A Socratic method for validation of measurement-based networking research

Balachander Krishnamurthy[a], Walter Willinger[a,*], Phillipa Gill[b], Martin Arlitt[c]

[a]*AT&T Labs-Research, Florham Park, NJ, USA*
[b]*University of Toronto, Toronto, ON, Canada*
[c]*HP Labs, Palo Alto, CA, USA and University of Calgary, Calgary, AB, Canada*

Abstract

Nearly three decades of Internet measurement has resulted in large-scale global infrastructures used by an increasing number of researchers. They have examined various Internet properties in areas such as network infrastructure (routers, links), traffic (measurement at packet, flow, and session level) and applications (DNS, Web, P2P, online social networks etc.) and presented results in diverse venues. Key related topics like security and privacy have also been explored. There is however a lack of clearly articulated standards that reduce the probability of common mistakes made in studies involving measurements, their analysis and modeling. A *community-wide* effort is likely to foster fidelity in datasets obtained from measurements and reused in subsequent studies. We present a Socratic approach as steps towards a solution to this problem by enumerating a sequence of questions that can be answered relatively quickly by both measurers and reusers of datasets. To illustrate the applicability and appropriateness of the questions we answer them for a number of past and current measurement studies.

Keywords: Internet measurement, topology, wireless, Web, Online Social Networks

1. Introduction

Although the Internet has been studied for decades with increasing diversity in the set of measurements collected and entities studied [20], there has been a notable lack of precisely articulated standards for such measurement-driven studies. Inherently the problem space is very large: the Internet is vast, constantly changing, reaches a significant fraction of the world's population, and is a key component in various aspects of daily life. At the same time, Internet researchers have diverse objectives ranging from performing highly specialized case studies to developing a theoretically sound foundation for the study of Internet-like systems. Thus, agreement on a single standard is unlikely to emerge quickly. We have a more modest goal in mind: raise the standards for validation of measurement-based networking research.

This paper expands on a HotMetrics'08 position paper [33] that argued for a practical approach to raising the bar for validating measurement-based networking research and to arriving at a prudent sense of just what the desired standards should be and may be able to achieve. Elaborating on the ideas discussed in [33], this paper outlines such an approach and illustrates it with a number of different examples. We fully realize that a commonly-accepted set of standards can only be established and implemented through a true community effort, and the main purpose of this work is to jump-start such an effort by advocating an approach that has the potential of triggering the necessary discourse within the community. By serving as a "strawman", the proposed approach is bound to meet objections and actively invites constructive criticism so standards that will ultimately emerge will generally be viewed as realistic and specific rather than as too idealistic or vague.

True to its "socratic" nature, our approach starts with a simple question: "Do the available measurements and their analysis and modeling efforts support the claims that are being made [in the paper at hand]?" Surprisingly, such an obvious question is typically not asked before efforts are expended. If the original measurers themselves do not ask this question, the subsequent users of the paper and data appear to fare no better. Often the key detraction is that a detailed recounting of all the potential pitfalls in carrying out measurements (*data hygiene*) is painful and severely under-appreciated; hence it is under-reported in papers (for two text-book examples that illustrate the meaning of "good" data hygiene, see [52, 58]). Issues relating to data hygiene may seem mundane and thus are rarely documented leading to the data being taken at face value. Rather than simply take researchers to task we start by refining the above question and advocate a *Socratic method*: asking researchers to answer a series of specific questions about the creation or reuse of data, and if applicable, about its statistical analysis, and validation of the proposed model. The purpose of these questions is to actively engage researchers to look at data closely, examining its hygiene, how it was analyzed, and what efforts were spent on modeling. We focus on the different roles played by the participants, such as those who produce the data and those who are the primary consumers of the data. Obviously if the original data gathering was unhygienic, the problem is compounded if

*Corresponding author
Email addresses: bala@research.att.com (Balachander Krishnamurthy), walter@research.att.com (Walter Willinger), phillipa@cs.toronto.edu (Phillipa Gill), arlitt@hpl.hp.com (Martin Arlitt)

the consumers were either unaware of it or did not take it into consideration. Even with properly gathered data it is possible for it to be misused by the consumers. It should be noted that producer and consumer groups may not intersect for a particular dataset but could easily overlap for a different one. Our proposals apply to users in either role. Our work is aimed at those who have some basic networking knowledge and have carried out or are interested in collecting measurements and/or using available data.

We are not the first to examine many of the issues above. For example, in the area of mobile ad-hoc network simulation, a plea for researchers to publish their data and meta-data along with their results, models, and statistical analysis has been made in [37]. The paper also shows that a general reluctance to do so has impeded a more open scrutiny of research in that area and has hurt the credibility of simulation as a research tool for the study of mobile ad-hoc networks. In the field of Internet measurements, researchers have tried to address the problem of improving the way in which data is gathered, shared, and used. For example, [53] enumerated a list of strategies, while [3] suggested proper ways for reusing data. Similarly, matching statistical rigor to the quality of the available data has been examined [61]. Others have examined modeling and validation efforts beyond just trivial data fitting exercises [42, 62]. Meta-data issues have been discussed [47, 53] and concerns about the treatment of shared measurements have been raised [2]. The brittleness of metrics have been examined in other contexts as well, e.g., in operating systems [46]. We however seek to place all of measurement-based research on a strong scientific substrate by a holistic examination of measurements, their use, analysis, modeling, and model validation.

The Internet research community has shown an increasing interest in having more datasets be shared. SIGCOMM and SIGMETRICS have a long history of encouraging empirical-based research, and conferences like IMC and PAM require datasets to be shared for a paper to be considered for the best paper award. As more and more datasets become available, the need for improved standards increases, as does the urgency for approaches advocating higher standards. To this end, our goal is to assist the measurement research community create, populate, and maintain a repository of meta-data associated with various datasets used in papers that they author. Such a repository would be similar to citation repositories. Ideally, the original measurer would participate and include enough information in their paper to enable consumers to easily glean answers to their questions about the resulting measurements. Failing that, any subsequent user has to answer the question and suggest changes/improvements to the meta-data in the repository. Participating in this process would help the consumer articulate their assumptions clearly and help future analysis.

The rest of the paper is divided as follows: Section 2 lists our initial set of rules and questions. Section 3 presents a detailed evaluation of diverse applications through the prism of our questions. Section 4 presents the inferred set of steps (the algorithm) so that any future measurer can follow the proposed standard. We conclude in Section 5 with a summary of our contributions and a look at future work.

2. Questions

What are the ways by which we can deconstruct the question we raised in Section 1: "Do the available measurements and their analysis and modeling efforts support the claims that are being made [in the paper at hand]?" We start by dividing this question into three broad sub-questions that deal with the issues of data hygiene, data analysis, and modeling efforts. Although we discuss these issues separately, it is understood that they are inter-related in the sense that data analysis and modeling are often useful tools for examining the hygiene of a given dataset. A schematic picture of our proposed Socratic approach is shown in Figure 1, and the different parts are discussed in more detail below.

2.1. Data hygiene

In deploying a measurement infrastructure for collecting data, the collector must list all known deficiencies associated with the measurement process and the measurements collected. Data hygiene is indicated by how carefully the quality of the measurements are checked and lies at the heart of any potential improvement to the situation at hand. The primary way by which hygiene can be ensured is the proper maintenance of meta-data associated with the measurements [53]. The meta-data should encompass all relevant information about the data and be examined at any subsequent date to assess the fidelity and applicability of the data. Typical components of meta-data in this context include: what measurement techniques were used, conditions of the network at the time of data gathering, and information about the location of the data gathering. For example, if the traffic mix at the location is heavily biased towards Web and P2P with only a tiny fraction of traffic from Online Social Networks (OSNs), then it is probably not a good candidate for reuse in examining the distribution of different OSNs.

