Project Summary – Critter-at-Home

Networking and cybersecurity research critically need publicly available, fresh and diverse *application-level* data, for data mining and for validation. For example, spam filtering, intrusion detection and traffic classification fields absolutely need application data to learn trends in legitimate network use, to establish ground truth, and for realistic evaluation of proposed systems.

There are very few publicly available network traces that contain application-level data. These suffer from being outdated, and from containing very specific data useful only to some researchers. The main reason for this lack of content-rich network data is the enormous privacy risk that sharing such data creates. Networking community has long worked to solve a simpler problem—sharing of content-less network data—with little progress. Sharing application-level data greatly increases privacy risks, because such data is rich with personal and private information, such as human names, social security numbers, phone numbers, usernames, passwords, credit card numbers, etc. that Internet criminals can monetize.

Intellectual Merit: We propose to develop a continuously updated archive of content-rich network data, contributed by volunteer users, called Critter-at-home. Users would join the Critter overlay whenever online, offering their data to interested researchers. Privacy of data contributors would be protected by several means. First, contributors may opt to host their own data on their machines, thus retaining full control over it. Second, we will process contributed data to modify all personal and private information (PPI) and we will encrypt it. Both actions will minimize its appeal to criminals. Third, no human apart from the contributor will ever access the raw, PPI-sanitized, data. Instead, researchers will query the data via our Critter-at-home framework, and they will receive aggregate statistics (counts, distributions, etc.) of the traffic features they query for. Four, all contact with a contributor will be at her discretion and will be done through an anonymous network, where contributor identities are hidden both from researchers and the Internet at large. Five, contributors that so desire will have full, fine-grained control over which traffic they share, which queries they are willing to answer and in which context.

Our proposed work relies in part on the secure query framework developed by PI Mirkovic under her other NSF-funded project, award number 0914780, for safe sharing of network data. This framework allows only for queries about aggregate features of the data, such as counts, distributions, etc. and preserves user privacy by applying k-anonymity and l-diversity principles. Our proposed research will focus on the following tasks: (1) investigating how to best identify and sanitize PPI information, (2) investigating how to provide both flexible and human-understandable framework and policies, (3) investigating how to assure contributors that their policies are honored by our framework, (4) modifying PI Mirkovic's existing secure query framework to support distributed operation as required by Critter-at-home, (5) building our proposed framework, and (6) attracting data contributors to Critter-at-home.

Broader Impacts: This research will result in a publicly accessible, diverse and fresh archive of content-rich network data. Such archive would greatly advance security research by providing necessary data for its validation and for data mining. This archive would further be valuable to a broader networking community and its data could be used e.g., for realistic traffic generation, as ground truth in traffic classification, and for many other purposes. Our sanitization and data protection approaches may further apply to and advance sharing of other sensitive data that is not related to networking. The PI will actively seek to involve undergraduates and students from underrepresented groups in research and educational activities within this project. All our code will be open source, and it and our findings will be made publicly available.