Martijn Gösgens
Centrum Wiskunde & Informatica
Amsterdam
research@martijngosgens.nl

Onderzoek Research

Community detection: between heuristics and statistics

In this article Martijn Gösgens explains the community detection method `modularity' for networks. He describes an interpretation of modularity in terms of Bayesian statistics and explains a phase transition in the corresponding prior distribution. Gösgens finished his PhD supervised by Remco van der Hofstad and Nelly Litvak. This article was written on the occasion of the VVS-OR Van Zwet PhD thesis award Gösgens received on 20 March 2025.

Introduction

Many networks contain groups of nodes that are better connected to each other than to the rest of the network. Examples include groups of friends on social media or Wikipedia pages on the same topic. *Community detection* is the task of partitioning the network nodes into such communities.

There exist many different algorithms for this task. The most widely used method for community detection is to maximize a quantity called *modularity* over the set of partitions. This method is popular because it is relatively simple and appears to find reasonable communities in practice. While there exist several interpretations of this method, from particle physics to random walks, there is no theoretical understanding of when this method actually works, and why it appears to work so well in practice.

More recently, a different community detection method based on *Bayesian statistics* has become popular. This method has a much stronger statistical motivation than the modularity heuristic. However, it is also much more complicated, making it nearly impossible to analyze this method theoretically.

In my thesis, we showed that the modularity heuristic can be interpreted in terms of Bayesian statistics. The corresponding prior distribution has a simple, but peculiar form. Studying this prior distribution helps us explain when the modularity method works, and when it doesn't. This interpretation forms a bridge between the modularity method and the Bayesian methods.

Community detection

Community detection has attracted attention from many disciplines. In *social sciences*, community detection helps identify divisions in social networks. Take the *Karate network* that is shown in Figure 1, for example. This network consists of members of a karate school that was studied by an anthropologist in the 70s [10]. During this study, a dispute broke out between the administrator and instructor of this club, which led to the club splitting into two groups. This small network is often used as a benchmark to demonstrate new community detection algorithms, where the goal is to predict this split as accurately as possible.

In *computer science and machine learning*, community detection algorithms are used to cluster networked data. For example, we can consider Wikipedia as a network of articles, where connections correspond to references between articles. If we apply a community detection algorithm on this network (like shown in

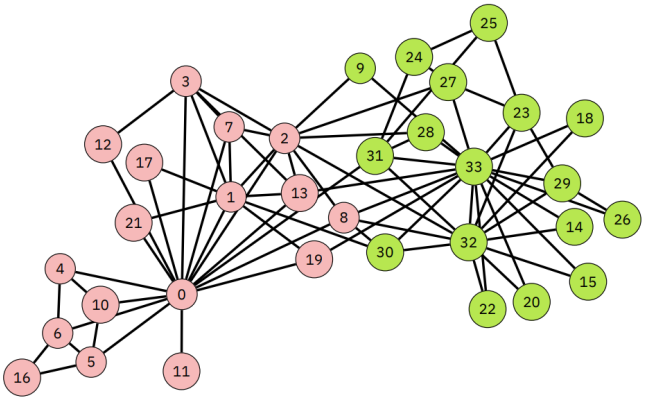

Figure 1 The social network of *Zachary's Karate Club* [10] that is often used to demonstrate community detection algorithms. After a dispute between the club's administrator and instructor (number 0 and 33), the club split up into the green and red groups.

Figure 2), we get a partition of the network articles that groups together articles on similar topics. This allows computer scientists to automate the categorization of such articles. The cool thing is that these algorithms can categorize articles by only looking at their connections among each other, without even looking at the text or titles of the articles. This allows one to categorize Chinese Wikipedia pages without even understanding a single Chinese character.

There is also a surprising number of *physicists* active in the field of community detection. Both the modularity method and the Bayesian method that we discuss here were actually introduced by physicists. Physicists like to view networks as interactive particle systems, where nodes correspond to particles and connections correspond to interactions.

For *statisticians*, community detection is an interesting inference problem, because the parameter to be estimated is not a simple scalar, but a partition, leading to all sorts of combinatorial challenges.