While it is easy to stress that all relevant information about the data gathering process should be recorded and stored, it is unlikely to be complete without a semi-structured schema describing all the records and fields of interest. The hardest part of measurement-based meta-data and one almost always overlooked is the need for the creators of the meta-data to include warnings and known limitations about the reuse of the data. For example, information about any known biases, concerns about degree of accuracy, or the duration of applicability and usefulness of the data should be an essential component in the meta-data description. In its presence consumers can quickly check the meta-data and decide if the data can be safely reused. In its absence there is a strong likelihood of consumers going astray. Without such meta-data, the consumer is likely to blindly assume that the data is of good quality. What exacerbates the problem is that producers and consumers tend to have different expertise or objectives, with the former producing measurement data typically for a particular purpose, and the latter using them often with a very different goal in mind. Providing meta-data with clear semantics using languages that are expressive but also appeal to producers and consumers alike would be one way to alleviate this issue.

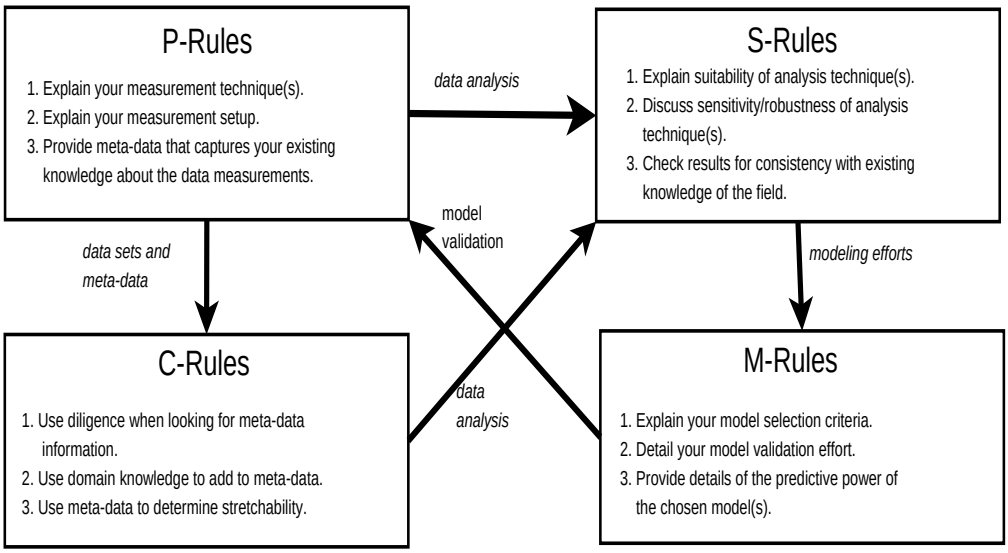

Figure 1: The Socratic approach in a nutshell.

Applying proper domain knowledge can help to fill in missing meta-data information. However, consumers cannot evade the responsibilities of proper secondary usage. If they intend to use the data for a different purpose then a detailed account of the assumptions made is essential. Internet measurement does not in general have a notion of canonical or benchmark datasets that is present in certain other scientific disciplines. Partly it is a result of a lack of longevity due to rapid churn in network conditions as well as application and traffic mix. The questions related to data hygiene focus on the need for a dataset's meta-data description so that meta-data availability becomes the norm rather than the exception. Using domain knowledge to check or enhance the description becomes the responsibility of any user of such data. Note that the questions are refined by listing descriptive key words/phrases such as *P-* (producer) and *C-* (consumer) rules.

P-rules for data producers: *Are the produced data of sufficient quality for the purpose for which they are used in the present study?*

1. Explain your measurement technique(s).
2. Explain your measurement setup.
3. Provide meta-data that captures your existing knowledge about the measurements.

The **P**-rules are essentially to ensure that producers of data clearly explain their knowledge about their measurements so the consumers can make an informed decision before using the dataset. So, for example, if the data producer used a particular measurement technique (say `traceroute`) they can indicate its inability to look into Layer-2 clouds. An example of a quality metric is associating detailed information in a packet trace dataset with the count of lost packets, various statistics about the burst-length of losses, and reasons for any such losses. One such paper that we recommend is [39]; by "measuring the measurer" the authors were able to provide minute details about packet losses for their data collection effort that resulted in the well-known Bellcore traces. By sharing information about the measurement setup, consumers may be able to glean enough information to decide whether it would be an appropriate reuse of the data for their application. If the measurement setup supported only unidirectional traffic gathering, involved middleboxes with caches, or local configuration that selectively blocked certain protocols or ports, then the traffic data could be affected. The extent of any deficiencies in the measurements and attempts taken to circumvent them must be explicit in the meta-data. Otherwise blind reuse of the data could result in false inferences.

There are several known problems in providing datasets. Some measurements are gathered in a closed environment where it is impossible to release data due to laws requiring privacy protection. Some data can only be made available in anonymized form. In the former case researchers are still obligated to provide a detailed answer to the questions posed by the **P**-rules and carefully document the schema of the data. In the latter case, several efforts have been made to suggest ways by which the anonymized data can still be useful for future studies (see Chapter 8 of [20] for a detailed discussion).

C-rules for consumers of data: *Are the available data of sufficient quality for the purpose for which they are used in the present study?*

1. Use diligence when looking for meta-data information.
2. Use domain knowledge to add to meta-data.
3. Use meta-data to determine stretchability.

The **C**-rules are best used at the start of the project that reuses data. The reuser must closely examine meta-data when they are available and if not reverse engineer them to the extent feasible and appropriate. The responsibility of proper use of existing datasets solely rests on the consumer. Examination of meta-data may reveal the expected lifetime of the data, the location

and prevalence of specific protocols in the traffic mix, and if it was associated too closely with the particular domain where it was originally created and used (e.g., WWW). *Stretchability* indicates how far an original dataset can be "stretched" and still be reused in a context for which the dataset may never have been intended to be used. Stretchability is a meta-property that signifies how applicable a qualitative property that has been derived from the original dataset is to the different usage of that dataset.

2.2. Data analysis

Often data analysis takes place in an atmosphere where the data may be unclean; yet extracting some useful information from it necessitates a data analytic approach that meshes well with the quality of the measurements. There is no point in using precise and highly sensitive statistical techniques when the datasets are known to have major deficiencies. What is needed instead are statistical tools that can tolerate known imperfections of the data. The resulting observed robustness properties of the data enhances the meta-data description and are potential candidates for measurement invariants providing critical information for consumers.

The key takeaways from measurement studies are often broad "rules of thumb" of the form of an observed Pareto-type principle or 80/20-type rule (i.e., 80% of the effects comes from 20% of the causes). If this is all that can be inferred from high-variability data of questionable quality then attempts at fitting a specific parameterized model (e.g., a power-law type or some other closed-form distribution) would be detrimental. The question related to analysis highlights key differences between analyzing high- and low-quality datasets and warns that ignoring the distinction is bad statistics and bad science: *Is the level of statistical rigor used in the analysis of the data commensurate with the quality of the available measurements?*

The **S**-rules or the statistical rules are:

1. Explain suitability of analysis technique(s).
2. Discuss sensitivity/robustness of analysis technique(s).
3. Check results for consistency with existing knowledge of the field.

A particular statistic of the data or statistical tool should not be so generic that it provides no information. An example of an unsuitable technique showing violation of such a non-informative methodology are "size-frequency" plots: log-log plots where the x-axis shows the value of some variable of the data (e.g., size, degree) and the y-axis depicts the frequencies with which the different values occur. The values on both axes are plotted on logarithmic scales. As illustrated in [42], these so-called "size-frequency plots" have a tendency to exhibit a straight-line behavior—a hallmark of apparent power-law relationships—even if the measurements are samples of an underlying low-variability distribution (e.g., exponential) and are therefore quite inconsistent with power-law behavior. To avoid making specious claims based on observed straight-line behavior in size-frequency plots, one just has to plot the same data cumulatively; i.e., consider plots where the x-axis shows the ranked values (e.g., smallest value first, largest value last) of the variable in question and the y-axis gives again the frequencies with which the different values occur. Simply examining the resulting "rank-frequency" plots (on double-logarithmic as well as on semi-logarithmic scales) is a significant improvement. Before applying a particular technique it is important to know the extent to which the statistics can vary as a function of the degree of imperfections present in the data. The bounds of biases in the results can often be explored by manipulations of the measurements and in-depth knowledge of the root causes of the errors/imperfections in the data. Sensitivity and bias knowledge will improve the meta-data of the dataset. The paper must provide sufficient evidence that the results based on the statistics are *not* artifacts of the measurements to meet the last of our **S**-rules.

