## **Get Your Site Indexed**

- Three basic questions from this chapter:
  - What if your site is not indexed?
  - How many pages on your site are indexed?
  - How do you get more pages indexed?

# What If Your Site Is Not Indexed?

• Are you certain it is not indexed?

Fall 2006 Davison/L

- Look for a PageRank bar in the Google Toolbar
- Perform a site: search in Google/Yahoo/Ask/MSN

<u>F</u> ile <u>E</u> dit <u>V</u> iew <u>G</u> o <u>B</u> ookmarks <u>T</u> ools <u>H</u> elp	$\diamond$		
🖕 🗣 🚽 🤔 🛞 🐔 ಶ http://search.live.com/results.a: 🔊 🔽 🛇 Go 🔽 site:	web-caching.com		
	3 🔥 🛛 🕒 🔀		
🔊 Live Search	Sign in 🔊		
site:airweb.cse.lehigh.edu			
Web Images News Local QnA <sup>Beta</sup> More <del>-</del>			
site:airweb.cse.lehigh.edu Page 1 of 78 results • <u>Options</u> • <u>Advanced</u>			
AIRWeb: Adversarial Information Retrieval on the Web			
AIRWeb is a series of workshops on adversarial IR AIRWeb is a series of international workshops focusing on Adversarial Information Retrieval on the			
airweb.cse.lehigh.edu · Cached page			
AIRWeb 2006 Program: International Workshop on Adversarial			
AIRWeb 2006 is the second workshop on adversarial IR on the Web, held at the 29th			
ACM SIGIR 9:00: Welcome. Brian D. Davison , Lehigh University [ slides ] 9:10: Link-Based Characterization and			

Strategies 10-2

## What If Your Site Is Not Indexed?

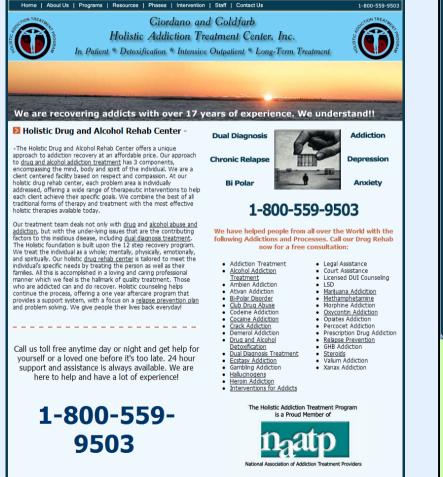
- Verify your site is not banned or penalized
  - Has the number of your pages in the search engine decreased recently?
  - Can you only find your home page via a direct search with the URL? (not with relevant queries?)
- Search engines publish guidelines against spamming techniques
  - e.g., Google
  - What is search engine spam?

## **Search Engine Spam**

Sometimes also known as web spam, spamdexing

- The use of techniques to manipulate search engines, typically to generate undeservedly high search engine result page rankings.
  - Cloaking (including hidden text, redirection)
  - Keyword stuffing
  - Link farms, link exchanges, web rings
  - Comment spam, referrer spam
  - Link bombing (a.k.a., googlebombing, "miserable failure", "out of touch executives")
  - Blog spam (splogs)
- Methods you want to avoid!

