Computer Sciences Department

Internet Multi-Resolution Analysis: A vision and framework in support of representing, analyzing, and visualizing internet measurements

Paul Barford
Craig Partridge
Walter Willinger

Technical Report #1646

September 2008

UNIVERSITY OF
WISCONSIN
MADISON

Internet Multi-Resolution Analysis: A vision and framework in support of representing, analyzing, and visualizing Internet measurements

Paul Barford
University of Wisconsin
pb@cs.wisc.edu

Craig Partridge
BBN Technologies
craig@aland.bbn.com

Walter Willinger
AT&T Labs Research
walter@research.att.com

ABSTRACT

Empirical analysis has been the foundation for a great deal of network research and has resulted in significant improvements to Internet systems, protocols and practices. Recent progress applying a constantly expanding set of increasingly more sophisticated statistical tools suggests the emergence of a new type of empirical network research that could benefit from a more principled approach for representing, analyzing, and visualizing a wide variety of Internet-related measurements. To this end, we propose and introduce in this position paper the concept of an Internet-centric multi-resolution analysis (MRA). Internet MRA is a structured approach to Internet data representation and establishes a framework for systematically applying statistical analysis, signal processing or machine learning techniques to provide critical insights into a number of challenging network research problems. Ultimately, the success of Internet MRA will be gauged by its ability to solve important problems in network research and operations and in the new lines of inquiry that it enables.

1. INTRODUCTION

Over the past few years, researchers have started to use an ever wider set of analytical techniques to discover and extract information from network measurement data. Some of these techniques are based on traditional time series analysis (*e.g.,* [32])or rely on classical signal processing tools such as Fourier or Wavelet transforms (*e.g.,* [1, 21, 16, 49, 22, 5]). Others make use of more sophisticated statistical concepts such as multi-dimensional scaling in the form of Principal Component Analysis (*e.g.,* [46, 25, 27]). Still others have started to borrow more heavily from ideas developed in the fields of machine learning and data mining (*e.g.,* [18, 19, 44]). Finally, tools are emerging that combine these analytic techniques with visualizations that enable new types of exploratory analysis and support for network operations [45, 14, 34].

An exciting feature of these techniques is that they appear to be pertinent to a range of important challenges facing empirical network research. On one end of the spectrum, they can be used to extract information that is not readily visible on inspection of largely semantic-free network measurements (*e.g.,* encrypted traffic). On the other end of the spectrum, they can be used to discover structure in seemingly unlimited and unstructured, semantic-rich measurements (*e.g.,* full packet traces). For example, recent studies have demonstrated the ability to trace traffic flows in encrypted wireless networks with high reliability [10]. At the same time, there have been significant advances in effectively combining intra-AS routing information with intra-AS SNMP measurements to estimate the intra-AS traffic matrix at different levels of granularity [30, 56, 57]. Both examples represent a compelling demonstration of combining information from network measurements with operational expertise to move a known but seldom realized concept into daily networking practice. In fact, there is a sense that we are on the verge of developing a suite of tools and algorithms that will allow us to peer far deeper into the behavior and structure of networks than ever before.

The initial successes of new analytical techniques can inspire the community to push ideas as fast as possible toward the horizon. There is, however, also a critical challenge – the challenge of convincing the network research community to rigorously test, examine, and validate newly proposed techniques so that we can be assured that they do, indeed, perform as expected or claimed. We have at times in the past become enamored with new and sophisticated-looking techniques (for instance, multi-fractal analysis), only to find out through experience that they did not live up to expectations.

The objective of this paper is to sketch a research vision that allows us to make the best possible use of new analytic techniques and increase their relevance for empirical networking research, while avoiding their pitfalls.

The challenge in sketching such a research vision is that it is far from clear exactly which questions the emerging new techniques and algorithms will be best-suited/well-suited to answer. So, rather than focusing solely on the questions, we aim in this paper at outlining a formal framework and practical implications of what we call an *Internet-centric Multi-Resolution Analysis (MRA)*.

Internet MRA is a structured approach to Internet data representation, specifically designed to accommodate the vertical (*e.g.,* layers) and horizontal (*e.g.,* domains) decompositions of the Internet architecture. It also captures in a systematic manner the "multi-scale" nature of the temporal (different time scale granularities), spatial (different

IP address space granularities), and functional (different layer-specific granularities) aspects of traffic flows across the network. We will describe how Internet MRA technology facilitates dealing with, *(i)* multi-scale representations of very large and diverse Internet-specific graph structures, *(ii)* dynamic processes on these structures, and *(iii)* aggregated spatio-temporal-functional network data representations and their associated analyses and visual representations. In short, we advocate the development of a mathematically inspired, practically useful, and computationally efficient framework for choosing, developing, using, and validating newly proposed analysis tools and techniques for representing, analyzing, and visualizing network measurement data of all kinds.

To help provide perspective and context we begin by discussing trends in network measurements which are the starting point for Internet MRA. These highlight an urgent need for a coherent and self-consistent framework for representing, understanding, and processing network-related data. We then describe the basic principles of the proposed Internet MRA technology. Analogous to wavelet techniques, these are based on the ideas of coarse-to-fine data representations in conjunction with decomposition and reconstruction algorithms. Finally, we describe a set of Internet MRA target problems to illustrate the Internet MRA methodology, and to highlight the critical problems of choosing, testing, calibrating, and validating statistical and data mining algorithms for the purpose(s) for which they have been proposed.

Our success in these efforts can be assessed in several ways. In the short term, it will be reflected in attempts by the network research community to refine the notion of Internet MRA, provide additional showcase examples, and organize existing or new measurement efforts in a MRA-like fashion. In the longer term, success will be measured by the ability of Internet MRA to solve existing hard problems and to open up new lines of inquiry. Of course, only time will tell, however we believe that a threshold has been reached at which a strong case can be made for the broad utility of Internet MRA methods and the viability of Internet MRA as a research domain in its own right.

2. TRENDS IN NETWORK MEASUREMENT

Our research vision for Internet MRA is shaped by the types of network measurements that we expect will emerge and dominate in the near future. They include the following three categories: semantically limited network-wide measurements, semantic-rich network-wide measurements, and opportunistic measurements. The common features between each class of measurements include the potentially vast quantity of diverse data collected from multiple sites over time. In each instance there are unique challenges that must be addressed before the data can be used to its fullest potential in research and operations.

2.1 A comment on measurements

Scientific data are often imperfect or contain errors, and Internet measurements are no exception. Although a great variety of different measurement capabilities and measurements appear to be readily available to the research community, as discussed in detail by Paxson in [37], conducting Internet measurement studies in a sound fashion is in general not as easy as it might first appear. The challenge is to know whether or not *"the results we derive from our measurements are indeed well-justified claims"* [37], and at issue are the quality of the measurements themselves as well as the quality of their analysis. To prevent measurement errors from adding up and measurement imperfections from tarnishing subsequent analysis or modeling efforts, it is essential to first assess and determine the quality of the available data within the context of its intended use.

Consider for example the case of traffic-related Internet measurements. Given a measurement tool has been tested and works correctly under normal operating conditions, assessing the quality of the data that result from a particular application of the tool typically involves checking for *(i) accuracy, i.e.,* imperfections incurred when using the tool; *(ii) precision, i.e.,* problems that may be inherent in the basic design of the tool; and *(iii) misconceptions, i.e.,* potential mismatches between what the tool is supposed to measure and what it actually does measure [37]. For example, when collecting packet-level traffic measurements via a packet recorder (*e.g.,* TCPDUMP [47]), it is well known that accuracy can be compromised in a number of different ways. For one, the packet recorder may not be able to keep up with the rate at which the packet filter (*e.g.,* BPF or libpcap) accepts packets; or the filter may not keep up with the rate at which the network tap processes the raw packet stream; and last but not least, the tap itself may fail to keep up with the raw bit rate on the link. Each of these situations will give rise to "drops" leading to discrepancies between the actual and the recorded traffic. All trace collection efforts, past and present, are prone to this type of inaccuracy, as well as reordering and duplication. For researchers to make informed decisions about whether or not such data sets are adequate for the type of analysis they have in mind, it is paramount to know how often such (or similar) conditions occur, when they happen, and what the precision of timing-related measurements is (*e.g.,* timestamps may not be synchronized between data sets gathered at the same time at different locations).

