

I know what you
(and your company)
did last summer...

FIRST 2007
Seville



PATERVA

Agenda

- About myself
- Think big / the dilemma with information
- The hacker in the dark corner...or not
- An information collection framework
- Evolution Demo
- Spying on your own + some more demos
- Data mining prevention and detection
- Using the information collected
- Conclusion



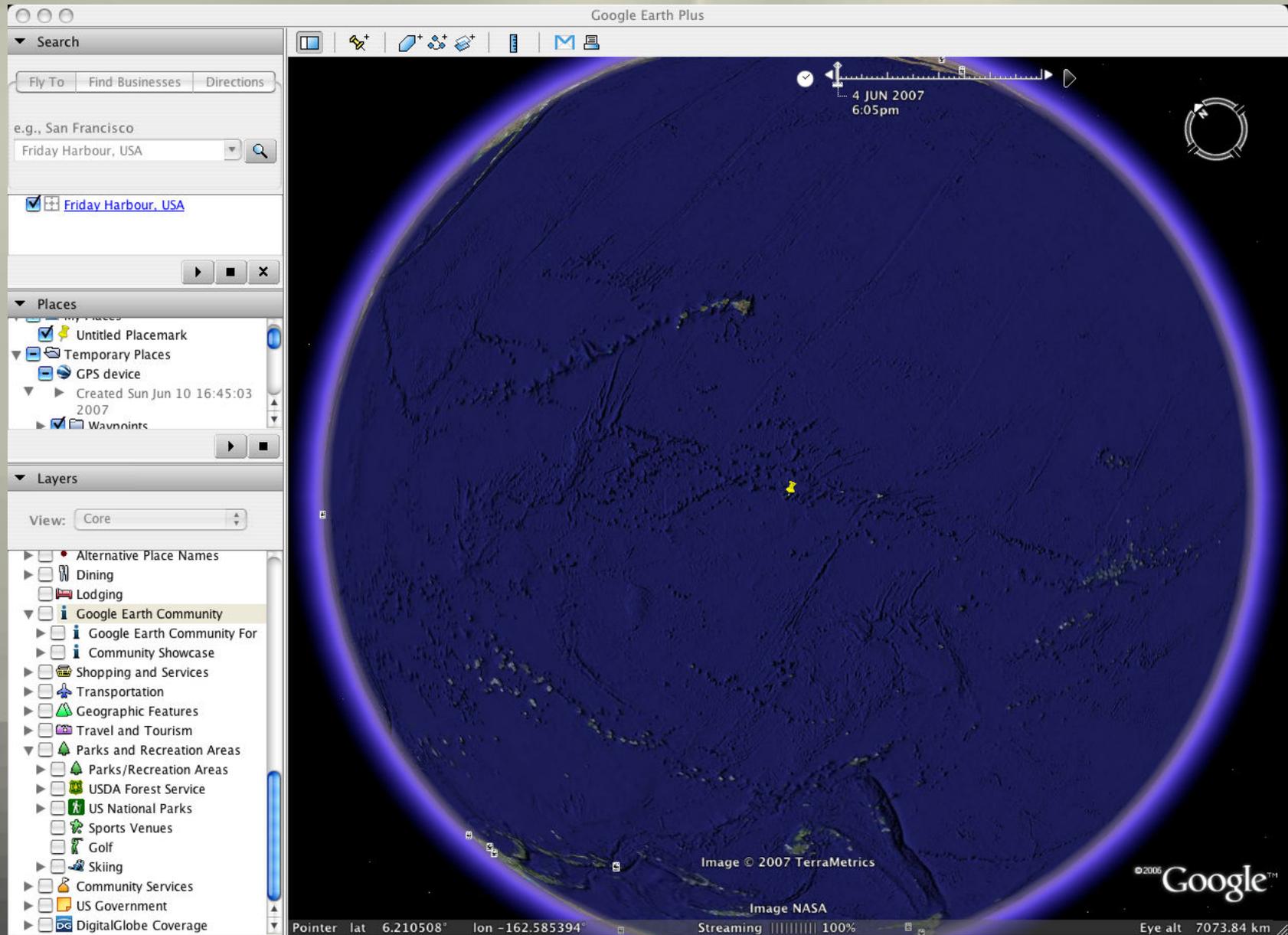
About myself

- Who I am?
- Why do we care??
 - Roelof Temmingh
 - Completed Bachelor's degree in Electronic Engineering
 - Started SensePost with friends in 2000
 - Talked/trained at BlackHat, RSA, Defcon, FIRST '03
 - Co-wrote some Syngress books
 - Built tools @ SensePost – Wikto, Suru, BiDiBLAH, Crowbar
 - Breaking into networks over the Internet, Web applications
- Today
 - Started Paterva beginning of 2007
 - Evolution
 - R&D

