NOTES
#6 - #8

B. MISHRA

FEBRUARY 29, 2012

Power and Randomness

From the 14th century to the 16th century, Florence dominated Europe as one of the most important cities – economically, politically, scientifically and culturally. The Medici Bank of Florence was one of the earliest financial institutions created by the Medici family in Italy during the 15th century. Machiavelli, a son of Florence, wrote his famous book "The Prince" there, which he thought described the Medici's goals; the book was dedicated to the Medicis, though not particularly favored by them. Galileo, under the patronage of the Medicis, discovered the moons of Jupiter (which he called Cosimo Sidera) and named these moons after Cosimo de' Medici's four children (Io, Europa, Ganymede and Callisto). The European Renaissance (Rinascimento in Italian, meaning rebirth) began in Tuscany (Central Italy), and was centered in the cities of Florence and Siena. Cosimo de' Medici played an important role in this renaissance through his support for education, establishing the Platonic Academy for the study of ancient works and spending more than 600,000 gold florins in support of architecture, scholarly learning, and other arts.

It is has been puzzling why the Medicis emerged as the most influential family in 15th century Florence. Even more interestingly, it is not immediately obvious why Cosimo de' Medici was ultimately able to form the most politically powerful and economically prosperous family in Florence, dominating Mediterranean trade. The Medicis started from an humble origin, coming from an agricultural region and were less powerful than most other important families in Florence, both politically and economically.

One explanation for their rise to power is provided in a paper by Padgett and Ansell (1993) "Robust Action and the Rise of the Medici," where the authors argue that Medicis' power derived from their situation in the social network of Florence. They suggested a new measure of power to take into account the "location" of the family with the network is the "betweenness" measure, as defined below.

Betweenness: Let $P(i,j)$ be the number of shortest paths (geodesics) connecting vertex i to vertex j. Let $P_k(i,j)$ be the number of shortest paths (geodesics) connecting the same two vertices i and j that include vertex k. The measure of betweenness (for

a network with n nodes) is then defined as

$$B_k \equiv \sum_{(i,j)\in E, i\neq j, k\notin\{i,j\}} \frac{P_k(i,j)/P(i,j)}{\binom{n-1}{2}},$$

(with the convention that $P_k(i,j)/P(i,j) = 0$ if $P(i,j) = 0$ (i.e., i and j are two distinct connected components).

Intuitively, this measure gives, for each pair of families, the fraction of the shortest paths that go through family k, suitably normalized.

It turns out that

$$B_{\text{Medici}} = 0.522, \quad \text{and} \quad B_k \leq 0.255, \forall k \neq \text{Medici}.$$

So the Medicis (in particular Cosimo) may be argued to have derived their unprecedented power simply by playing a central role in the social network of influential families in Florence.

The arguments above seem to indicate that the "bridge" persons, who connect two different cohesive societies and can operate in both the social networks relatively easily, may accrues a significant advantage.

Properties of Networks

So far, we have focused our attention on the graph structures induced by a social network, and various statistical properties (e.g., summary statistics or quantitative performance measures). Thus we can now compare two networks and make guess about which one is more effective. Furthermore, we can also tell what position in the network-topology is more advantageous. A critical component of designing a good social network would be those mechanisms (e.g., recommendations, discovery processes, ability to perform preferential attachments, etc.) that induce better topologies and allow a "motivated" individual to situate himself advantageously.

A short list of desirable properties:

1. Degree distributions (hubbiness) and Densities
2. Clustering (Cliques, Clans and Clubs)
3. Diameter (degrees-of-separation) and average path length
4. Centrality (Betweenness)

For instance, we may focus on just the degree distribution, $P(d)$, of a network, which quantifies relative frequencies of nodes that have different degrees $0 \leq d < n$. For instance, given a graph, we may describe its $P(d)$ by a histogram, i.e., $P(d)$ is just the fraction of nodes with degree d. For a random graph model (to be described shortly), $P(d)$ is a probability distribution. We will focus on two types of degree distributions:

1. $P(d) \leq ce^{-\lambda d}$, for some $\lambda > 0$ and $c > 0$: The tail of the distribution falls off faster than an exponential (thus large degrees are unlikely).
2. $P(d) = cd^{-\alpha}$, for some $\alpha > 0$ and $c > 0$: *Power-law distribution*: The tail of the distribution is fat, (thus there will be many more nodes with very large degrees).

Power-laws appear in a wide variety of settings including networks describing incomes, city populations, WWW, and the Internet – it is also known as a scale-free distribution: a distribution that is unchanged (within a multiplicative factor) under a rescaling of the variable. Power laws are often interpreted in a log-log plot, where it appears linear:

$$\ln \text{freq}(d) = k - \alpha \ln d.$$

Random Graphs

There are two ways of describing random graphs, and are closely related variants of the Erdös-Rényi (ER) random graphs.