2.3. Modeling efforts

Typical network-related modeling work accepts a given dataset blindly, often infers some first-order distributional properties of the data and determines the "best-fitting" model (e.g., distribution, temporal process, graph) along with parameter estimates. A visual assessment of the quality of the fit or an apparently more objective evaluation involving some commonly-used goodness-of-fit criterion is then done. The distributional properties of the data inferred is seen as reproduced in the model and thus the model is claimed to be valid. However, if the data often cannot be taken at face value, an accurate description (i.e., model) of the data at hand is no longer of interest.

We have to move past the simple and guaranteed exercise in data fitting. For the same set of distributional properties there are many diverse models that fit the data equally well. Models are often considered valid if they reproduce the same statistics of the data that played a key role in selecting the model in the first place! Both model selection and model validation through the same dataset poses serious statistical problems.

Our radical suggestion is to make matching particular statistics of the data a non-issue and eliminate the arbitrariness associated with determining which statistics of the data to focus on. Next, we seek to carefully examine the model 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. New implies entirely new types of data, with different semantic content, that have not played any role in the entire modeling process up to this point. The resulting measurements are only used for the purpose of model validation.[1] Such a statistically clean separation between the data used for model selection and the data used for model validation is alien to most of today's network-related models. This brings us to the modeling related question and the keywords covering the corresponding modeling rules: *Does model validation reduce to showing that the proposed model is able to reproduce a certain statistic of the available data, and if so, what criteria have been used to rule out alternate models that fit the given data equally well?*

The **M**-rules are:

[1] This re-iterates the "closing-the-loop" argument in [62].

1. Explain your model selection criteria.
2. Detail your model validation effort.
3. Provide details of the predictive power of the chosen model(s).

The **M**-rules try to ensure that modeling approaches respect the designed nature of the system, the engineering intuition that exists about its parts, and are fully consistent with available measurements (e.g., see the first-principles approach to modeling the Internet's router-level topology described in [41]). Just as the analytic techniques discussed above, the produced models must have strong robustness properties against the known shortcomings of the data. Being insensitive to the conditions under which the data was collected, its size, and duration is essential. As discussed in [61], this is especially important in situations where the size of the data or duration of the data collection effort are somewhat arbitrary and hence should play no role in the model selection process—having access to more or less data should primarily impact the confidence intervals associated with the estimates of the model parameters, but not the choice of the model.

3. Evaluation

We now present actual papers as exemplars of our examination of standards of measurement. We try to cover a reasonable span of areas choosing datasets that are reasonably well known and have had somewhat significant impact. The areas chosen typically involve datasets we are familiar with – we were either consumers (Section 3.1), interested observers (Section 3.2), or original producers (Section 3.3) – and a number of the papers discussed below include one or more of the authors of this study as co-authors. Our goal is to provide the readers with concrete guidelines of how to carry out a similar analysis in their area of interest. Our goal is **not** to discredit any of the papers or authors cited but to use specific aspects of their work as illustrations of the usefulness and appropriateness of our list of questions in search for improved standards for measurement-driven networking research.

We explore two different paths to evaluate our set of questions. The first one is a view from a topic point of view; in terms of how measurements in a particular important area have been carried out over the years and the impact of primary datasets. We chose two areas: topology modeling and wireless. Topology modeling is one of the most studied areas with multiple datasets and approaches, and one that has spawned numerous sub-areas of research in routing and architecture. It is important to note that obtaining accurate Internet connectivity-related measurements is generally hard except for those researchers who have access to large ISPs. Wireless was chosen due to its dramatic increase in importance just in the last few years.

The second path that we take is to pick a popular dataset and do a forward traversal tracing all the reuse of that dataset. The dataset in question has been reused in over a hundred publications. Although we don't examine all of the papers, we select a subset among them as good and bad examples of how well they have reused the data and made proper inferences.

Finally, we use an evolving new area, that of Online Social Networks (OSNs), as a different kind of example. As this area is still in its early stages, our intent is that our proposed rules can have a prescriptive value as measurements and analyses on OSNs are carried out. It is thus discussed separately in the next section.

3.1. Internet topology modeling

Internet topology modeling has been a very active area of measurement-based research for more than a decade, and during that time it has spawned numerous sub-areas of research in routing and architecture. Much of the published work in this area relies on a few publicly available data sources that have resulted from a small number of large-scale measurement efforts, which in turn have deployed either of the following two measurement techniques: `traceroute` or BGP table information. While the datasets typically depend on the date and size or extent of the measurement study, the key features of these measurement techniques have largely remained the same. This makes them interesting examples for examining their original use and reuse in some of the seminal subsequent studies.

One such case study concerns the use of `traceroute`-based measurements for inferring and modeling the Internet's router-level topology and is described in detail in [33] (see also [59]). It demonstrates why in view of the **P**-rules, the original measurement and data collection effort by Pansiot and Grad reported in [51] is a commendable early example of a paper in the area of measurement-based Internet research that provides a thorough and very detailed meta-data description and has stood the test of time. In particular, [51] states as explicit purpose for collecting this dataset a desire *"to get some experimental data on the shape of multicast trees one can actually obtain in [the real] Internet ..."* and says nothing about its use for inferring the Internet's router-level topology. In this sense, [51] shows why in terms of the **C**-, **S**-, and **M**-rules, some of the seminal papers in this area (e.g., [23] and [1]) have become text-book examples of how errors can add up and produce completely unsubstantiated claims, even though they may look quite plausible to non-networking experts. In fact, by consulting the meta-data description given in [51], applying the **C**-rules highlights some basic limitations that prevent a `traceroute`-based measurement effort from revealing the Internet's router-level connectivity to any reasonable degree. In a nutshell, and as discussed in more detail in [59], what makes the available `traceroute`-based measurements in general useless for inferring router-level connectivity are: (i) systematic errors due to an inability to resolve IP aliases and trace through opaque Layer-2 clouds; (ii) potential bias caused by oversampling some nodes while undersampling others; and (iii) inherent difficulties caused by the limited numbers and locations of vantage points from where `traceroute`-probes can be launched. In view of this, it is very unfortunate that starting with [23], the meta-data description provided in [51] has been largely ignored and forgotten; in fact, the majority of later papers in this area typically only cite [23], but no longer [51]. Although such secondary citations are a well-known problem, as our example demonstrates, 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.

A second case study discussed in [33] involves the BGP-based measurements and their use for inferring and modeling the Internet's AS-level topology. The original datasets are tied to an organization called *The National Laboratory for Applied Network Research (NLANR[2])*, an early NSF-funded effort to characterize the behavior of high performance connection networks. The NLANR project relied on full BGP routing tables collected by the *Route Views Project at the University of Oregon*[3] for the clearly articulated original purpose – "*to respond to interest in the part of operators in determining how the global routing system viewed their prefixes and/or AS space.*" Here, the relevant datasets come with essentially no meta-data information that would help subsequent users in deciding whether reusing these datasets for some alternative purpose such as inferring the Internet's AS-level topology is justified. As such, the burden of proof rests solely with the researchers who use the data for this purpose. Unfortunately, the early seminal papers in this area (e.g., [23] and [1]) have advocated an "as is" use of these BGP-based datasets, even though readily available domain knowledge says otherwise—BGP is *not* a mechanism by which networks distribute their connectivity, but instead, 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. As discussed, for example, in [49, 6], using these BGP data for the purpose of inferring and modeling the Internet's AS-level topology is completely unjustified due to the high degree of incompleteness, inaccuracy, and ambiguity that the data exhibit and impacts all aspects of a careful investigation of the Internet's AS-level connectivity structure. Recent studies have also shown that this problem cannot be rectified by augmenting BGP-based studies of the AS-level Internet with the available `traceroute`-based measurements [15, 68].

