CS-E4075 - Special Course in Machine Learning, Data Science and Artificial Intelligence D: Signed graphs: spectral theory and applications

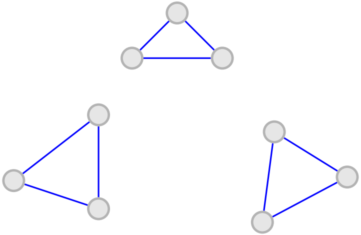
Clustering under the stochastic block model

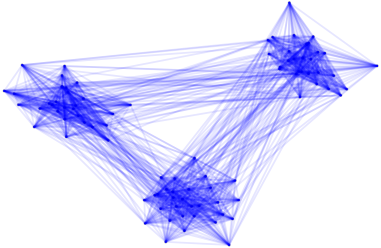
Bruno Ordozgoiti

Aalto University 2021

Outline

- We know that the “bottom” eigenvectors of the graph Laplacian reveal “easy” clusters.
- However, real-world graphs are noisy.
- When does spectral clustering work?





The stochastic block model

The stochastic block model

Consider a graph $G = (V, E)$, and a partition $V = B_1 \cup \cdots \cup B_k$.

Define a matrix $B \in \mathbb{R}^{k \times k}$, $B_{ij} = B_{ji} \in [0, 1]$.

Define a mapping $c : V \to \{1, \ldots, k\}$

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(B_{c(i)c(j)}) & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

The stochastic block model

Population matrix: $\mathbb{E}[A] = P$.

Example:

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

3 communities of size 3.

$$P = \begin{pmatrix} 0 & p & p & q & q & q & q & q & q \\ p & 0 & p & q & q & q & q & q & q \\ p & p & 0 & q & q & q & q & q & q \\ q & q & q & 0 & p & p & q & q & q \\ q & q & q & p & 0 & p & q & q & q \\ q & q & q & p & p & 0 & q & q & q \\ q & q & q & q & q & q & 0 & p & p \\ q & q & q & q & q & q & p & 0 & p \\ q & q & q & q & q & q & p & p & 0 \end{pmatrix}$$

The stochastic block model

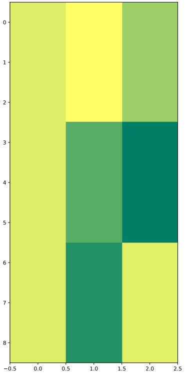
$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{3, 3, 3\}$,
$p = 0.8, q = 0.2$

$$\mathcal{L} = \begin{pmatrix} \delta & -p & -p & -q & -q & -q & -q & -q & -q \\ -p & \delta & -p & -q & -q & -q & -q & -q & -q \\ -p & -p & \delta & -q & -q & -q & -q & -q & -q \\ -q & -q & -q & \delta & -p & -p & -q & -q & -q \\ -q & -q & -q & -p & \delta & -p & -q & -q & -q \\ -q & -q & -q & -p & -p & \delta & -q & -q & -q \\ -q & -q & -q & -q & -q & -q & \delta & -p & -p \\ -q & -q & -q & -q & -q & -q & -p & \delta & -p \\ -q & -q & -q & -q & -q & -q & -p & -p & \delta \end{pmatrix}$$

$\delta = 2p + 6q$.

Bottom eigenvectors of $\mathcal{L}$:



The stochastic block model

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{3, 3, 3\}$,
$p = 0.8, q = 0.2$

The stochastic block model

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{3, 3, 3\}$,
$p = 0.8, q = 0.2$

- What are the eigenvectors of P?

The stochastic block model

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{3, 3, 3\}$,
$p = 0.8, q = 0.2$

- What are the eigenvectors of P?
- What are the eigenvectors of $\mathcal{D}^{-1/2}\mathcal{L}\mathcal{D}^{-1/2}$?
  - $\mathcal{D} = diag(\delta, \delta, \ldots, \delta)$.

The stochastic block model

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{3, 3, 3\}$,
$p = 0.8, q = 0.2$

- What are the eigenvectors of P?
- What are the eigenvectors of $\mathcal{D}^{-1/2}\mathcal{L}\mathcal{D}^{-1/2}$?
  - $\mathcal{D} = diag(\delta, \delta, \ldots, \delta)$.

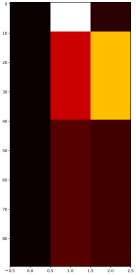
In the *planted communities model* (two probabilities, p, q), if the communities are of the same size, all these matrices have the same eigenvectors.

The stochastic block model

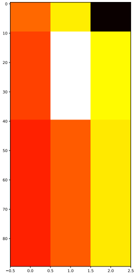
$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

Community sizes: $\{10, 30, 50\}$,
$p = 0.65, q = 0.35$

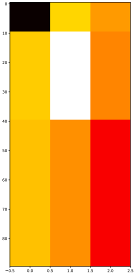
Bottom eigenvectors of $\mathcal{L}$:



Bottom eigenvectors of $\mathcal{L}_n$:

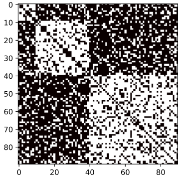


Top eigenvectors of P:



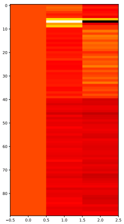
The stochastic block model

Sample from the SBM:

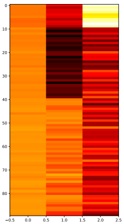


Community sizes: $\{10, 30, 50\}$,
$p = 0.65, q = 0.35$

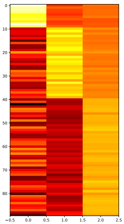
Bottom eigenvectors of L:



Bottom eigenvectors of L_n:



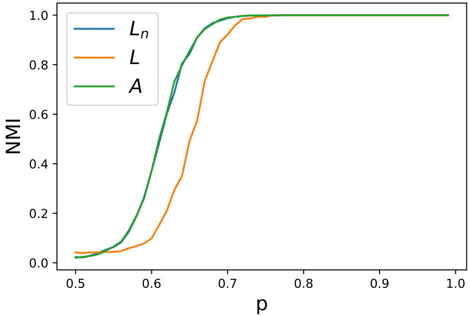
Top eigenvectors of A:

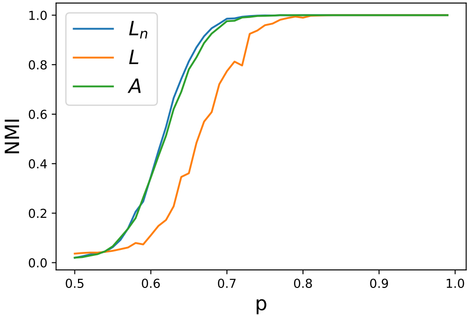


The stochastic block model

Clustering performance using eigenvectors of different matrices: L_n, L, A.

We vary the value of $p \in [0.5, 1], q = 1 - p$.





Theory

The stochastic block model

The eigenvectors of $\mathcal{L}_n$ are perfect for clustering.

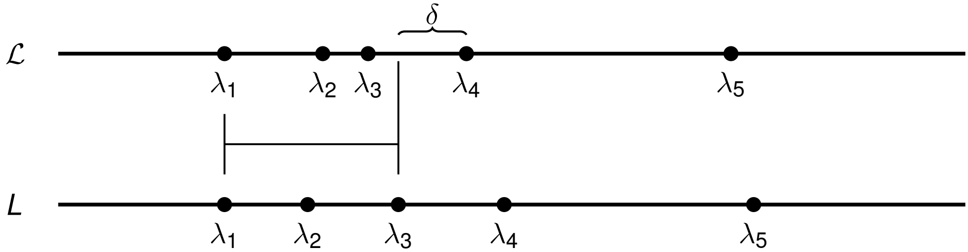
However, we will not be clustering this graph, but a graph sampled from the corresponding SBM.

