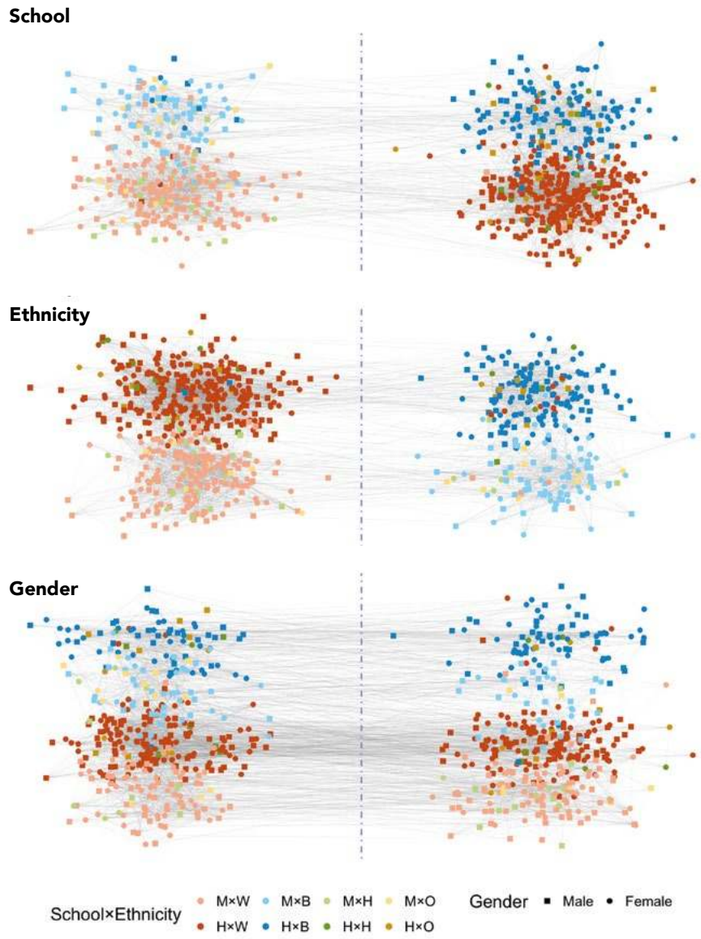
A needle in a haystack – the future of big data

Dr Yang Feng *is Associate Professor of Statistics at Columbia University. His research aims to structure, into a useful form, the voluminous data available from many areas of science, humanity, industry and governments, like social networks, the study of the genome, understanding economics or finance and health sciences. Using network modelling, he has focused on novel ways of detecting "communities" more accurately by using available nodal information. Dr Feng's approach is underpinned by rigorous theory and its effectiveness is demonstrated using simulated and real networks.*

A network is a way to represent information and is underpinned by mathematical methods that are well understood. Networks are groups of nodes interconnected by links, or edges, that can be directed (from one node to another) or undirected (two way). Web pages are examples of directed networks with the page representing a node and a hyperlink as an edge. Dr Feng uses networks to find "communities" more accurately. These are nodes that are densely connected as a group but have few connections to other groups, like people in social networks with similar interests or researchers collaborating within a scientific field. Of interest to Dr Feng are "covariates" from the data under study as they may help to improve the accuracy for identifying communities.

Figure 1: Community detection results when considering school, ethnicity, and gender as the ground truth. Predicted communities are separated by the middle dash line.



COMMUNITIES AND NODAL INFORMATION

Recognising communities within networks clarifies their structure, offering practical benefits; for example, social network groups share similar interests so recommendations can be better targeted. Broadly, current methods for identifying communities within data sets are either algorithmic, relying on derived computer programs, or model based, using statistical methods, a common one being the stochastic block model. This is a model that assumes the nodes inside the same community behave identically when interacting with other nodes. For example, if persons A and B belong to the same community, they would exhibit similar behaviour when communicating with any other person C.

In real networks, nodes contain properties that can help pinpoint community structures within the data. As examples, social networks have their user profiles attached to nodes and cited scientific papers contain author information, keywords and abstracts. Dr Feng considered that this kind of covariate information, combined with edges, could better infer the existence of communities, through the two different relationships described in Figure 1.

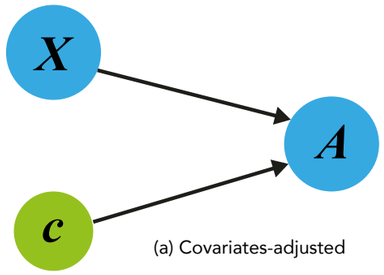
ASYMPTOTIC APPROACH

Dr Feng's work introduces a flexible statistical model, tuned to identify communities using the structure of the network, its nodes and edges, and nodal properties that represent the covariate information. The model uses a grounding of network mathematics to create matrices for the network and its connections, plus the nodal properties (the covariates) and uses an iterative approach to identify the communities. One challenge was that a pure likelihood-based approach was sensitive to the initial solutions so an alternative had to be developed. This involved finding well-behaved initial values for the model using optimisation techniques. These worked better than random initialisation.