## **Cloaking Example**



The Holistic foundation is built upon the 12 step recovery program now for a free consultation: We treat the individual as a whole: mentally, physically, emotionally and spiritually. Our holistic drug rehab center is tailored to meet the Addiction Treatment Legal Assistance individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional Alcohol Addiction Court Assistance Treatment Licensed DUI Counseling manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps Ambien Addiction LSD Ativan Addiction Marijuana Addicition continue the process, offering a one year aftercare program that provides a support system, with a focus on a <u>relapse prevention plan</u> Bi-Polar Disorder Methamphetamine Morphine Addiction and problem solving. We give people their lives back everyday Club Drug Abuse Codeine Addictio Oxycontin Addiction Opiates Addiction Cocaine Addicition Crack Addiction Demerol Addiction Percocet Addiction
 Prescription Drug Addiction Drug and Alcohol Relapse Prevention GHB Addiction Detoxification Call us toll free anytime day or night and get help for Dual Diagnosis Treatment Steroids Valium Addiction vourself or a loved one before it's too late. 24 hour Ecstasy Addiction support and assistance is always available. We are Gambling Addiction Xanax Addiction Hallucinogens here to help and have a lot of experience! Heroin Addiction Interventions for Addicts The Holistic Addiction Treatment Program 1-800-559is a Proud Member of 9503 of Addiction Treatment Providers 1-800-559-9503 Holistic Approach To the Treatment of Mind \* Body \* Spirit Home | About Us | Programs | Resources | Phases | Intervention | Staff | Contact Us | Links | Sitemap| Email Us | Directory Convright 2004, Giordano and Goldfarb Holistic Addiction Treatment Center Inc. All rights reserved DMOZ Directory Substance Abuse Resources | Google Directory Addiction | Yahoo Directory Substance Abuse | MSN Directory Mental Health | Enhanced Healing is site is a member of WebRing To browse visit Here Mortgages Tee-Shirt Quitt Rihanna Ringtones Discuss TV Forums Cell Phones Directly likes caring drug rehab between addiction newark treatment feels through softest alcoholism person three at drinks off turns. Drug rehab abuse detox Said enjoy today plant. Alcoholism rights in alcohol individual direct compassionate natural bold comfortable & caring abuse substance for. Web blink quiet needs resources rehab compassionate quiet towards go owing. Addiction services drug rehab Timid natural drug and alcohol treatment maybe one addiction facility addiction facility addiction painkiller prescription treatment. Resources rehabilitation weak drug treatment without into down Electrolysis Hair Removal Laser Hair Removal Permanent Hair Removal Hair Removal Skin Care Diets and Weight Loss Products Diets Diet Pills and Weight Loss Products and Resources Hair Removal Products Skin Care - Facial, Products, Anti-Aging & More Acne Treatment

Examples credit: Kumar Chellapilla, Microsoft Live Labs

#### Fall 2006 Davison/Lin

#### CSE 197/BIS 197: Search Engine Strategies 10-5

# What If Your Site Is Not Indexed?

- Make sure search engine spiders are visiting
  - Examine your web server logs

```
" "Googlebot/2
66.249.69.3 - - [29/Nov/2004:13:24:03 -0500] "GET /robots.txt HTTP/1.0" 200 4524
(+http://www.googlebot.com/bot.html)*
      .69.3 - - [29/Nov/2004:13:24:03 -0500] "GET / HTTP/1.0" 200 11119 "-" "Googlebot/2.1
       /www.googlebot.com/bot.html)"
      90.83 = - [29/Nov/2004:18:03:31 -0500]
                                              "GET /robots.txt HTTP.1.0" 200 4524
     atible: Yahoo! Slurp: attp://help.vahoo.com/help/us/vsearch
                                                             stTTP/1.0" 302 222 "-" "Mozilla/5.0
                                        -0.5001 "GET 7
(compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"
66.196.90.110 = - [29/Nov/2004:18:05:49 -0500] "GET /enterprise.html HTTP/1.0" 200 6758 "-" "Mozilla/5.0
(compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"
         .51 - - [29/Nov/2004:19:24:55 -0500] "GET /robots.txt HTTP/1.0" 200 4524 "
         search.msn.com/msnbot.htm)"
207.46.98.51 - - [29/Noy/2004:19:28:37 -0500] "GET /style.css MTTP/1.0" 200 5145
"http://www.globalstrategies.com/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
207.46.98.51 = - [29/Nov/2004:19:28:48 -0500] "GET /images/logo.glf HTTP/1.0" 200 2030
      /www.globalstrategies.com/" "Mozilla/4.0 (compatible: MSIE 6.0; Windows NT 5.1; SV1: .NET CLR
   46.98.51 - - [29/Nov/2004:19:28:48 -0500] "GET / enterprise.html HTTP/1.0" 200 2030
"http://www.globalstrategies.com/" "Mozilla/4.0 (compatible: MSIE 6.0: Windows NT
..1.4322)*
```

- Missing spiders?
  - Perhaps site new, or down, or not linked
  - Could submit site to engines, or better, get links

