This chapter presents a mathematical construction that serves both as a *generative* model for the analysis of algorithms for community detection in the context of statistical learning theory, and as a *hypothesized* underlying model in the context of community detection on a graph whose generative mechanism is unknown.

0.1 The stochastic block model

0.1.1 The general SBM

The canonical probabilistic model for graphs with community structure is called the stochastic blockmodel, or SBM for short. For a lengthier discussion on the origins and variants of this model, refer to [**?**].

Definition 1 (Stochastic blockmodel)**.** Let $n \in \mathbb{N}$, $k \in \mathbb{N}$, $\pi = (\pi_1, \ldots, \pi_k)$ be a probability vector on $[k] := \{1, \ldots, k\}$ and Γ be a $k \times k$ symmetric matrix with entries $\gamma_{ij} \in [0, 1]$. A pair (Z, G) is said to be *drawn under a* SBM(n, π, Γ) if

- $Z = (Z_1, \ldots, Z_n)$ is an n-tuple of $\mathbb{N}^k$-valued random variables $Z_i \sim \mathcal{M}(1, \pi)$,
- G is a simple graph with n vertices whose symmetric adjacency matrix has zero diagonal and for $j > i$, $A_{ij}|Z \sim$ Ber(γ_{Z_i, Z_j}), the lower triangular part being completed by symmetry.

Remark. The quantity n should be thought of as being the number of nodes in G, k should be thought of as being the number of communities in G, π should be thought of as being a prior on the community assignments Z, and Γ should be thought of as a matrix of intra-cluster and inter-cluster connectivities. The random variables of the model are the n community assignments Z and the $\binom{n}{2}$ entries A_{ij} of the adjacency matrix.

Remark. Although each community assignment Z_i is a vector, the same Z_i can be used to denote the *number* of the community that node i is assigned to, to make notation lighter.

It is important to emphasize that although intuition frequently refers to the assortative case, such as in Figure 1, the SBM is versatile and can reproduce many other characteristics of graphs with communities. For instance, the SBM can generate bipartite graphs as a model for the example given in the introduction, where couples dance in a party; it can also generate graphs with “stars”, and reproduce the “core-periphery” phenomenon. See Figure 2.

