# Web Crawling

Introduction to Information Retrieval
Informatics 141 / CS 121
Donald J. Patterson

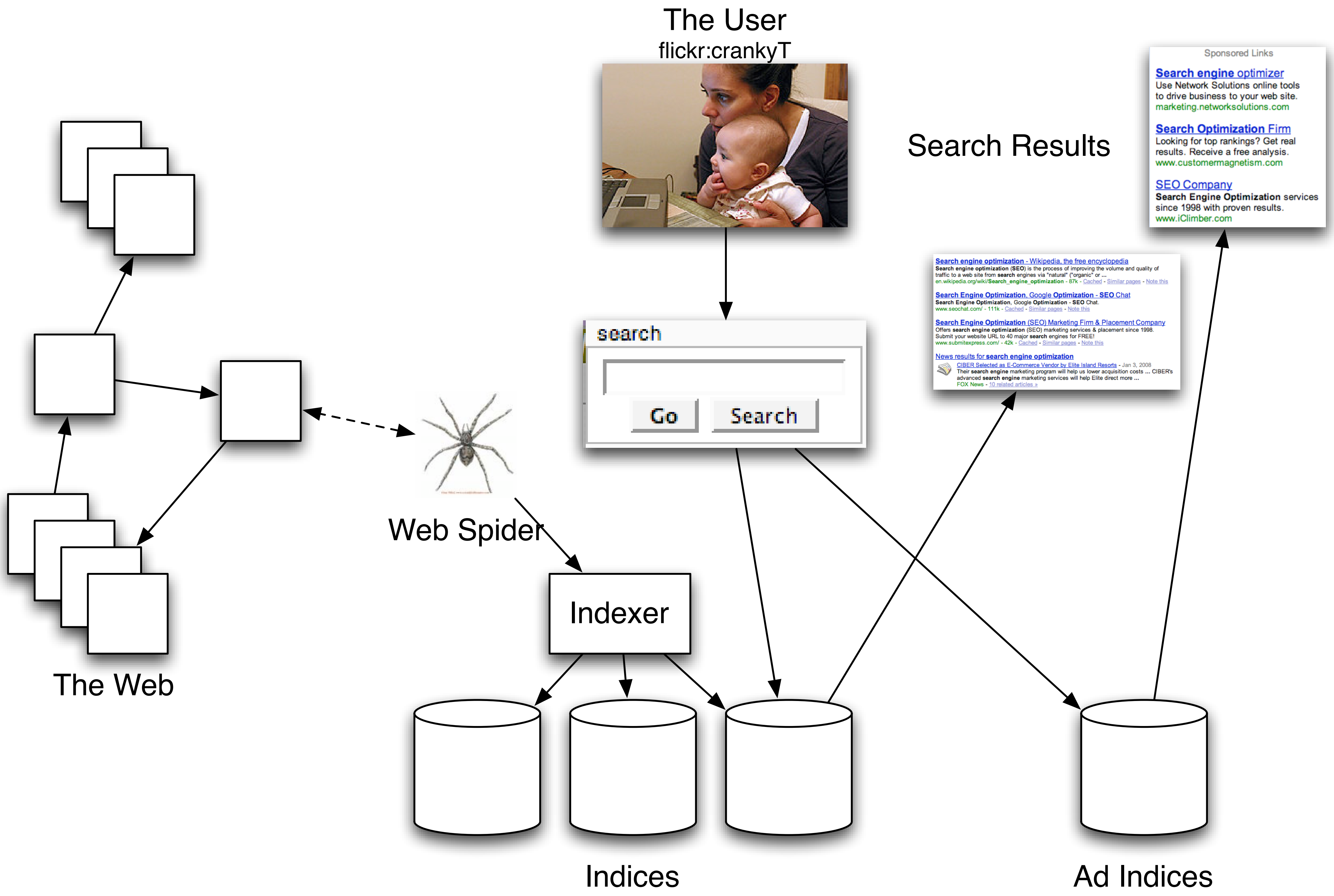
Content adapted from Hinrich Schütze
http://www.informationretrieval.org

# Web Crawling Outline

## Overview

- Introduction
- URL Frontier
- Robust Crawling
    - DNS

# Introduction



Sponsored Links

Search engine optimizer
Use Network Solutions online tools to drive business to your web site.
marketing.networksolutions.com

