Search in Social and Navigation Networks

Denis Helic

TU Graz, KTI

May 13, 2013
Navigation in Social Networks

- Nebraska stockholders
- Nebraska random
- Boston random
- Target
- Boston stockbroker
Navigation in social networks

- Two experimental findings
- There are short paths in social networks
- People find those short paths efficiently
- Local search
- Watts and Kleinberg explained the phenomenon
Network Navigability

Definition

Network is navigable if and only if there is a short path between all or almost all pairs of nodes in the network.

Formally:

1. There exist a giant component
2. The effective diameter is low – bounded by $\log(n)$, where $n$ is the number of nodes in the network
Example 1:

![Network Diagram]

**Figure:** Network is not navigable because there is no giant component, i.e., the network is not connected.
Example 2:

\[
\text{Not navigable: giant component, BUT eff. diam: } 7 > \log_2(8).
\]

Figure: Now, there is a giant component, i.e. the network is connected. However the network is not navigable because \(\text{eff. diam} = 6\), and \(6 > \log_2(8)\).
Example 3:

Figure: The network is navigable because there is a giant component and $\text{eff. diam} = 2$. Effective diameter is bounded by $\log_2(10)$. 
We discussed so far global network navigability.
In case we have global knowledge of network.
Easy to design efficient navigation procedures to find an arbitrary target node from an arbitrary start node.
E.g., breadth-first search with time complexity $O(n + m)$. 

What is with *local* network navigability?
- In case we have only local knowledge of network
- We know only outgoing links from the current node
- We do not know links of the neighbors of the current node
- Is it now easy to design efficient navigation procedures?
- These procedure are typically called decentralized search
Local Network Navigability

- How we are able to find other people efficiently?
- What is the structure of a social network that make efficient decentralized search possible?
- There are structural cues in the network which allows us to design sub-linear algorithms
- What are these structural clues?
Local Network Navigability

Example:

Figure: A is start node and D is target node.
Local Network Navigability

Step 1:

Figure: There are two possible paths from A. Obviously, the optimal path leads to B. What is the structural property that can guide us in selecting B?
Local Network Navigability

Step 1:

Figure: There are two possible paths from A. Obviously, the optimal path leads to B. What is the structural property that can guide us in selecting B?

Degree
Utilizing high-degree nodes in power-law networks (Search in power-law networks by Adamic, 2003)
Adapt the breadth-first search by first investigating nodes with higher degrees.
Reduces time complexity to range $O(n^{\frac{1}{2}})$ to $O(\log(n))$ for $2 \leq \alpha \leq 3$
Local Network Navigability

Step 2:

A network is efficiently navigable iff:
If there is an algorithm that can find a short path with only local knowledge, and the delivery time of the algorithm is bounded polynomially by $\log^k(n)$.

Efficiently navigable, if the algorithm knows it needs to go through $A \in B \in C$.


Figure: There are seven possible paths from B. Obviously, the optimal path leads to C. What is the structural property that can guide us in selecting C?
Local Network Navigability

Step 2:

![Network diagram]

**Figure:** There are seven possible paths from B. Obviously, the optimal path leads to C. What is the structural property that can guide us in selecting C?

**Local clustering**
Summarizing, local network navigability requires:

1. Existence of network hubs that are connected to many nodes
2. Existence of network clusters where nodes are highly interlinked
Local Network Navigability

We base our strategy on probability:

1. The first requirement increases the probability that a hub has a link to the target node or to the target node cluster.
2. The second requirement increases the probability that there is a link to the target node from the cluster it belongs to.
Local Network Navigability

Formally:

1. Power-low degree distribution with exponent $\gamma$
2. High clustering coefficient $C$
Local Network Navigability

![Diagram showing navigable and non-navigable regions of clustering and degree distribution.]

Figure: Navigable networks in $\gamma$, $C$ space.
Revisiting Step 2:

Figure: There are seven possible paths from B. Obviously, the optimal path leads to C. What is an additional hint that can guide us in selecting C over E?
Local Network Navigability

Revisiting Step 2:

\[
\text{A network is efficiently navigable iff: If there is an algorithm that can find a short path with only local knowledge, and the delivery time of the algorithm is bounded polynomially by } \log k(n).\]

Figure: There are seven possible paths from B. Obviously, the optimal path leads to C. What is an additional hint that can guide us in selecting C over E?

Node similarity
Local Network Navigability

- Similarity between nodes is external to the network
- It is derived from some additional information that we have about network nodes
- In Millgram’s experiment people selected the next person according to their occupation or geography
- E.g., a friend in Boston, or a friend working in financial sector
All this information, i.e. degrees, clustering, similarity is our background knowledge about the network.

We use this background knowledge to guide us in our search for a target node.

At each step we consult the background knowledge.