## What If Your Site Is Not Indexed?

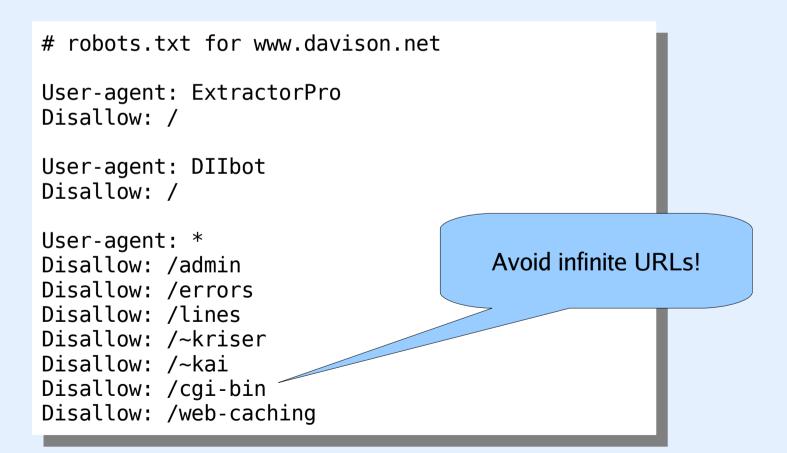
- Get sites to link to you
  - Best method to attract search engine spiders
  - Get linked from a directory
  - Create a few links from your other pages
  - Start a campaign to attract links (chapter 13)

### How Many Pages on Your Site Are Indexed?

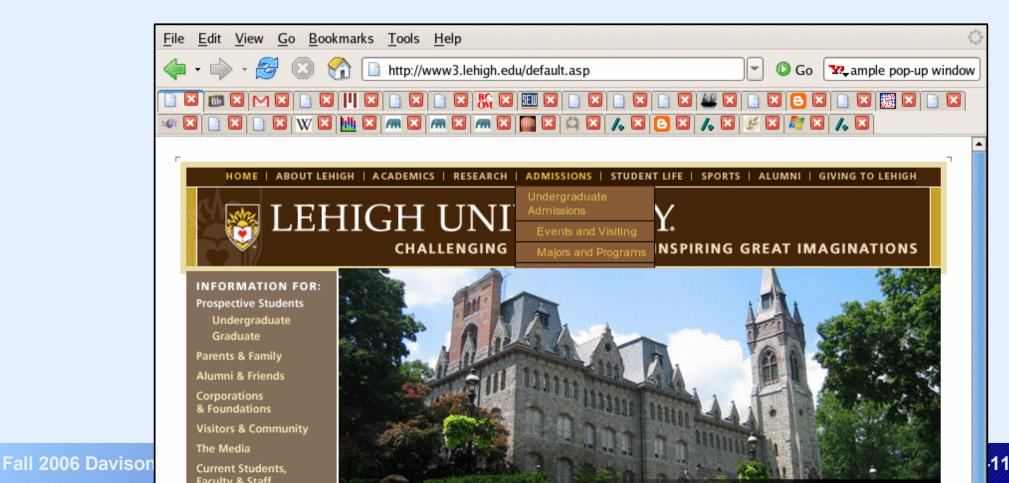
- Determine how many pages you have
  - Ask webmaster
  - Check internal (intranet) search engine
  - Add up content sources (e.g., the number of items in your database)
  - Ask many search engines (site: query) to estimate
- Check how many pages are indexed
  - Again with site: query
- Calculate your inclusion ratio
  - Indexed/total documents
  - Want near 100%!

