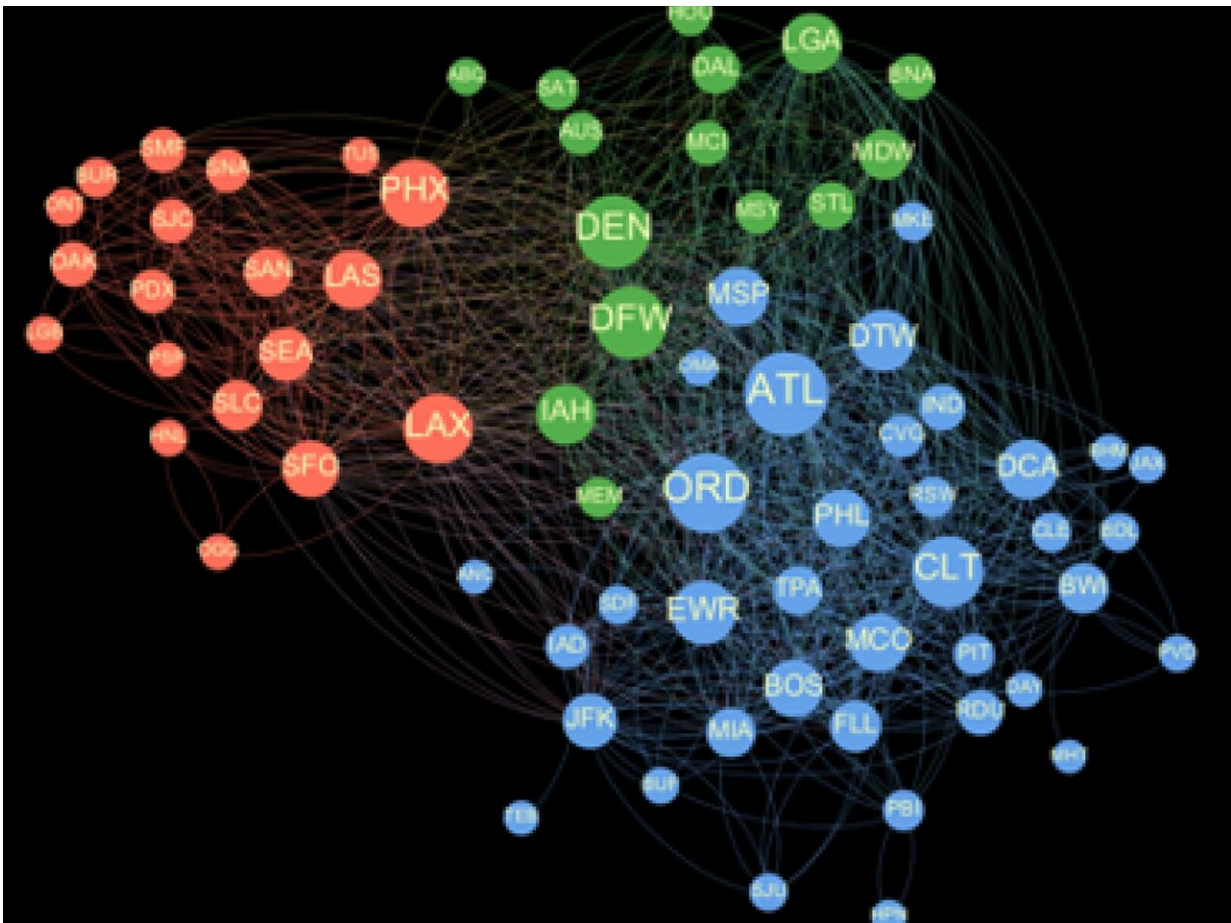# Protecting essential connections in a tangled web

February 20 2020, by Karyn Hede

Airport influence maximization network. Credit: Arun Sathanur, PNNL

It's winter. And as any frequent traveler knows, winter can mean airport

weather delays. A blizzard in Minneapolis, a major airport hub, can quickly lead to delays in balmy Miami or foggy London.

To minimize disruptions, air traffic control analysts work to prioritize recovery efforts. But with so many variables, it's hard for them to make confident recommendations. But this is just the kind of data-driven problem that a computer can be programmed to solve. The issue is time. Current methods aren't fast enough to provide solutions in real time.

Now, a research team led by computer scientists at PNNL has developed a new Graph tool, called Ripples, that can solve a complex graph analytics problem like airport disruption analysis in less than one minute on a supercomputer. The best comparable tool might take a whole day on a regular computer to solve the same problem. One day, the computing milestone may make analysis of network effects like air traffic disruptions available to real-time decision makers.

"Our approach leverages a rigorous social network analysis methodology, formally known as the influence maximization problem, and scales it to run on highly efficient parallel computing platforms," said Arun Sathanur, a PNNL computer scientist who led the airport modeling work. "These models excel at finding influential entities, analyzing the impact of connectivity, and pointing out where disruptions have the largest cascading ripple effect."

The research team, which also includes researchers from Northeastern University and the Department of Transportation's Volpe National Transportation Systems Center, presented their airport network analysis at the IEEE International Symposium on Technologies for Homeland Security in November 2019.