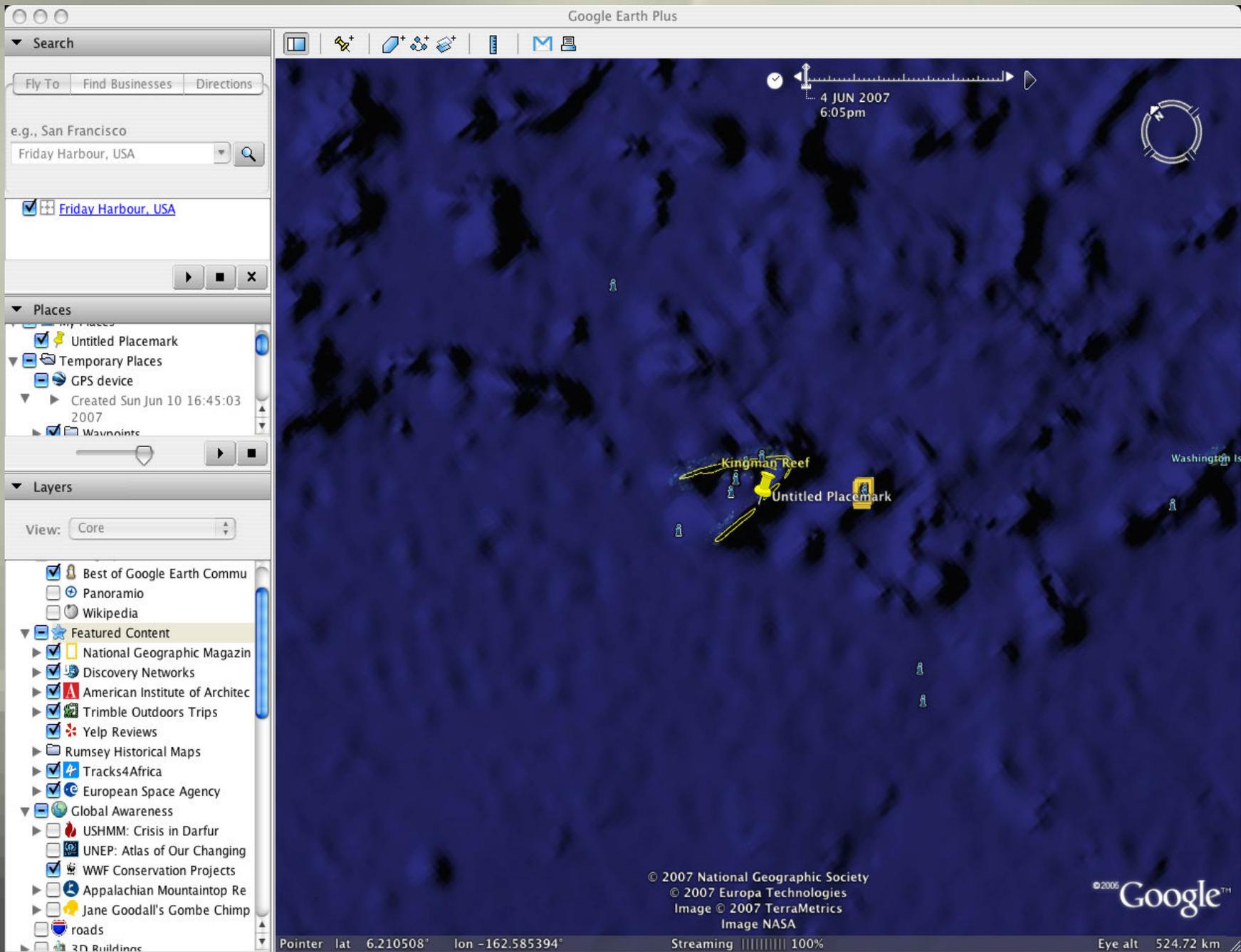


Think big

- Have you ever heard about the Palmyra Island(s)?



Think big



Think big

Palmyra Atoll @ nationalgeographic.com

http://www.nationalgeographic.com/ngm/0103/1

The Spy Who Billed...
Paterva Evolution [Beta!]
- [uncon] Rootkitted By Default
Palmyra Atoll @ nationalgeograph...

NATIONAL GEOGRAPHIC .COM

Support The Fresh Air Fund
Help spirits grow!

the Fresh Air Fund
serving children since 1877

NATIONAL GEOGRAPHIC MAGAZINE

HOME CONTACT US FORUMS SUBSCRIBE NGM Site Index

ZOOM IN

More photos from Palmyra

<< Back to Feature Page

View exclusive photographs and get the facts behind the frame.



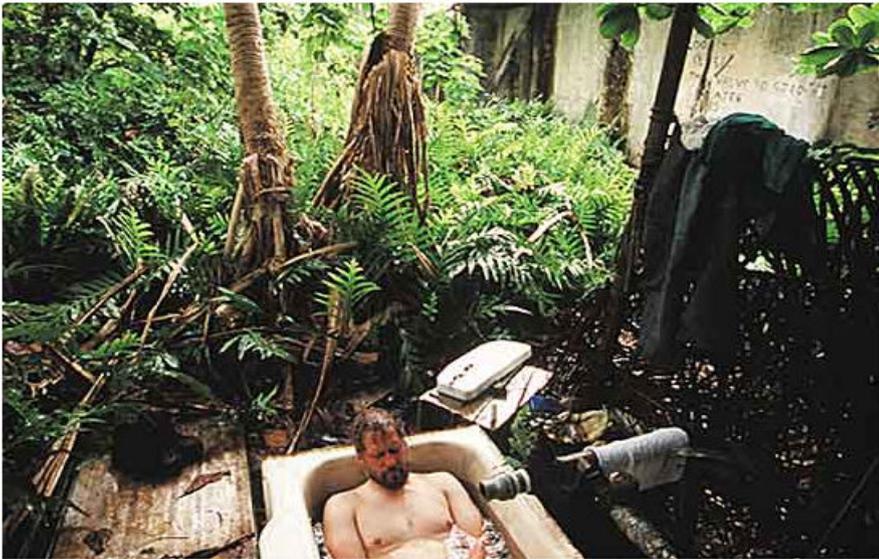
Click to ZOOM IN >>



Click to ZOOM IN >>



Click to ZOOM IN >>



A Coveted Spot

Photograph by Randy Olson

Nirvana! The atoll's sole bathtub is a daily retreat for engineer David Johnson. Water is furnished by a jury-rigged line leading from a concrete catchment basin, at upper right, installed by a U.S. military unit during World War II. The legendary soaking spot was eagerly anticipated by transient cruising sailors short on fresh water and grown grungy during long, tiring passages. Johnson, a patent attorney with a sideline in electronics, set up a communications system for the atoll, a crucial service for pilots homing in on this tiny speck of terrain in the vast Pacific.

