Lecture 9: Scale + Evolution
(bonus)

How does time modify large networks?
(i.e., how will FB look in 10 years?)

COMS 4995-1: Introduction to Social Networks
Tuesday, October 4th
Practice: Make your own power law

What would it take to modify today’s apple policy
... and obtain a power law?
1. A lot of apple
2. A reinforcement rule in apple’s distribution
Assume infinite amount of students
- Each one asks a question (random uniform)
- A student with a question receives 2 apples
  * One for her (red), One to distribute (green)
  * Send the green apple to a “friend” chosen randomly
  * With proba $p$, this friend keeps it
  * With proba $1-p$ this friend sends to the person who already received a green apple from her.
Example of time evolution
Outline

* The ubiquitous power-law

* The three sources of power law
  o reinforcement
  o optimization
  o artefacts of random stopping

* A bit more on graph evolution
A generalization of our previous construction

- States that the increase/decrease of an entity should be made proportional to its size.

Discrete version:

- Let $X(t)$ be a variable (e.g. your income) discretized $Y(t) = j$ if and only if your income $X(t)$ is in $[m \cdot e^i; m \cdot e^{i+1}]$.
- Proportional evolution: $P[Y(t+1) = j \mid Y(t) = i] = f(j-i)$.
- E.g.: $f(1) = p; f(-1) = 1-p$, then $P[X(\infty) > x] = \frac{x}{m} \ln\left(\frac{p}{1-p}\right)$.

A generalization of our previous construction
- States that the increase/decrease of an entity should be made proportional to its size.

Continuous version:
- Let $X(t)$ be a variable (e.g. your income)
- Proportional evolution: $X(t+1) = F(t) X(t)$
- For large $t$, $X(t)$ approaches lognormal distribution (i.e. $\log(X(t)) \sim$ normal dist; $X(t)$ heavy tailed, not PL)

Skewed distribution
- Can be explained by reinforcement, propor. effect.
- This also applies to degree distribution in graph, that also exhibit small-world and “hub” structure

Be careful
- Lognormal and power law may both be obtained
- Graphs obtained do not exhibit clustering coefficient
- There are other sources explaining power-laws!
Outline

- The ubiquitous power-law

- The three sources of power law
  - reinforcement
  - optimization
  - artefacts of random stopping

- A bit more on graph evolution
First, show that power law are optimal
  o Uses resource in the most efficient manner
  o Creates the optimal topology
Second, show that system are lead to optimal
  o Following Darwinian argument: highest fitness
  o Greedy behavior of peers (Game-theoretic appr.)
To convey information, you can use
- Short words frequently, longer words rarely

A measure of information:
- Average cost (in characters) \[ H = - \sum_j p_j \log_2 p_j \]
- Ideally, \( p_j \) decreasing and \( c_j = \log_d(j) \)

Thm: Choose Max(H/C) obtained when
- For any \( j \), derivative w.r.t. \( p_j \), \( H'_j/C'_j = H/C \)
- Implies \( p_j \) follows power law in \( i \)

An informational theory of the statistical structure of language.
B. Mandelbrot, Comm. Theory. (1953)
Assume that nodes join sequentially

- at a location chosen uniformly on the plane
- Choose a neighbor $j$ to connect to minimize its cost
- Cost $(i) = \alpha d(i,j) + h_j$; $d(i,j)$ eucl. dist., $h_j$ netw. dist.

**Example 2:** Internet topology

Assume peers search for items using random walk
  o Ask a node, if it does not have ask its neighbors, etc.
  o Let \( q_i \) denote the query rate of item \( i \)
* Each node keeps some item in cache
  o Let \( p_i \) denote the fraction of cache devoted to item \( i \)
* How to minimize the search time (i.e. \( \sum q_i / p_i \))?
  o Caching more popular items (blockbusters)
  o Preserving some diversity of content
Thm: *Square root* allocation of cache is optimal

- It is defined as $p_i = \sqrt{q_i} / \sum_j \sqrt{q_j}$
- If $q_i$ follows a power-law, then $p_i$ as well

Thm: This allocation can be obtained by a simple distributed mechanism (*path replication*)

- After finding the item, give a copy it to all nodes previously contacted, discarding a random items.
- Since search size $\sim 1/p_i$ new copy produced at rate $\sqrt{q_i}$
- Copy erased at rate $p_i \sim \sqrt{q_i}$

