What are our standards for validation of measurement-based networking research?

Balachander Krishnamurthy and Walter Willinger
AT&T Labs-Research
Florham Park, NJ 07932
{bala, walter}@research.att.com

ABSTRACT

Standards? What standards?

1. INTRODUCTION

Measurement-based studies have become an increasingly important component of Internet research. One naturally wonders if - as a research community - we have adhered to or established certain standards for performing such studies. We argue in this position paper that there really exist no such standards, and researchers tend to repeat common errors and are by and large oblivious to the many pitfalls that stem from taking Internet measurements at face value. One could simply attribute this situation to "laziness" on the side of networking researchers and leave it at that. Alternatively, one could confront the community with a stringent set of standards, but this may feel like asking to "live a life without biblical sin." Neither solution will help much and will change the current lack of standards. Instead, by posing a number of specific questions and discussing possible answers, we advocate here a practical approach to arrive at a prudent sense of just what these desired standards should be and may be able to achieve. However, we fully realize that a commonly-accepted set of standards can only be established and implemented through a true community effort, and we outline some initial steps at initiating such an effort within the networking research community.

With the proliferation of measurement-based Internet research efforts (e.g., see [3] and references therein) has come an increasing awareness of the often limited nature and poor quality of the measurements that can be collected from decentralized and distributed large-scale system such as the Internet. Clearly, the responsibility to fully acknowledge and detail the main limitations, shortcomings, and pitfalls that are exhibited by some of the most widely available and used measurements and that result from the inherent inability of the Internet to efficiently and effectively support large-scale third-party measurements lies squarely with the networking researchers. They are either the producers, owners, consumers, or users of the existing data and as such, can be expected to be intimately familiar with the instrumentation, measurement process, or collection effort. In particular, they ought to be knowledgeable about deciding whether or not to use the data at hand for purposes for which the original measurements were never intended. Unfortunately, as a community, we have by and large set a bad example, either by ignoring these responsibilities all-together or by only giving lip service to them.

To address this problem, we pose in this paper a number of questions that we believe are pertinent to raising the bar for measurement-based Internet research. The questions' sole purpose is to raise some critical issues that any scientist involved in measurement-based research should mull over and should be able to address in the context of his or her work. The dilemma we face posing such questions is that they have to be high-level and flexible enough to be broadly applicable and at the same time detailed and specific enough to avoid being viewed as "mom and apple pie" questions. We try to overcome this dilemma by illustrating answers to our questions in the context of a specific application area where measurements have played an important role and where the answers show the potential that our questions have for raising the standards.

The key question for any measurement-based research effort is **Q: "Do the available measurements and their analysis and modeling efforts support the claims that are made?"** We argue that a satisfactory answer requires a thorough examination of all or a subset of the following issues: (1) quality of the available measurements, (2) appropriateness of the statistical analysis of the measurements, (3) scientific value of the modeling approach, and (4) thoroughness of the validation effort. In turn, such an examination can be facilitated by a set of sub-questions that deal with issues (1)-(4) and concern aspects of data hygiene and data usage, data analysis, and the purpose of modeling and meaning of model validation, respectively. In the rest of the paper, we discuss these issues in more detail and illustrate the breadth of criteria that, when applied judiciously, are bound to raise the standards for measurement-based Internet research. We also discuss our approach in the context of parallel efforts aimed at improving strategies and practices for sound Internet measurements [10], enabling sharing of data via anonymization techniques [8], establishing an etiquette for using someone else's measurement data [2], enforcing a level of statistical rigor that is commensurate with the quality of the available data [12], and advocating modeling and model validation efforts that move beyond trivial data fitting exercises [13, 4]. While clearly inspired by these parallel efforts, the objective of this position paper is more holistic—to improve measurement-based research as a whole (including measurements, their use and analysis, modeling, and model validation) to the point where its reputation as a scientific discipline is no longer an issue.

