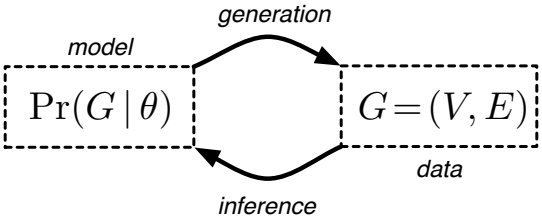
# 1 More stochastic block model

Recall that the stochastic block model (SBM) is a generative model for network structure and thus defines a probability distribution over networks $\Pr(G \,|\, \theta)$, where $\theta$ represents the SBM's parameters. If we choose the values of $\theta$, we can draw instances from the corresponding distribution. On the other hand, if we are given a network $G$, we can use techniques from statistical inference to estimate the values of $\theta$ that best explain or reproduce the observed pattern of connectivity.




For now, we will assume that the number of groups $k$ is fixed.

## 1.1 Model definition

Recall that in its most basic version, the SBM is defined by a scalar value and two simple data structures:

- $k$ : a scalar value denoting the number of groups or modules in the network,
- $\vec{z}$ : a $n \times 1$ vector where $z_\ell$ [or $z(\ell)$] gives the group index of vertex $\ell$,
- $M$ : a $k \times k$ stochastic block matrix, where $M_{ij}$ gives the probability that a vertex of type $i$ is connected to a vertex of type $j$.

## 1.2 Fitting the model to data

Given a choice of $k$ and an observed network $G$, we can use the SBM to infer the latent community assignments $z$ and stochastic block matrix $M$. There are several ways of doing this, the simplest of which is to use maximum likelihood. That is, we aim to choose $z$ and $M$ such that we maximum the likelihood of generating exactly the edges observed in $G$. The likelihood of the data, given the model is

$$\mathcal{L}(G \,|\, M, z) = \prod_{u,v} \Pr(u, v \,|\, M, z)$$

$$\mathcal{L}(G \,|\, M, z) = \prod_{(u,v) \in E} \Pr(u, v \,|\, M, z) \prod_{(u,v) \notin E} 1 - \Pr(u, v \,|\, M, z) \ ,$$

where we have separated the terms corresponding to edges we observe $(u, v) \in E$ and those we do not $(u, v) \notin E$. Thus, every pair $u, v$ appears in the likelihood function, and the function contains $O(n^2)$ terms.[1]

**Maximum likelihood choice**.
In general, $z$ and $M$ can assume any values and the likelihood remains well defined. However, because we aim to maximize the probability of the SBM generating $G$, only one particular choice of $M$ corresponds to this choice, which is the maximum likelihood choice of $M$ conditioned on the partition $z$. This simplifies the inference considerably.

Observe that each pair of group labels $i, j$ identifies a "bundle" of edges, i.e., edges that run from group $i$ to group $j$.[2] Under the SBM, each of these edges is iid with parameter $M_{ij}$, implying that the number of edges we actually observe in this bundle is binomial distributed. For simplicity, let $N_{ij}$ be the number of possible edges between groups $i$ and $j$, and let $N_i$ be the number of vertices with label $i$.

In the case where $i \neq j$, the number of possible edges is always $N_{ij} = N_i N_j$.

The case of $i = j$ depends on they type of network we are modeling: if it is a directed network with self-loops, then $N_{ii} = N_i^2$, while if we prohibit self-loops then it is $N_{ii} = N_i(N_i - 1)$. If the network is simple (no self-loops and undirected), then $N_{ii} = \binom{N_i}{2}$. In the equations below, we will use the more general term $N_{ij}$, but this should be replaced with the appropriate expression when the model is applied to real data.

Suppose that for some particular choice of $i$ and $j$, we observe $E_{ij}$ edges in the $i, j$ bundle. Because this number is binomially distributed, the maximum likelihood choice for the probability $M_{ij}$ that any particular edge in this bundle exists is simply the MLE for a binomial with expected value $E_{ij}$, that is, $\hat{M}_{ij} = E_{ij}/N_{ij}$, which can easily be derived by counting the edges in a bundle, given the network $G$ and a particular partition $z$.

