CS 683: Advanced Design & Analysis of Algorithms March 26, 2008

Lecture 25

Lecturer: John Hopcroft *Scribe: Anand Bhaskar, Yookyung Jo*

1 Catalan Numbers

One way to define the n-th Catalan number, C_n, is the number of valid/balanced parenthesizations with n open and n closed parentheses. C_n can be computed using several ways, and two of them are described here.

1.1 Generating function approach

For a balanced parenthesization of length $2n$ (n '(' and n ')' symbols), the first symbol should be a '('. Let i be the position of the ')' symbol corresponding to the first '(' symbol. i must be even, so let $i = 2k$. Then, the substring from position 2 to $2k-1$ must be a valid parenthesization of length $2k-2$, and also the substring from $2k+1$ to $2n$ must be a valid parenthesization of length $2n-2k$. Ranging k over all possible values along the string, we have the recurrence,

$$C_n = C_0C_{n-1} + C_1C_{n-2} + ... + C_{n-2}C_1 + C_{n-1}C_0 = \sum_{k=1}^{n} C_{k-1}C_{n-k}$$

Above, $C_0 = 1$. Let $C(x) = \sum_{k\geq 0} C_k x^k$. Then, from the above recurrence, we see that

$$C(x) = x[C(x)]^2 + 1$$

Solving the quadratic in $C(x)$, we get

$$C(x) = \frac{1 - \sqrt{1-4x}}{2x}$$

The other root is discarded because the C_i need to be positive.
Then, by the binomial theorem, for $n \geq 1$,

$$\begin{aligned} C_n &= \frac{(1 \cdot 3... \cdot (2n-1))4^{n+1}}{2^{n+2}(n+1)!} \\ &= \frac{(1 \cdot 3... \cdot (2n-1))2^n n!}{n!(n+1)!} \\ &= \frac{(2n)!}{n!(n+1)!} \\ &= \frac{1}{n+1}\binom{2n}{n} \end{aligned}$$

1.2 Combinatorial argument (using the Reflection principle)

The total number of length $2n$ strings with n '(' and n ')' symbols is just $\binom{2n}{n}$.
Consider an unbalanced parenthesization with n '(' and n ')' symbols. Then, there is a prefix of the string such that there is one more ')' symbol than '(' symbol. Let there be k '(' symbols and $k+1$ ')' symbols in this

prefix. Flip all symbols of the string after the prefix and we now have a new string with $k+n-(k+1) = n-1$ '(' symbols and $k+1+n-k = n+1$ ')' symbols. Example,

$$\underbrace{(...))}_{\text{first unbalanced prefix with } k \text{ '(' and } k+1 \text{ ')'}} \qquad \overbrace{()...)}^{n-k \text{ '(' and } n-(k+1) \text{ ')'}}$$

After flipping the second portion, it becomes

$$\underbrace{(...))}_{\text{first unbalanced prefix with } k \text{ '(' and } k+1 \text{ ')'}} \qquad \overbrace{)(...(}^{n-k \text{ ')' and } n-(k+1) \text{ '('}}$$

We claim that there is a bijection between unbalanced parenthesizations with n '(' and n ')' symbols and strings with $n-1$ '(' and $n+1$ ')' symbols. If we have an unbalanced parenthesization with n '(' and n ')' symbols, we can apply the flipping procedure above to produce a string with $n-1$ '(' and $n+1$ ')' symbols. This procedure is reversible and if we have a string with $n-1$ '(' and $n+1$ ')' symbols, there is a first prefix in which there are two more ')' symbols than '(' symbols, ie. k '(' and $k+2$ ')' symbols in the first such prefix. If we flip all the symbols starting from the rightmost ')' symbol in this prefix, we will have a string with $k+(n+1)-(k+1) = n$ '(' symbols and $k+1+n-1-k = n$ ')' symbols, and this parenthesization will be unbalanced since the first $2k+1$ letters of the prefix will have k '(' and $k+1$ ')' symbols.
Using this bijection, we find that there are $\binom{2n}{n-1}$ unbalanced parenthesizations with n '(' and n ')' symbols.

Define
$C_n = \#$ balanced parenthesizations of length $2n$ (the n^{th} Catalan number)
$A_n = \#$ total parenthesizations of length $2n = \binom{2n}{n}$
$B_n = \#$ unbalanced parenthesizations of length $2n = \binom{2n}{n-1}$
Then,

$$\begin{aligned} C_n &= A_n - B_n \\ &= \binom{2n}{n} - \binom{2n}{n-1} \\ &= \frac{1}{n+1}\binom{2n}{n} \end{aligned}$$

Refer to the excellent Wikipedia article `http://en.wikipedia.org/wiki/Catalan_number` for other definitions and derivations.

2 High level review of Wigner's theorem

To show that the probability distribution of the normalized eigenvalues of an $n \times n$ symmetric matrix (with entries 1 and -1 with equal probability) approaches the probability distribution $\frac{2}{\pi}\sqrt{1-\lambda^2}$ over $-1 \le \lambda \le 1$, we show that the k'th moments of both these distributions are same in the limit for n.

