Use of the Internet in Investigations

Massachusetts Office of the Attorney General
September 18, 2009

Ming Chow
Tufts University
mchow@cs.tufts.edu
http://www.cs.tufts.edu/~mchow

Introduction

- Unless you have been living in the woods (literally) for the last decade, fragments of your and other people's information is available on the Internet.
- It is not a matter of *if* the information is available, but *how and where* can you find it.
- Despite the plethora of legislation, federal laws, and regulations, it has been easier to obtain information online.
- The Internet can be a daunting resource for investigators because of information overload. Are there any tools to facilitate investigations and analysis?
- The goals of this training:
  - Hone your information discovery skills
  - Identify and effectively use the resources that matters today

About Myself

- Day: Work at Harvard University
- Night: Instructor at Tufts University
- Taught the course *Security, Privacy, and Politics in the Computer Age* in Spring 2005 and 2007
- Taught *Cyber Security* at Middlesex Community College in Spring 2008 and Spring 2009
- Taught *Use of the Internet in Fraud Investigations* at NEAIFI Annual Trainings in 2007 and in 2008
- SANS / GIAC Certified Incident Handler (GCIH)

What We Will Cover

- Google hacking
- Information quality
- Other search engines, databases, and investigation tools
- Social networking and microblogging
- Network information tools
- Privacy, including anonymizers
- NOTE: All tools that we will cover are free to use (i.e., no paid subscriptions required)

What We Will Not Cover

- Protect yourself
- Search and seizure
- Eradication and recovery
- Forensics
- Underground economy
- Cybercrime
    - Phishing
    - Identity theft
    - Auction fraud
    - Insurance fraud
    - Scams (e.g., Nigerian Money Orders, Lottery Club)

Reconnaissance

- Information is gold.
  - If I was a bad guy, this is what I want.
- *Online* reconnaissance includes web searching, website analysis, and resource mapping.
- Elements of personnel discovery:
  - Full name
  - User and account names
  - Physical and e-mail addresses
  - Telephone numbers
  - Employment history
  - Personal websites
  - Hobbies
  - Recent whereabouts
  - Relationships

Caveat

- The plethora of tools and information on the Internet can amount to information overload.
- Some information you will find may not be accurate.
- You must verify your results and use your common sense.

Google

- Google searching is an invaluable skill for webmasters, executives (e.g., human resources, marketing), attackers, and investigators.
- The power of Google is underutilized
- What we will discuss:
    - Basic and advanced search query building
    - Special Google functions and patterns
    - How to find sensitive information, including vulnerabilities and even accessing security cameras
    - Google’s privacy policy
- What we will NOT discuss:
    - Google’s Page-Rank Algorithm
    - Google’s services and tools (e.g., Google Earth)
    - Google’s infrastructure (databases and servers)

Google Search Basics

- *All search queries are case insensitive*
- *Default AND queries*; use `OR` (capitalized) to find results with at least one of the words
- *I’m Feeling Lucky* – returns the most popular search result for the query
- *Exact phrases* – put quotation marks around query
  - Example: `"white collar crime"`
- *Word inclusion*: use `+` before search word in phrase (no space between the `+` and the word)
  - Example: `insurance fraud +auto`
- *Word exclusion*: use `–` before the search word in phrase (no space between the `–` and the word); used to weed out results
  - Example: `auto theft -grand –game`
- *Wildcards*:
  - `*` (asterisk): Match one or more whole words
    - Example: `reclaiming * dollars`
  - `.` (period): Match one single character (including whitespace)
    - Example: `index.of`

Advanced Search

- URL: **http://www.google.com/advanced_search?hl=en**
- Custom specifications:
    - Language
    - File format
    - Domain
    - Usage rights (e.g., free to use or share, free to use share or modify)
    - SafeSearch

Language Tools

- URL: **http://www.google.com/language_tools?hl=en**
- Translate text from one language to another
- Translate an entire page
- Supported languages:
    - Arabic (ar)
    - Chinese, Simplified (zh-CN)
    - German (de)
    - Japanese (ja)
    - Spanish (es)
    - Elmer Fudd (xx-elmer)
    - Hacker (xx-hacker)
    - Klingon (xx-klingon)
    - Over 100 others