Another example are connectivity-related measurements that are notorious for their ambiguities, inaccuracies, and incompleteness. Many of these can, at best, be described as being of "limited quality." This is true at the physical layer, where to date traceroute-based measurements have formed the basis for inferring router-level connectivity, even though traceroute was never intended to be used in this context (see for example the discussion in [2]). It remains true at the higher layers of the protocol stack (where Internet connectivity becomes more virtual), but for very different reasons. For example, as far as measurements for inferring AS-level connectivity are concerned, ASs are generally reluctant to disclose information regarding their peering relationships with other ASs and routing policies; they typically consider such information to be proprietary. This makes it practically impossible to measure AS connectivity and peering relationships directly, and requires the collection of alternative or "surrogate" measurements that are feasible and can shed light on the quantities of interest. For example, the measurements that are often used for inferring AS-level maps consist of BGP routing table snapshots such as those collected by the University of Oregon Route Views Project [51]. To illuminate the degree of ambiguity in the inferred AS connectivity data, note for example that due to the way BGP routing works, snapshots of BGP routing tables taken

at a few vantage points on the Internet over time are unlikely to uncover and capture all existing connections between ASs. Indeed, [8] reports that AS graphs inferred from the Route Views data typically miss between 20-50% or even more of the existing AS connections. This is an example of the general problem of *vantage point* mentioned in [37], whereby the location(s) of exactly where the measurements are performed can significantly skew the interpretation of the measurements, often in quite non-intuitive ways. Other problems that are of concern in this context have to do with ambiguities that can arise when inferring the type of peering relationships between two ASs or, more importantly, with the dynamic nature of AS-level connectivity, whereby new ASs can join and existing ASs can leave, merge, or split at any time. As before, before embarking on any analysis or modeling effort, it is paramount to understand the types of ambiguities inherent in these data, as well as their severity.

2.2 When information is scarce: Inference and encrypted traffic

The needs of many commercial and military applications drive the deployment and use of more secure forms of data communication in wired and wireless networks. The result is increasingly ubiquity of payload encryption in network packet traffic. While this traffic is typically straightforward to record and collect, measurements may provide little more than simple timestamps and sizes of the individual packets seen on one or more links. This would certainly be the case for traffic traversing an encrypted tunnel *e.g.,* via IPSec, which is common in virtual private networks. In the case of application-level encryption *e.g.,* via SSL, packet header information would also be available. Encrypted traffic traces are an important example of network measurements with low semantic content.

One vital problem is to determine how much information about the original applications can be extracted from such semantically poor data sets. This can be viewed as a straightforward security problem ("how much information is encryption successfully hiding") or as a more subtle network management question ("are the applications on the encrypted virtual private network getting adequate quality of service?"). In either case, the challenge is to develop techniques that enhance the data and enable recovery of as much information as desired or required to reconstruct traffic flows with high reliability and high fidelity. One approach is to rely on domain-specific knowledge or auxiliary meta-data that is available about the network at the time the measurements were collected.

Another vital problem is to understand how much or what sort of noise (*i.e.,* jitter) in measurements a given technique can tolerate and still be effective. Consider the example of applying Internet MRA techniques in a network management context. When the network becomes overloaded and there is substantial interaction between traffic, will the algorithms still be able to extract enough information to operate the encrypted network effectively? Similarly, understanding specific failure modes of Internet MRA algorithms will enable them to become the foundation of robust and predictable systems.

A third problem is having confidence in our results. Unlike the other measurement situations described below, we often cannot confirm our results by looking at the original data because that data is encrypted. So we need to take extra care to test our algorithms and their results so that actions based on those results can be taken with confidence that they reflect underlying reality.

2.3 When information is plentiful: Inference and network-wide measurements

The efficient management, secure operations, and effective control of a large Internet Service Provider or content distribution network requires a network-wide measurement infrastructure capable of supporting real-time decision making processes. Monitors in these infrastructures often include SNMP MIB-based systems, end-to-end active probe-based systems, and highly sophisticated hardware-based systems (*e.g.,* [24, 39]). Management decisions are ideally made by having a comprehensive perspective of an infrastructure which means collection and analysis of voluminous, high-dimensional data sets typically with very rich semantic content.

One challenge in this case is that even though available information is plentiful, the input data for some key decision making processes may not be directly observable or measurable but must be *inferred* or *estimated* from the right subsets of relevant measurements contained in the often huge sets of available network data. Thus, the problem becomes one of purposefully mining the network-wide measurements and developing techniques that ultimately provide the desired inputs to the tools and algorithms that the network operator considers essential.

To achieve this objective requires first and foremost a clear understanding of the available measurements. Most important is understanding what information the measurements can and cannot provide (with or without the use of domain-specific knowledge), and which class of techniques is appropriate for mining the desired information. The overall process can also be enhanced by understanding what sort of additional measurements would be feasible to collect in support of rigorous tool calibration and validation. As mentioned earlier, there is also the issue of data quality, and techniques that are more or less robust to the known deficiencies in the data are highly prized.

A salient feature of the proposed Internet MRA approach is its focus on effective data representations that ideally support an informative mapping of the data to the underlying architectural structure of the Internet. In the context of voluminous network-wide measurements, such a mapping typically reveals the measurements' value (*i.e.,* information regarding directly observable quantities versus quantities that need to be inferred) and quality, and identifies other measurement experiments that either complement the existing measurements or play a critical role in the validation phase.

2.4 When information is spontaneous: Inference and opportunistic measurements

A different type of semantically rich measurements arise when aspects of the network are measured at times when it is under stress or duress due to events such as hardware failures, software errors, misconfigurations, congestion, or nefarious intrusions and attacks. These "opportunistic" measurements [7] are often relatively easy to collect (once the decisions of what, where, and when to measure have been made), and have the potential of providing unique insights into the structure, dynamics, and operations of the Internet as a whole, typically well beyond the confines of what is vis-

ible to individual network operators. For example, it would have been difficult to predict the inability of many of the major news portals to handle the observed volume of traffic during the 9/11 attacks [43]. Challenging problems in this context concern the features of large-scale networks that can be discovered via opportunistic measurements, and the class of statistical/machine learning, or other analysis techniques that are amenable to these measurements.

3. OUR VISION: INTERNET MULTI- RESOLUTION ANALYSIS

Past experience has taught us that important aspects of the dynamics and structure of the Internet often manifest themselves in very different empirical phenomena that are observed in data measured at different scales in time and space. Also, while many analysis, inference and visualization techniques were originally developed for Internet data measured at a specific scale, they often led to methods that were more multi-scale in nature, ultimately enabling a more accurate analysis and a more comprehensive perspective.

It is clear that there are many challenges in understanding Internet phenomena at different scales, and we believe that these efforts would benefit from a more principled and rigorous approach. In the past, once the first demonstrations of a technique have been reduced to a set of exciting graphs, it has often been unclear what the next step should be. We argue that a "conceptual roadmap" would help make progress in this regard. Furthermore, practical considerations in dealing with massive and complex data sets from emerging/envisioned measurement tools suggest a set of requirements for future empirical analysis that even by themselves would be a tremendous asset.