We ask which link leads with the highest probability to a given target node.
Decentralized search

- Model of navigation with local knowledge only
- Intuitions about the network nodes
- Distance metric
- Algorithmically people are greedy
Decentralized Search

- Background knowledge defines a notion of *distance* between nodes.
- It is a *metric space* assigning to each node unique coordinates.
- A distance metrics allows to calculate the distance between nodes.
- Background knowledge is a black-box executing a simple function: `getDistance(node, target_node)`.
Figure: Informed greedy decentralized search.
Figure: Informed greedy decentralized search.
Figure: Informed greedy decentralized search.
Figure: Informed greedy decentralized search.
Decentralized search

Figure: Informed greedy decentralized search.
Decentralized search

**Figure:** Informed greedy decentralized search.
Decentralized search

Figure: Informed greedy decentralized search.
Figure: Informed greedy decentralized search.
Figure: Informed greedy decentralized search.
Decentralized search

Figure: Informed greedy decentralized search.
What is the nature of background knowledge?
It is a metric space, e.g. 1-D spaces, 2-D vector spaces, 3-D Euclidean spaces, hyperbolic spaces, ...
Example with a circle as a metric space by Boguna: Navigability of Complex Networks, 2007
Example with a hyperbolic space by Krioukov: Hyperbolic Geometry of Complex Networks, 2010
Hierarchical Background Knowledge

- A hierarchy of nodes is also a metric space
- Watts: Identity and Search in Social Networks, 2002
- Adamic: How to search a social network, 2005
Hierarchy as a Metric Space

Figure: Node distances in a hierarchy.
Hierarchy as a Metric Space

- Height of least common ancestor in the hierarchy
- Parent and siblings are at $d = 1$
- From there on recursively: the parent’s siblings are at $d = 2$, the children of the parent’s siblings are at $d = 3$, and so on
Hierarchy as a Metric Space

If neither of the nodes are parents, children, grandparents, grandchildren, etc:

\[ d(i, j) = h(i) + h(j) - 2h(lca(i, j)) - 1 \]

Otherwise:

\[ d(i, j) = \text{abs}(h(i) - h(j)) \]
### Navigation in Information vs. Social Networks

<table>
<thead>
<tr>
<th></th>
<th><strong>Social Networks</strong></th>
<th><strong>Information Networks</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Agents per search</td>
<td>multiple agents</td>
<td>single agent</td>
</tr>
<tr>
<td>Type of routing</td>
<td>decentralized</td>
<td>centralized</td>
</tr>
<tr>
<td>Knowledge of network</td>
<td>local</td>
<td>local</td>
</tr>
<tr>
<td>Local knowledge</td>
<td>rich</td>
<td>limited</td>
</tr>
<tr>
<td>Searcher</td>
<td>endogenous</td>
<td>exogenous</td>
</tr>
<tr>
<td>Routing decisions</td>
<td>social intuitions</td>
<td>topical intuitions</td>
</tr>
<tr>
<td>Candidate consultation</td>
<td>costly</td>
<td>cheap</td>
</tr>
</tbody>
</table>

**Notes:**
- **Social Networks** vs **Information Networks**
- Comparison of agents per search, type of routing, knowledge of network, local knowledge, searcher, routing decisions, and candidate consultation.
- **Social Intuitions** vs **Topical Intuitions**
- **Costly** vs **Cheap**
Human Navigation on the Web

- Human Wayfinding in Information Networks
- Robert West and Jure Leskovec
- A study of a large collection of human click paths
- Statistical analysis of the click paths and strategies humans apply when navigating information networks
Users play a navigation game in Wikipedia.

You get a randomly selected starting page and a randomly selected target page.

In a short time period you have to reach the target page only by clicking on links.

Similar as Wikigame: http://thewikigame.com/
Human Navigation

Figure: Sample game
Human Navigation

- Start page: DIK-DIK
- Target page: Albert Einstein
- Using back button is allowed (electron-atom)
- Full path is a path including back clicks
- Effective path is a path without back clicks
Shortest path to the target in squares
If a click decreases the shortest path it is called lucrative
In the example not every click is lucrative
At first there is progress and then a cycle
Water is important because it connects two parts of the network → it is a hub
Human Navigation

Figure: Clicks histogram
Human Navigation

- Black circles show shortest paths $\rightarrow$ small-world network
- Blue X’s show human effective paths
- Red dots show human complete paths
- Green +’s complete human paths corrected for drop-outs
Human Navigation

<table>
<thead>
<tr>
<th>path-length metric</th>
<th>mode</th>
<th>median</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>shortest possible paths</td>
<td>3</td>
<td>3</td>
<td>2.9</td>
</tr>
<tr>
<td>human, effective</td>
<td>4</td>
<td>4</td>
<td>4.9</td>
</tr>
<tr>
<td>human, incl. back-clicks</td>
<td>4</td>
<td>5</td>
<td>5.8</td>
</tr>
<tr>
<td>human, drop-out–corrected</td>
<td>4</td>
<td>6</td>
<td>8.9</td>
</tr>
</tbody>
</table>

**Figure:** Click summaries
Effective human paths are typically not much longer than optimal shortest paths.
They differ typically by one click.
Full path lengths with back-clicks are one click more than effective path lengths.
Typically, one back-click per game.
Still very efficient.
The variance in human paths is higher than for optimal solutions → heavy tail.
Human Navigation

Two questions

- What is the reason for the large variance?
- Why is human search still so efficient on average?

