New technique for dataset cluster detection

November 27 2013

(Phys.org) —A persistent problem for mathematicians trying to understand the structures of networks – in datasets representing relationships among everything from galaxies to people – is community detection: finding groups of related data points, or nodes. A network that contains three groups of nodes is fundamentally different from a network that contains two groups.

"If you're looking at who eats whom in a food web, you need to understand how groups of predators depend on groups of prey. In an epidemic model, you need to know how fast a disease will spread in one community and how likely it is for it to cross to another community," says SFI Professor Cris Moore.

In a paper published this week in *PNAS*, Moore and collaborators at UC Berkeley and in Paris offer a twist on past approaches to cluster detection that seems to address weaknesses that traditional techniques have in dealing with sparse data—networks where most nodes have just a few links.

Traditionally, mathematicians find communities in one of two ways: statistical inference, a highly iterative method that reassesses network-wide probabilities at each step, and spectral analysis, a faster "random walk" technique that groups nodes by focusing on the flow of information or probability through a network.

To understand the latter approach, imagine taking a walk in a city. At each intersection you flip a coin to decide which way to turn. As time

goes on, the likelihood that you are at another given point in the city can be expressed as a probability. If the city is divided by a river into two parts with just a few bridges between them, it will take a long time for your probability to flow from one side of the river to the other. This slow "flow of probability" helps reveal that the city is divided.

Both techniques – statistical inference and spectral analysis – work well for networks with many links between nodes. But in sparse networks where most nodes are related to only a few others – as in the case in many real-world networks – classic spectral techniques fall short.

Statistical inference, in fact, finds groupings as well as any algorithm can – all the way down to a theoretical limit revealed by Moore and collaborators in a 2011 paper. And it can handle networks with millions of nodes in minutes. But spectral techniques are even faster, partly because they rely on simpler linear equations to update their information rather than more complex and difficult-to-crunch nonlinear ones.

Further, statistical methods tend to slow down when a network features a large number of clusters. And statistical techniques can go off on wild tangents in situations where they get an early but erroneous picture of network structure; often, in fact, spectral techniques are used to provide a rough snapshot of a network that statistical methods can refine.

The challenge has been, then, to find a spectral method that is computationally efficient, but that also finds groupings in networks down to the theoretical limit.

In the paper, aptly titled "Spectral Redemption," the researchers try out a spectral method they call the "non-backtracking operator." Put simply, it specifies that during analysis, information flowing from node to node may not immediately return from whence it came.

Think of a pinball machine, with the ball as information and bumpers as nodes. Often the ball gets stuck dinging around inside a group of bumpers.Eventually it frees itself and rolls across the board, only to get trapped temporarily inside another localized group of bumpers.

"Traditional spectral methods get stuck on highly connected nodes, rattling back and forth between those nodes and their neighbors," Moore says. "They get confused by localized structures in the network rather than finding the large-scale structures we care about."

In the pinball analogy, the non-backtracking rule would prohibit the ball from returning to a bumper it had just bounced from. In network clustering, this seemingly minor modification is enough to ensure that information doesn't get hung up in one area of the network for very long, and sooner paints a picture of the various clusters present in the system.

"Each hub gets the attention it deserves, but not more," he says.

The researchers tested their non-backtracking technique on several network datasets commonly used to benchmark clustering methods, including several real-world networks. They found that their spectral method is reliable down to the theoretical limit.

"This work shows how crucial it is to build connections between scientific communities," says Elchanan Mossel, a key collaborator at UC Berkeley. "Bringing together concepts, methods and points of view from statistical physics [Santa Fe and Paris] and mathematics [Berkeley] gave us a whole that is much greater than the sum of the parts."

More information: "Phase transition in the detection of modules in sparse networks." Aurelien Decelle, Florent Krzakala, Cristopher Moore, Lenka Zdeborová. *Phys. Rev. Lett.* 107, 065701 (2011). DOI: 10.1103/PhysRevLett.107.065701

Provided by Santa Fe Institute