These observation show why domain knowledge in the form of `traceroute-` or BGP-specific "details" matters when dealing with issues related to data hygiene, statistical rigor, and model validation. Both case studies are also perfect examples for illustrating that via a combination of our **C**-, **S**-, and **M**-rules, the main sources of errors and their cumulative effect can be largely eliminated. However, the efforts to succeed in this endeavor can be expected to be significant and will typically require (i) developing an alternative modeling approach that makes good use of the available datasets despite their known shortcomings and limitations, (ii) no more arguing for the validity of a proposed model simply because it is capable of matching a particular statistic of the data, and (iii) putting forward substantial and convincing validation arguments and procedures (e.g., see [21]). For a related example involving un-sanitized vs. sanitized BGP data, see the discussion in [17] and [65].

[2]http://www.nlanr.net

[3]http://www.routeviews.org

3.2. Measurements of Wireless Networks

There has been a dramatic increase in measurement of wireless networks in recent years. The considerable resources required to establish and maintain a measurement infrastructure for wireless networks has resulted in many studies of wireless network characteristics reusing data collected in previous studies. Development of a Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) [30] has helped address this demand. This section illustrates how our questions and proposed rules can inform measurement-based research activities in the wireless area and how the wireless domain may contribute to a broader interpretation of our questions and rules. To this end, we consider the collection and reuse of the most popular dataset in the CRAWDAD repository: the data collected at Dartmouth College [31].

3.2.1. Production of a Wireless Dataset

Wireless networks pose many challenges to network measurement. These include interference caused by other wireless networks and the importance of spatial characteristics such as the location of users, buildings and access points (APs). These challenges, as well as ambiguities and limitations of measurement techniques, need to be addressed by producers of wireless datasets. Measurement studies such as [25, 29] illustrate the rigor needed when measuring wireless networks, and we focus here on [29] to check the relevance of our **P**-rules in the context of the production of a wireless dataset that has been reused numerous times.

While [29] does not explicitly demonstrate that the produced data is of sufficient quality for the purposes for which it is used, the study goes to great lengths to ensure the accuracy of the measurements. This is done by developing a wired-side methodology that combines SNMP and `syslog` measurements. By periodically polling the APs using SNMP, the authors are able to gather information on the amount of data transferred by each AP as well as the list of cards currently associated with the AP. Since relying on SNMP polling alone would place limits on the granularity of the mobility information, the authors also use `syslog` to monitor mobility of the wireless clients. Syslog data was gathered by configuring the access points to send `syslog` messages every time a card authenticated, associated, reassociated, disassociated, or deauthenticated (definitions in [29]). As a result, the authors were able to collect much more detailed information on the interactions between client cards and the APs than would have been possible if they had only used SNMP. When placing network monitors running `tcpdump`, the authors were not able to place them so as to capture packet-level traffic for the entire wireless network (due to the configuration of the network). To avoid bias, the authors attempted to place monitors in buildings that would be representative of a wide variety of campus users (e.g., dorms, library, student center). Wireless-side monitors would have been an alternate way to avoid being limited by configuration of the wired network.

Choosing a combination of measurement techniques (SNMP, `syslog`, and `tcpdump`) supports by and large the authors' arguments that their findings are valid and not simply artifacts

of the available measurements, despite some limitations of the measurements which the authors describe in detail. Specific limitations include ambiguities which arise when using a MAC address to identify a user, and holes in the data which may introduce bias into results. These limitations are discussed in the papers that characterize this dataset [25, 29] as well as in the CRAWDAD repository [31]. Documenting such limitations of the data collection effort can benefit both future consumers of the dataset as well as future producers of wireless datasets. Specifically, the authors notice frequent association events in the `syslog` datasets. These are caused by network cards aggressively searching for the best signal. While making note of such behaviors may seem tedious and orthogonal to the characteristics the authors sought to measure, this information can be used by both consumers of their `syslog` data and others who may use `syslog` to collect data in the future. Other limitations not mentioned in [29] include the absence of concurrent RSSI measurements that could have helped to understand the aggressive searching of APs. Also not documented are the power settings for the different APs, preferably with power maps. Such power maps constitute critical meta information for wireless datasets that consumers could use to select appropriate data.

In terms of the statistical analysis of the data produced in their measurement study, [29] relies predominantly on simple statistics such as CDFs and histograms and takes care to minimize the impact of the noted limitations of the measurements. Where the quality of their data is questionable the authors take care not to over-analyze. Specifically, the frequent card associations in the `syslog` data affects their observations of user sessions causing them to observe a large number of short sessions. This limitation is noted in the discussion and knowledge of this artifact in their data enables the authors to draw appropriate conclusions about session behavior (such as stating that sessions tend to be very short). When considering traffic per day and per hour, error bars are used to illustrate the variation between daily and hourly measurements. In this sense, [29] is an example that adheres to our prescribed **S**-rules without shedding new light on their interpretation or possible limitations.

3.2.2. Reuse of a Wireless Dataset

Data collection at Dartmouth continued long after the original study was published and the majority of this data has been made available to other researchers. The dataset now includes more than 5 years of data collected from the campus WLAN at Dartmouth College. Trace data that has been made available includes SNMP, `syslog`, and `tcpdump` traces. These provide information on data transfer of wireless cards and access points, interactions between wireless cards and APs, and packet headers, respectively.

This data has been made available on the CRAWDAD repository which provides methods for data hygiene related tasks. Specifically, a meta-data format is provided where authors of datasets can provide detailed information about the environment, network, methodology, sanitization, and other relevant features that impact the measurements. For example, information about network deployment can be especially beneficial when determining if the measurements are appropriate for reuse in another situation. Recently, a method for evaluating the completeness of wireless traces after the initial trace collection has been developed in [57]. This work is an example of relevant meta-data being elicited from measurements after they have been produced.

The data collected at Dartmouth College has been a valuable resource for researchers in diverse areas of wireless networking. This data has been applied to a wide range of topics including congestion control at APs [9], network security [55], and delay tolerant networking (DTNs) [14, 28, 38]. Such widespread usage underscores the need for consumers of data to ensure that the data they use is indeed stretchable to their desired application and state any assumptions made when applying data to a new domain. Stretchability in the wireless domain is affected by several factors including when and where the measurements were made (e.g., wired-side vs. wireless-side), the type of network technology (e.g., WLAN vs. Bluetooth), and types of access devices.

For example, one of the most popular measurements from the Dartmouth dataset has been the `syslog` traces. These traces have been used for many studies where information about user mobility is required. While many studies use the `syslog` data in the context of user mobility in a WLAN (e.g., [9, 55]), an interesting application of this data has been in the field of DTNs [14, 28, 38]. The application of the Dartmouth dataset to DTNs is an example of a wireless dataset from one domain being *stretched* for use in a different application and, as part of our **C**-rules, begs for an explanation. To illustrate, we focus on one of these DTN studies that uses the data from Dartmouth college [14]. In addition to the **C**-rules, the **S**- and **M**-rules also apply to this particular study where the authors focus on characterizing the time between contacts of the pairs of devices (inter-contact time) and use several datasets in addition to the Dartmouth dataset. These datasets included a second WLAN trace and a trace of Bluetooth-enabled PDAs. Additionally, new measurements using iMotes were made.

The measurements from the Bluetooth-enabled PDAs are clearly applicable to the study of DTNs as the traces show when the PDAs were in range of each other. However, the traces of WLAN mobility needed to be converted into mobility patterns in an ad hoc network. To make this conversion the authors assume that clients within range of the same access point could potentially connect with each other. This conversion has three main limitations that the authors enumerate. The conversion can be optimistic in the case of clients that are at opposite ends of a cell who may not be able to connect with each other. It may also be pessimistic for clients that are in neighboring cells who may actually be close enough to make a connection. Finally, laptops are the most common device used in the Dartmouth WLAN trace [25]. The type of mobility observed with a laptop which is not always with its owner and powered off at times, may differ from the mobility observed for PDAs and iMotes which are generally always with their owner. In this sense, the stretchability of the Dartmouth dataset to the DTN domain remains somewhat questionable and would require further investigation. A similar conclusion is reached when examining the stretchability of the Dartmouth dataset to the study of

congestion control at APs [9], but the arguments are different and involve the simple scaling up or down of observed total offered load that ignores the reactive nature of end-to-end TCP connections that make up the total load.