Let G be a sampled graph, and L_n its Laplacian.

The SBM analysis literature aims to show how and when the eigenvectors of L_n resemble those of $\mathcal{L}_n$.

The stochastic block model

Davis-Kahan theorem: How similar are the eigenvectors of L and $\mathcal{L}$?



Given two matrices $L, \mathcal{L}$, the difference between their eigenvector spaces can be bounded as follows:

$$\|X - X'\|_F^2 \leq \frac{2\|L - \mathcal{L}\|_F^2}{\delta^2}.$$

The stochastic block model

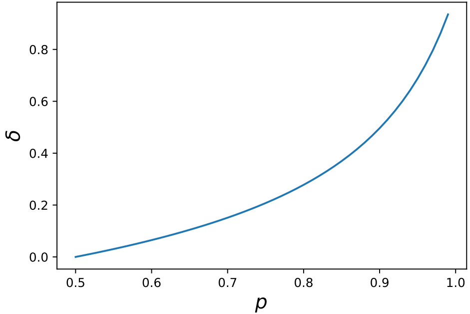
Example of analysis[1].

- $\|L - \mathcal{L}\|$ does not converge, but $\|LL - \mathcal{L}\mathcal{L}\|$ converges as the graph size increases.
- If $\|LL - \mathcal{L}\mathcal{L}\|$ is small, the eigenvectors of L and $\mathcal{L}$ "should" be close.
- Davis-Kahan theorem bounds how close, based on
  - $\|LL - \mathcal{L}\mathcal{L}\|$,
  - the gap between the eigenvalues of interest and the rest.
- The eigenvalue gap is related to $|p - q|$.
- If $|p - q|$ is "large", spectral clustering will work on sufficiently large, dense graphs.

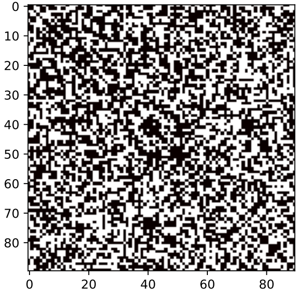
[1] Rohe, Karl, Sourav Chatterjee, and Bin Yu. "Spectral clustering and the high-dimensional stochastic blockmodel." Annals of Statistics 39.4 (2011): 1878-1915.

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



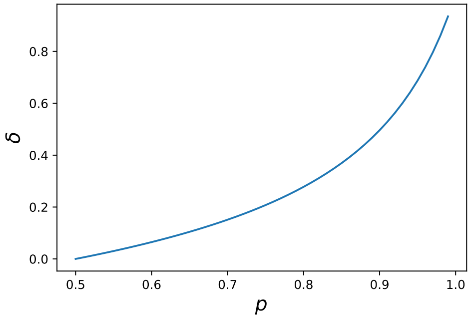
Sample from the SBM:



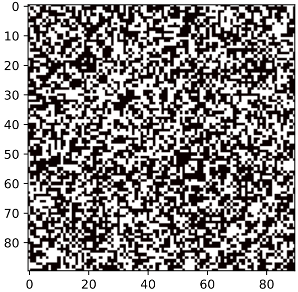
$p = 0.50, q = 1 - p$
$\delta = 0.000$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

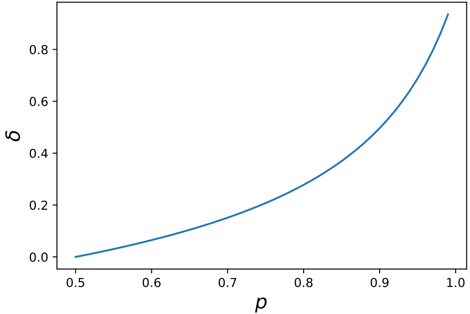


$p = 0.51, q = 1 - p$

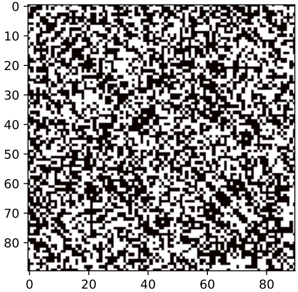
$\delta = 0.006$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



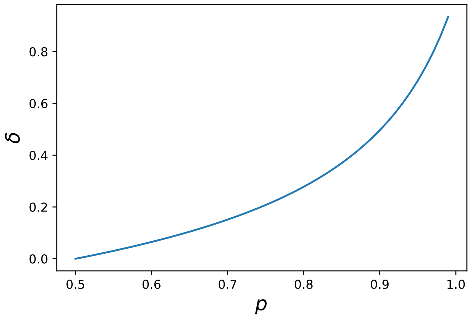
Sample from the SBM:



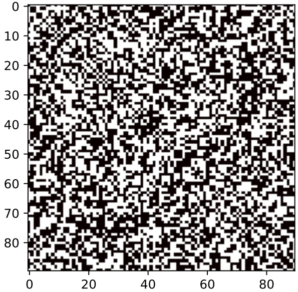
$p = 0.52, q = 1 - p$

$\delta = 0.012$

$|p - q|$ vs. eigenvalue gap:



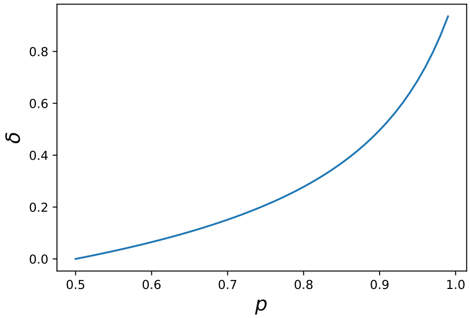
Sample from the SBM:



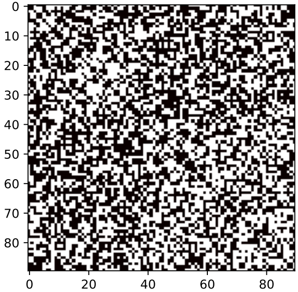
$p = 0.53, q = 1 - p$
$\delta = 0.018$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

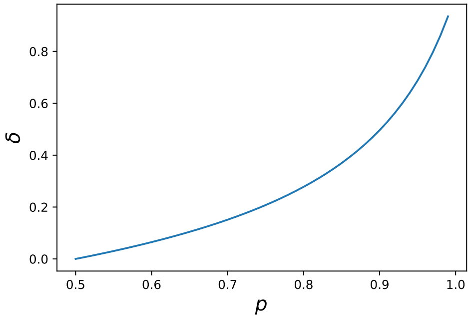


$p = 0.54, q = 1 - p$

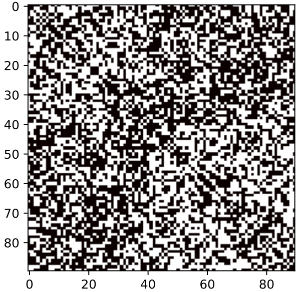
$\delta = 0.024$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

