高维概率

High-Dimensional Probability

十一、应用：随机矩阵的若干应用

@滕佳烨

- 上节课说了啥

四种情形下的Concentration

Coordinate, row, symmetry

sub-Gaussian?

期望形式?

双边?

- 这节课要说啥

  应用：纠错编码

  应用：网络中的社区检测

- **应用：纠错编码**

我们想要通过微信发送“高维概率”四个字。于是我们在屏幕上打出“高维概率”，然后点击发送。

然而，这种传输可能有失误，比如“维”被误传成“哔”，那么对方只能收到“高哔概率”。我们怎么解决这样的问题呢？

或者说，我们能不能找到一个传输方法，能够抵挡住这样的失误呢？

- **应用：纠错编码**

  一个简单的尝试：

  如果我们传输的是
  “高维概率高维概率高维概率高维概率高维概率高维概率高维概率”

  即使其中几个字变成“哔”，我们也仍然能够得到原始信息
  “高哔概率高维概率哔维概率高维概哔哔维概哔高维概率高维哔率”

  ——通过频数统计就能做到

- **问题：效率太低**

- **应用：纠错编码**

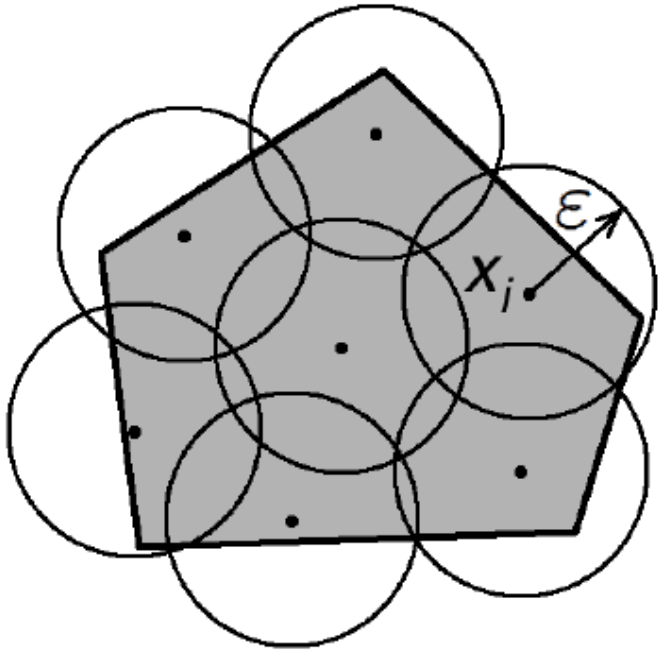
最短能用几个字符，来传播一个长度为n的字符串呢？

首先假设传输的误差是**不大**的。
我们转换一下这个问题。假设字符的分布是一个高维空间。这个空间中有互不相交的一堆球，并且球心是有效的字符。
这样，我们将传输后的字符转成球心字符，则能得到原始信息。

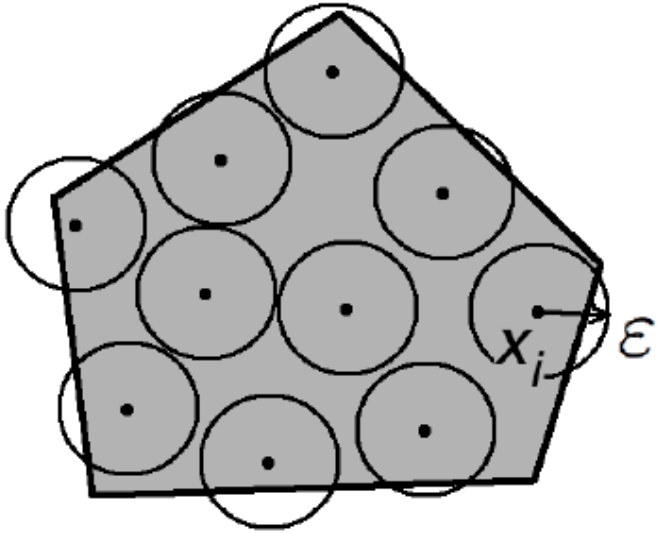
所以问题就变成，在多大的高维空间内，才能有n个不相交的半径为d的球呢？
---Packing number

- 应用：纠错编码

Packing number 与 Covering number



(a) This covering of a pentagon K by seven ε-balls shows that $\mathcal{N}(K, \varepsilon) \leq 7$.