Search Operators

- **IMPORTANT: All operators must follow with a ":" and no space between the operator and the first query word**
- `cache` – Shows the last version of the web page that Google has in its cache
  - Example: `cache:www.myspace.com/barackobama`
  - Pages may not necessarily be cached from the night before, possibly from a week ago
  - Very useful to look at the old information (e.g., links, photos, and information no longer available on current site)
  - Link to cached page should also appear for most search results
  - Only text is cached by Google. No images and stylesheets are cached. Images on cached pages are retrieved from the source (hopefully the images are not deleted).
- `inurl` – Shows web pages that only have the specified search words in its URL
  - Example: `inurl:"pub"`
- `intitle` - Shows web pages that only have the specified search words in its title (on the web browser bar)
  - Example: `intitle:"index of"`
- `intext` – Shows web pages that only have the specified search words in the body of the document
  - Example: `intext:senior citizens scam`
- `filetype` – Shows only files with a particular file format (e.g., .doc, .pdf, .jpg)
  - Example: `intext:myspace filetype:pdf`
- `site` – Shows only web pages in a particular domain
  - Example: `site:livejournal.com hacked`

Search Operators (continued)

- `link` – Shows all web pages that link to the specified web page
  - Example: `link:www.ifb.org`
- `related` – Lists web pages that are “similar”
  - Example: `related:www.acfe-boston.org`
- `phonebook`
  - `rphonebook` – Residential phonebook
  - `bphonebook` – Business phonebook
  - Acceptable combinations:
    - first name (or first initial), last name, city (state is optional)
    - first name (or first initial), last name, state
    - first name (or first initial), last name, area code
    - first name (or first initial), last name, zip code
    - phone number, including area code
    - last name, city, state
    - last name, zip code
    - To have your residential phone a
- `define` – Returns the definition of a word
  - Example: `define:omnipotent`
- `info` – Shows information about the web page
  - Example: `info:www.mass.gov`

Anatomy of the Search URL

- `http://www.google.com/search?hl=en&q=inurl%3A%22insurance+fraud%22+%2Bauto&btnG=Search`
- Identify the parameter/value pairs:
  - `hl` = Language
  - `q` = Query (with search words concatenated with "+")
  - `as_q =` Advanced search query
  - `num` = Maximum number of results
  - `btnG` = Google search button; `btnI` = "I'm Feeling Lucky Button"
  - Special characters:
    - `%22` = space
    - `%2B` = "+"
    - `%3A` = ":"

Special Patterns

- Built-in calculator and conversion tool
  - `sqrt(128)`
  - `e^2`
  - `(8 * pi) /3`
  - `2007 in roman numerals`
  - `40 yards to kilometers`
- UPS and USPS tracking numbers
- Addresses
- UPC codes
  - `045496900083`

Special Patterns (continued)

- Vehicle Identification Numbers (VINs)
- Stock quotes
  - `GOOG`
  - `AAPL`
  - `C`
- Spell-checking
  - `berkshare haithway`
- Area codes
  - `718`
  - `213`

Johnny Long’s Google Hacking Database (GHDB)

- Database of Google search queries that reveal sensitive information and vulnerabilities
- Initially a joke, now a major tool, and is even used in several major security products
- URL: **http://johnny.ihackstuff.com/ghdb/**
- What some of the searches can reveal:
  - User names and passwords
  - Financial spreadsheets
  - Web pages containing vulnerable data or way too much information
  - Online shopping information, including credit card numbers
  - Access to security cameras
- Tools based on the GHDB
  - Foundstone’s SiteDigger (version 2.0)
  - Gooscan (written by Johnny Long)

Google and Privacy

- What does Google knows about you (i.e., information stored in their databases)?
  - Source IP address
  - Timestamp
  - The search query
  - Browser and operating system used
  - Cookie ID
  - Example of an entry in their server log:

    -
      ```
      123.45.67.89 - 25/Mar/2003 10:15:32 -
      http://www.google.com/search?q=cars -
      Firefox 1.0.7; Windows NT 5.1 -
      740674ce2123e969
      ```

Google and Privacy (continued)

- Google’s new privacy policy (updated on September 8, 2008):
  - Server logs will be anonymized after 9 months (i.e., search queries are not linked to individuals)
  - See FAQs for more details: **http://googleblog.blogspot.com/2008/09/another-step-to-protect-user-privacy.html**

Google and Privacy (continued)

- Removing content and search results off of Google:
  - If you are the owner of the website or content, go to http://www.google.com/support/webmasters/bin/answer.py?answer=35301 for instructions
  - If you *are not* the owner of the website or content:
    1. Google: Kindly ask the webmaster or owner of the content to remove the information.
    2. Google: If that does not work, use the web page removal tool at https://www.google.com/webmasters/tools/removals?pli=1
    3. Tough luck!