Using publicly available data provided by the Department of Transportation's Federal Aviation Administration, they grouped airports

into clusters of influence and showed which airports are the most influential, as well as how the most important "influencer" list changes throughout the calendar year.

The findings provide a proof-of-principle, which could eventually be used to manage airport network disruptions, Sathanur added.

"Ripples provides a powerful tool for proactive strategic planning and operations, and has broad applicability across networked transportation infrastructure systems," said Sam Chatterjee, an operations research scientist at PNNL and principal investigator for the airport modeling work led by Sathanur.
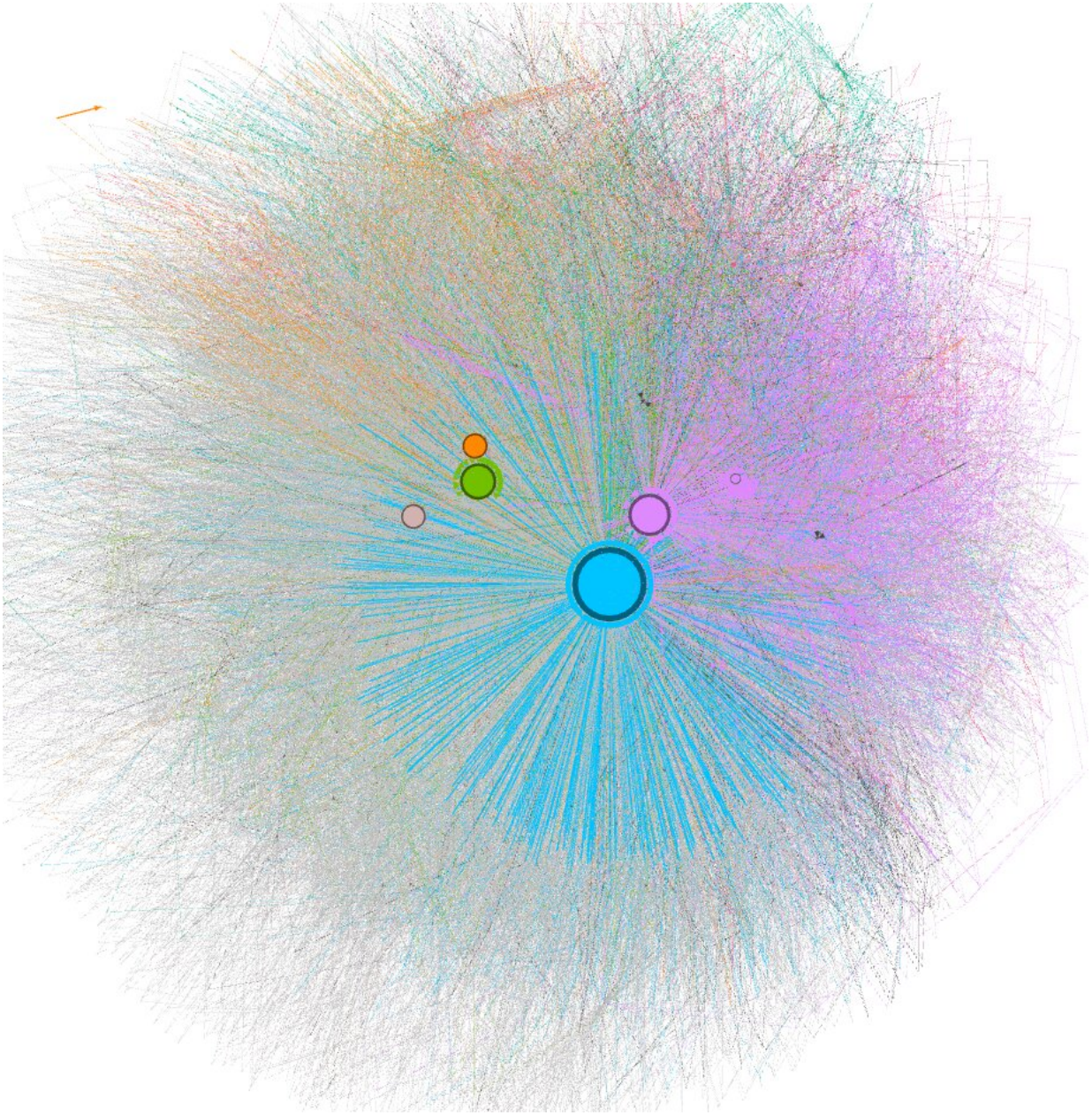
Representations of a complex system of atmospheric chemical reactions. Credit: Pacific Northwest National Laboratory

The ultimate logistics

In an increasingly congested world, being able to quickly restore service after accidental systems malfunctions or cybersecurity breaches would be a huge benefit. This is the realm of network analysis, which was first developed to understand how people in social networks are connected to one another. Increasingly, network analysis and visual analytics are being used to do things like spot unauthorized access to computer networks, detect relationships among proteins in cancerous tumors, and solve transportation congestion dilemmas like the airport network congestion problem.