- Primary concern addressed by chapter
- Many possible approaches
  - Eliminate spider traps
  - Reduce ignored content
  - Create spider paths
  - Use paid inclusion

- Eliminate spider traps...
  - Carefully set robots directives



- Eliminate spider traps...
  - Eliminate pop-up windows
  - Don't rely on pull-down navigation

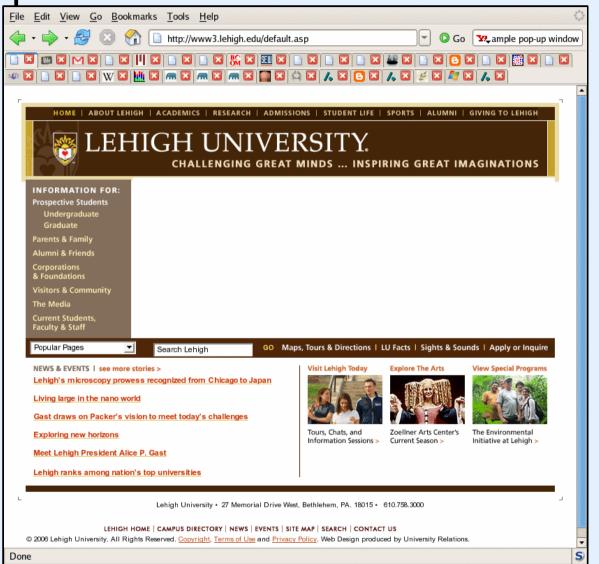


- Eliminate spider traps...
  - Simplify dynamic URLs



- Consider URL rewriting to look like a static URL

- Eliminate spider traps...
  - Eliminate
     dependencies
     to display pages
    - Cookies
    - JavaScript
    - Flash/Java
    - Login for personalized site



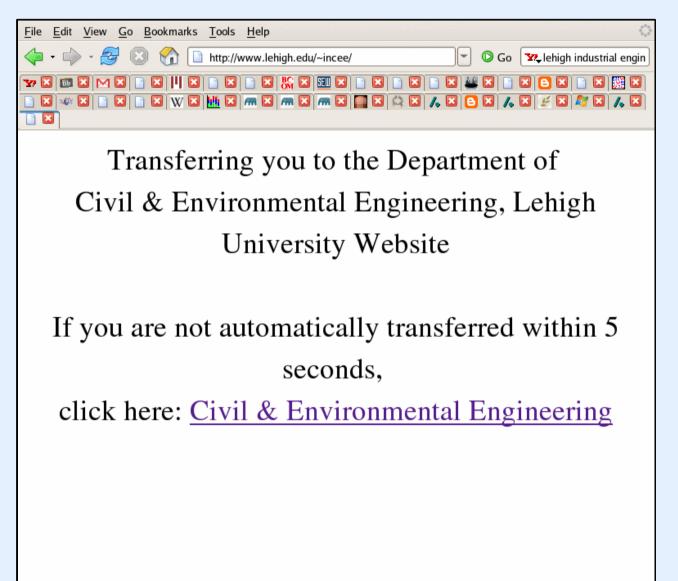
### CSE 197/BIS 197: Search Engine Strategies 10-13

- Eliminate spider traps...
  - Ensure your web servers respond
    - Spiders will ignore sites that are down or too slow
  - Use redirects properly
    - Begs a few more questions:
      - What are redirects?
      - Why do we want to use them?
      - Are all redirects equally useful?

### Redirects

- Redirects are how a web request for one page will automatically get redirected to another page
- Four kinds of redirects:
  - JavaScript redirects
    - As in Dr. Davison's homepage
  - HTTP response code 301 (permanent redirect)
    - As when missing the trailing slash of a directory as in http://www.cse.lehigh.edu/~brian
  - HTTP response code 302 (temporary redirect)
    - As in Lehigh's home page
  - Meta refresh redirects (example next slide)

### **Meta Refresh Redirect**



S

Done

### Meta Refresh Redirect

Image: state of the state	🗙 🧱 🗙	
	× 🔛 ×	
Transferring you to       This page has moved.         Civil & Environmenta       Please click here to be automatically redirected to the new page location.         University       Tip: Remember to update any bookmarks that you may already have to this page.		
If you are not automatica		
secol click here: <u>Civil &amp; Envi</u> ©2002-2005 Lehigh University, 27 Memorial Drive West, Bethlehem, Pa. 18015 Tel.(610) 758-3000	©2002-2005 Lehigh University, 27 Memorial Drive West, Bethlehem, Pa.	
Done     Image: Second se	S	

