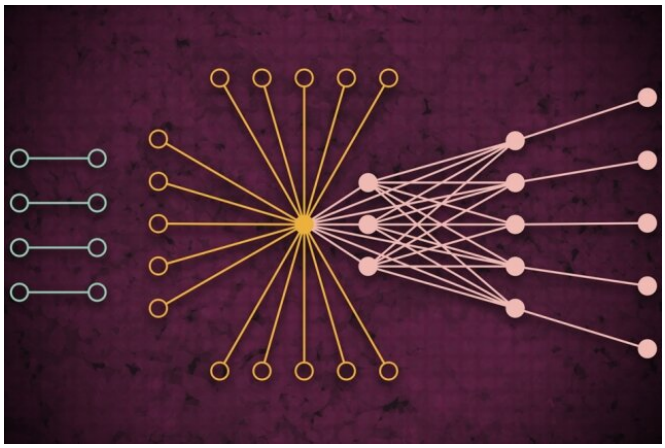
Supercomputer analyzes web traffic across entire internet

28 October 2019, by Rob Matheson



Using a supercomputing system, MIT researchers developed a model that captures what global web traffic could look like on a given day, including previously unseen isolated links (left) that rarely connect but seem to impact core web traffic (right). Credit: MIT News

Using a supercomputing system, MIT researchers have developed a model that captures what web traffic looks like around the world on a given day, which can be used as a measurement tool for internet research and many other applications.

Understanding web traffic patterns at such a large scale, the researchers say, is useful for informing internet policy, identifying and preventing outages, defending against cyberattacks, and designing more efficient computing infrastructure. A paper describing the approach was presented at the recent IEEE High Performance Extreme Computing Conference.

For their work, the researchers gathered the largest publicly available internet traffic dataset, comprising 50 billion data packets exchanged in different locations across the globe over a period of several years.

They ran the data through a novel "neural network" pipeline operating across 10,000 processors of the MIT SuperCloud, a system that combines computing resources from the MIT Lincoln Laboratory and across the Institute. That pipeline automatically trained a model that captures the relationship for all links in the dataset—from common pings to giants like Google and Facebook, to rare links that only briefly connect yet seem to have some impact on web traffic.

The model can take any massive network dataset and generate some statistical measurements about how all connections in the network affect each other. That can be used to reveal insights about peer-to-peer filesharing, nefarious IP addresses and spamming behavior, the distribution of attacks in critical sectors, and traffic bottlenecks to better allocate computing resources and keep data flowing.

In concept, the work is similar to measuring the cosmic microwave background of space, the near-uniform radio waves traveling around our universe that have been an important source of information to study phenomena in outer space. "We built an accurate model for measuring the background of the virtual universe of the Internet," says Jeremy Kepner, a researcher at the MIT Lincoln Laboratory Supercomputing Center and an astronomer by training. "If you want to detect any variance or anomalies, you have to have a good model of the background."

Joining Kepner on the paper are: Kenjiro Cho of the Internet Initiative Japan; KC Claffy of the Center for Applied Internet Data Analysis at the University of California at San Diego; Vijay Gadepally and Peter Michaleas of Lincoln Laboratory's Supercomputing Center; and Lauren Milechin, a researcher in MIT's Department of Earth, Atmospheric and Planetary Sciences.

Breaking up data

In internet research, experts study anomalies in web traffic that may indicate, for instance, cyber threats. To do so, it helps to first understand what normal traffic looks like. But capturing that has remained challenging. Traditional "traffic-analysis" models can only analyze small samples of data packets exchanged between sources and destinations limited by location. That reduces the model's accuracy.

The researchers weren't specifically looking to tackle this traffic-analysis issue. But they had been developing new techniques that could be used on the MIT SuperCloud to process massive network matrices. Internet traffic was the perfect test case.

Networks are usually studied in the form of graphs, with actors represented by nodes, and links representing connections between the nodes. With internet traffic, the nodes vary in sizes and location. Large supernodes are popular hubs, such as Google or Facebook. Leaf nodes spread out from that supernode and have multiple connections to each other and the supernode. Located outside that "core" of supernodes and leaf nodes are isolated nodes and links, which connect to each other only rarely.

Capturing the full extent of those graphs is infeasible for traditional models. "You can't touch that data without access to a supercomputer," Kepner says.

In partnership with the Widely Integrated Distributed Environment (WIDE) project, founded by several Japanese universities, and the Center for Applied Internet Data Analysis (CAIDA), in California, the MIT researchers captured the world's largest packet-capture dataset for internet traffic. The anonymized dataset contains nearly 50 billion unique source and destination data points between consumers and various apps and services during random days across various locations over Japan and the U.S., dating back to 2015.

Before they could train any model on that data, they needed to do some extensive preprocessing. To do so, they utilized software they created previously, called Dynamic Distributed Dimensional Data Mode (D4M), which uses some averaging techniques to efficiently compute and sort "hypersparse data" that contains far more empty space than data points. The researchers broke the data into units of about 100,000 packets across 10,000 MIT SuperCloud processors. This generated more compact matrices of billions of rows and columns of interactions between sources and destinations.

Capturing outliers

But the vast majority of cells in this hypersparse dataset were still empty. To process the matrices, the team ran a neural network on the same 10,000 cores. Behind the scenes, a trial-and-error technique started fitting models to the entirety of the data, creating a probability distribution of potentially accurate models.

Then, it used a modified error-correction technique to further refine the parameters of each model to capture as much data as possible. Traditionally, error-correcting techniques in machine learning will try to reduce the significance of any outlying data in order to make the model fit a normal probability distribution, which makes it more accurate overall. But the researchers used some math tricks to ensure the model still saw all outlying data—such as isolated links—as significant to the overall measurements.

In the end, the neural network essentially generates a simple model, with only two parameters, that describes the internet traffic dataset, "from really popular nodes to isolated nodes, and the complete spectrum of everything in between," Kepner says.

The researchers are now reaching out to the scientific community to find their next application for the model. Experts, for instance, could examine the significance of the isolated links the researchers found in their experiments that are rare but seem to impact web traffic in the core nodes.

Beyond the internet, the neural network pipeline can be used to analyze any hypersparse network, such as biological and social networks. "We've now given the scientific community a fantastic tool for people who want to build more robust networks or detect anomalies of networks," Kepner says. "Those anomalies can be just normal behaviors of

what users do, or it could be people doing things you don't want."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology