C. PROJECT SUMMARY

Traffic generation plays an integral part of cyber-security defense testing in network testbeds. Generating just any traffic is easy, but generating **realistic** traffic is hard. The first challenge lies in diversity of networks, which then leads to diversity of their traffic. Thus what is realistic for one network may not be realistic for another. The models used in traffic generation must thus be tailored to the networks that the cyber-security defense targets for its deployment, and are usually mined from traffic logs collected in similar networks. The second challenge lies in the shortage of public traffic logs from which to mine these models. The DHS-sposored PREDICT project aims to address this issue by simplifying network data sharing and it has already made a large number of traffic logs available to researchers.

But the key reason why realistic traffic generation is hard is that "realistic" means different things to different people. A researcher testing a defense, which detects largest hitters, may care only that the traffic volume per source resembles values seen in real networks. Another researcher testing a DDoS defense requires congestion-responsive traffic generation but may not care about address distribution or traffic contents. All existing traffic generators have a **fixed definition of realism**. This definition – a set of traffic dimensions that users supposedly care about – is hard-coded in the generator's code. The generator then mines information about only these dimensions from traffic traces and it generates traffic that fits the mined values. The generated traffic is "realistic" but only along those fixed dimensions. A researcher caring about a different dimension set must change the generator's source code, which is often a huge effort.

We propose to build a traffic generator whose definition of realism can be fully specified by a user. We will first build the generator where users can choose from a large, fixed set those dimensions that should be reproduced realistically. The generator will extract models for these dimensions out of traffic traces or application logs, and will then reproduce the traffic that fits those models. We will next improve the generator by coupling it with a high-level language that can be used to specify custom dimensions. This will enable users to fully customize the definition of reality to fit their research goals.

The generator will consist of three modules: (1) a module that mines values for user-specified dimensions from traffic logs, (2) a module that generates random traffic that fits the model mined in the previous step, (3) a module that replays traffic from a log so that it exactly matches the logged traffic along the user-specified dimensions.

Intellectual Merit: The key novelty of our approach lies in the customizable definition of realism that our generator will support. By allowing users to specify their own reality dimensions our traffic generation tool will be generic enough to meet the evaluation needs of any cyber security researcher. Further, integration of the traffic generation from models and traffic replay in a single tool is novel — existing tools support only one of these generation approaches. Finally, our tool will support traffic generation at human, application, transport or network level while existing tools support it only at one select level.

Broader Impact: Our proposed work will advance cyber-security defense research by supporting rigorous and realistic evaluation of its products. It will do that by both fitting researchers' needs well and by being extremely portable and easy to deploy and use. Because users will be able to customize the definition of realism as they desire, the evaluation will properly stress the cyber-security defenses and its results will be predictive of the defenses' performance in real deployment. The traffic generator's capabilities to both generate traffic from learned models and to replay it from network logs enable a wide range of testing strategies and support thorough exploration of problem space. Better evaluation strategies will lead to better cyber-security defenses.

Unlike much of existing traffic generation software that is of research quality, our software will be developed by a full-time programmer and will be of production quality. We will ensure that it is portable to different hardware and operating systems, and self-contained. This will make its deployment easy on private testbeds, eliminating time researchers invest today in modification of existing generators to make them run on their platform and/or to make them fit the realism definition of the researcher. We will further integrate our traffic generator with the DETER testbed for cyber security experimentation. This open, free testbed has more than 2,000 users from academia, industry and government and is used extensively both in cyber security research and education. Such a wide and diverse user base will help drive our development and it will also benefit directly from our progress. We will further work closely with seven DETER researchers to have them feed our design process and test-drive prototypes of our software. All our software will be released as open-source under the GNU GPL v3 license.