Search Optimization Firm
Looking for top rankings? Get real results. Receive a free analysis.
www.customermagnetism.com

SEO Company
**Search Engine Optimization** services since 1998 with proven results.
www.iClimber.com

Search engine optimization - Wikipedia, the free encyclopedia
**Search engine optimization** (**SEO**) is the process of improving the volume and quality of traffic to a web site from **search** engines via "natural" ("organic" or ...
en.wikipedia.org/wiki/**Search_engine_optimization** - 87k - Cached - Similar pages - Note this

Search Engine Optimization, Google Optimization - SEO Chat
**Search Engine Optimization**, Google **Optimization** - **SEO** Chat.
www.seochat.com/ - 111k - Cached - Similar pages - Note this

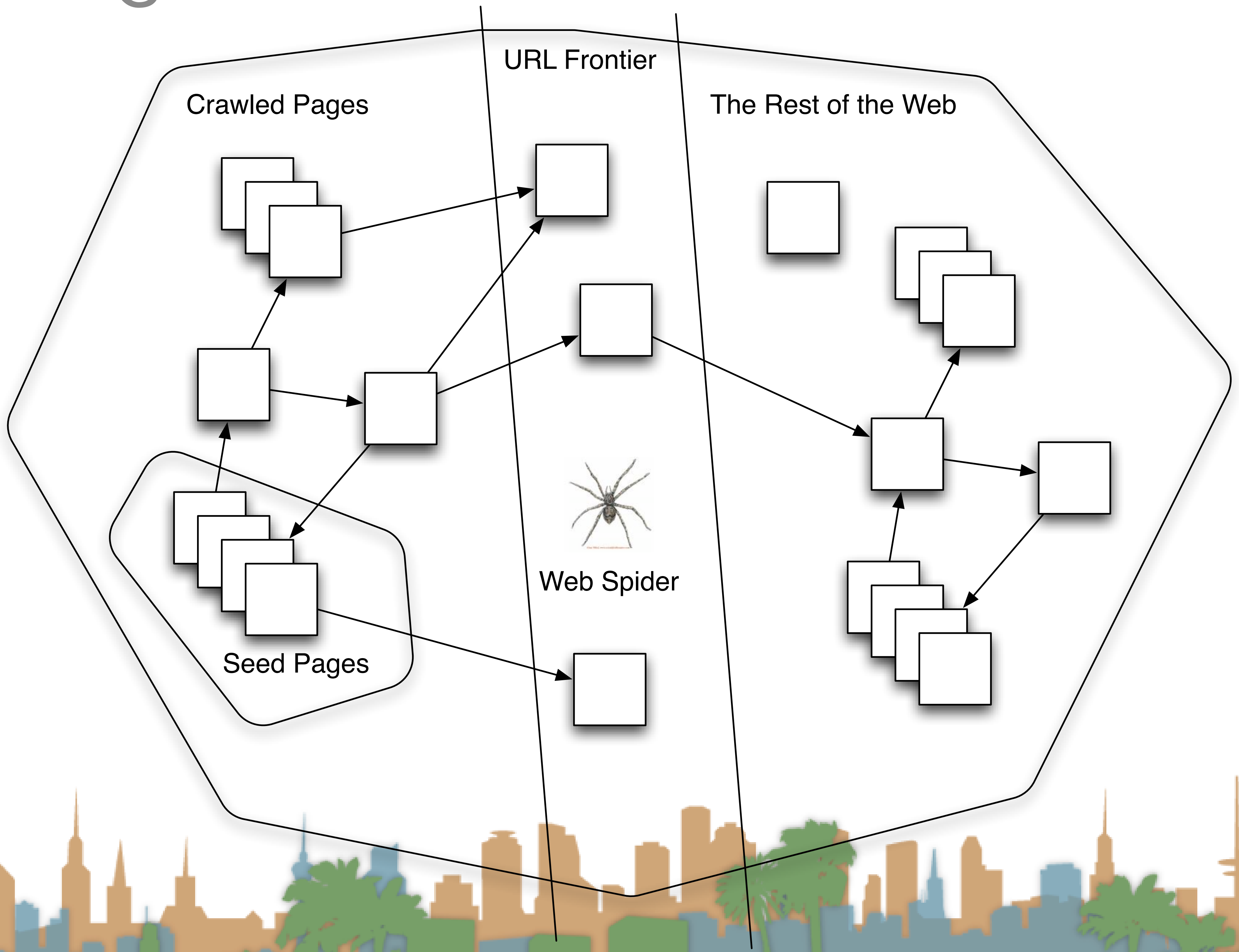
Search Engine Optimization (SEO) Marketing Firm & Placement Company
Offers **search engine optimization** (SEO) marketing services & placement since 1998. Submit your website URL to 40 major **search** engines for FREE!
www.submitexpress.com/ - 42k - Cached - Similar pages - Note this

