1 More configuration model

In the last lecture, we explored the definition of the configuration model, a simple method for drawing networks from the ensemble, and derived some of its mathematical properties. This time, we'll finish up a few more mathematical properties, and explore using it to study empirical networks. Recall that the fundamental property of the configuration model is the probability of an edge between i and j:

$$p_{ij} = \frac{k_i k_j}{2m} \ , \tag{1}$$

which holds in the large-m limit.

1.1 More mathematical properties

Expected number of common neighbors.
Given a pair of vertices i and j, with degrees k_i and k_j, how many common neighbors n_{ij} do we expected them to have?

For some ℓ to be a common neighbor of a pair i and j, both the (i, ℓ) edge and the (j, ℓ) edges must exist. As with the multi-edge calculation above, the correct calculation must account for the reduction in the number of available stubs for the (j, ℓ) edge once we condition on the (i, ℓ) edge existing. Thus, the probability that ℓ is a common neighbor is the product of the probability that ℓ is a neighbor of i, which is given by Eq. (1), and the probability that ℓ is a neighbor of j, given that the edge (i, ℓ) exists, which is also given by Eq. (1) except that we must decrement the stub count on ℓ.

$$\begin{aligned} n_{ij} &= \sum_{\ell} \left(\frac{k_i k_\ell}{2m}\right)\left(\frac{k_j(k_\ell - 1)}{2m}\right) \\ &= \left(\frac{k_i k_j}{2m}\right) \sum_{\ell} \frac{k_\ell(k_\ell - 1)}{\langle k \rangle n} \\ &= p_{ij} \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \ . \end{aligned} \tag{2}$$

Thus, the probability that i and j have a common neighbor is proportional to the probability that they themselves are connected (where the constant of proportionality again depends on the first and second moments of the degree sequence).

The excess degree distribution.
Many quantities about the configuration model, including the clustering coefficient, can be calculated using something called the *excess degree distribution*, which gives the degree distribution of a

randomly chosen neighbor of a randomly chosen vertex, excluding the edge followed to get there. This distribution also shows us something slightly counterintuitive about configuration model networks.

Let p_k be the fraction of vertices in the network with degree k, and suppose that following the edge brings us to a vertex of degree k. What is the probability that event? To have arrived at a vertex with degree k, we must have followed an edge attached to one of the $n\,p_k$ vertices of degree k in the network. Because edges are a random matching conditioned on the vertex's degrees, the end point of every edge in the network has the same probability $k/2m$ (in the limit of large m) of connecting to one of the stubs attached to our vertex.

Thus, the degree distribution of a randomly chosen neighbor is

$$\begin{aligned} p_{\text{neighbor has } k} &= \frac{k}{2m} n\,p_k \\ &= \frac{k\,p_k}{\langle k \rangle} \ . \end{aligned} \tag{3}$$

Although the excess degree distribution is closely related to Eq. (3), there are a few interesting things this formula implies that are worth describing.

From this expression, we can calculate the average degree of such a neighbor, as $\langle k_{\text{neighbor}} \rangle = \sum_k k\,p_{\text{neighbor has } k} = \langle k^2 \rangle / \langle k \rangle$, which is strictly greater than the mean degree itself $\langle k \rangle$ (do you see why?). Counterintuitively, this means that your neighbors in the network tend to have a greater degree than you do. This happens because high-degree vertices have more edges attached to them, and each edge provides a chance that the random step will choose them.

Returning to the excess degree distribution, note that because we followed an edge to get to our final destination, its degree must be at least 1, as there are no edges we could follow to arrive a vertex with degree 0. The excess degree distribution is the probability of the number of other edges attached to our destination, and thus we substitute $k+1$ for k in our expression for the probability of a degree k. This yields

$$q_k = \frac{(k+1)p_{k+1}}{\langle k \rangle} \ . \tag{4}$$

Expected clustering coefficient.
The clustering coefficient C is the average probability that two neighbors of a vertex are themselves neighbors of each other, which we can calculate now using Eq. (4). Given that we start at some vertex v (which has degree $k \geq 2$), we choose a random pair of its neighbors i and j, and ask for the probability that they themselves are connected. The degree distribution of i (or j), however, is

exactly the excess degree distribution, because we chose a random vertex v and followed a randomly chosen edge.