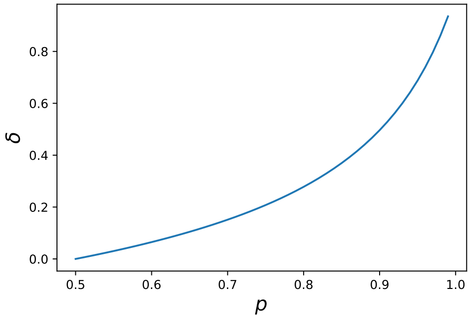


$p = 0.55, q = 1 - p$

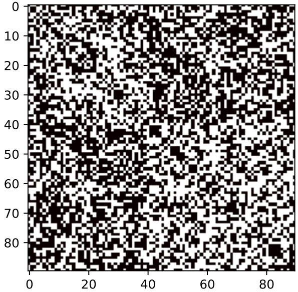
$\delta = 0.030$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

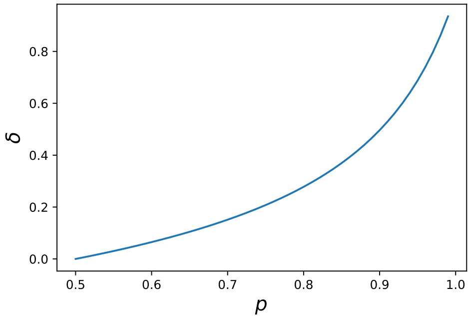


$p = 0.56, q = 1 - p$

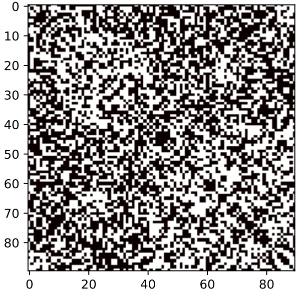
$\delta = 0.037$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

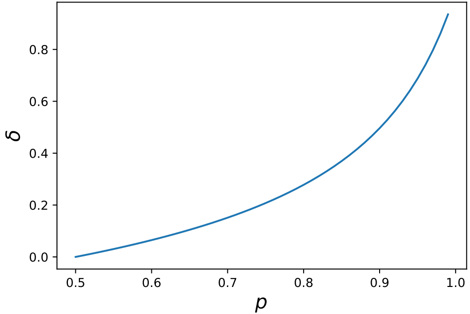


$p = 0.57, q = 1 - p$

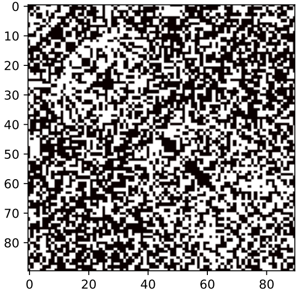
$\delta = 0.044$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

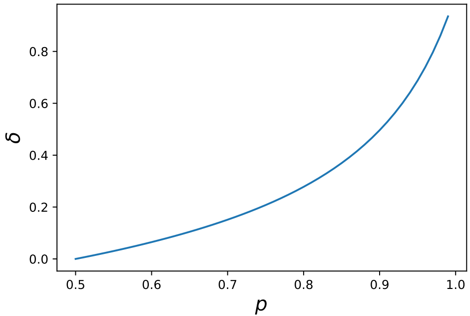


$p = 0.58, q = 1 - p$

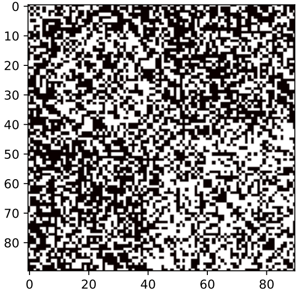
$\delta = 0.051$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

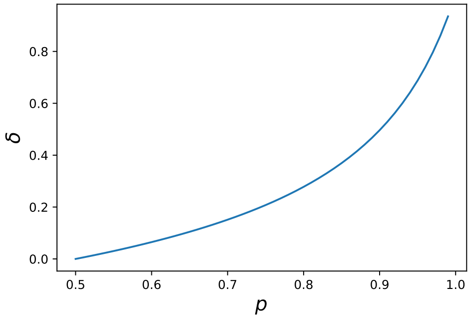


$p = 0.59, q = 1 - p$

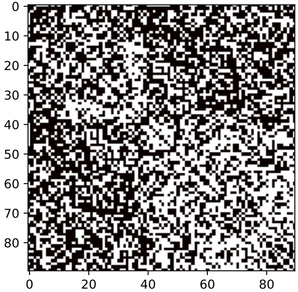
$\delta = 0.058$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

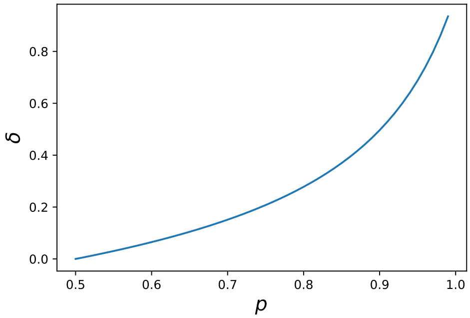


$p = 0.60, q = 1 - p$

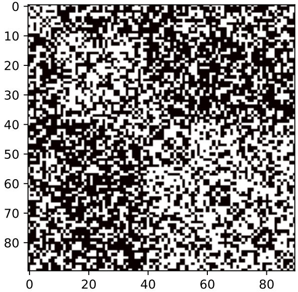
$\delta = 0.065$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

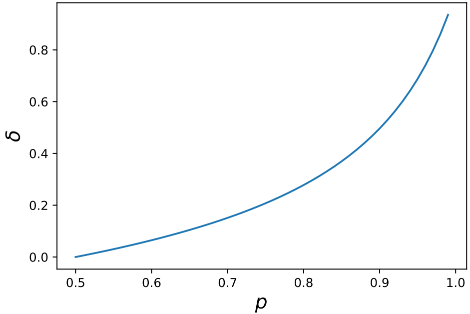


$p = 0.61, q = 1 - p$

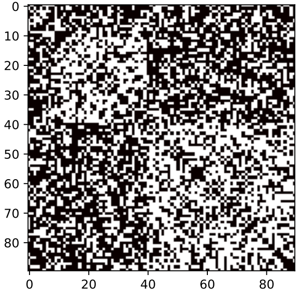
$\delta = 0.072$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

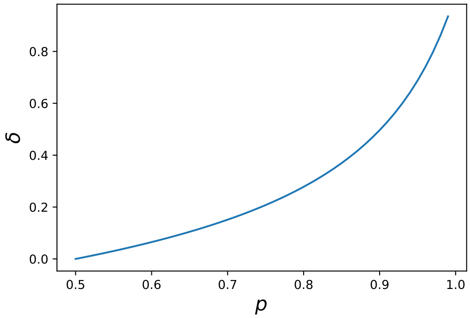


$p = 0.62, q = 1 - p$

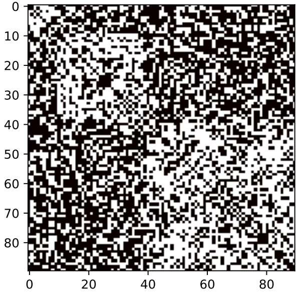
$\delta = 0.080$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

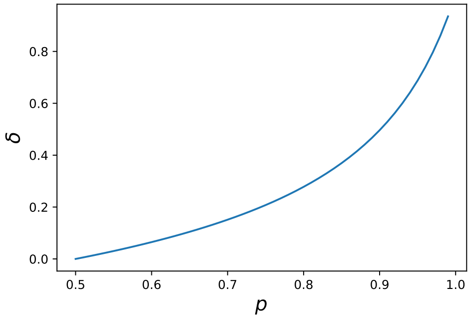


$p = 0.63, q = 1 - p$

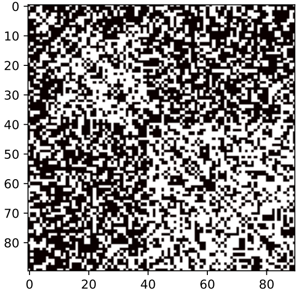
$\delta = 0.088$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