This choice allows us to simplify the likelihood function by substituting the MLE for $M_{ij}$ into the

[1] Whether it is $n^2$ or $n^2 - n$ or $\binom{n}{2}$ terms depends on whether we are modeling a directed network with self-loops, directed without self-loops, or a simple network. That is, if an edge cannot exist between a pair of vertices by definition of the network type, then that pair mustn't contribute to the total likelihood of the observed data. Furthermore, if the existence of an edge is already accounted for by some other pair of groups, e.g., $j, i$ and $i, j$ when the network is undirected, then it cannot contribute a second time to the likelihood.

[2] Note: the pair $j, i$ only denotes a distinct edge bundle from that of $i, j$ if the network is directed. If the network is undirected, the calculation only runs over the $\binom{k}{2}$ unique pairs of group labels.

previous equation:

$$\begin{aligned}\mathcal{L}(G \,|\, M, z) &= \prod_{(u,v)\in E} M_{z_u z_v} \prod_{(u,v)\notin E} (1 - M_{z_u z_v}) \\ &= \prod_{i,j} M_{ij}^{E_{ij}} (1 - M_{ij})^{N_{ij} - E_{ij}} \qquad (1) \\ &= \prod_{i,j} \left(\frac{E_{ij}}{N_{ij}}\right)^{E_{ij}} \left(1 - \frac{E_{ij}}{N_{ij}}\right)^{N_{ij} - E_{ij}} \quad . \qquad (2)\end{aligned}$$

Taking the logarithm yields the log-likelihood function

$$\ln \mathcal{L} = \sum_{i,j} E_{ij} \ln \frac{E_{ij}}{N_{ij}} + (N_{ij} - E_{ij}) \ln \left(\frac{N_{ij} - E_{ij}}{N_{ij}}\right) \quad .$$

Applying the rules of logarithms and collecting like terms yields

$$\begin{aligned}\ln \mathcal{L} &= \sum_{i,j} E_{ij} \ln E_{ij} - E_{ij} \ln N_{ij} + (N_{ij} - E_{ij}) \left(\ln (N_{ij} - E_{ij}) - \ln N_{ij}\right) \\ &= \sum_{i,j} E_{ij} \ln E_{ij} - E_{ij} \ln N_{ij} + N_{ij} \ln (N_{ij} - E_{ij}) - N_{ij} \ln N_{ij} - E_{ij} \ln (N_{ij} - E_{ij}) + E_{ij} \ln N_{ij} \\ &= \sum_{i,j} E_{ij} \ln E_{ij} + N_{ij} \ln (N_{ij} - E_{ij}) - N_{ij} \ln N_{ij} - E_{ij} \ln (N_{ij} - E_{ij}) \\ &= \sum_{i,j} E_{ij} \ln E_{ij} + (N_{ij} - E_{ij}) \ln (N_{ij} - E_{ij}) - N_{ij} \ln N_{ij} \quad , \qquad (3)\end{aligned}$$

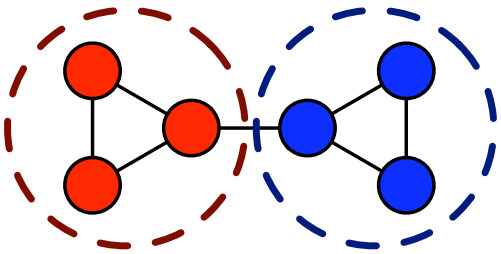
which is a function that depends only on the counts induced by $z$.[3]

### 1.3 An example

Returning to the simple network of two triangles joined by a single edge, we can tabulate the maximum likelihood stochastic block matrix $M$ for each of the usual two partitions, and thus compute their likelihoods. To do so, we use the version of the SBM that generates simple networks, i.e., we restrict the summation to exclude self-loops and we count edges between groups only once. Applying Eq. (3) to the corresponding matrices shows that the "good" partition is about 177 times more likely to generate the observed data than the "bad" partition.

