C. PROJECT SUMMARY

Traffic traces are essential for networking research, for data mining and validation purposes. But while networking has been blooming for more than forty years, there are less than ten public trace archives to date! This indicates that many organizations resist trace sharing because of privacy concerns. A traditional remedy for privacy risk is to *sanitize* trace data by removing packet contents and anonymizing IP addresses in headers. Many researchers have shown, however, that sanitization can be broken, and private data revealed. *Passive attacks* use data from sanitized traces that were previously thought privacy-safe, such as packet timing or length, to break anonymization. *Active attacks* inject well-crafted packets into a trace during a collection time, and identify them in the sanitized trace thus breaking the anonymization. While passive attacks can be handled by removing more data during sanitization, at a great expense of utility, active attacks cannot. All these threats lead to reduced trace sharing.

One could argue that some privacy risks are justified if trace sharing promotes research. But current usage of public traces indicates that sanitized traces lack data needed for high-quality research. Out of 144 papers total, published in SIGCOMM and IMC in 2006 and 2007, 49 used some network trace, but only 10 used a public trace and the rest used privately collected traces. Limited access to high-value trace data seriously limits networking research, hinders validation of prior results, and favors researchers who work at organizations with unrestricted access to network data, such as large ISPs. Thus trace sanitization fails in both its goals: it cannot provide useful data to many researchers and it cannot protect privacy.

Intellectual Merit: We propose to explore a novel direction to address the privacy/utility tradeoff of trace sharing: secure queries on original traces under their owner's control. The data owner publishes a query language and an online portal, allowing researchers to submit sets of queries to be run on data. Only certain operations are allowed on certain data fields, and in specific contexts. This policy is specified by the provider and enforced by the language interpreter. The interpreter analyzes the queries, runs those that are permitted and returns the results to the researcher. The results consist of aggregate information such as counts, histograms, distributions, etc. and not of individual packets. PI is aware of extensive prior work in privacy-preserving data mining in databases. We reuse a lot of findings from this research, but argue that trace data mining has a different set of challenges from databases, as discussed in the proposal.

Secure queries address privacy/utility tradeoff much better than sanitization. Privacy is protected by finer-grain control given to data owner, which permits us to detect many passive attacks and to minimize information leakage from active attacks. Future attack vectors can be handled by adding new constraints on the query language. Secure queries also show a potential to reveal more data to researchers than it was possible with sanitization. Fine-grain control via query language enables processing of many fields in the application header, and even in sensitive application content, while satisfying the owner's privacy concerns. This is likely to increase utility of public traces for application and security research.

Our work will first investigate research utility of network trace data, and the relationship of known and novel attacks to combinations of packet fields, operations on those fields, and contexts that pose privacy risk. Based on our findings, we will develop a secure query language Trol, and an interpreter for this language Patrol. Trol will support common operations on traces, needed for networking research, and Patrol will prohibit queries and contexts that pose a privacy risk as specified by the provider's privacy policy. Both the language and policies will be extensible by data owners to accommodate future discoveries.

Trol and Patrol will be deployed at USC/ISI and will run on publicly available, sanitized trace archives and on synthetically generated, full packet traces. This deployment will help us test expressiveness and privacy protection of Trol operations. We will frequently solicit feedback from researchers using our systems and from data owners to refine and improve them. We will also publicize our work among data owners, to motivate the shift from sanitization to protection of traces via secure queries.

Broader Impact: Trol and Patrol will advance networking research by enabling sharing of previously discarded data, promoting novel research that could previously be conducted only with private traces. They will also increase trace sharing by providing data owners with strong privacy guarantees, enabling them to exert fine-grained control over data access and to monitor data usage. A secondary benefit of these tools will be to provide an easy to use, high-level environment for trace processing. We further expect that our findings will improve privacy and promote sharing of other network data, such as firewall or access logs. All results of our research will be made publicly available.

Keywords: Privacy; trace sharing; anonymization; secure queries