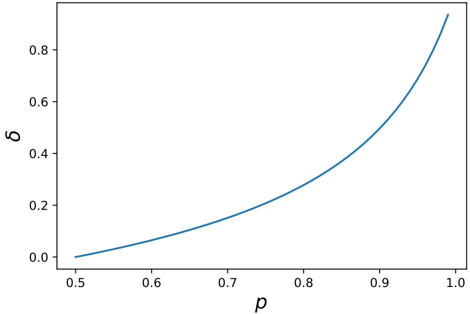


$p = 0.64, q = 1 - p$

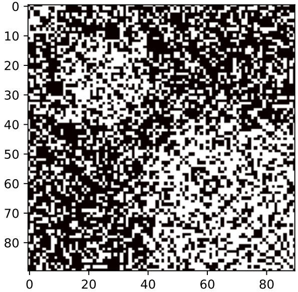
$\delta = 0.096$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

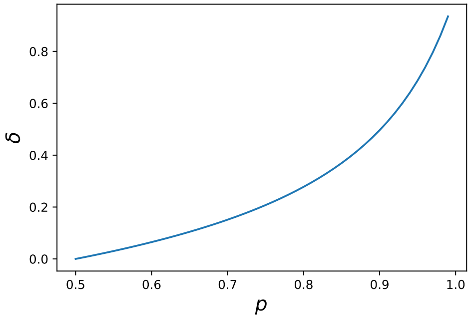


$p = 0.65, q = 1 - p$

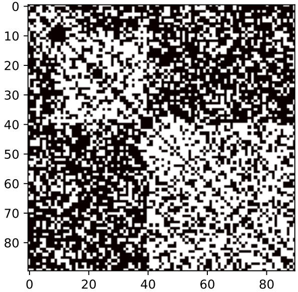
$\delta = 0.105$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

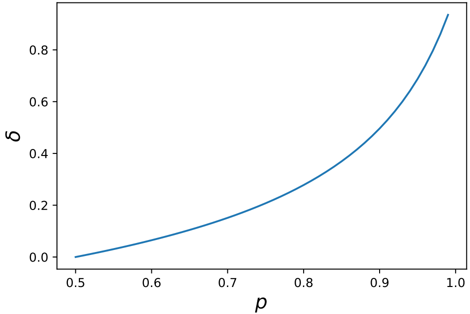


$p = 0.66, q = 1 - p$

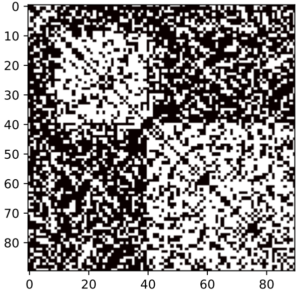
$\delta = 0.113$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

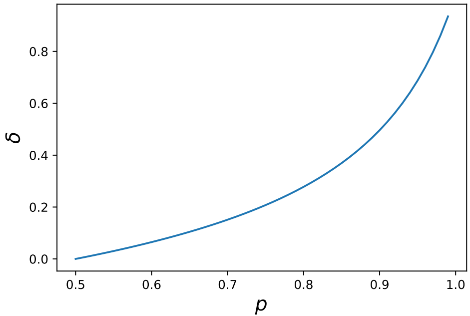


$p = 0.67, q = 1 - p$

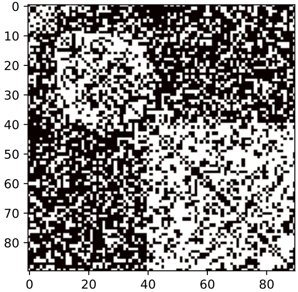
$\delta = 0.122$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

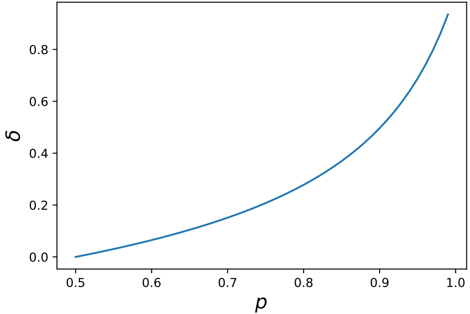


$p = 0.68, q = 1 - p$

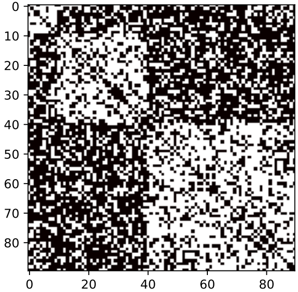
$\delta = 0.132$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

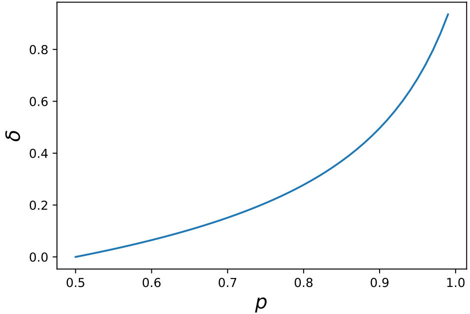


$p = 0.69, q = 1 - p$

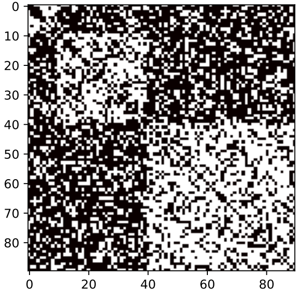
$\delta = 0.141$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

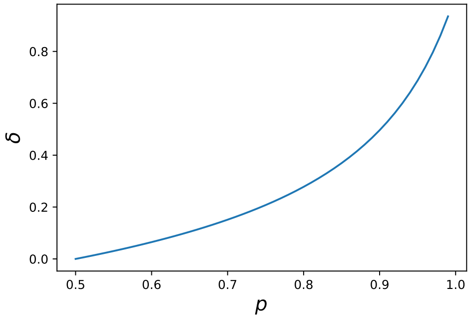


$p = 0.70, q = 1 - p$

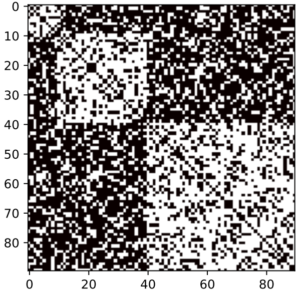
$\delta = 0.151$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

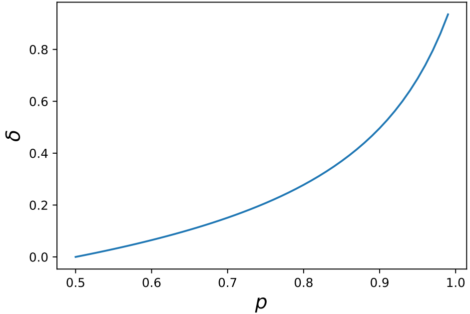


$p = 0.71, q = 1 - p$

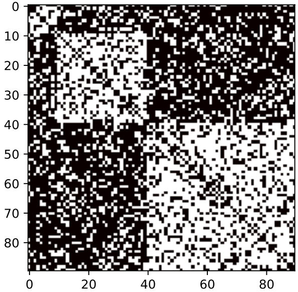
$\delta = 0.162$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