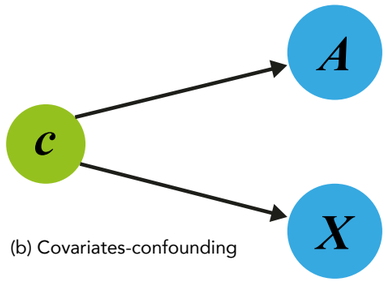
Recognising communities within networks clarifies their structure, offering practical benefits, like better recommendations in web searches

NOW, TO REAL LIFE

It was time to test the model on some real data so an example was chosen comprising a research team of 77 employees working in a manufacturing company. To create a network, consider the employees as nodes with their links, or edges, being how much they interact to allow them to do their work. These



(a) Covariates-adjusted



(b) Covariates-confounding

Figure 2: Two different relationships among nodal information X, community information c and the observed adjacency matrix A.

links are weighted; if Helen co-operates with Joseph then the weight might be based upon their interaction: 0: none, 1: very infrequent, 2: infrequent, 3: somewhat infrequent, 4: somewhat frequent, 5: frequent and 6: very frequent. The dataset contained other attributes about each employee, in particular their country location and level in the organisation.

The data represented a weighted, directed network and needed to be converted to a binary undirected network. The final model used the frequency of communication to isolate those who didn't communicate often and included properties from the database to seed it. The attribute "location" was a "ground truth" as disparate location implies lower interaction and this could be used to test the effectiveness of the process. Dr Feng found that by incorporating the nodal properties, community detection accuracy was improved and semi-definite programming performed as well as two of the newly proposed likelihood-based methods.

Another more complex example used a 'friend network' at a USA high school, taken from the National Longitudinal Study of Adolescent to Adult Health comprising 795 students between nine and 12 years of age at the high school and between seven and eight years at the feeder middle school. The collection had multiple covariate properties of grade, gender, ethnicity and number of friends nominated (up to ten). In communities like these the nodal information (like age or ethnicity) can often infer a 'ground truth' for the community identifier.

In real networks, nodes contain properties that can help pinpoint community structures within the data

The final dataset, after removing those with missing covariates, had 777 nodes and 4,124 edges. Dr Feng and his team used their model with this dataset as it contained multiple category variables; anyone could be considered as the community of interest with the other variables controlled to make a prediction. School, race and gender were used as the ground truth (students at the same school are more likely to be friends, for example) and the other two properties were controlled when detecting a community. The results presented in Figure 2 suggested an accurate community detection for school and ethnicity but did not perform as well when using gender as the true label. Dr Feng concluded that the gender result was no better than random and that there was the possibly of another covariate available that had not been identified. Finally, Dr Feng examined community detection of this dataset using standard network models (like the stochastic block model) and concluded that detection is poor; naively applying them to detect communities would lead to unreliable findings.

LOOKING AHEAD

Dr Feng's work has demonstrated the viability of using statistical models to detect communities in networks that are built from network data that includes attributes provided at each node, like a person's interests or location. It has real world applications in many disciplines or industries, like forensics or drug selection from genetic data, and holds promise when using a wider family of node properties or a greater number of them and in networks that have low densities of communities.

Behind the Bench

Dr Yang Feng

E: yang.feng@columbia.edu **T:** +1 (212) 851 2139 **W:** http://www.stat.columbia.edu/~yangfeng/

Research Objectives
Dr Feng's research focuses on proposing statistical models for networks and developing efficient estimation methods. He is currently interested in finding out how the available nodal information can help with community detection in networks.

Funding
• NSF

Collaborators
• Sihan Huang
• Haolei Weng

Bio
Yang Feng is an associate professor of statistics at Columbia University. His current research interests include high-dimensional statistical learning, network models, nonparametric and semiparametric methods and bioinformatics. He is currently an associate editor for Statistica Sinica, Computational Statistics & Data Analysis, and Statistical Analysis & Data Mining.

Contact
Dr Yang Feng
Associate Professor
Department of Statistics
Columbia University
1255 Amsterdam Ave.
10th Floor, MC 4690
New York, NY 10027
USA

Q&A

What would you say was the strength of the approach taken by your team in comparison to that taken by others in your field?
Compared with the existing approaches, the proposed approach is intuitive, can be computed efficiently, and has solid theoretical justification of its performance.

Why did you choose semi-definite programming and how did you ensure that the computational load was achievable?
The semi-definite programming (SDP) approach is a popular method for relaxing a NP hard problem to a convex one. By using SDP, the computation becomes feasible through a well-known algorithm called ADMM. At the same time, we provide theoretical justification on the solution of SDP. Empirically, we observe using the SDP solution as the initial solution to our likelihood-based methods can improve the estimation accuracy significantly.

How do you see your research being used in health care and its services?
I see this research framework to be potentially useful in the health care domain, where precision medicine is the current trend to ensure everyone receives the personalised treatment that is best for each individual. If we can collect the network information among different patients along with their personal information, the proposed method may be used to detect different communities among patients. It is possible that we may want to use different treatments for patients that are in different groups.

There's a great deal of mathematics in your papers. Could you summarise how the mathematics helped you develop your solution and test the accuracy of your findings?
Indeed, a lot of math was used in this research project. We use likelihood-based approaches to detect communities and this naturally leads to the study for the maximisers of the likelihood functions. Quantifying the theoretical properties of those MLEs requires various techniques from mathematics and statistics.

How would you like to see your research developed further and to what practical benefit?
Currently, I am working to further develop this project by integrating the network information to improve prediction. This would require us to study the regression problem under dependent sample where the dependency is characterised by the network. I hope this research will lead to improvements in personalised recommendation and advertisement targeting.

I hope this research will lead to improvements in personalised recommendation and advertisement targeting