CSC 5450 Randomness and Computation

Week 14: Social Networks

Plan ① Power law, preferential attachment

② Small world phenomenon, decentralized search

③ Epidemics, branching process

④ Quick review of the course

Schedule

HW 3: due May 14, will be posted by Apr 26 (Fri)

Project: due May 21, please send a softcopy to chi@cse.cuhk.edu.hk.

Power Law [1, chapter 18]

In internet, it is observed that the fraction of web pages that have k in-links is approximately proportional to $1/k^2$.

This kind of power laws seem to dominate when the quantity being measured is a type of popularity.

For example, the fraction of telephone numbers that receive k calls per day is roughly proportional to $1/k^2$, the fraction of books that are bought by k people is roughly proportional to $1/k^3$, the fraction of scientific papers that receive k citations in total is roughly proportional to $1/k^3$, etc.

The normal distribution is very different from a power law distribution. In a normal distribution, the probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c.

By the central limit theorem, any quantity that can be viewed as the sum of many small independent random effects will be well approximated by the normal distribution.

To understand why power laws are so widespread in "real networks", it would be useful to have a simple model that generates these distributions, to give insights about the underlying processes.

Clearly, any model based on independent random variables is not going to work.

It turns out that there is a simple and natural model to generate power law distributions.

Rich-Get-Richer Models

In the following model, we assume that people have a tendency to copy the decisions of people who act before them.

① Pages are created in order, and named $1, \ldots, N$.

② When a new page j is created

(a) with probability p, page j creates a link to a uniform random page from $1, \ldots, j-1$.

(b) with probability $1-p$, page j chooses a page i uniformly at random from among all earlier pages and creates a link to the page that i points to.

(c) one can repeat this process to create multiple independent links from page j.

Observe that the process is "rich-get-richer", as step 2(b) is equivalent to:

2(b): with probability $1-p$, page j chooses a page l with probability proportional to l's current number of in-links, and creates a link to l.

This phenomenon is also called <u>preferential attachment</u>, as pages that are more popular are more preferred. This model also explains some phenomenon without human decision making, e.g. the fraction of cities with population K, the number of copies of a gene in a genome, etc.

Analysis [1, Chapter 18.7]

Let $X_j(t)$ be the number of in-links to a node j at a time step $t \geq j$.

- Initially, $X_j(j) = 0$ for all j.
- At time $t+1$, there are t nodes. The probability that node $t+1$ links to node j is $\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$.

To simplify the analysis, we consider a deterministic approximation of this random process.

We approximate $X_j(t)$ by a continuous function of time $x_j(t)$.

- Initially, $x_j(j) = 0$ for all j.
- When node $j+1$ arrives, the number of in-links to node j increases with probability $\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$.

  We model this by the differential equation $\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$.

Solving the Differential Equation

Let $q = 1-p$. Then $\frac{dx_j}{dt} = \frac{p+qx_j}{t} \Rightarrow \frac{1}{p+qx_j}\frac{dx_j}{dt} = \frac{1}{t}$.

Integrating both sides $\int \frac{1}{p+qx_j}\frac{dx_j}{dt}\,dt = \int \frac{1}{t}\,dt \Rightarrow \ln(p+qx_j) = q\ln t + c$ for a constant c.

Exponentiating, we get $p+qx_j = e^c \cdot t^q$, and thus $x_j(t) = \frac{1}{q}(e^c t^q - p)$.

Using the initial condition that $x_j(j)=0$, this implies that $0 = e^c j^q - p$, and so $e^c = p/j^q$.

Therefore, $x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$.

What is the fraction of all nodes having at least k in-links at time t?

A node j satisfies $x_j(t) \geq k$ iff $x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k \Leftrightarrow j \leq t\left[\frac{q}{p}\cdot k + 1\right]^{-\frac{1}{q}}$.

Since there are t nodes a time t, the fraction is just $\left[\frac{q}{p}\cdot k+1\right]^{-\frac{1}{q}}$.

Finally, to compute the fraction of nodes with exactly k in-links, we can differentiate the above expression and get the answer as $\frac{1}{q}\cdot\frac{q}{p}\cdot\left[\frac{q}{p}k+1\right]^{-1-\frac{1}{q}} = \frac{1}{p}\left[\frac{q}{p}k+1\right]^{-1-\frac{1}{q}} = \Theta(k^{-1-\frac{1}{q}})$.

So, the exponent in the power law is $1+\frac{1}{q} = 1+\frac{1}{1-p} \geq 2$.

See [2] for a more formal analysis.