Modularity maximization

Modularity is the most widely used community detection method in practice. Consider a graph G consisting of n nodes, which we number from 1 to n. We write $i \sim j$ if there is a connection between i and j. We denote the number of connections adjacent to i (its *degree*) by d_i and denote the total number of edges by m. Let C denote a partition of $\{1, \dots, n\}$ into communities and let

$$\text{Pairs}(C) = \{1 \le i < j \le n : i \text{ and } j \text{ are in the same community in } C\},$$

denote the set of *intra-community* pairs. Modularity is given by

$$\text{Modularity}_\gamma(C, G) = \sum_{(i,j)\in \text{Pairs}(C)} \left(\mathbb{1}\{i \sim j\} - \gamma \frac{d_i d_j}{2m} \right). \quad (1)$$

That is, we count the number of connections that are *inside* communities, minus a *penalty* term for every intra-community pair of C. To detect communities with modularity, we maximize it over the set of partitions C.

The parameter γ controls how heavily we penalize for missing connections. For large values of γ, the partition C that maximizes (1) will consist of a large number of small, densely connected communities. There is no clear guideline of how this parameter should be chosen. The communities in the Wikipedia network were detected using $\gamma = 1$, which is the default choice. When maximizing modularity with $\gamma = 1$ in the Karate network from Figure [1], the resulting partition only misclassifies node 8.

There are many different ways to interpret the maximization of modularity. Modularity was initially motivated [6] as the number of connections inside communities minus the expectation if we were to randomly reshuffle the connections. In addition, modularity is related to how often a random walker jumps between communities [1]. In my thesis, we also proved another interpretation of modularity in terms of hyperspherical geometry [3].

There is also a simpler variant of modularity:

$$M_\gamma(C, G) = \sum_{(i,j)\in \text{Pairs}(C)} \left(\mathbb{1}\{i \sim j\} - \gamma \frac{m}{\binom{n}{2}} \right), \quad (2)$$

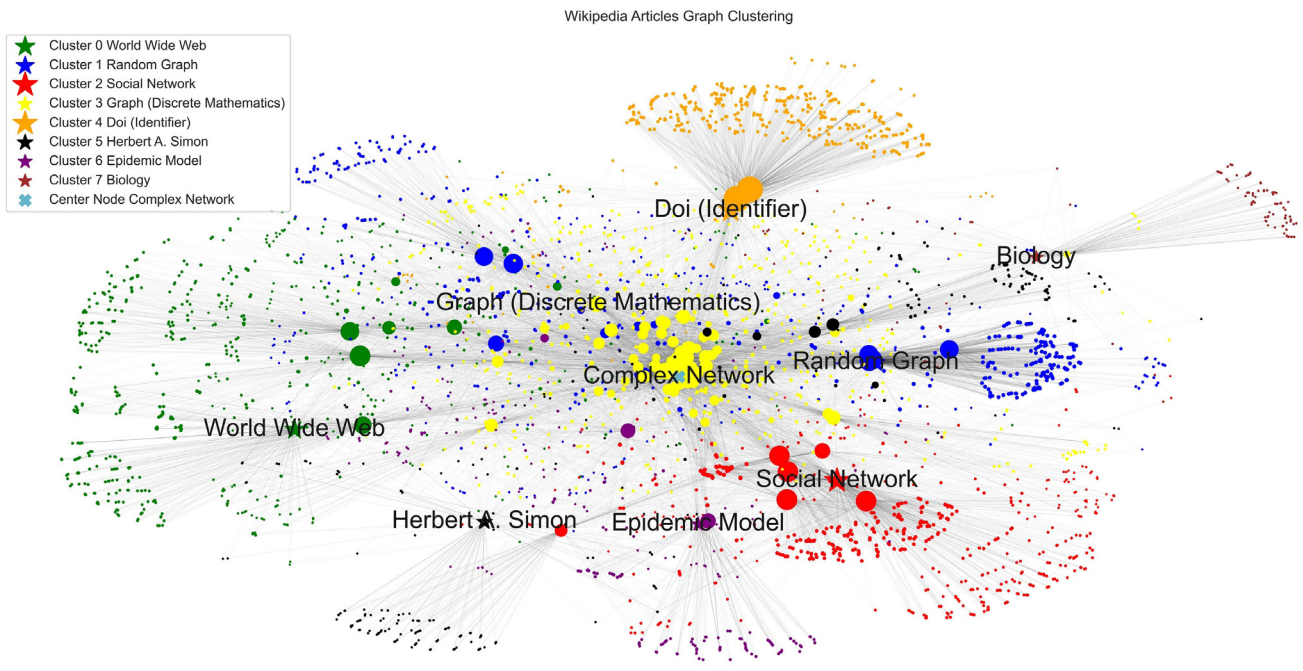

Illustration: Danie Danilo Cuevas Mendez

Figure 2 A network of Wikipedia articles related to networks [4]. The communities are obtained by maximizing modularity and roughly correspond to subject areas.

which arises when we reshuffle the connections differently. This form is related to the Potts model in physics [9] that is used to study ferromagnetism.

