such as citation graphs and the Autonomous Systems graph, may be increasing over time [16]. If the same holds in the graph of Web pages, the answer may be no. We do not at this point know enough about Web graph to know. Further, from the practical point of view, one would like to take the model and use it to control the system. Suppose Google could somehow encourage people to create more high quality links, and it wanted to determine if it could do so in a way that was worth the cost? This would require not only validating the underlying model, but understanding how the model would respond to control mechanisms, such as advertising, contests, or direct payments, that might increase links.

2 New Directions

I can happily say that I believe the community is currently moving toward validation. A key step in this direction, perhaps between modeling and validation, is work on *invalidation*. Invalidation can take multiple forms, including showing that a model is insufficient to explain a power law by providing a reasonable or more compelling alternative, or considering additional properties beyond the power law behavior suggested by a proposed model and showing these properties do not fit the data. For example, an insightful recent paper by Lakhina, Byers, Crovella, and Xie [15] questions the methodology of using *traceroute*-based maps (as in, for example the highly cited [9]) of the Internet to conclude that the degree distribution of the Internet-connectivity graphs follows a power law. With traceroute-based maps, one attempts to obtain a snapshot of the edges in the network by tracing the route of packets from one or more specialized sources and various destinations in the network. Unfortunately, as shown in [15], this leads to a very biased sample of the edges in the network; they show that under this methodology the edges found even from random regular graphs would yield a subgraph that also exhibits power law behavior. This is because there is an inherent correlation between the observed degree of a vertex in the resulting subgraph and the proximity of the source of the traceroute queries. This paper does not conclude that the Internet does not have such a degree distribution, but does provide reason to believe the methodology needs to be re-examined. Similarly, the paper by Chen et al. [7] considers several different characteristics of the router graph, and finds that models based on preferential attachment do not have the appropriate properties that this graph seems to have. It then proposes a new model which appears to have these additional properties.

The article in this month's journal, by Li, Alderson, Doyle, and Willinger [17], is in part a paper on invalidation, although it is also a great deal more. A major part of the paper is to clarify that there are many different types of graphs that yield power law distributions on the degree sequences, and these graphs may appear extremely different with respect to other properties. They argue that many people have applied certain types of graphs, such as graphs created preferential attachment, to describe real objects, such as the Internet connectivity graph, based on the fact that both have a power law degree distribution. They show that this characterization does not match the real world by an examination of the data. Beyond such invalidation arguments, they also provide a large framework and foundations that will certainly play a role in future invalidation and validation arguments. In particular, they attempt to give a simple but useful way of distinguishing various types of graphs with heavy-tailed indegree and/or outdegree distributions, and to build a proper vocabulary for making appropriate distinctions. I believe this paper should be and will be read by everyone working in the area. It will provide a healthy jumping-off point for a richer, more refined view of the various power law models and their relation to real structures.

This ad hoc approach based on invalidation, whereby we repeatedly find alternatives or flaws in existing models and adapt accordingly, has already been useful in developing models with appropriate properties for specific experiments and in bringing important variations to light. It is, however, incomplete and unsatisfying both theoretically and practically. It is never clear that one has captured all relevant statistics appropriately. Invalidation leads us naturally to consider whether validation is possible from the outset.

What other approaches to validation are there? A compelling possibility is to make use of *time series analysis*, with the goal that one observes the system over time to judge the underlying assumptions of the proposed model. That is, instead of using the model to generate samples and see if they appear to have the same features from some limited feature list as the object being studied, one examines the dynamics of the object being studied to see if it fits the model. Naturally, another advantage of this approach is that time series analysis might suggest variations or outright changes in the underlying model that would lead to better modeling. While time series analysis is certainly at least implicit in other works, I am not aware of any clear theoretical formulation in the setting of analysis of power law graphs. An appropriate theoretical framework is imperatively needed.

In most cases, time series analysis for computer network problems will require some type of *trace-based analysis*, obtained by monitoring the system in some way. Unfortunately, trace-based analysis is often difficult and time-consuming, with many potential problems. The large quantity of data that must be gathered is one obvious drawback. Heterogeneity in the network, including heterogeneity over users and heterogeneity in behavior over time, is another difficulty. Inherent noise may be a challenge. Finally, when dealing with existing networks, relevant past data may not have been captured, making it difficult to determine if the model was appropriate from the beginning.

Sampling provides efficient means of coping with some of the problems of large traces. There are two natural sampling approaches in this setting. The first is to consider global traces at discrete time steps, or *snapshots*. The second approach is to capture more detailed information at a subset of the sites. Which type of sampling is appropriate may depend on the object being studied; in some cases both would be appropriate, and they may reinforce each other.

Coping with these various problems will require a rich collaboration between theorists and system builders, as implementing appropriate sampling, recording, and analysis tools for such analyses will require sophisticated understanding of both the underlying theory and systems.

While time-series analysis with trace-based analysis may be one path toward validation, there should certainly be others. The underlying question is whether or not we can break the systems we are studying down sufficiently that we can successfully map actions being taken to steps in the model. It seems a goal within reach, and certainly a challenge that should be tackled.

There is still the further issue of finding ways to control system behavior. In many natural sciences, simply observing a power law is a significant result, and there is little hope of changing the underlying resulting power law. For example, the size of earthquakes roughly follows a power law distribution, and without significant human intervention in the movements of the earth's crust, there is little hope that there is much we can do to affect it. In computer systems, however, we might expect that in many cases we can change the system behavior, potentially modifying the causes that give rise to the power law in order to better engineer the system. In our world, we can to some extent control the laws of nature, and we need to know how to use that power effectively.

