Preferential Attachment and Network Evolution

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
Task: (HW3 is optional)
- Find node correspondences between two graphs

Incentives:
- European chocolates!
- Fame!
- Up to 10% extra credit

Due:
- Monday Nov 14
- No late days!
Random network
(Erdos-Renyi random graph)
Degree distribution is Binomial

Scale-free (power-law) network
Degree distribution is Power-law
We will analyze the following model:

- Nodes arrive in order $1, 2, 3, \ldots, n$
- When node $i$ is created it makes a single link to an earlier node $i$ chosen:
  - 1) With prob. $p$, $i$ links to $j$ chosen uniformly at random (from among all earlier nodes)
  - 2) With prob. $1-p$, node $i$ chooses node $j$ uniformly at random and links to a node $j$ points to.

[Mitzenmacher, ‘03]
**Claim:** The described model generates networks where the fraction of nodes with degree $k$ scales as:

$$P(d_i = k) \propto k^{-(1 + \frac{1}{q})}$$

where $q = 1 - p$

**Consider deterministic and continuous approximation to the in-degree of node $i$ as a function of time $t$**

- $t$ is the number of nodes that have arrived so far
- In-degree $d_i(t)$ of node $i$ ($i=1,2,\ldots,n$) is a continuous quantity and it grows deterministically with time $t$
Initial condition:
- \( d_i(t) = 0 \), when \( t = i \) (node \( i \) just arrived)

Expected change of \( d_i(t) \) over time:
- Node \( i \) gains an in-link at step \( t+1 \) only if a link from a newly created node \( t+1 \) points to it.

What’s the probability of this event?
- With prob. \( p \) node \( t+1 \) links randomly:
  - Links to our node \( i \) with prob. \( 1/t \)
- With prob. \( 1-p \) node \( t+1 \) links preferentially:
  - Links to our node \( i \) with prob. \( d_i(t)/t \)

So: Prob. node \( t+1 \) links to \( i \) is:
\[
p \frac{1}{t} + (1 - p) \frac{d_i(t)}{t}
\]
What is the rate of growth of $d_i$?

- Expected change of $d_i(t)$ over time

\[
\frac{dd_i(t)}{dt} = p \frac{1}{t} + (1 - p) \frac{d_i(t)}{t} = \frac{p+q d_i(t)}{t}
\]

\[
\frac{1}{p+q d_i(t)} \frac{dd_i(t)}{dt} = \frac{1}{t} dt
\]

\[
\int \frac{1}{p+q d_i(t)} dd_i(t) = \int \frac{1}{t} dt
\]

\[
\frac{1}{q} \ln(p + q d_i(t)) = \ln t + c
\]

\[
q d_i(t) + p = A t^q \Rightarrow d_i(t) = \frac{1}{q} (A t^q - p)
\]
What is the constant A?

- We know: \( d_i(i) = 0 \)

- So: \( d_i(i) = \frac{1}{q} (A_i^q - p) = 0 \)

- \( \Rightarrow A = \frac{p}{iq} \)

- \( \Rightarrow d_i(t) = \frac{p}{q} \left( \left( \frac{t}{i} \right)^q - 1 \right) \)
What is $F(d)$ the fraction of nodes that has degree at least $d$ at time $t$?

- How many nodes $i$ have degree $> t$?
  
  \[ d_i(t) = \frac{p}{q} \left( \left( \frac{t}{i} \right)^q - 1 \right) > d \]
  
  - then: $i < t \left( \frac{q}{p} d - 1 \right)^{-\frac{1}{q}}$

- There are $t$ nodes total at time $t$ so $F(d)$:
  
  \[ F(d) = \left[ \frac{q}{p} d + 1 \right]^{-\frac{1}{q}} \]
What is the fraction of nodes with degree exactly $d$?

Take derivative of $F(d)$:

- $F(d)$ is CDF, so $F'(d)$ is the PDF

$$F'(d) = \frac{1}{p} \left[ \frac{q}{p} d + 1 \right]^{-1 - \frac{1}{q}} \quad \Rightarrow \quad \alpha = 1 + \frac{1}{q}$$
Two changes from the $G_{np}$
- Groth + Preferential attachment
- Do we need both? Yes!
- Add growth to $G_{np}$ (assume 1 edge is added at each step)
  - $X_j = \text{degree of node } j \text{ at the end}$
  - $X_j(u) = 1$ if node $u$ links to $j$, else 0
  - $X_j = X_j(j + 1) + X_j(j + 2) + \cdots + X_j(n)$
  - $E[X_j(u)] = P[u \text{ links to } j] = \frac{1}{u-1}$
  - $E[X_j] = \sum_{u=j}^{n} \frac{1}{u-1} = H_{n-1} - H_j$
  - $E[X_j] = \log(n-1) - \log(j) = \log \left( \frac{n-1}{j} \right) \neq \left( \frac{n}{j} \right)^{\alpha}$
Preferential attachment gives power-law degrees

