Around the Web in Six Weeks: Documenting a Large-Scale Crawl

Sarker Tanzir Ahmed, Clint Sparkman, Hsin-Tsang Lee, and Dmitri Loguinov

Internet Research Lab
Department of Computer Science and Engineering
Texas A&M University

April 29, 2015
Agenda

- Introduction
- Background
- Crawl Analysis
  - Page-Level
  - Server-Level
- Internet Coverage
- Extrapolation
- Conclusion
Introduction

• Web crawling is a challenging experiment
  – Its perceived difficulty hinders non-commercial endeavors

• Industry has been the major player
  – Reluctant to disclose actual methodology

• Academic endeavors are limited
  – Popular belief that a Internet-wide requires huge hardware setup
  – Most published crawls are rather limited in size and span in the Internet and lack useful details about the crawl
  – No standard methodology to compare different crawls
Introduction (2)

• Our IRLbot crawl experiment in 2007 is the largest non-commercial crawl of the Internet to this date
  – Collected 7.3B pages in 41 days using a single crawler node
• Here the objective is to dissect the collected data
  – Analyze Internet wide coverage, spam avoidance etc
  – Compare to commercial search engines using a novel method
Agenda

• Introduction
• Background
• Crawl Analysis
  – Page-Level
  – Server-Level
• Internet Coverage
• Extrapolation
• Conclusion
**Background - Inside a Web Crawler**

- Forms a cycle where each component has to keep up to persist the crawl rate $S$
- Example: IRLbot’s duplicate elimination rate was over 100K/s with peak rate $S=3K$ pps, $m=h=1$

### Term Description

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td># of downloaded pages</td>
</tr>
<tr>
<td>$q$</td>
<td>Fraction of HTML pages</td>
</tr>
<tr>
<td>$l$</td>
<td>Links/page</td>
</tr>
<tr>
<td>$p$</td>
<td>Fraction of unseen links</td>
</tr>
<tr>
<td>$h$</td>
<td># of crawler nodes</td>
</tr>
<tr>
<td>$S$</td>
<td>Crawl rate (pages/sec)</td>
</tr>
</tbody>
</table>
Background – Crawler Design (2)

- Crawl design boils down to a trade-off \( \{D/h, S/h, q, l\} \)
  - Increase in one typically results in decrease in others
- Different methods of scaling \( S \) in existing literature
  - Clear trade-off between \( D \) and \( S \)
  - Reduce \( q \) by crawling non-HTMLs (Mercator)
  - Eliminate dynamic URLs to reduce \( l \) (ClueWeb09)
  - Eliminate disk-based duplicate elimination by RAM-based method (UbiCrawler, WebBase), or by revisiting same pages (Internet Archive)
- None of at least 50M page crawls have real-time spam avoidance or global frontier prioritization
  - IRLbot uses real-time frontier prioritization
Background – Crawler Design (3)

- Among distributed crawlers, one of the most prominent is ClueWeb09
  - Parallelized Apache Nutch to 1600 processors in Google-NSF-IBM cluster and discarded all dynamic links (i.e., dropping \( l \) by 84%)
  - Crawled 1B pages in 52 days at average rate 222 pps

- Some IRLbot Configuration and Features
  - Used \( m=h=1 \), (i.e., one single crawler node, seeded from only www.tamu.edu)
  - Highest \( q \) and unrestricted \( l \)
  - Used real-time frontier prioritization based on the PLD graph
  - Rate \( S \) and \( D \) determined by factors outside our control (i.e., university bandwidth)
  - Collected \( D=7.3B \) pages in 41 days at average rate 2K pps
Agenda

• Introduction
• Background
• Crawl Analysis
  – Page-Level
  – Server-Level
• Internet Coverage
• Extrapolation
• Conclusion
Page-Level Analysis – URL Cycle