Our proposed **S**-rules come into play when examining the statistical analysis of the inter-contact times in DTNs [14]. The complementary cumulative distribution function (CCDF) used for characterizing the tail of the inter-contact time distribution across the datasets can be affected by the quality of the measurements at hand. Specifically, the granularity and the duration of the measurements impact the low and high values of the distribution, respectively. The authors discuss these issues in relation to their analysis. The inter-contact time distribution is also a statistic which may be sensitive to the type of network considered. This makes analyzing the inter-contact time distribution of the WLAN traces problematic if the desired application is, for example, mobile ad hoc networks of iMotes or PDAs. Leveraging the different types of datasets allows the authors to observe the differences between these networks. They are able to observe that while the value of the inter-contact time is sensitive to the network type, robustness is observed in the tail characteristics of the inter-contact time distribution.

The authors also present a model for the inter-contact time in DTNs based on their datasets. They propose that the distribution of the inter-contact times is heavy tailed and decays more slowly than the exponential distribution proposed in previous studies. This trend is observed across all datasets, with different parameters for the WLAN and iMote datasets. However, it is unclear whether the observed differences in the parameter estimates are genuine or a result of the limited quality of the underlying measurements. The fact that the model selected is able to capture behavior between the various datasets despite different access devices and data collection methodologies makes a convincing case for the model selected. There are, however, some weaknesses in the modeling approach taken in this study. For instance, the primary motivation behind the model selection is finding a model that is able to reproduce characteristics of the observed data rather than finding a model that is able to capture the underlying behavior that generates the distribution.

3.3. ClarkNet dataset

Next, taking a different path, we explore the use and reuse of a specific trace collected 15 years ago. After presenting the origins and motivation for the trace we examine a subset of the subsequent studies that used the dataset.

3.3.1. Background

One of the first Web server workload characterization studies ([4]; presented in June 1996) examined six different Web server workloads varying in intensity and duration. The primary contribution of the paper was the identification of ten characteristics common to all of the datasets. A challenge for the 1996 study was obtaining appropriate datasets. This experience, coupled with requests from other researchers, motivated the authors to make these datasets publicly available. They obtained permission to release four of the six datasets used in their study (Calgary, ClarkNet, NASA, Saskatchewan) and made them available in the Internet Traffic Archive [26] in April 1996. Here, we focus on consumption of the ClarkNet dataset, as it had the most intense workload (measured by average request rate) of the four datasets.

3.3.2. A History of Consumption

After an extensive search, we identified 139 research publications that utilized the ClarkNet dataset. These included 112 papers in workshops, conferences and journals, three books, eight theses, six technical reports and ten papers in non-English venues. Figure 2 shows the breakdown of peer-reviewed papers over time. The original authors used the ClarkNet dataset in four different papers (including [4]) between 1995 and 1997. The first use of this dataset by *other* authors occurred in 1997, and surprisingly, has continued through 2010, when the ClarkNet dataset was more than 15 years old.[4] Although use peaked in 2003 and has generally been declining since, the second largest use occurred in 2007. The 108 papers written by other authors were published in 90 unique venues; some authors wrote multiple papers with some venues publishing multiple articles. Roughly a quarter of these papers were published in non-systems domains (e.g., Artificial Intelligence, data mining, software engineering).

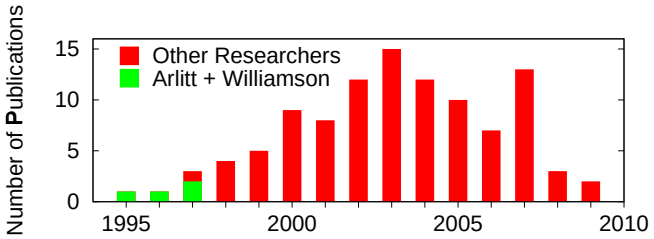

Figure 2: Publication timeline for ClarkNet dataset.

3.3.3. Observations and Implications

With over 100 papers reusing the ClarkNet dataset, we picked a subset and provide examples of how the authors could have benefited by answering our questions. We start by examining the adherence to the **P**-, **C**-, **S**- and **M**-rules.

Gathering and reviewing more than a decade's worth of research publications that utilized a familiar dataset provides additional insights from the producer's perspective. First, the consumption of the data as shown in Figure 2 lasted much longer than expected, and one can only speculate about the underlying reasons. Second, the data was used in a much broader range of venues and domains, and by a large number of researchers. Both of these reasons support the **P**-rule requirements for thorough documentation of meta-data about the dataset. As at least some fraction of the consumers will be affected by the weaknesses of the data, alerting them to known weaknesses would

[4]The use of the dataset has ironically outlasted ClarkNet itself as all ClarkNet products and services were sold off or dismantled by 2003.

be beneficial. Some of the meta-data may be forgotten over time if it is not documented.

With the ClarkNet dataset, the **P**-rules were followed to a degree. This happened by producers completing the template Web page at the Internet Traffic Archive for the dataset. However, the **P**-rules suggest additional meta-data, which in hindsight may have been useful to some consumers. For example, the measurement technique used in this case was simply the gathering of *access logs* from the ClarkNet Web server. A missing piece of meta-data is the version of Web server used to collect the logs. This might have been useful for tracking any bugs discovered in the logging mechanism, which might have affected the collected data.

Perhaps a more significant observation is that the meta-data may need to be revised over time, as more is learned about the dataset. For example, the meta-data on the ClarkNet dataset only alerts (potential) users to [4]; in 1997 an extended version [5] of this work included knowledge that had been determined by the authors such as limitations of the datasets. Simply tracking use of the dataset (e.g., a wiki that allows users to list their own publications that use the dataset) would assist researchers in learning of any additional meta-data discovered by others. The contributors of datasets will be able to see the tangible benefits of making datasets publicly available. A third insight is that the meta-data should be packaged with the data (in addition to being available on a Web page). Numerous publications indicated they retrieved the dataset from a location other than the ITA; it is unclear if these sites that replicated the dataset also replicated the meta-data. This issue also arises when the data is exchanged directly between consumers.

As for the **C**-rules, we provide an example of inappropriate re-use of a dataset. The ClarkNet dataset does not contain a definitive identifier for distinct users. However, some studies assume that there is a one-to-one mapping between a client IP address and a user. In the ClarkNet dataset, there are 142,993 unique fully qualified domain names (FQDNs) and IP addresses. Of these, 69 have the term 'proxy' in the FQDN (less than 0.05%). However, when the number of requests per host identifier is considered, 31 of the 100 busiest hosts have the term proxy in their name (31%), as do 24 of the top 30 hosts (80% - all from AOL). This suggests that care is needed in utilizing this data set for studying user behavior. Nanopoulos et al. [48] use the ClarkNet dataset in a comparison of prefetching algorithms. They indicate that a first step in preparing the data is the identification of user sessions, and refer to Cooley et al. [16] for the method to do this. Cooley et al. [16] correctly identify the presence of proxies as an issue to address in the identification of user sessions. They provide two methods (use cookies or client-side agents) and two heuristics (based on user-agent or referer header information) for distinguishing between users that are utilizing the same client machine. While all of these are valid in general, none of them are applicable to the ClarkNet dataset, as it does not contain per-user identifiers like cookies, nor does it contain user-agent or referer headers. Thus, the Cooley technique would not have correctly identified all individual users in the ClarkNet data, and therefore the dataset should not have been used in [48] (unless the issue could have been addressed utilizing a different technique).

Reflecting on the **C**-rules, several issues arise. First, consumers need to be disciplined in their use of meta-data. For example, consumers should maintain the original label associated with the dataset, to ensure that readers (or reviewers) are aware that a dataset used in one study is the same as in other studies that used the same dataset. The distributors of the ClarkNet data labeled it as ClarkNet; most consumers maintained this label but a few referred to it as C.Net, CNet, or Balbach. Also, researchers who use a dataset multiple times should apply the **C**-rules every time they utilize the dataset, to aid in avoiding problems encountered in one study from contaminating follow-on studies (particularly if one or more new participants are involved). Finally, since some venues publish multiple papers that use the same dataset, we suggest that reviewers should also apply the **C**-rules in their reviews (e.g., refer authors to a particular rule that they have failed to meet). One reason for doing this is that the reviewers are often in a better position to assess the longevity of the dataset for that particular domain than are the producers of the dataset.