2. INTERNET MEASUREMENT ISN'T EASY

Designing and deploying techniques and tools for measuring the Internet and for collecting original data sets comes with responsibilities that the networking research community has not taken all that serious. The key problem faced by nearly anyone wanting to perform measurements on the Internet is that its decentralized and distributed architecture does not support third-party measurements. As a result, measurement efforts that involve multiple ISPs or ASes become non-trivial, often rely on engineering hacks, typically require innovative new approaches that are rarely validated in practice, and generally may not yield the originally desired data. However, the networking community has been slow in accepting the fact that more often than not, *what we can measure in an Internet-like environment is typically not the same as what we really want to measure (or what we think we actually measure)*, and this basic observation can have serious and wide-ranging implications for the analysis and modeling of the resulting measurements, as well as for the validation of the claims that are based on them.

Ignoring this fact is clearly not a solution as it has the potential of preventing us from deriving results from our measurements that we can trust. At the same time, detailing each and every known problem and deficiency is typically a painstakingly arduous process (e.g., [10, 11]) that generally garners little or no acclaim or appreciation. However, if we as domain experts don't make the effort to list all known deficiencies associated with the measurement techniques we develop, the measurement tools we deploy, and the measurements we collect, who will? Also, if we as domain experts don't make this activity which we call "data hygiene" a focal point of any future measurement-based research effort, how will we ever be able to understand our data to the point where we can trust the results? Finally, if we don't take data hygiene more seriously, how can we criticize non-networking researchers who tend to take publicly available data sets at face value and often make claims that subsequently collapse under scrutiny of the data and/or when checked by experts?

To ensure that we can trust the results that we derive from our data to the point that they withstand detailed scrutiny of the underlying measurements by domain experts, we argue that any work in the area of measurement-based Internet research ought to be able and answer in the affirmative the above-posed question Q. It also should be able to support the answer with rigorous and verifiable arguments. In the following, we illustrate how outlining a broad strategy for answering this basic question has the potential of providing a prudent sense of just what sort of standards measurement-based Internet research may need to become a respected scientific discipline.

3. TO GET STARTED: FOUR QUESTIONS

Measurement-based networking research and its validation is largely a lesson in how errors of various forms occur and can add up. Some of these errors' main sources are the measurement process itself, the analysis of the resulting data, the modeling work that is informed by this analysis, and the model validation effort. Any proposed framework that helps keeping those errors in check to the extent possible can be viewed as raising the bar for measurement-based networking research and as defining an initial set of standards of the type we envision the networking community will embrace, and–over time–refine, improve, and ultimately adhere to. Paraphrasing Paxson [10] slightly, the focus should be emphatically on *"developing confidence that the results derived from [the measurements at hand] are indeed well-justified claims"* and argues for posing a set of questions that try to expose the likely sources for errors and are concerned with issues related to data hygiene, data analysis, and modeling, including model validation, respectively. As illustrated in Section 4, depending on the scope of the work in question, measurement-based networking research efforts have to deal with all or a subset of these issues.

3.1 Issue #1: Data hygiene

Q1: Are the available measurements of good enough quality for the purpose for which they are used in the present study?

Answering this question typically depends on whether the author of the present study is the producer of the available measurements or simply a consumer (i.e., user of data collected and made available by someone other than the author). In the author-as-producer case, the list of suggestions presented in [10] includes the commendable one to maintain *meta-data* associated with any newly collected set of Internet-related measurements, but in practice, this suggestion is seldom honored. We believe it is paramount to revisit the *meta-data* idea, expand it, and develop it to the point where its utility becomes obvious and is critical to any serious data hygiene effort. While there is in general no "best practice" for collecting and maintaining meta-data associated with a newly collected dataset, in practice, any meta-data description should aim to include as much of the information that is pertinent to the collection of the measurements and their future use by third-parties. Ideally, this information should provide details about the measurement technique used, its shortcomings and limitations (if any), and alternative techniques considered. It should spell out in detail any issues concerning bias, completeness, accuracy, or ambiguity of the data that are known as a result of the producer's in-depth understanding of the measurement and data collection effort. And if at all possible, it should include any information about the operating conditions of the network at the time the measurements were made (e.g., relevant infrastructure- or protocol-specific aspects, network usage and application mix) and that might impact proper use of the data in subsequent studies (either by the author or other users).

