Preferential Attachment and Network Evolution
Exponential vs. Power-Law Tails

Model: $G_{np}$
**Model: Preferential attachment**

- **Preferential attachment**
  [Price ‘65, Albert-Barabasi ’99, Mitzenmacher ‘03]
  - Nodes arrive in order \(1,2,\ldots,n\)
  - At step \(j\), let \(d_i\) be the degree of node \(i < j\)
  - A new node \(j\) arrives and creates \(m\) out-links
  - Prob. of \(j\) linking to a previous node \(i\) is proportional to degree \(d_i\) of node \(i\)

\[
P(j \rightarrow i) = \frac{d_i}{\sum_k d_k}
\]
New nodes are more likely to link to nodes that already have high degree

Herbert Simon’s result:
- Power-laws arise from “Rich get richer” (cumulative advantage)

Examples [Price ‘65]
- Citations: New citations to a paper are proportional to the number it already has
The Exact Model

We will analyze the following model:

- Nodes arrive in order $1, 2, 3, \ldots, n$
- When node $j$ is created it makes a single out-link to an earlier node $i$ chosen:
  - 1) With prob. $p$, $j$ links to $i$ chosen uniformly at random (from among all earlier nodes)
  - 2) With prob. $1 - p$, node $j$ chooses $i$ uniformly at random and links to node $l$ that $i$ points to
    - This is same as saying: With prob. $1 - p$, node $j$ links to node $l$ with prob. proportional to $d_l$ (the in-degree of $l$)
- Our graph is directed: Every node has out-degree 1

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

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

where $q=1-p$.

So we get power-law degree distribution with exponent:

$$\alpha = 1 + \frac{1}{1-p}$$
Consider deterministic and continuous approximation to the 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 $i$ ($i = 1, 2, ..., n$) is a continuous quantity and it grows deterministically as a function of time $t$.

Plan: Analyze $d_i(t)$ – continuous in-degree of node $i$ at time $t > i$. 
Continuous Degree: What We Know

- **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. \( d_i(t)/t \)

- **Prob. node** \( t + 1 \) **links to** \( i \) **is:**
  \[ p \frac{1}{t} + (1 - p) \frac{d_i(t)}{t} \]
At $t = 4$ node $i = 4$ comes. It has out-degree of 1 to deterministically share with other nodes:

<table>
<thead>
<tr>
<th>Node $i$</th>
<th>$d_i(t)$</th>
<th>$d_i(t+1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$=0 + p \frac{1}{4} + (1 - p) \frac{0}{4}$</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>$=2 + p \frac{1}{4} + (1 - p) \frac{2}{4}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$=0 + p \frac{1}{4} + (1 - p) \frac{1}{4}$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>$=1 + p \frac{1}{4} + (1 - p) \frac{1}{4}$</td>
</tr>
<tr>
<td>4</td>
<td>/</td>
<td>0</td>
</tr>
</tbody>
</table>

$d_i(t) - d_i(t - 1) = \frac{dd_i(t)}{dt} = p \frac{1}{t} + (1 - p) \frac{d_i(t)}{t}$

How does $d_i(t)$ evolve as $t \to \infty$?
What is the rate of growth of $d_i$?

- **Expected change of $d_i(t)$:**
  
  $d_i(t + 1) - d_i(t) = p \frac{1}{t} + (1 - p) \frac{d_i(t)}{t}$

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

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

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

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

  $p + q d_i(t) = e^c t^q \Rightarrow d_i(t) = \frac{1}{q} (At^q - p)$
What is the value of constant A?

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

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

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

- And so \(\Rightarrow d_i(t) = \frac{p}{q} \left( \frac{(t)^q}{i^q} - 1 \right)\)

**Observation:** Old nodes (small \(i\) values) have higher in-degrees \(d_i(t)\)
What is $F(k)$ the fraction of nodes that has degree less than $k$ at time $t$?

- How many nodes have degree $< k$?
  
  $d_i(t) = \frac{p}{q} \left( \left( \frac{t}{i} \right)^q - 1 \right) < k$

- Solve for $i$ and obtain: $i < t \left( \frac{q}{p} k + 1 \right)^{-\frac{1}{q}}$

There are $t$ nodes total at time $t$ so the fraction $F(k)$ is:

$$F(k) = \left[ \frac{q}{p} k + 1 \right]^{-\frac{1}{q}}$$
What is the fraction of nodes with degree exactly $k$?