Among the **S**-rules, the third rule, that of "checking results for consistency", is perhaps the most important here. We illustrate this by considering several papers that used the ClarkNet dataset that we have insight into. The authors of [4] indicated that the analysis techniques used in the paper were suitable and robust (the first two **S**-rules). However, subsequent studies suggested several improvements with respect to the analysis. A simple example is that with high-variability data (such as the file sizes in the ClarkNet data), the mean is largely uninformative. Instead, it was suggested that the median be reported, as it is more meaningful than the mean and also more robust to inaccuracies in the data. This is an example of how additional scrutiny (the third **S**-rule) would have improved the quality of the analysis results in [4]. Another study that deals with checking for inconsistency and appeared in subsequent papers published by others is by Downey [22] who re-analyzed the file size distribution in the ClarkNet dataset. He concluded that the evidence to support the Pareto model (as reported in [4]) is "weak and mixed," and suggested the lognormal distribution as a more appropriate model for file sizes. However, as discussed in detail in [61], favoring the higher-parameterized lognormal model over the parsimonious Pareto model comes at the cost of extreme sensitivity of the lognormal parameters to the size of the dataset (i.e., duration of data collection) which seriously questions the usefulness of the lognormal alternative in practice.

Lastly, we consider two examples for the **M**-rules. Use of an autocorrelation model to model Web server traffic is suggested in [43]. The work determines the model parameter settings by analyzing Web server traces (including ClarkNet). The model for each workload is validated by comparing the mean square error between the empirical autocorrelation function and the theoretical autocorrelation function of the model. Applying the **M**-rules reveals that the paper is strictly an exercise in data-fitting, demonstrates little creativity in building the model, does not demonstrate the predictive power of the model, and validates the model against the data used to parameterize the

model. In effect, the **M**-rules question the main purpose of this particular modeling effort.

Another modeling paper [10] extended an existing multifractal model, to reduce the complexity of the model from O(N) to O(1). The predictive power of the model is demonstrated, by comparing its accuracy in choosing files to cache against two other existing models. To validate their model, they examine how accurately their model captures the temporal and spatial locality of the empirical data. This paper more closely adheres to some of our **M**-rules. However, since no model selection criteria are provided and alternative models that fit the data equally well are not considered, the validity of the proposed model remains questionable, especially in the absence of any meaningful and network-centric explanation.

4. Discussion

Having presented three example evaluations across diverse areas, we now illustrate how our Socratic method for evaluating measurement-based research applies in a new and emerging area. We use Online Social Networks (OSN) as an example area for several reasons. For one, OSNs have recently, dramatically gained in popularity with a corresponding increase in interest in measuring them. OSNs are now the most popular application since the World Wide Web began in 1992. Users are first class objects in the sense that they are the primary creators of content and a significant part of the communication in OSNs stems from interactions between users. Along with Web 2.0 technologies (such as AJAX and mashups), thousands of OSNs have sprung up including MySpace and Facebook which recently reached a user base of half a billion users. A large number of external applications use the distribution platform of OSNs to enable new forms of inter-user interaction [18].

Moreover, given that we are still in the early days of OSNs and OSN research, we fully expect to see rapid and possibly drastic changes in the design and functionality of future OSNs. Thus the predictive value of initial measurement studies are not likely to be very high, unless they are accompanied by useful meta-data information. To this end, readily available domain knowledge about the design and operation of OSNs ought to guide measurement efforts in this area. However, because of the newness of the field, there is still a lack of any organized effort to collect OSN-specific data, and the number of consumers of such data has remained small. Since this situation can be expected to change with time, there exists a unique opportunity to start a dialog on establishing proper meta-data in this domain and ensure that any data released will meet some basic criteria. Thus as the number of publications in this area increases, our expectation is that there will be a prescriptive value in applying our Socratic method.

4.1. The **P**-*rules and OSN measurements*

Two of the most popular techniques that have been used to measure OSNs are active crawls [45] and passive measurements in the form of packet traces [24, 56] or click stream data [11]. OSNs do not expose their link structure partly due to legitimate privacy concerns, but also in a deliberate attempt to prevent external crawlers from gathering the connection matrix of the OSN. Active data gathering runs into limitations in the form of acceptable use policies and restrictions on the number of requests. Also, an active crawl that uses a particular technique (e.g., flooding, certain types of random walk-based crawling, sampling) to discover an essentially unknown structure is likely to miss portions of the OSN graph, especially the loosely connected regions. Passive data gathering also has limitations and will certainly miss information about users who did not communicate during the measurement period. Measurers of OSNs should provide necessary meta-data information to indicate the limitations imposed on their measurements as a result of the techniques used and policies encountered so consumers of their data could re-examine the techniques and policies at the time of reuse.

Although many OSNs provide an 'open' API for access to portions of their network, as yet there is no single API that can help gather data across multiple OSNs. In the absence of generic crawlers, most studies to date have been on a small scale. Crawlers have to parse and extract a wide variety of links: navigation, friend, group etc. In the presence of Javascript and asynchronous ability a crawler may have to simulate user clicks. One way to probe sites like Facebook that reveal only portions of the connection information is to create external applications via the OSN APIs that can collect anonymized data about users who use the application. As pointed out in [18], the community needs general purpose tools that can be customized to crawl and parse a particular OSN site. Such tools will expose commonalities across OSNs and highlight generic technical issues for measuring OSNs. Agreeing on a class of measurement techniques and tools will help future measurers in OSNs.

With regard to the measurement setup, gathering data in an OSN typically involves significant overhead in the form of gaining access to different portions (e.g., regional networks) of the unknown structure to study global patterns or derive results that are valid for the OSN as a whole. This is further complicated by the scale and differences between cultures, languages, and geographic regions. Moreover, as a recent privacy study [34] showed, what makes performing OSN-wide inferences even more difficult is the fact that the changes internal to an OSN are non-uniform; significant asymmetrical changes within regional networks in Facebook were observed within a two-month period. With the phenomenal growth in the number of users joining popular OSNs such as Facebook, we expect such changes to become both broader and even more non-uniform.

Many examinations of individual OSNs have been carried out [7, 35, 36]. These have included studies of properties like rankings, geographical popularity [13], object sizes, access patterns, rate of change [24], degree and cluster coefficient, and difficulty in finding backward links [45]. Properties such as connectivity, content, and technology are common to most OSNs and thus can be part of a comparative study [18, 36, 45].

There are many different ways to study OSNs. For example, studies have examined YouTube both from campus edge networks [24, 67] and using crawling techniques [45, 13]. More

concerning are seemingly minor differences in methodology that can lead to divergent results between studies. An early paper [27] on Twitter that tried to mine the words used in communication to extract communities and also examined the friendship relationship and different classes of users is an example of how the sample size and duration of the data can affect the findings. The underlying dataset consisted of a two-month long collection of random recent Twitter messages that is available in Twitter's public timeline. This passive data gathering was followed by fetches of friends information about the users. A subsequent study [32] which included two different active crawls, in addition to gathering the public data, paints a broader picture of the Twitter user graph. In particular, passive users are better represented in this study, as portions of the full graph may never have been discovered if they were not reachable from those who happened to be active during the earlier study. The effect of the dependence on only active users is a difficult parameter to estimate. A false inference about sequential growth of user IDs also creeps into [27] and was pointed out in [32]. Encountering such a diverse set of techniques used to measure OSNs stresses the importance of understanding how the gathered data might be affected as a result of the measurement setup and techniques. Ideally, all of the relevant information will be captured in the meta-data associated with OSN measurements, but if current datasets are an indication, we are still far from this ideal scenario.

