Structure and Dynamics of Complex Networks (2024/25 Winter)
Homework 1

Mihaly Hanics

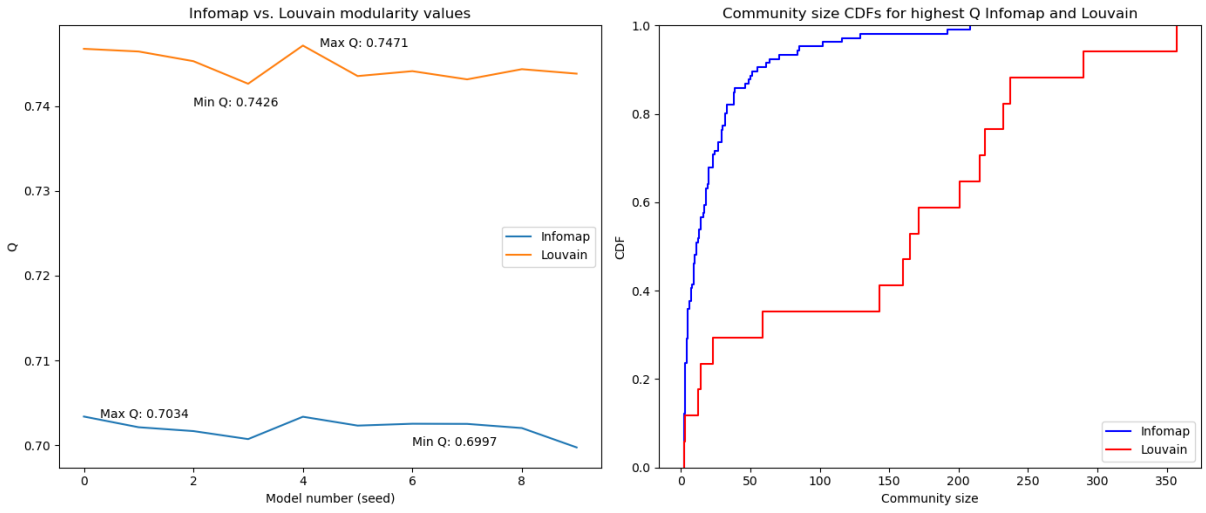
This report is about exploring the community structure of a painter network, with different models and measures. The report is written for the “Structure and Dynamics of Complex Networks” course at the Central European University.

Data

The network explored is a painter network based on locations of painters throughout times, collected from the PainterPalette dataset. Painters are connected if (based on the original dataset) they lived 5 years at the same location; periods are approximated. This is a network with extensive temporal information, with artists dating back to the middle ages all the way to the 20th century. Therefore, the network is longitudinal and connections only span to short range (painters in the same lifetime), making the network's diameter rather large (26) despite its size (2503 nodes). The network is undirected, weights are not considered. The network is available here.

Using Infomap and Louvain to detect communities

To detect communities in the network, two different algorithms were used: Infomap and Louvain. Both algorithms were ran 10 times, with modularity score being measured. The modularity values across runs of one method did not change much, but between the methods, there is a difference of 0.04 across any runs. This could signal that using modularity as measure (which is not necessarily the best measure of goodness), the Louvain method produces better partitions for this network. However, Louvain is a greedy algorithm that optimizes modularity, unlike Infomap, which optimizes description length of random walk flows in the system.



The cumulative distributions of community sizes for the highest modularity score partitions of each method were plotted. As seen on the diagram on the previous page, Infomap captures rather "continuous" community sizes, whereas Louvain captures a heterogenous set of sizes, but more rare, splitting the network into much less communities than Infomap does (only 17 in comparison).

Metrics on Infomap partitions: Rand index and Jaccard index

The Rand index for two partitions (X,Y) is calculated as:

$$\text{Rand(X,Y)} = \frac{a_{00} + a_{11}}{a_{00} + a_{11} + a_{01} + a_{10}}$$

where:

- a_{00}: number of pairs of nodes where the two nodes are in **different communities in both** X and Y

- a_{11}: number of pairs of nodes where the two nodes are in the **same community in both** X and Y

- a_{10}: number of pairs of nodes where the two nodes are in the **same community in X but in different communities in Y**

- a_{01}: number of pairs of nodes where the two nodes are in **different communities in X but in the same community in Y**

The Jaccard index for two partitions (X,Y) is a modified version, to account for many pair of nodes being in different communities in any partition (a_{00} being typically too large in comparison to other terms):

$$\text{Jaccard(X,Y)} = \frac{a_{11}}{a_{11} + a_{01} + a_{10}}$$

The **Rand index** between the best and lowest modularity InfoMap partitions: **0.993**

Jaccard index for best and lowest modularity Infomap partitions: **0.763**

The indices between the best Infomap and Louvain partitions: 0.937, 0.309.