The probability that i and j are themselves connected is $k_i k_j/2m$, and the clustering coefficient is given by this probability multiplied by the probability that i has excess degree k_i and that j has excess degree k_j, and summed over all choices of k_i and k_j:

$$\begin{aligned}
C &= \sum_{k_i=0}^{\infty}\sum_{k_j=0}^{\infty} q_{k_i} q_{k_j} \frac{k_i k_j}{2m} \\
&= \frac{1}{2m}\left[\sum_{k=0}^{\infty} q_k k\right]^2 \\
&= \frac{1}{2m\langle k\rangle^2}\left[\sum_{k=0}^{\infty} k(k+1)p_{k+1}\right]^2 \\
&= \frac{1}{2m\langle k\rangle^2}\left[\sum_{k=0}^{\infty} k(k-1)p_k\right]^2 \\
&= \frac{1}{2m\langle k\rangle^2}\left[\sum_{k=0}^{\infty} k^2\,p_k - \sum_{k=0}^{\infty} k\,p_k\right]^2 \\
&= \frac{1}{n}\frac{\left[\langle k^2\rangle - \langle k\rangle\right]^2}{\langle k\rangle^3} \quad . \qquad (5)
\end{aligned}$$

where we have used the definition of the mth moment of a distribution to reduce the summations. Like the expression we derived for the expected number of multi-edges, the expected clustering coefficient is a vanishing fraction $O(1/n)$ in the limit of large networks, so long as the second moment of the degree distribution is finite.

Expected clustering coefficient (alternative).
It should also be possible to calculate the expected clustering coefficient under the configuration model by starting with the expected number of common neighbors n_{ij} for some pair i, j, which we derived in Eq. (2). In particular, given the result derived in Eq. (5), we can express the clustering coefficient in terms of n_{ij} and p_{ij}:

$$C = \frac{1}{2m}\left(\frac{n_{ij}}{p_{ij}}\right)^2 \quad . \qquad (6)$$

(Can you explain why this formula is correct?)

The giant component, and network diameter.
Just as with the Erdős-Rényi random graph model, the configuration model also exhibits a phase transition for the appearance of a giant component. The most compact calculation uses generating functions, and is given in Chapter 13.8 in *Networks*. The result of these calculations is a simple formula for estimating when a giant component will exist, which, like all of our other results, depends only on the first and second moments of the degree distribution:

$$\langle k^2 \rangle - 2\langle k \rangle > 0 \ . \tag{7}$$

Unlike our previous results, however, this equation works even when the second moment of the distribution is infinite. In that case, the requirement is trivially true.

A corollary of the existence of the giant component in this model is the implication that the diameter of the network grows logarithmically with n, when a giant component exists. As with $G(n,p)$, the configuration model is locally tree-like (which is consistent with the vanishingly small clustering coefficient derived above), implying that the number of vertices within a distance ℓ of some vertex v grows exponentially with ℓ, where the rate of this growth again depends on the first two moments of the degree distribution (which are themselves related to the number of first- and second-neighbors of v).

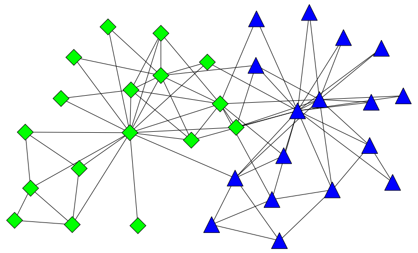
1.2 Directed random graphs

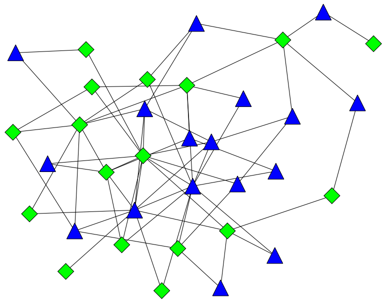
All of these results can be generalized to the case of directed graphs, and the intuition we built from the undirected case generally carries over to the directed case, as well. There are, of course, small differences, as now we must concern ourselves with both the in- and out-degree distributions, and the results will depend on second moments of these. (The first moments of the in- and out-degree distributions must be equal. Do you see why?)