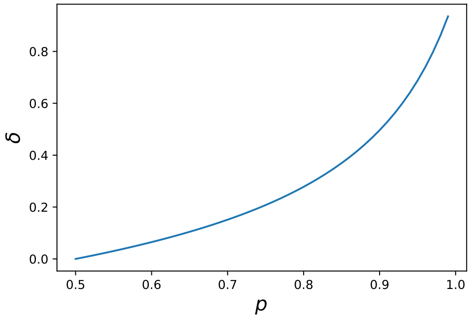
Sample from the SBM:



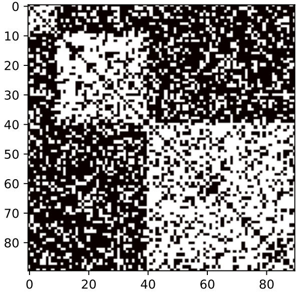
$p = 0.72, q = 1 - p$
$\delta = 0.172$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

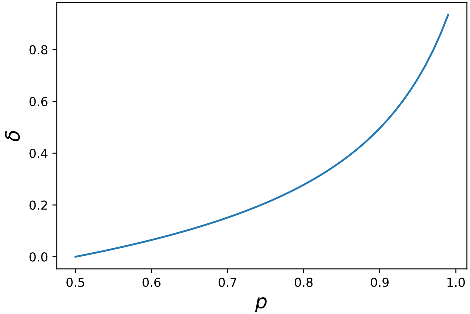


$p = 0.73, q = 1 - p$

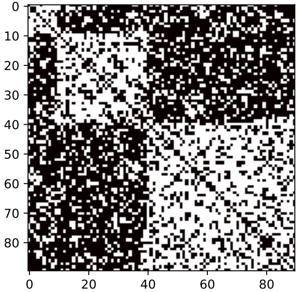
$\delta = 0.184$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



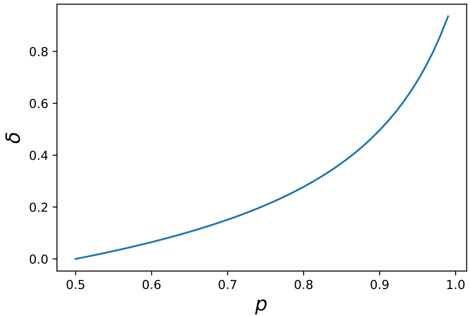
Sample from the SBM:



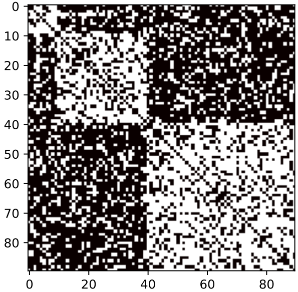
$p = 0.74, q = 1 - p$
$\delta = 0.195$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

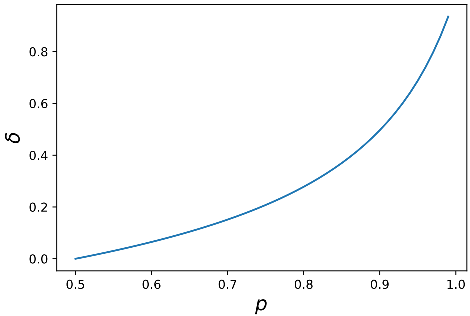


$p = 0.75, q = 1 - p$

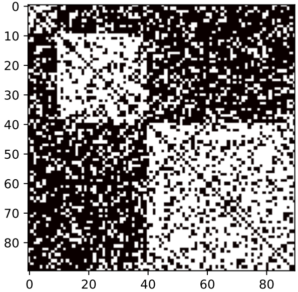
$\delta = 0.208$

The stochastic block model

$|p-q|$ vs. eigenvalue gap:



Sample from the SBM:

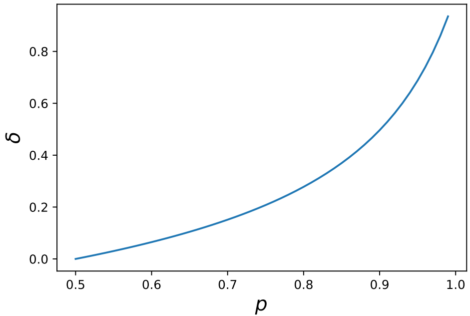


$p = 0.76, q = 1 - p$

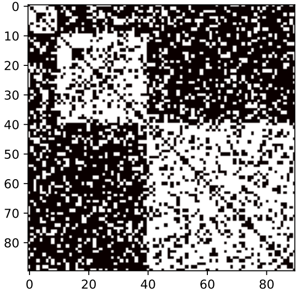
$\delta = 0.220$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

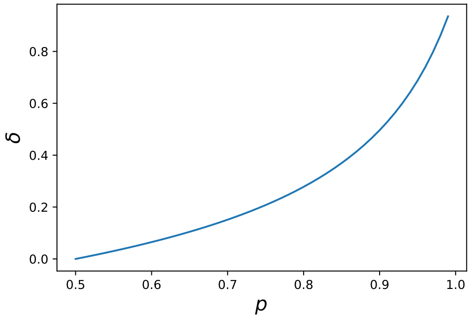


$p = 0.77, q = 1 - p$

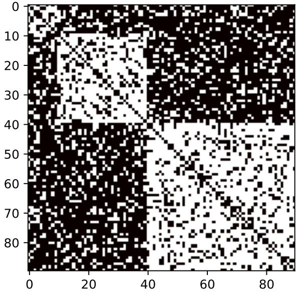
$\delta = 0.234$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

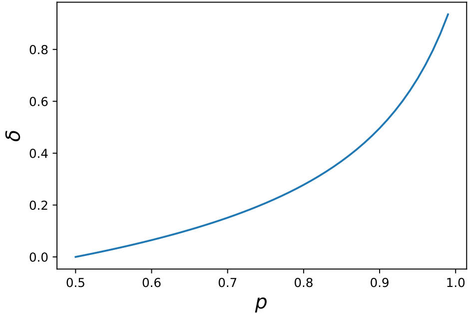


$p = 0.78, q = 1 - p$

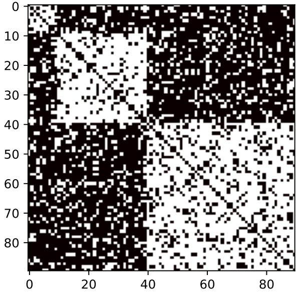
$\delta = 0.248$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

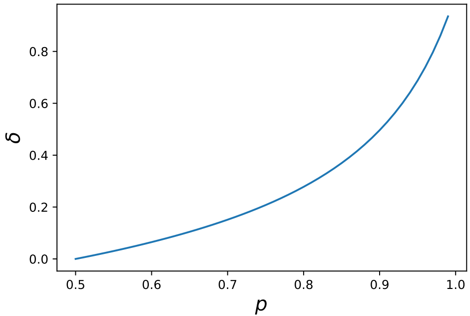


$p = 0.79, q = 1 - p$

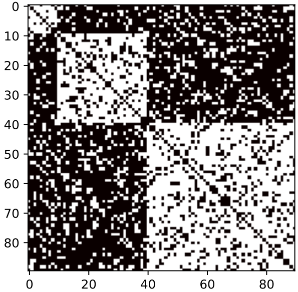
$\delta = 0.262$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