- **Connect** 7,606,109,371
- **Valid replies** 7,437,281,300 (98.9%)
- **Non-text/html** 86,476,067 (1.1%)
- **Network errors** 162,057,287 (2.1%)
- **HTTP errors** 6,770,784 (0.1%)
- **URLs with valid host** 8,058,635,011 (94.4%)
- **Admitted URLs** 8,267,316,148
- **Robot errors** 449,243,979 (5.6%)
- **Blacklisted** 3,281,661 (0.04%)
- **No DNS** 208,681,137
- **Unique URLs** 41,502,195,631
- **Unique edges** 310,361,986,596
- **Crawlable URLs** 377,995,369,202

**Robot error type**
- **Robots.txt disallow** 296,966,591
- **Network error on robots.txt** 106,638,856
- **HTTP error on robots.txt** 24,221,293
- **Robots.txt forbidden** 20,621,185
- **Robots.txt loop** 612,160
- **Robots.txt over 64KB** 183,894
- **Total** 449,243,979

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Full download** 7,437,281,300

**HTML 200 OK** 6,380,051,942

**Found URLs** 394,619,023,142

**HTTP errors** 6,770,784

**Network errors** 162,057,287

**HTTP errors** 6,770,784

**Bad ext** 3,332,296,97

**Admitted URLs** 8,267,316,148

**Found URLs** 387,605,655,905

**Total found URLs** 394,619,023,142

**Fail checks** 13,291,356,961

**Crawlable URLs** 377,995,369,202

**URLs with valid host** 8,058,635,011

**Robot errors** 449,243,979

**Blacklisted** 3,281,661

**No DNS** 208,681,137

**Robots.txt disallow** 296,966,591

**Network error on robots.txt** 106,638,856

**HTTP error on robots.txt** 24,221,293

**Robots.txt forbidden** 20,621,185

**Robots.txt loop** 612,160

**Robots.txt over 64KB** 183,894

**Total** 449,243,979

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067

**Connect** 7,606,109,371

**Valid replies** 7,437,281,300

**Non-text/html** 86,476,067
**Page-Level Analysis – URL Statistics**

**Admitted URLs**
- Total: 8,267,316,148
- With valid host: 8,058,635,011
- Blacklisted: 3,281,661
- Robot errors: 449,243,979
- No DNS: 208,681,137

**Non-text/html**
- Total: 86,476,067

**Connect**
- Total: 7,606,109,371
  - Valid replies: 7,437,281,300
  - Non-text/html: 86,476,067

**Full download**
- Total: 7,350,805,233
  - HTML 200 OK: 6,380,051,942

**Found URLs**
- Total: 387,605,655,905

**Network error type**

<table>
<thead>
<tr>
<th>Error Type</th>
<th>URLs Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect fail</td>
<td>124,752,717</td>
</tr>
<tr>
<td>Receive fail</td>
<td>36,534,906</td>
</tr>
<tr>
<td>Slow download</td>
<td>421,427</td>
</tr>
<tr>
<td>Page over 4MB</td>
<td>338,872</td>
</tr>
<tr>
<td>Send fail</td>
<td>9,365</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>162,057,287</td>
</tr>
</tbody>
</table>

**Connect**
- 97.8%

**Valid replies**
- 98.9%

**Full download**
- 86.8%

**Found URLs**
- 13.2%

**Network errors**
- Total: 162,057,287

**HTTP errors**
- 6,770,784

**Bad ext**
- 3,332,296,975

**Unique URLs**
- Total: 41,502,195,631

**Crawlable URLs**
- Total: 377,995,369,202

**Pass checks**
- 97.8%

**Fail checks**
- 2.1%

**URLs with valid host**
- 94.4%

**Blacklisted**
- 0.04%

**Robot errors**
- 5.6%

**No DNS**
- 2.5%

**Unique edges**
- Total: 310,361,986,596
**Page-Level Analysis – URL Statistics**

- **Admitted URLs**
  - Total: 8,267,316,148

- **Non-text/html**
  - Total: 86,476,067

- **Connect**
  - 7,606,109,371 (97.8%)

- **Valid replies**
  - 7,437,281,300 (98.9%)