These different interpretations make it less surprising that modularity maximization works well in practice. Nevertheless, these motivations feel a little ad-hoc: we still have no statistical justification of using modularity to estimate communities.

The closest thing to a statistical justification of modularity came from a relation with *likelihood maximization* [5]. Likelihood maximization is a standard method in statistical inference, where we estimate a model parameter by maximizing the probability of the observed data. In the case of community detection, the observed data is the graph G and the parameter that we want to estimate is the community partition C.

The simplest model for graphs with community structure is the *Planted Partition Model* (PPM). In this model, we are given a partition of the nodes into communities and two parameters $p_{\text{in}}, p_{\text{out}} \in [0,1]$. For any two nodes, the probability that they are connected is p_{in} if they are part of the same community, and p_{out} if they are part of different communities. Typically $p_{\text{in}} > p_{\text{out}}$, so that communities are more densely connected than the remainder of the graph. An example of a PPM graph can be seen in Figure 3.

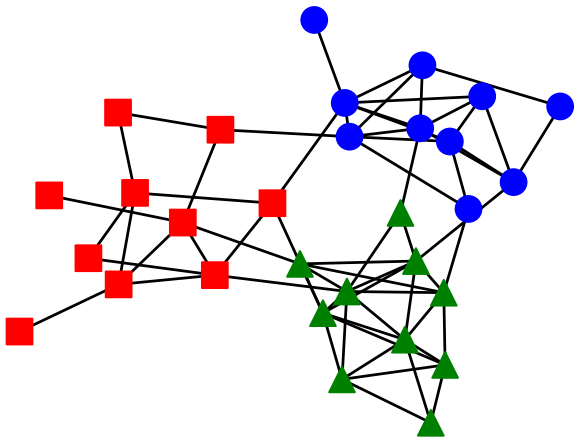

Figure 3 A graph sampled from the Planted Partition Model (PPM) with $p_{\text{in}} = 1/2$ and $p_{\text{out}} = 1/40$.

It turns out [5] that for a particular value $\gamma = \gamma_{\text{MLE}}(p_{\text{in}}, p_{\text{out}})$ maximizing (2) is equivalent to maximizing the likelihood of a PPM with connection probabilities $p_{\text{in}}, p_{\text{out}}$. So for this particular value γ, modularity maximization is a statistically justified method. Of course, the parameters $p_{\text{in}}, p_{\text{out}}$ also need to be estimated, but we ignore this for simplicity.

Given this equivalence between (2) and PPM likelihood, we would expect that if our graph is sampled from a PPM with connection probabilities $p_{\text{in}}, p_{\text{out}}$, then the parameter $\gamma = \gamma_{\text{MLE}}(p_{\text{in}}, p_{\text{out}})$ leads to the best estimator. Frustratingly so, it appears that the method works better if we choose this parameter to be *slightly larger* or *slightly lower* than γ_{MLE}. This is of course very unsatisfying: we just found a statistically justified way to pick our parameter, but now it turns out that it still requires some ad-hoc tweaking.

Bayesian statistics

Since the only statistical justification of modularity leads to disappointing results, network scientists have developed new community detection methods based on statistics. In particular, *Bayesian* methods are popular because they offer a lot of flexibility in modeling.

In Bayesian statistics, we assume a particular *prior* distribution over the parameter that we want to estimate. Then, after observing the data, we use *Bayes' rule* to derive the *posterior distribution* of the parameter: the distribution of the parameter conditioned on the observed data. That is,

$$P(C \mid G) = \frac{P(G \mid C) \cdot P(C)}{P(G)} \propto P(G \mid C) \cdot P(C). \tag{3}$$

Here, $P(C \mid G)$ is the *posterior probability* that we want to maximize over C for the observed graph G, $P(C \mid G)$ is the *likelihood*, $P(C)$ is the *prior* probability of the community partition C and the normalizing quantity $P(G)$ is irrelevant for the optimization since it does not depend on C.