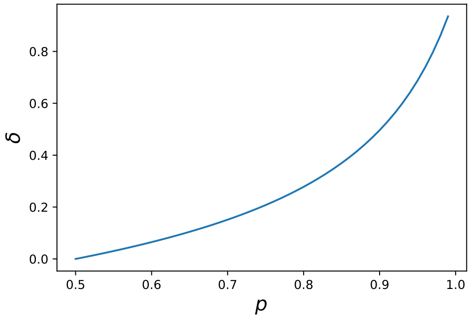
Sample from the SBM:



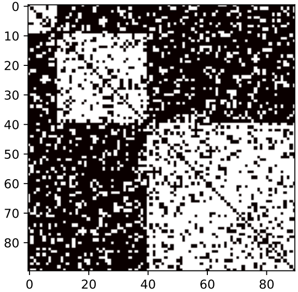
$p = 0.80, q = 1 - p$
$\delta = 0.278$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

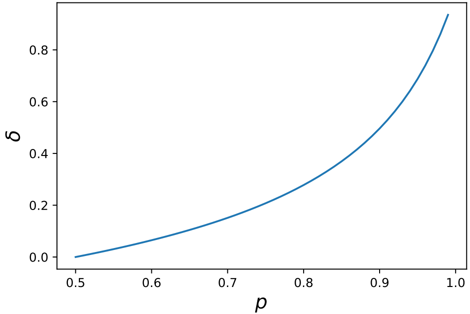


$p = 0.81, q = 1 - p$

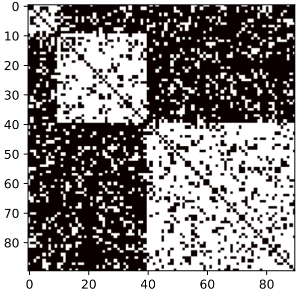
$\delta = 0.294$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



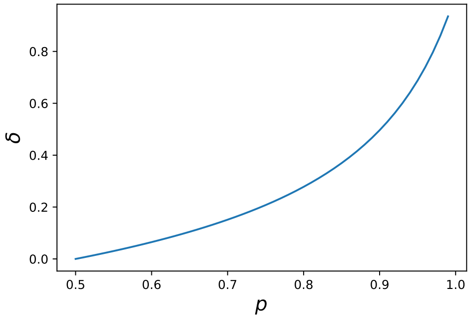
Sample from the SBM:



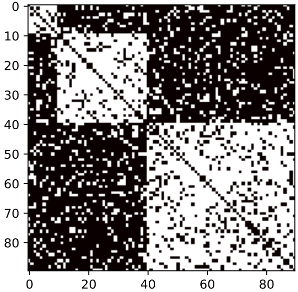
$p = 0.82, q = 1 - p$
$\delta = 0.311$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

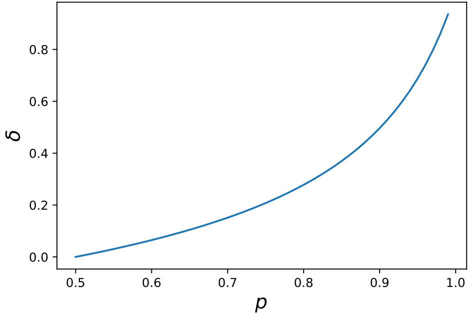


$p = 0.83, q = 1 - p$

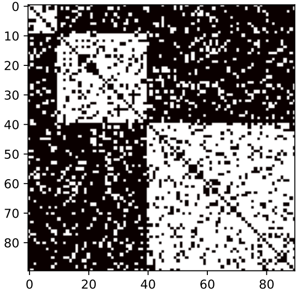
$\delta = 0.330$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

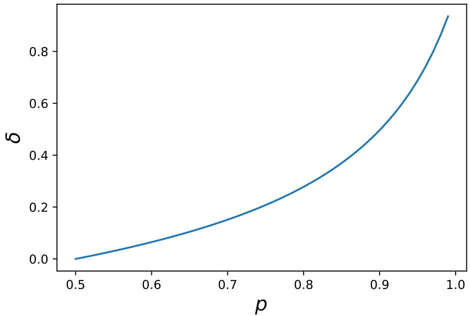


$p = 0.84, q = 1 - p$

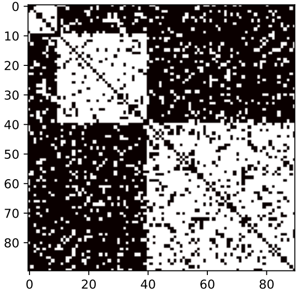
$\delta = 0.349$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



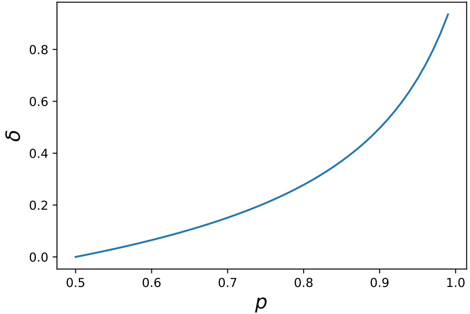
Sample from the SBM:



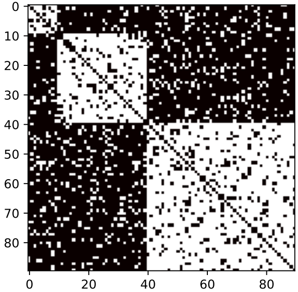
$p = 0.85, q = 1 - p$
$\delta = 0.370$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



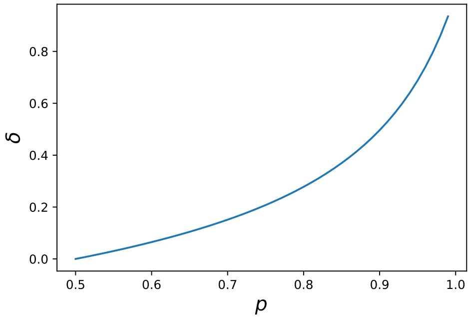
Sample from the SBM:



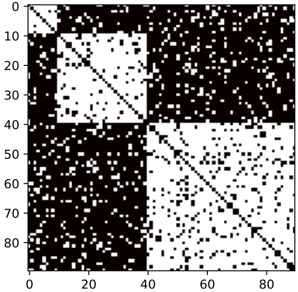
$p = 0.86, q = 1 - p$
$\delta = 0.391$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

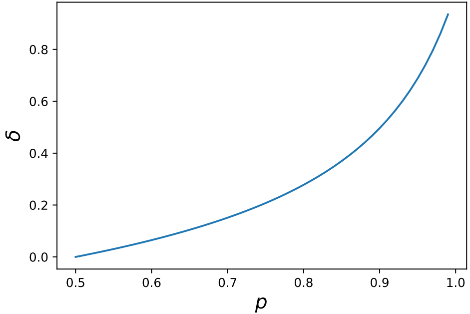


$p = 0.87, q = 1 - p$

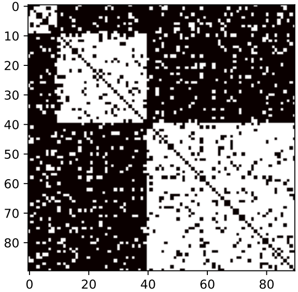
$\delta = 0.415$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

