Search in structured networks
How do we search?

Mary

Who could introduce me to Richard Gere?

Bob

Jane

Richard Gere – spaceodissey, Flickr; http://creativecommons.org/licenses/by/2.0/deed.en

Friends collage – luc, Flickr; http://creativecommons.org/licenses/by/2.0/deed.en
power-law graph

number of nodes found

- 94
- 67
- 63
- 54
- 6
- 2
number of nodes found

Poisson graph
How would you search for a node here?
What about here?
gnutella network fragment
Gnutella network

50% of the files in a 700 node network can be found in < 8 steps
And here?
here?
here?

Source: http://maps.google.com
here?

Source: http://maps.google.com
Small world experiments review

Milgram (1960’s), Dodds, Muhamad, Watts (2003)

Given a target individual and a particular property, pass the message to a person you correspond with who is “closest” to the target.

Short chain lengths – six degrees of separation

Typical strategy – if far from target choose someone geographically closer, if close to target geographically, choose someone professionally closer
Is this the whole picture?

Why are small worlds navigable?

How are people able to find short paths?

How to choose among hundreds of acquaintances?

**Strategy:**
Simple greedy algorithm - each participant chooses correspondent who is closest to target with respect to the given property

**Models**

gеography
Kleinberg (2000)

hierarchical groups
Watts, Dodds, Newman (2001), Kleinberg(2001)

high degree nodes
Reverse small world experiment

- Killworth & Bernard (1978):
- Given hypothetical targets (name, occupation, location, hobbies, religion...) participants choose an acquaintance for each target
- Acquaintance chosen based on
  - (most often) occupation, geography
  - only 7% because they “know a lot of people”
- Simple greedy algorithm: most similar acquaintance
- two-step strategy rare

How many hops actually separate any two individuals in the world?

- Participants are not perfect in routing messages
- They use only local information
- “The accuracy of small world chains in social networks”
  Peter D. Killworth, Chris McCarty, H. Russell Bernard & Mark House:
  - Analyze 10920 shortest path connections between 105 members of an interviewing bureau,
  - together with the equivalent conceptual, or ‘small world’ routes, which use individuals’ selections of intermediaries.
  - This permits the first study of the impact of accuracy within small world chains.
  - The mean small world path length (3.23) is 40% longer than the mean of the actual shortest paths (2.30)
  - Model suggests that people make a less than optimal small world choice more than half the time.
review: Spatial search


“The geographic movement of the [message] 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”

S.Milgram ‘The small world problem’, Psychology Today 1,61,1967

nodes are placed on a lattice and connect to nearest neighbors

additional links placed with $p_{uv} \sim d_{uv}^{-r}$
When $r=0$, links are randomly distributed, $\text{ASP} \sim \log(n)$, $n$ size of grid
When $r=0$, any decentralized algorithm is at least $a_0 n^{2/3}$

no locality

$p \sim p_0$

When $r<2$, expected time at least $\alpha_r n^{(2-r)/3}$
Overly localized links on a lattice

When $r>2$ expected search time $\sim N^{(r-2)/(r-1)}$
Links balanced between long and short range

When $r=2$, expected time of a DA is at most $C \,(\log N)^2$

$p \sim \frac{1}{d^2}$
how does the probability of long-range links affect search?

http://projects.si.umich.edu/netlearn/NetLogo4/SmallWorldSearch.html
Testing search models on social networks

**advantage:** have access to entire communication network and to individual’s attributes

**Use a well defined network:**
HP Labs email correspondence over 3.5 months

Edges are between individuals who sent at least 6 email messages each way

450 users
median degree = 10, mean degree = 13
average shortest path = 3

**Node properties specified:**
degree
geographical location
position in organizational hierarchy

Can greedy strategies work?
the network otherwise known as sample.gdf
Strategy 1: High degree search

