CS 6604: Data Mining Large Networks and Time-series

B. Aditya Prakash

Lecture #4: The Watts-Strogatz Model and Decentralized Search
Recap: A Small World

- What is the typical shortest path length between two people?
  - Milgram’67 : 6.2 (using letters to a Boston stock-broker)
  - Dodds et. al.’03 : 7 (using email to multiple targets)

- Implications
  - On the clustering
  - On the way people route and find the target
IMPLICATIONS ON THE CLUSTERING
Small World From Exponential Growth

Assuming fan-out = 100
# people 1 step away: 100
# people 2 steps away: 100²
# people 3 steps away: 100³

your friends
friends of your friends
But, real networks are locally clustered!

Triadic Closure reduces the growth rate
How to have both? Short Paths and Local Structure

<table>
<thead>
<tr>
<th></th>
<th>Random Graphs $G(n,p)$</th>
<th>Regular Lattices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree Distribution</strong></td>
<td>Binomial(n,p)</td>
<td>Constant (co-ordination number)</td>
</tr>
<tr>
<td><strong>Clustering Co-efficient</strong></td>
<td>$\approx p$ (SMALL)</td>
<td>Constant (e.g. 6/15 for interior nodes in 2-d)</td>
</tr>
<tr>
<td><strong>Diameter (Average Shortest Path)</strong></td>
<td>$O(\log n)$</td>
<td>$\approx N^{1/D}$ (e.g. $\sqrt{N}$ for 2-d)</td>
</tr>
</tbody>
</table>

Benchmark 1: regular lattices

One-dimensional lattice:
- $\frac{2}{N}$ nodes inside for $1 \leq k \leq N_l$,
- $\frac{C}{N}$ nodes inside for $6 \leq k \leq N_l$.

The average path-length varies as:
- Constant degree (coordination number),
- Constant clustering coefficient.

\[ f(P(k)) = G(k - 4) \]
\[ f(C) = \frac{1}{2} \] for all nodes.

\[ f(\text{Path length}) = \] The average shortest path.

So, we have:
- Constant degree,
- Constant avg. clustering coeff.
- Linear avg. path-length.

Note about calculations:
- We are interested in quantities as graphs get O(DUJH1Č).
- We will use big-O: $I[\frac{2}{7}]DV[\frac{1}{2}]$ if $f(x) < g(x)c$ for all $x > x_0$ and some constant $c$. 

9/27/2012

# How to have both? Short Paths and Local Structure

<table>
<thead>
<tr>
<th></th>
<th>Random Graphs $G(n,p)$</th>
<th>Real Networks</th>
<th>Regular Lattices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree Distribution</strong></td>
<td>$\text{Binomial}(n,p)$</td>
<td>$\text{Skewed}$</td>
<td>Constant (co-ordination number)</td>
</tr>
<tr>
<td><strong>Clustering Coefficient</strong></td>
<td>$\approx p \ (\text{SMALL})$</td>
<td>$\text{Large}$</td>
<td>Constant (e.g. 6/15 for interior nodes in 2-d)</td>
</tr>
<tr>
<td><strong>Diameter (Average Shortest Path)</strong></td>
<td>$O(\log n)$</td>
<td>$\text{Small-world}$</td>
<td>$\approx N^{1/D} \ (\text{e.g. } \sqrt{N} \text{ for 2-d})$</td>
</tr>
</tbody>
</table>

---

Prakash 2013  
CS 6604: DM Large Networks & Time-Series
Watts-Strogatz Model [Watts, Strogatz’98]

K neighbors

Regular

Random

1. Start with a (low-dimensional) lattice

2. Rewire edges with probability p

\[ I = \frac{N}{2K} \]
\[ C = \frac{3(K-2)}{(K-1)} \]

\[ I \approx \log \frac{N}{\log K} \]
\[ C \approx \frac{K}{N} \]

Is there a regime with small I and large C?
Transition: little randomness can create a small world

Broad interval of $p$ over which $C(p)/C(0) \approx 1$ but $L(p)/L(0) \approx 0$
W-S Model Properties

- **Diameter** \( l(N, p) \approx \frac{N^{1/d}}{K} f(pKN) \)
  
  \[
  f(x) = \begin{cases} 
  \text{const} & \text{if } x \ll 1 \\
  \ln x / x & \text{if } x \gg 1 
  \end{cases}
  \]

  - Lattice-like
  - Random-graph-like

- **Clustering Coefficient** \( C(p) = C(0)(1 - p)^3 \)

- **Degree Distribution** \( E[d_v] = K \)
  
  (dist. \( P[k] \) depends on \( p \), but is concentrated on the expectation)
W-S Model: Summary