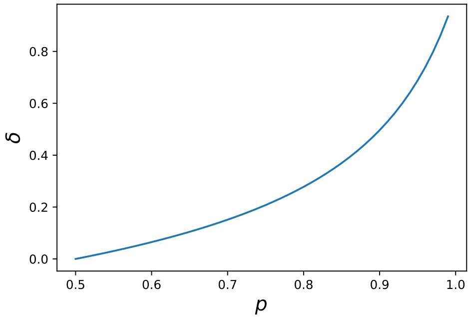


$p = 0.88, q = 1 - p$

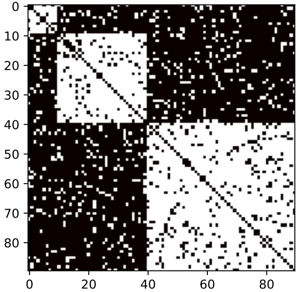
$\delta = 0.440$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

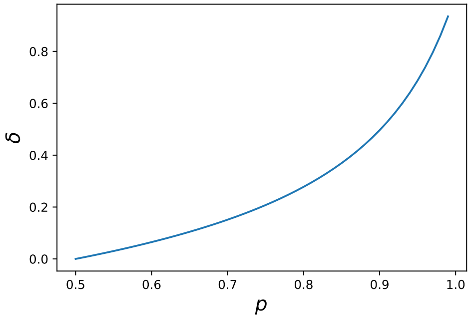


$p = 0.89, q = 1 - p$

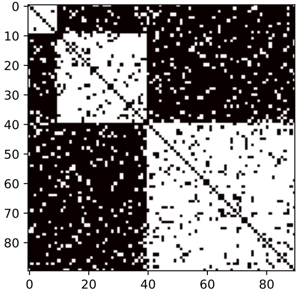
$\delta = 0.467$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

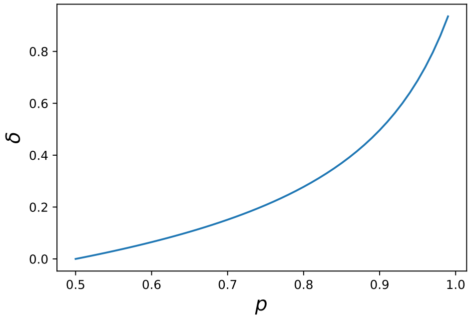


$p = 0.90, q = 1 - p$

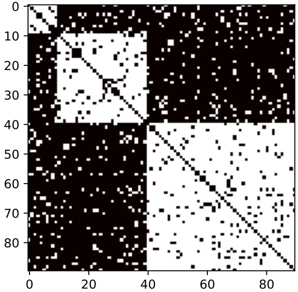
$\delta = 0.496$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



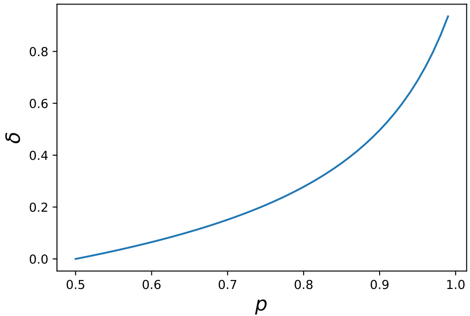
Sample from the SBM:



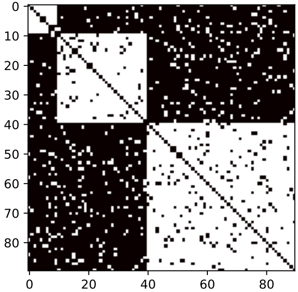
$p = 0.91, q = 1 - p$
$\delta = 0.528$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

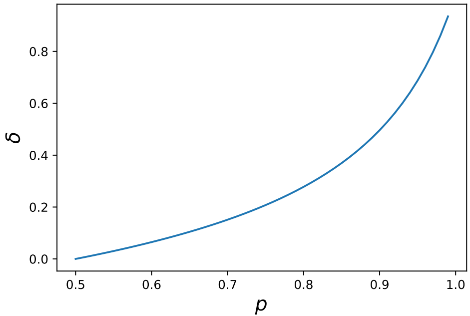


$p = 0.92, q = 1 - p$

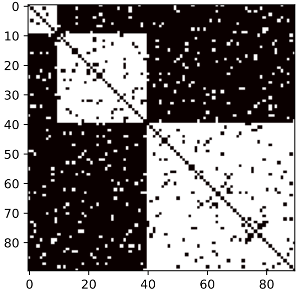
$\delta = 0.563$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

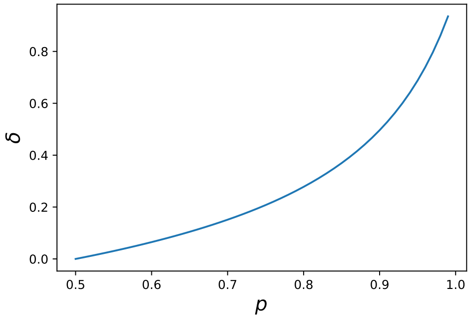


$p = 0.93, q = 1 - p$

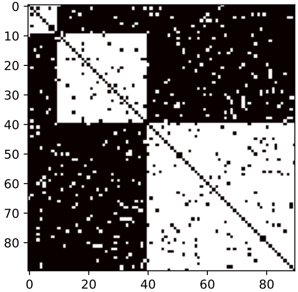
$\delta = 0.600$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

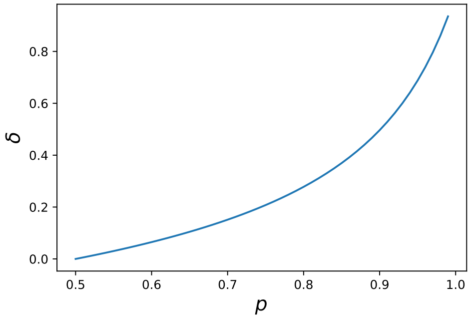


$p = 0.94, q = 1 - p$

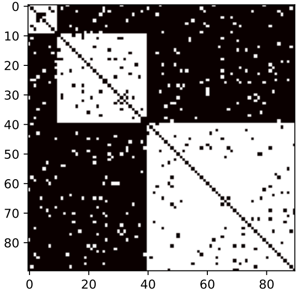
$\delta = 0.642$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

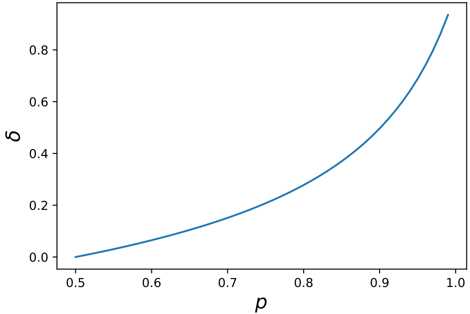


$p = 0.95, q = 1 - p$

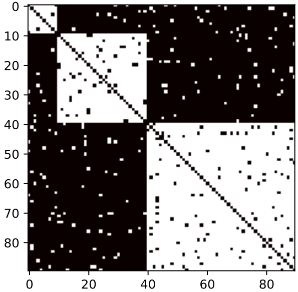
$\delta = 0.688$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

