MS&E 233 Lecture 9 & 10: Network models

Ashish Goel, scribed by Riley Matthews

April 30, May 2

- Review: Application of PR to Netflix movie suggestions
  - Client x:

$$\text{sim}(i) = \begin{cases} (1-\epsilon) \sum_{m:\, i \text{ likes } m} \left(\frac{\text{rel}(m)}{\#\text{ of users who liked m}} \right) & \text{if } x \neq i \\ \epsilon + (1-\epsilon) \sum_{m:\, i \text{ likes } m} \left(\frac{\text{rel}(m)}{\#\text{ of users who liked m}} \right) & \text{if } x = i \end{cases}$$

    where

$$\text{rel}(m) = \sum_{i:\, i \text{ likes } m} \left(\frac{\text{sim}(i)}{\#\text{ of moives i liked}} \right)$$

- Power-Iteration
  - $\text{sim}_0 = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{if } x \neq i \end{cases}$
  - Let the matrix M be the encoding of the user movie preferences, with $M_{i,j} = \begin{cases} 1 & \text{if user j likes movie i} \\ 0 & \text{otherwise} \end{cases}$
  - Let the "forward matrix", F_M, be the matrix in which the entries of M are normalized by their column sum, i.e.,

$$\left(F\_M\right)_{i,j} = \frac{M_{i,j}}{\sum_k M_{k,j}}$$

  - Similarly let the "reverse matrix", R_M, be the matrix in which the entires of M are normalized by their respective row sum, i.e.,

$$\left(R\_M\right)_{i,j} = \frac{M_{i,j}}{\sum_k M_{i,k}}$$

  - Notice that $\text{rel}_r = (F\_M)\,\text{sim}_r$

- Initialize: $\text{sim}_0(i) = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{if } x \neq i \end{cases}$ (a vector)
- Iteration: $\text{sim}_{r+1} = \epsilon\,\text{sim}_0 + (1-\epsilon)(R\_M^{\mathrm{T}}\text{rel}_r)$
- With large dataset, the method often gives excellent results
- **Recap:** We have now seen a basic method for ranking search results, personalizing search, ad targeting, and recommendation systems.

- **Virality**
  - Examples of companies that grew virally: Dropbox, Hotmail, Facebook, Myspace, Youtube, Piazza
  - Definition of viral growth: In the natural course of using the product, the user "infects" others (i.e., others start using the product). Viral growth IS NOT when users go out and actively promote the product, e.g., referral programs.
  - History
    - YouTube: YouTube let you embed videos (videos that can be displayed on *other* websites) which offered immediate value to the user. One person can get value from YouTube regardless of whether others use it or not. And, by embedding videos (normal use of the product) the user is likely to "infect" others.
    - Hotmail: Hotmail made it was easy to set up an email account where you could send emails to and receive emails from anybody who had an email address. At the end of every email sent from a Hotmail account, there was a signature along the lines of "sent from my Hotmail account"
  - Story so far: Key components of viral growth are
    - product must deliver value to first user
    - organic growth
  - What about, Myspace and Facebook? They are not useful to the first user.
    - Facebook: Started in a close-knit community
    - Myspace: Started in active community of musicians who wanted to share music with eachother
    - Piazza: Has class as captive audience
  - So most companies with initial viral growth have followed one of two paths: (1) value is delivered to first user and users infect others in the course of normal use of the service (2) they have started with small close-knit communities.

- Quantitative analysis of virality: we need to understand models of networks and growth

- Consider two networks A and B in which there are $N_A = 10^4$ and $N_B = 10^8$ individuals, respectively.
- Let's say that the probability that any two individuals are friends is p.
- Question: If we want to infect an entire network, how many friends, on average, should people have? (Np)
  - What is the probability that a given individual has no friends?

$$(1-p)^{N-1} \approx (1-p)^N$$

  - What is the expected number of individuals with no friends?

$$N(1-p)^{N-1} \approx N(1-p)^N$$

END OF LECTURE 9

- Erdos-Renyi model
  - $G_{N,p}$, an undirected network of N nodes where any pair of nodes (i, j) are connected with probability p.
  - What is the probability that a given node i has no friends? $= (1-p)^{N-1} \approx (1-p)^N$
    - $\Rightarrow$ Expected number of nodes with no friends $\approx N(1-p)^N$
  - Goal: We would like $N(1-p)^N < 1$, i.e., we would like the expected number of users with no friends to be < 1:

$$\begin{aligned} N(1-p)^N &< 1 \\ \Rightarrow (1-p)^N &< N^{-1} \\ \Rightarrow N\log(1-p) &< -\log N \\ \Rightarrow -Np &< -\log N \\ \Rightarrow p &> \frac{\log N}{N} \end{aligned}$$

    where the second-to-last step uses the approximation: $\log(1-p) \approx -p$ when p is small.
  - So if we are at this threshold, the expected degree of a node (number of edges incident to that node) is $> \log N$. (Since the expected degree is Np)
  - Moreover, it turns out that this $p > \frac{\log N}{n}$ threshold is also the point at which the graph is almost certainly connected.
  - Despite the simple assumptions of this model, the "threshold phenomena" it predicts is a durable insight.
  - Example:

    - Network A with 10^4 individuals: Each individual must have $\log(10^4) \approx 10$ friends on average for the network to be connected
    - Network B with 10^8 individuals: Each individual must have $\log(10^8) \approx 19$ friends on avergage for the network to be connected
  - Diameter of a network
    - In network A: how many friends (at the threshold) does each individual have? $= 10$. How many friends-of-friends? $\approx 10^2$. Firends-of-friends-of-friends? $\approx 10^3$.
    - The diameter, Δ, is the number of such layers or degrees of (friend) separation needed to cover every node in the Network:

$$(pN)^{\Delta} \approx N \Rightarrow \Delta \log(pN) \approx \log N \Rightarrow \Delta \approx \frac{\log N}{\log(pN)} \approx \log N$$

  - Main takeaway: tipping points/thresholds are a durable insight about network connectivity
  - If each user infects one more user, the resulting process will not cover every node, but will still cover a large population.
  - Shortcomings of the E-R model
    - In the real world, if two people are friends, their respective sets of friends are very likely to overlap (the process of acquiring friends is not random)
    - The dynamics of loners, socialites etc are absent from this model

- Preferential Attachment
  - At time t:
    - New person arrives
    - Makes k friends: selects each of these k new friends with probability proportional to their current # of friends – "the rich get richer"
    - We are assuming undirected networks, though similar ideas can be applied to directed networks
  - Analysis is similar to earlier discussion of heavy tailed phenomena:
    - $d_i(t)$ = Number of friends that the i^{th} person has at time t.
    - $d_i(i) = k$. (i^{th} person arrived at time i, has only made the initial k friends)
    - Total # of friendship edges at times t $\approx kt$

    Now, using the same approximation as we did for heavy-tails, we get:

$$\frac{\mathrm{d}}{\mathrm{d}t} d_i(t) = \frac{k d_i(t)}{2kt} \Rightarrow \frac{\mathrm{d}d_i(t)}{d_i(t)} = \frac{\mathrm{d}t}{2t}$$

$$\Rightarrow \ln d_i(t) = \frac{1}{2} \ln t + C$$

Given the initial condition $d_i(i) = k$, we find

$$\ln k = \frac{1}{2}\ln i + C \Rightarrow C = \ln k - \frac{1}{2}\ln i$$

yielding

$$\begin{aligned} \ln d_i(t) - \ln k &= \frac{1}{2}\ln t - \frac{1}{2}\ln i \\ \Rightarrow \ln \frac{d_i(t)}{k} &= \ln \sqrt{t/i} \\ \Rightarrow d_i(t) &= k\frac{\sqrt{t}}{\sqrt{i}} \end{aligned}$$

- Again, this model is not perfect and there are phenomena that it does not capture, but the insight that preferential attachment results in a heavy-tail of degrees is durable. Thus, real-life networks (the web, twitter, the email graph, etc) often have heavy-tailed degrees. In particular, there are nodes with disproportionately large "influence".
- How to find influencers?
  - Fix $t = N$, and therefore $d_i = k\frac{\sqrt{N}}{\sqrt{i}}$
  - Experiment 1: Pick a random user u
    - Expected$[d(u)] = 2k$
    - (Result of in-class experiment: a few random twitter users all have < 10 followers)
  - Experiment 2: Pick a random endpint f of a random edge:

$$\begin{aligned} \text{Expected}[d(f)] &= \sum_i \big(d_i \cdot \text{Prob}(i = f)\big) \\ &\approx \sum_i \big(k\frac{\sqrt{N}}{\sqrt{i}} \cdot \frac{d_i}{2kN}\big) \\ &= \sum_i \big(k\frac{\sqrt{N}}{\sqrt{i}} \cdot k\frac{\sqrt{N}}{\sqrt{i}}\frac{1}{2kN}\big) \\ &= \frac{k}{2}\sum_i \frac{1}{i} \approx \frac{k}{2}\ln N \end{aligned}$$

  - For large N, the latter number, $(k/2)\ln n$, is much larger than $2k$ (they are both approximations, so we shouldn't use them for small N).
    - (Result of in-class experiment: a twitter user randomly chosen from a random user's "following" list has >> 10 followers
  - Insight: This phenomenon is called the "Inspection paradox": random user is not likely to be influential, but a random friend of a random user if likely to be more influential