News results for **search engine optimization**
CIBER Selected as E-Commerce Vendor by Elite Island Resorts - Jan 3, 2008
Their **search engine** marketing program will help us lower acquisition costs ... CIBER's advanced **search engine** marketing services will help Elite direct more ...
FOX News - 10 related articles »

## The basic crawl algorithm

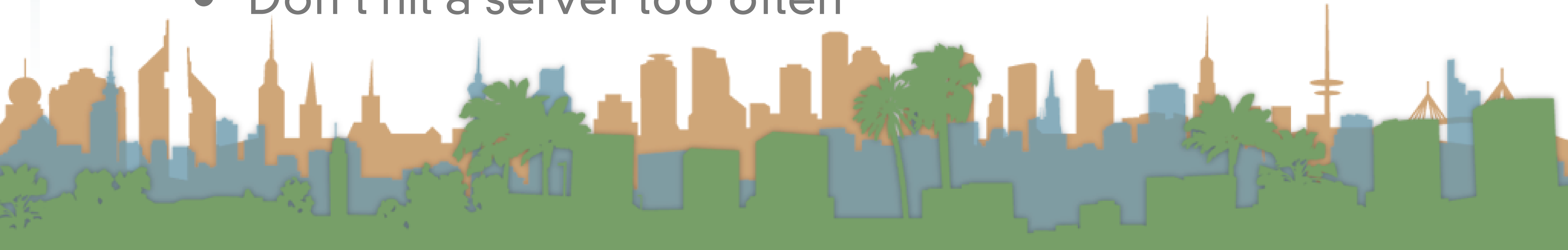
- Initialize a queue of URLs (“seed” URLs)
- Repeat
    - Remove a URL from the queue
    - Fetch associated page
    - Parse and analyze page
    - Store representation of page
    - Extract URLs from page and add to queue

# Crawling the web



# Basic Algorithm is not reality...

- Real web crawling requires multiple machines
    - All steps distributed on different computers
- Even Non-Adversarial pages pose problems
    - Latency and bandwidth to remote servers vary
    - Webmasters have opinions about crawling their turf
        - How “deep” in a URL should you go?
    - Site mirrors and duplicate pages
- Politeness
    - Don’t hit a server too often



# Basic Algorithm is not reality...

- Adversarial Web Pages
  - Spam Pages
  - Spider Traps

# Minimum Characteristics for a Web Crawler

- Be Polite:
  - Respect implicit and explicit terms on website
  - Crawl pages you're allowed to
  - Respect "robots.txt" (more on this coming up)
- Be Robust
  - Handle traps and spam gracefully