PHOTO FAST FACTS

Camera: Leica M6	Weather Conditions: Broken clouds
Film Type: Fujichrome Velvia 50	Time of Day: 5 p.m.

Done



Think big

- This year alone 1500 000 000 GB of new, unique information would be created .. > the last 5000 year's info.
- Technical info doubles every 2 years...predicted to be doubling every 72h in 2010
- There are 2.7 billion searches done on Google every month.
- The number of text messages sent & received per day > world's population.
- There are 5 times more words in the English language than in Shakespear's time.
- 3000 books are published daily.
- One week of New York Times > the info in a life of someone in the 18th century.



Think big

Yeah right....just how do they know...and who are 'they'

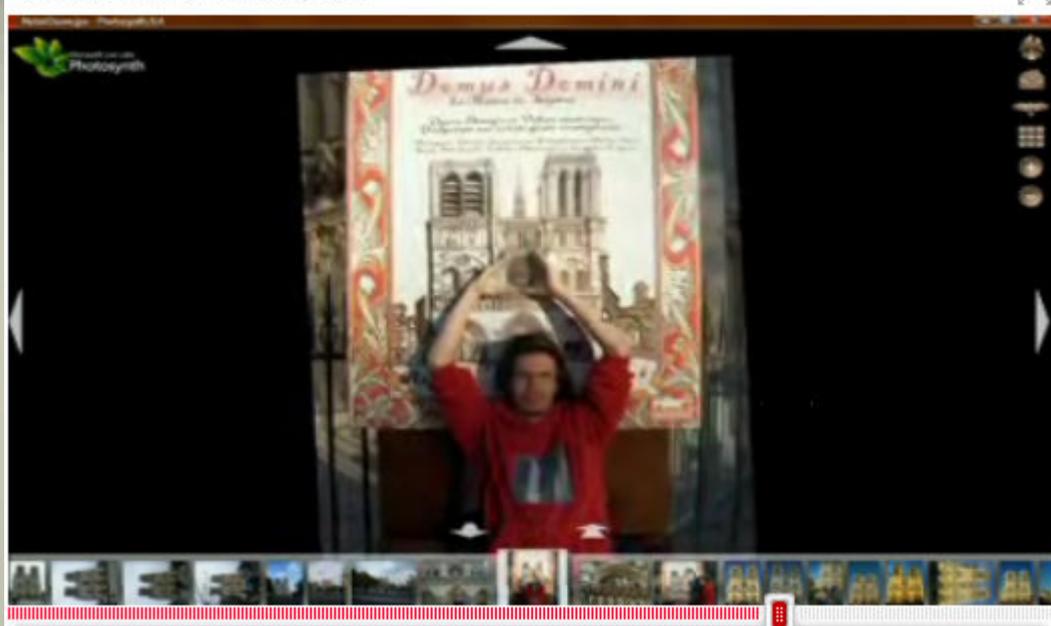
But even at 50% of that we are still dealing with a heck of lot of bits and bytes.

- Recently people started getting uneasy with privacy and Google Maps/Earth.
- Microsoft created/bought PhotoSynth – that allows a 3D environment to be built using a variety of photos.
- See the Notre Dame in 3D from photos they scraped from Flickr...
- More and more camera phones are equipped with GPS these days
- Google Flat Earth?





Filmed Mar 2007; Posted May 2007



Filmed Mar 2007; Posted May 2007



“But, much though I hate to bring it up, this has a **dark side** as well. If it can separate out your head and hands from a poster of Notre Dame and correlate that with a zillion other photos, what's to prevent it from separating out YOU? And **tracking you** (both temporally and spatially, thanks to the mountain of metadata that accumulates) wherever you are or were”



Think big

It seems that we have a dilemma.

On the **one hand** we want instant access to information (Google etc), and we want it to be linked for us (Photosynth, Facebook, LinkedIn)

On the **other hand** we get upset when people use this linked, instant information in a 'bad way'.

It seems that information is like a 1024 bit key, split in 16 bit chunks – you can look at each part individually, but if you piece it together in the right order it leads to trouble.



Think deep...introspection

Why do real hackers hack systems?

No...I mean...really...

“See if it's safe because I bank there” - yeah...OK if you say so.

The pleasure from seeing the lock open == the info appearing on your screen. Control over the lock.

It's being able to do things normal people can't, knowing things that normal people don't.

It's less about applying the information->knowledge



The assessment that made me think

- External (over the Internet) assessment for Company ABC in 2004
- Big company, loads of Internet facing infrastructure
- Very juicy data, control over interesting things
- In 5 years– never had a security assessment.

- Results – no surprise - complete compromise
- 7 different vectors into the very internal network
- Low hanging fruit (on the ground really)



The assessment that made me think

- Many boxes compromised by (different) worms
- But - no signs of human break in
- No rootkits
- No temp files, scanners
- It appeared that nobody went in there....

- Or did they?
- If they did they covered their tracks REALLY well.
- For all the 'hackers in dark corners' it's pretty quiet out there.
- Seems that if you really have skill you are working for an assessment company – less risky and better pay.



Call me disillusioned

- Speaking to a friend at Commercial Crime Forensics
- If they are not hacking into the system and taking money, how do they then?
- There seem to be a disconnect between what pen testing companies are doing and what's happening in real life.
- It's nice to show your skill in papers and conferences.
- And it's nice to work on complex technical problems.
- But real criminals does not write buffer overflows and their attacks almost never involve compromising computers
- It does however involve tech...and human nature.