- Take derivative of $F(k)$:
  - $F(k)$ is CDF, so $F'(k)$ is the PDF!

\[
F'(k) = \frac{1}{p} \left[ \frac{q}{p} k + 1 \right]^{-1 - \frac{1}{q}} \Rightarrow \alpha = 1 + \frac{1}{1 - p}
\]

q.e.d.
Preferential attachment gives power-law degrees!

Intuitively reasonable process

Can tune $p$ to get the observed exponent

- On the web, $P[\text{node has degree } d] \sim d^{-2.1}$
- $2.1 = 1 + 1/(1-p)$ $\Rightarrow p \sim 0.1$
Preferential attachment is not so good at predicting network structure

- **Age-degree correlation**
  - **Solution:** Node fitness (virtual degree)

- **Links among high degree nodes:**
  - On the web nodes sometime avoid linking to each other

**Further questions:**

- What is a reasonable model for how people sample through network node and link to them?
  - Short random walks
Many models lead to Power-Laws

- **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 and attach to both endpoints of this edge

- **Node duplication**
  - Duplicate a node with all its edges
  - Randomly prune edges of new node
Two changes from the $G_{np}$

- The network grows
- Preferential attachment

Do we need both? Yes!

Add growth to $G_{np}$ (assume $p = 1$):

- $X_j = \text{degree of node } j \text{ at the end}$
- $X_j(u) = 1$ if $u$ links to $j$, else 0
- $X_j = X_j(j + 1) + X_j(j + 2) + \cdots + X_j(n)$
- $E[X_j(u)] = P[u \text{ links to } j] = 1/(u - 1)$
- $E[X_j] = \sum_{j=1}^{n} \frac{1}{u-1} = \frac{1}{j} + \frac{1}{j+1} + \cdots + \frac{1}{n-1} = H_{n-1} - H_j$
- $E[X_j] = \log(n - 1) - \log(j) = \log((n - 1)/j)$ \textbf{NOT} $\binom{n}{j}^\alpha$

Extra!

$H_n$...$n^{th}$ harmonic number:

$H_n = \sum_{k=1}^{n} \frac{1}{k} \approx \log(n)$
## Distances in Preferential Attachment

<table>
<thead>
<tr>
<th>Degree Exponent</th>
<th>Small_world</th>
<th>Ultra_small_world</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\alpha = 3$</td>
<td>$\frac{\log \log n}{\log (\alpha - 1)}$ $2 &lt; \alpha &lt; 3$</td>
</tr>
<tr>
<td>Avg. path length</td>
<td>$\log n$</td>
<td>$\log \log n$</td>
</tr>
<tr>
<td>Degree exponent</td>
<td>$\alpha &gt; 3$</td>
<td></td>
</tr>
</tbody>
</table>

Size of the biggest hub is of order $O(N)$. Most nodes can be connected within two steps, thus the average path length will be independent of the network size.

The average path length increases slower than logarithmically. In $G_{np}$ all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network vast majority of the path go through the few high degree hubs, reducing the distances between nodes.

Some models produce $\alpha = 3$. This was first derived by Bollobas et al. for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.
Summary: Scale-Free Networks

- \( \alpha = 1 \)
  - Second moment \( \langle k^2 \rangle \) diverges

- \( \alpha = 2 \)
  - Average \( \langle k \rangle \) diverges
  - Ultra small world behavior

- \( \alpha = 3 \)
  - \( \langle k^2 \rangle \) finite
  - \( \langle k \rangle \) finite
  - Small world
  - The scale-free behavior is relevant
  - Behaves like a random network

Extra!
Consequence of Power-Law Degrees
Consequence: Network Resilience

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

- Nodes can be removed:
  - Random failure:
    - Remove nodes uniformly at random
  - Targeted attack:
    - Remove nodes in order of decreasing degree

- This is important for robustness of the internet as well as epidemiology
Real networks are resilient to random failures

$G_{np}$ has better resilience to targeted attacks

- Need to remove all pages of degree $>5$ to disconnect the Web
- But this is a very small fraction of all web pages
Lessons Learned

- There is no universal degree exponent characterizing all networks
- We need growth and the preferential attachment for the emergence of scale-free property
  - The mechanism is domain dependent
    - Many processes give rise to scale-free networks
- Modeling real networks:
  - Identify microscopic processes that occur in the network
  - Measure their frequency from real data
  - Develop dynamical models that capture these processes
  - If the model is correct, it should predict the observations
Evolution of Social Networks
Network Evolution: Observation