- **Full download**
  - 7,350,805,233 (86.8%)

- **HTML 200 OK**
  - 6,380,051,942

- **HTTP errors**
  - 6,770,784

- **HTTP error type**
  - Bad HTTP response: 4,139,148
  - Decompression fail: 1,110,272
  - Bad HTTP status: 682,613
  - Invalid base URL: 593,941
  - Bad chunking: 242,858
  - Header over 64 KB: 1,952
  - **Total**: 6,770,784

- **HTTP errors (200%)**
  - Bad ext: 3,332,296,975

- **Network errors**
  - 162,057,287

- **Unique edges**
  - 310,361,986,596

- **Unique URLs**
  - 41,502,195,631

- **Crawlable URLs**
  - 377,995,369,202

- **URLs with valid host**
  - 8,058,635,011 (94.4%)

- **Bad ext**
  - 3,332,296,975

- **Robot errors**
  - 449,243,979

- **No DNS**
  - 208,681,137

- **Blacklisted**
  - 3,281,661

- **Robot errors**
  - 449,243,979

- **No DNS**
  - 208,681,137

- **Unique edges**
  - 310,361,986,596
Page-Level Analysis – URL Statistics

- **Admitted URLs**: 8,267,316,148
- **Unique URLs**: 41,502,195,631
- **Crawlable URLs**: 377,995,369,202

- **Connect**: 7,606,109,371
  - **Valid replies**: 7,437,281,300
    - **Non-text/html**: 86,476,067
      - **HTTP errors**: 6,770,784
        - **Network errors**: 162,057,287
          - **Bad ext**: 3,332,296,975
            - **Unknown ext**: 2,561,590,507
              - **Missing ext**: 78,938,236,200
                - **Dynamic/HTML**: 296,495,542,495
                  - **Total found URLs**: 394,619,023,142
                    - **Found URLs**: 387,605,655,905

- **Fail checks**: 13,291,356,96
  - **Pass checks**: 381,327,666,177

- **HTTP errors**: 970,753,291
  - **Valid replies**: 7,350,805,233
    - **Full download**: 7,437,281,300
      - **URLS with valid host**: 8,058,635,011
        - **Admitted URLs**: 8,267,316,148
          - **Blacklisted**: 3,281,661
            - **Robot errors**: 449,243,979
              - **No DNS**: 208,681,137
                - **Unique edges**: 310,361,986,596
                  - **Unique URLs**: 41,502,195,631
                    - **Crawlable URLs**: 377,995,369,202

Page-Level Analysis - A Few Notes

• Countering Spam
  – Did real-time PLD ranking on the current web graph
  – Treated 301/302 as regular links (processed through cycle)
  – Detected slow downloads (no data for 60 sec or takes more than 180 sec)
  – Detected infinite data stuffing and cut off after 4 MB

• Avoid non-HTMLs
  – Only processed pages with “Content-type: text/html” (86.5M discarded objects would take 346 TB in the worst case)
  – Transmitted “Accept: text/html” header field, but resulted in only 6.6% reduction, while extension filtering leads to 0.37% (not very effective!)

• The result is 8.3 KB per object
A Few Notes (2)

• URL Processing
  - Processed `a-href`, `frame-src` and `meta-refresh` tags. Did not follow `img` tags
  - Checked URLs for correctness and syntax
  - Used a black list of non-HTML extensions, resulted in 0.37% saving in bandwidth (note for future crawlers)

• Web graph
  - Constructed a web graph with 3 TB web graph with 310B edges and 41B nodes
  - Average crawl depth 12 (compare to 1.8 of ClueWeb09)
  - 60% of downloaded pages were dynamic (i.e. contains “?”)
Agenda

• Introduction
• Background
• Page-Level Analysis
• Server-Level Analysis
  – DNS and Robots
  – Bandwidth
• Internet Coverage
• Extrapolation
• Conclusion
Server-Level Analysis – DNS and Robots