Small World Phenomenon [1, chapter 20]

In an experiment, random people in a town are asked to send letters to random people. They are given the target's name, occupation, address, and some other information, and they need to forward the letters through people whom they know. The result of the experiment is that roughly a third of the letters eventually arrived, in a median of six steps, and this is known as "six degree of separation".

The experiment seems to suggest that ① there are many short paths in the network, and

② there is a decentralized procedure to find such a path.

We are interested in a simple probabilistic model that has these two properties.

One possible model is to have a 2-dimensional grid, where each node is a person and people close-by know each other (say two nodes have an edge if they are of distance d in the grid).

Furthermore, each node is connected to a random node in the network.

A graph generated by this simple model is an expander graph with high probability, and so there

is a short path between two vertices.

While this provides a simple model for the first property, it does not provide an answer to the second property, that a short path can be found in a decentralized manner, since the random edges are too random.

We consider a variant of the above simple model, where there is some structure on the random edges.

Again, we have a 2-dimensional grid, where nodes close-by are connected to each other.

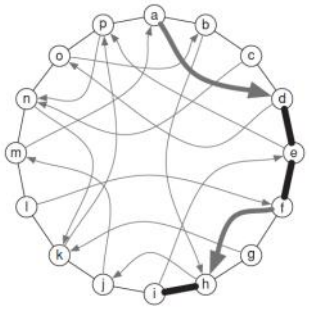
Each node v has a random edge vw, where the probability of having vw is proportional to $\frac{1}{d(v,w)^q}$.

When $q=0$, this is just the above model with completely random edges, and it is not easy to find a short path in a decentralized manner. When q is large, most edges are short and there won't be short path between far-away nodes.

It turns out that the optimal exponent is when $q=2$, in which case one can find a path of length $O(\text{polylog}(n))$ in a decentralized manner, where n is the number of nodes in the grid.

Analysis [1, chapter 20.7]

We do an analysis in a one-dimensional grid, and leave the 2-dimesional case in homework.



picture from [1]:

a random network and

a decentralized short path.

Searching: The decentralized searching algorithm is simple: Just send to the neighbor that is closer to the target.

Let X be the number of steps in this path.

We say that a node is in phase j of the path if its distance from the target is between 2^j and 2^{j+1}. Let X_j be the number of steps in phase j.

Then $E[X] = \sum_{i=1}^{\log n} E[X_i]$.

We will show that $E[X_i] = O(\log n)$, and thus $E[X] = O(\log^2 n)$.

We said that the probability of having an edge between v and w is proportional to $\frac{1}{d(v,w)}$.

First, we calculate this probability precisely.

First, we calculate this probability precisely.

In a cycle of n nodes, $\sum_{w \in V} \frac{1}{d(v,w)} \leq 2\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n/2}\right) \leq 2\log_2 n$.

So, the probability of having the edge vw is at least $\frac{1}{2\log n} \cdot \frac{1}{d(v,w)}$.

Now, suppose v is in phase j, and its distance to the target is d.

Note that phase j will end if v jumps to a node with distance at most d/2 from the target.

Let S be the set of nodes with distance at most d/2 from the target. So, $|S| = d+1$.

Each node in S is of distance at most $3d/2$ from v.

So, the probability that the random neighbor of v is in S is at least $d \cdot \frac{1}{2\log n} \cdot \frac{1}{3d/2} = \frac{1}{3\log n}$

Hence, $E[X_j] = 3\log n$, as it is just a geometric variable with $p = \frac{1}{3\log n}$.

Therefore, $E[X] \leq 3\log^2 n$.

Epidemics [1, chapter 21]

It is of interest to study how ideas or diseases spread in networks. While spreading ideas in networks may relate more to human decision making, the process of spreading (computer) virus is really more like a random process.

The simplest model of contagion is called the <u>branching process</u>.

Consider the setting where there is a k-regular infinite rooted tree.

Initially, the root is infected. Then, it will infect its neighbor with probability p, independently.

And then the virus will spread to the lower levels and so on, until all nodes in one level are not infected, and then the disease dies out.

Let $R = p \cdot k$ be the reproductive number of this process.

We will prove the following basic result.

<u>Theorem</u> If $R < 1$, then with probability one, the disease dies out after a finite number of steps.

If $R > 1$, then with positive probability, the disease will survive forever.

Before we see the proof, let's see another model of epidemic model: the <u>SIR epidemic model.</u>

In this model, we are given a directed graph.

Each node is in one of the three states:

Susceptible : Before the node has caught the disease, it is suspectible to infection from its neighbors.

Infectious : Once the node is infected, it will stay infected for a fixed number of steps t_I, and in each step it will pass the disease to each of its suspectible neighbor with probability p.