In this section we suggest a framework in support of (*i*) providing effective data representations, (*ii*) developing efficient analysis and visualization techniques, and (*iii*) establishing rigorous calibration and validation procedures. To this end, we are inspired by the mathematically appealing properties of multi-resolution analysis (especially wavelet decomposition [13, 1]) and advocate development of a broader multi-resolution framework that can be applied to Internet data. The Internet MRA framework includes the ability to incorporate the multi-scale nature of the underlying physical network structure, and the generic ability to support aggregated spatio-temporal network data representations across protocol layers as well as the their analysis and visualization.

3.1 Basic Principles of Classic MRA

By providing general methodologies and mathematical tools for studying complex objects such as high-dimensional, semantic-rich data sets, or large-scale network structures, MRA is a technology that supports the representation and processing of *scientific* data. It has had enormous impact in many areas of science including Signal Processing, Computer Graphics, and Geometric Modeling. Assuming that the data sets are defined on some domain whose elements are indexed by a set of homogeneous or heterogeneous coordinates, it organizes them into strata of coarse-to-fine resolutions, where fast and compact, low-resolution descriptions that have approximately the correct large-scale behavior can be successively refined by adding finer details either locally or globally. Key principles of any MRA are *scale invariance* – in

the sense that algorithms are largely independent of the resolution level – and *efficiency* in the sense that it favors the design of very fast and robust implementation algorithms.

One of the successes of MRA research has been the development of a quantitative mathematical theory for comparative evaluation of a large class of MRA algorithms known as 'wavelet representations' [13]. A key reason for this success has been the fact that for data defined on domains whose points are indexed by a set of homogeneous, regularly-spaced coordinates (*e.g.*, in the traditional MRA for images or signals, the coordinates are typically pixel coordinates, or the time values in a time series of observations, respectively), the construction of appropriate coarse-to-fine resolution domains is relatively straightforward. More formally, the following three concepts figure prominently in any MRA construction:

1. Coarse-to-fine resolution domains,

2. Decomposition algorithms,

3. Reconstruction algorithms.

The first concept refers to a methodology for extracting from the data's original domain two new domains: the 'aggregation' domain, and the 'detail' domain. The key principle is that the 'aggregation' domain will be, loosely speaking, a coarser version of the original domain of the data. The latter is now split by the *same* methodology, thereby leading to a hierarchy of 'detail' domains complemented by 'aggregation' domains, where the number of scales or levels in the corresponding hierarchy of domains is flexible.

The two other concepts are complementary. By using a 'decomposition' algorithm, one derives from the data defined on the original domain two new data sets: one defined on the 'aggregation' domain and one on the 'detail' domain. Continuing iteratively (*i.e.,* decomposing the aggregate data further) the original data is decomposed into a hierarchy of fine-to-coarse 'detail' scales. The resulting representations enable consistent analysis by highlighting key features that exist in the data at each of the different resolution domains.

In contrast, a 'reconstruction' algorithm inverts the decomposition process by starting with the data defined on the 'detail' domains and reconstructing the data defined on the original domain. This process may be best understood in terms of *coarse-to-fine prediction/residuals.* Starting with the data at the coarsest level, this algorithm combines that data with the data from the corresponding level of detail resulting in "predicted" values at the new grid points of the next finer level. By carrying this prediction from level to level, one obtains a description on the original (finest) domain that is derived solely from the coarse approximation. The true object typically has more detail than given by this smoothed description. This means that going from one level of description to the next finer one, the values at the new grid points are in general not identical to those given by the coarse-to-fine prediction rule; the residual difference at every scale is given by the detail data stored at that level. Reconstruction is critical in MRA for validating the robustness of the decomposition.

It is important to stress that for a given data set, there are typically many different MRA decomposition/reconstruction algorithms each leading to a potentially different representation of the same original data set. The choice of the 'right' or 'correct' representation is often critical for the success of

a particular application. Unfortunately, finding the right representation for a given application and data set is often a time consuming trial and error process.

3.2 Towards an MRA for the Internet

The goal of Internet MRA is to apply the experience and insight from classical MRA in a broader context. This is done by viewing Internet data as being defined on some domain whose elements are indexed by a set of coordinates and organizing them into strata of coarse-to-fine resolutions. In this case, compact, low-resolution descriptions with approximately correct large-scale behavior can be successively refined by adding finer details either locally or globally. In this process, the key requirements of any mathematical MRA (*i.e.,* scale invariance in the sense that algorithms are largely independent of the resolution level, and efficiency in the sense that it favors the design of fast and robust implementation algorithms) may have to be reinterpreted to allow for efficient, but possibly scale-dependent algorithms that will exploit layer-specific aspects of Internet data. We now look at how that reinterpretation might be done.

The first challenge is in the structure of the data. Classic MRA assumes that data is indexed by a set of homogeneous, regularly-spaced coordinates or can be transformed in a straightforward fashion into such a representation. Unfortunately, Internet data are notorious for their highly heterogeneous and irregular domains. For example, "points" can represent geographical or topological router or link location; time; packet header information such as protocol type, source/destination IP address, source/destination port number; flow attributes; and other possible variables; or any combination thereof. General MRA representations applicable to data with such irregular domains are by and large non-existent, and even the construction of coarse-to-fine resolution domains is a formidable challenge, typically representing a significant research effort by itself.

Another challenge is classical MRA's focus on invertible decomposition and reconstruction algorithms as a method of validation. Irregular domains make invertibility difficult to achieve. For instance, one way to get irregular data into a form for MRA is to (re)sample it to make it regular – if we wish to use invertibility to check our result, we would like to invert back to the original (irregular) data but some information is likely lost/modified in the resampling. Validation is, of course, essential. For Internet MRA, however, we need to find acceptable methods of validation that do not depend on invertibility.

Given that the current state-of-the-art MRA technology cannot readily cope with the fascinating new challenges posed by the available and anticipated Internet data, our primary focus is on sketching an Internet MRA technology appropriate for dealing with multi-scale representations of large graph structures, dynamic processes on them, and for aggregated spatio-temporal network data representations across multiple network layers and the analysis and visual representations of them.

Whenever necessary, we favor hand-crafted representations that capture the underlying Internet structure more faithfully over alternatives that are more classically MRA-like from a strictly mathematical perspective. At the same time, we seek to adopt the mind set of classical MRA, with its emphasis on changing or adjusting resolutions in verifiable ways to provide the necessary intellectual focus.

3.3 A comment on algorithms

Before presenting details of our Internet MRA framework, a short discussion of the analysis algorithms is useful to set context.

The analysis techniques that are currently available for Internet MRA are generally statistical mechanisms. Many of these are familiar to anyone with a background in signal processing or numerical analysis, however others come from newer fields of study such as machine learning. A feature they all tend to share is that they *transform* their data, that is, they take input data in one domain and produce an output in another domain. One example is the Fourier Transform, which converts a time series into a frequency map. Another is Principal Components Analysis, which seeks to convert multiple, interrelated, measurements into a smaller, more coherent, space that reflects the underlying drivers of the phenomena being measured.

Three benefits of these algorithms are, *(i)* they are fairly robust to their input type (one could equally well feed the Fourier Transform a series of packet arrival times or an hourly count of the number of routing updates received and get extract frequency information), *(ii)* they are robust to a wide range of variability in measurement quality (it often tends to wash out as white noise), and *(iii)* they shine a brilliant light on the features they are designed to measure.

The biggest challenge in working with these algorithms is that they produce results in a transformed space where it is often difficult to determine or interpret what aspect of the measured space is reflected in the transformed result.

