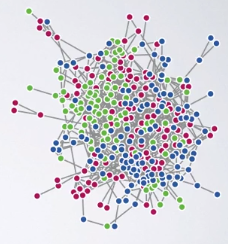
**Computational Methods for Modeling and Analyzing Complex Networks**

**by Aaron Clauset**

As the ‘big’ data buzzwords and trends sweep the CS industry, researchers struggle to make sense of the mass data being made available. Inspired by the availability of this new ‘big’ data, Aaron Clauset explains and explores the applications of modeling complex networks to discover and refine existing processes and analyze them further. By linking the unique components of data with relational vectors, the graphs can be organized and divided to produce pragmatic results.

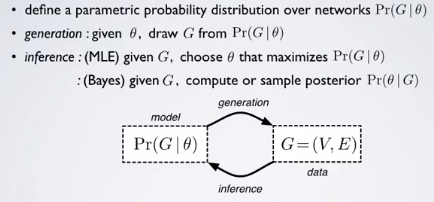
How can one construct a graph? Clauset describes *edges* to be the link between any two paired things and *vertices* as the nodes or entities which the relationship is formed upon. This provides the basis of a graph, with additional values and groupings providing further levels of distinction. The pairing of two nodes is the most basic form of a graph.



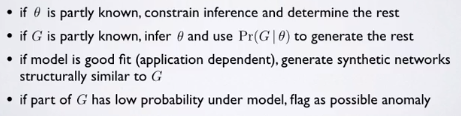
*(The coloured nodes act as the unique vertices with the grey links between them as the edges; the colours add an additional weight to each vertex to further subgroup them)*

The main goal behind modeling the datasets is to acquire a graph with relational structured data instead of random noise; a graph of data that clearly and correctly relates the vertices to one another through which we can discover insights to the myriad of data sets we have. The advantage of such networks is that they allow us to discover relationships between one part of the network to another and even between networks of the same type. To flesh this out, let’s take an example of a wood working community: whilst it’s expected that wood workers would have a tight and consistent interaction with lumberyards and tool workers, when modeling their interaction corpus an unexpected strong relationship could be identified with acid producers who are needed for the restoration and repair of specialty tools such as rasps and abrasives. Identifying such relations would help in marketing and integrating the processes between these two disparate markets. Furthermore, it allows us to view relations on a smaller scale vs. a larger scale. We can see how specific individuals and nodes interact with one another as well as how large groups interact with other groups (effective in understanding for e.g. intercultural communication, military divisions and organized/unorganized groups interactions within themselves and outsiders). Utilizing these inferences and predictions, we can interpolate and extrapolate on past and future occurrences to either make sense of them or to predict their occurrence beforehand.

To create these models, we follow a principled approach as well as making use of generative models to analyze networks. By drawing instances from distributions that depend on Bayes or likelihood models, we can predict the probability of specific models and fine tune them. To verify the accurateness of the model, we can generate data (once we have created the initial models) to check if the synthetic data produced actually creates a similar network that we can infer information from and whether we’re getting decent predications and extrapolations from our models.



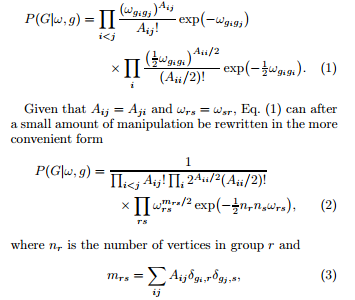
An advantage of utilizing these models is that we can infer the structures by extrapolating on some of the partly known data; we can complete grids and networks. This allows us to do otherwise dangerous or unethical research on existing systems (like testing a rewiring the electric grids of a nation more efficiently or testing disease control techniques without human testing).



*(methods of extrapolating on partial data using the probability distribution models)*

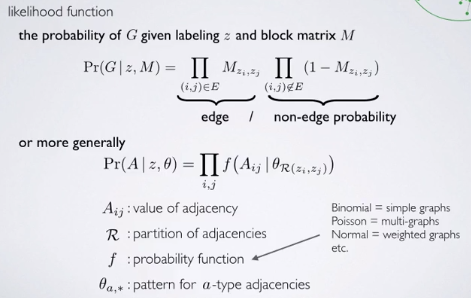
There are many advantages that come with being able to infer from partial data using our models. Using these models allows us to learn with auxiliary info by augmenting it to our data and running our models. Furthermore, these models are scalable for future expansion or contraction of the size of data. This also allows us to differently model the selected data (communities etc) by changing the variable factors (frequency of interaction vs. length of interaction). Lastly, this model would allow ranking our vertices and edges to identify anomalies and change-points, as well as give us the ability to clean up noisy models.