- A tiny amount of long-range weak ties (randomness) is enough to make a small world
- Gave insights about the interplay between clustering and short paths
- Does not lead to skewed degrees
- Does not enable navigation
### Other ‘Small World’ Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>size</th>
<th>av. shortest path</th>
<th>Shortest path in fitted random graph</th>
<th>Clustering (averaged over vertices)</th>
<th>Clustering in random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film actors</td>
<td>225,226</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>1,520,251</td>
<td>4.6</td>
<td>4.91</td>
<td>0.56</td>
<td>1.8 x 10^-4</td>
</tr>
<tr>
<td>E.Coli substrate graph</td>
<td>282</td>
<td>2.9</td>
<td>3.04</td>
<td>0.32</td>
<td>0.026</td>
</tr>
<tr>
<td>C.Elegans</td>
<td>282</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**Source:** L. Adamic
Recap: A Small World

- What is the typical shortest path length between two people?
  - Milgram’67 : 6.2 (using letters to a Boston stock-broker)
  - Dodds et. al.’03 : 7 (using email to multiple targets)

- Implications
  - On the clustering
  - On the way people route and find the target
IMPLICATIONS ON NAVIGATION
Decentralized Search

- Navigate to the target

Setting:
- source s knows the locations of its friends and target t
- s does not know the entire graph
- **Geographic routing** (not shortest paths!)
  - Navigate to the node closest to target t
- Interested in **Search Time** T
  - Number of steps to reach t
Main results

- “Searchable”
  - Low search time, more precisely $O((\log n)^\beta)$
  - Kleinberg’s Model: $O((\log n)^2)$ (beta = 2)

- Not searchable
  - High search time, more precisely $O(n^\alpha)$
  - Watts-Strogatz Model: $O(n^{2/3})$ (alpha = 2/3 for 2-d grid)
  - Erdos-Renyi Model: $O(n)$ (alpha = 1)

Note: All are asymptotic results as $n \to \infty$
Not all small worlds are alike!

- **Watts-Strogatz Model:**
  - d-lattice with each node has ONE random edge
  - Small world (i.e. paths of $O(\log n)$ length exist)
  - BUT a decentralized algorithm needs $O(n^{1/2})$ steps!
  - For a d-dim it is $O(n^{d/(d+1)})$: so for a grid it is $O(n^{2/3})$

- But for Kleinberg’s model it is $O((\log n)^2)$ for the 1-d case
Kleinberg’s Model

- Intuition: real long-range links are not random
  - They follow some ‘geography’

- Model: (for 2-d grid)
  - Node has one long-range link
    \( P(u \rightarrow v) \propto d(u, v)^{-\alpha} \)
  - \( d(u, v) = \) grid distance between \( u \) and \( v \)
  - \( \alpha = \) parameter \( \geq 0 \)
Kleinberg’s Model in 1-d: Proof

- **Result:** for $\alpha = 1$, $T$ is $O((\log n)^2)$ steps

- **Assume:** $d(i, t) = d$

- **Set interval $I$ around target:**
  
  $I = d$ (so $x = d/2$)

\[
\text{Prob} \begin{cases} 
\text{Long range link from } i \\
\text{points to a node in } I
\end{cases} = ?
\]
Search time Kleinberg’s Model in 1-d: Proof

- \( P(i \rightarrow w) = \frac{d(i,w)^{-1}}{\sum_{u \neq i} d(i,u)^{-1}} \)

(by def.)

Now:

\[
\sum_{u \neq i} d(i,u)^{-1} = \sum_{d=1}^{n/2} 2d^{-1} = 2 \sum_{d=1}^{n/2} \frac{1}{d} \leq 2 \log_2 n
\]

Harmonic numbers

\[
H(n/2) = \sum_{d=1}^{n/2} \frac{1}{d} \leq 1 + \int_{1}^{n/2} \frac{dx}{x} = 1 + \ln \left( \frac{n}{2} \right) \leq 1 + \log_2 \left( \frac{n}{2} \right)
\]
Proof Contd.

- **P(i points to a node in I)**
  
  \[ P(i \text{ points to a node in I}) = \sum_{w \in I} P(i \rightarrow w) \]
  
  \[ \geq \frac{1}{2 \log_2 n} \sum_{w \in I} \frac{1}{d(i, w)} \]
  
  \[ \geq \frac{1}{2 \log_2 n} \times d \times \frac{2}{3d} \]
  
  \[ = O\left(\frac{1}{\log_2 n}\right) \]

*Max d(i, w) = d + x = 3d/2*
Proof Contd.

- **Strategy:** use myopic search. Take the long-range link if it brings you closer to $t$ otherwise use a local link.

- Just proved that:
  \[
  \text{Prob}(\text{Long range link from node } i \text{ points to a node in } I) = O(1/\log_2 n)
  \]

- So, in expected steps = $O(\log_2 n)$
  - we get from node $i$ to interval $I$ around $t$
  - we halved the distance to $t$!
Proof Contd.