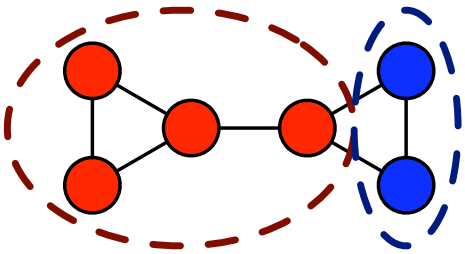
Although the qualitative results are the same as for using the modularity function—the good partition is better—this likelihood-based approach provides additional information in the form of

[3]Where we define $0^0 = 1$. This choice is necessary to prevent the numerical calculation from failing when either 0 edges run between, or 0 edges do not run between, a pair of groups.



$\mathcal{L}_{\text{good}} = 0.043304\ldots$
$\ln \mathcal{L}_{\text{good}} = -3.1395\ldots$

| $M_{\text{good}}$ | red | blue |
|---|---|---|
| red | 3/3 | 1/9 |
| blue | 1/9 | 3/3 |



$\mathcal{L}_{\text{bad}} = 0.000244\ldots$
$\ln \mathcal{L}_{\text{bad}} = -8.3178\ldots$

| $M_{\text{bad}}$ | red | blue |
|---|---|---|
| red | 4/6 | 2/8 |
| blue | 2/8 | 1/1 |

the likelihood ratio. That is, we now know just how much better, in probabilistic terms, the better partition is.

### 1.4 Choosing the number of groups $k$

Recall that we fixed $k$ the number of groups. In many applications, we would like to allow $k$ to vary and thus decide whether some choice $k' > k$ is better.

Because $k$ determines the "size" of the model, allowing $k$ to vary presents a difficulty: the larger a value of $k$ we choose, the more parameters we have in $M$, which may lead to over fitting. In the limit of $k = n$, every vertex is in a group by itself, the matrix $M$ becomes identical to the adjacency matrix $A$ and the likelihood is maximized at $\mathcal{L} = 1$. That is, the model has *memorized* the data exactly. Thus, as we increase $k$, the SBM distribution over networks becomes increasingly concentrated around the empirically observed network $G$.

Thus, the SBM has a downside relative to modularity maximization, which had no free parameter controlling its model complexity.[4] There are, however, statistically principled ways of choosing $k$, but these require additional steps as increasing $k$ directly increases the number of parameters in the model, which increases the risk of over fitting the data. A method for *regularization* or complexity control is thus necessary in order to penalize larger models for their additional flexibility. That is, we would only want to use a larger model (larger $k$) if the additional flexibility was statistically

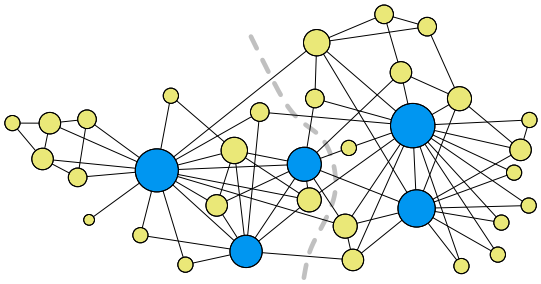
[4]The downside is not as great as you may imagine, however, as the modularity function has a built in preference for modules with certain characteristics, which the SBM lacks.

warranted. Popular choices for regularization include Bayesian marginalization, Bayes factors, various information criteria (BIC, AIC, etc.), minimum description length (MDL) approaches, and likelihood ratio tests. We will not cover any of these techniques here.

### 1.5 Correcting for degree heterogeneity

In recent years, the SBM has become a popular model upon which to build more sophisticated generative models of community structure, with many variations.