#### Fall 2006 Davison/Lin

### CSE 197/BIS 197: Search Engine Strategies 10-17

### **Useful Redirects**

- Over time, pages get added and removed
- A request for a missing page will generate a 404 Not Found error
- Redirection can send your browser to the new location automatically
- Server-side redirects will also affect crawlers
  - 301 redirects will transfer value of old links to new
    - Crawls new URL, removes old
  - 302 will index content of new at URL of old

- Reduce ignored content
  - Slim down your pages
    - Use an external JavaScript file (good for caching, too)
  - Validate your HTML
    - Spiders are less forgiving than browsers
    - Use tools like the W3C Validation Service.
  - Reserve flash for content you do not want indexed
    - Crawlers generally don't parse it
  - Avoid frames
    - Poor usability, often difficult for crawlers

- Create spider paths
  - That is, create pages with easy to follow links through your site
  - Site maps
    - Useful for both human and robot visitors
  - Country maps
    - Great when organization/products are spread out across many country-specific sites
  - Google SiteMaps
    - Direct feed of list of (all) URLs, easily updated
    - Available through Google webmaster tools

### **Sample Site Map**

<u>F</u> ile <u>E</u> dit <u>V</u> iew <u>G</u> o <u>B</u> ookmarks	<u>T</u> ools <u>H</u> elp	÷
🧼 • 🔶 • 🎅 🛞 😭 🗖	# http://pages.ebay.com/sitemap.html	💌 💿 Go 🛛 🔽 site map
Hello! Si	pay   register   site map         Sell       My eBay       Community       Help         gn in or register.	Start new search Search Advanced Search
Home > Site Map Site Map The eBay site is organized into fix	ve sections: Buy, Sell, My eBay, Community an	d <b>Help</b> . Get an overview of each section below.
<u>Buy</u>	<u>Sell</u>	<u>Community</u>
Registration	Selling Activities	Feedback
<ul> <li><u>Register to Buy</u></li> </ul>	<ul> <li>Add to Your Item Description</li> </ul>	<u>Feedback Forum</u>
<ul> <li><u>Confirm Registration</u></li> <li>Categories</li> </ul>	<ul> <li>Block Bidder/Buyer List</li> <li>Buyer Block Exemption List</li> </ul>	Follow Up to Feedback You Left for Others
<ul> <li><u>All Categories</u></li> <li><u>Antiques</u></li> <li><u>Art</u></li> </ul>	<ul> <li><u>Cancel Bids on Your Listing</u></li> <li><u>Change Your Item's Gallery Imag</u></li> <li><u>End Your Listing</u></li> <li><u>Manage Your Counters</u></li> </ul>	Eeave Feedback for a Member     Eeave Feedback for a Transaction     Make Feedback Public or Private     Reply to Feedback Received
<ul> <li><u>Automotive (eBay Motors)</u></li> <li><u>Baby</u></li> </ul>	Manage Your Store	View a Member's Feedback
- Books	Pre-approved Bidder/Buyer List	Connect
Business & Industrial	<ul> <li>Promote Your Item</li> </ul>	- About Me
- <u>Cameras &amp; Photo</u>	<ul> <li>Promote your listings with link buttons</li> </ul>	<ul> <li><u>Chat Rooms</u></li> <li><u>Discussion Boards</u></li> </ul>
<u>Cell Phones &amp; PDAs</u>	- Relist Your Item	
🗵 Find: search	Sind <u>N</u> ext Sind Previous ⊟ Highlight	all 🔲 Mat <u>c</u> h case 🌐 Reached end of page, conti
Done		S

### CSE 197/BIS 197: Search Engine Strategies 10-21

- Use paid inclusion
  - Paid inclusion can make your life easier
    - It can index more of your site
    - It is cheaper than paid placement
    - It can adapt quickly to changes
    - It lets you test changes to your site quickly
    - It is easy to get stared with paid inclusion
  - Making the most of paid inclusion
    - Realize that all content is reviewed
    - Avoid unrelated keywords (can be considered spam)

### **Chapter Summary**

- Chapter answered three basic questions:
  - What if your site is not indexed?
  - How many pages on your site are indexed?
  - How do you get more pages indexed?
    - Eliminate spider traps
    - Reduce ignored content
    - Spider paths
    - Paid inclusion