Power-law degree distribution of all senders of email passing through HP labs

![Graph showing a power-law distribution with the x-axis labeled as 'number of recipients sender has sent email to' and the y-axis labeled as 'proportion of senders'.]
Filtered network
(at least 6 messages sent each way)

Degree distribution no longer power-law, but Poisson

It would take 40 steps on average (median of 16) to reach a target!
Strategy 2: Geography
Communication across corporate geography

87% of the 4000 links are between individuals on the same floor

Cubicle distance vs. probability of being linked

Livejournal

LiveJournal provides an API to crawl the friendship network + profiles
  - friendly to researchers
  - great research opportunity

basic statistics
  - Users (stats from April 2006)
    - How many users, and how many of those are active?
    - Total accounts: 9980558
    - ... active in some way: 1979716
    - ... that have ever updated: 6755023
    - ... updating in last 30 days: 1300312
    - ... updating in last 7 days: 751301
    - ... updating in past 24 hours: 216581
Predominantly female & young demographic

- Male: 1370813 (32.4%)
- Female: 2856360 (67.6%)
- Unspecified: 1575389

Age distribution:

- 13: 18483
- 14: 87505
- 15: 211445
- 16: 343922
- 17: 400947
- 18: 414601
- 19: 405472
- 20: 371789
- 21: 303076
- 22: 239255
- 23: 194379
- 24: 152569
- 25: 127121
- 26: 98900
- 27: 73392
- 28: 59188
- 29: 48666
Geographic Routing in Social Networks

- David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins (PNAS 2005)

- data used
  - Feb. 2004
  - 500,000 LiveJournal users with US locations
  - giant component (77.6%) of the network
  - clustering coefficient: 0.2
The broad degree distributions we’ve learned to know and love

- but more probably lognormal than power law

Results of a simple greedy geographical algorithm

- Choose source $s$ and target $t$ randomly
- Try to reach target’s city – not target itself
- At each step, the message is forwarded from the current message holder $u$ to the friend $v$ of $u$ geographically closest to $t$

13% of the chains are completed

stop if $d(v,t) > d(u,t)$

80% of the chains are completed

pick a neighbor at random in the same city if possible, else stop

the geographic basis of friendship

- $\delta = d(u,v)$ the distance between pairs of people
- The probability that two people are friends given their distance is equal to
  - $P(\delta) = \varepsilon + f(\delta)$, $\varepsilon$ is a constant independent of geography
  - $\varepsilon$ is $5.0 \times 10^{-6}$ for LiveJournal users who are very far apart

the geographic basis of friendship

- The average user will have ~ 2.5 non-geographic friends
- The other friends (5.5 on average) are distributed according to an approximate 1/distance relationship
- But 1/d was proved not to be navigable by Kleinberg, so what gives?

Navigability in networks of variable geographical density

- Kleinberg assumed a uniformly populated 2D lattice
- But population is far from uniform
- population networks and rank-based friendship
  - probability of knowing a person depends not on absolute distance but on relative distance (i.e. how many people live closer) \( \Pr[u \rightarrow v] \sim 1/\text{rank}_u(v) \)

what if we don’t have geography?
does community structure help?
review: hierarchical small world models

Individuals classified into a hierarchy, \( h_{ij} = \) height of the least common ancestor.

\[
p_{ij} \sim b^{-\alpha h_{ij}}
\]

**Theorem:** If \( \alpha = 1 \) and outdegree is polylogarithmic, can \( s \sim O(\log n) \)

Group structure models:
Individuals belong to nested groups
\( q = \) size of smallest group that \( v,w \) belong to

\[
f(q) \sim q^{-\alpha}
\]

**Theorem:** If \( \alpha = 1 \) and outdegree is polylogarithmic, can \( s \sim O(\log n) \)

---

Why search is fast in hierarchical topologies

\[ \lambda^2 |R| < |R'| < \lambda |R| \]

\[ k = c \log^2 n \]

calculate probability that \( s \) fails to have a link in \( R' \)
individuals belong to hierarchically nested groups

\[ p_{ij} \sim \exp(-\alpha x) \]

multiple independent hierarchies \( h=1,2,\ldots,H \) coexist corresponding to occupation, geography, hobbies, religion…

Social distance—Bipartite networks:

Identity and search in social networks
Watts, Dodds, Newman (2001)

Message chains fail at each node with probability $p$
Network is ‘searchable’ if a fraction $r$ of messages reach the target

$$q = \langle (1 - p)^L \rangle_L \geq r$$

Small World Model, Watts et al.