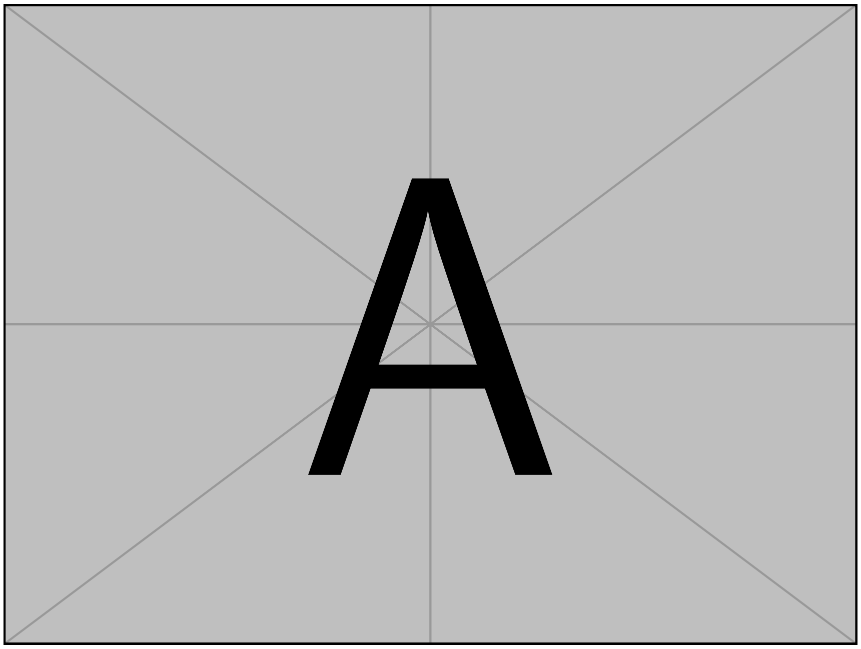
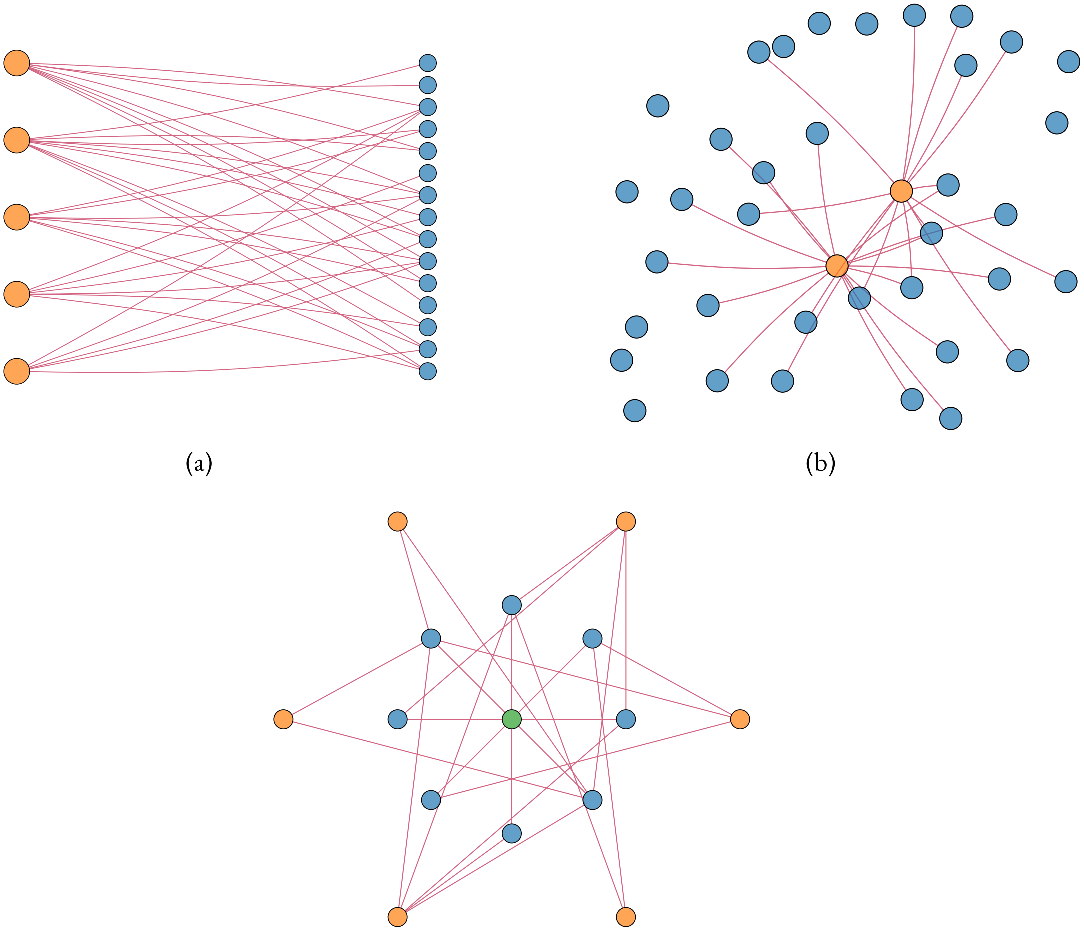


Figure 1: An example of an SBM graph

0.1.2 The symmetric SBM

Even though the SBM is a simple and intuitive model for graphs with communities, the calculations associated with it can already yield long expressions and present subtleties. For this reason, it is desirable to have a yet simpler version of the SBM where one can test intuitions and perform preliminary calculations. The symmetric SBM is precisely such a model.

Definition 2 (Symmetric SBM)**.** The pair (X, G) is drawn from SSBM (n, k, p, q) if it is drawn from an SBM model with $\pi = \frac{1}{k}\mathbf{1}_k$ and Γ taking values p on the diagonal and q outside the diagonal.



(a)

(b)

(c)

Figure 2: The SBM is versatile and can give rise to different features, such as (a) bipartite structures, (b) star structures, (c) core-periphery structures.