One particularly nice variation is the so-called "degree corrected" SBM, by Karrer and Newman, which was motivated by the fact that when you apply the SBM to networks with skewed degree distributions, the model tends to group vertices by degree. For instance, consider the karate club network. Setting $k = 2$ yields the following labeling of the vertices, which does not correspond to the "true" or socially observed groups.



karate club, with SBM $k = 2$

It is not hard to compute the stochastic block matrices corresponding to this division, which places the five highest-degree vertices in one group and all other vertices in the other group, and to the socially observed division. Given these, we can then compute the log-likelihood scores for the two partitions. The following tables show the results, indicating that, indeed, the SBM division is more likely (more positive log-likelihood), by a substantial margin. In fact, the SBM division is $\exp(198.50-179.39) \approx 10^8$ times more likely to generate the observed edges than the social division.

| $M_{\text{social}}$ | A (17) | B (17) |
|---|---|---|
| A (17) | 35/136 | 11/289 |
| B (17) | 11/289 | 32/136 |
| A (17) | 0.2574 | 0.0381 |
| B (17) | 0.0381 | 0.2353 |

social division, $\ln \mathcal{L} = -198.50$

| $M_{\text{SBM}}$ | A ( 5) | B (29) |
|---|---|---|
| A ( 5) | 5/10 | 54/145 |
| B (29) | 54/145 | 19/406 |
| A ( 5) | 0.5000 | 0.3724 |
| B (29) | 0.3724 | 0.0468 |

SBM division, $\ln \mathcal{L} = -179.39$

Why does the SBM do this? Recall that the SBM can only decompose a network into combinations of random bipartite graphs and Erdős-Rényi random graphs, each of which has a Poisson degree distribution with mean $M_{ij}N_i$. Thus, if a network exhibits a more skewed degree distribution, as in the case of the karate club, the model correctly recognizes that in order to reproduce this pattern, it should place those high-degree vertices in a small group together with a large probability of connecting to other larger groups. In short, the likelihood function is maximized when each $M_{ij}$ value is close to either 0 or 1, and the SBM thus prefers partitions that produce this kind of pattern. When a graph is sparse, a large but very weakly connected group is better, from the SBM's perspective, than two moderately sized but denser groups. Hence, the observed SBM division.

The downside of this tendency of the SBM is that a skewed degree distribution, like the one we see in the karate club, is a kind of violation of the SBM's particular assumption of edge independence, and the SBM seeks to explain it over other kinds of latent group structure that we might care about.

The degree-corrected SBM modifies the generative model in a way that allows vertices to have arbitrary degrees without having to force the combination of $z$ and $M$ to produce them. In addition to the usual SBM parameters, we add to each vertex a "propensity" parameter $\gamma_u$ that controls the expected degree of vertex $u$. Recall that the model for the number of edges between a pair of vertices $u$ and $v$ in the SBM is a Bernoulli distribution. In the degree-corrected SBM, we simply replace this Bernoulli distribution with a Poisson distribution with mean $\gamma_u\gamma_v M_{z_u z_v}$.

The probability of observing the network $G$ with adjacency matrix $A$ is then

$$\begin{aligned} P(G \,|\, \gamma, M, z) &= \prod_{u,v} \text{Poisson}(\gamma_u\gamma_v M_{z_u z_v}) \\ &= \prod_{u<v} \frac{(\gamma_u\gamma_v M_{z_u z_v})^{A_{uv}}}{A_{uv}!} \exp\left(-\gamma_u\gamma_v M_{z_u z_v}\right) \\ &\quad \times \prod_{u} \frac{\left(\frac{1}{2}\gamma_u^2 M_{z_u z_u}\right)^{A_{uu}/2}}{(A_{uu}/2)!} \exp\left(-\frac{1}{2}\gamma_u^2 M_{z_u z_u}\right) \,, \end{aligned} \tag{4}$$

where the composite likelihood appears because we assume an undirected network, and thus need to count edges within groups differently from edges between groups. When the Poisson mean $\gamma_u\gamma_v M_{z_u z_v}$ is typically very small, i.e., close to 0, we get a network that is sparse. This reformulation also implies that the networks produced are no longer simple, but are instead undirected multigraphs, as occasionally we will produce multiple edges between a pair $u, v$.