Looking ahead and recognizing that dynamism is an integral part of most OSNs, the current crop of single-snapshot datasets is clearly insufficient. What is needed are multiple snapshots and associated meta-data information. Given the rate at which OSNs are evolving, meta-data attributes that are necessary so that plausible inferences can be drawn include the dates of the individual snapshots and the locations where they were gathered, the rate and manner of growth in user population and activity level, and timing information related to individual users or their activities. However, even in the presence of multiple snapshots, there are issues related to the meta-data and the quality of the data (e.g., missing events). Consider for example the recent work on various link prediction models [40, 64] that have been proposed to examine the evolution of OSNs. Meta-data about OSN-specific peculiarities and the potential for missing or inaccurate data can easily skew inferences. To illustrate, the methodology used to predict growth of friends in OSNs with symmetric friend relationships (like Facebook and MySpace) will not work for asymmetric OSNs like Twitter. On Facebook two users have to become mutual friends while on Twitter a large number of users can "follow" another user without the latter following any of them. Furthermore, OSN aggregators like FriendFeed [44] consist of only users who are present on multiple OSNs and are thus a skewed subset of OSN users.

Another topic where the current crop of single-snapshot datasets is limiting OSN research and where the availability of new semantic-rich OSN data is critical is inferring user interactions in OSNs. Clickstream data or packet traces (assuming they are made public) would be a perfect source. However, without a variety of additional attributes, such as user mix, local popularity of the OSN features, and nature of and reason for communication, inferences drawn could be incorrect. For example, it is well known that while two users may be "friends", the depth of their "friendship" is better reflected by the frequency and nature of communication which would typically not be present in packet traces. Thus to gain a basic understanding of how users or groups of users interact in an OSN will require information that can be gleaned from a combination of packet traces, clickstream data, and active crawls, and the "fusion" of these different data sources and corresponding meta-data information looms as an important open problem.

4.2. The **C***-rules and OSN measurements*

As in other areas of measurement-based networking research, producers of OSN-specific measurements are constantly being asked to make the crawled portion of the OSN graphs available and some have admirably done so already. At least two recent papers have made their datasets available: YouTube data in [13] and the crawled graph in [45]. The former's meta-data is better explained; the latter's anonymized data is likely to be less useful as it is just a description of the graph structure of their crawl. In abstract, the **C**-rules for OSNs are to ensure stretchability keeping in mind the key differences between the various OSNs. Similarities already observed between various OSNs at the macro level are a risky foundation for blind reuse. Data gathered in one OSN may be skewed due to the presence of certain features that are absent in the OSN to which the data is being applied. Data collected initially for the purpose of characterization is often a poor candidate for reuse as it is typically gathered in a single venue with a limited reflection of the overall distribution. The lifetime of early data is also limited in the fast changing OSN world. Given the considerable restrictions and other obstacles in gathering data in OSNs, any available data is likely to lack representativeness, and for any associated meta-data to be useful and informative, it must provide precise information about the collection methodology and any limitations in place at the time of data collection.

In general, there has been surprisingly little or no reuse of the data, and so statistical and modeling analysis from a reuse point is largely premature.[5] An important reason for this observed lack of reuse of OSN data is that current OSN research is slowly moving away from treating OSNs as static graphs and performing simple graph-based characterization of OSNs. Increasingly, researchers have recognized the need to look past just (static) friendship relationship and deal with dynamism as an integral part of real-world OSNs [60]. The evolving nature and observed structure of (some) OSNs have motivated researchers to focus more on issues relating to internals of OSNs and their distributed architectures, user interactions within and across OSNs, role and usage of external applications, new economic models, and algorithms that can cope with the large-scale nature and dynamics of OSNs. Clearly, for any in-depth studies of these and related issues, having access to a collection of generic nodes and links is insufficient. What these newer areas

[5] Authors of [45] and [13] were not aware of external publications that included reuse of their datasets.

of OSN research require are not just (static) friendship graphs but crawled data with a substantial amount of meta-data information that reflects the high semantic content associated with individual users and their activities within the OSNs [19]. However, in contrast to crawled data that results in generic friendship graphs, the type of crawled data required for these newer areas of OSN research has instantly raised serious privacy concerns that have effectively ruled out any reuse of such data by other researchers.

To deal with this problem and ensure the reuse and wider availability of such data, the topic of anonymizing evolving and annotated graphs has attracted recent attention. Initially the work was in anonymizing network data in the form of packet traces. It is useful to contrast anonymization of packet traces, where there have been considerable efforts [63, 50, 54] to the new ongoing work in OSNs. For payload-free packet trace data, the principle focus was to anonymize IP addresses. However, the absence of appropriate IP address information could negatively impact the ability to naturally group packets or recover the communication "graph" data, leading to work on prefix-preserving anonymization. However, in the OSN context, there are many more parameters that could result in re-identification. As recent work in OSN anonymization [12] shows, in the presence of analytic guarantees of privacy and anonymity, OSNs may be willing to release anonymized versions of snapshots and associated meta-data. It had been shown [8] earlier that attackers with background knowledge can learn information about some individuals on an OSN from an unlabeled graph by planting new nodes and linking them to legitimate users. Thus, we need to know the time of addition of nodes to distinguish original nodes and new ones. In a passive version of an attack, an adversary can learn about a large close-knit group and thus properties like stronger connections need to be known. Some defensive techniques to prevent re-identification have led to the use of adding and removing edges from the graph being anonymized. But the resulting graph will be different and may not be as useful to studying the same properties as in the original graph [66].

5. Conclusion

Early Internet measurement projects involving datasets of traffic-related quantities (e.g., packet traces, Web server workloads) have led to a general belief that Internet measurements are of high quality and that subsequent data analysis and modeling efforts can take the collected data at face value. However, more recent measurement efforts that concern Internet connectivity-related quantities (e.g., router-level connections, AS-level links) have highlighted the fact that in the Internet, it is more often than not the case that *what we can measure is in general not what we want to measure (or what we think we actually measure).* This realization has serious and wide-ranging implications, not only for the analysis and modeling of the resulting measurements, but also for the validation of claims that are derived from such data or the proposed models.

Motivated by an ever-increasing number of measurement-based studies in the area of Internet research, we have argued in this paper that it is time to examine how we can validate our research process; that is, *developing confidence that the results derived from [the measurements at hand] are indeed well-justified claims* [53]. A lack of specific standards has led to repetition of errors in various aspects of measurement-based networking research, and we have outlined a Socratic method to help correct this problem. As a first step we have proposed a set of key questions and rules for producers and consumers of data, as well as those who are involved in analysis and modeling efforts. However, we believe that trying to reach agreement on some basic standards requires a much broader effort than just our (likely biased) views and needs the involvement of the community as a whole to encourage an ongoing dialog between measurers, modelers, and experimenters. One of our long-term goals is to initiate and encourage a *community-wide* effort that tracks meta-data associated with different datasets that are gathered and reused in studies. Although we have not delved into the specifics of meta-data formats for different types of datasets here, we plan to do that in follow-up work or, better yet, look towards the community to discuss and adopt one.

There is no denying that raising the bar for measurement-based networking research creates more work. While maintaining adequate meta-data is especially important for rapidly evolving and changing systems such as the Internet for which the value of a given dataset is bound to change over time, in practice, this property should make researchers think twice before investing a lot of time and effort setting up accurate measurements of phenomena that may or may not exist over a longer period. Arguing for a more prominent role of the meta-data idea seems to strike a healthy balance between aiming for "perfect" data that may take an unreasonable time and effort to collect and may have only a short shelf time and producing "useful" data where the required effort/time is more commensurable with the data's generally short shelf life and typically limited usage.

Acknowledgments

We would like to thank the reviewers for their helpful comments. We also thank the anonymous reviewers of multiple Internet measurement conferences to which this paper was submitted—although they did not deem a critical discussion of papers that appeared in their and other venues suitable for publication, some of their comments were very constructive and helped to improve the paper.

References

[1] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, pages 406 (378–382), 2000. http://www.nature.com/doifinder/10.1038/35019019.

[2] M. Allman. What ought a program committee to do? In *USENIX WOWCS*, April 2008.