0.2 Model likelihood

0.2.1 Complete model likelihood

The community assignment variable Z is latent, in the sense that in practice it is not observed. However, it *is* a random variable of the model that is determined at the moment of sampling of the graph observed, and thus different possible assignments have different probabilities associated to them. What is called the *complete* model likelihood is the likelihood taking Z into account as a variable. It writes

$$p(A, Z) = \prod_{i=1}^{n} \pi_{Z_i} \prod_{\substack{i=1 \\ j>i}}^{n} \gamma_{Z_i Z_j}^{A_{ij}} (1 - \gamma_{Z_i Z_j})^{1-A_{ij}}. \tag{1}$$

Of course it cannot be calculated from an observation, since the Z are unknown.

Remark. Equation (1) is an example of the abuse of notation described in the remarks below Definition 1.

0.2.2 Observed model likelihood

In order to have a likelihood associated to an observation, one needs to marginalize the latent variables present in Equation (1). That is,

$$p(A) := \sum_{Z \in K^n} p(A, Z). \tag{2}$$

This sum over the whole latent space is intractable, since it has an exponential number of terms and it cannot be analytically evaluated to an useful simpler form. Therefore, it will be strictly necessary to approximate this observed likelihood in order to perform estimation and inference on SBMs from a probabilistic point of view.

0.3 Agreement and degrees of recovery

0.3.1 Agreement

Before discussing specific algorithms to estimate the parameters and communities, consider the task of evaluating the answer returned by any such algorithm. Assume knowledge of the true vector of community assignments $Z^\star$. Given an estimate $\hat{Z}$ for $Z^\star$ how can one measure the quality of such an estimation? The most intuitive metric for this is the *agreement*, which is simply the average number of nodes whose communities were estimated correctly, up to an arbitrary relabeling of the communities. This relabeling must be taken into account since the choice of the integer associated to a community is arbitrary.

Definition 3 (Agreement)**.** Let $Z^\star$ and $\hat{Z}$ be, respectively, the true and an arbitrary vector of community assignments. Let also S_k denote the group of permutations of $[k]$. Define the *agreement* between $Z^\star$ and $\hat{Z}$ to be

$$A(Z^\star, \hat{Z}) := \frac{1}{n} \max_{\sigma \in S_k} \sum_{i=1}^{n} \mathbf{1}(Z_i^\star = \sigma \hat{Z}_i). \tag{3}$$

However, when studying weaker forms of recovery (described below) under general (asymmetric) SBMs, a *normalized* version of this metric is actually needed.

Definition 4 (Normalized agreement)**.** Let $Z^\star$ and $\hat{Z}$ be, respectively, the true and an arbitrary vector of community assignments. Let also S_k denote the group of permutations of $[k]$. Define the *normalized agreement* between $Z^\star$ and $\hat{Z}$ to be

$$\tilde{A}(Z^\star, \hat{Z}) := \frac{1}{k} \max_{\sigma \in S_k} \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \mathbf{1}(Z_i^\star = \sigma \hat{Z}_i) \mathbf{1}(Z_i^\star = k)}{\sum_{i=1}^{n} \mathbf{1}(Z_i^\star = k)}. \tag{4}$$

0.3.2 Degrees of recovery

Definition 3 can be used to measure different degrees of performance of algorithms whose task is to estimate the communities $Z^\star$. The degree to which an algorithm is capable of recovering the communities is captured in what are called the different (asymptotic) *degrees of recovery*.

Definition 5 (Degrees of recovery)**.** Let $(Z^\star, G) \sim \text{SBM}(n, \pi^\star, \Gamma^\star)$, and $\hat{Z}$ be the output of an algorithm taking $(G, \pi^\star, \Gamma^\star)$ as input. Then, the following *degrees of recovery* are said to be *solved* if, asymptotically on n, one has

- Exact recovery $\leftrightarrow \mathbb{P}(A(Z^\star, \hat{Z}) = 1) = 1 - o(1)$,
- Almost exact recovery $\leftrightarrow \mathbb{P}(A(Z^\star, \hat{Z}) = 1 - o(1)) = 1 - o(1)$,
- Partial recovery $\leftrightarrow \mathbb{P}(\tilde{A}(Z^\star, \hat{Z}) \geq \alpha) = 1 - o(1), \alpha \in (1/k, 1)$,
- Weak recovery (also called detection) $\leftrightarrow \mathbb{P}(\tilde{A}(Z^\star, \hat{Z}) \geq 1/k + \Omega(1)) = 1 - o(1)$.