DNS expiration

- DNS lookups: 297,184,517
- Hosts: 641,982,061
- PLDs: 89,652,630
- Queued URLs: 41,502,195,631

- Unique hosts: 260,113,628
- DNS found: 171,101,737
- Live hosts: 156,236,808
- Crawlable hosts: 146,158,035

- DNS error: 89,011,891
- Dead hosts: 14,864,929
- Robots problem: 10,078,773

- Name error: 65,241,643
- Refused: 241,555
- Reserved IP: 128,029

- Receive fail: 3,641,064
- Bad HTTP: 2,976,378

- Robot loop: 87,318
- Send fail: 229

- Robots 404: 73,776,294
- Robots 2xx: 72,381,741

- text/plain: 54,747,900
- No MIME: 17,513,363
- Other MIME: 120,478

- Invalid pkt: 513,556
- Other: 1,868

- Size > 64KB: 1,895

- 1.5 GB

- 96% of Crawlable hosts are Live hosts.
- 91% of Live hosts are correctly found.
- 96% of Queued URLs are valid.
- 66% of Unique hosts are found.

- DNS lookups: 297,184,517
- Hosts: 641,982,061
- PLDs: 89,652,630
- Queued URLs: 41,502,195,631

- Unique hosts: 260,113,628
- DNS found: 171,101,737
- Live hosts: 156,236,808
- Crawlable hosts: 146,158,035

- DNS error: 89,011,891
- Dead hosts: 14,864,929
- Robots problem: 10,078,773

- Name error: 65,241,643
- Refused: 241,555
- Reserved IP: 128,029

- Receive fail: 3,641,064
- Bad HTTP: 2,976,378

- Robot loop: 87,318
- Send fail: 229

- Robots 404: 73,776,294
- Robots 2xx: 72,381,741

- text/plain: 54,747,900
- No MIME: 17,513,363
- Other MIME: 120,478

- Invalid pkt: 513,556
- Other: 1,868

- Size > 64KB: 1,895

- 1.5 GB
## Server-Level Analysis - Bandwidth

### Outbound Traffic
- **DNS** 23 GB
- **Robots** 33 GB
- **HTTP GET** 1.8 TB
- **TCP/IP** 1.1 TB

### Inbound Traffic
- **DNS** 37 GB
- **Robots** 254 GB
- **HTTP error** 718 GB aborted
- **HTML 200 OK** 143 TB full d/l

### Server-Level Analysis - Bandwidth

- **Compressed Bytes**: 65,867,784
- **HTTP error bytes**: 2,258,814,609,583
- **HTTP header bytes**: 2,571,309,459,132
- **Full download bytes**: 143,589,123,572,029
- **TCP/IP header bytes**: 3,985,467,341,092

### Bandwidth
- **Outbound Traffic**: 1.8% of 83,929,337,155 bytes
- **Inbound Traffic**: 96% of 6,614,286,502,692 bytes

### Bandwidth Comparison
- **7 Mbps**: 4% out, 96% in
- **320 Mbps**: 95% out, 5% in

### HTTP GET Bandwidth
- **DNS** 23 GB
- **Robots** 33 GB
- **HTTP GET** 1.8 TB
- **TCP/IP** 1.1 TB

### HTML 200 OK Bandwidth
- **DNS** 37 GB
- **Robots** 254 GB
- **HTTP error** 718 GB aborted
- **HTML 200 OK** 143 TB full d/l

### TCP/IP Header Bandwidth
- **DNS** 5,329,096,697 bytes
- **Robots** 2,258,814,609,583 bytes
- **HTTP GET** 6,614,286,502,692 bytes
- **TCP/IP** 3,985,467,341,092 bytes
Agenda

• Introduction
• Background
• Page-Level Analysis
• Server-Level Analysis
• Internet Coverage
• Extrapolation
• Conclusion
Internet Coverage

• Can use different measures
  - Collection of crawled 200 OK pages
  - Constructed web graph size

• Not much available information in standardized fashion
  - Mercator uses img tags, while UbiCrawler removes frontiers
  - WebBase considers robots.txt as crawled page