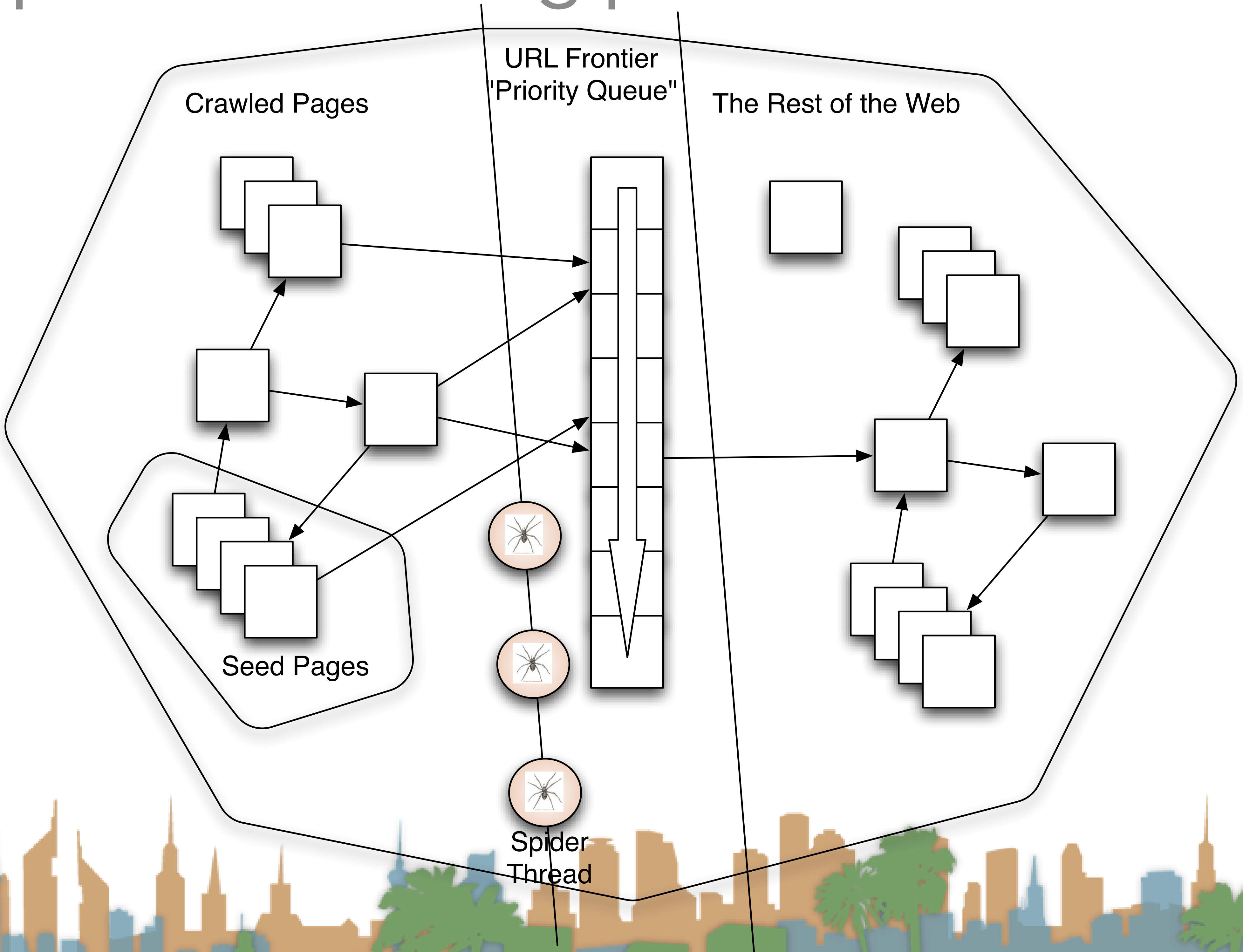
# Desired Characteristics for a Web Crawler

- Be a distributed systems
    - Run on multiple machines
- Be scalable
    - Adding more machines allows you to crawl faster
- Be Efficient
    - Fully utilize available processing and bandwidth
- Focus on “Quality” Pages
    - Crawl good information first

## Desired Characteristics for a Web Crawler

- Support Continuous Operation
  - Fetch fresh copies of previously crawled pages
- Be Extensible
  - Be able to adapt to new data formats, protocols, etc.
  - Today it’s AJAX, tomorrow it’s SilverLight, then....

# Updated Crawling picture



# URL Frontier

- Frontier Queue might have multiple pages from the same host
  - These need to be load balanced (“politeness”)
- All crawl threads should be kept busy

# URL Frontier

## Politeness?

- It is easy enough for a website to block a crawler
- Explicit Politeness
  - "Robots Exclusion Standard"
    - Defined by a "robots.txt" file maintained by a webmaster
    - What portions of the site can be crawled.
      - Irrelevant, private or other data excluded.
    - Voluntary compliance by crawlers.
    - Based on regular expression matching