$G(n, M)$ **Model:** In the $G(n, M)$ model, a graph $G = (V, E)$ is chosen uniformly at random from the collection of all graphs, which have $|V| = n$ nodes and $|E| = M$ edges. For example, in the $G(3, 2)$ model, there are exactly three possible graphs on three vertices and two edges, each assigned a probability $\frac{1}{3}$.

$G(n, p)$ **Model:** In the $G(n, p)$ model, a graph $G = (V, E)$, is constructed by connecting every pair of nodes uniformly randomly. For every pair of vertices $u, v \in V$, an edge $(u, v) \in E$ is included in the graph with probability p independent from every other edge.

Equivalently, all graphs with n nodes and M edges have equal probability of

$$p = \frac{M}{\binom{n}{2}}.$$

The parameter p in this model is exactly the density of the graph; as p increases from 0 to 1, the model produces denser

graphs with higher likelihood than sparser graphs. Thus, at $p = \frac{1}{2}$, all graphs on n vertices are chosen with equal probability (no bias for any density).

Random graphs are often studied in the asymptotic case, as $|V| = n$, the number of vertices, tends to infinity.

The *expected number of edges* in $G(n, p)$ is

$$\langle |E| \rangle = \binom{n}{2} p,$$

and the *expected degree* is

$$\bar{d} = \langle d \rangle = \text{density}(n-1) = (n-1)p.$$

The degree of a vertex in a graph $G \in G(n, p)$ is distributed as a Binomial: $d(v) \sim \text{Bin}(n-1, p)$.

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

A Poisson approximation (for large n and np = const) is given as $d(v) \sim \text{Poisson}(np)$

$$Pr[d(v) = k] \to \frac{(np)^k \exp[-np]}{k!}, \quad n \to \infty, np = \text{const.}$$

In their original pair, Erdös and Rényi showed how "connectedness" properties of $G(n, p)$ can change dramatically sharply as p crosses certain precise threshold values. For instance:

Small p: If $p < \frac{(1-\epsilon)\ln n}{n}$, then a graph in $G(n, p)$ will almost surely contain isolated vertices, and thus disconnected.

Large p: If $p > \frac{(1+\epsilon)\ln n}{n}$, then a graph in $G(n, p)$ will almost surely be connected.

§§§

Next, we will briefly talk about the so-called *0-1 Laws*, which describes a phenomena where an event either occurs or does not occur – almost surely. These are also seen as *tipping points* or *phase transitions*, as with a small increase in a critical parameter, the event of interest very quickly goes from probability 0 (almost never) to probability 1 (almost sure).

Imagine sending a friend-request randomly to n other individuals in the network. We make the assumption (rather ideal) that if the recipient is already a friend, he

simply ignores the request, but otherwise (he received your message for the first time), he accepts you as a friend – never ignores, declines or unfriends you.

It turns out that after $\Theta(n \ln n)$ requests, one will have a.s. (almost surely) befriended all the n individuals. This bound is rather sharp, in the sense that with fewer (by a small constant) requests one would have missed someone; with more (also, by a small constant) requests one would have wasted requests.

Of course, if everyone in the network does this with only $\Theta(n^2 \ln n)$ the social network would be complete (K_n) achieving the maximum density of 1.

To understand this phenomenon, we will look at a classical problem, called "*Coupon Collectors' Problem*" – which is related to the following "Collect All Coupons and Win" contest.

Coupon Collector's Problem:

***Problem Statement*:** Suppose there are n coupons, from which coupons ae being collected *with replacement*.

What is the probability that more than t sample trials are needed to collect *all* coupons.

More precisely, *given n coupons, how many coupons are* expected *to be drawn with replacement, before each coupon has been drawn at least once.*

An Example: If $n = 52$, a sharp bound occurs at $t = 225$. That is, if you draw a card randomly (with replacement) from a deck of cards, after 225 draws you would have seen every card at least one *almost surely*.

In general $t = \Theta(n \ln n)$. WHY?

Let

$t_i =$ time to collect ith coupon after collecting $(i-1)$th coupon.

$$\begin{aligned} T &= \sum_{i=1}^{n} t_i = \text{time to collect all coupons} \\ & \quad t_i\text{'s are independent} \\ E(T) &= \sum_{i=1}^{n} E(t_i) \\ Var(T) &= \sum_{i=1}^{n} Var(t_i) \end{aligned}$$

Note that the probability of collecting a new coupon after the $(i-1)$th is:

$$p_i = \frac{n-i+1}{n},$$

and thus $t_i \sim$ Geometric$(\frac{1}{p_i})$. That is,

$$Pr[t_i = k] = (1-p_i)^k p_i :$$

First k attempts fail to get the "new" coupon with the $(k+1)$th succeeding.

$$\begin{aligned} E(t_i) &= \frac{1}{p_i} = \frac{n}{n-i+1} \\ Var(t_i) &= \frac{1-p_i}{p_i^2} = \frac{(i-1)n}{(n-i+1)^2} \end{aligned}$$