4. INTERNET MRA IN PRACTICE

One of the premises of the envisioned Internet MRA is that despite the Internet's size and complexity, it may nevertheless be possible to analyze and visualize it as a whole at suitably chosen coarse levels, and to zoom in on areas of specific interests and consider them in greater detail. A number of observations and findings obtained to date from analyzing a wide range of Internet data strongly suggest that thinking more carefully in terms of appropriate MRA representations can provide a powerful framework for illuminating the often rich information contained in Internet data, for developing novel techniques for uncovering this rich information, and for adhering to a rigorous standard when it comes to calibrating and validating the newly proposed techniques and algorithms. In many ways, this can be considered a road map for the "day in the life of the Internet" metaphor described in [42], which aims at a global, empirical understanding of Internet behavior. In the following, we illustrate these three topic areas with examples that highlight some of the key MRA concepts and principles.

4.1 Internet Data and MRA

Early Internet data gathering was heavily focused on packet traffic measurements, and typically consisted of high-precision packet arrival times and individual packet sizes (*e.g.,* [28]). Recorded on a single physical (uni- or bi-directional) link and at a particular layer in the TCP/IP protocol stack (link- or network-layer), this type of Internet data has formed the basis of classical studies of the temporal dynamics of Internet traffic, with the 1-D time series representing the number of packets or bytes per time unit attracting most of the attention. Dedicated hardware has traditionally been used to

gather this data, and current technology enables lossless collection at multiple points in a network over extended periods of time. For these data, the most simple notion of Internet MRA refers to classical wavelet decompositions that support the efficient analysis of the traffic rate process at a given link and a specific layer within the larger Internet structure across a range of different time scales.

When aggregated over sufficiently large time intervals (*e.g.,* 5 or 10 minutes), there exists operational support in the form of SNMP measurements to collect these 1-D time series from all the interfaces in an ISP's network. The resulting data provide a broad view of traffic activity on all links in the network. Together with detailed information about the network's physical infrastructure and its internal routing matrix, these 1-D time series (2 per interface, one for inbound traffic, one for outbound traffic) have been critical for recent work on intra-AS traffic matrix estimation which has immediate application in the network management domain [41]. Traffic matrices offer a first glimpse at the spatio-temporal behavior of Internet traffic within the confines of an ISP or AS, albeit at rather coarse time scales and strictly from a link-layer perspective. For these data, no known MRA technology is readily available, and they serve as an example that the development of a practically relevant Internet MRA demands new innovations, representing a dramatic widening of MRA technology as it is known and used today.

Aggregating even further in time (*e.g.,* over hours or days), but also in space, we can, in theory, consider objects such as the Internet's inter-AS traffic matrix – a coarse-scale global spatio-temporal view of how raw traffic volume (*e.g.,* bytes per day or week) is exchanged over the Internet at the level of individual ASs. The practical difficulties in obtaining such a global picture are that the Internet data necessary for estimating an inter-AS traffic matrix are either highly ambiguous or simply not available. An example is understanding AS-level topology as described in Section 2.1. At the same time, no measurement infrastructure at present is capable of collecting raw traffic volumes on peering links on a global scale.

4.2   Internet Data Analysis and MRA

For the 1-D time series representing the number of packets or bytes transmitted over a given link per time unit, elementary MRA technology in the form of simple time domain aggregation led to the discovery of the *self-similar scaling behavior of Internet traffic* over medium-to-large timescales [28, 38, 11]. More recently, wavelet-based MRA techniques have been successfully applied to the analysis of similar data sets and have contributed significantly to an improved understanding of the scaling properties associated with the temporal dynamics of Internet traffic as observed on a single link [32]. Critical to the success of these wavelet-based MRA techniques was the efficiency and compactness of the underlying large datasets whose domains simply consist of regularly-spaced coordinates on the time axis. In fact, because all of the different detail data can be organized along the same basic coordinate (*i.e.,* time), the original data can be reconstructed *selectively* by choosing detail coefficients only in areas of interest. This provides a concrete illustration of the powerful time-localization or adaptive "zoom-in" capabilities of wavelet-based MRA techniques. These capabilities have already been put to good use in the analysis of Internet traffic traces, where they have aided the detec-

tion and identification of structural properties of network traffic that are localized in time such as volume anomalies (*e.g.,* [3]). The various findings obtained to date from these applications of wavelet-based MRA techniques to the analysis of the temporal dynamics of Internet traffic have demonstrated that by relying on a combination of empirical studies and mathematical tools, the underlying data can be exploited to a degree where success is no longer measured by how well a technique can be used to describe the data in a statistical sense, but rather by how useful and accurate it is in seeking and providing *physical explanations* of observed phenomena and relating them to elementary networking mechanisms or concepts. In other words, establishing genuine cause-effect relationships for important observed phenomena is another key objective of Internet MRA.

Unfortunately, existing MRA technology cannot yet cope effectively with the more general situations that Internet data demand and where the basic coordinates along which the data are organized are more complex and heterogeneous. Consider for example the data set consisting of one days worth of SNMP data from every interface in an ISP's network. This data can be thought of as being defined on a highly irregular domain representing the ISP's physical network infrastructure and consisting of nodes representing routers (with their interfaces) and links representing physical connections between two interfaces (where each link is considered to be uni-directional and associated with its corresponding time series of SNMP measurements). To develop an Internet MRA of this data in support of traffic matrix estimation requires the construction of coarsened versions of the original data domain, preferably in such a way that each coarsened domain reflects some new graphical/spatial structure with either an intrinsic geographical (*i.e.,* metropolitan areas, regions, countries) or networking-specific (*e.g.,* router-, PoP-, AS-level) meaning.

4.3   MRA and Validation/Calibration

A compelling aspect of Internet research is that we often have the ability to measure characteristics of the network and thereby obtain data (at least in principle) necessary to support or refute claims based on either some empirical analysis or some proposed mathematical model. Here we emphasize how being rigorous about Internet MRA is essential for examining newly proposed statistical techniques so that we can be assured that they do, indeed, perform as claimed. Several recent papers outline important aspects of this issues, in particular [37, 52].

Consider again the case of the 1-D time series representing the number of packets or bytes per time unit. To validate the findings of a classical wavelet-based or some other analysis of these data, we typically rely on packet traces collected at the same link, consisting of partial or full packet headers and containing perhaps some amount of the payload. With these measurements, we can, among other things, evaluate TCP/IP flow sizes, and more basic path characteristics such as RTTs. While the former figures prominently in a networking-centric explanation of the main underlying cause for the observed self-similar scaling behavior of the traffic rate process over medium-to-large time scales, the latter is a major cause for the observed deviations from self-similar scaling of the measured traffic rate process over small-to-medium time scales.

Next, consider the case of the network-wide SNMP data

and their use for traffic matrix estimation. To validate findings derived from recently developed traffic matrix estimation techniques (*e.g.,* good agreement with certain types of gravity models [56]), access to network-wide packet traces would certainly do, but is clearly unrealistic/unreasonable. Instead, in large ISPs, there exists widely-deployed operational support for flow export (*e.g.,* Cisco's NetFlow) measurements. With NetFlow data collected from every router inside an ISP's network, we can, for example, obtain the actual traffic flows between every pair of origin-destination or ingress-egress points in the ISP's network and thus check the fidelity and accuracy of any estimated gravity-type model.

Finally, consider the case where the desire for validation is currently hampered by a lack of adequate/appropriate publicly available data. Such a situation arises, for example, in the context of inter-AS traffic matrix estimation. However, far from being a show-stopper, such situations can often outline whole new research agendas in their own right. For example, in the case of inter-AS traffic matrix estimation, challenging open questions are "What are good "surrogate" data?" or "How to measure/collect such "surrogate" data?" or "How good/useful are these data in view of the original problem?" or "Do these data contain other interesting information that can be mined?"

