Mathematical Lacks in Network Analysis

Jennifer Webster and Stephen Young
Pacific Northwest National Laboratory
Report Number PNNL-SA-124003

Historians, anthropologists, epidemiologists, and many others have been interested in social networks and community interactions for centuries – struggles of political groups, the spread of disease, the birth of new ideologies. Some of the most common tasks are to identify communities within a larger population, to identify new communities as they emerge from the existing structures, and to identify influential individuals within the network. Trained social scientists with sufficient evidence can answer many of these questions, but with the growth of the Internet and the rise of social media, we are now being asked to meet these needs in real time with limited data. As a result, network analysis as a computational discipline has grown as well and many mathematical formulations[1] have been proposed to identify the structures of social networks (Kadushin; Newman). However, the methods proposed by graph theorists, physicists, and computer scientists, among others, are sensitive to how the data is collected, the choice of network representation, and the precise definitions of "community" and "influential". In addition, electronic data sources provide a wealth of primary source material that can easily tax the algorithms developed by this field, the visualization software resources, and classic human-driven analysis techniques as the size of the population to be studied and the amount of information produced by each subject increases. While there is definitely an ongoing interest in the analysis of these networks structures, the mathematical gaps in network analysis should be addressed to facilitate a more rigorous analysis of the networks produced in this large data environment.

[1] Over 1000 documents on arXiv for Computer Science – Social and Information Networks topic area in 2016

The mathematical shortcomings of network analysis can be lumped into three primary areas: uncertainty and stability, incorporation of expert data, and temporal evolution.

Uncertainty and Stability: With the ubiquitous nature of online social networks (OSNs), there is a substantial amount of data regarding various communities and their relationship. However, this proliferation of data is not without cost, as many of the data sources do not accurately report the sociological flavor of the relationships present within the OSNs. For instance, a "friendship" on Facebook can easily have a variety of sociological interpretations, representing professional, social, familial, or other networks. Furthermore, snapshots of these networks have several sources of noise, including "spoofed" accounts, spam accounts, and the time decay of relationships. This uncertainty in the meaning and accuracy of the underlying data will naturally propagate to any subsequent network analysis. However, there has been relatively little mathematical analysis dedicated to understanding how the presence of uncertainty influences the accuracy and stability of the identified communities.

Expert Data: In many social systems, there exist multiple levels of granularity for community structures within the system, and these structures need not form a hierarchy when comparing different levels of resolution. For instance, in the US political system there are clearly two top-level communities of Republicans and Democrats, but there are also finer grained communities defined by state and regional issues, which can be crosscutting or entirely contained within the top-level communities. However, modern social network analysis (SNA) techniques give analysts rather poor tools to explore communities at a given level of granularity or within a given relationship. Typically there are a few poorly understood hyper-parameters that can be modified randomly or there is a course-grained measure such as the number of communities that can be fed into the algorithm. Both of these options neglect to fully exploit the subject matter expertise regarding the social system and thus produce inferior results.

Temporal Evolution: Some of the most interesting analysis of communities focus on the dynamic evolution of the system as driven by it's individual members,

e.g., tracking the development of the US political system from the founding to the current party divisions. The majority of current SNA techniques view the network as a single snapshot in time and do not consider the ongoing dynamics. The temporally aware techniques that are available are typically ad-hoc methods that attempt to stitch to together a series of snapshots into a coherent whole by aligning the identified communities across snapshots (Hartmann, Kappes and Wagner). However, these ad-hoc techniques fail to capture the dynamics that generate the evolution of communities and hence are not predictive. It is worth noting that capturing the temporal evolution of social networks is likely to be an extremely challenging problem as the drivers for community behavior are in the core/static part of the community and the evolution of the community typically occurs on the periphery/dynamic portion of the community.

Although these are acknowledged issues in network analysis, they have been for the most part unaddressed in the academic literature. In part, this is because of a lack of relevance to the financial interest of the large technology companies, such as Facebook, Yahoo!, and Google, who provide a significant source of funding for research in network analysis. In the case of technology companies, the goal of the SNA is (typically) to sell targeted advertisements using the full scope of private data gleaned from the network. Community assignment algorithms facilitate this in two ways, by providing "friendship" recommendations and by defining large groups that may be addressed with targeted advertisements. Because of the relatively small "cost" (opportunity or otherwise) associated with displaying a targeted advertisement to someone who is only weakly a member of a community (or only looks like they are member, but in fact, are not), the issues of uncertainty and stability are relatively unimportant to advertising decisions. Certainly, the gains possible by considering uncertainty and stability are small in comparison with the identification of large, likely communities, and thus are only of minor importance. Similarly, the precise relationship between communities and their evolution over time are of limited relevance to the advertising market, as there is again limited benefit to advertising strategies with the exploitation of these ideas. In many cases,

advertising agencies can accomplish their goals without needing to predict how communities evolve over time.

This is in stark contrast to the needs of the IC, where the robustness of results under uncertainty and the ability to predict the manner in which various communities will evolve is of the utmost importance, and where the wealth of expert knowledge is a vital resource that must be exploited to enhance the community identification schemes. Currently, the IC is compensating for the uncertainty in data with the experience of individuals and as such is restricted in the amount and type of behaviors it can reliably detect. In many ways, industry has an easier problem to solve, as it is sufficient to find correlations in behaviors. However, the IC and the broader social science community are concerned with motivating factors and the growth of new ideas and organizations. By developing robust and purpose-driven mathematical algorithms, which can address the uncertainty of data, the abundance of expert information, and the temporal nature of communities, one can better describe and predict the motivations of a community and it's members. Industrial partners have already invested in the mathematical and computational science communities to strengthen network analysis algorithms for their purposes, making it an ideal time for strategic partnerships from the IC to leverage the existing research capabilities for the development of algorithms tailored to the unique goals of the IC.

Bibliography

Hartmann, Tanja, Andrea Kappes and Dorothea Wagner. "Clustering Evolving Networks." arXiv:1401.3516 [cs.SI] (2014): 46p.

Kadushin, Charles. Understanding Social Networks. Oxford: Oxford University Press, 2012.

Newman, M.E.J. Networks, An Introduction. Oxford: Oxford University Press, 2010.