Using Community Detection Algorithms for Sustainability Applications

Sucheta Soundarajan and Carla Gomes

Cornell University, Ithaca NY 14853, USA
sucheta@cs.cornell.edu

In computer science, social network analysis is becoming an increasingly important way of understanding the links and relationships between individuals. Although classically a social science area, the field has more recently attracted attention from mathematical scientists such as physicists, statisticians, and computer scientists, who have contributed powerful methods for analyzing large networks, with important consequences for fields as diverse as politics, epidemiology, and economics. Social network analysis has generally not been applied to topics in computational sustainability; however, it likely has the potential to dramatically alter our perspectives and approaches to a multitude of sustainability problems.

Although the term "social network" is often used in reference to online social networking sites, such as Facebook or Twitter, the field of social network analysis studies a far broader set of networks: indeed, any system containing individuals and interactions (or, more generally, relationships of any sort) can be represented by a network. In standard social networks, these individuals are people, and the relationships might be friendships. In a different network, the individuals could be fruit fly genes, and we might say that two genes are related if they act toward the same biological function. Networks can be used to study the social interactions of wolves, the migration of birds, the impact of humans on fish populations, or the spread of an arboreal disease through a forest. Study of these networks, then, might assist in developing strategies to preserve wolf populations, identifying locations for wind turbine placement so as to not interfere with bird travel, calculating the number of fish that an individual fisherman should be allowed to harvest, or determining which trees ought to be inoculated against the disease.

As an example of how network analysis methods may be useful to sustainability applications, consider the contrast between network components and network communities. A connected component in a network is a set of vertices such that every vertex in that set is reachable from any other vertex in the set. In contrast, while there is little consensus on a mathematical definition for community, resesarchers have generally understood it to be a set of vertices that are well connected to each other (and, in some cases, poorly connected to the rest of the network) [1][2]. Consider, for example, two large sets of vertices, each of which is internally well-connected (i.e., vertices in one set tend to be connected to other vertices in the same set), that are connected by a single edge. These vertices form a single connected component, but consist of two communities.

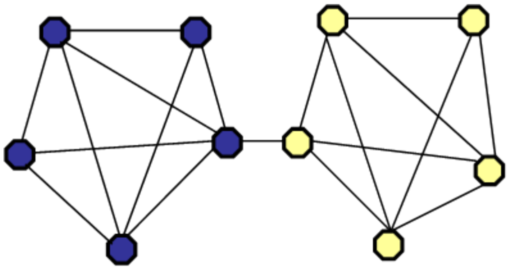

Fig. 1. A Network Containing One Component with Two Communities

In the rest of this abstract, we give specific examples of sustainability problems that may benefit from the application of social network analysis, and end with a brief discussion of preliminary results and future work.

1 Habitat Preservation

Consider an animal species whose habitat has become fractured, perhaps because humans have encroached upon its environment, or a species in which groups of animals tend to congregate around certain points (such as breeding areas), with few individuals in the areas between those points. For the purposes of habitat preservation, one might wish to identify which of these locations is most crucial (or "central") to the network. A typical approach to this problem might consist of first determining the distance D that typical animal will travel from its 'home' location and then creating a network in which the vertices represent locations, and two locations are connected by an edge if they are within D distance of each other. One could then identify connected components in this network and then find vertices whose removal would split a previously connected component into multiple pieces. These vertices are most important to the overall connectivity of the network, and so we would conclude that they ought to be preserved (as opposed to vertices near the outside of a component, whose loss would affect only the animals at that single location).

However, this method's focus on connected components is a somewhat coarse approach, as it does not take into account the strength or quantity of connections between different locations. Using more sophisticated community detection algorithms will give us a more finely-grained understanding of actual animal movement, and will likely give us deeper insight into developing effective conservation policies. For example, in the above example of two well-connected vertex sets connected by a single edge, it is likely in many cases that animals will stay in their own set, and so it may be more prudent to focus on connectivity within each of the two vertex sets rather than connectivity between the two sets.

2 Animal Genetics

Researchers may have access to genetic data from a population of animals. Using such data, one can create a network of genetic relationships by setting some threshold of genetic similarity (i.e., if animals exceed a certain amount of genetic similarity, then we consider them ‘related’). In this network, we can identify communities of genetically similar individuals, and then analyze the genetic interactions between these groups. For instance, we might consider whether they are mostly separate, suggesting that there is little genetic intermingling between family groups, or whether there is gene flow between the groups, indicating some degree of interbreeding. Are these results supported by biological and ecological studies? What are the implications for the health of the species overall: Is there sufficient gene flow to maintain genetic diversity? Are any groups particularly isolated from the rest of the population?

3 Wildlife Corridors

For many threatened or endangered species, the surviving population is fractured into separate groups. This separation may lead to loss of genetic diversity and decreased resistance to disease. Thus, ecologists and biologists create “wildlife corridors,” paths of land that connect the isolated populations. A common solution to this problem is to find the cheapest corridor that connects all populations. If each community is equally likely to interact with every other community, then it makes sense to create connections between all populations. On the other hand, if the network of communities demonstrates, for example, a hub-and-spoke structure with central and fringe communities, then this information can be taken into account when designing the corridors. As described in the previous section, we may be able to identify isolated groups of animals. These groups, in some cases, might appear to be geographically close to the rest of the population, but their genetic isolation may demonstrate a need for additional corridors to and from that group.

4 Social Network Methods

While existing social network analysis methods have the potential to be broadly applicable to various topics in sustainability, it is probable that these applications may help direct the development of new social network analysis tools. Current social network analysis methods and knowledge have been developed primarily through study of online networks such as Facebook. These networks have given us deep insight into how certain people act, particularly those who are young and live in a First World country, but it is unclear how well they can be generalized to other networks. To make existing analysis tools more useful to specific sustainability topics, we will need to quantify the differences between these “classical” networks and those networks that arise from sustainability domains.

5 Conclusion and Future Work

The development and use of social network analysis techniques has revolutionized our understanding and use of many human social networks. To apply such methods to topics in computational sustainability, we must first identify important features of various sustainability networks, and use these features to design appropriate analysis tools. Initial efforts at applying such tools to real data, such as genetic networks, has indicated that use of community detection methods may provide dramatically better results than simpler methods. For instance, early results suggest that a population can be clearly split into distinct communities, even when those communities have a great deal of geographic overlap.

A main challenge in this area is for computer scientists to obtain access to and permission to use large quantities of data, particularly because data from many sustainability domains (e.g., genetic information) are often difficult and expensive to collect. It may thus be useful for researchers to create generative models for producing artificial data. A good generative model will create a synthetic network that is similar to real networks in important ways, and may be particularly useful for understanding how and why networks evolve.

References

1. Newman, M.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103, 8577–8582 (2006)
2. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814–818 (2005)