In contrast to the author-as-producer case, the author-as-user case is generally concerned with datasets that the researchers who originally collected the data of interest have released and made available to other interested parties. While commendable, the main problem is that many of the currently available/produced datasets come without adequate meta-data descriptions. Consequently, a typical user of such data embarks almost immediately on an analysis of the data whose quality is largely taken for granted even though vital details about the measurement tools, data collection process, and networking conditions at the time of the data collection remain unknown. However, using such data for, say, the purpose of repeating the experiment in question and validating earlier results or, more importantly, for a purpose the original data were never intended to be used, needs to account for the ambient changes that may have occurred since the original data were collected. Clearly, without adequate meta-data, it is generally difficult to completely pinpoint those changes and understand the full extent of their impact, but applying proper domain knowledge can go a long way filling in missing meta-data information. In any case, the absence of adequate meta-data should not be an excuse for users of such data to neglect their responsibilities and largely ignore all issues related to proper secondary usage of original data. Providing convincing evidence that an existing dataset can be used for a very different purpose than it was originally intended is the sole responsibility of the user of such data and requires, at a minimum, a detailed account of the assumptions that are made about the data and a list of issues that a carefully crafted *meta-data* description of the measurements should/would address. While reliance on canonical datasets (common in other areas in science) would be clearly useful, such situations are rare in the Internet measurement field where the underlying conditions tend to undergo constant changes. Q1's sole purpose is to increase the focus on a dataset's meta-data description to the point where its critical role for efforts related to data hygiene becomes obvious, where its availability becomes the rule rather than the exception, and where using domain knowledge to check or enhance the description becomes the responsibility of any user of such data.

3.2 Issue #2: Data analysis

Q2: Is the level of statistical rigor used in the analysis of the data commensurate with the quality of the available measurements?

After assessing the overall quality of available measurements, the next step towards improving measurement-based research as a whole concerns the quality of the analysis of the data. At issue is how to analyze datasets that are in general tarnished by various documented or undocumented types of errors and imperfections, yet contain some amount of useful information. Mining that information is at the heart of the problem and requires a data analytic approach that matches well with the quality of the measurements.

Clearly, it makes little sense to apply very sophisticated statistical analysis techniques that are highly sensitive to inaccuracies in the data if the datasets have been identified to exhibit major deficiencies. Instead, what would be desirable here is an emphasis on statistics of the data and on statistical tools that are largely resilient to most of the known imperfections of the data. When accompanied by such strong robustness properties, the statistics and tools in question would be informative and useful despite the problematic nature of the data. In turn, such observed robustness properties of the data are extremely valuable, not only because they feed back to and enhance the *meta-data* description, but also because they become potential candidates for measurement invariants and as such provide critical information for future users of these measurements.

The biggest take-away points from measurement studies are often in "broad rules of thumb" and not in details. For example, an observed Pareto-type principle or 80/20-type rule (i.e., 80% of the effects comes from 20% of the causes) is often all that can be reliably and robustly inferred from high-variability data of questionable quality, and any attempt at fitting a specific parameterized model (e.g., a power-law type or some other gobbledygook distributions) would be statistical "overkill." In this sense, the question concerning statistical rigor cuts both ways – reliance on statistically sophisticated-looking methods in situations where the data don't justify their application should be as much frowned upon as avoidance of statistically rigorous approaches in cases where the data at hand justify a detailed and more elaborate analysis. Q2 is intended to raise the general awareness that there are important differences between analyzing high- and low-quality datasets, and that approaching the latter the same way as the former is not only bad statistics but also bad science.

3.3 Issue #3: Modeling

Q3: Have alternative models that are also consistent with the available data been considered, and what criteria have been used to rule them out?

Q4: Does model validation reduce to showing that the proposed model is able to reproduce certain statistics of the available data?