# URL Frontier

## Politeness?

- Explicit Politeness
  - “Sitemaps”
    - Introduced by Google, but open standard
    - XML based
    - Allows webmasters to specify:
      - Location of pages (URL islands)
      - Importance of pages
      - Update frequency of pages
    - Sitemap location listed in robots.txt

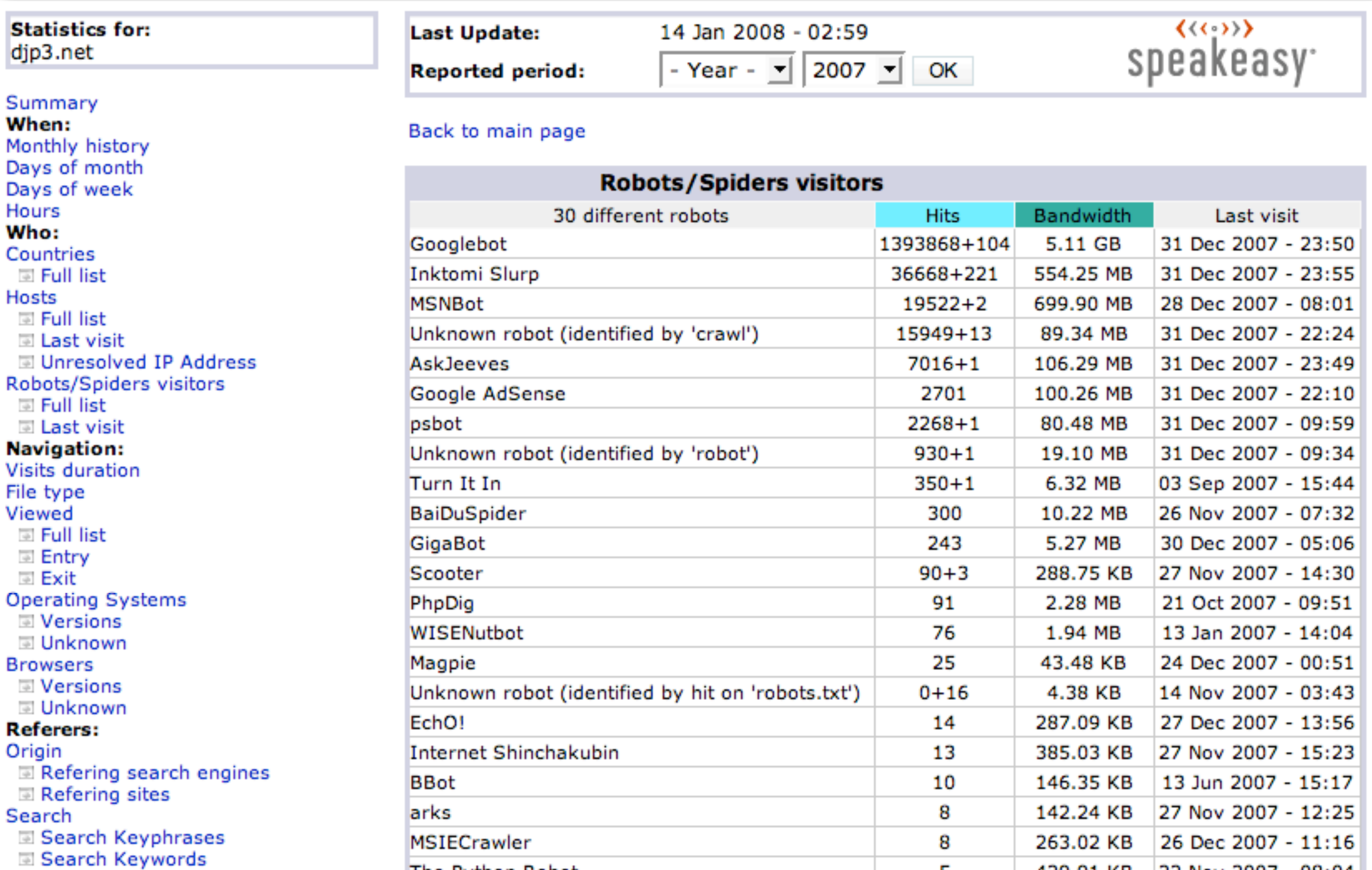
# URL Frontier

## Politeness?

- Implicit Politeness
  - Even without specification avoid hitting any site too often
  - It costs bandwidth and computing resources for host.