We can see that the Rand index is above 99%, in comparison the Jaccard index is only 76%. The Rand index, is however typically in any case very high, because of most pair of nodes not being in one community in any case (which is especially true with high number of communities). Therefore, the Jaccard index is more insightful, and it being only 76% shows

that even between two partitions of the same method, with very similar modularity can be significant differences in partitions (along with a lot of similarities). The only ~31% value for the Jaccard index means that there are quite significant differences between the two partitions, which is partially coming from the difference in distributions.

Random configured partitions

We can also compare these metrics to the expected values in a random configuration, where the community sizes are the same, but nodes are randomly assigned to communities. (This differs from partition to partition, it depends on the community size distribution). This can be done e.g. by randomly swapping nodes between communities, or by shuffling the community assignments across nodes (which is what I implemented). Comparing the previous values to the on expected values of the same metrics in a random configuration can give insight into how well can these metrics separate good partition similarities from "random".

We can try to calculate the expected values by hand: Let's take two partitions (X,Y), we take Y to be have all nodes having a random community, but *size distributions are the same as for X*. (This is an important simplification.)

Let's say we have L communities (N nodes), with community sizes $l_1, l_2, \dots, l_L$. Just for simplification, let's say we have 200 nodes in the first community.

What is the expected value of a_{11}^{C1} (pair of nodes that are both in the first community for both partitions) in this random configuration? Well, there are $\binom{200}{2}$ pairs of nodes in the first community. For any pair, to be both in community C_i in the other partition, the probability is $\binom{l_i}{2}$ divided by $\binom{N}{2}$, or $\frac{l_i(l_i-1)}{N(N-1)}$.

So, the probability of being in the same community in partition Y for the two nodes is: $p_{\text{same community}} = \sum_{i=1}^{L} \frac{l_i(l_i-1)}{N(N-1)}$. Therefore, the expected value of a_{11}^{C1} is $\binom{200}{2} \cdot p_{\text{same community}}$.

From this, we can derive that the expected value of a_{11}, which is:

$E[a_{11}] = \sum_{i=1}^{L} \binom{l_i}{2} \cdot p_{\text{same community}}$.

We can also simply derive the expected value of a_{10} in a similar way: this time, we look for the two nodes not to be in the same community in partition Y, so the expected value of a_{10} is $E[a_{10}] = \sum_{i=1}^{L} \binom{l_i}{2} \cdot \left(1 - p_{\text{same community}}\right)$.

As we can just swap the roles of a_{10} and a_{01} (we can do this as the distributions are the same, for any such (X,Y) partitions, $a_{10} = a_1 - a_{11} = \sum_{i=1}^{L} \binom{l_i}{2} - a_{11} = a_1' - a_{11} = a_{01}$, where a_1 and a_1' are the amount of pairs in the same community, for partition X and Y respectively, and they both equal $\sum_{i=1}^{L} \binom{l_i}{2}$). Therefore $E[a_{01}] = E[a_{10}]$, of course.

We could also derive $E[a_{00}]$ similarly, but we can simply derive it from the other values, as $a_{00} = \binom{N}{2} - a_{11} - a_{10} - a_{01}$. Therefore $E[a_{00}] = \binom{N}{2} - E[a_{11}] - E[a_{10}] - E[a_{01}]$.

Making these calculations on the data, we $E[a_{11}] = 2671$ and $E[a_{10}] = E[a_{01}] = 88778$. If we run a random shuffling on the nodes of the highest Q Infomap partition and look into the a_{ij} values, we can see indeed a_{11} is close to our expected value of 2671, and a_{10} and a_{01} are equal, close to the estimate of 88778 - so our initial calculations seem to be correct. With these values, the Rand index is 0.943, and the Jaccard index is 0.014, which shows that the Rand index even with random configurations is very high, and the Jaccard index is a better measure.

To show by how much there is an increase in the indices, let's make a random shuffling of the lowest Q-value InfoMap partition, then calculating the Rand and Jaccard indices with the highest Q-value InfoMap partition. We can then compare these values to the original two partition indices, seen before. In this case, as the distributions are different, a_{10} and a_{01} are not equal, and calculations are more complicated.

The Rand index Rand(X, R(Y)) = 0.946, the Jaccard index Jaccard index Jaccard(X, R(Y))=0.013. We see similar values to the previous random configuration. Let's calculate the ratios of the original indices to the random configuration indices: $\frac{\text{Rand}(X,Y)}{\text{Rand}(X,R(Y))} = 1.05$ and $\frac{\text{Jaccard}(X,Y)}{\text{Jaccard}(X,R(Y))} = 57.02$. (X is the highest Q Infomap partition, Y is the lowest Q infomap partition, and R(Y) is a shuffling of the nodes of the Y partition but keeping community sizes.)

We can see that the Rand index increase is small, but the Jaccard index increase 57-fold. This again shows that the Jaccard index is a better metric for comparing partitions, because against a random configuration, the value is much larger than for the Rand index.

Conclusion

The two methods partition many pairs of nodes into the same community, however significant differences come from Infomap finding more communities in a less broad range. To evaluate the partition differences, the Jaccard index proved to be more valuable, than the Rand index.