For measurement-based research studies that include a substantial modeling component, much of the current network-related modeling work can be succinctly summarized as follows. Start with a given dataset and take the available data at face value. Next infer some distributional properties of the data (mainly first-order properties, sometime second-order properties) and determine the "best-fitting" model (e.g., distribution, temporal process, graph) and corresponding parameter estimates. Here, "best-fitting" refers either to a subjective or "eyeballing" assessment of the quality of the fit or to an apparently more objective evaluation involving some commonly-used goodness-of-fit criterion. Lastly, argue for the validity of the chosen model by virtue of the fact that it reproduces the distributional properties of the data examined in the second step. However, given that more often than not, the available measurements cannot be taken at face value, providing an accurate description (i.e., model) of the data at hand is precisely no longer the point and largely counterproductive.

The commonly-used recipe for network-related modeling described above has reduced this activity to a large degree to an exercise in data fitting, a mostly uninspiring activity that creates little excitement and is generally detrimental to scientific advances. The reasons for this are all too clear. For one, there is no surprise in this approach as the recipe is guaranteed to produce a model. In fact, for one and the same set of distributional properties, there are in general many different models that fit the data equally well. Even worse, depending on the distributional properties of interest, the resulting models are likely to be different, and rarely do there exist solid guidelines for ruling out equally well-fitting models. The area that has been especially neglected by this widely-accepted approach is model validation. Models are generally declared to be valid by virtue of the largely predictable fact that they reproduce the very same statistics of the data that played a key role in selecting the model in the first place. How can we be confident that the results that we drive from such models are valid? Not only does the use of the very same dataset for both model selection and model validation pose serious statistical problems, but being able to reproduce some statistics of the data should be a simple and uninteresting by-product of a good model.

To develop a more scientifically grounded and constructive model validation methodology, a radical suggestion is to make matching particular statistics of the data a non-issue. While seemingly extreme and non-constructive, there are good reasons to consider this idea. For one, given the known deficiencies in the data, matching a particular statistics of the data may precisely be the wrong approach, unless that statistics has been found to be largely robust to these deficiencies. Moreover, it eliminates the arbitrariness associated with determining which statistics of the data to focus on. Indeed, it treats all statistics equally. A model that is approximately right can be expected to implicitly match most statistics of the data (at least qualitatively). Another concrete suggestion that would increase our confidence in a proposed model is to carefully examine it in terms of what new types of measurements it identifies that are either already available (but have not been used in the present context) or could be collected and used to check the validity of the model. Here, by "new" we do not mean "same type of measurements as before, just more." What we mean are completely new types of data, with very different semantic content, that have played no role whatsoever in the entire modeling process up to this point. A key benefit of such an approach is that the resulting measurements are only used for the purpose of model validation.[1] This way, there is a statistically clean separation between the data used for model selection and the data used for model validation, a feature that is alien to most of today's network-related models. Questions Q3 and Q4 reflect our envisioned standards for network-related modeling in general and network-specific model selection and validation in particular, and are intended to outline the new role for modeling when obtaining the "best-fitting" model for a dataset of questionable quality is precisely no longer the ultimate goal.

4. ILLUSTRATIVE EXAMPLES

Initial reactions to our attempts at raising the bar for measurement-based networking research through questions like Q1-Q4 have focused on a lack of specific examples and a general vagueness and open-endedness of the

[1] We re-iterate the "closing-the-loop" argument in [13].

questions that leaves researchers without precise guidance and is unlikely to lead to a consistent approach. To address both of these criticisms, we discuss in the following two papers in the area of Internet topology modeling which in the last decade has been an especially active field in measurement-based networking research. Specifically, we comment in the following on two of the most influential and highly cited papers in this field that have appeared in first-tier research venues and critique their main methodologies/claims in the process of answering some or all of the questions Q1-Q4. By "naming names", our intention is not to criticize particular authors or single out some of their work. The sole purpose is to demonstrate through concrete examples why and how errors in measurement-based research can occur and accumulate, how they can be exposed, and why questions like Q1-Q4 help us to distinguish between well-justified and specious claims.

4.1 On Power-Law Relationships of the Internet Topology [5]

The paper's claim to fame is that it is the first work in the networking area that reports on observed power-law distributions for the node degrees of inferred router-level and inferred AS-level topologies of the Internet. The paper relies on existing datasets, cites the sources for the data, and immediately embarks on an analysis of the data that takes the available measurements at face value. As the paper deals with existing datasets and is mainly concerned with their statistical analysis, the relevant questions in this case are Q1 and Q2.