While (3) looks simple, the expressions for the likelihood function $P(C \mid G)$ and the prior distribution $P(C)$ that are commonly used are quite complicated [7]. The expressions involve many factorials and binomial coefficients. I will not give these expressions here, since introducing the notation would require several pages, and the expressions wouldn't fit the margins anyway.

Despite the complicated setup, this Bayesian method can be implemented surprisingly efficiently, and it works well in practice [8]. For us mathematicians, however, the situation is somewhat disappointing: the expressions are too messy to prove elegant theorems about it. Ideally, we would want to prove performance guarantees for this method whenever G is sampled from a known distribution with communities. Unfortunately, this is completely hopeless.

Bayesian interpretation of modularity

From (3), we see that the only relevant difference between Bayesian methods and likelihood maximization is the multiplication with the prior $P(C)$. If $P(C)$ is constant (i.e., not depending on C), then this Bayesian method is entirely equivalent to likelihood maximization. This already helps explain the disappointing behavior of modularity with parameter γ_{MLE}: if $P(C)$ is constant, then this means that the prior distribution is *uniform* over all partitions. Hence, modularity with parameter γ_{MLE} is equivalent to Bayesian inference with uniform prior. While in many cases, a uniform prior is reasonable, the uniform distribution over partitions is somewhat peculiar: a uniform community partition has community sizes that are highly concentrated around $\log n$. If we want to detect communities of different sizes than $\log n$ (which is a rather atypical community size), we need to pick γ slightly above or below γ_{MLE}.

But by how much should we change the parameter γ? And what is the statistical interpretation of the resulting method? Again, we turn to Bayesian statistics. It turns out that we can view (2) in terms of Bayesian statistics for any value γ. This in itself is not surprising: indeed, for any function $f(G, C)$, we can come up with a likelihood function $P(C \mid G)$ and a prior distribution $P(C)$ so that

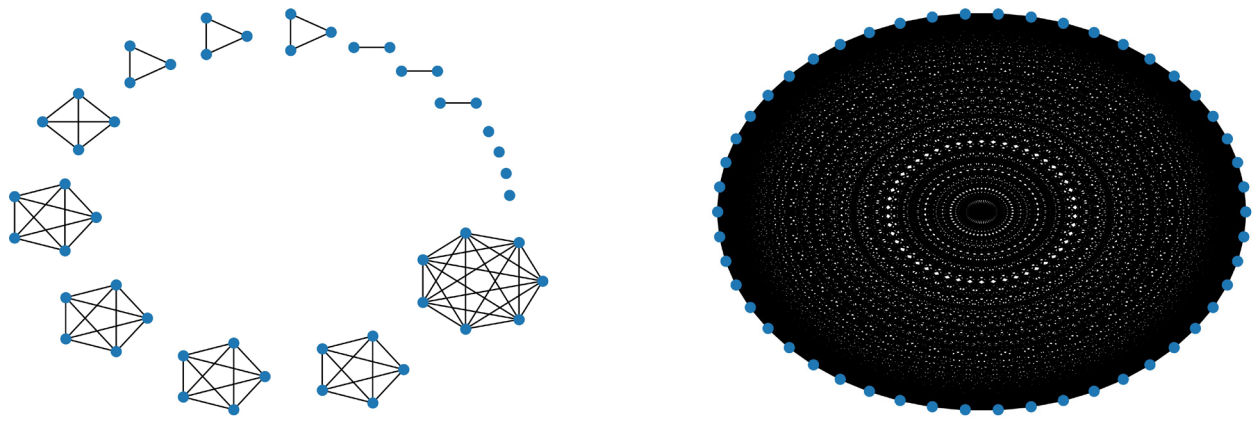

Figure 4 Left: samples of $P^*(C)$ on $n = 50$ nodes, sampled with $p = 0.51$ and $p = 0.53$ respectively. Right: the complete graph, where every two nodes are connected.

maximizing $f(G, C)$ over C is equivalent to maximizing a posterior of the form (3). However, in this case, the corresponding likelihood and prior functions are simple and interpretable. The likelihood function is the PPM likelihood that we discussed earlier, while the prior distribution is of the form

$$P^*(C) = \frac{w^{|\mathrm{Pairs}(C)|}}{Z_n(w)}, \tag{4}$$

where $w > 0$ is a parameter and

$$Z_n(w) = \sum_C w^{|\mathrm{Pairs}(C)|} \tag{5}$$

is the normalization constant.