Before we really start

It seems that we can classify security technology in 4 categories:

- Attack – e.g. Port scanner
- Prevention – e.g. Firewall
- Detection – e.g. Intrusion Detection System
- Cure – e.g. Patches

Which came first – the port scanner or the firewall?

Technology is driven by one of these factors

Let's try to apply this to information mining



Hacker's 6th sense

People regularly ask me - what can you find out about...?

... without touching/hacking them

... without them knowing

Most of the time it turns out there is a lot to be found.

“Why couldn't you do this??”

- Because hackers know the fabric of the 'net
- Because hackers think different thoughts
- Because hackers are have built-in deviousness
- Because hackers know who and how to ask

“But why would you like to know this?”

Hackers are rarely interested in the application of data,
but know how to get to it.



People always ask me...

Non-tech people ask “what can you tell me about this:

- Email address
- Company
- Person
- Word/phrase
- Website

...and us tech people think:

- IP addresses, virtual hosts
- Netblocks / AS routes
- Affiliations (Linkedin, Myspace, Orkut, Zoominfo, Facebook etc)
- Microformats - XFN, vCards, hCards etc. etc.
- Forward and reverse DNS – MX/NS records
- Whois records / rWhois and referring registrars
- Google fu / Deep web search
- Services like ip2geo, wayback machine, Google Earth, Zoominfo
- Meta information in documents



Information collection framework

In order to make sense of collected information we need to

- give meaning to raw data by
- categorizing information and
- linking information together by
- means of relationships

We thus need a framework that have a lot of tentacles in many data sources and that can shuffle data from one type to another in a intelligent way.

Evolution is such a framework...hopefully..



Disclaimer

Many systems out there (SiloBreaker etc.) that does a GREAT job at collecting general information and news and processing it.

This is a bit different:

- Looking at specific information by
- Using open source information,
- Which is free,
- That's generic for
- Global use and try to
- Find relationships between them.

E.g. Your database of Eastern Congolese surnames are not going to feature (but it could if you really want to)



Part I – relationship collection

The thinking behind the framework

Step 1: What can you tell me about...

- How would you do it as a human? =>
- Information transforms in the framework

Step 2: Collect information

- Searching, surfing
- Deep web & services

Step 3: Parse/convert information to **new entities**

Step 4: **Goto 1**

Find the 'hidden' relationships -

A -> B -> C and

X -> Y -> C

..then $A \sim X$



Evolution Entities

- Entities are things – like a person, an email address.
 - So, its about organising and linking “things”.
 - We have microformats (www.microformats.org)
 - Seems to be interesting, but not extensive enough
 - What about IP numbers, phrases etc?
-
- Entities have properties and methods (sounds a lot like a Java class eh?)
 - Methods==transforms that convert to other entities
 - Properties of parent entity get populated during the transform
 - Generated entities (children) can have their properties populated by the parent transform.



Evolution Entities

- First we need a shell that surrounds all entities.
- The shell looks like this:
 - Entity (the core)
 - Weight (distance from source Entity)
 - Date start
 - Date end
 - Generator
 - Parent*



Evolution Entities

- Properties marked with a * is a generated entity (or a pointer to other entities)
- Person Entity:
 - FirstName, LastName
 - Additional information
 - Country
 - Email addresses*
 - URLs (sightings)*
 - Telephone numbers*
 - Dates* (first seen, born, death etc.)
 - Persons*
 - Phrases*
 - Affiliations* (IM, social networks)



Evolution Entities

- Telephone Entity:
 - Country code
 - City code
 - Area code
 - Rest
 - URLs* (sightings)
 - Email addresses*
 - Telephone numbers*
 - **Persons*** (reverse white,yellow,red pages)
 - Location*
 - Phrases*



Evolution Entities

- Location Entity
 - Long
 - Lat
 - Country
 - City
 - Area
 - Full address
- Email address Entity:
 - Email addresses*
 - Telephone numbers*
 - **Persons***
 - Domain*
 - Location*
 - Phrases*



Evolution Entities

- Website (from URL) entity
 - Dates* (first seen, changed dates)
 - DNS name*
 - Websites* (incoming, outgoing links)
 - Documents* (Google)
 - Files* (Google & brute force)
- Domain Entity
 - Domain name
 - Email addresses* (whois)
 - Telephone numbers* (whois)
 - Persons* (whois)
 - DNS Names* (brute force, MX, NS)
 - Sub domains* (Google)



Evolution Entities

- DNS Name entity
 - DNS Name
 - Domains*
 - IP Addresses*
- IP address entity
 - IP address
 - Email addresses* (whois)
 - Telephone numbers* (whois)
 - Persons* (administrator, technical)
 - Websites* (Virtual hosted)
 - Location*
- Document entity
 - Title
 - Location (URL)
 - Persons* (author)



Evolution Entities

- Phrase entity
 - Telephone numbers*
 - Email addresses*
 - URLs* (sightings)
- Date entity
 - Date
 - Time
- Affiliation Entity
 - Type
 - Direct link
 - HTML source



Evolution Transforms

- Entity or Entities -> **Transform** -> Entity or Entities
- Repeat (& filter)

Examples:

Email address -> Person(s) [PGP] -> Email addresses [PGP] -
> Person(s)[PGP] -> goto 1

Domain -> sub domains -> DNS names -> IP numbers -> IP
blocks -> Geo locations



Evolution Transforms

- Transforms - “who can do anything with a ...?”
- Entities open and expandable
- Transforms uses plugin architecture
- Thus, you are only limited by your own imagination
- Framework independent of the entities/transforms.