Summary: Google

- Google is so far ahead of the game in the “search engine war.” It does very well in terms of accuracy (without the spam too).
- It has reached a point where if you cannot find what you are looking for in Google, then it is a really bad thing.
- Engineers at Google tinker with the algorithm every day to ensure the best results.
- Google receives many government and law enforcement subpoenas each year, and the amount is expected to grow.
- Google’s ubiquitous computing and omnipresence are troubling to many people.

Other Search Engines and Databases

- Google is a very powerful search engine, but it is not the only search engine around.
- Sometimes, Google components such as the phonebook act erratically.
- Many public records are now searchable (e.g., local, state, and federal government).
- There is a wealth of search engines tailored for investigative purposes.

Other Well-Known Search Engines

- Yahoo!: **http://www.yahoo.com/**
- Bing: **http://www.bing.com/**
- MSN: **http://www.msn.com/**
- Lycos: **http://www.lycos.com/**
- AltaVista: **http://www.altavista.com/**
- Ask: **http://www.ask.com/**
- HotBot: **http://www.hotbot.com/**

Meta-Search Engines

- Utilize multiple search engines at the same time (e.g., Yahoo!, MSN, Cuil, Bing):
  - Dogpile: **http://www.dogpile.com/**
  - Mamma: **http://www.mamma.com/**
  - Metacrawler: **http://www.metacrawler.com/**
  - Rollyo (Roll Your Own Search Engine): **http://www.rollyo.com/**

People Search with ZoomInfo

- **http://www.zoominfo.com/**
- Focused on people, companies, and relationships
- Over 4.5 million users a month
- Crawls the web and uses natural language processing and artificial intelligence to parse information

Reverse Lookups

- Addresses
  - 411.com: http://www.411.com/10671/reverse_address
  - InfoSpace: http://www.infospace.com/info/redirs_all.htm?pgtarg=reve
  - Reverse Address Directory: http://www.reverseaddress.com/
- Telephone Numbers
  - Directory Assistance Plus: http://www.daplus.us/
  - AnyWho (AT&T): http://www.anywho.com/rl.html
- Cell Phone Numbers (very difficult to do)
  - IAF: http://phonenumbers.iaf.net/phone.php

Internet Archive

- Internet Archive (a.k.a. Wayback Machine)
  - One of the largest digital libraries in the world, and is recognized by the American Library Association.
  - You can view pages (snapshots) of some of the most popular websites from years ago on different dates.
  - **http://www.archive.org/**

Public Records Databases

- Unfortunately, most, if not all, of databases require a subscription account:
  - LexisNexis: http://www.lexisnexis.com/
  - ChoicePoint: http://www.choicepoint.com/
  - Intelius People Search: http://www.intelius.com/
  - Public Records: http://www.publicrecordsfinder.org/
  - Gov-Records: http://freerecordsregistry.com/
  - BRB Publications: http://www.brbpub.com/pubrecsites.asp
  - PeopleFinders: http://www.peoplefinders.com/

What is “Web 2.0”?

- *“Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform”* – Tim O’Reilly, founder of O’Reilly Media; 2006
- Elements of Web 2.0:
  - Creativity
  - Information sharing
  - Collaboration
  - Web functionality

What is “Web 2.0”? (continued)

- Products epitomizing Web 2.0:
  - Wikipedia
  - Blogs
  - Social networking
  - RSS and XML feeds
  - YouTube
  - Google Maps API (the programming interface that is used in thousands of products)
  - Web-based applications and databases
  - Image sharing services

What is “Web 2.0”? (continued)

- Why?
  - Less expensive and more powerful computers
  - Inexpensive storage and web hosting
  - Digital photography
  - Computer programming skills
  - Venture capitalists
  - Google
  - Creativity

A Dangerous Shift

- Then: Mass media
    - By in large, still a reputable source
- Now: Personal media; user-driven content
    - Ask yourself, do you trust everything that you read in a Wikipedia entry?