Constructing directed random graphs using the configuration model is also analogous, but with one small variation. Now, instead of maintaining a single array v containing the names of the stubs, we must maintain two arrays, v_{in} and v_{out}, each of length m, which contain the in- and out-stubs respectively. The uniformly random matching we choose is then between these arrays, with the beginning of an edge chosen from v_{out} and the ending of an edge chosen from v_{in}.

2 A null model for empirical networks

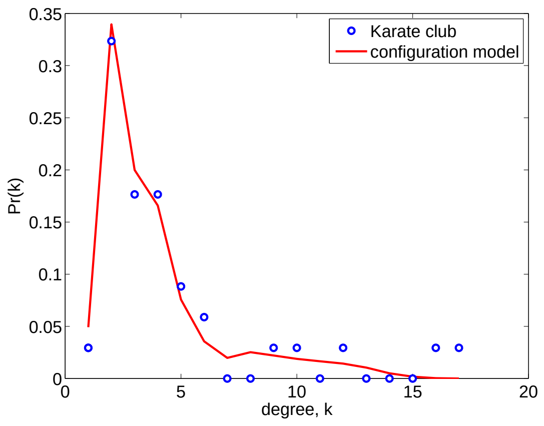
The most common use of the configuration model in analyzing real-world networks is as a null model, i.e., as an expectation against which we measure deviations. Recall from the last lecture our example of the karate club, and an instance drawn from the corresponding configuration model. Using the configuration model to generate many such instances, we can use each network as input to our structural measures. This produces a distribution of measures, which we can then compare

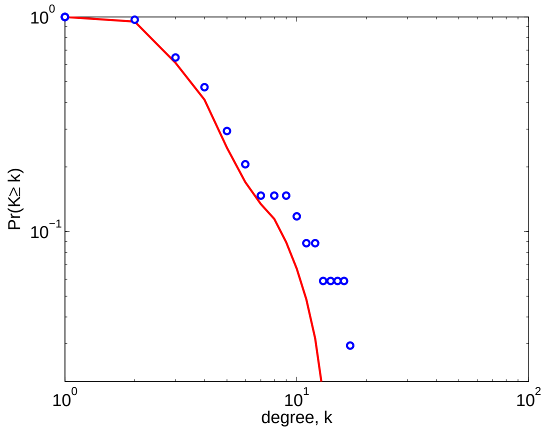




directly to the empirical values. Each of the mini-experiments below used 1000 instances of the configuration model, and where multi-edges were collapsed and self-loops discarded.

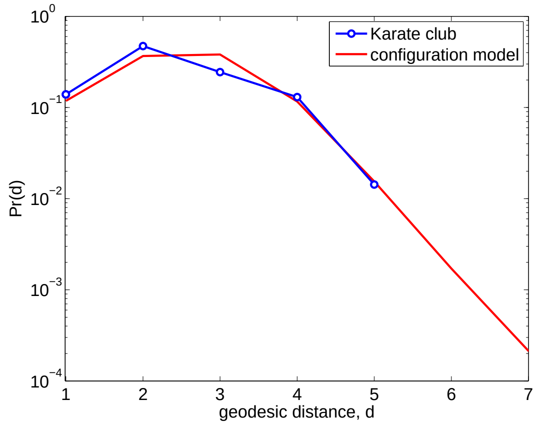
The degree distribution (shown below as both pdf and ccdf) is very similar, but with a few notable differences. In particular, there the highest-degree vertices in the model have slightly lower degree values than observed empirically. This is reflects the fact that both multi-edges and self-loops have a higher probability of occurring if k_i is large, and thus converting the generated network into a simple network tends to remove edges attached to these high-degree vertices. Otherwise, the generated degree distribution is very close to the empirical one, as we expect.[1]

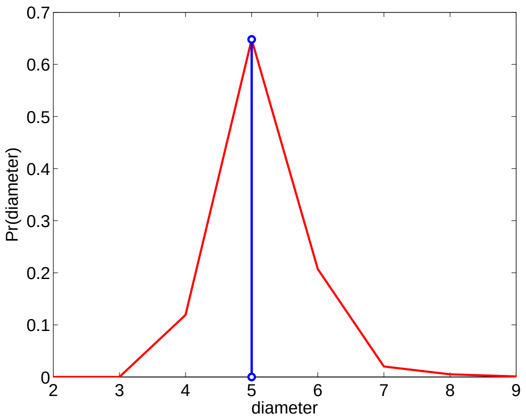




[1] It is possible to change the configuration model slightly in order to eliminate self-loops and multi-edges, by flipping a coin for each pair i, j (where $i \neq j$) with bias exactly $k_i k_j / 2m$. In this model, the expected degree we generate is distributed as $\hat{k}_i \sim \text{Poisson}(k_i)$, which assumes the specified value in expectation.