5. INTERNET MRA TARGET PROBLEMS

In the prior sections we have outlined a vision for an Internet MRA that provides multi-scale representations of large graph structures and the spatio-temporal traffic activity that takes place on these structures. In this section, we provide three examples of challenging open problems that are amenable to multi-resolution analysis as a means for demonstrating how to map a problem to our Internet MRA framework. In each instance we describe the available measurements, proposed analysis techniques including domain-specific knowledge used in addressing the problem, and calibration/validation methodology.

5.1 Network Traffic Analysis for a Single Link in the Wired Internet

As mentioned earlier, for the 1-D time series of traffic rates derived from the full packet traces measured on a particular link within an ISP's physical infrastructure, the notion of Internet MRA typically refers to classical wavelet decompositions that support the efficient (temporal) analysis of these time series across a range of different time scales. However, the utility of full packet traces goes well beyond this use. In fact, they naturally extend the information content contained in the 1-D packet rate process time series by revealing much of the vertical decomposition (*i.e.,* layering) of the traffic in the sense that individual packets passing through the given link can now be associated uniquely with their corresponding higher-layer entities such as flows, TCP connection, and possibly even application-layer sessions [15, 35, 53]. They also extend the information content contained in the 1-D packet rate process time series by illuminating the traffic's horizontal decomposition in terms of the packets' IP source and destination addresses. While all the packets traverse the common link where they are recorded, they arrive at that link coming, in general, from a number of different places (in IP space, and/or geography), and they depart from that link heading generally towards a huge variety of different destinations. In this sense, these semanti-

cally rich packet traces motivate the development of a richer Internet MRA that goes beyond the link- and layer-specific temporal analysis of the 1-D packet rate processes and accounts for their more complex structure that becomes apparent once the particular link is properly placed within the vertical and horizontal components of the Internet architecture. Clearly, they provide a rich source of data that invites a deeper look into the behavior and structure of the Internet, and by aggregating across layers and over all IP source and destination addresses, we fully recover the original 1-D packet rate processes. In effect, their analysis requires the development of new techniques and tools, thereby repeating the whole data representation-analysis-validation cycle, albeit with data that extends beyond the link of interest – for example, flow data.

Like packet traces, flow data are semantically rich and can be used for many purposes other than validating estimated traffic matrices. Flow data require in general the development of problem-specific analysis techniques and algorithms, but since they don't provide any information about within-flow packet dynamics, they are of limited use in situations where these dynamic effects are critical. Due to their network-wide availability, flow data sets are extremely helpful for analyzing or validating various aspects of the spatio-temporal behavior of traffic on a network-wide scale. They naturally augment our hand-crafted Internet MRA above for a single link by supporting network-wide constructions of the coarse-to-fine resolution domains via the use of flow attributes such as source/destination IP addresses, prefixes, or AS number. Moreover, by exploiting (where appropriate) the flows' port numbers, we can support the sort of multi-layer MRA representation alluded to earlier. Other applications of this extended MRA technology (or slightly modified versions thereof) loom as real possibilities and have already been pursued to some degree in the context of detecting and identifying anomalous network traffic [25, 26].

5.2 Understanding Encrypted Traffic

Encryption is widely used in today's Internet to protect sensitive traffic. The use of encryption has lead to two (complimentary) challenges. First, many people wonder what information can be extracted from an encrypted data stream. Second, ISPs are finding themselves challenged to manage customer's encrypted traffic streams and are trying to determine what information they can glean from examining the encrypted traffic stream. (The issue of managing encrypted traffic has also led to proposals to expose parts of traffic headers while encrypting others, to give ISPs more information).

5.2.1 MRA issues

From the perspective of Internet MRA, these challenges tend to reduce to variants of a single challenging question: *How much information can be gleaned from one or more discrete event time series representing packet arrival times?* Each time series typically represents the traffic between two intermediate points in a network and can optionally be enhanced with additional information such as packet sizes.

Work on this question is still preliminary, but the early results are exciting. Published work has demonstrated the ability to reconstruct the topology of wireless networks, to extract traffic round-trip times and other signatures [33], to identify individual traffic sources and sinks in the net-

work [10], and to infer whether a network denial of service attack is originating from more than one source [22]. (Note that the denial of service attack traffic is not encrypted but suffers a comparable problem – the attack streams' contents all look alike).

All of this work originates in simple time series measurements: packet traces. And, with the exception of the work deriving sources and sinks (which uses a causal probability model to generate a traffic matrix), all this work uses MRA or MRA-like methods such as Fourier transforms and wavelet decomposition to extract frequency information from the traces. There has been work that seeks to split the data into aggregates and details and use that split to advantage: in particular, the denial of service attack problem benefits from examining the detail domain.

As such, it might seem that understanding encrypted network traffic could be handled straightforwardly using Classical MRA and does not need an Internet-centered MRA approach. But the central problem in this research is one of validation. When running an MRA algorithm over an encrypted packet trace, how does one know that the result of the algorithm actually says anything about the underlying traffic stream(s)? Expressing this problem somewhat more formally, what these techniques typically do is *transform* the traffic trace into another dimension (*e.g.,* frequency) rather than aggregate or decompose the trace and the challenge is to clearly relate the results in the new dimension with the original trace. Aggregation and decomposition may help in this process, but are not the central feature.

5.2.2 Validation

At this point, the development and validation of techniques is still one of trial and error. One typically takes an unencrypted traffic trace (either from simulation or a real network), converts it to a time series (as if the traffic were encrypted) and then transforms the time series using the proposed algorithm.

The algorithm has been chosen with the expectation it will reveal certain properties of the traffic in the trace. Once the transformation is performed, one seeks to see if the properties are, indeed, present and also to explain any additional information that may not have been expected but is present in the transformed result.

If, indeed, the algorithm appears to reveal important properties of the trace the next step is to validate the algorithm using a two-party process. One party takes the trace and extracts the time series. The second party (which has not seen the trace), applies the algorithm to the time series and then interprets the result. Only after the interpretation is finalized do the two parties meet and compare interpretation with the original trace. If the interpretation proves correct, the algorithm can be deemed sound.

5.2.3 Likely Future Work

The next step in the evolution of Internet MRA for understanding encrypted traffic appears to be to construct decomposition techniques. Based on the output of the initial algorithms, one reprocesses the input time series to extract new time series which are then processed again (either with the same algorithm or a new one). This technique has already been used for denial of service attacks [22].

In summary, the current research trajectory for understanding encrypted traffic is the validation of algorithms to interpret time series at various levels of resolution, and the development of informed decomposition techniques that add in clarifying data for finer grained interpretation.

5.3 Traffic Matrix Estimation

5.3.1 The case of a single ISP

An important responsibility of network managers is to insure that their infrastructure operates within specified performance bounds. This is a very challenging problem in large service provider networks whose physical infrastructures are global and diverse, and where the traffic dynamics are complex and unpredictable. One aspect of enabling effective network management is to be able to accurately estimate in (near) real-time the volume of traffic flowing across the network at different levels of granularity of the underlying network infrastructure (*e.g.,* routers or PoPs) This is the challenge in traffic matrix estimation, and is in many ways a prototypical problem for Internet MRA.

To illustrate a concrete Internet MRA framework for the traffic matrix estimation problem, we consider in this section a single service provider and note that large service provider networks support network-wide SNMP measurements as part of normal network operations. The resulting data can be viewed as being defined on a domain A that is a (directed multi-) graph and represents the ISP's physical (*i.e.,* layer-2) network connectivity structure: nodes are routers/switches and links are uni-directional physical connections associated with a physical interface at the nodes on both ends – the outbound interface on one router/switch for the outgoing traffic and the inbound interface on the other router/switch for incoming traffic. The domain's "points" or "coordinates" are these links, and the data value associated with each point consists of the time series of SNMP measurements associated with that link (*i.e.,* number of bytes/time unit, where the time unit is typically between 5-60 minutes).