# URL Frontier

## Politeness?



**Statistics for:**
djp3.net

[Summary]
**When:**
Monthly history
Days of month
Days of week
Hours
**Who:**
Countries
- Full list

Hosts
- Full list
- Last visit
- Unresolved IP Address

Robots/Spiders visitors
- Full list
- Last visit

**Navigation:**
Visits duration
File type
Viewed
- Full list
- Entry
- Exit

Operating Systems
- Versions
- Unknown

Browsers
- Versions
- Unknown

**Referers:**
Origin
- Refering search engines
- Refering sites

Search
- Search Keyphrases
- Search Keywords

**Last Update:** 14 Jan 2008 - 02:59
**Reported period:** - Year - 2007 OK

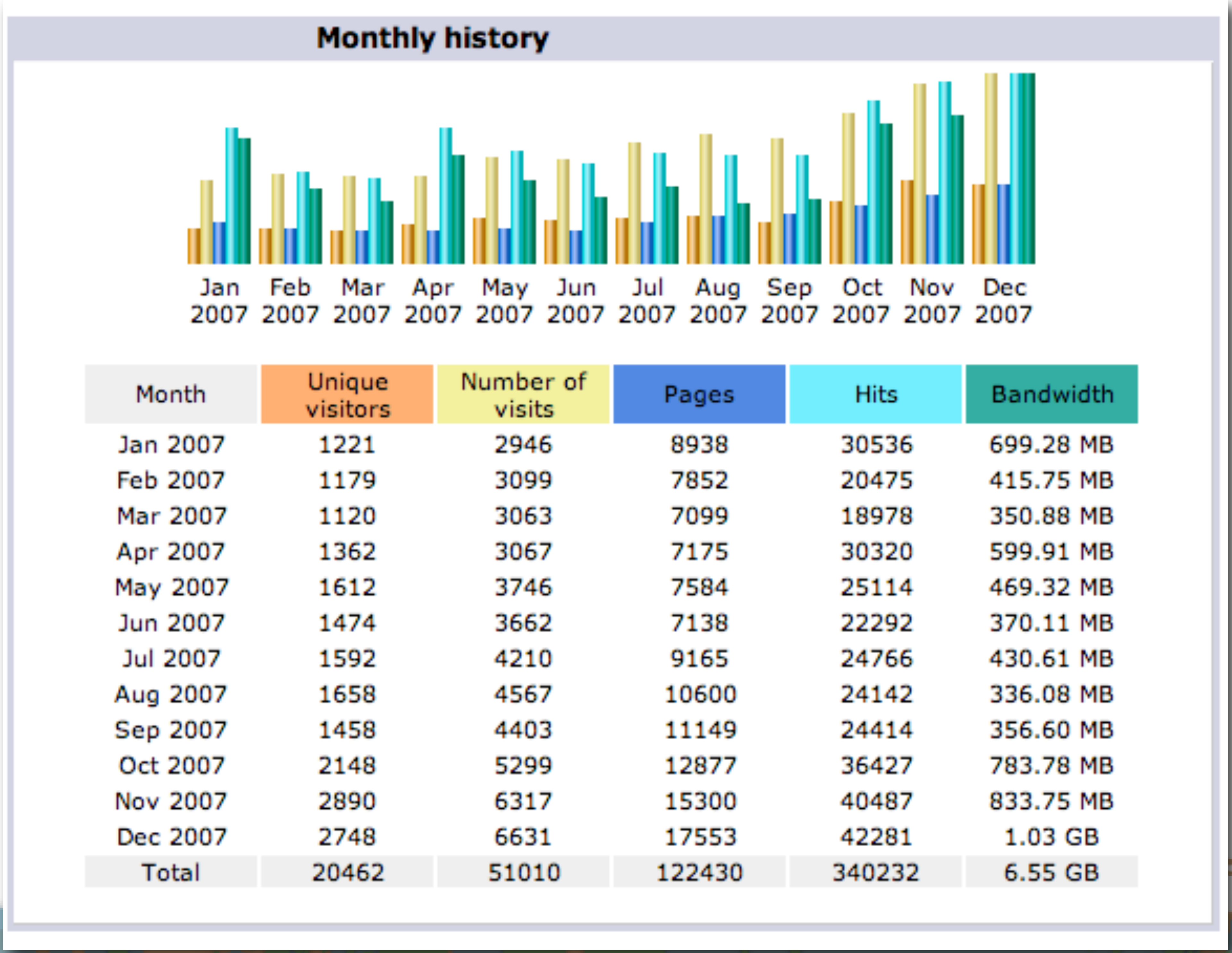
Back to main page

**Robots/Spiders visitors**

| 30 different robots | Hits | Bandwidth | Last visit |
|---|---|---|---|
| Googlebot | 1393868+104 | 5.11 GB | 31 Dec 2007 - 23:50 |
| Inktomi Slurp | 36668+221 | 554.25 MB | 31 Dec 2007 - 23:55 |
| MSNBot | 19522+2 | 699.90 MB | 28 Dec 2007 - 08:01 |
| Unknown robot (identified by 'crawl') | 15949+13 | 89.34 MB | 31 Dec 2007 - 22:24 |
| AskJeeves | 7016+1 | 106.29 MB | 31 Dec 2007 - 23:49 |
| Google AdSense | 2701 | 100.26 MB | 31 Dec 2007 - 22:10 |
| psbot | 2268+1 | 80.48 MB | 31 Dec 2007 - 09:59 |
| Unknown robot (identified by 'robot') | 930+1 | 19.10 MB | 31 Dec 2007 - 09:34 |
| Turn It In | 350+1 | 6.32 MB | 03 Sep 2007 - 15:44 |
| BaiDuSpider | 300 | 10.22 MB | 26 Nov 2007 - 07:32 |
| GigaBot | 243 | 5.27 MB | 30 Dec 2007 - 05:06 |