- **BUT**
  - Distance can be halved max. of \( \log_2 n \) times
  - So expected time to reach \( t \)
  \[
  = O(\log_2 n \times \log_2 n)
  \]
  \[
  = O((\log_2 n)^2)
  \]

Milgram’67: “The geographic movement of the [letter] from Nebraska to Massachusetts is striking. There is a progressive closing in on the target area as each new person is added to the chain”
Parameter $\alpha$

- $\alpha = 0 ::: \text{Watts-Strogatz} ::: O(\sqrt{n})$ search steps
- $\alpha = 1 ::: O((\log n)^2)$ search steps

Best when $\alpha = 2$

The “Inverse-Square” Network
Parameter $\alpha$

SMALL $\alpha$: Too many long links

LARGE $\alpha$: Too many short links
\[ P(u \rightarrow v) \sim d(u,v)^{- \text{dim}} \text{ works} \]

- Approx. uniform over all scales of resolution 
  \((d, d^2, d^3, \ldots)\)

\[ N = \# \text{ of points at distance } d \sim d^{\text{dim}} \]
\[ P = \text{so if prob. goes down as } d^{-\text{dim}} \]
\[ \Rightarrow \]
\[ N \ast P = \text{Constant probability of a link to a node at every } d \]
P(\(u \rightarrow v\)) \sim d(u,v)^{-\text{dim}} \text{ works}

- How the argument works for 2-d grid

P(u \text{ points to a node in } I)

> 1/(\text{normalization}) \ast \text{size}(I) \ast \text{Prob. on each node}

= \log_2 n \ast d^2 \ast d^{-2}
Other models: Hierarchies

- $h(u, v) = \text{tree-distance}$
  (height of least-common ancestor)
- $P(u \rightarrow v) \sim b^{-\alpha} h(u,v)$
- # nodes at dist. $h = (b-1) b^{h-1} \sim b^h$
  – (again uniform at all scales)
- Start at $s$, want to go to $t$
  – Only see links out of node you are at
  – You know where target $t$ is in the tree
Extensions [Watts et. al.’02]

- Multiple hierarchies: geography, profession...
- Generate separate random graph in each hierarchy
- Superimpose the graphs
- Search: choose a link that gets closest in any hierarchy

- Analysis: typically simulations
  - “Too many” hierarchies hurt
Empirical Studies for navigation

- Small-World HP [Adamic-Adar’05]
  - Email logs from HP-labs (436 people)
  - Link if u and v exchanged > 5 emails each way
  - Organization

Q: How many edges cross groups?
A: $P(u \rightarrow v) \sim \frac{1}{(social \ dist.)^{3/4}}$
**Small-World in LiveJournal** [Liben-Nowell et al ’05]

- LiveJournal data
  - Bloggers + Zip-codes
  - 0.5M bloggers, 4M links

- $P(u, v) = \delta^{-\alpha} :: \text{What is } \alpha$?

- **Problem:** Non-uniform population density
- **Solution:** Rank-based friendship

Link length in a network of bloggers
Small-World in LiveJournal [Liben-Nowell et al. ’05]

Non-uniform density

Uniform Density

\[ \text{rank}(v,w) = | \{ u : d(v,u) < d(v,w) \} | \]
Small-World in LiveJournal [Liben-Nowell et. al ’05]

- $P(u,v) \sim \text{rank}(u,v)^{-\alpha} :::: \text{What is } \alpha?$
- If uniform density then $\alpha = \text{dim.}$
- In this special case, $\alpha = 1$ is the best for search

No East-West Coast difference
Small-World in LiveJournal [Liben-Nowell et. al ’05]

- Geographic navigation
  - Decentralized search in LiveJournal: 12% hops finished, average 4.12 hops
Applications of search: Finding Files in P2P network

- Cycle with node ids 0 to $2^{m-1}$
- File key (k) is assigned to node $a(k)$ with ID $\geq k$

**Search**

- If every node knows its immediate neighbor, use sequential search
- Otherwise?
Applications of search: Finding Files in P2P network

- **Chord Algorithm [Stoica et.al. 2001]**
  - A node maintains a table of \( m = \log(N) \) entries
  - i-th entry of a node \( n \) contains the address of \( n+2^i \) neighbor
  - Take the longest link that does not overshoot
    - With each step we halve the distance to the target

\[ \begin{align*}
N_{8+1} &= N_{14} \\
N_{8+2} &= N_{14} \\
N_{8+4} &= N_{14} \\
N_{8+8} &= N_{21} \\
N_{8+16} &= N_{32} \\
N_{8+32} &= N_{42} \\
N_{42+1} &= N_{48} \\
N_{42+2} &= N_{48} \\
N_{42+4} &= N_{48} \\
N_{42+8} &= N_{51} \\
N_{42+16} &= N_{51} \\
N_{42+32} &= N_{51}
\end{align*} \]

\( O(\log N) \) steps!