Replication strategies in unstructured peer-to-peer networks.
Outline

* The ubiquitous power-law

* The three sources of power law
  o reinforcement
  o optimization
  o artefacts of random stopping

* A bit more on graph evolution
Assume that you produce a large corpus randomly
- The space bar is hit with probability $p$
- Each other character has same probability $q = (1-p)/d$

What is the frequency of words obtained?
- All words of length $j$ has same probability $pq^j$
- There are $d^j$ such words
- Hence $P[\text{word with rank } i] = p \cdot i^\alpha$ with $\alpha = (-\ln(q)/\ln(d))$

Random corpus also exhibits power law!

Some Effects of Intermittent Silence.
Reality test #2: Internet AS graph

Internet routers degrees follow power law.

How was this measured?

Traceroute: pick a source compute shortest path to many destinations.

Reality test #2: Internet AS graph

What about running Traceroute on uniform random graphs?

After sampling degree resembles power law!

Sampling biases in IP topology measurements.
A Lakhina, JW Byers, M Crovella, P Xie. Infocom (2003)
Reality test #2: Internet AS graph

* Traceroute does not discover all edges
  o As it needs to be on the shortest paths from a destination to the source
  o Moreover, this reinforces the presence of edges near the source

* Thm: For general degree dist. random graphs,
  o the degree distribution changes radically
  o It may exhibit heavy tail/power law artificially

On the bias of traceroute sampling.
D Achlioptas, A Clauset, D Kempe, C Moore. J. ACM (2009)
Frequency of items, popularity of objects and distributions of many variables are highly skewed

- This proves that it cannot be the sum of independent effect (Central limit theory would predict Normal).
- It could be reinforcement, or the result of a bias (typically random stopping),
- Is some cases (caching, topology) it can relate to optimal efficiency
Outline

* The ubiquitous power-law

* The three sources of power law
  o reinforcement
  o optimization
  o artefacts of random stopping

* A bit more on graph evolution
Microscopic graph evolution

- Graphs evolve over time
  - Densification and shrinking distance/diameter
  - Some theoretical justifications

- What about a direct likelihood comparison?
  - Empirical data vs. predicted events at this time
  - Advantages: (1) no focus on specific property, (2) quantitative comparison between models
Graph Evolution: Densification and Shrinking Diameters

JURE LESKOVEC
Carnegie Mellon University
JON KLEINBERG
Cornell University
and
CHRISTOS FALOUTSOS
Carnegie Mellon University

How do real graphs evolve over time? What are normal growth patterns in social, technological, and information networks? Many studies have discovered patterns in static graphs, identifying properties in a single snapshot of a large network or in a very small number of snapshots; these include heavy tails for in- and out-degree distributions, communities, small-world phenomena, and others. However, given the lack of information about network evolution over long periods, it has been hard to convert these findings into statements about trends over time.

Here we study a wide range of real graphs, and we observe some surprising phenomena. First, most of these graphs densify over time with the number of edges growing superlinearly in the number of nodes. Second, the average distance between nodes often shrinks over time in contrast to the conventional wisdom that such distance parameters should increase slowly as a function of the number of nodes (like $O(\log n)$ or $O(\log(\log n))$).

Existing graph generation models do not exhibit these types of behavior even at a qualitative level. We provide a new graph generator, based on a forest fire spreading process that has a simple, intuitive justification, requires very few parameters (like the flammability of nodes), and produces graphs exhibiting the full range of properties observed both in prior work and in the present study.

This material is based on work supported by the National Science Foundation under Grants No. IIS-0209107 SENSOR-0329549 EF-0331657 IIS-0326322 IIS-0534205, CCF-0325453, IIS-0329064, CNS-0403340, CCR-0122581; a David and Lucile Packard Foundation Fellowship; and also by the Pennsylvania Infrastructure Technology Alliance (PITA), a partnership of Carnegie Mellon, Lehigh University and the Commonwealth of Pennsylvania's Department of Community and Economic Development. Additional funding was provided by a generous gift from Hewlett-Packard. J. Leskovec was partially supported by the Microsoft Research Graduate Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

Author's address: J. Leskovec, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, 15213 Pittsburgh PA, USA; email: jure@cs.cmu.edu.