It turns out that if we constrain the expected number of intra-community pairs $|\mathrm{Pairs}(C)|$, then (4) is a *maximal entropy* distribution [2]. This means that (4) is the *most random* distribution among all partition distributions with a given expected number of intra-community pairs. In a way, this means that assuming the prior $P^*(C)$ is a relatively modest assumption.

The prior (4) also has a nice probabilistic interpretation: let us view a community partition as a graph where two nodes are connected if they are in the same community. Next, we consider a random graph distribution where every node pair is connected with probability $p(w) = \frac{w}{1+w}$ independently and condition on the event that this graph corresponds to a community partition. It turns out that this conditional distribution is identical to our prior distribution (4).

A graph corresponds to a partition whenever each of its connected components are cliques. This is an extremely unlikely event to condition on. For reference, if $p = \frac{1}{4}$ and $n = 50$, then the probability that this event occurs is $1.57 \cdot 10^{-125}$.

To study the asymptotics of this prior distribution (4), we need a surprising amount of *complex analysis*. The asymptotics of (4) is largely determined by the normalizing quantity (5). The sum over all partitions and the difficult exponents in (5) make it challenging to determine the asymptotics of $Z_n(w)$. But, using some tools from the field of *analytic combinatorics*, it turns out that the *generating function* of $Z_n(w)/n!$ has a surprisingly simple form:

$$\sum_{n=0}^{\infty} \frac{Z_n(w)}{n!} \cdot z^n = \exp\left(\sum_{s=1}^{\infty} w^{\binom{s}{2}} \frac{z^s}{s!}\right).$$

This allows us to use *Cauchy's integration formula* to represent the normalizing quantity by a contour integral over the complex plane:

$$Z_n(w) = \frac{n!}{2\pi i} \oint \exp\left(\sum_{s=1}^{\infty} w^{\binom{s}{2}} \frac{z^s}{s!}\right) \frac{dz}{z^{n+1}}.$$

This integral can then be approximated using the *saddle point method*, which gives us the asymptotics of the normalizing quantity.

The asymptotics reveal that this prior distribution has some peculiar behavior: for $w < 1$, the community sizes are of the order $\sqrt{\log n}$ and highly concentrated. For $w = 1$, (4) is the uniform distribution, leading to communities of size $\log n$. For any $w > 1$ our partition will consist of a single community of size n with probability tending to 1. This tells us that the *typical community size* grows slowly to $\log n$ for $w \leq 1$, before suddenly jumping to n for $\mathrm{w} > 1$. This dramatic *phase transition* is demonstrated in Figure 4. The asymptotics also show that the community sizes are highly *concentrated*, i.e., they are all close to their typical size.

Putting everything together, for every $\gamma > 0$ there are $p_{\mathrm{in}}, p_{\mathrm{out}}, w$ so that the maximizing modularity from (2) is equivalent to a Bayesian method with PPM likelihood and a prior given by (4). This suggests that the method works well whenever the communities that we try to detect resemble a typical partition sampled from (4). This explains why modularity maximization works relatively well when the communities are small and equally-sized. The phase transition in the model leads to unstable performance when the communities are larger than $\log n$.

References

1 J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, **107**(29):12755–12760, 2010.

2 Martijn Gösgens, Lukas Lüchtrath, Elena Magnanini, Marc Noy, and ´Elie de Panafieu. The Erdős-Rényi random graph conditioned on every component being a clique. arXiv preprint *arXiv*:2405.13454, 2024.

3 Martijn Gösgens, Remco van der Hofstad, and Nelly Litvak. The hyperspherical geometry of community detection: Modularity as a distance. *Journal of Machine Learning Research*, **24**(112):1–36, 2023.

4 Danie Mendez. Wikipedia's articles graph analysis. https://danieldcm. medium. com/wikipedias-articles-graph-analysis-320d8630a46b, 2021.

5 Mark EJ Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315, 2016.

6 Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E* **69**(2):026113, 2004.

7 Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in Network Clustering and Blockmodeling*, 289–332, 2019.

8 Tiago P Peixoto. *Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths*. Cambridge University Press, 2023.

9 Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, **74**(1):016110, 2006.

10 Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**(4):452–473, 1977.