The k'th moment of the probability distribution $\frac{2}{\pi}\sqrt{1-\lambda^2}$ where $-1 \le \lambda \le 1$ is given by,

$$c(k) = \int_{-1}^{1} \lambda^k \frac{2}{\pi}\sqrt{1-\lambda^2}d\lambda$$

Making the substitution of $\lambda = \sin\theta$, integrating by parts, and solving the recurrence relation involving the k'th and $(k-1)$'th moments, we get

$$c(k) = \frac{1}{2^{k-1}(k+2)}\binom{k}{\frac{k}{2}}$$

The k'th moment of the probability distribution of the normalized eigenvalues is given as,

$$\begin{aligned} \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\lambda_j}{2\sqrt{n}}\right)^k\right) &= \frac{1}{2^k}\frac{1}{n^{1+k/2}}\mathrm{E}(\sum_{j=1}^{n}\lambda_j^k) \\ &= \frac{1}{2^k}\frac{1}{n^{1+k/2}}\mathrm{E}(\text{trace of } A^k) \end{aligned}$$

where we have used the fact that λ_j^k is the eigenvalue of the matrix A^k because λ_j is the eigenvalue of the matrix A, and the sum of all eigenvalues of a matrix is equal to the trace (the sum of the diagonal entries) of the matrix.

When we interpret A as the adjacency matrix representation of a random graph, the (i, j)' th element of A^k represents all paths of length k from the vertex i to the vertex j. The trace of A^k is the sum of values of all these paths from vertex i back to vertex i of length k for each vertex i in the graph.

The value of each edge is 1 if the edge is present, -1 otherwise. Since edges are generated statistically independently and the expected vaue of each edge is 0, only the paths where each edge in the path appears even number of times have non-zero value (actually the value of such a path is 1).

Asymptotically, the kind of paths that visits $k/2$ number of vertices with each edge visited twice dominates. The number of such paths is given as $\frac{1}{k/2+1}\binom{k}{k/2}\cdot n^{k/2}\cdot n$ where $n^{k/2}$ is the number of assignments of $k/2$ vertices to visit, $\frac{1}{\frac{k}{2}+1}\binom{k}{k/2}$ is the number of possible paths given the vertices (the same as the number of possible DFS paths or possible nested parentheses or the catalan number $C_{k/2}$), n is multiplied because the starting point of such paths could be any of n vertices.

Thus, the resulting k'th moment of the normalized eigenvalue is equivalent to the k'th moment of the distribution $\frac{2}{\pi}\sqrt{1-\lambda^2}$, indicating that the probability distribution of the normalized eigenvalue is $\frac{2}{\pi}\sqrt{1-\lambda^2}$.

3 Finding the structure of a random graph

Given a symmetric matrix, we want to find out whether it is random or there is some structure embedded in it.

Our claim is that one can project the structure down to a lower dimensional space.

However, Mihail and Papadmitriou has shown in their 2002 paper "On the Eigenvalue Power Law" that for power law graphs, the eigenvalues are associated with high degree vertices, not the structure in the graph.

Does this imply that the eigenvalue method fails to discover the structure of power law graphs? We have shown that we could still discover the structure of power law graphs by properly normalizing the adjacency matrix entries.

When we write the matrix as

$$\tilde{M} = M + N$$

where the M matrix is the signal (structure) and the N matrix is the noise, we want to discover the signal. But the main difficulty is that the eigenvalues are dominated by the maximum variance elements. If the method is stopped by the maximum variance elements, then our idea is why not multiply the low variance elements by constants until their variance equals that of the max variance elements.

In the power law graphs or any graphs with a given degree distribution, the edge between vertex i and j exists with probability proportional to $d_i d_j$, where d_i is the expected degree of a vertex i. When we look at the adjacency matrix A of such a graph, each entry has different probability to be 1, thus each entry has different variance. To rescale the variance of entries in A uniformly, what constants should we multiply to each element?

The following simple example gives us a hint on how to find such constants. If a random variable y has the following distribution,

$$y = \begin{cases} 1 & \text{with probability } cp \\ 0 & \text{with probability } 1-cp \end{cases}$$

its variance $\sigma^2(y) = cp(1-cp) \simeq cp$ for small values of cp.

On the other hand, a variable z that has value c with probability p

$$z = \begin{cases} c & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

has its variance $\sigma^2(z) \simeq c^2 p$. The variance of y and z are different. But, if the value of z is rescaled as

$$z = \begin{cases} \sqrt{c} & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

then, $\sigma^2(z) \simeq cp$, and y and z have asymptotically same variance.

The (i,j)'th element a_{ij} of a power law graph adjacency matrix A has value 1 with probability proportional to $d_i d_j$. Thus, a_{ij}'s variance is also proportional to $d_i d_j$. But, if we were to scale the value of a_{ij} by $\frac{1}{\sqrt{d_i d_j}}$, the variance of each element in A becomes uniform. Then, the eigenvalues are no longer dominated by the maximum variance elements, and we could discover the structure. The following transformation of A produces the desired rescaling.

$$\begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \cdots & \\ & & & d_n \end{pmatrix}^{-\frac{1}{2}} A \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \cdots & \\ & & & d_n \end{pmatrix}^{-\frac{1}{2}}$$