Permission to make digital or hard copies part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from the Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.


ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 2, Publication date: March 2007.

Impact of degree

- Normalized proba of edge with destination degree

\[ p_e(d) = \frac{\sum_t [e_t = (u, v) \land d_{t-1}(v) = d]}{\sum_t \{u : d_{t-1}(u) = d\}}. \]

\[ e = (u, v) \] degree of \( v \) at time \( t \)
Impact of age

- Normalized proba of edge with age of source

\[ e(a) = \frac{|\{e = (u, v) : t(e) - t(u) = a\}|}{|\{t(u) : t_e - t(u) \geq a\}|} \]
Impact of distance

- Distance distribution computed before edge, by convention h=0 denotes disconnected
- Normalized Proba of edges with distance

Main findings: triangle closing edges are many!
65% Flickr, 50% LinkedIn
What about time

- Node’s active life
  Time where it adds edges exponential distribution

- Time gap between two edges
  Follows Hybrid Power-Law+Exp. tail
A triangle closing model

1. Arrival of nodes follows function $N(t)$.
   - Each one connects with an edge following degree
   - A la copying model/preferential attachment

2. Nodes adds triangle closing edges
   - According a sequence following time gaps ($\alpha, \beta$)
   - For a “lifetime” distributed exponentially ($\lambda$)
   - Each triangle closing edge chosen according to random 2-step neighbor-of-neighbor choices.
Numerical Validation

Clustering coef.
Degree distrib.
Distance distrib.
All seems to match well

(c) Geodesic distance
Additional slides:
Other papers on evolution
Graph evolution

<table>
<thead>
<tr>
<th>DATA SET: Temporal graphs from Flickr Yahoo! 360</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMPIRMICAL RESULT</td>
</tr>
<tr>
<td>1. Reciprocation is fast</td>
</tr>
<tr>
<td>2. Densification + shrinking diameter after a certain phase</td>
</tr>
<tr>
<td>3. Outside Giant Connected Comp (GCC) isolated nodes + star which grows and merge to GCC</td>
</tr>
</tbody>
</table>

| MODEL 3 types Passive-Inviters-Linkers with fixed ratio of population |
| Edge originates in L or I (chosen by deg) |
| Edge from I ends in new node P |
| Edge from L ends in L or I with a trumping factor within L |
| ANALYSIS |
| Numerical results show that it reproduces size distribution of isolated communities accurately. |

How do real graphs evolve over time? What are normal growth patterns in social, technological, and information networks? Many studies have discovered patterns in static graphs, identifying properties in a single snapshot of a large network or in a very small number of snapshots; these include heavy tails for in- and out-degree distributions, communities, small-world phenomena, and others. However, given the lack of information about network evolution over long periods, it has been hard to convert these findings into statements about trends over time.

Here we study a wide range of real graphs, and we observe some surprising phenomena. First, most of these graphs densify over time with the number of edges growing superlinearly in the number of nodes. Second, the average distance between nodes often shrinks over time in contrast to the conventional wisdom that such distance parameters should increase slowly as a function of the number of nodes (like $O(\log n)$ or $O(\log(\log n))$).

Existing graph generation models do not exhibit these types of behavior even at a qualitative level. We provide a new graph generator, based on a forest fire spreading process that has a simple, intuitive justification, requires very few parameters (like the flammability of nodes), and produces graphs exhibiting the full range of properties observed both in prior work and in the present study.

**Microscopic Graph evolution (2)**

<table>
<thead>
<tr>
<th>DATA SET: Flickr</th>
<th>EMPIRICAL RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Reciprocation is fast</td>
</tr>
<tr>
<td></td>
<td>2. Links created/received according to out-degree/in-degree</td>
</tr>
<tr>
<td></td>
<td>3. Proximity bias</td>
</tr>
</tbody>
</table>

| MODEL None | ANALYSIS None |

Microscopic Graph evolution (3)

<table>
<thead>
<tr>
<th>DATA SET: FriendFeed</th>
<th>EMPIRICAL RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Old nodes receive new edges in proportion of their degree</td>
</tr>
<tr>
<td></td>
<td>2. Proximity bias</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each node picks a degree first</td>
</tr>
<tr>
<td>Then pick the closest node with given degree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical results shows it creates proximity bias</td>
</tr>
</tbody>
</table>