Intuitively reasonable process

Can tune $p$ to get the observed exponent

- On the web, $P[\text{node has degree } d] \sim d^{-2.1}$
- $2.1 = 1 + 1/(1-p) \Rightarrow p \sim 0.1$

There are also other network formation mechanisms that generate scale-free networks:

- Random surfer model
- Forest Fire model
PA-like Link Formation

- **Copying mechanism** (directed network)
  - select a node and an edge of this node
  - attach to the endpoint of this edge

- **Walking on a network** (directed network)
  - the new node connects to a node, then to every first, second, ... neighbor of this node

- **Attaching to edges**
  - select an edge
  - attach to both endpoints of this edge

- **Node duplication**
  - duplicate a node with all its edges
  - randomly prune edges of new node
Preferential attachment is not so good at predicting network structure

- Age-degree correlation
- Links among high degree nodes
  - On the web nodes sometime avoid linking to each other

Further questions:

- What is a reasonable probabilistic model for how people sample through web-pages and link to them?
  - Short+Random walks
  - Effect of search engines – reaching pages based on number of links to them
Network resilience

- How does the connectivity of the network change as the vertices get removed? [Albert et al. 00; Palmer et al. 01]

- **Vertices can be removed:**
  - Uniformly at random
  - In order of decreasing degree

- It is important for **epidemiology**
  - Removal of vertices corresponds to vaccination
Real-world networks are resilient to random attacks

- You need to remove all web-pages of degree > 5 to disconnect the web
- But this is a very small fraction of all web pages

Random network has better resilience to targeted attacks
Evolution of Social Networks
Preferential attachment is a model of a growing network

What governs network growth and evolution?

- P1) Node arrival process:
  - When nodes enter the network

- P2) Edge initiation process:
  - Each node decides when to initiate an edge

- P3) Edge destination process:
  - The node determines destination of the edge
Let’s Look at the Data

- **4 online social networks with exact edge arrival sequence**
  - For every edge \((u, v)\) we know exact time of the appearance \(t_{uv}\)

- **Directly observe mechanisms leading to global network properties**

<table>
<thead>
<tr>
<th>Network</th>
<th>(T)</th>
<th>(N)</th>
<th>(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flickr</strong> (03/2003–09/2005)</td>
<td>621</td>
<td>584,207</td>
<td>3,554,130</td>
</tr>
<tr>
<td><strong>Delicious</strong> (05/2006–02/2007)</td>
<td>292</td>
<td>203,234</td>
<td>430,707</td>
</tr>
<tr>
<td><strong>Answers</strong> (03/2007–06/2007)</td>
<td>121</td>
<td>598,314</td>
<td>1,834,217</td>
</tr>
<tr>
<td><strong>LinkedIn</strong> (05/2003–10/2006)</td>
<td>1294</td>
<td>7,550,955</td>
<td>30,682,028</td>
</tr>
</tbody>
</table>

[Leskovec et al., KDD ’08]

and so on for millions…
P1) When are New Nodes Arriving?

Flickr: Exponential

\[ N(t) \approx e^{0.25t} \]

Delicious: Linear

\[ N(t) = 16t^2 + 3 \times 10^3 t + 4 \times 10^4 \]

Answers: Sub-linear

\[ N(t) = -284t^2 + 4 \times 10^4 t - 2.5 \times 10^3 \]

LinkedIn: Quadratic

\[ N(t) = 3900t^3 + 7600t - 1.3 \times 10^5 \]
How long do nodes live?

Node life-time is the time between the 1st and the last edge of a node.

How do nodes “wake up” to create links?

Edge creation events
P2) What is Node Lifetime?

Node lifetime is exponentially distributed:

\[ p_l(\alpha) = \lambda e^{-\lambda \alpha} \]

- **Lifetime \( \alpha \):** time between node’s first and last edge
How do nodes “wake up” to create edges?

**Edge gap** $\delta_i(d)$: time between $d^{th}$ and $(d+1)^{st}$ edge of node $i$:
- Let $t_i(d)$ be the creation time of $d$-th edge of node $i$
- $\delta_i(d) = t_i(d + 1) - t_i(d)$

$\delta(d)$ is a distribution (histogram) of $\delta_i(d)$ over all nodes $i$.
P2) When Do Nodes Create Edges?

Edge gap $\delta(d)$: inter-arrival time between $d^{th}$ and $d+1^{st}$ edge

For every $d$ we get a separate histogram

$$p_g(\delta(1)) \propto \delta(1)^{-\alpha} e^{-\beta}$$
How do $\alpha$ and $\beta$ change as a function of $d$?