To derive "aggregation" domains and related "detail" domains that are associated with the data's original domain A and are meaningful for the TM estimation problem, recall that a (point-to-point) traffic matrix represents the traffic demands between pairs of source/ingress and destination/egress nodes in the network. Depending on what these source and destination nodes represent, we obtain different kinds of TMs and rely in the following on the taxonomy used in [57] and introduced in [30]. The latter can be viewed as a first attempt at formalizing the hierarchical nature inherent in TM estimation and is closest in spirit to what we call Internet MRA. We refer in the following to nodes and links that are wholly internal to the service provider's network as "backbone" or "core" routers and links, and call the others "edge" or "non-core" routers and links. Edge links are further categorized into "access" links, connecting to customers, and "peering" links connecting to other (non-customer) ASs. For simplicity, we assume that all access and peering links terminate at edge routers and that backbone links only terminate at backbone routers.

We first define a fine-scale "aggregation" domain A_1 (and related "detail" domain D_1) associated with the data's original domain A as follows. A_1 is given by a (simple directed) graph, where the nodes represent routers and where two nodes are connected by a directed link if there are uni-directional physical connections between the two routers in question. The domain's "points" or "coordinates" are these

links, and the data value associated with each point in A_1 is given by the superposition of those time series that are associated with the points in the data's original domain and map to the link in question (*i.e.,* the superposition of one or more SNMP time series associated with the uni-directional physical connections between the given router pair). We do not formally describe the corresponding detail domain D_1, but require that it contains all information necessary to determine which points in the data's original domain map to which element in A_1 and how much of the corresponding data value (*i.e.,* router-to-router traffic demand) is due to each to its constituent data values (*i.e.,* individual SNMP time series). A_1 and its associated data set represent a coarsened version of the original domain/data and are tailor-made for estimating the "router-level" TM that represents the traffic demands exchanged between pairs of routers (*i.e.,* core or access routers). Note that these demands aggregate the traffic on all the links between a given router pair.

To obtain a coarser scale "aggregation" domain A_2 (and related "detail" domain D_2), we again consider a (simple connected) graph, where now the nodes represent PoPs and (directed) links indicate that a given pair of PoPs is connected by one or more backbone links. The links in this graph represent the points in A_2, and each point is associated with a data value that gives the (time series of) traffic volume originating in one PoP and destined for the other PoP. As before, the corresponding "detail" domain D_2 is required to contains all the information needed to disambiguate A_2 and the data defined on A_2 in terms of the data's original domain. Clearly, A_2 is a coarsened version of A_1 (and A) and is well-suited for investigating "PoP-level" TMs. A PoP-level TM represents the traffic demands exchanged between pairs of PoPs (*e.g.,* cities) in a network. PoPs are composed of a collection of routers (core and access routers), and PoP-level traffic demands are typically aggregates of demands between all the customers, peers, router, and links associated with a given pair of PoPs.

When trying to infer a service provider's router- or PoP-level TMs, the basic problem is that there exists an inherent mismatch between the data needed (*e.g.,* ingress-egress or source-destination demands) and the data most readily available (*e.g.,* SNMP measurements). Since the number of ingress-egress pairs is typically much larger than the number of links in the network, this mismatch results in formulations of the TM estimation problem that involve a highly under-constrained system of equations. Recent research efforts have focused on solving this problem using a number of different methods, including linear programming [17], Baysian estimation [48], expectation maximization [6], and tomogravity [56, 57]. Most of these methods are heuristic in nature, but the tomography approach can be justified on information-theoretic grounds [57]. Validation in the sense of comparing an inferred TM against the "ground truth" has been largely hampered by a lack of adequate data, mainly because establishing the "ground truth" in this context would require access to Netflow data from every router in the service provider's network and detailed intra-domain routing information.

5.3.2 The Internet's ISP-level ecosystem

By taking a more Internet-wide perspective and viewing a service provider's network as part of the Internet's ISP-level ecosystem, we can construct TMs at even coarser scales than PoPs. In fact, since large ISPs often use a number of different Autonomous Systems (ASs) to implement and execute their business model, we can define the aggregation domain A_3 to be the (simple directed) graph whose nodes are ASs and where two ASs are connected by a directed link if one AS sends traffic to the other AS (or vice versa) in accordance with an existing peering relationship between the two ASs. These links define the "points" of A_3, and each point is associated with a data value that corresponds to the total traffic sent on that (directed) link. As before, the corresponding "detail" domain D_3 is required to contains all the information needed to disambiguate A_3 and the data defined on A_3 in terms of the data's original domain. By collapsing those ASs that belong to one and the same ISP into a single node, we obtain the aggregation domain A_4 – a (simple directed) graph that represents traffic exchanges at the level of individual ISPs – and corresponding "detail" domain D_4 (defined in a similar manner as before).

Clearly, the aggregation domains A_3 and A_4 (together with their detail domains D_3 and D_4) are well-suited for exploring the Internet's AS- and ISP-level TMs. However, in contrast to estimating the router- or PoP-level TMs for a single ISP, due to the competitive environment in today's ISP market, neither the AS-specific domains A_3 and D_3 (and associated data sets) nor the ISP-specific domains A_4 and D_4 (and associated data sets) are directly measurable and have remained by and large unexplored. In fact, when trying to infer AS- or ISP-level TMs, the situation regarding available measurements is highly precarious: as mentioned above, inferred AS graphs are known to be incomplete and ambiguous [8] and estimating inter-AS traffic demands is known to be a difficult and largely unresolved problem [9] (similar observations apply to ISP-level graphs and inter-ISP traffic demands). As a result, inference for AS- or ISP-level TMs has remained by and large an open problem, and the design of innovative experiments for measuring AS/ISP-level connectivity and traffic demands, and the development of novel validation techniques are also challenging open problems.

5.3.3 Discussion

Recently developed techniques have been very successful in accurately estimating intra-AS or ISP-specific traffic matrices, but they have largely focused on the data's original domain (*i.e.,* inferring traffic demands between every pair of ingress-egress routers in the network), with little or no concern for how they fare when applied across multiple "scales." In particular, little or no attention has been paid to how these techniques combine with our hand-crafted decomposition/reconstruction algorithms to produce iterative methods that exploit the hierarchical structure inherent in our proposed MRA framework for the TM estimation problem.

The proposed MRA framework with its coarse-to-fine resolution domains appears to be rich enough to support a coherent approach for studying TMs at different levels of (spatial) granularity and to provide adaptive "zoom-in" capabilities for exploring structural properties of traffic matrices that are localized in "space" (*e.g.,* traffic matrix for a single router or PoP which may be useful in diagnosing performance problems or anomalous behavior). As a result, a unified approach to traffic matrix estimation looms as a real possibility, where at the finer scales, we deal with physical domains that consist of a single router (and its inter-

faces with adjacent links) or PoPs and support the estimation of router- and PoP-level TMs, respectively. At coarser scales, we are concerned with more logical or virtual constructs associated with the Internet-wide ecosystem (*e.g.,* ASs or ISPs). Moreover, it is easy to envision augmenting the proposed framework with additional scales to represent yet other physical or logical structures with intrinsic network-specific meaning (*e.g.,* prefix-prefix TMs) or refining it to allow for the treatment of, for example, demand or point-to-multipoint TMs [58].