Two possible answers to the first question:

- Some missions are harder than the others
- Some users are better than the others
Human Navigation

- Two questions
  - What is the reason for the large variance?
  - Why is human search still so efficient on average?
- Two possible answers to the first questions
  - Some missions are harder than the others
  - Some users are better than the others
Some missions have longer optimal paths
Necessarily, some of them are harder than the others
Even missions with the same optimal paths can be different because e.g. links are more or less lucrative
Four selected missions with shortest path of 3 have been posted more often to users
Two questions arise: First, what is the reason for the large variance in human search time? Second, why is human search still so efficient on average?—several answers are conceivable. One might pose to 4/5/5.8 for all games). Of course, even among missions with a SPL of 3 clicks, the mode/mean/median is 4/5/6.0, as opposed to 3.2/3.9. A harder mission allows for shorter games on average than others. This leads us to conclude that both hardness of mission and individual skill play a role in explaining the large search time variance. Second, why is human search still so efficient on average?—several answers are conceivable. One might pose to 4/5/5.8 for all games). Of course, even among missions with a SPL of 3 clicks, the mode/mean/median is 4/5/6.0, as opposed to 3.2/3.9. A harder mission allows for shorter games on average than others. This leads us to conclude that both hardness of mission and individual skill play a role in explaining the large search time variance.

Figure: Specific search missions
Human Navigation

- There is considerable search time variance
- Some missions allow for lower average search time
- Hardness of mission plays a role
- Individual skill play plays also a role
Human Navigation

- Why humans are able to navigate efficiently?
- Is there a sampling bias?
- There are many drop-outs, e.g. 54% of all games are canceled before reaching the target page.
- This might introduce a bias towards observing shorter chains than we would observe by forcing users to finish.
- This might produce longer chains.
Human Navigation

Figure: Drop-out rates
Human Navigation

- At each step players give up with 10% probability
- Using those drop-rates we can correct for that bias
- We compute an ideal histogram assuming that the users never give up
- Although longer games are now more frequent the distributions are similar qualitatively
- Median e.g. 1 click higher
Human Navigation

Figure: Clicks histogram
Conjecture: Wikipedia networks is efficiently navigable (even without global knowledge)

Humans have an intuition of what to expect behind a link

Probability of two articles being linked is higher if they are similar
Click path analysis
Games with shortest path 3
Only effective paths analyzed
Analysis of how some quantitative properties change along the paths with the game progress
Curves combined from games having the same effective path length
Human Navigation

Figure: Properties evolution along the click paths
Degree and similarity are most important factors in human navigation.

In the early phase, degree is the most important.

In the endgame, similarity is the most important.

To analyze the importance of features, train a logistic regression model.

The feature weights allow us to infer the importance of a particular feature.
Figure: Feature weights
Human Navigation

- Both features are positive everywhere
- Both high degree and high similarity are characteristics of the human clicks
- Degree dominates in the early game
- As the game progress similarity is increasingly important
- Similarity starts dominating earlier in more efficient games
Decentralized search as a model of navigation in information networks?
Limited topical knowledge
Cheap candidate consultation
High variation in user navigation
We model uncertainty probabilistically
# Action Selection

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Definition</th>
<th>Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>$\min(d(j, t))$</td>
<td>Always follows intuitions</td>
</tr>
<tr>
<td>$\epsilon$-greedy</td>
<td>$\min(d(j, t), P = 1 - \epsilon$ random, $P = \epsilon$)</td>
<td>Follows intuitions with probability $1 - \epsilon$; acts randomly with probability $\epsilon$</td>
</tr>
<tr>
<td>Softmax</td>
<td>$p(j) \propto e^{cf(j)}$, $f(j) = 1 - \frac{d(j, t)}{\max_{k,l \in V} d(k, l)}$</td>
<td>Follows intuitions to different extents (controlled by $c$)</td>
</tr>
<tr>
<td>Inv. distance</td>
<td>$p(j) \propto f(j)^{-c}$, $f(j) = \frac{d(j, t)}{\max_{k,l \in V} d(k, l)}$</td>
<td>Follows intuitions to different extents (controlled by $c$)</td>
</tr>
</tbody>
</table>
Figure: Stochastic decentralized search.
$\epsilon$-greedy