Both the distribution of pairwise geodesic distances and the network's diameter are accurately reproduced under the configuration model, indicating that neither of these measures of the network are particularly interesting as patterns themselves. That is, they are about what we would expect for a random graph with the same degree distribution. One nice feature of the configuration model's pairwise distance distribution is that it both follows and extends the empirical pattern out to geodesic distances beyond what are observed in the network itself.





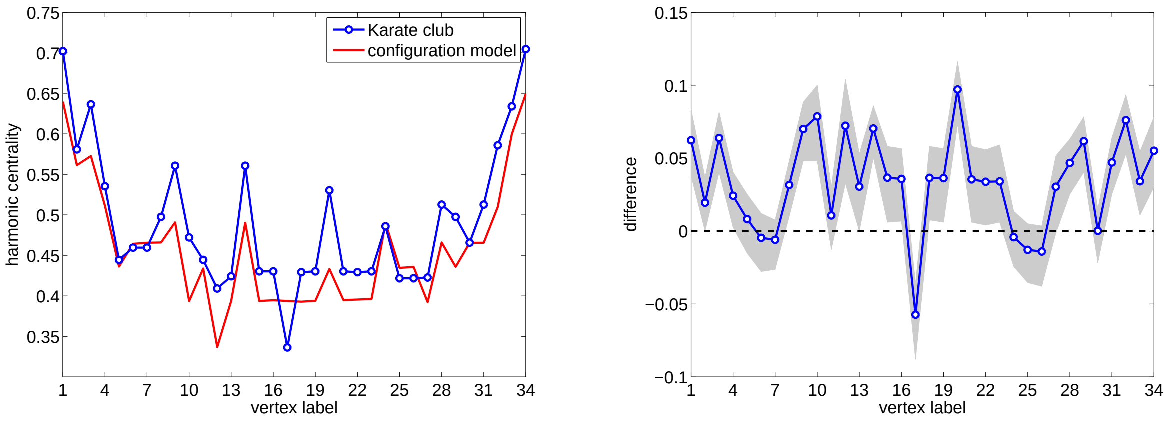
We may also examine vertex-level measures, such as measures of centrality. From the geodesic distances used in the previous figures, we may also estimate the mean harmonic centrality of each vertex. The first figure below plots both the empirical harmonic centralities (in order of vertex label, from 1 to 34) and the mean values under the configuration model. The various centrality scores are now placed in context, showing that their scores are largely driven by the associated vertex degree, as demonstrated by the similar overall pattern seen in the configuration model networks.[2]

But, not all of the values are explained by degree alone. The second figure plots the difference between the observed and expected centrality scores Δ, where the line $\Delta = 0$ indicates no difference between observed and expected values. If an observed value is above this line, then it is more central than we would expect based on degree alone, while if it is below the line, it is less central.

When making such comparisons, however, it is important to remember that the null model defines a distribution over networks, and thus the difference is also a distribution. Fortunately, however, computing the expected centrality scores by drawing many instances from the configuration model also produces the distribution of centrality scores for each vertex, which provides us with a quan-

[2] Recall also that the Pearson correlation coefficient for harmonic centrality and degree was large $r^2 = 0.83$, a fact that reinforces our conclusion here.

titative notion of how much variance is in the configuration model value. The grey shaded region shows the 25 and 75% quantiles on the distribution of centrality scores for each vertex. When the $\Delta = 0$ line is outside of this range, we may claim with some confidence that the observed value is different from the expected value.



This analysis shows that the main vertices (1 and 34, the president and instructor) are somewhat more central than we would expect just based on their degree alone. In fact, most vertices are more central than we would expect, one is less central than we expect, and about a third of the vertices fall in line with the expectation.

3 At home

1. Read Chapter 13.3–13.11 (pages 445–483) in *Networks*
2. Next time: guest lecture on dynamic social networks