The propensity parameters in the Poisson mean are arbitrary to within a multiplicative constant, which we can absorb into the stochastic block matrix $M$. This observation allows us to normalize

the propensity scores

$$\sum_u \gamma_u \delta_{i,z_u} = 1 \ , \tag{5}$$

for all group labels $i$, and where $\delta(i,j) = 1$ if $i = j$ and 0 otherwise. This constraint implies that $\gamma_u$ is equal to the probability that an edge emerging from the group $z_u$ will connect to vertex $u$, and allows us to simplify Eq. (4) as

$$P(G \,|\, \gamma, M, z) = C \prod_u \gamma_u^{k_u} \prod_{i,j} M_{ij}^{E_{ij}/2} \exp\left(-\frac{1}{2} M_{ij}\right) \ , \tag{6}$$

where $k_u$ is the degree of vertex $u$, $C$ is a constant (see below), and $E_{ij}$ is the total number of edges between groups $i$ and $j$ or twice that number for $i = j$ (again, because we assume an undirected network),

$$E_{ij} = \sum_{uv} A_{uv} \delta_{i,z_u} \delta_{j,z_v} \ . \tag{7}$$

The constant $C$ depends only on the adjacency matrix $A$, and includes the various factorials that come out of the Poisson distributions in Eq. (4)

$$C = \left( \prod_{u<v} A_{uv}! \prod_u 2^{A_{uu}/2} (A_{uu}/2)! \right)^{-1} . \tag{8}$$

Taking derivatives of the log-likelihood function (derived from Eq. (6)) allows us to write down the maximum likelihood estimators[5] for the model parameters, given a partition $z$. These have a particular nice form:

$$\hat{\gamma}_u = \frac{k_u}{\kappa_{z_u}} \qquad\qquad \hat{M}_{ij} = E_{ij} \ , \tag{9}$$

where $\kappa_i$ is the sum of the degrees in group $i$, i.e., the total degree of the community. As with the SBM, we can further simplify the form of the likelihood function by substituting these MLEs into its form. The result is a fairly compact expression that again depends only on the counts induced by the choice of partition $z$:

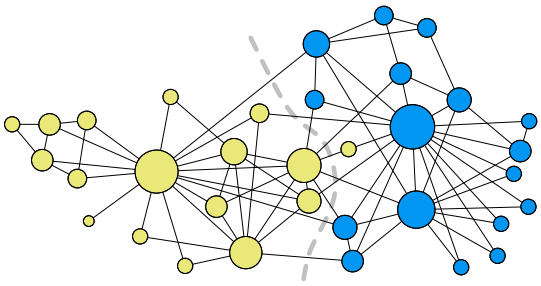
$$\ln \mathcal{L}(G \,|\, z) = \sum_{ij} \frac{E_{ij}}{2m} \ln \frac{E_{ij}/2m}{(\kappa_i/2m)(\kappa_j/2m)} \ . \tag{10}$$

Notably, this form is similar in some ways to the modularity function, which includes terms for the expected number of edges between a pair of groups, conditioned on the fraction of all edges

[5] Try this at home.

attached to those groups. Thus, the degree corrected SBM (or DC-SBM, if you like acronyms) can be thought of as seeking a partition that maximizes the "information" contained in the labels relative to a random graph with a given degree sequence, while the SBM seeks the same relative to a different null model, the Erdős-Rényi random graph.

Applying the DC-SBM to the karate club allows the propensity parameters to generate more skewed degrees within communities, and yields an inferred division that is much closer to the truth.[6]



with degree correction

## 2 At home

1. Read Chapter 8 (pages 359–418) in *Pattern Recognition*
2. Next time: hierarchical block models

[6]There is a fun story associated with the one misclassified vertex, which has an equal number of connections to each of the two groups. In the original club, this person apparently had a karate exam coming up soon, and when the club split in two, they chose to go with the instructor's faction instead of the president's in order to be better prepared for the exam.