CS 683 Lecture 1
Date: 01/21/08
Scribe: Seth Marvel

Preliminaries:

- Course mechanics: no book, no final, potentially a midterm. Grade is based primarily on a portfolio containing worked homework problems.
- Motivation for course: academic CS is shifting focus from traditional CS topics (e.g. compilers, programming languages, operating systems) to topics on the scale of the Internet (e.g. random graphs of various types, graph theoretic algorithms). Students that master the content of this course will be positioned well to lead CS into this new area of study.
- This course will serve as an introduction to CS for the next 30 years, making use of probability and statistics rather than discrete math and formal logic.

Course topics may include:

(1) Random graphs
(2) Large, high dimensional data sets
    (. . . which have fundamentally different properties than low dimensional data sets. E.g., although points randomly selected in a compact planar region are *not* evenly spaced out, points selected by the same method in a high-dimensional compact space are.)
(3) Dimension reduction
(4) Spectral techniques
(5) Sketches, shingles
(6) Vapnik-Chervonenkis dimension
(7) Collaborative filtering

Lecture

$G(n,p)$: a graph of n vertices in which each edge is present with an independent probability p. E.g., $G(n,1/2)$ has expected degree $n/2$, and $G(n,d/n)$ has expected degree d.

The probability of a single configuration in which a given vertex v will bond to k of the n–1 remaining vertices has probability $p^k(1–p)^{n–k–1}$ and there are n–1 choose k such configurations, so

$$\Pr[\deg(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-k-1} \tag{1}$$

In the limit $n \to \infty$, (1) goes to the Poisson distribution if we hold $d = np$ constant. For finite d and k, we have the asymptotic relation:

$$\binom{n-1}{k} = \frac{(n-1)(n-2)\cdots(n-k)}{k!} \sim \frac{n^k}{k!} \tag{2}$$

Likewise, for $G(n,d/n)$,

$$\Pr[\deg(v)=k]=\binom{n-1}{k}\left(\frac{d}{n}\right)^k\left(1-\frac{d}{n}\right)^{n-k-1}\sim\frac{n^k}{k!}\frac{d^k}{n^k}\left(1-\frac{d}{n}\right)^{n-k-1} \quad (3)$$

However,

$$\lim_{n\to\infty}\left(1-d/n\right)^{n-k-1}=e^{-d} \quad (4)$$

So,

$$\lim_{n\to\infty}\Pr[\deg(v)=k]=\frac{d^k}{k!}e^{-d} \quad \text{(Poisson, as claimed)} \quad (5)$$

However, many random graphs in the real world have power-law degree distributions rather than Poisson distributions, indicating that these graphs are *not* formed in the way described above. Rather, as we will see, the power-law distribution arises as a consequence of adding both edges and vertices in time.

Currently, there are about 100 billion static webpages on the Internet. Suppose we apply a page ranking algorithm to x billion of these pages and retain 1 billion of the ranked results. How large should x be? If x is too small, important pages may not be ranked. If x is too large, utilities may be unnecessarily high (e.g. Google's electric bill last year was $200,000,000).

Even if it is unlikely that a given vertex of $G(n,1/n)$ is highly ranked, it may be likely that *one* of the n vertices in $G(n,1/n)$ is highly ranked, for n sufficiently large. In particular, there is likely a vertex of degree $k = \log n / \log\log n$. To see this, first note

$$\log k^k = k\log k=\frac{\log n}{\log\log n}[\log\log n-\log\log\log n]\sim\log n \quad (6)$$

since the second term in brackets is far less than the first for large n. So, $k^k \sim n$. Since $k! < k^k$,

$$\Pr[\deg(v)=k]>e^{-1}/n \quad (7)$$

Then,

$$\Pr[\exists v:\deg(v)=k]=1-(1-\Pr[\deg(v)=k])^n>1-(1-e^{-1}/n)^n=1-e^{-1/e}\approx 0.31 \quad (8)$$

Caveats: For *extremely* large n, the logarithm base is indeed arbitrary, but for moderately large n, say $n = 10^5$, different bases can give very different k. Similarly, log log log n is not infinitesimal relative to log log n until n is extremely large.

Suggested homework: What is $(1 - f(n)/n)^n$?

Other interesting questions:
(1) How many triangles are present in $G(n,d/n)$?
(2) How should we characterize the largest connected component of a random graph (think of the common bowtie or butterfly picture of the Internet with small in/out wings and a large "strongly connected component" in the center)?