[3] M. Allman and V. Paxson. Issues and etiquette concerning use of shared measurement data. In *Proc. IMC'07*, 2007.

[4] M. Arlitt and C. Williamson. Web server workload characterization: The search for invariants. In *Proc. of ACM SIGMETRICS*, 1996.

[5] M. Arlitt and C. Williamson. Internet web servers: Workload characterization and performance implications. In *IEEE/ACM Trans. on Networking*, 1997.

[6] B. Augustin, B. Krishnamurthy, and W. Willinger. IXPs: Mapped? In *Proc. IMC'09*, 2009.

[7] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.

[8] L. Backstrom. C. Dwork, and J. Kleinberg. Wherefore art thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *Proc. 16th Intl. World Wide Web Conference*, 2007.

[9] P. Bahl, M. Hajiaghayi, K. Jain, S. Mirrokni, L. Qui, and A. Saberi. Cell breathing in wireless lans: Algorithms and evaluation. *IEEE Transactions on Mobile Computing*, June, 2007.

[10] A. Balamash and M. Krunz. Modeling Web requests: a multifractal approach. *Computer Networks*, 43, 2003.

[11] F. Benvenuto, T. Rodrigues, M. Chaa, and V. Almeida. Characterizing user behavior in Online Social Networks, In *Proc. IMC'09*, 2009.

[12] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Class-based graph anonymization for social network data. In *Proc. VLDB'09*, 2009.

[13] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes. In *Proc. IMC'07*, 2007.

[14] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *INFOCOM*, April 2006.

[15] K. Chen, D. R. Choffnes, R. Potharaju, Y. Chen, F. E. Bustamante, D. Pei, and Y. Zhao. Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users. In *Proc. CoNext*, 2009.

[16] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide Web browsing patterns. *Knowledge and Information Systems*, 1(1), 1999.

[17] J. Cowie, A. T. Ogielski, BJ Premore, and Y. Yuan. Internet worms and global routing instabilities. In *Proc. SPIE*, July/August 2002.

[18] G. Cormode and B. Krishnamurthy. Key differences between Web 1.0 and Web 2.0. *First Monday*, 13(6), June 2008.

[19] G. Cormode, B. Krishnamurthy, and W. Willinger. A manifesto for modeling and measurement in social media. In *First Monday* 15(9), 6 September 2010.

[20] M. Crovella and B. Krishnamurthy. *Internet Measurement: Infrastructure, Traffic, and Applications.* John Wiley&Sons, 2006.

[21] J. C. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. The "robust yet fragile" nature of the Internet. In *Proc. of National Academy of Science*, 102(41):14497–14502, 2005.

[22] A. Downey. Lognormal and pareto distributions in the Internet. *Computer Communications*, 28, 2005.

[23] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. of ACM SIGCOMM*, pages 251–262, 1999.

[24] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *Proc. IMC'07*, 2007.

[25] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Proc. of MobiCom'04*, pages 187–201, September 2004.

[26] Moderated repository of network traffic traces. http://ita.ee.lbl.gov.

[27] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *KDD*, 2007.

[28] E. Jones, L. Li, and P. A. Ward. Practical routing in delay-tolerant networks. In *Proc. of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, August 2005.

[29] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. In *Proc. of MobiCom'02*, pages 107–118, 2002.

[30] D. Kotz and T. Henderson. CRAWDAD - a community resource for archiving wireless data at dartmouth. http://crawdad.cs.dartmouth.edu, Apr. 2008.

[31] D. Kotz, T. Henderson, and I. Abyzov. CRAWDAD data set dartmouth/campus (v. 2004-12-18). Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/campus, Dec. 2004.

[32] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *ACM SIGCOMM Workshop on Online Social Networks*, August 2008.

[33] B. Krishnamurthy and W. Willinger. What are our standards for validation of measurement-based networking research? In *Performance Evaluation Review (Proc. HotMETRICS'08 Workshop)*, 36(2):64–69, 2008.

[34] B. Krishnamurthy and C. Wills. Characterizing Privacy in Online Social Networks. In *ACM SIGCOMM Workshop on Online Social Networks*, 2008.

[35] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW Conference*, 2005.

[36] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, 2006.

[37] S. Kurkowski, T. Camp, and M. Colagrosso. MANET simulation studies: The incredibles. In *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(4):50–61, 2005.

[38] J. Leguay, T. Friedman, and V. Conan. Evaluating mobility pattern space routing for DTNs. In *INFOCOM*, April 2006.

[39] W. Leland and D. Wilson. High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnections. In *INFOCOM*, 1991.

[40] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, 2006.

[41] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the Internet's router-level topology. In *ACM SIGCOMM*, 2004.

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

[43] M. Li, W. Jia, and W. Zhao. Modeling www-traffic data by autocorrelations. In *International Conference on Distributed Multimedia Systems*, 2001.

[44] S. Garg, T. Gupta, N. Carlsson, and A. Mahanti. Evolution of an online social aggregation network: An empirical study. In *Proc. IMC'09*, 2009.

[45] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. IMC'07*, 2007.

[46] J. Mogul. Brittle metrics in operating systems research. In *Workshop on Hot Topics in Operating Systems*, 1999.

[47] S. Moon and T. Roscoe. Metadata management of terabyte datasets from an IP backbone network: Experience and challenges. In *Workshop on Network-Related Data Management.*

[48] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. A data mining algorithm for generalized Web prefetching. *IEEE Transactions on Knowledge and Data Engineering*, 15(5), 2003.

[49] 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. In *Proc. of ACM SIGMETRICS*, 2008.

[50] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization. In *ACM SIGCOMM Computer Communication Review*, 36(1):29–38, 2006.

[51] J. Pansiot and D. Grad. On routes and multicast trees in the Internet. *ACM SIGCOMM Computer Communication Review*, 28(1):41–50, Jan 1998.

[52] V. Paxson. End-to-end routing behavior in the Internet. In *Proc. of ACM SIGCOMM*, 1996.

[53] V. Paxson. Strategies for sound Internet measurement. In *Proc. IMC'04*, 2004.

[54] B. Ribeiro, W. Chen, G. Miklau, and D. Towsley. Analyzing Privacy in Enterprise Packet Trace Anonymization. In *Proc, 15th NDSS*, 2008.

[55] S. Sarat and A. Terzis. On using mobility to propagate malware. In *Proc. of the 5th Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2007)*, April 2007.

[56] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding Online Social Network usage from a network perspective, In *Proc. IMC'09*, 2009.

[57] A. Schulman, D. Levin, and N. Spring. On the fidelity of 802.11 packet traces. In *Proc. PAM'08*, 2008.

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

[59] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the Internet: A source of enormous confusion and great potential. In *Notices of the AMS*, volume 56, pages 586–599, 2009.

[60] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. Research on Online Social Networks: Time to face the real challenges.

In *Performance Evaluation Review (Proc. HotMETRICS'09 Workshop)*, 37(3):49–54, 2009.
In *Proc. HotMETRICS'09 Workshop*, 2009.

[61] W. Willinger, D. Alderson, and L. Li. A pragmatic approach to dealing with high-variability in network measurements. In *Proc. IMC'04*, 2004.

[62] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. Scaling phenomena in the Internet: Critically examining criticality. In *Proc. of National Academy of Science*, volume 99, pages 2573–2580, 2002.

[63] J. Xu, J. Fan, M. Ammar, and S. B. Moon. On the Design and Performance of Prefix-Preserving IP Traffic Trace Anonymization, In *Proc. IMW'01*, 2001.

[64] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu. Scalable proximity estimation and link prediction in online social networks. In *Proc. IMC'09*, 2009.

[65] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. Observation and analysis of BGP behavior under stress. In *Proc. IMC'02*, 2002.

[66] E. Zeleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data *ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD*, 2007.

[67] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: Youtube network traces at a campus network - measurements and implications. In *IEEE MMCN*, 2008.

[68] Y. Zhang, R. Oliveira, H. Zhang, and L. Zhang. Quantifying the pitfalls of traceroute in AS connectivity inference. In *Proc. PAM'10*, 2010.