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Crawled (HTML 200 OK)</th>
<th>Web graph</th>
<th>Host graph</th>
<th>PLD graph</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pages hosts PLDs TLDs</td>
<td>nodes edges</td>
<td>nodes edges</td>
<td>nodes edges</td>
</tr>
<tr>
<td>AltaVista [9]</td>
<td>– – – –</td>
<td>271M 2.1B</td>
<td>– –</td>
<td>– –</td>
</tr>
<tr>
<td>Polybot [36]</td>
<td>121M 5M – – –</td>
<td>– –</td>
<td>– –</td>
<td>– –</td>
</tr>
<tr>
<td>Google [6]</td>
<td>– – – –</td>
<td>1.3B 19.5B</td>
<td>12.8M 395M</td>
<td>– –</td>
</tr>
<tr>
<td>Mercator [10]</td>
<td>429M ~ 10M – – –</td>
<td>– 18.3B</td>
<td>– –</td>
<td>– –</td>
</tr>
<tr>
<td>WebFountain [20]</td>
<td>1B – – – –</td>
<td>4.75B 37B</td>
<td>19.7M 1.1B</td>
<td>– –</td>
</tr>
<tr>
<td>WebBase [16]</td>
<td>98M 51K – – –</td>
<td>– 4.2B</td>
<td>– –</td>
<td>– –</td>
</tr>
<tr>
<td>ClueWeb09 [19]</td>
<td>1B – – – – –</td>
<td>4.8B 7.9B</td>
<td>– –</td>
<td>– –</td>
</tr>
<tr>
<td>IRLbot</td>
<td>6.3B 117M 33M 256</td>
<td>41B 310B</td>
<td>641M 6.8B</td>
<td>89M 1.8B</td>
</tr>
<tr>
<td>UbiCrawler .uk [7]</td>
<td>105M 114K – – 1</td>
<td>105M 3.7B</td>
<td>114K –</td>
<td>– –</td>
</tr>
<tr>
<td>IRLbot .uk</td>
<td>197M 2.8M 1.2M 1</td>
<td>1.3B 9.5B</td>
<td>5M 54M</td>
<td>1.5M 18M</td>
</tr>
<tr>
<td>IRLbot .cn</td>
<td>209M 3.3M 539K 1</td>
<td>1.1B 11.9B</td>
<td>8.4M 103M</td>
<td>711K 19.7M</td>
</tr>
</tbody>
</table>
Internet Coverage – TLD Level

• A novel method of comparing crawls
  - Reveals crawler budget on different parts of the Internet

• Use site queries (i.e., “site:domain”) to obtain Google and Yahoo’s (now part of Bing) index size
  - In 1/2008, they contained 30B and 37B pages, respectively

<table>
<thead>
<tr>
<th>TLD</th>
<th>Google</th>
<th>Yahoo</th>
<th>IRLbot</th>
<th>WebBase</th>
<th>ClueWeb</th>
</tr>
</thead>
<tbody>
<tr>
<td>.com</td>
<td>46.7%</td>
<td>38.3%</td>
<td>43.3%</td>
<td>31.2%</td>
<td>54.8%</td>
</tr>
<tr>
<td>.net</td>
<td>6.9%</td>
<td>7.7%</td>
<td>6.9%</td>
<td>2.2%</td>
<td>6.7%</td>
</tr>
<tr>
<td>.de</td>
<td>6.6%</td>
<td>6.8%</td>
<td>7.4%</td>
<td>3.8%</td>
<td>3.8%</td>
</tr>
<tr>
<td>.org</td>
<td>5.5%</td>
<td>6.3%</td>
<td>6.6%</td>
<td>17.8%</td>
<td>6.6%</td>
</tr>
<tr>
<td>.cn</td>
<td>3.7%</td>
<td>4.6%</td>
<td>3.3%</td>
<td>0.2%</td>
<td>5.6%</td>
</tr>
<tr>
<td>.jp</td>
<td>3.4%</td>
<td>5.2%</td>
<td>1.2%</td>
<td>1.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>.ru</td>
<td>2.3%</td>
<td>4.6%</td>
<td>3.3%</td>
<td>0.6%</td>
<td>0.1%</td>
</tr>
<tr>
<td>.uk</td>
<td>2.2%</td>
<td>3.0%</td>
<td>3.1%</td>
<td>4.9%</td>
<td>1.7%</td>
</tr>
<tr>
<td>.pl</td>
<td>1.6%</td>
<td>1.9%</td>
<td>1.3%</td>
<td>0.2%</td>
<td>0.3%</td>
</tr>
<tr>
<td>.nl</td>
<td>1.4%</td>
<td>1.4%</td>
<td>2.0%</td>
<td>0.5%</td>
<td>0.1%</td>
</tr>
<tr>
<td>TLDs</td>
<td>255</td>
<td>256</td>
<td>256</td>
<td>174</td>
<td>254</td>
</tr>
</tbody>
</table>
TLD Coverage – Google Order