Remark. There is an intuition for why $\alpha > 1/k$ in the definitions of partial and weak recovery above. If one assumes knowledge of the true parameters $(\pi^\star, \Gamma^\star)$ of the model when designing an estimation algorithm, then the trivial algorithm of simply assigning each node a random community according to $\pi^\star$ will achieve an agreement of $\|\pi\|_2^2$, by the law of large numbers. In particular, in the case where the communities are uniform, $\pi = \mathbf{1}_k/k$ and the trivial agreement reached will be $A = 1/k$.

0.4 Asymptotic topologies

0.4.1 The case of the Erdős-Rényi model

The first random graph model was the Erdős-Rényi model, denoted $G(n, p)$, over graphs with n vertices [**?**]. Under this model, the presence of an edge between each pair of nodes is determined by a Bernoulli random variable of parameter p. See Figure 3.

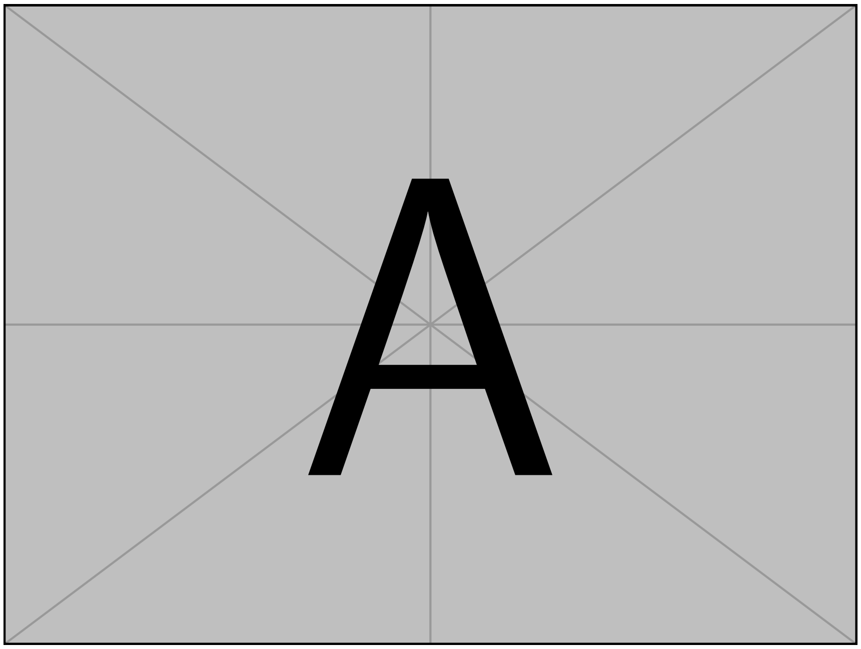


Figure 3: An observation from an Erdős-Rényi model $G(30, 0.2)$

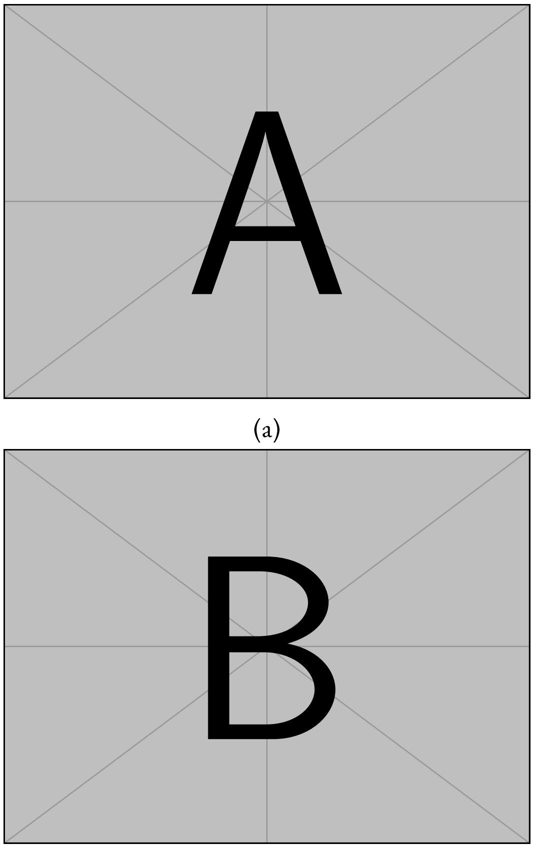