Concerns (Besides Information Overload)

- Entity disambiguation (Bob Jones vs. Bob Jones)
- Who or what is the data source?
- Purpose / motive behind data being available
    - To spread opinion?
    - To attract business clients?
    - To apply for a loan?
    - To be malicious? (e.g., defame someone?)
        - Example: spoofed website; malware is downloaded onto your machine just by accessing the website. Computer is taken over by attacker.
- When was the data updates?
- Actual individual?
    - Web traces created as person lives
    - *Under age 30 are more likely to have a greater online presence (e.g., Facebook)*
- Fake individual?
    - Identities created only for the purpose of fraudulent activities

Distinguishing Information Quality

- Use your common sense, and you must always verify your results.
- From Stanford University:
  - The five types of elements that increased credibility perceptions were real-world feel, ease of use, expertise, trustworthiness, and tailoring.
    - Example: Washington Post
  - The two types of elements that hurt credibility were commercial implications, and amateurism. A few red flags include advertisements, lots of images, and popups.
  - (Source: **http://captology.stanford.edu/pdf/p61-fogg.pdf**)

Social Networking

- Build communities to share information and interests: from events and activities to personal details
- Largely “open” environments (rules differ for various services)
- Uses:
  - Reconnect with old classmates and friends
  - Meet people with similar hobbies and professions
  - Download and share files (e.g., videos, pictures)
  - Connect with new people
- Services of a social networking website or software normally include:
  - User profiling
  - Messaging (e.g., chat and e-mail)
  - File (image and video) sharing
  - Blogging
  - Message boards
  - Privacy controls

Social Networking Services

- **Facebook (~250,000,000 users)**
  - **Over 85% of college students have a Facebook account**
- **MySpace (~124,000,000 users and shrinking)**
- Ning
- Bebo
- Orkut
- Adult FriendFinder
- Friendster
- Classmates
- Xanga
- …and the list goes on

LinkedIn

- http://www.linkedin.com/
- Over 40 million users
- For business and professional networking, and to enhance one’s reputation
- Site features:
  - Connections (contact network)
  - Professional experience profile and resume posting
  - Profile views
  - Professional Q&A
- Profiles are searchable via Google, Monster.com, and other search engines

LinkedIn (continued)

- Information that can be retrieved:
  - Job history and companies
  - Recommendations
  - E-mail address
  - Education history
  - Professional organization memberships
  - Interests
  - Details on connections and relationships
  - Who looked at the person’s profile and when?
- Similar service: Plaxo (http://www.plaxo.com/)

Rapleaf

- http://www.rapleaf.com/
- Recent “Web 2.0” startup
- As advertised:
  - Lookup your information
  - Manage your online privacy
  - Manage your online reputation
- As of July 10, 2008, you can no longer search for information on others, however…

E-Mail Address Lookup and Verification

- …if you are a programmer, you can apply for and use the Rapleaf API
- Give one input: *an e-mail address*
- *The results:*
  - E-mail address verification
  - Social networking profiles that are registered under the e-mail address
  - Location (city and state)
  - Full name (potentially)
  - Online "reputation"
- Here it is: **http://www.cs.tufts.edu/~mchow/rapleaf_search.php**
- Alternative service: CentralOps (**http://centralops.net/**)

Social Networking + Blogging + Mobile Device = Microblogging

- Texting or text messaging on most mobile devices (e.g., cell phones) uses the Short Message Service (SMS) protocol; 160 character message length.
- Microblogging – short messages or updates; usually at most 140 characters.
- The mantra: simplicity
- Other ways to send short messages:
  - Online form
  - Instant Messaging
  - E-mail
- Why microblogging?
  - News
  - Networking
  - Updates (e.g., from conferences)
  - Lifecasting / lifestreaming

Twitter

- **http://www.twitter.com/**
- Over 10 million users
- Very lightweight: messages up to 140 characters long
- How messages can be sent:
  - Directly on the website
  - SMS / mobile phone
  - Facebook
- Messages displayed on user’s page
- *Value:* instant messages; can pin-point what user is currently doing or even whereabouts
- April 2008: A student “twittered” his way out of an Egyptian jail (http://www.cnn.com/2008/TECH/04/25/twitter.buck/index.html)
- Standalone iPhone application available: “Twitteriffic” (free)

The Lingering Problems with Social Networking

- Plenty of legitimate business (and criminal) opportunities still available.
- User education on social networking still lax.
- Exhibitionism and amateurism
  - Pornography
  - Gang-related activities
- Sexual predators
- Prone to fraud and exploits (due to poor programming and development practices)
  - Identity theft
  - Cracked user accounts
  - Malware (e.g., viruses, spyware) delivery

Privacy

- *“Privacy is dead, deal with it.”* --Scott McNealy, CEO of Sun Microsystems
- Under Justice Department guidelines, anything posted online is “fair game” (and many young people feel the same way).
- Possible hearsay

Summary: Web 2.0 and Social Networking

- Social networking websites and Web 2.0 services have changed the landscape of the Internet, and have redefined privacy.
- Plenty of legitimate and illegitimate business opportunities still available.
- User education on social networking still lax.