As far as Q1 is concerned, we focus first on the dataset that is used to derive the reported power-law claim for the inferred router-level graph of the Internet denoted by ROUT-95 in [5]. To this end, a careful reading of the cited paper [9] that describes the original measurements is both educational and illuminating. For one, the explicit purpose for collecting the dataset underlying the ROUT-95 graph was *"to get some experimental data on the shape of multicast trees one can actually obtain in [the real] Internet ..."* [9], and the tool-of-choice was `traceroute`. Clearly, [5] and subsequent studies that have relied on this dataset have used it for a purpose for which it was not intended, namely inferring the Internet's router-level topology. Furthermore, the paper provides an early example of a meta-data description that, while somewhat informal and terse, is surprisingly detailed, informative, and useful. For example, in addition to numerous experiment-specific details such as route collection and coverage of hosts by domains and geography, the meta-data description in [9] also spells out in detail numerous problems and limitations due to the reliance on `traceroute`, including IP aliasing resolution and how it was handled, as well as `traceroute`'s inability to penetrate opaque Layer-2 clouds and its likely consequences, both for interpreting the data at hand and for future such collection efforts if Layer-2 technology becomes more prevalent. In view of this, it is very unfortunate that starting with [5], this meta-data description has been largely ignored and forgotten; in fact, the majority of later papers in this area typically only cite [5], but no longer [9]. Although such secondary citations are a well-known problem, in the measurement arena their impact tends to be magnified as critical information available in the primary citation is often obscured to the point where it is no longer visible in the cited work.

When combined with a basic understanding of the capabilities and limitations of the `traceroute` measurement tool, a careful examination of the meta-data description associated with the dataset reported in [9] results in a simple but **negative** answer to question Q1—the dataset at hand is inadequate for studying the Internet's router-level topology, and the main reason is it's sole reliance on the `traceroute` tool, which was never intended to be used to map the connectivity of the Internet at the router-level. In view of `traceroute`'s key limitations—the high-degree nodes it detects in the network core are fictitious and represent entire opaque Layer-2 clouds, and if there actual high-degree nodes in the network, existing technology relegates them to the edge of the network where no generic `traceroute`-based measurement experiment will detect them—our answer should come as no surprise and shows why domain knowledge in the form of such tool-specific "details" matters when dealing with issues related to data hygiene.

Next, we consider the datasets that have been used to derive the reported power-law claim for three inferred AS-level graphs of the Internet, denoted in [5] by INT-11-97, INT-04-98, and INT-12-98, respectively. As source for these datasets, [5] refers to *The National Laboratory for Applied Network Research (NLANR)*[2], which in turn relied on full BGP routing tables collected by the *Route Views Project at the University of Oregon* to generate the data for constructing the inferred AS connectivity maps. Largely unaware of the project's clearly articulated original purpose – *"to respond to interest on the part of operators in determining how the global routing system viewed their prefixes and/or AS space"* – starting with [5], the research community has started to rely on the resulting datasets for a purpose (i.e., inferring the Internet AS-level topology) for which they were not intended. However, in stark contrast to the dataset underlying the ROUT-95 router graph, these datasets come with essentially no meta-

[2] In the meantime, the NLANR project has officially ended, and the operational stewardship for all of its machines and data has been taken over by the *Cooperative Association for Internet Data Analysis (CAIDA)* at UCSD in July 2006.

data information that would help in deciding whether using these datasets for this "new" purpose is legitimate or problematic.