The standard generative model is the stochastic block model where a community equals some vertices with the same pattern of intercommunity connections. The advantages of the SBM are remarkable: when we test the model on real-world models where we already know the results, it tends to return accurate correct results[[1]](#footnote-1). Furthermore, it naturally models many large-scale patterns and is highly effective in practice (Newman) and has nice mathematical functions particularly that it allows you to augment with auxiliary data which allows for easy usage in science since it naturally allows you to quantify uncertainty.



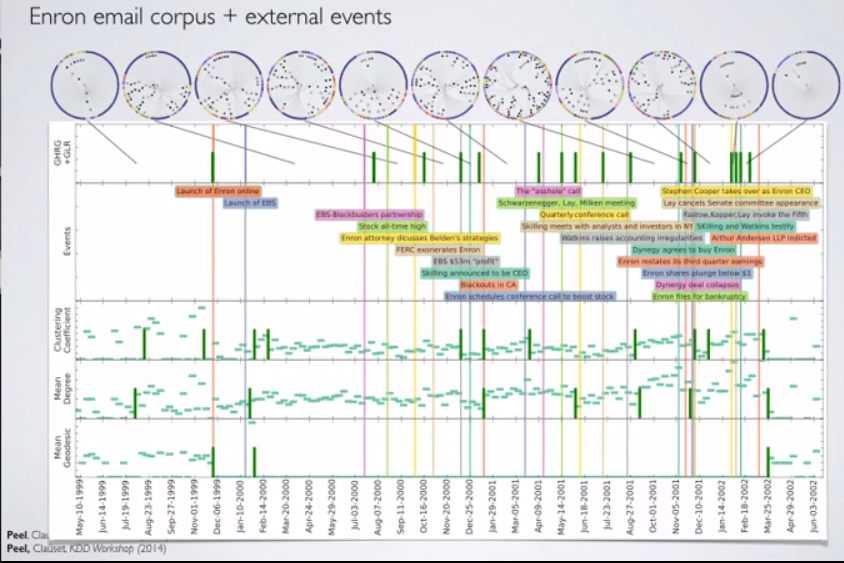
*(Taken from the referenced Karrer & Newman’s ‘Stochastic block models and community structure in networks’)*

Clauset also briefly mentions the likelihood function, an alternative to the SBM; however he doesn’t further explore it and leaves us with the general probability distribution form:



Whilst SBM models are quite effective, we can apply weights (such as interaction frequency, duration, capacity, strength, character and outcome) to better emphasize the ranking and value of certain relations and filter out noise to produce a Weighted Stochastic Block Model. To produce the WSBM, we threshold the weights and apply unweighted methods to the dataset. A few problems arise when doing this, most importantly that by thresholding we are discarding some information and obscuring the structure. To counter this, Clauset utilizes model edge existence and edge weights along with a few machine learning functions to ‘fudge’ the models into the accurate direction (the accurate direction is guessed by consulting with experts on the topic).

There are many examples where WSBM have pragmatic results such as using time series techniques to distinguish norms from noise to identify change points.



*(The Enron scandal visualized with an application of a WSBM to the email corpus released to identify the change-points with the progression of time)*

Challenges present themselves; what data can we apply this to since we have a boundless amount of data? How much auxiliary info can we add? How do we interpret these models? What are the dynamics of the networks? How can we directly model network dynamics?

I was particularly interested in the computational social science projects Clauset described (such as the global statistical patterns in terrorism and war (guessing rare events); competition in sports and online games; human social dynamics in online environments) in relation to my online stock day trading. I postulated that WSBM could be applied to the network of Market Makers (major brokers that handle all major stock transactions) and their interactions with each other. Hypothetically, the network would have edges that can act as the direction of sale with weights such as the stock price of the transaction against the market price and volume. This would open up an unparalleled level of stock analysis and detection of market-rigging practices practiced by top firms.

Whilst professional traders are all very aware of HFT (high frequency trading) and automated fraud to manipulate stock prices, authorities have often struggled to monitor and detect it (also due to being vastly underfunded). By visualizing who was selling to who and at what price and the frequency of these trades, we could easily detect instances where major brokers spoof the market price by intentionally undercutting the market price by spamming low-amount stock sales to partners below market values to drive down investor confidence then capitalizing on the panic sell to profit.

Furthermore, outside of illegal activity, we could use this to better understand the market dynamics and the collective changes and inflections in group behavior on a very individual level. This would give us insight to group psychology in a very high-risk survival context that goes between fear and joy. It also has the very high reward of making the first person to accurately analyze the graphs and build a model that runs in real time a *very* wealthy man.

# Works Cited

Newman, Brian Karrer and M. E. J. "Stochastic blockmodels and community structure in networks." *Phys. Rev. E 83, 016107* 23 August 2011: 1-2.

1. Based on modeling examples such as a karate social networks given in the Keller/Newman paper on SBM [↑](#footnote-ref-1)