Fit to each plot of $\delta(d)$: $p_g(\delta(d)) \propto \delta(d)^{-\alpha(d)} e^{-\beta(d)}$
P2) Evolution of Edge Gaps

- \( \alpha \) is const, \( \beta \) linear in \( d \) – gaps get smaller with \( d \)

\[
p_g(\delta(d)) \propto \delta(d)^{-\alpha} e^{-\beta \cdot d}
\]
P3) How to Select Destination?

- Source node $i$ wakes up and creates an edge
- How does $i$ select a target node $j$?
  - What is the degree of the target $j$?
    - Do preferential attachment really hold?
  - How many hops away is the target $j$?
    - Are edges attaching locally?
Are edges more likely to connect to higher degree nodes?

\[ p_e(k) \propto k^\tau \]

<table>
<thead>
<tr>
<th>Network</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gnp</td>
<td>0</td>
</tr>
<tr>
<td>PA</td>
<td>1</td>
</tr>
<tr>
<td>Flickr</td>
<td>1</td>
</tr>
<tr>
<td>Delicious</td>
<td>1</td>
</tr>
<tr>
<td>Answers</td>
<td>0.9</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>0.6</td>
</tr>
</tbody>
</table>
How “far” is the Target Node?

- Just before the edge \((u,w)\) is placed how many hops are between \(u\) and \(w\)?

![Graph showing edge probability against hops]

<table>
<thead>
<tr>
<th>Network</th>
<th>% Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr</td>
<td>66%</td>
</tr>
<tr>
<td>Delicious</td>
<td>28%</td>
</tr>
<tr>
<td>Answers</td>
<td>23%</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>50%</td>
</tr>
</tbody>
</table>

Real edges are local! Most of them close triangles!
Focus only on triad-closing edges

New triad-closing edge \((u,w)\) appears next

Model this as 2 independent choices:

1. \(u\) chooses neighbor \(v\)
2. \(v\) chooses neighbor \(w\) and connect \(u\) to \(w\)

E.g.: Under Random-Random:

\[
p(u, w) = \frac{1}{5} \cdot \frac{1}{2} + \frac{1}{5} \cdot 1 = \frac{3}{10}
\]

Under a particular pair of “strategies”: Likelihood of the graph = \(\prod_{(u,w) \in E} p(u, w)\)
**Triad Closing Strategies**

- **Improvement over the baseline:**
  - Baseline: Pick a random node 2 hops away

- **Strategy to select v (1st node):**

<table>
<thead>
<tr>
<th>Flickr</th>
<th>Strategy</th>
<th>random</th>
<th>deg^{0.2}</th>
<th>com</th>
<th>last^{-0.4}</th>
<th>comlast^{-0.4}</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>random</td>
<td>13.6</td>
<td>13.9</td>
<td>14.3</td>
<td>16.1</td>
<td>15.7</td>
</tr>
<tr>
<td>deg^{0.1}</td>
<td>deg</td>
<td>13.5</td>
<td>14.2</td>
<td>13.7</td>
<td>16.0</td>
<td>15.6</td>
</tr>
<tr>
<td>last^{0.2}</td>
<td>last</td>
<td>14.7</td>
<td>15.6</td>
<td>15.0</td>
<td>17.2</td>
<td>16.9</td>
</tr>
<tr>
<td>com</td>
<td>com</td>
<td>11.2</td>
<td>11.6</td>
<td>11.9</td>
<td>13.9</td>
<td>13.4</td>
</tr>
<tr>
<td>comlast^{0.1}</td>
<td>comlast</td>
<td>11.0</td>
<td>11.4</td>
<td>11.7</td>
<td>13.6</td>
<td>13.2</td>
</tr>
</tbody>
</table>

- **Strategies to pick a neighbor:**
  - random: uniformly at random
  - deg: proportional to its degree
  - com: prop. to the number of common friends
  - last: prop. to time since last activity
  - comlast: prop. to com*last
### Summary of the Model

**The model of network evolution**

<table>
<thead>
<tr>
<th>Process</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P1) Node arrival</strong></td>
<td>• Node arrival function is given</td>
</tr>
<tr>
<td><strong>P2) Edge initiation</strong></td>
<td>• Node lifetime is exponential</td>
</tr>
<tr>
<td></td>
<td>• Edge gaps get smaller as the degree increases</td>
</tr>
<tr>
<td><strong>P3) Edge destination</strong></td>
<td>Pick edge destination using random-random</td>
</tr>
</tbody>
</table>
Analysis of the Model

- **Theorem**: Exponential node lifetimes and power-law with exponential cutoff edge gaps lead to power-law degree distributions.

- Interesting as temporal behavior predicts structural network property.

[Leskovec et al., KDD '08]
Proof sketch

- Node lifetime: $p_l(a) =$
- Node of life-time $a$, what is its final degree $D$?

- What is distribution of $D$ as a func. of $\lambda, \alpha, \beta$?