Although this model does not present clusters of nodes, it is nevertheless of great interest, since it reveals a key phenomenon: there exist tightly defined and distinct "asymptotic topologies" for random graphs arising from this model as $n \to \infty$. Which one arises is a function of the growth of p with respect to n.

Theorem 1. *content...*

In the SBM, essentially the same phenomenon happens, and this affects directly the performance of algorithms, and has led to ideas of how to increase their robustness.

0.4.2 The case of the SBM

0.4.3 Why does this matter?



(a)

(b)

Figure 4: (a) At $np = 0.8 < 1$, there are some small trees of size at most $O(\log(n))$. (b) At $np = 1.33 > 1$ a giant component appears, of size $O(n^{2/3})$.

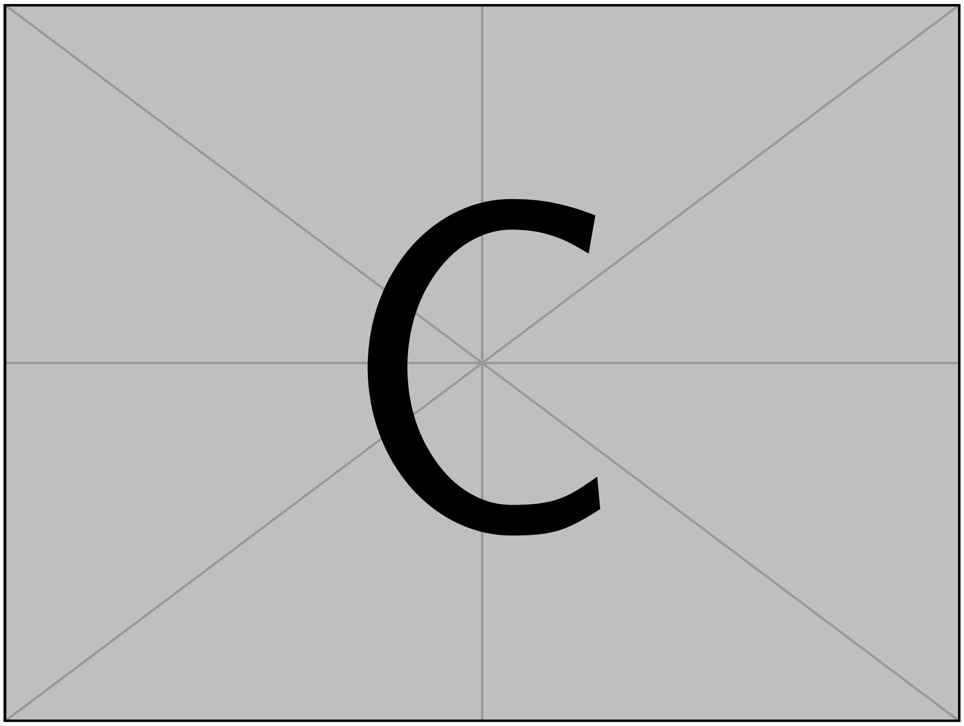


Figure 5: At $p = 0.011 < 0.012$ there exists almost surely an isolated vertex, and the graph is disconnected. When $p = 0.013 > 0.012$, isolated vertices disappear almost surely, and the graph finally becomes connected.