An obvious benefit from providing an appropriate MRA framework for the TM estimation problem is that it illuminates the different roles played by the different data sets by associating them with specific components of the proposed MRA framework. At the ISP level, the SNMP data sets play a fundamental role, and are defined on a domain whose points are the links of a particular graph. The various TMs are inferred from more or less coarsened versions of these original data, and validation of an inferred TM requires access to NetFlow data from every node (*i.e.,* router) in the data's original domain. At the Internet-wide level, the situation is very different, because neither the basic graph structure underlying AS- or ISP-level TM estimation nor the associated traffic flow information are directly observable/measurable and need to be inferred from auxiliary data sets. The relevance, adequacy, and usefulness of the latter can be judged by how well it reflects the key components of the given MRA framework at the coarse resolution levels.

5.4 Network Intrusion Detection

Securing networks from malicious attacks and intrusions is an extremely challenging problem. A basic principle of network security is to have accurate and timely information on all scans and attacks so that their effects can be mitigated and countermeasures can be deployed. Network intrusion detection is the task of accurately identifying malicious traffic that flows across a link. This is done using two different methods. *Anomaly detection* is done by establishing a baseline of "normal" behavior and then looking for statistical deviations from this baseline (*e.g.,* [3, 20]). *Misuse detection* is done by comparing all traffic on a link against a library of malicious signatures and generating an alert when there is a match. Standard examples of systems that employ misuse detection are SNORT [40] and Bro [36]. Both of these methods are complicated by the inherent variability and diversity of benign traffic and by the ingenuity and persistence of malicious adversaries.

Intrusion detection systems (IDS) typically monitor all packet traffic on a link (often an ingress/egress link) online, or are used on packet traces collected for off-line analysis. In either case, packet headers, payloads as well as meta-data such as timestamps on packets can be considered in the detection process. Recent studies have also suggested the use of additional measurement systems such as network honeypots – systems used to monitor routed but otherwise unused IP addresses – as a means for augmenting the perspective of the security analyst [55]. Further sources of data include summaries of aggregate firewall/NIDS logs from sources such as Dshield.org [50]. Additionally, forensic off-line analysis of attacks and intrusions frequently draws from a variety of data sources including system logs from target hosts.

From the most local perspective, intrusion detection nat-

urally lends itself to multi-scale representation. As mentioned above, multi-resolution methods have already shown promise when applied to the problem of anomaly detection [3]. Similarly, the alerts generated by an IDS on a single link are typically organized in a hierarchy related to their type, their scope vis-a-vis the attack target(s) within a network, and their volume over time. This suggest an MRA-based approach to assessing and visualizing alerts [29, 23].

More broadly, the task of intrusion detection resembles the problem of traffic matrix estimation since IDS are often deployed at key vantage points through out a given ISP. Both the traffic data itself and the alerts generated by each of the individual IDS form the basis for aggregation and detail domains as described in Section 5.3.1. In this case, however, the problem is not to identify flows between source-destination pairs, it is to highlight the most important aspects of attacks and intrusions as they encounter a network's defense perimeter. Distributed alert organization and fusion is an open and active area of research (*e.g.,* [4, 12, 59]) that we believe will be enhanced by the systematic application of Internet MRA methods.

While not well studied, it seems clear that extending the notion of intrusion detection beyond the boundaries of a single network to use data from a diverse set of IDS throughout the Internet offers some compelling possibilities. For example, if new types of attacks are detected in isolated portions of the Internet, countermeasures may be able to be deployed broadly before these attacks can spread widely. This vision is the primary motivation for attack data sharing protocols (*e.g.,* [54, 31]) . This general idea is a different but clearly related dimension of the Internet ISP-level ecosystem described above.

Finally, the tasks of calibration and validation in network intrusion detection are frequently costly and time consuming since they are primarily ad-hoc. Signature sets for intrusion detection systems are usually calibrated by hand as are the thresholds or anomaly detection systems. This process is typically conducted by experienced security analysts who can study and interpret individual signature details. Validation is closely related to the calibration process and is often done by trial and error in live deployments. The focus of the calibration/validation cycle is most often to reduce the number of false positive alerts, thereby increasing the utility of the IDS deployment.

6. SUMMARY

Recent advances in networking research based on the application of analytic techniques from signal processing, statistics and applied mathematics suggest significant opportunities for addressing important empirical networking problems. In this paper we present a vision called Internet MRA in support of a principled approach toward data-driven networking research. Our motivation for this work is to call attention to the opportunities of Internet MRA to provide a framework for its use designed to increase the potential for successful application and limit misuse.

The key characteristic of Internet MRA is a structured approach to representing, analyzing, and visualizing Internet-related measurements that respects the critical design aspects of today's Internet architecture (*e.g.,* layering, horizontal decomposition). We argue that this approach is essential for the effective treatment of the spatial, temporal and functional decomposition of the traffic that flows across

the network and for the successful application of a broad range of analysis methods to the existing and anticipated types of network measurement data. We reinforce this argument by providing examples of the application of Internet MRA methods to several key problems.

It is our hope that this position paper will catalyze the use of Internet MRA methods in research that targets the problem domains that we describe as well as others. We also posit that this approach to network data analysis serves as a touchstone for communication and collaboration between network researchers and traditional data analysis communities. To that end, in the long term, we believe that the true value of Internet MRA will be realized not only in its successful application in networking, but also in the new research directions that it opens across research communities.

7. REFERENCES

[1] P. Abry and D. Veitch. Wavelet Analysis of Long Range Dependent Traffic. *IEEE/ACM Transactions on Information Theory*, 44(1), 1998.

[2] D. Alderson, L. Li, W. Willinger, and J. Doyle. Understanding Internet Topology: Principles, Models and Validation. *IEEE/ACM Transactions on Networking*, 13(6), 2005.

[3] P. Barford, J. Kline, D. Plonka, and A. Ron. A Signal Analysis of Network Traffic Anomalies. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, Marseilles, France, November 2002.

[4] T. Bass. Intrusion Detection Systems Multisensor Data Fusion: Creating Cyberspace Situational Awareness. *Communications of the ACM*, 43(1), 2000.

[5] A. Broido, E. Nemeth, and K. Claffy. Spectroscopy of DNS Update Traffic. In *Proceedings of ACM SIGMETRICS '03*, San Diego, CA, June 2003.

[6] J. Cao, D. Davis, S. Vander Weil, and B. Yu. Time-Varying Network Tomography. *Journal of the American Statistical Association*, 2000.

[7] M. Casado, T. Garfinkel, W. Cui, V. Paxson, and S. Savage. Opportunistic Measurement: Extracting Insight from Spurious Traffic. In *Proceedings of ACM Fourth Workshop on Hot Topics in Networks (HotNets '05)*, College Park, MD, November 2005.

[8] H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. Towards Capturing Representative AS-level Internet Topologies. In *Proceedings of ACM SIGMETRICS '02*, Marina Del Rey, CA, June 2002.

[9] H. Chang, S. Jamin, Z. Mao, and W. Willinger. An Empirical Approach to Modeling Inter-AS Traffic Matrices. In *Proceedings of ACM Internet Measurement Conference '05*, Berkeley, CA, October 2005.

[10] D. Cousins, C. Partridge, K. Bongiovanni, A. Jackson, R. Krishnan, T. Saxena, and W. Strayer. Understanding Encrypted Networks Through Signal and Systems Analysis of Traffic Timing. In *Proceedings of IEEE Aerospace Conference*, Big Sky, MT, March 2003.

[11] M. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6), December 1997.

[12] F. Cuppens and A. Miege. Alert Correlation in a Cooperative Intrusion Detection Framework. In *Proceedings of IEEE Symposium on Security and Privacy (Oakland '02)*, Oakland, CA, May 2002.

[13] I. Daubachies. *Ten Lectures on Wavelets*. SIAM Journal, 1992.

[14] C. Estan, S. Savage, and G. Varghese. Automatically Inferring Patterns of Resource Consumption in Network Traffic. In *Proceedings of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.