Network Information Tools

- **whois** - Provides information about domain names and their registrants (technical contacts, expiration date, etc.)
- **nslookup / dig** - Queries Internet domain name servers (DNS)
- **ping** - See if a host is reachable
- **traceroute** - Sends trace packets for determining information; can determine a server that is slowing down transmission
- These tools are available either from a UNIX / Linux command line, MS Windows Command Prompt, or from the web
  - **http://www.centralops.net/**

Information Discovery Tools

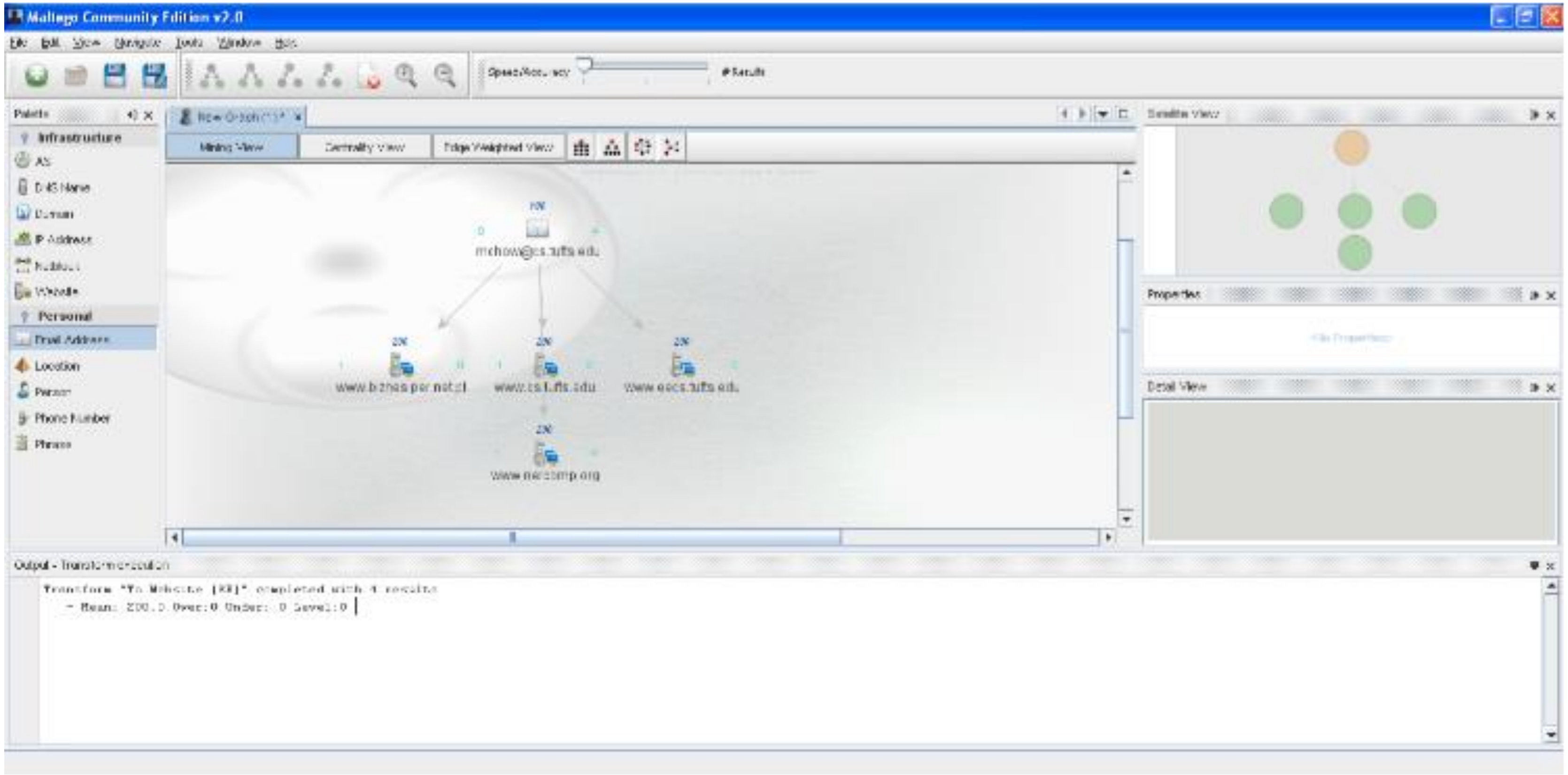
- *Quick review:* list some of the services and tools that we have discussed and demonstrated?
- *Problem:* information overload
- ***Question: What tools are available to conduct investigations using the Internet, and how reliable are the tools?***
  - Answer: So far, we have concentrated on using only one computer program: the web browser.
  - The approach of using a web browser and finding the needle-in-the-haystack is a burden for investigators.
  - The goal: *connect-the-dots* given the plethora of information available on the Internet.