- If you are in the Lexis/Nexis club...
- If you work for the phone company
- If you have access to any other juicy databases etc.
- The current version -> using global, generic, free, open sources



Evolution Transforms

Some really easy examples

- DNS name -> IP number(s)
- IP number -> DNS name
- Domain (whois) -> email address(es)
- IP address (whois) -> telephone number
- Telephone number -> Geo location
- IP number -> Geo location



Transforms tel -> email

Some more interesting examples:

Consider Telephone number -> email address

How would a human do it?

- Step 1: Google the telephone number
- Step 2: Look at the snippet/Surf to the result
- Step 2: Look for email addresses
- Step 3: See which are clearly connected to the telephone number

But it gets nasty...



Transforms: tel- > email

Assume the number is +27 83 448 6996

- =~ 083 448 6996
- =~ (0)83 448 6996
- =~ 083 4486996
- =~ +2783 4486996
- =~ etc

Option 1: only look for 448 6996.

- Your friendly plumber in Burundi

Option 2: Try combinations

- But “+27 83 448 6996” is more likely to be correct than 083 448 6996



Transforms: email -> XXX

Assume the email address is roelof@paterva.com

- roelof@paterva.com
- roelof at paterva.com
- roelof [at] paterva.com
- roelof at paterva dot com
- roelof at removethis paterva dot com
- roelof_at_paterva_dot_com
- And then some...



Transforms – confidence levels

Same goes for First Name/Last Name

- Results on search query for “Roelof Temmingh” is more likely to contain the correct email address for me than a query for R Temmingh.

So – let's do all the queries, but give each a 'confidence index'

Are there other parameters we can use to increase the quality of our results?



Transforms: factors when sorting by relevance

- Frequency of the parsed result
If, after parsing, I get roelof.temmingh@gmail.com 100 times it's likely to be more relevant than something that appears 2 times.
- Significance of the site where the term is located
Especially necessary when working with phrases.
- Correlation to the original search term.
roelof.temmingh@gmail.com is more likely to be Roelof Temmingh's email address than kosie.kramer@yahoo.com.
- Proximity to the search term.



Transforms : using Google

So, for each “fuzzy” search (where relationships are not 1:1) I can create an confidence index.

Using Google and only looking at snippets I get:

- Speedy results
- Control over amount of results returned
- Country specific results
- Significance of site -> Page Rank
- Proximity -> Anything in the snippet is already in close proximity to the search term

Confidence index=

(Query confidence * Frequency across all pages * correlation to search term) + (sum of page ranks)

Google helps even more

Numrange – for finding telephone numbers in a certain area.



Applications a.k.a So what??

For conventional security:

- Stock standard footprinting (DNS, IPs, domains etc)
- Nice for finding war dialing (?!) ranges
- Targets for social engineering and client side attacks
- Alternative email addresses for content attacks
 - “if the attacker can get the victim to visit...etc..”
- Finding alliances with weaker security postures
- Understanding business drivers and sensitivities



Applications a.k.a So what??

For conventional security [while we're at it]

- Easily extended to do a lot more in the conventional space:
- IP entity (already have)
- Portscanner transform ->
- Port entity ->
- Service banner transform ->
- Phrase (the banner itself)
- Or...vulnerability scanner transform ->
- Vulnerability entity (e.g. NessusID...take your pick)

- It really doesn't matter what the data and the connection between them is



Applications a.k.a So what??

- Is abc.com a phishing site?

Domain -> email addresses at domain

-> telephone numbers 'connected'

-> related DNS names -> IPs

-> Website

-> Whois -> email addresses

-> whois -> telephone numbers

Telephone numbers -> Geo location(s)

Telephone numbers -> Alternative emails

Telephone numbers -> Related tel. numbers

IPs -> Geo location(s)

IPs -> co-hosted websites on same IP

IPs -> whois -> owners / tel / email

Website -> First seen date

Website -> Phrases



Applications : more interesting stuff

- Who at the NSA uses Gmail?

Domain -> telephone numbers

Telephone number (area) -> Email addresses

- Which NASA employees are using Myspace/LinkedIn?

Domain -> email addresses

Email addresses -> social network

- Which people in Kabul are using Skype?

Telephone number (area) -> email addresses

Email addresses -> Affiliation



Even more applications

- In which countries do the USMC have bases in?

Domain -> Email addresses

Email addresses -> domain = sub domains

Sub domains -> DNS names

DNS names -> IP addresses

IP addresses -> Geo location

Other uses even...

- What are the names and email addresses of single, young woman in my neighbourhood who are straight (or not)?

phone area -> email

email -> social myspace

[filter]



Demo: Evolution web
<http://www.paterva.com>



This is almost like...

- Surfing the web
- Sites are explicitly linked (by people)
- s/sites/entities/g
- Linked by some association.
- And you can get lost as easy...



Limitations of the web interface:

- Direction is single entity, many transforms
- Not – multiple entities, single transform
- Does not show more than one degree of separation
- No $A \rightarrow B \rightarrow C$ and $X \rightarrow Y \rightarrow C$, thus $A = \sim X$
- After 2 rounds of testing you'd lose your train of thought.

So what we really need is a GUI that remembers the paths/connections.



Demo: Evolution GUI
<http://www.paterva.com>



In the future

- Automatically grow entities.
- Automatically grow entities until a keyword is found.
- Two entities -> grow automatically until connected.
- Multiple entities -> find most common entity.

- Better layout...
- Multi-threaded transforms = speed
- Hidden connections



Hold your horses ...a.k.a 'but this is BS'

- “...my mother/grandma can't even operate a mouse...”
- Not everyone on the net gives out their info
- Not all systems readily give info to computers

The info is kept closer...but also wider..

- Whois info details on new domains are now removed
- But look at Myspace / LinkedIn / Facebook
- Web 2.0 ... new generation of 'WHAT privacy??'
- “Nowadays the kids have their photos on flickr, their profiles and friends on Facebook or Myspace, and their personal thoughts on LiveJournal/Blogger/Twitter, and future-cringeworthy-moments be damned. Concern about privacy is so last century. :o)”

What this “Deep Web” thing anyhow?