| Scooter | 90+3 | 288.75 KB | 27 Nov 2007 - 14:30 |
| PhpDig | 91 | 2.28 MB | 21 Oct 2007 - 09:51 |
| WISENutbot | 76 | 1.94 MB | 13 Jan 2007 - 14:04 |
| Magpie | 25 | 43.48 KB | 24 Dec 2007 - 00:51 |
| Unknown robot (identified by hit on 'robots.txt') | 0+16 | 4.38 KB | 14 Nov 2007 - 03:43 |
| EchO! | 14 | 287.09 KB | 27 Dec 2007 - 13:56 |
| Internet Shinchakubin | 13 | 385.03 KB | 27 Nov 2007 - 15:23 |
| BBot | 10 | 146.35 KB | 13 Jun 2007 - 15:17 |
| arks | 8 | 142.24 KB | 27 Nov 2007 - 12:25 |
| MSIECrawler | 8 | 263.02 KB | 26 Dec 2007 - 11:16 |
| The Python Robot | 5 | [illegible] | [illegible] |

# URL Frontier

## Politeness?



**Monthly history**

| Month | Unique visitors | Number of visits | Pages | Hits | Bandwidth |
|---|---|---|---|---|---|
| Jan 2007 | 1221 | 2946 | 8938 | 30536 | 699.28 MB |
| Feb 2007 | 1179 | 3099 | 7852 | 20475 | 415.75 MB |
| Mar 2007 | 1120 | 3063 | 7099 | 18978 | 350.88 MB |
| Apr 2007 | 1362 | 3067 | 7175 | 30320 | 599.91 MB |
| May 2007 | 1612 | 3746 | 7584 | 25114 | 469.32 MB |
| Jun 2007 | 1474 | 3662 | 7138 | 22292 | 370.11 MB |
| Jul 2007 | 1592 | 4210 | 9165 | 24766 | 430.61 MB |
| Aug 2007 | 1658 | 4567 | 10600 | 24142 | 336.08 MB |
| Sep 2007 | 1458 | 4403 | 11149 | 24414 | 356.60 MB |
| Oct 2007 | 2148 | 5299 | 12877 | 36427 | 783.78 MB |
| Nov 2007 | 2890 | 6317 | 15300 | 40487 | 833.75 MB |
| Dec 2007 | 2748 | 6631 | 17553 | 42281 | 1.03 GB |
| Total | 20462 | 51010 | 122430 | 340232 | 6.55 GB |