Web traffic: analysis of navigation data and modeling at single user level.

José Javier Ramasco
Outline

- Internet and the Web
- Navigation traces
- Data analysis at an aggregate level
- Individual-level data: navigation trees
- Models of Web navigation
Internet and the WWW (Web)

The Internet in 1969 (ARPA)
Internet and the WWW (Web)

The Internet today
Internet and the Web

Internet Users in the World by Geographic Regions

Source: Internet World Stats - www.internetworldstats.com/stats.htm
Estimated Internet users are 1,596,270,108 for March 31, 2009
Copyright © 2009, Miniwatts Marketing Group
Web navigation & navigation traces

http://www.a.edu
http://www.b.edu
Navigation traces

CLIENT

Buddy's Computer

192.168.65.33

NETWORK

SERVER

Vulgar Entertainment, Inc.

10.99.205.122

Port 13029

Port 80
Navigation traces (Web requests)

- Source MAC: 03:5a:66:17:90:5e
- Dest. MAC: 10:99:19:3f:51:2f
- Source IP: 192.168.39.190
- Dest. IP: 127.100.251.3
- Source Port: 9421
- Dest. Port: 80
- GET /index.html HTTP/1.1
- Agent: SuperCrawler-2009/beta
- Referer: http://www.grumpy-puppy.com/
- Host: www.happy-kitty.com
Why to study navigation traces?

- Privacy concerns
- High potential benefits
Why to study navigation traces?

- We are interested in navigation traces from the point of view of the study of human activity and human interaction with an information-based system.

- Our final aim would be to be able to model this navigation processes in a realistic way.

- All our data is properly anonymized, and we comply with the laws or rules for privacy protection.
Databases

**Emory University**

- Students: 12,300
- Faculty: ~3,200

**Indiana University, Bloomington**

- Students: 42,000
- Faculty: ~5,000
**Databases (Emory University)**

- The database is formed by the weblogs of Emory University from Apr. 1st 2005 to Jan. 17th 2006 (41 weeks).
- Each click in a web of the university is registered at the time resolution of 1 second.

<table>
<thead>
<tr>
<th>Number of IPs</th>
<th>( N_{IP} )</th>
<th>3,179,671</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of URLs</td>
<td>( N_{URL} )</td>
<td>2,562,398</td>
</tr>
<tr>
<td>Total Number of page requests (weight)</td>
<td>( \Omega )</td>
<td>53,582,121</td>
</tr>
<tr>
<td>Average number of IPs introduced per day</td>
<td>( n_{IP} )</td>
<td>10,742</td>
</tr>
<tr>
<td>Average number of URLs introduced per day</td>
<td>( n_{URL} )</td>
<td>8,396</td>
</tr>
<tr>
<td>Average number of edges introduced per day</td>
<td>( e )</td>
<td>77,569</td>
</tr>
<tr>
<td>Average weight increment per day</td>
<td>( \Omega^\dagger )</td>
<td>186,350</td>
</tr>
</tbody>
</table>
Databases (Indiana University)

- The database is formed by the Web requests from a dorm of the University.
- Data collected from March 5, 2008 through May 3, 2008
- 408 million HTTP requests
- 1083 unique MAC addresses (Computers).
- 29.8 million page requests
- 967 unique users
- 630,000 Web servers
- 110,000 referring hosts
Aggregate results
Aggregate results

![Graph showing pageviews per day with data points and trend lines labeled as Thanksgiving and Average. The x-axis represents days from Mon, 21 to Tue, 29, and the y-axis represents pageviews per day in 10^5. There are vertical lines at Thu, 24, Sun, 27, and Mon, 28, indicating possible anomalies or special days.]
Aggregate results
Aggregate results

\[ P\text{(Requests)} \]

\[ Requests \]

exponent = 1.75
Aggregate results

![Graph showing the distribution of Users and P(Users) with an exponent of 2.0](image)
Aggregate results

\[ P(\tau_v) \sim \tau_v^{-1} \]

\[ P(\tau_c) \sim \tau_c^{-1.25} \]
Individual users results

\[ P(\text{Interval}) \]

\[ 10^0 \quad 10^{-2} \quad 10^{-4} \quad 10^{-6} \quad 10^{-8} \]

\[ \text{Interval (sec)} \]

\[ 10^0 \quad 10^{2} \quad 10^{4} \quad 10^{6} \]

\[ 1.45 \]

\[ 1.6 \]

\[ 3.5 \]

\[ 3 \]

\[ 2.5 \]

\[ 2 \]

\[ 1.5 \]

\[ 1 \]

\[ 0.5 \]

\[ 0 \]

\[ 1 \quad 1.2 \quad 1.4 \quad 1.6 \quad 1.8 \quad 2 \]

\[ \text{Power Law Exponent (t)} \]
Individual users results (Sessions)

www.mia-movies.com/cgi-bin/swiftclicks/out.cqi
Models: PageRank

PageRank is the simplest navigation model. The basic rules are:

- The users perform a random walk in the Web.
- With a certain probability $p$, they “teletransport”.
- Each of these “teletransportation” events mark the beginning of a new session.
Models: BookRank

We added a further detail in order to mimic the user Web surfing activity:

- Each user keeps a list of pages visited (bookmarks)
- They are ordered according to the number of times he/she visited the pages (rank $r$)
- Each time a user starts a new session, starting page selected from the bookmark list with $prob \sim r^{-\alpha}$
- Back bottom, $p$
Models: bookmarks + topicality (ABC)

In order to reproduce the single user traces we needed to add yet another ingredient related to pages topicality:

- Each trace starts with a E level
- There is a cost for each action $E_t = E_{t-1} - C$
- For each new page visited the $E_t = (1-\Delta \eta) E_{t-1}$
- If $E_t < 0$, new session
Simulation vs empirical data
Simulation vs empirical data

The graph compares the distribution of some metric $P(\omega)$ across different models and datasets. The x-axis represents the parameter $\omega$ on a logarithmic scale, while the y-axis shows the probability $P(\omega)$ also on a logarithmic scale. Different models and datasets are represented by distinct markers and colors:

- **Click stream** (black dots)
- **BookRank** (red line)
- **PageRank** (blue dashed line)

For empirical data:
- **Empirical** (black line)
- **BookRank** (white square)
- **ABC (G1)** (black triangle)
- **ABC (G2)** (black inverted triangle)

The graph illustrates how the empirical data aligns with the simulation models, highlighting the dissimilarities and similarities in their distributions.
Simulation vs empirical data

\[ P(D_s) \sim 10^{-\frac{D_s}{2}} \]

\[ P(N_s) \sim 10^{-\frac{N_s}{2}} \]

- **Empirical**
- **PageRank**
- **BookRank**
Simulation vs empirical data

Data

Model

\[ P(D_s) \] vs. \( D_s \)

\[ P(N_s) \] vs. \( N_s \)
Conclusions

- We have studied the Web navigation traces of a large number of users.
- Some of the features seem to be relatively universal despite natural user-user variability.
- We have proposed a family of models able to reproduce deeper and deeper characteristics of the users’ navigation patterns.
- How far should we go? Do this last simple model implement topicality satisfactorily? And what about real time dynamics? ...
Collaborators & papers

Mark Meiss  Bruno Gonçalves

Sandro Flammini  Fil Menczer

- Remembering what we like: Toward an agent-based model of Web traffic, WSDM 2009 Late Breaking Results
- What's in a Session: Tracking Individual Behavior on the Web, Hypertext 2009
- Agents, Bookmarks and Clicks: A topical model of Web traffic, Hypertext 2010