Removed : After a node has been infected, it will recover and never be infected again.

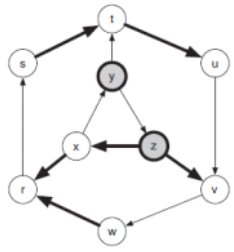
One can also considered a more general model, when one node is recovered, it will be immured to the process for a period t_R, and after that it will be suspectible again.

But this can be reduced to the SIR model, by considering the "time expanded" graph.

One can also view the SIR process as a static process, by flipping the coin of each edge ahead. That is, for each edge, with probability p it stays in the graph, and it is removed from the graph with probability $1-p$.

Then, whether a node will be infected is equivalent to whether there is a directed path from some infected node at the beginning to this node.

This is closely related to percolation theory, a well-studied area in probability theory.



picture from [1].

Analysis of Branching Process [1, chapter 21.8]

The number of node in level n is K^n.

Let X_n be the number of infected nodes at level n.

For the j-th node in level n, let Y_{nj} be the indicator variable whether j is infected.

Then $E[X_n] = \sum_{j=1}^{K^n} E[Y_{n,j}] = \sum_{j=1}^{K^n} p^n = (Kp)^n = R^n$.

When $R<1$, $E[X_n] \to 0$ as $n \to \infty$.

In fact, by Markov's inequality, $Pr(X_n \geq 1) \leq E[X_n] = R^n$.

This shows that the disease dies out in finite steps with probability one

The other direction

The other direction

Let q_n be the probability that the virus survives in the n-th level.

When $R>1$, then $E[X_n] \to \infty$ as $n \to \infty$, but it does not imply that $q_n > 0$.

To prove $q_n > 0$, we write a recursive formula for q_n.

The probability that the virus survives through one particular neighbor of the root is $p \cdot q_{n-1}$.

So, it fails to survives in all branches with probability $(1-p\cdot q_{n-1})^k$, and hence

$1-q_n = (1-p\cdot q_{n-1})^k$, and it follows that $q_n = 1-(1-p\cdot q_{n-1})^k$.

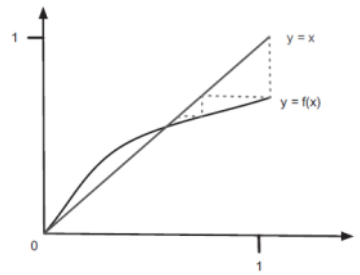
Let $f(x) = 1-(1-px)^k$. Then $q_n = f(q_{n-1})$. And we want to study whether the sequence

$1, f(1), f(f(1)), f(f(f(1))), \ldots$ will converge to a positive value.

Here are some properties of the function f.

- $f(0)=0$, $f(1) = 1-(1-p)^k < 1$.
- $f'(x) = kp(1-px)^{k-1}$, and it is strictly positive and monotonically decreasing when $x \in [0,1]$.
  This means that f is increasing and concave.
- $f'(0) = kp = R$. So, if $R>1$, then the curve goes above the line $y=x$ initially.

So, the plot of f looks like the following figure (from [1]):



And the sequence will converge to a positive value as shown in the figure.

References

[1] Easley, Kleinberg. Networks, crowds, and markets.

Review

Week 1-3: basic, expectation, Markov, Chebyshev, Chernoff, applications

week 4-5: probabilistic methods: first moment, second moment, local lemma, random graphs.

week 6-7: random walk, stationary distribution, hitting time, mixing time, Markov chain, coupling.
. (Lovasz-Simonovits curve, Cheeger's inequality)

week 8-10 : algebraic techniques, polynomials, hashing, k-wise independence, eigenvalues.

week 11-14 : topics: graph, approximation, complexity, social networks.

We've have many settings where randomness are (or seem to be) indispensable.

- distributed computing : routing, network coding.
- parallel computing : matching.
- probabilistic method : Ramsey graphs, algorithms from local lemma
- sublinear algorithms : streaming, local graph partitioning, PCP.
- counting and sampling : Markov chain

There are some nice topics and techniques that we did not have time to discuss, e.g.,

online algorithms, martingales, derandomization, etc.

After this course, I hope you can :

- understand and feel comfortable with basic concepts and techniques from probability theory, and in particular can read research papers using these ideas, and appreciate and enjoy them.
- have some intuition when randomness is useful and can do some basic probabilistic reasoning.
- have the background and interest to learn more advanced topics and techniques.
- apply and create some ideas in your research area. please let me know if you do. hope the course project is your first step.

THE END. THANK YOU (especially those who attended all the lectures :)