However, for the analysis results to be trustworthy, a sequence of calculations to compute the influence spread must be performed. This turns out to be a computationally hard problem, said Mahantesh Halappanavar, senior scientist at PNNL and the principal investigator of ExaGraph, an applications co-design center funded by the Department of Energy's (DOE's) Exascale Computing Project.

"For many real-world scenarios, it is not always clear how to assign accurate weight to the strength of connections between individual entities in the network," he said. "We, therefore, repeat simulations with multiple settings to increase confidence of computed solutions." Even when the weights are well known, the method still relies on performing a large number of simulations to identify influential entities.

They estimate the most important influencers in any group by running these repeated simulations of an influence cascade model until they arrive at an accurate estimate. This approach is what makes it daunting to find even a small set of important influencers in a moderately large network, taking days to complete.

That's why Ripples' dramatic improvement in speed-to-solution is so significant.

"Zeroing in on the most influential entities in large networks can quickly become time consuming," said Ananth Kalyanaraman, a co-developer of Ripples and Boeing centennial chair in computer science at the School of Electrical Engineering and Computer Science, Washington State University, in Pullman. "Ripples, and its newer variant cuRipples, uses a strategy of exploiting massive amounts of computing power, including those in modern graphics processing units to seek the 'next most influential' entity during its search."
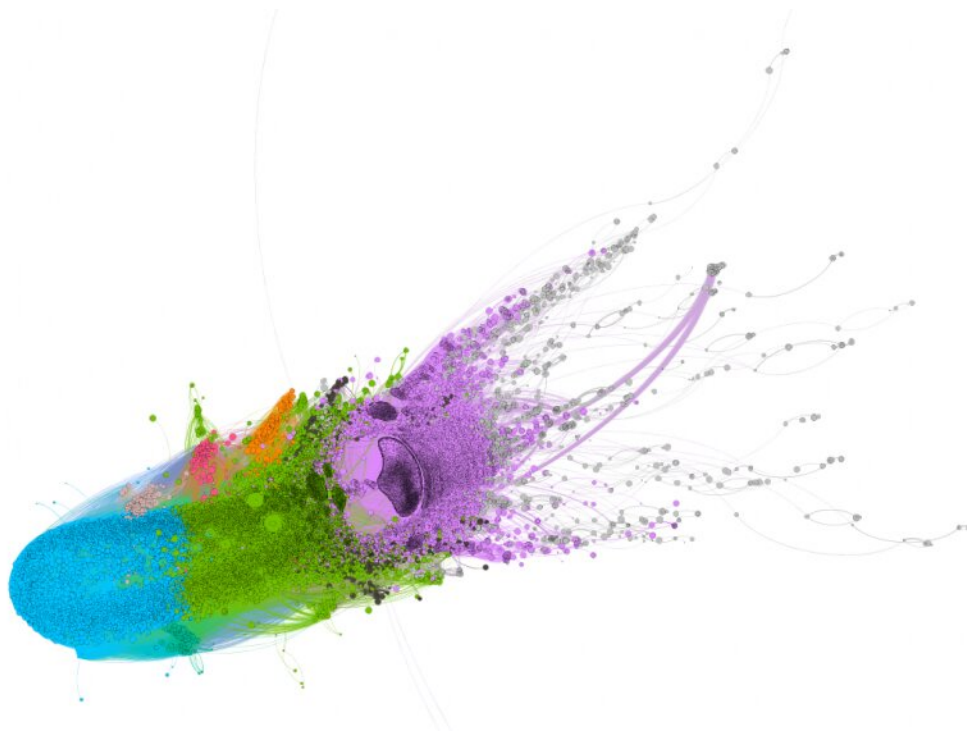
Protein similarity analysis using Ripples. Credit: Pacific Northwest National Laboratory

Reliable answers

Further, Ripples is based on the solution that comes with what's called an "approximation guarantee," which allows the user to trade off the quality of solution with the time to compute a solution, while also having the ability to judge the quality of the solution computed. The PNNL- and WSU-based teams worked closely together to scale the Ripples tool efficiently on the fastest supercomputers managed by DOE.

This strategy allows Ripples to efficiently converge on a higher-quality solution, up to 790 times faster than previous methods not designed for parallel systems.

"If we could converge on a solution in under a minute, we can begin to use this as an interactive tool," says Marco Minutoli at PNNL, the lead developer of Ripples. "We can ask and answer new questions in close to real time."

PNNL scientists are already doing just that. They have started to use Ripples to crunch massive amounts of data and find the most important influencers in:

- Identifying the most important species within a community of soil microorganisms as it responds to changes in moisture;
- Tracking the spread of infectious diseases and suggesting containment strategies to control the spread of an epidemic; and
- Identifying the most important components in air samples for inclusion in detailed climate models to study their influence in air pollution.

"To the best of our knowledge, this is the first effort in parallelizing the influence maximization operation at scale," said Minutoli.

The research team has made the method available for the research community on Github. They are planning the next major advance

(cuRipples), which will be to optimize the method on Summit, the world's fastest supercomputer.

Provided by Pacific Northwest National Laboratory