(a) Yahoo (all)

(b) IRLbot (all)

(c) Yahoo (top 40)

(d) IRLbot (top 40)

.info (#15)

.edu (#12)

.gov (#24)
Agenda

• Introduction
• Background
• Page-Level Analysis
• Server-Level Analysis
• Internet Coverage
• Extrapolation
• Conclusion
Extrapolation

• Assume that the crawl is a stochastic process \( \{(X_t, Y_t)\} \) on the Internet, a web graph \( G(V,E) \), where the process terminates at \( t = N \leq |E| \) edges.

• Define \( p(t) \) as the probability that URL \( Y_t \) has not been seen before.

• Objective: In a larger crawl, can we estimate the number of
  - Unique URLs \( L_N \)
  - Crawled pages \( C_N = L_N \frac{\# \text{ of links/page}}{} \)
Extrapolation (2)

• Assume that the reference crawl (IRLbot) has \( K \) links, \( U \) unique links. The unknown crawl (e.g., Google) has \( N \) links \((r=N/K)\). What is \( L_N \) and \( C_N \)?

• Also assume \( z=t/K \) and a new function \( \tilde{p}(z) = p(zK) \). Thus, the unknown crawler has:

\[
E[L_N] = K \int_0^r \tilde{p}(z) dz = U + K \int_1^r \tilde{p}(z) dz
\]

• With Pareto fit (i.e., \( \tilde{p}(z) = \beta z^{-\alpha} \)), we get:

\[
E[L_N] \approx U + \frac{K \beta (r^{1-\alpha} - 1)}{1 - \alpha}
\]
Extrapolation - Results

<table>
<thead>
<tr>
<th>Crawl</th>
<th>Ratio $r$</th>
<th>Crawled Links $N$</th>
<th>Crawled Pages $C_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRLbot 2007</td>
<td>1</td>
<td>394B</td>
<td>6.3B</td>
</tr>
<tr>
<td>Google 2008 ($E[L_N] = 1T$)</td>
<td>40</td>
<td>12T</td>
<td>256B</td>
</tr>
<tr>
<td>Google 2012 ($E[L_N] = 30T$)</td>
<td>1,981</td>
<td>592T</td>
<td>12T</td>
</tr>
</tbody>
</table>

Using 20B pages/day (@41 Gbps), takes 50 months of crawling

- How about Hots/PLD level graphs in Google 2012?
  - With $r=1981$, Google has 5.2B unique hosts (IRLbot has 641M), and 90.6M unique PLDs (IRLbot has 89M)
Conclusion

• Presented IRLbot implementation and experiment in detail
  – Discussed the impact of various design choices
  – Provided guidelines for future crawlers
  – Exposed weird/effective spamming techniques
• Developed new methods for capturing crawl coverage
• Outlined a simple extrapolation mechanism to infer proprietary and undocumented crawls
  – A simple model for crawl growth rate
Thank you!
Questions?