Project Summary

Troubleshooting undesirable network events, such as poor connectivity or performance experienced by users, is a vital undertaking for any Internet Service Provider (ISP). This undertaking has never been easy, and now becomes ever more challenging due to factors such as higher user expectations, deployment of new protocols and services, and the often subtle or "silent" nature of such events. It is however important to identify the root causes of these network events, and hopefully prevent them from happening again, because the underlying errors and bugs, left undetected, will come back to haunt us in the future.

Unfortunately, current network measurement and monitoring infrastructures offer grossly inadequate support for such root cause analyses, making them extremely difficult to carry out. This inadequacy is clear when we draw an analogy to software debugging. Network operators would love to have a "core dump" that would give them a clear idea as to "who" (which flow) went "where" (which network path or output link) at "what speed" around the time a network event happened and allow them to troubleshoot the network in a principled manner. Yet, the information network operators have at hand for diagnosis and troubleshooting purposes is nowhere close to the debugging information available to software engineers.

Intellectual Merit: The objective of this project is to build a network diagnosis support infrastructure that allows for speedy troubleshooting and root cause analysis of a wide range of undesirable network events such as routing failures, traffic anomalies, and configuration errors. We hope to provide the essential debugging information network operators need to troubleshoot a network like what is provided to software engineers by the "core dump". We are however not suggesting to perform a full core dump of a network, which a large ISP can ill afford. Rather, our objective is to extract, from high-speed packet streams on individual network links, an approximate and highly compressed core dump that is orders of magnitude smaller in size (compression requirement) yet can provide information almost as rich and accurate (fidelity requirement) as what we can obtain from a full core dump for networking troubleshooting purposes.

While this objective sounds daunting given the stringent compression and fidelity requirements that seem inherently in conflict with each other, a set of network *data streaming* and intelligent sampling algorithms developed in the past decade by the PIs and others will indeed bring us one step closer to this objective. To close the remaining gap, we propose to investigate how to sample a small yet representative set of flows that are most valuable to network troubleshooting tasks, and to design novel decoding techniques that can significantly improve the coding efficiencies of existing data streaming algorithms and hence greatly reduce their costs, without compromising estimation accuracies. Also related to this cost issue is our observation, during our collaborations with AT&T, that subtle network events may be detected or reported at a much later time than when they occurred. Therefore, we propose to study how to compress and unavoidably decay, in the smoothest manner possible, "core dumps" stored in persistent storage (i.e., disks), so that network operators can "travel back in time" as much as possible in order to diagnose and troubleshoot such events, given the finite amount of disk space we have for keeping such data.

Broader Impact: This project will engage both graduate and undergraduate students through integrated classroom curriculum and research training that span multiple disciplines, from fundamental mathematics and algorithm design to networking. The results will be broadly disseminated through publications, invited talks, tutorials, and open-sourcing of software developed for this project in accordance with the policies of each institution. The PIs will work closely with leading networking and systems solution providers, such as AT&T, to facilitate the transfer of technology from the research environment to actual commercial deployments. Last but not least, the PIs are committed to outreach efforts at our corresponding campuses to broaden the participation of under-represented groups in research and higher education.

Keywords: Network data streaming, intelligent sampling, diagnosis, troubleshooting, root-cause analysis.