Thus

$$\begin{aligned} E(T) &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = nH_n \\ &= n\ln n + \gamma n + \frac{1}{2} + o(n) \\ &\qquad \gamma = 0.577 = \text{Euler's Const.} \end{aligned}$$

$$\begin{aligned} Var(T) &\leq \frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + \cdots + \frac{n^2}{1} \leq \frac{\pi^2}{6} n^2 \\ \sigma(T) &\leq \frac{\pi n}{\sqrt{6}}. \end{aligned}$$

By Chebyshev Inequality:

$$Pr[|T - \mu_T| \geq k\sigma_T] \leq \frac{1}{k^2}.$$

Thus

$$\begin{aligned} & Pr[|T - nH_n| \geq c \cdot n] \\ &\leq Pr[|T - nH_n| \geq \frac{c\sqrt{6}}{\pi} \cdot \frac{\pi n}{\sqrt{6}}] \\ &\leq \frac{\pi^2}{6c^2}. \end{aligned}$$

Thus we have the fopllowing 0-1 Law:

- If the number attempts so far, $T < (1-\epsilon)nH_n$, then you are almost surely missing some coupons.

- If the number attempts so far, $T > (1+\epsilon)nH_n$, then you almost surely have all the coupons.

The following is a nice generalization: T_k = First time k copies of each coupon has been collected. T_k also has a 0-1 Law (a phase transition) at its expected value:

$$T_k = n\ln n + (k-1)n\ln\ln n + O(n), \quad \text{as } n \to \infty.$$

Threshold Function for Conectivity: Erdös-Rényi 1961

Theorem: *A threshold funstion for the connectivity of Erdös-Renyi model $G(n,p)$ occurs at*

$$p(n) = \frac{\ln n}{n}.$$

That is for a graph $G(n, \lambda\frac{\ln n}{n})$,

- *if $\lambda < 1$, $Pr(\text{connectivity}) = 0$;*
- *if $\lambda > 1$, $Pr(\text{connectivity}) = 1$.*

Proof: Consider the following indicator random variable $I_i \sim$ Bernoulli(π)

$$I_i = \begin{cases} 1, & \text{if node } i \text{ is isolated;} \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$\begin{aligned} \pi &= Pr[I_i = 1] \\ &= (1-p)^n = e^{-pn} = e^{-\lambda \ln n} = n^{-\lambda}. \end{aligned}$$

The total number of isolated nodes is thus

$$X = \sum I_i, \text{ where } I_i = \text{Bernoulli}(n^{-\lambda}).$$

We claim that

$$E(X) \approx Var(X) = n \cdot n^{-\lambda}.$$

But note that since

$$Var(X) \geq (0 - E(X))^2 Pr[X = 0],$$

we have

$$Pr[X = 0] \leq E(X)^{-1} = n^{\lambda - 1}.$$

If $\lambda < 1$, then

$$\lim_{n\to\infty} Pr[X = 0] \to 0,$$

and $\exists_i I_i \neq 0$, and the graph is a.s. disconnected.

Note however, when $\lambda > 1$, we cannot conclude from the fact that $Pr[X = 0] \neq 0$ that the graph is in fact connected. TOO BAD.

Going back to our claim:

$$E(X) = \sum_{i=1}^{n} E(I_i) = n\pi = n \cdot n^{-\lambda}.$$

But since I_i's are not indepemdent,

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{n} Var(I_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} Cov(I_i, I_j) \\
&= nVar(I_1) + 2\binom{n}{2} Cov(I_1, I_2) \\
&= n[E(I_1^2) - E(I_1)^2] + n(n-1)[E(I_1 I_2) - E(I_1)E(I_2)] \\
&= n[\pi - \pi^2] + n(n-1)[(1-p)^{2n-3} - \pi^2] \\
&= n\pi(1-\pi) + n(n-1)\left[\frac{\pi^2}{1-p} - \pi^2\right] \\
&\approx n\pi + n^2\pi^2 p \\
&\approx n \cdot n^{-\lambda} = E(X).
\end{aligned}$$

Graph is Disconnected: We will describe this property as follows:

$$\exists_{V' \subset V, |V'| \leq n/2} CUT[V', V \setminus V'];$$

that is, the graph has been "cut" with no edge going from V' to its complement $V \setminus V'$. Let $|V'| = k$, then

$$\begin{aligned}
&Pr[\text{Graph Disconnected}] \\
&\quad\equiv\quad Pr[\exists_{k \leq n/2} \exists_{V', |V'|=k} CUT[V', V \setminus V']] \\
&\quad\leq\quad \sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)}
\end{aligned}$$

A tedious calculaion reveals that

$$Pr[\text{Graph Disconnected}] \to 0, \quad \lim n \to \infty,$$

when $p = \frac{\lambda \ln n}{n}$, $\lambda > 1$. $\square$