Spying on your own

You are :

- the information you publish
- the information others publish about you
- your associations, and
- **the information you search for.**

The holy grail of information collection – your search terms

How can we know what other people are searching for?



Collecting search terms

Recently AOL 'lost' a couple of search terms (well OK 20 million of them)

You can search search terms at data.aolsearchlogs.com. Searches are connected by userID. People are really weird - hours of fun...

- User #13324924 searched on 'help navada a resident of yours is harrassing me and threatening to have me killed by a guy named ronnie and she will not give me my belongings she tookk them and refuses to return my nesseceties' at 2006-04-29 20:11:47
- User #13324924 searched on 'cheap flights' at 2006-05-16 05:43:55



Collecting your search terms

If we control the infrastructure of a network we can

- Redirect outgoing traffic to port 80 into a Squid proxy
- Change two settings in the default Squid config to log MIME headers (cookie) and show parameters (search terms)
- Copy logs into a database
- Extract search terms per Google cookie (which expires in 2038)

...and have exactly the same...(plus we get all the sites the user went to as a bonus)

What prevents your company/government to do the same?
Or ... publish your proxy in open proxy list / become a TOR node



Collecting your search terms

Redirecting traffic transparently to Squid is easy [BSD]:
ipfw add 10 fwd 127.0.0.1,3128 tcp from any to any 80

Consider the following Squid log file entry for a Google search:

```
1181178854.617 985 196.22.177.60 TCP_MISS/200 6674 GET
http://www.google.com/search?hl=en&q=FIRST+Spain&btnG=Google+Search - DIRECT/64.233.183.147 text/html [Host:
www.google.com\r\nUser-Agent: Mozilla/5.0 (Macintosh; U; Intel Mac OS X; en-US; rv:1.8.0.12) Gecko/20070508
Firefox/1.5.0.12\r\nAccept:
text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5\r\nAccept-Language: en-
us,en;q=0.5\r\nAccept-Encoding: gzip,deflate\r\nAccept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7\r\nKeep-Alive: 300\r\nProxy-
Connection: keep-alive\r\nReferer: http://www.google.com/\r\nCookie:
PREF=ID=189373c0b59355a1:TM=1180932598:LM=1181068190:GM=1:S=Qa9d1y0RVZnvTbAM; GTZ=-120; TZ=-120;
S=gmail=pnzCF1YctjOSsnR8X7zBpA:gmail_yj=eDI35jgqL12M4mbpx-
gWzw:gmproxy=QO1qaNVM_gc:gmproxy_yj=EGyldkRPhq8:gmproxy_yj_sub=8j24bYO2SDY:dasher_cpanel=6ue4_G2-
qJU\r\n
```

The Cookie, IP number and search terms have been highlighted:

- Cookie: **PREF=ID=189373c0b59355a1**
- IP number: 196.22.177.60
- Search term: **FIRST+Spain**



Demo: Collecting your search terms (PollyMe)



Collecting your search terms

Next, we log into Gmail...after a while we see requests like:

```
1181179254.749 938 196.22.177.60 TCP_MISS/200 436 POST http://mail.google.com/mail/channel/bind?at=dcab728835011ea6-113027c61bd&VER=2&SID=53382F9A57445B1E&RID=10472&zx=7vdr9uobvgxy&it=19680 - DIRECT/66.249.91.18 text/html
[Host: mail.google.com\r\nUser-Agent: Mozilla/5.0 (Macintosh; U; Intel Mac OS X; en-US; rv:1.8.0.12) Gecko/20070508
Firefox/1.5.0.12\r\nAccept:
text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5\r\nAccept-Language: en-
us,en;q=0.5\r\nAccept-Encoding: gzip,deflate\r\nAccept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7\r\nKeep-Alive: 300\r\nProxy-
Connection: keep-alive\r\nContent-Type: application/x-www-form-urlencoded\r\nContent-Length: 69\r\nCookie:
__utma=173272373.1530872424.1181157683.1181157683.1181157683.1; __utmb=173272373; __utmc=173272373;
__utmz=173272373.1181157683.1.1.utmccn=(direct)|utmcsr=(direct)|utmcmd=(none);
GX=DQAAAHoAAAACDTIMPAKZE4vUWYQFbx0qX_QTbdP7mX3LaJF948
k4qTOaBmiUjmNhZyrTD6r4UkjtULJ1tSH_Wwexng9SemjNrzQzgsWVgfNK -KyyczJTUZped_dFIQ
Y1JrULqophfw2KvNuizeiNzxc R6kedcu 2QB EI76thpmtR2S2Se_A;
S=gmail=s6d4CAfZiiZzcXy5jem8iw:gmail_yj=iLyYmZW7SF_2P2IHWSF83Q:gmproxy=ayDL1lsc6L4:gmproxy_yj=3LplGSHA_
GA:gmproxy_yj_sub=1unO3s5vdEc:dasher_cpanel=6ue4_G2-qJU; GMAIL_AT=dcab728835011ea6-113027c61bd;
gmailchat=roelof.temmingh@gmail.com/505680;
GMAIL_LOGIN=1181157680818/1181157680818/1181157690363/1181157695313/1181157696991/1181157701802/118115
7703163/false/false;
PREF=ID=189373c0b59355a1:TM=1180932598:LM=1181068190:GM=1:S=Qa9d1y0RVZnvTbAM; TZ=-
120;
S=gmail=s6d4CAfZiiZzcXy5jem8iw:gmail_yj=iLyYmZW7SF_2P2IHWSF83Q:gmproxy=ayDL1lsc6L4:gmproxy_yj=3LplGSHA_
GA:gmproxy_yj_sub=1unO3s5vdEc:dasher_cpanel=6ue4_G2-qJU; GMAIL_RTT=699;
GMAIL_LOGIN=T1181157680818/1181157680818/1181157690363; SID=DQAAAHgAAAC8aa3M-
yPO21ybqncosZx_mpu6BDmx9hEqZsdqvCsjpSKIJ0ARuGDP-
WZQzvpDLqLqg4Ksg5BD417oZzEkJ3q0RG0RutrX_HAwJRNJX-TasfHIG5cWvPAm-
RKkp2A9vj5ZZG23CEtPLxgBTWNWjszs8kKc_EAtU88sW5a1DhPIHQ; GMAIL_HELP=hosted:0\r\nPragma: no-
cache\r\nCache-Control: no-cache\r\n
```



