nutonian

CASE STUDY

University of Vermont Researchers Hunt Down Hidden Network Nodes with the Help of Eureqa™

Ever since the Internet and its offspring brought networks into the limelight, network science has been galloping forward, spurred on by an explosion of new interest and newly available data.

The field's increasingly powerful array of analytic and predictive tools have found application in social sciences, biology, physics, marketing, and just about every other field that casts its eye on collections of creatures or objects that interact with each other. The tools of network science, however, can only be relied upon to the extent to which the network in question is represented accurately. Getting the most out of those tools, therefore, depends on finding reliable ways to assess a network's accuracy and determine where any inaccuracies lie.

Much progress has been made in the detection of unrecorded relationships between nodes (i.e., people or objects) in a network. The trickier problem of finding entirely missing nodes is what drew the attention of James Bagrow, professor of mathematics and statistics, and Josh Bongard, professor of computer science, both at the University of Vermont. Bagrow and Bongard work together as members of the school's Complex Systems Center, and Bongard also directs the school's Morphology, Evolution, and Cognition Lab.

The Experiment

"We wanted to know if we could find a signature within incomplete data that would represent something that's missing," says Bagrow. "For example, I'm not connected on Facebook to my grandmother, but can we tell that my grandmother exists based on the Facebook activities of my brothers and I?"

Step one towards finding an answer was accomplished with the aid of a dozen hackathon participants. The team developed a method for simulating networks with properties similar to online social networks, both in terms of structure and in terms of how often each node "chose" to either create new content or pass along content created by another node. Next, Bagrow and Bongard created 100 of these networks, activated them, and measured the average time it took for content to travel between each pair of nodes within each network. Eureqa was then used to generate a model expressing node-to-node information transmission time as a function of a variety of statistics characterizing the individual nodes, their relative positioning, and global network properties.

That information-flow model was then used to predict transmission times for 200 new networks, but in half of those, one or more nodes were hidden; all mention of the node or nodes was erased from data seen by the model.

Bagrow and Bongard had hypothesized that their model's predictions would be significantly less accurate for networks with masked nodes, and that hypothesis was borne out, thus validating the use of information-flow models as tools in the detection of missing nodes. Furthermore, predicted and actual transmission times were found to deviate most strongly for pairs of nodes that surrounded a hidden node, so those local deviations could be used to determine a hidden node's most likely location.

Why Eureqa?

1. Portable models

Bagrow and Bongard were familiar with various machine learning approaches that could have been used in their experiment. Portability of output provided one reason to go with Eureqa. “Because Eureqa’s models are mathematical equations,” says Bongard, “once we had created a model, it was easy to then apply that model to new networks. With an artificial neural network, or some other ‘black box’ machine learning tool, that would have been a much more complicated process.”

2. Deep Transparency that Fosters Insight

Any mathematical equation is, of course, transparent in the sense that it is readable. But what attracted Bagrow and Bongard was transparency in a deeper sense. Bongard explains: “We could have used a support vector machine or some other state-of-the-art linear or nonlinear regression method to get a low order polynomial approximation, but that wouldn’t have given us any real insight into the nature of the relationship between information flow and network structure. We had some intuitions about what that relationship would be, and we were hoping to see some of those components in the equations that Eureqa gave us. And, indeed, we did see that.”

Eureqa delivered an equation expressing transmission time as a function of a dominant variable, length (the shortest topological path between two nodes), plus the log of several additional terms. “It made sense that transmission time was strongly correlated to length,” says Bagrow, “and a dominant variable plus a logarithmic correction is the kind of thing you see quite a bit in network dynamics.”

Close inspection of that logarithmic correction provided some food for thought. For example, Bagrow could see that, in many circumstances, an increase in the number of links to the nodes in question would increase the predicted transmission time. “That was a bit counterintuitive,” he says. “You would think that a big hub with lots of links should get information pretty quickly. At the same time, it could be that, if I’m a node with lots and lots of links, I can’t listen to everybody who is sending information at the same time.”