Figure: Stochastic decentralized search.
Figure: Stochastic decentralized search.
$\epsilon$-greedy

Figure: Stochastic decentralized search.
Figure: Stochastic decentralized search.
$\epsilon$-greedy

Figure: Stochastic decentralized search.
$\epsilon$-greedy

Figure: Stochastic decentralized search.
Figure: Stochastic decentralized search.
$\epsilon$-greedy

**Figure:** Stochastic decentralized search.
ε-greedy

Figure: Stochastic decentralized search.
Figure: Stochastic decentralized search.
## Action Selection

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Definition</th>
<th>Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>$\min(d(j, t))$</td>
<td>Always follows intuitions</td>
</tr>
<tr>
<td>$\epsilon$-greedy</td>
<td>$\min(d(j, t), P = 1 - \epsilon$ random, $P = \epsilon$</td>
<td>Follows intuitions with probability $1 - \epsilon$; acts randomly with probability $\epsilon$</td>
</tr>
<tr>
<td>Softmax</td>
<td>$p(j) \propto e^{cf(j)}$, $f(j) = 1 - \frac{d(j, t)}{\max_{k, l \in V} d(k, l)}$</td>
<td>Follows intuitions to different extents (controlled by $c$)</td>
</tr>
<tr>
<td>Inv. distance</td>
<td>$p(j) \propto f(j)^{-c}$, $f(j) = \frac{d(j, t)}{\max_{k, l \in V} d(k, l)}$</td>
<td>Follows intuitions to different extents (controlled by $c$)</td>
</tr>
</tbody>
</table>
Figure: Stochastic decentralized search.
Figure: Softmax probabilities.
Figure: Softmax probabilities.
Figure: Softmax probabilities.
Softmax

Figure: Softmax probabilities.
Figure: Softmax probabilities.
Figure: Softmax probabilities.
Experimental Setup

- The WikiGame
- A randomly selected start page and a randomly selected target page
- Log of 250,000 successfully completed games
- Wikipedia network with 10 million nodes and 250 million links
- Distance space: a hierarchy of Wikipedia pages
Results

(a) $\epsilon$-greedy

(b) Softmax

(c) Inverse distance

Figure: Success rate and stretch of different navigation models.
Discussion

- Greedy strategy is too efficient
- People are not always greedy and deterministic
- Human navigation is stochastic in nature
- Appropriately configured any probabilistic model is good
- For simplicity and intuitive interpretation we consider only $\epsilon$-greedy
**Discussion**

- \( \epsilon \)-greedy: With \( P = 1 - \epsilon \) humans are greedy and follow their intuitions and with \( P = \epsilon \) they navigate randomly.
- In our dataset \( \epsilon = 0.15 \).

**Our interpretation**

- Humans are greedy when they *exploit* what they know.
- They act “randomly” when they *explore* what they do not know.
- Exploration: incomplete knowledge and cheap cost of consultation.
Results

(a) $\epsilon$-greedy

(b) Softmax

(c) Inverse distance

Figure: Hop length distributions.
Results

(a) $\epsilon$-greedy  
(b) Softmax  
(c) Inverse distance

Figure: Similarity of hop length distributions between models & humans, and models & shortest path.
Discussion

- Mode of human distribution is 4
- Mode of all simulators is 3
- Simulators are more efficient in the early game phases
- $\epsilon$ is not constant
- It is a function of the navigation progress
Figure: Rates of exploration vs. navigation progress.
Decaying $\epsilon$-greedy

- As humans make progress they orient them better in the network
- Ratio of exploitation to exploration increases

$$\epsilon(t) = \epsilon_0 \lambda^{-t}$$ (1)
Decaying $\epsilon$-greedy

(a) Success rate and stretch  (b) Hop length distributions  (c) KL divergence

Figure: Humans are matched with $\lambda = 2$, $\epsilon_0 = 0.9$. 
Decaying $\epsilon$-greedy

- General approach adaptable to other navigational settings
- E.g. a related starting page: pick a smaller $\epsilon_0$
- E.g. users orient them faster: pick a greater $\lambda$
- Fit the model instead of using it as a generative model
- Use another background knowledge, e.g. a smaller hierarchy
Decaying $\epsilon$-greedy

(a) Success rate and stretch  
(b) Hop length distributions  
(c) KL divergence

**Figure:** Humans are matched with $\lambda = 2$, $\epsilon_0 = 0.7$. 
Conclusions

- **Social networks**
  - Good knowledge about ego-networks
  - people *exploit* that knowledge
- **Information networks**
  - Incomplete knowledge
  - Humans *exploit* but they also tend to *explore*
Conclusions

- “Randomness” aspect of navigation
- Ratio of exploitation and exploration is not constant
- The rate of change and the initial ratio depend on navigation settings
Thank You!

Questions?