Paterva Maltego

- **http://www.paterva.com/**
- Paterva is a South African company
- *“An open source intelligence and forensics application. It allows for the mining and gathering of information as well as the representation of this information in a meaningful way. Coupled with its graphing libraries, Maltego allows you to identify key relationships between information and identify previously unknown relationships between them.”*
- Apply a set of *transforms* to an initial node, and will create new nodes, branches, and relationships
- In other words: “determine the relationships and real world links between a vast set of information” (e.g., e-mail address, phone numbers)

Demo of Paterva’s Maltego

- Two versions
  - Community Edition (Free)
  - Commercial Edition



```
Transform "To Website [SE]" completed with 4 results
   - Mean: 200.0 Over:0 Under: 0 Level:0
```

A Word on Being Stealth

- Remember, you are not invisible or truly anonymous when performing online investigations.
- You may not want to leave your IP address in various logs.
- We have used one method to be stealth: throwaway e-mail accounts
- Other methods to enhance your anonymity online:
  - Tor (download at http://tor.eff.org/)
  - Change your MAC address
  - Use public web proxies
  - Use an anonymizer service (e.g., http://www.webwarper.net/)
  - Use your own proxy machines
  - Use a public Internet café or connection
- It is possible to use Google as a proxy (but not a good one)
- Still, you have to be careful what you use

Problems With New Services

- Spawning too fast
- Long-term viability
  - Largely advertising revenue-based
  - Will it cease to exist tomorrow?
- Reliability and security
  - Scalability issues; that is, server overload when there are too many users
  - Suffers from poor programming and development (e.g., Twitter’s insecure login by default)
  - Rushed out; buggy
  - Many tools that utilizes a service’s API (e.g., Rapleaf, Twitter) are at the mercy of the company

Conclusion

- The feeling of information overload when using the Internet for investigations is quite common…
- …but if you can be *tactical* and narrow down your resources, you *will* find what you are looking for.
- Because innovation on the Internet grows at such a rapid pace, it is crucial to keep pace and learn new relevant technologies and tools.
- You can only believe and use so much of what is on the Internet. Thus, you must always verify your results, and use your common sense.
- The human network is even more critical now to verify information.
- Do I see limits being placed on the ease of access to information? That would conflict with the underlying premise of the Internet: it a public good.

Conclusion

- The feeling of information overload when using the Internet for investigations is quite common…
- …but if you can be *tactical* and narrow down your resources, you *will* find what you are looking for.
- Because innovation on the Internet grows at such a rapid pace, it is crucial to keep pace and learn new relevant technologies and tools.
- You can only believe and use so much of what is on the Internet. Thus, you must always verify your results, and use your common sense.
- The human network is even more critical now to verify information.
- Do I see limits being placed on the ease of access to information? That would conflict with the underlying premise of the Internet: it a public good.

Resources

- *Googling: I'm Feeling (un)Lucky* by Gregory Conti. Presented at DEFCON 14 in Las Vegas, NV (August 2006).
- *Internet Tools For Investigations* by Fred Howell. Presented at the NEAIFI 2nd Annual Training (June 2007).
- *Maltego Part I - Intro and Personal Recon* by Chris Gates: http://www.ethicalhacker.net/content/view/202/1/
- *More Than MySpace: A Social Networking Websites Guide for Investigators* by Ryan Kapaun (May 2007).
- *Safety of MySpace* by Hemanshu Nigam. Published in the Police Chief Magazine (March 2007 issue). http://policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=1140&issue_id=32007
- *Search Engine Security Auditing* by Russ McRee. Published in the June 2007 ISSA Journal. http://holisticinfosec.org/toolsmith/docs/june2007.pdf

Resources (continued)

- *Security Warrior* by Peikari and Chuvakin. Published by O'Reilly (2004).
- *Tactical Exploitation* by H.D. Moore and Valsmith. Presented at DEFCON 15 in Las Vegas, NV (August 2007). http://www.metasploit.com/confs/blackhat2007/tactical_blackhat2007.pdf
- *Working the Network* by Maggie Jackson: http://www.boston.com/jobs/news/articles/2008/09/14/working_the_network/