Fits Milgram’s data well

Model parameters:
- $N = 10^8$
- $z = 300$
- $g = 100$
- $b = 10$
- $\alpha = 1$, $H = 2$

$L_{\text{model}} = 6.7$
$L_{\text{data}} = 6.5$

more slides on this:
does it work in practice? back to HP Labs: Organizational hierarchy
Email correspondence superimposed on the organizational hierarchy

Example of search path

- Hierarchical distance = 5
- Search path distance = 4
Probability of linking vs. distance in hierarchy

in the ‘searchable’ regime: $0 < \alpha < 2$ (Watts, Dodds, Newman 2001)
Results

<table>
<thead>
<tr>
<th>distance</th>
<th>hierarchy</th>
<th>geography</th>
<th>geodesic</th>
<th>org</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>mean</td>
<td>5.7 (4.7)</td>
<td>12</td>
<td>3.1</td>
<td>6.1</td>
<td>57.4</td>
</tr>
</tbody>
</table>

Expt 2

Searching a social networking website

Source: ClubNexus - Orkut Buyukkokten, Tyler Ziemann
Profiles:

status (UG or G)
year
major or department
residence
gender

Personality (choose 3 exactly):

you funny, kind, weird, ...
friendship honesty/trust, common interests, commitment, ...
romance -“- 
free time socializing, getting outside, reading, ...
support unconditional accepters, comic-relief givers, eternal optimists

Interests (choose as many as apply)

books mystery & thriller, science fiction, romance, ...
movies western, biography, horror, ...
music folk, jazz, techno, ...
social activities ballroom dancing, barbecuing, bar-hopping, ...
lnd sports soccer, tennis, golf, ...
water sports sailing, kayaking, swimming, ...
other sports ski diving, weightlifting, billiards, ...
Differences between data sets

<table>
<thead>
<tr>
<th>HP labs email network</th>
<th>Online community</th>
</tr>
</thead>
<tbody>
<tr>
<td>• complete image of communication network</td>
<td>• partial information of social network</td>
</tr>
<tr>
<td>• affinity not reflected</td>
<td>• only friends listed</td>
</tr>
</tbody>
</table>
Degree Distribution for Nexus Net
2469 users, average degree 8.2

Problem: how to construct hierarchies?

Probability of linking by separation in years

Hierarchies not useful for other attributes:

Geography

Other attributes: major, sports, freetime activities, movie preferences…

Strategy using user profiles

prob. two undergrads are friends (consider simultaneously)

• both undergraduate, both graduate, or one of each
• same or different year
• both male, both female, or one of each
• same or different residences
• same or different major/department

Results

<table>
<thead>
<tr>
<th>strategy</th>
<th>median</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>133</td>
<td>390</td>
</tr>
<tr>
<td>high degree</td>
<td>39</td>
<td>137</td>
</tr>
<tr>
<td>profile</td>
<td>21</td>
<td>53</td>
</tr>
</tbody>
</table>

With an attrition rate of 25%, 5% of the messages get through at an average of 4.8 steps, => hence network is barely searchable
conclusions

- Individuals associate on different levels into groups.
- Group structure facilitates decentralized search using social ties.
- Hierarchy search faster than geographical search
- A fraction of ‘important’ individuals are easily findable
- Humans may be more resourceful in executing search tasks: making use of weak ties using more sophisticated strategies