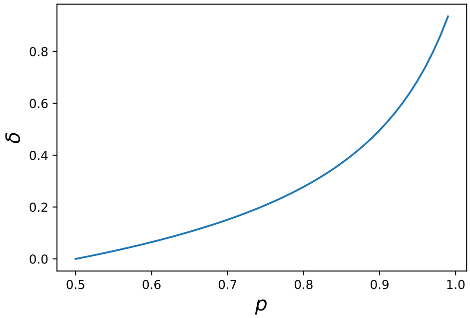


$p = 0.96, q = 1 - p$

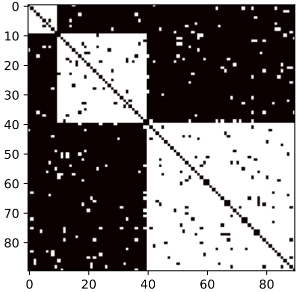
$\delta = 0.739$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

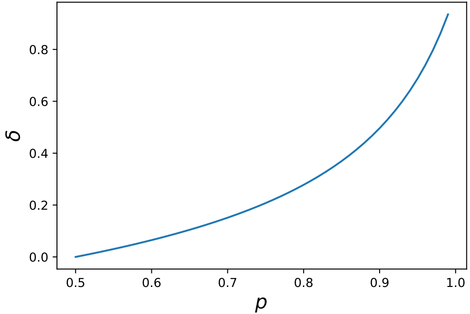


$p = 0.97, q = 1 - p$

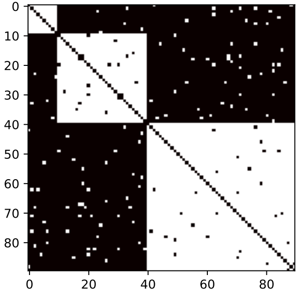
$\delta = 0.796$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:

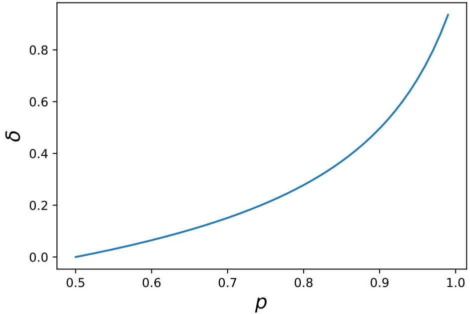


$p = 0.98, q = 1 - p$

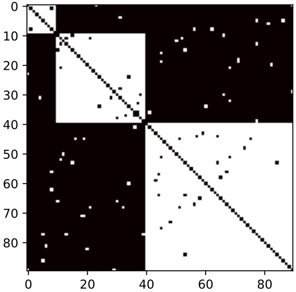
$\delta = 0.861$

The stochastic block model

$|p - q|$ vs. eigenvalue gap:



Sample from the SBM:



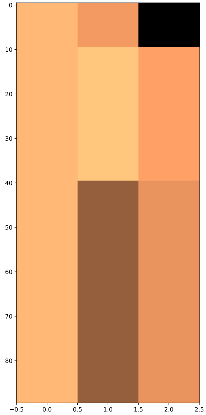
$p = 0.99, q = 1 - p$

$\delta = 0.935$

The stochastic block model

Note: even if $p = 0.51, q = 0.49$, (community sizes: $\{10, 30, 50\}$)

Bottom eigenvectors of $\mathcal{L}$:



The signed stochastic block model

The stochastic block model

Based on the work by Mercado et al.[2]

Consider a signed graph $G = (V, E^+, E^-)$.

The Signed SBM (SSBM) has four probability parameters: p_+, p_-, q_+, q_-.

Edges are sampled independently, so an edge can be positive and negative.

Define

- $\mathcal{A}^+ = \mathbb{E}[A^+]$,
- $\mathcal{A}^- = \mathbb{E}[A^-]$,
- $\mathcal{D}^+$ so that $\mathcal{D}^+_{ii} = \sum_j \mathcal{A}^+_{ij}$,
- $\mathcal{D}^-$ so that $\mathcal{D}^-_{ii} = \sum_j \mathcal{A}^-_{ij}$,

- $\mathcal{L}_n = {\mathcal{D}^+}^{-1/2}(\mathcal{D}^+ - \mathcal{A}^+){\mathcal{D}^+}^{-1/2}$,
- $\mathcal{Q}_n = {\mathcal{D}^-}^{-1/2}(\mathcal{D}^- - \mathcal{A}^-){\mathcal{D}^-}^{-1/2}$,
- $\mathcal{L}_p = \left(\frac{\mathcal{L}_n^p + \mathcal{Q}_n^p}{2}\right)^{1/p}$.

[2] Mercado, Pedro, Francesco Tudisco, and Matthias Hein. "Spectral clustering of signed graphs via matrix power means." International Conference on Machine Learning. PMLR, 2019.

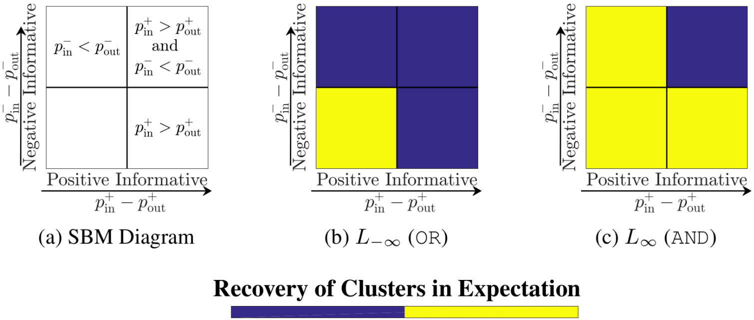
The stochastic block model

Power means generalize other means. For $a, b \in \mathbb{R}$, $m_p(a,b) = \left(\frac{a^p + b^p}{2}\right)^{1/p}$:

$p \to -\infty$	$m_p(a,b) = \min\{a,b\}$	
$p = -1$	$m_p(a,b) = 2\left(\frac{1}{a} + \frac{1}{b}\right)$	(harmonic mean)
$p \to 0$	$m_p(a,b) = \sqrt{ab}$	(geometric mean)
$p = 1$	$m_p(a,b) = (a+b)/2$	(arithmetic mean)
$p \to \infty$	$m_p(a,b) = \max\{a,b\}$	

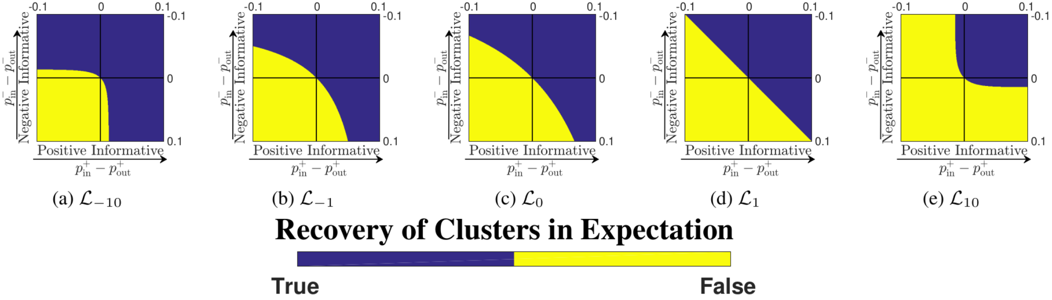
The stochastic block model

The conditions for recovery depend on p_+, p_-, q_+, q_- and p.



(a) SBM Diagram (b) $L_{-\infty}$ (OR) (c) L_{∞} (AND)

The stochastic block model



Take-aways from this lecture:

- The stochastic block model.
- Analysis of the SBM.
- The Davis-Kahan theorem (eigenvector perturbation).
- The signed stochastic block model.