- The 2 exp. funcs. “cancel”. Power-law survives
Evolving the Networks

- Given the model one can take an existing network continue its evolution

- Compare true and predicted (based on the theorem) degree exponent:

<table>
<thead>
<tr>
<th></th>
<th>Flickr</th>
<th>Delicious</th>
<th>Answers</th>
<th>LinkedIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>0.0092</td>
<td>0.0052</td>
<td>0.019</td>
<td>0.0018</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.84</td>
<td>0.92</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.0020</td>
<td>0.00032</td>
<td>0.0038</td>
<td>0.00036</td>
</tr>
<tr>
<td>true</td>
<td>1.73</td>
<td>2.38</td>
<td>1.90</td>
<td>2.11</td>
</tr>
<tr>
<td>predicted</td>
<td>1.74</td>
<td>2.30</td>
<td>1.75</td>
<td>2.08</td>
</tr>
</tbody>
</table>
How do networks evolve at the macro level?

What are global phenomena of network growth?

Questions:

What is the relation between the number of nodes $n(t)$ and number of edges $e(t)$ over time $t$?

How does diameter change as the network grows?

How does degree distribution evolve as the network grows?
Network Evolution

- $N(t)$ ... nodes at time $t$
- $E(t)$ ... edges at time $t$
- Suppose that
  \[ N(t+1) = 2 \times N(t) \]
- Q: what is
  \[ E(t+1) = 2 \times E(t) \]
- A: over-doubled!
  - But obeying the Densification Power Law
Q1) Network Evolution

- What is the relation between the number of nodes and the edges over time?
- First guess: constant average degree over time
- Networks are denser over time
- Densification Power Law:

\[ E(t) \propto N(t)^a \]

\( a \) ... densification exponent \((1 \leq a \leq 2)\)
Densification Power Law

- the number of edges grows faster than the number of nodes – average degree is increasing

$E(t) \propto N(t)^a$

or equivalently

$$\frac{\log(E(t))}{\log(N(t))} = \text{const}$$

- densification exponent: $1 \leq a \leq 2$:
  - $a=1$: linear growth – constant out-degree (traditionally assumed)
  - $a=2$: quadratic growth – clique

[Leskovec et al. KDD 05]
Prior models and intuition say that the network diameter slowly grows (like $\log N$, $\log \log N$)

- Diameter shrinks over time
  - as the network grows the distances between the nodes slowly decrease
Diameter of a Densifying $G_{np}$

Densifying random graph has increasing diameter $\Rightarrow$ There is more to shrinking diameter than just densification.

Is shrinking diameter just a consequence of densification?

[Diamantov et al. TKDD 07]
Is it the degree sequence?

Compare diameter of a:

- True network (red)
- Random network with the same degree distribution (blue)

Densification + degree sequence give shrinking diameter
How does degree distribution evolve to allow for densification?

**Option 1)** Degree exponent $\gamma_n$ is constant:

- **Fact 1:** For degree exponent $1 < \gamma < 2$: $a = \frac{2}{\gamma}$

A consequence of what we learned in last class:

- Power-laws with exponents $<2$ have infinite expectations.
- So, to maintain constant degree exponent $\gamma$ distribution with the average degree needs to grow.
How does degree distribution evolve to allow for densification?

Option 2) Exponent $\gamma_n$ evolves with graph size $n$:

- Fact 2:
  \[
  \gamma_n = \frac{4n^{a-1} - 1}{2n^{a-1} - 1}
  \]

Remember, expected degree is:

\[
E[x] = \frac{\gamma - 1}{\gamma - 2} x_{\text{min}}
\]

So $\gamma$ has to decay as a function of graph size for the avg. degree to go up.
Want to model graphs that density and have shrinking diameters

Intuition:
- How do we meet friends at a party?
- How do we identify references when writing papers?
The Forest Fire model has 2 parameters:
- $p$ ... forward burning probability
- $r$ ... backward burning probability

The model:
- Each turn a new node $v$ arrives
- Uniformly at random chooses an “ambassador” $w$
- Flip 2 geometric coins to determine the number of in- and out-links of $w$ to follow
- “Fire” spreads recursively until it dies
- New node $v$ links to all burned nodes
Forest Fire Model

- Forest Fire generates graphs that **densify** and have **shrinking diameter**

\[
E(t) = 10^{1.32} N(t)
\]

\[
\text{diameter} = 5.2 \times 10^{-1} t^{1.32}, \quad R^2 = 1.00
\]
Forest Fire Model

- Forest Fire also generates graphs with power-law degree distribution

log count vs. log in-degree  
log count vs. log out-degree

11/2/2011
Fix backward probability $r$ and vary forward burning prob. $p$

Notice a sharp transition between sparse and clique-like graphs

Sweet spot is very narrow