Introduction to Computational Social Science assignment 2:

Why CSS? Complexity explorables

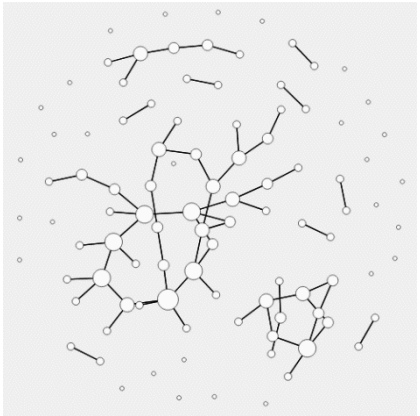
Mihaly Hanics, MS in Social Data Science

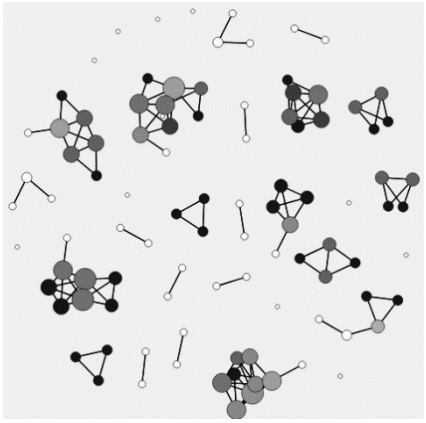
This report is about exploring Jujujajáki networks, a community structure emergence model, my interpretation of the results, describing the emergence of complex behaviour, plus critique.

The work tries to model how from a (loosely) connected society eventually communities form and tries to understand the structures. For this, they designed an algorithm to rewire any (sparse) graph with the following loop: Randomly choose a node, then with predefined P_0, P_E, P_L probabilities, respectively, the node can isolate itself from all contacts, it can connect a link to an unlinked node (edge weight equals default w_0), and can create a link to finish a triangle, increasing the triangle's other two edges' weights by δ, the reinforcement increment parameter. On the website is a playground, we can tune these parameters and run the algorithm on a network. Visualization includes increasing size of a node based on the more connections they are, and the darker the node is, the higher the local clustering coefficient (the ratio of the number of links between your friends over the maximum amount).

Playing with the explorable, I observed some things. Sadly, I don't have the exact parameter limits, it wasn't described, but the minimum exploration probability is bigger than 0, because even if it is minimal, we get new "cluster-connecting" edges. If you tune up isolation rate to the maximum, then no matter what the initial network is, after some time you'll end up in a non-dynamic totally disconnected network. If it is not at the maximum, at a high rate we still get a totally empty network or near-empty network, but time-and-time again some connections form. (This means that the maximum isolation rate is 100%.) I found that playing with all the parameters, the biggest impact a difference in parameter makes is when the parameter is the isolation rate. This may be because as I see it ranges from 0 to 1, and other parameters range in a smaller interval, but it could simply indicate that it is more significant than other parameters. It could make sense, isolation could remove many edges at a time. When isolation is 0, eventually the network becomes totally connected (1 component).

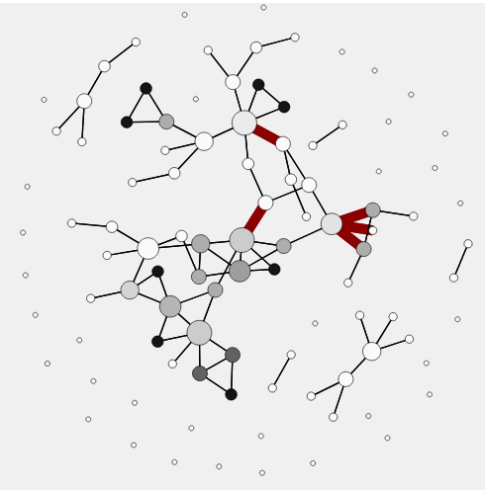
When local search probability is minimal, the links that get created contain much less triangles (3 nodes that form a K_3 complete graph), and this makes sense, as we exclude the help of forming triangles via connecting someone with a friend of a friend. These graphs (a snapshot) to me resemble something similar to an Erdős-Rényi graph, and there indeed shall be some correlation, the mechanism to generate links is determined by a probability value (not the same mechanisms, but they are similar). (The authors call these types of connections focal closure.) See the graph on the left.





When local search probability is high (and exploration is low), connected components start turning into more dense components, pseudo-complete graphs. This is what they call cyclic closure. The only way to connect two different components is via exploration links, which happen rarely and it takes time to "familiarize" the two connected components with each other. We could also call these "weak links", as they did in the paper, regarding back to Granovetter. After enough iterations, the components all have a high local clustering coefficient as long as exploration is kept low, and the number of disconnected clusters (components) depends on the rate of isolation. The isolation of a node in a totally connected component would still remain totally connected, and its local clustering coefficient stay 1. Low isolation rate leads to 1 big component in most cases, high isolation rate leads to something like on the graph on the right. If you're creative enough, you could say this network has "large distances" even though most distances in this network are 1-2 or infinity. If you would connect the *m* components with *m-1* edges, the average distance would be huge no matter how you connect the components. Look how more "reaching" the network on the left is, with less edges. The reinforcement parameter helps further increase the clustering effect: because at each local connection creation, the most likely "picked middleman" is a friend that has the strongest connection (highest edge weight), which was previously reinforced likely multiple times, meaning the two friends have taken part in creating many triangles.

When both local search probability and exploration is high enough, and isolation is kept on a normal level we start to experience small-world behaviors. This is due to having the high-level regularity, clusteredness of the network on the right, and the "crossing" weak ties, that speed up information (disease) spread, we get the best of the two models. This is what Watts and Strogatz argued in their original paper as well. The isolation is a very sensitive parameter in this case, I found a value around which the network stayed fairly "stable", below that value the nodes formed only one big cluster, above it the network became way too sparse and components appear and disappear too suddenly. With "optimal" isolation, this is something we can get:



While none of the cases were mentioned to particularly create both cluster-like communities and weak ties, with good tuning of the parameters, similarly to the last case, we can balance well between the strong clusteredness and the weak links connecting clusters, the isolations thinning the paths from one node to the other, making the structure more realistic.

Critique: Landing page states: “A mechanism that may explain community structures in networks”, well... Even if it is designed to specifically target this problem “compactly” (the inputs are solely about parameterizing link connections/disconnections, nothing external), I find it hard to believe that running this algorithm countless times gives any new information about how communities are formed. It gives you information on how communities *can* form, under very strict circumstances. How they are formed when based on the factors of your predefined (most importantly, constant, non-changing) chances to meet someone totally new, the constant chances to meet someone through your friends. Why are probabilities constant? The link reinforcement is linear (increment is constant, again). Why? Why is the model only discrete, not continuous? Of course other opportunities are possible, the answer I believe would be that this way, the model can be kept simple and parameters can be computed easily, and even this more restricted set of parameter-possibilities includes one (or more) parameter-setting where the outcome of the model with these settings is accurate enough (people don’t notice that this is basically just more efficient brute-forcing, but that’s not important now). To understand what I mean: there are phenomena in networks like small-world networks, power law networks, community forming, but algorithms that the respective researchers created are just a method of obtaining or simulating this phenomenon, they don’t explain the phenomenon. Replicating a phenomenon with a specific process of course doesn’t mean that this exact process is what happens everytime the phenomenon appears.

If I were to do such research, I would go from data on how connections were formed in a network (let’s say, online connections) and try to predict “backwards” to find a precise algorithm, not create an algorithm that closely describes how we logically think we make friends.