Like [5], most papers in this area have ignored this issue and have taken the available data at face value, despite early warnings by domain experts that these datasets may provide only a very sketchy picture of the Internet AS-level connectivity structure. As recent studies have documented (e.g., see [7] and references therein), these warnings were fully warranted and an "as-is" use of these BGP-derived datasets for studying AS-level connectivity is seriously flawed because of the high degree of incompleteness, inaccuracy, and ambiguity that the data exhibit and that impacts all aspects of a careful investigation of the Internet's AS-level topology. Thus, in the case of these BGP-derived datasets, the answer to question Q1 is again simple and **negative**, but it is largely accumulated domain knowledge and not readily available meta-data that leads to this answer. In short, BGP is *not* a mechanism by which ASes distribute their connectivity. Instead, BGP is a protocol by which ASes distribute the reachability of their networks via a set of routing paths that have been chosen by other ASes in accordance with their policies. Naturally, each AS can only see the subset of existing AS connections formed by these policy-influenced routes, and it is again these "details" that prevent an "as-is" use of these datasets beyond the purpose for which they were originally collected.

In view of our negative response to question Q1, answering question Q2 becomes straightforward. Indeed, having listed some of the more severe problems with `traceroute`-derived router-level graphs, it should be clear that any fitting of a particular parameterized distribution (e.g., power-law distribution with index α as in [5]) is statistical "overkill." In the case of the dataset underlying the inferred ROUT-95 router-level graph, even broad "rules-of-thumb" type claims like a Pareto-type principle for node degrees cannot be justified in view of the fundamental problems of `traceroute` with respect to the high-degree nodes. On the other hand, for the datasets underlying the inferred AS-level graphs, power-law claims with specific α-values for the inferred node degree distributions cannot be supported, and all that can be concluded with sufficient confidence from the BGP-derived AS maps are Pareto-type principles; that is, a small number of nodes have many neighbors, while most nodes are connected to only a small number of neighbors. In this sense, the answer to question Q2 is also negative and follows directly from the data hygiene issues that were raised in the process of answering question Q1.

4.2 Error and Attack Tolerance of Complex Networks [1]

From an Internet perspective, this paper owns its popularity to an appealingly simply model of the Internet AS-level topology that yields surprising discoveries resulting in significant claims. In particular, using a scale-free network model of the preferential attachment type, [1] claims that the AS-level Internet exhibits a surprisingly high degree of tolerance against random node failure, but that this error tolerance comes at a high price in that the AS-level Internet is extremely vulnerable to targeted attacks (i.e., the selection and removal of the high-degree nodes), a property that has become known as the "Achilles' heel of the Internet".

To arrive at these conclusions for the AS-level Internet, [1] relies on an existing dataset, cites the source for the data and the source for the assumed power-law node degree distribution, and embarks on a modeling effort that takes the data at face value and the power-law claim as given. As such, the relevant questions in this case are Q1–Q4, but since the data source is the same as in [5] (i.e., NLANR), and since [5] is also cited as source of the power-law claim, we already know that the answers to Q1 and Q2 are negative. As for questions Q3 and Q4, the answers are easy and explicitly given in [1]: "no" for Q3 and "yes" for Q4. Thus, in terms of our overall quest for determining whether the available measurements and their analysis and modeling efforts support the claims that are made in this paper with respect to the Internet, the result is a picture-book example of how errors can add up and produce completely unsubstantiated claims, even though they may look quite plausible to non-networking experts. The key results are derived from a model which is argued to be valid for the sole reason that it exhibits a power-law node degree distribution, an assumed property of the AS-level Internet that is not supported by the used dataset in the first place. And even if it could be supported, there exist many different types of network models with identical power-law connectivity distribution but radically different structural features (e.g., see [4] and references therein).

5. CONCLUSION AND OUTLOOK

Trying to raise the bar for measurement-based networking research clearly adds more work. However, while for example, maintaining adequate meta-data is especially important for rapidly evolving and changing systems such as the Internet for which the value of a given set of measurements is bound to change over time, in practice, this property also makes researchers think twice before investing a lot of time and efforts setting up accurate measurements of phenomena that may or may not exist over a longer period. Arguing for a more prominent role for the meta-data idea seems to strike a healthy balance between aiming for "perfect" measurements that may take an unreasonable time and effort to collect and may have only a short shelf time and producing "useful" measurements where the required effort/time is more commensurable with the data's generally short shelf life and typically limited usage. Only time will tell if researchers will be willing to spend the time and effort needed to equip their measurements with adequate meta-data, but as illustrated in Section 4, evidence is accumulating that ignoring this issue is detrimental for measurement-based networking research.

