Research Statement

Nick Feamster

feamster@cc.gatech.edu
http://www.cc.gatech.edu/∼feamster/

Overview

The goal of my research is to help network operators run their networks better, and to enable users of these networks (both public and private) to experience high availability and good end-to-end performance; I call this area "network operations". I tackle practical problems using a "first principles" approach, design systems based on these principles, and implement and deploy these systems in practice.

My research comprises two complementary approaches: (1) designing and deploying real-world systems, tools, and algorithms that are immediately applicable to today's networks and (2) developing fundamental network primitives and architectures that could dramatically improve network operations, even if the deployment cost is substantial. Deploying "bottom up", real-world solutions to network operations problems gives me a unique perspective on real-world network operations and an appreciation for details (e.g., this bottom-up allowed me to build strong domain expertise in both BGP routing and spam). Great strides in networking research are often inherently "top down", but such fundamental work must be informed by real-world problems and details. My two-fold approach allows my research strike a balance between operational relevance and long-term impact.

My current research applies this approach to two areas: (1) distributed algorithms and systems for network monitoring and security; and (2) routing protocols for improving network connectivity in federated, competitive environments.

Network Monitoring and Security

Users of computer networks demand high availability and good performance in the face of continually changing network conditions. Most communications networks are kept afloat by the vigilance of network operators, who tune network configurations in response to changes in available resources, failures of network elements, fluctuations in traffic demands, or the onset of malicious or otherwise unwanted traffic (e.g., spam).

Responsiveness to changing network conditions requires increasingly sophisticated detection and mitigation strategies as networks become saddled with more features and subject to new classes of unforeseen threats, and as users' demands from the network become increasingly stringent. Network monitoring has long been a critical piece of this puzzle: mitigating network disruptions depends, first and foremost, on tools and techniques to quickly and accurately detect these disruptions and determine their causes.

Network monitoring must be accurate and robust. Accurate monitoring techniques should detect disruptions when they occur (with a negligible number of false alarms), and, to the extent possible, identify the cause of the disruption (e.g., the faulty network element, the source of unwanted traffic). Robust monitoring should be able to detect disruptions when measurements may be noisy, incomplete, or when attackers are actively trying to disguise their presence. Network monitoring strategies are most accurate when they are distributed; that is, when they draw upon observations from a large number of vantage points.

Unfortunately, fast distributed detection of network anomalies is challenging. Information about network conditions is voluminous and noisy. Network operations needs better techniques for quickly detecting—and diagnosing—network faults and anomalies, which requires fundamental advances both in data mining algorithms and in distributed systems for processing this data. My research in network monitoring and security focuses on both the application of data mining algorithms (and the development of new algorithms), and, perhaps most importantly, the implementation and evaluation of these algorithms in real-world net-

works. I am applying this method to spam filtering and to network fault diagnosis.

Problem 1: Spam Filtering

Spam filtering today is, by and large, reactive: network operators observe spammers and generate content filters and IP-based blacklists based on this observed activity. Content-based filtering is rapidly becoming ineffective as spammers develop increasingly sophisticated techniques (e.g., image-based spam) to evade these filters. To make matters worse, recent trends in spam activity suggest that reactive techniques such as IP-based blacklisting are becoming ineffective as an increasing fraction of spam is being sent from "fresh" IP addresses for which operators have little or no information about the reputation.

Next-generation spam filters must be proactive and predictive. My research focuses on developing techniques that can help network operators identify spam based solely on network-level patterns. I call this approach "behavioral blacklisting". Our research involves both developing new spam filtering techniques and algorithms and evaluating these algorithms on real networks.

Problem 2: Network Fault Diagnosis

Network faults and disruptions—changes in network conditions that are caused by underlying failures of routing protocols or network equipment—have a significant impact on network performance and availability. Operators today have myriad datasets (e.g, traffic statistics, SNMP, routing data, "syslogs") at their disposal to monitor for network disruptions. All of these datasets have proven difficult to use for extracting actionable events from "background noise". Indeed, network data is voluminous, complex and noisy; network operators are not at a loss for network data; rather, they lack efficient algorithms and systems for analyzing and processing this data to quickly detect network faults while at the same time maintaining a low overall false alarm rate.

My research focuses on developing algorithms and systems to help network operators quickly detect and locate network faults. I am primarily focused on developing techniques for network-wide data analysis; techniques that incorporate independent observations of the same phenomenon (e.g., a network fault) from multiple vantage points can provide clues as to the severity, location, and cause of the fault. Network-wide monitoring, however, mandates both new algorithms for efficiently extracting actionable network disruptions from distributed, voluminous, heterogeneous data and a system for efficiently processing this data so that disruptions can be detected scalably and quickly. I am working on algorithms and systems for distributed network fault detection that I hope to ultimately deploy in operational networks, from campus networks such as Georgia Tech to possibly even large transit networks.

Improving Network Connectivity

One of the current Internet's biggest pitfalls is that today's end hosts cannot make efficient use of available connectivity that exists in the underlying physical network. This inability to leverage the underlying rich connectivity creates both inefficient routing (e.g., unnecessarily long paths) and fragility (as evidenced by the continual de-peering debacles, fiber cuts, etc., all of which cause significant disruptions to connectivity). These inefficiencies arise from two characteristics of today's routing protocols: (1) single-path routing and (2) restrictive, bilateral business policies. Both of these characteristics prevent end hosts from using available network capacity to a destination. In my research, I am investigating how new routing protocols can scalably achieve sufficient path diversity by making use of multiple parallel paths and how a new routing framework could create a more efficient market for connectivity. Although I am exploring both of these areas in the context of "future Internet architectures", I am also exploring how facets of these designs can be deployed in today's Internet routing fabric.

Problem 1: Scalable Path Diversity

To make the best use of available capacity, and to gracefully degrade in the event of network faults, routing protocols should expose multiple routes for each destination to each end host. Unfortunately, scalably providing such path diversity has proven difficult in practice, since computing and storing additional routes at each node implies (often prohibitive) increases in complexity and state. I am developing solutions that use network virtualization—a primitive that allows multiple virtual networks to operate in parallel on a single, shared physical infrastructure—to scalably offer path diversity to end hosts.

I am designing and implementing these architectures on VINI (virtual network infrastructure), a distributed testbed for new network protocols and architectures that I am developing with my students, as well as colleagues at Princeton University. VINI is aligned with a larger effort at the National Science Foundation (specifically, the GENI and FIND research programs) to help network researchers and operators design, deploy, and evaluate new network protocols and architectures in realistic settings.

Problem 2: Efficient Markets for Connectivity

Much of the inefficiency in today's interdomain routing (i.e., routing between the Internet's thousands of independently operated, competing networks) results from the fact that interdomain connectivity is cobbled together from bilateral contracts between pairs of networks. This restrictive contracting model not only "hides" a significant amount of connectivity from end hosts but it also introduces fragility and inefficiency on end-to-end paths. I am designing (and ultimately plan to implement) a new routing architecture that is centered around a small number of aggregators, which serve as clearinghouses for long-haul connectivity between "edge" networks. This AGregator-Oriented Architecture (AGORA) should create both more efficient end-to-end paths and a way for end hosts to transfer value for these more efficient paths to the networks that offer this connectivity.