Collecting your search terms

- Note the same PREF=ID
- Note the (GoogleTalk) email address is displayed
- Someone monitoring the traffic only needs the GX cookie to be able to access my mailbox...and also...
- The cookie is not destroyed when you log out (it takes about 20 minutes)....unless...
- You have the 'remember me' setting which makes it last longer.



Demo: Spying on your own (Gmail)



Collecting your search terms

Let's think about this for just a bit:

- My evil (ISP / company / government) can easily transparently proxy my HTTP traffic.
- They can read my mail (but they can do that anyway right?) True, but they can now also:
 - search all my mail
 - search/download my contact list
 - see my browsing history (...!...)
 - send mail as me [not spoofed]
 - Adswords/Groups/Analytics/etc <?>
 - Start their own social network :)
 - Scare Google and imitate a CSS worm...:>

It's not really a Gmail issue – any web mail system – Gmail just has many nice features.

How do they fix this?



A different thought

Your life story in no more than 5 pages
...A.k.a your resume'

Once you get someone's resume' you know all about the
person

You can search for it ...or...

You can get people to send it to you

Recruitment is easy:

Post a job ad and wait for people to send their life story

You can even specify which types of people...:)

*“Looking for nuclear scientist/engineer with experience in
Uranium enrichment and military background. Earn top
dollar, 401K plan, dental coverage, 25days leave. Flexi time
Apply within...”*



Data mining - Prevention

Trick question...

Who is better of?

John Doe

- Has 4 email addresses
- Is on FaceBook, Myspace and Orkut
- His CV is on the 'net and he tells his life story
- Has his own domain [johndoe.com]
- Lists his personal telephone number and address in whois

Sue Nabrixy:

- Uses kloekkloek@hushmail.com
- Never use her real name
- Use throw-away email addresses when registering



Demo: throw away email addresses



Data mining Prevention

Sometimes it's good to be anonymous...

The techniques for individuals and organisations are very different...especially when the org needs 'technical services'.

Consumers vs. Producers on Web 1.0 and how that has changed today.



Prevention – organisations

What if you have to have a web site & infrastructure?

Web sites:

- CAPTCHAs
- Spider traps on the web site
- roelof<a>paterva dot com and generic email addresses
- Using images for telephone numbers(?)
- Keep those XLS, CSVs off your site!
- Robots.txt / Sitemap.xml (?)

But, all of these things seems to be counter-intuitive for making an easily accessible site.

Again – we want to protect against data mining, but we want instant access to information...see slide 1



Prevention – organisations

Infrastructure etc:

- Use generic address to register domains – e.g. domains@abc.com, and where possible, use different ones
- Beware of email bounce discovery!
- Keep your fwd DNS zone as generic as possible (ip-127.0.0.1.abc.com). Make sure you control zone transfers!
- Keep your rev DNS zone as clean as possible.
- Keep as much away from your real network - NS/MX/www
- Have policies in place regarding the use of your domain, blogging, using office telephones for private use.

This can be transparent to the end user, but it is only preventing discovery of your infrastructure.



Prevention – individuals

If we are not talking passive...and in real time:

Spying on your own LAN

- IP address / mac address
- Broadcasts / multicasts / unicasts
- Software updates
- IM
- ...and then some...POF..etc..resistance is futile...

On the Internet [tracking you...if you come to me]

- Your IP address / network range
- Via Proxies [HTTP_X_FORWARDED_FOR]
- Cookies that never grow stale
- Browser leaks [pick a number, 1. CSS history stealing]
- Your IP inside headers from (web) mail services

Use disposable email addresses

...but keep in mind...

Use TOR / Privoxy

And of course...don't give out your details



How do you surf/search anonymously?

Use a proxy

- Hides your IP address [errr..really?]
- Does not hide your cookie

Clear your cookies regularly

- Doesn't prevent them saving your terms
- Can still track on IP address

Use something like Scroogle(.org)

- “we don't use cookies
- we don't save your search terms
- logs are deleted every 48h”

TOR [tor.eff.org] & Privoxy [privoxy.org]

- TOR = different IP address every time
- Privoxy = manage cookies, DNS requests, and other



Demo: Browser leaking demo (MrT)



Data mining - Detection

It's called passive for a reason...but there are tell tale signs:

- Monitor your DNS servers for signs of brute force
- Check your web server logs for mirroring
- Some social networks shows you your profile has been visited (Orkut for one)
- Inspect the referrers in your web server logs for referral from search engines...and the search term.
- Set up Adswords honey words..



Honey words

I run a super secret project called **Sookah**.

I don't ever want people to know about it.

When someone search for the word Sookah I want to know it
leaked out somehow

I don't want them to find out that I know

I register an Adword...isn't Google wonderful?