There are several basic means of implementing control. For example, we might find additional constraints to add to the system. Imagine, for example, that we introduced a geometric restriction, allowing links only of a certain length. This is the idea behind the geometric preferential attachment model studied in [12]; links arise only between nodes sufficiently close in some suitable metric space. Rather than adding direct constraints on the users or the system, one might achieve similar effects through the softer approach of adding incentives or costs to the system. Such constraints arise in the various HOT (Highly Optimized Tolerance, Heuristically Optimized Tradeoffs) proposals [5, 8], but the effects of trying to control the system through such constraints remains open. The study of how to affect system behavior would also fit nicely into the growing area of *distributed algorithmic mechanism design*, as introduced in [10, 11].

The truth is, however, that at this point our understanding of the actual processes that generate power laws in networks is sufficiently limited that we do not know what kinds of control mechanisms are suitable. The foundations of this area are still wide open. In addition, by gaining a better understanding of mechanisms that underly power laws and how they might be modified in the computer network domain, we might also find approaches that work in other domains, including social networks, economics, or even biological systems. In this way, our insights might have a powerful impact outside the field of computer networks as well.

3 Conclusion

The proposed agenda is that we must move from observation, interpretation, and modeling toward validation and control in power law research. Validation holds the promise of making power law research more scientifically sound; control holds the promise of making it more directly useful. Recent signs suggest that we have already started moving along this path, as people have begun to realize that claims being made for many power law models go beyond what can reasonably be shown, and papers invalidating previous models are coming to the forefront. The article in this issue by Li, Alderson, Doyle, and Willinger is an outstanding example of this, introducing both what has been done in the past and research directions for the future [17].

The idea behind this agenda yields many interesting possible corollaries on which the community must build a consensus. For quite a while, an NP-completeness result has generally not been considered worthy of publication, unless the result was important for a specific field, especially challenging mathematically, or led to new mathematical or computational insights. I would suggest that the same standard should now hold for a large part of the research in power laws. Already, observing a power law in itself no longer seems sufficient for publication in networks, except in rare circumstances; usually some proposed model is expected. With the plethora of papers on modeling available, perhaps the bar will start to rise for such papers. At the very least, I would argue that modeling papers should include as a matter of course a proposal for how the model could potentially be validated, even if the authors do not themselves tackle the challenge of gathering and processing the data to perform the validation. Or perhaps papers giving new variations of existing models should be required to explicitly invalidate the old models, using real data.

While this proposed agenda certainly raises a number of challenges, it creates many opportunities and research problems. Research in power laws has already brought together computer scientists from theory and systems in remarkable ways, as well as researchers from many other communities. The agenda of validation and control promises to continue and possibly even enhance these collaborations. The agenda also raises the possibility that this research can begin actively affecting real systems, making them more efficient and better designed. The past decade has been very exciting, with many remarkable discoveries in the theory and practice of power laws. I believe that moving the agenda toward validation and control will lead to continuing excitement from this research area.

Acknowledgments

Thanks to David Parkes for several interesting discussions on this subject, and thanks to Scott Aaronson, John Byers, and Fan Chung for comments on earlier drafts.

References

[1] A.-L. Barabási. *Linked: How Everything is Connected to Everything Else and What It Means.* Perseus Publishing, 2002.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, pages 509-512, 1999.

[3] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, vol. 272, pages 173-189, 1999.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web: experiments and models. In *Proc. of the 9th World Wide Web Conference*, 2000.

[5] J. M. Carlson and J. Doyle. Highly optimized tolerance: a mechanism for power laws in designed systems. *Physics Review E*, 60(2):1412-1427, 1999.

[6] S. Chakrabarti, A. Frieze, and J. Vera. The influence of search engines on preferential attachment. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 293-300, 2005.

[7] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, W. Willinger. The Origin of Power Laws in Internet Topologies Revisited. In *Proceedings of IEEE INFOCOM*, pp. 608-617, 2002.

[8] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically Optimized Tradeoffs: A new paradigm for power laws in the Internet. In *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, pp. 110-122, 2002.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM 1999 Conference*, pages 251-261, 1999.

[10] J. Feigenbaum and C. Papadimitriou and R. Sami and S. Shenker. A BGP-based Mechanism for Lowest-Cost Routing. In *Proc. of the 2002 ACM Symposium on Principles of Distributed Computing*, pp. 173–182, 2002.

[11] J. Feigenbaum and S. Shenker. Distributed Algorithmic Mechanism Design: Recent Results and Future Directions. In *Proc. 6th Int. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, pp. 1-13, 2002.

[12] A. Flaxman, A. Frieze, and J. Vera. A geometric preferential attachment model of networks. In *Proc. of 3rd International Workshop on Algorithms and Models for the Web-Graph*, pp. 44-55, 2004.

[13] B. A. Huberman and L. A. Adamic. Evolutionary dynamics of the World Wide Web. Technical Report, Xerox Palo Alto Research Center, 1999. Appears as a brief communication in *Nature*, 401, p. 131, 1999.

[14] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: measurements, models, and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, 1999.

[15] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling Biases in IP Topology Measurements. In *Proceedings of IEEE INFOCOM*, pp. 332-341, 2003.

[16] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceedings of the 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 177-187, 2005.

[17] L. Li, D. Alderson, C. Doyle, and W. Willinger. Towards a Theory o Scale-Free Graphs: Definition, Properties, and Implications. This issue.

[18] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, vol 1, No. 2, pp. 226-251, 2004.

[19] M. E. J. Newman. Power laws, Pareto distributions and Zipf's Law. *Contemporary Physics*, 46, 323-351 (2005).

[20] D. J. Watts. *Six Degrees: The Science of a Connected Age.* W. W. Norton and Company, 2003.