3. No Sympathy for Parameters that Don’t Pull Their Weight

One more reason the Vermonters chose Eureqa: parametric parsimony. They had come up with 18 statistics which they felt might be related to transmission time. They knew there was redundancy within the 18 in that many were strongly correlated, but, knowing Eureqa’s approach, they knew that wouldn’t be a problem. They fed Eureqa all 18 parameters and weren’t surprised to find only 6 used in the optimal model.

Coming Attractions: Germs and Brains

Bagrow and Bongard believe their approach will prove equally effective in the detection of other kinds of network misrepresentations — false nodes or merged nodes, for example. As they explore these possibilities they are branching out into new subject domains.

First, there’s epidemiology. They’re looking at a century’s worth of census data and data describing the spread of various diseases within the US. “I think modeling disease transmission with Eureqa could help us get some intuition as to why outbreaks of particular diseases occurred at particular times and locations,” says Bagrow.

And then there’s the big daddy of networks — the human brain. “We’ve begun using Eureqa for brain imaging studies,” Bongard says. “It’s similar to the social networks case in that we’re still talking about the movement of information. In this case, information takes the form of electrical patterns that appear in one part of the brain and a short time later in another. We can use Eureqa right out of the box to look at terabytes of brain-scan data and model that information flow. Deviations from the model can then tell us which parts of the brain need to be re-scanned at a higher resolution.”

About Eureqa

Eureqa is breakthrough technology that uncovers and explains the intrinsic relationships hidden within complex data. For more information or to get started with a free trial, visit **www.nutonian.com**.

THE RESEARCHERS

James P. Bagrow [1,2,3,4]
Suma Desu [1,2,3,4]
Morgan R. Frank [1,2,3,4]
Narine Manukyan [5,2,3]
Lewis Mitchell [1,2,3,4]
Andrew Reagan [1,2,3,4]
Eric E. Bloedorn [6]
Lashon B. Booker [6]
Luther K. Branting [6]
Michael J. Smith [6]
Brian F. Tivnan [6,2,3,4]
Christopher M. Danforth [1,2,3,4]
Peter S. Dodds [1,2,3,4]
Joshua C. Bongard [5,2,3]

ARTICLE

Shadow Networks: Discovering Hidden Nodes with Models of Information Flow.
http://arxiv.org/abs/1312.6122

RESEARCH QUESTION

Can models of information flow across networks be used to detect hidden or undiscovered network nodes?

CHALLENGE

The researchers wanted a model relating information flow to network structure that would not only be accurate but would also provide insight into the relationships being modeled.

SOLUTION

Eureqa delivered an accurate model with a general structure that resonated with the researchers' expert intuitions and which contained details that led to new insights.

RESULTS

The researchers have demonstrated a viable approach to the detection of hidden network nodes and have taken an important step toward the development of a general method for detecting and locating inaccuracies in network mappings.

KEY FEATURES FOR THE RESEARCHERS

- Eureqa's models are simple equations and can therefore easily be hardcoded into scripts and applied in novel situations.
- Model transparency provides insight into the structure of relationships within the data set.
- Eureqa's search process filters out unimportant or redundant parameters, thus allowing a liberal approach to the inclusion of parameters for consideration.
- Inclusion or exclusion of particular mathematical "building blocks" during a formula search allows leveraging of expert knowledge for increased search efficiency.
- Outliers in the data set can be easily toggled into and out of consideration.

NOTES

[1] Department of Mathematics & Statistics, The University of Vermont, Burlington, VT 05401, USA
[2] Vermont Complex Systems Center, The University of Vermont, Burlington, VT 05401, USA
[3] Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401, USA
[4] Computational Story Lab, The University of Vermont, Burlington, VT 05401, USA
[5] Department of Computer Science, The University of Vermont, Burlington, VT 05401, USA
[6] The MITRE Corporation, McLean, VA 22102, USA