[15] A. Feldmann. BLT: Bi-Layer Tracing of HTTP and TCP/IP. In *Proceedings of the World Wide Web Conference*, Toronto, Canada, May 1999.

[16] A. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. QuickSAND: Quick Summary and Analysis of Network Data. Technical Report 2001-43, DIMACS, November 2001.

[17] O. Goldschmidt. ISP Backbone Traffic Inference Methods to Support Traffic Engineering. Internet Statistics and Metrics Analysis (ISMA) Workshop, December 2000.

[18] M. Goldszmidt and E. Kiciman. First Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML'06). http://research.microsoft.com/workshops/sysml, June 2006.

[19] M. Goldszmidt and E. Kiciman. Second IEEE International Workshop on Networking Meets Databases (NetDB'06). http://www.cs.brown.edu/research/db/netdb06, April 2006.

[20] Y. Gu, A. McCallum, and D. Towsley. Detecting Anomalies in Network Traffic Using Maximum Entrophy Estimation. In *Proceedings of ACM Internet Measurement Conference '05*, Berkeley, CA, October 2005.

[21] P. Huang, A. Feldmann, and W. Willinger. A Non-intrusive, Wavelet-based Approach to Detecting Network Performance Problems. In *Proceedings of ACM Internet Measurement Workshop '01*, San Francisco, CA, November 2001.

[22] A. Hussain, J. Heidemann, and C. Papadopoulos. A Framework for Classifying Denial of Service Attacks. In *Proceedings of the ACM SIGCOMM Conference*, Karlsuhe, Germany, August 2003.

[23] H. Koike and K. Ohno. SnortView: Visualization System for Snort Logs. In *Proceedings of ACM Workshop on Visualization and Data Mining for Computer Security*, Washington, DC, 2004.

[24] Sprint Labs. The IPMon Project. http://ipmon.sprint.com/, March 2006.

[25] A. Lakhina, M. Crovella, and C. Diot. Diagnosing Network-Wide Traffic Anomalies. In *Proceedings of ACM SIGCOMM '04*, Portland, OR, August 2004.

[26] A. Lakhina, M. Crovella, and C. Diot. Mining Anomalies Using Traffic Feature Distributions. In *Proceedings of ACM SIGCOMM '04*, Philadelphia, PA, August 2005.

[27] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural Analysis of Network Traffic Flows. In *Proceedings of ACM SIGMETRICS '04*, New York, NY, June 2004.

[28] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the Self-Similar nature of Ethernet Traffic (extended version). *IEEE/ACM Transactions on Networking*, pages 2:1–15, 1994.

[29] Y. Livnat, J. Agutter, S. Moon, R. Erbacher, and S. Foresti. A Visualization Paradigm for Network Intrusion Detection. In *Proceedings of IEEE Workshop on Information Assurance and Security*, West Point, NY, June 2005.

[30] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. In *Proceedings of ACM SIGCOMM '02*, Pittsburgh, PA, August 2002.

[31] D. Nojiri, J. Rowe, and K. Levitt. Cooperative Response Strategies for Large Scale Attack Mitigation. In *Proceedings of the 3rd DARPA Information Survivability Conference and Exposition (DISCEX'03)*, April 2003.

[32] K. Park and W. Willinger, editors. *Self-Similar Network Traffic Analysis and Performance Evaluation.* Wiley, 2000.

[33] C. Partridge, D. Cousins, A.W. Jackson, R. Krishnan, T. Saxena, and W.T. Strayer. Using Signal Processing to Analyze Wireless Data Traffic. In *ACM Workshop on Wireless Security (WISE)*, Atlanta, GA, September 2002.

[34] N. Patwari, A. Hero, and A. Pacholski). Manifold Learning Visualization of Network Traffic Data. In *Proceedings of the ACM SIGCOMM Workshop on Mining Network Data (MineNet'05)*, Philadelphia, PA, August 2005.

[35] V. Paxson. *Measurements and Analysis of End-to-End Internet Dynamics.* PhD thesis, University of California Berkeley, 1997.

[36] V. Paxson. BRO: A System for Detecting Network Intruders in Real Time. In *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, January 1998.

[37] V. Paxson. Strategies for Sound Internet Measurement. In *Proceedings of ACM Internet Measurement Conference '04*, Taormina, Italy, October 2004.

[38] V. Paxson and S. Floyd. Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3(3), June 1995.

[39] AT&T Labs Research. The Gigascope Project. http://public.research.att.com/, March 2006.

[40] M. Roesch. The SNORT Network Intrusion Detection System. http://www.snort.org, 2006.

[41] M. Roughan, M. Thorup, and Y. Zhang. Performance of Estimated Traffic Matricies in Traffic Engineering. In *Proceedings of ACM SIGMETRICS '03*, San Diego, CA, June 2003.

[42] Computer Science and National Research Council Telecommunications Board. *Looking Over the Fence at Networks, A Neighbor's View of Networking Research.* National Academies Press, November 2002.

[43] Computer Science and National Research Council Telecommunications Board. *The Internet Under Crisis Conditions: Learning from September 11.* National Academies Press, November 2002.

[44] S. Sen and S. Sahu. Second Workshop on Mining Network Data (MineNet'06). http://www.acm.org/sigs/sigcomm/sigcomm2006, September 2006.

[45] J. Sommers, P. Barford, and W. Willinger. SPLAT: A Visualization Tool for Mining Internet Measurements. In *Proceedings of the Passive and Active Measurement Conference (PAM'06)*, Adelaide, Australia, March 2006.

[46] L. Tang and M. Crovella. Virtual Landmarks. In *Proceedings of ACM Internet Measurement Conference '03*, Miami, FL, October 2003.

[47] tcpdump. *http : //ftp.ee.lbl.gov/tcpdump.tar.Z*.

[48] C. Tebaldi and M. West. Bayesian Inference of Network Traffic Using Link Count Data. *Journal of the American Statistical Association*, June 1998.

[49] X. Tian, H. Wu, and C. Ji. A Unified Framework for Understanding Network Traffic Using Independent Wavelet Models. In *Proceedings of IEEE INFOCOM '02*, New York, NY, June 2002.

[50] J. Ullrich. Dshield.org. http://www.dshield.org, March 2006.

[51] Route Views. University of oregon. http://www.antc.uoregon.edu/routeviews.

[52] W. Willinger, D. Alderson, and L. Li. A Pragmatic Approach to Dealing with High-variability in Network Measurements. In *Proceedings of ACM Internet Measurement Conference '04*, Taormina, Italy, October 2004.

[53] V. Paxson Y. Zhang, L. Breslau and S. Shenker. On the Characteristics and Origins of Internet Flow Rates. In *Proceedings of ACM SIGCOMM '02*, Pittsburgh, PA, August 2002.

[54] V. Yegneswaran, P. Barford, and S. Jha. Global Intrusion Detection in the DOMINO Overlay System. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, February 2004.

[55] V. Yegneswaran, P. Barford, and V. Paxson. Using Honeynets for Internet Situational Awareness. In *Proceedings of ACM Fourth Workshop on Hot Topics in Networks (HotNets '05)*, College Park, MD, November 2005.

[56] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads. In *Proceedings of ACM SIGMETRICS '03*, San Diego, CA, June 2003.

[57] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An Information-Theoretic Approach to Traffic Matrix Estimation. In *Proceedings of ACM SIGCOMM '03*, Karlsruhe, Germany, August 2003.

[58] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. Estimating Point-to-point and Point-to-multipoint Traffic Matrices: an Information-theoretic Approach. *IEEE/ACM Transactions on Networking*, 13(5), 2005.

[59] C. Zou, L. Gao, W. Gong, and D. Towsley. Monitoring and Early Warning for Internet Worms. In *Proceedings of the 10th ACM Conference on Computer and Communications Security*, Washington, DC, October 2003.