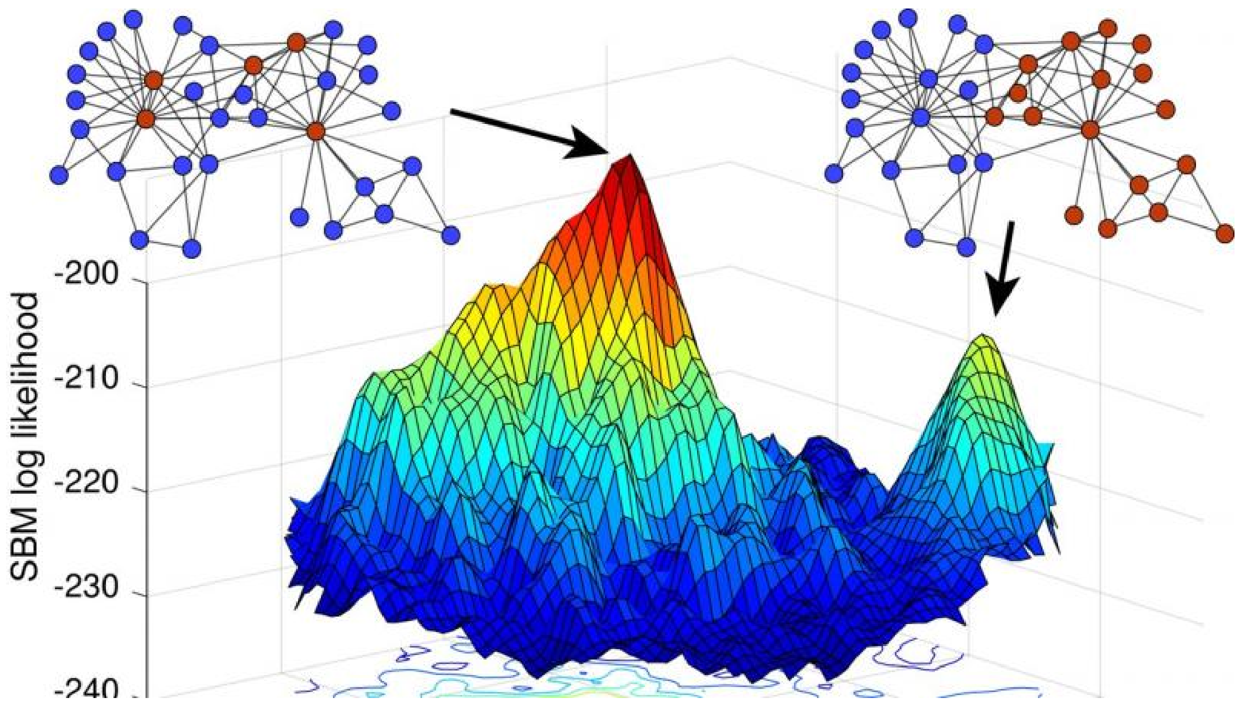
Does my algorithm work? There's no shortcut for community detection

May 4 2017, by Jenna Marshall



Metadata is not ground truth. In the space of all possible partitions of a real-world social network, the lower peak corresponds to the social group partition given by the metadata. The higher peak corresponds to a leader-follower partition within the network. Image courtesy Peel, Larremore, and Clauset. Credit: Santa Fe Institute

Community detection is an important tool for scientists studying networks. It provides descriptions of the large-scale network by dividing its nodes into related communities. To test community detection algorithms, researchers run the algorithm on known data from a real-world network and check to see if their results match up with existing node labels—metadata—from that network.

But a new paper published this week in *Science Advances* calls that approach into question.

Real-world networks are large and complex. Food webs, social networks, or genetic relationships may consist of hundreds, or even millions, of nodes. To understand the overarching layout of a large network, scientists design algorithms to divide the network's nodes into significant groups, which make the network easier to understand. In other words, community detection allows a researcher to zoom out, seeing big patterns in the forest, instead of being caught up in the trees. In the past, researchers have used metadata as a sort of answer key or "ground truth" to verify that their community detection algorithms are performing well.

"Unfortunately, tempting as this practice is, with real-world data, there is no answer key, no ground truth," explains Daniel Larremore, one of two lead authors of the paper and an Omidyar Fellow at the Santa Fe Institute. "Our research rigorously shows that using metadata as ground truth to validate algorithms is fundamentally problematic and introduces biases without telling us what we really need to know: does my algorithm work?"

When scientists use metadata to validate algorithms, they limit the types of communities they can validate. Larremore likens this to a teacher leading a class discussion, and only responding to students who raise points the teacher is already familiar with.

"If we want creative algorithms that can handle all kinds of challenges, then restricting the answers to one set of "ground truth" metadata means we're pushing our algorithms through this bottleneck of low diversity, and low creativity," he says. "We'll only ever get algorithms that solve a small and restricted set of problems."

Having exposed the shortcomings of metadata as a test for community detection, Larremore and co-authors Leto Peel (Université Catholique de Louvain) and Aaron Clauset (SFI, CU Boulder) go on to quash any hope of creating a universal algorithm for detecting communities by their network structures. The paper mathematically proves the first No Free Lunch Theorem for community detection: any algorithm that's exceptionally good at finding communities in one type of network must be exceptionally bad at finding communities in another.

David Wolpert, also of the Santa Fe Institute, first posited a No Free Lunch Theorem for machine learning algorithms in 1997.

The authors hope that by mathematically proving the futility of universal detection algorithms, they can, according to Larremore "free people up to work on specialist algorithms."

The new paper curbs enthusiasm for finding any single, universally optimal approach to understanding complex network datasets. Still, the authors do see a constructive side to their findings. In the final section of their paper, they reverse the usual script. Instead of using metadata to validate an algorithm's performance, as in the past, they introduce two new statistical approaches that use metadata in conjunction with the network itself to probe the more fundamental questions of network science: what are the deeper patterns between the nodes, links, and metadata alike, and how can we use these to learn about the system that the network represents?

More information: Leto Peel et al. The ground truth about metadata and community detection in networks, *Science Advances* (2017). DOI:

10.1126/sciadv.1602548

Provided by Santa Fe Institute