(b) This packing of a pentagon K by ten ε-balls shows that $\mathcal{P}(K, \varepsilon) \geq 10$.

- **应用：网络中的社区检测**

假设有社区甲和社区乙，每个社区有n/2个人。社区内人群认识的比例为p，社区间人群认识的比例为q。这样形成的随机图被叫做stochastic block model，写作G(n,p,q)

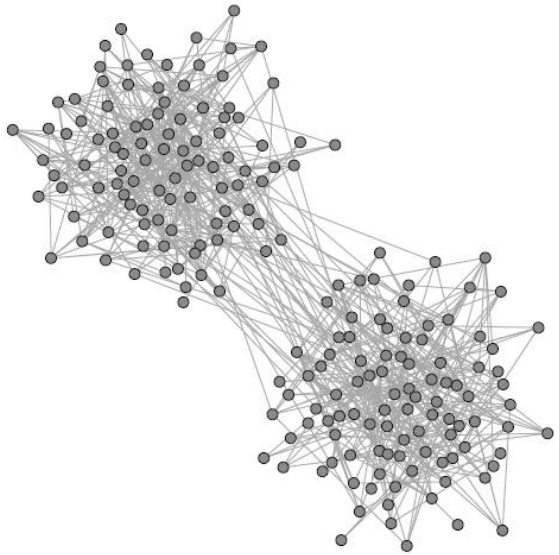


Figure 4.4 A random graph generated according to the stochastic block model $G(n, p, q)$ with $n = 200$, $p = 1/20$ and $q = 1/200$.

- **应用：网络中的社区检测**

A：连接矩阵，n*n大小，定义$D = EA$，则$A = D + R$
我们的目标是：**若给出A，我们要分辨哪些人属于一个社区！**
注意到：D中含有很重要的信息！倘若我们已知D，则完全可以分辨人群的归属。例如，$\mathbf{u_2}$（特征向量）对应的就是社区归属情况：

$$D = \mathbb{E}\, A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right].$$

Exercise 4.5.2. ☕☕ Check that the matrix D has rank 2, and the non-zero eigenvalues λ_i and the corresponding eigenvectors u_i are

$$\lambda_1 = \Big(\frac{p+q}{2}\Big)n, \quad u_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \quad \lambda_2 = \Big(\frac{p-q}{2}\Big)n, \quad u_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}. \qquad (4.17)$$

- **应用：网络中的社区检测**

  但是我们不知道D啊！怎么办？

  没关系，注意到$A - D = R$能够通过之前学的矩阵Concentration对最大特征值进行bound，也就是说，R不会很大。因此在高维情况下**可以用A去近似D**。

  具体的近似结果细节，可以参考教材。

- 应用：网络中的社区检测

Spectral Clustering Algorithm

Input: graph G

Output: a partition of the vertices of G into two communities

1: Compute the adjacency matrix A of the graph.
2: Compute the eigenvector $v_2(A)$ corresponding to the second largest eigenvalue of A.
3: Partition the vertices into two communities based on the signs of the coefficients of $v_2(A)$. (To be specific, if $v_2(A)_j > 0$ put vertex j into first community, otherwise in the second.)

Theorem 4.5.6 (Spectral clustering for the stochastic block model). *Let* $G \sim G(n, p, q)$ *with* $p > q$, *and* $\min(q, p-q) = \mu > 0$. *Then, with probability at least* $1 - 4e^{-n}$, *the spectral clustering algorithm identifies the communities of* G *correctly up to* C/μ^2 *misclassified vertices.*

- 应用：纠错编码

Packing number 与 Covering number

- 应用：网络中的社区检测

在高维情况中利用本身去近似均值

谢谢！

@滕佳烨