- Preferential attachment is a model of a growing network
- Can we find a more realistic model?
- What governs network growth & 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

[Leskovec, Backstrom, Kumar, Tomkins, 2008]
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 creation $t_{uv}$

- Directly observe mechanisms leading to global network properties

<table>
<thead>
<tr>
<th>Network</th>
<th>Start Date</th>
<th>End Date</th>
<th>$T$</th>
<th>$N$</th>
<th>$E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F) Flickr</td>
<td>03/2003–09/2005</td>
<td></td>
<td>621</td>
<td>584,207</td>
<td>3,554,130</td>
</tr>
<tr>
<td>(D) Delicious</td>
<td>05/2006–02/2007</td>
<td></td>
<td>292</td>
<td>203,234</td>
<td>430,707</td>
</tr>
<tr>
<td>(A) Answers</td>
<td>03/2007–06/2007</td>
<td></td>
<td>121</td>
<td>598,314</td>
<td>1,834,217</td>
</tr>
<tr>
<td>(L) LinkedIn</td>
<td>05/2003–10/2006</td>
<td></td>
<td>1294</td>
<td>7,550,955</td>
<td>30,682,028</td>
</tr>
</tbody>
</table>
P1) When are New Nodes Arriving?

- **Flickr:** Exponential
  - Equation: $N(t) \approx e^{0.25t}$
  - Graph shows nodes increasing exponentially over time.

- **Delicious:** Linear
  - Equation: $N(t) = 16t^2 + 3e3t + 4e4$
  - Graph shows nodes increasing linearly over time.

- **Answers:** Sub-linear
  - Equation: $N(t) = -284t^2 + 4e4t - 2.5e3$
  - Graph shows nodes increasing sub-linearly over time.

- **LinkedIn:** Quadratic
  - Equation: $N(t) = 3900t^2 + 7600t - 1.3e5$
  - Graph shows nodes increasing quadratically over time.
P2) When Do Nodes Create Edges?

- **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?**
P2) What is Node Lifetime?

LinkedIn

- **Lifetime $a$:** Time between node’s first and last edge

Node lifetime is exponentially distributed:

$$p_l(a) = \lambda e^{-\lambda a}$$
How do nodes “wake up” to create edges?

**Edge gap** $\delta_d(i)$: time between $d^{th}$ and $d + 1^{st}$ edge of node $i$:

- Let $t_d(i)$ be the creation time of $d$-th edge of node $i$
- $\delta_d(i) = t_{d+1}(i) - t_d(i)$

- $\delta_d$ is a distribution (histogram) of $\delta_d(i)$ 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 is distributed by a power-law with exponential cut-off

For every $d$ we make a separate histogram

$p_g(\delta_1) \propto \delta_1^{-\alpha} e^{-\beta}$
How do \( \alpha \) and \( \beta \) change as a function of \( d \)?

To each plot of \( \delta_d \) fit:

\[
p_g(\delta_d) \propto \delta_d^{-\alpha_d} e^{-\beta_d}
\]

\( \alpha \) is constant!

\( \beta \) linearly increases!
P2) Evolution of Edge Gaps

- $\alpha$ const., $\beta$ linear in $d$. What does this mean?
- 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 $j$?
    - Does 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? YES!

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

<table>
<thead>
<tr>
<th>Network</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_{np}</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>
Just before the edge \((u,w)\) is placed how many hops are between \(u\) and \(w\)?

<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!
How to Close the Triangles?

- Focus only on triad-closing edges
- New triad-closing edge \((u,w)\) appears next
- 2 step walk model:
  - \(u\) is about to create an edge
  1. \(u\) chooses neighbor \(v\)
  2. \(v\) chooses neighbor \(w\)
    and \(u\) connects to \(w\)
- One can use different strategies for choosing \(v\) and \(w\): Random-Random works well. Why?
  - More common friends (more paths) helps
  - High-degree nodes are more likely to be hit

[Leskovec et al., KDD ’08]
# Triad Closing Strategies

- **Improvement in log-likelihood over baseline:**
  - **Baseline:** Pick a random node 2 hops away

### Strategies to select $v$ (1st node)

<table>
<thead>
<tr>
<th>Flickr</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>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>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>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>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>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

Extra!
## 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>