Similarly troubling is the rather precarious current state of network-related modeling where the same underlying dataset can give rise to very different, but apparently equally "good" models, which in turn can give rise to completely opposite scientific claims and theories concerning one and the same observed phenomenon. Clearly, model validation has to mean more than being able to match the data well. Also, who is to say that matching a few statistics of the data determines a model? It does not, and two models that match the data well with respect to some statistics can still be radically different in terms of other properties, their structures, or their functionality. While remembering G.E.P. Box's observation *"All models are wrong, some models are useful,"* without being more specific about which models are deemed useful and why, the comment is of little practical value. A more constructive piece of advice aligned with what we envision with respect to model validation is from B.B. Mandelbrot [6], who observed *"If exactitude is elusive, it is better to be approximately right than certifiably wrong."* A driving force behind this break with traditional modeling has been the realization that because of its engineered architecture, a thorough understanding of its component technologies, and the availability of extensive (but not necessarily very accurate) measurement capabilities, the Internet provides a setting in which most claims about its properties, structure, and functionality can be unambiguously resolved, though perhaps not without substantial efforts. In turn, models that result in incorrect, misleading, or wrong claims can and will be identified and labeled accordingly, but it may take considerable time (and efforts) to expose them. This motivates the development of modeling approaches that respect the designed nature of the system, reflect the engineering intuition that exists about great many of its parts, and that are fully consistent with as many measurements as possible.

While we view questions Q1–Q4 as a first step towards raising the bar for measurement-based networking research, we also believe that trying to get agreement on basic standards requires a much broader effort than just our (likely biased) views and needs the involvement of the community as a whole. Such a community effort should encourage an ongoing dialogue between measurers, modelers, and experimenters. It would have the additional benefit of creating a compliance/verification toolkit so that the researchers have a way of ensuring that they have met certain standards or rules. Currently, in ongoing work we are constructing a more extensive list of examples that cover areas other than Internet topology modeling. We believe that once as a community, we have developed a canonical set of examples in various categories, the applicability and usefulness of a suggested set of rules or standards in the context of a *specific* measurement experiment will become clear and easy to answer. When done right, imposing reasonable standards can define new research directions in statistics, data analysis, and mathematical modeling and can contribute to a scientifically more viable modeling paradigm. However, for networking researchers, the ultimate promise is that when executed diligently and fully consistent with the proposed (or extended) standards, measurement-based research is capable of providing an unprecedented understanding of complex, large-scale, engineered systems such as the Internet and of more virtual systems associated with it.

6. REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks, *Nature*, 406, 2000.

[2] M. Allman and V. Paxson. Issues and etiquette concerning use of shared measurement data, *Proc. IMC*, 2007.

[3] M. Crovella and B. Krishnamurthy. *Internet Measurement: Infrastructure, Traffic, and Applications.* J. Wiley&Sons, New York, 2006.

[4] L. Li, D. Alderson, J.C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definitions, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology, *ACM Computer Communication Review*, 29(4), 1999.

[6] B.B. Mandelbrot. *Fractals and Scaling in Finance.* Springer-Verlag, New York, 1997.

[7] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang. In search of the elusive ground truth: The Internet's AS-level connectivity structure, *Proc. ACM SIGMETRICS*, 2008.

[8] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization, *ACM Computer Communication Review*, 36(1), 2006.

[9] J.-J. Pansiot and D. Grad. On routes and multicast trees in the Internet, *ACM Computer Communication Review*, 28(1), 1998.

[10] V. Paxson. Strategies for sound Internet Measurement. *Proc. IMC*, 2004.

[11] D. Stutzbach and R. Rejaie. Understanding churn in peer-to-peer networks. *Proc. IMC*, 2006.

[12] W. Willinger, D. Alderson, and L. Li. A pragmatic approach to dealing with high-variability in network measurements. *Proc. IMC*, 2004.

[13] Willinger et al, Scaling phenomena in the Internet: Critically examining criticality. *Proc. Nat. Acad. Sci.*, 99:2573–2580, 2002.