Getting Started Latest Headlines
Google sookah Search [Advanced Search](#) [Preferences](#)
 Search: the web pages from South Africa

Web Results 1 - 10 of about 449 for sookah. (0.14 seconds)

SoundClick song info: Sookah by Hadjee and the Wartones - Modern ...
 SoundClick band page for Hadjee and the Wartones: band bio and Traditional Arabic MP3 music downloads.
www.soundclick.com/bands/songInfo.cfm?bandID=471024&songID=4380477 - 16k - Supplemental Result - [Cached](#) - [Similar pages](#)

Lyrics for "Sookah" by Hadjee and the Wartones - SoundClick song info
 About "Sookah": the sun rose and so did she, with a renewed vigor and energy, determined to meet her daily destiny thru the trials and tribulations of the ...
www.soundclick.com/bands/Lyrics.cfm?BandID=471024&songid=4380477 - 13k - Supplemental Result - [Cached](#) - [Similar pages](#)

sookah's Favorites » Arab-Zone.com
 Arab-Zone.com. Meet Arab People, Make new friends... Blogs, Friends, Groups, Emails and alot more!
www.arab-zone.com/sookah/favorites/ - 16k - [Cached](#) - [Similar pages](#)

sookah's Profile » Arab-Zone.com

Sponsored Links

Register now
 Best priced domain and web hosting. Register your domain today!
www.register.com

Getting Started
 Google Ad
 Campaign Account Snapshot
 All Campaigns
 + Create a new campaign

<input type="checkbox"/>	Campaign Name	Current Status	Current Budget [?]	Clicks	Impr.	CTR	Avg. CPC	Cost
<input type="checkbox"/>	Sookha	Active	R15.00 / day	0	1	0.00%	-	R0.00
Total - all 1 campaigns		-	R15.00 / day active campaigns	0	1	0.00%	-	R0.00

[Learn how your account settings affect your ad performance.](#)

Reporting is not real-time. Clicks and impressions received in the last 3 hours may not be included here. Time zone for all dates and times in data tables, reports, and billing: (GMT+02:00) Johannesburg. [Learn more.](#)

Starting page is: **Campaign Summary (this page).**
[Make Account Snapshot my starting page.](#)

Part II : Using the information

Hackers are not good at applying information.
They are devious – but not outright criminal.

I'll give it a shot though...



Using the information

Hit & run

- Spoof email from the FD to employees (& Bloomberg) stating the CEO has resigned / company is insolvent / sell your shares / etc.
- Register a site in the name of the holding company's director. Mirror a porn site. Populate. Spoof mail from a techie at the sister company to everyone about the “discovery”. Make it look like a CC that went wrong.
- Spoof SMS from the FD's mobile phone to a high profile investor about corruption in the company. Watch the share price drop, buy low, sell high.

But this is kid's stuff and is easy to spot. Timing however is everything...



Using the information

Let's assume we create a 'information footprint' of company X using the framework.

After a couple of hours we know who are directors at company X and

- Their email addresses – their private ones too
- Their hobbies
- Their social network – clubs they belong to, their friends
- The location of their blogs

We also have a list of email addresses of their employees

So...



Using the information

Personal online identity theft

If these people don't have a strong online presence we will help them out:

- Create a Gmail (or pick any other widely used free email) address in their name
- Register them (and their friends) on LinkedIn
- Register them on MSN/Skype/Google Talk
- Register them on MySpace (or pick appropriate social network)
- Create a blog for them on Blogspot/WordPress
- Subscribe them to a couple of mailing lists
- Oh...and you might register a similar domain for them (like...phishing you know)

Can you automated this ? (even with CAPTCHAs?)

We've collected all the info to re-create them very life like



Using the information: identities are grown, not born

People don't appear from nowhere

We need to give their new identities credibility

If you Google someone...they magically exists

Blogs and posts are time stamped

- Unless we control the underlying infrastructure (although its more effort and a single point of failure)

Thus – we need to grow these identities because we can't manipulate the time line

- Arb postings to mailing lists/blogs
- Sign a “guestbook” (remember those?)

Basically get the identity out there on anything that's indexed – and will thus show up later.

It can take months...perhaps years..if you do it right

If we investigate it – it must look real.



Using the data

Starting a campaign

- Now that I am you and your board I need a mission
- It actually boils down to plain old marketing

- People are flock animals
- So first create a virtual flock of people.
 - Grow many more identities (we know the players)
 - Manipulate counters (its SO 90s)
 - Manipulate Votes / polls
 - Generate automated comments / forum
 - ...and others will follow

Seems like a lot of hard work, but
...watch this space...



Coming soon to blog near you: automated comments

/code/autosent roeloftemmingh\$ perl generate.pl data-disagree/

- I don't understand this. The ideas you spoke about are honestly bullshit. I have never seen such a load of bullshit.
- Our company basically do not figure your statement. The fundamentals here are wrong! I have not experienced this type of situation! Do you believe this nonsense? This is totally fabricated.
- I have surely never seen this type of situation. Get a life. You honestly believe in this nonsense? This is F^%#ING appalling.
- We do not get this thinking.
- This idea is unbelievably bad! You consider this waste? This is nonsense!
- F\$%& THAT!
- This posting is bad! The thinking here is full of shit.
- I have never ever seen any of the things listed here. Dude, you should be spending time on other stuff!
- This is f#@ing ridiculous. Do you consider this bull? C'mon - think about it.
- Your company ought to be researching more interesting things.
- We utterly do not figure this article. This argument is rubbish.

CAPTCHAS??



Go watch “Wag the dog” again...

...but think of the Internet of today.



Detecting and preventing these attacks

- Detecting:
 - Search the Internet for yourself, your company
 - Sweep social networks for your name
- Preventing:
 - Become a celebrity - :)
 - Have a really common name / surname
 - Remember the domain race of 5 years ago?

And for service providers:

- How about the ability to truly identify someone?
- Resists the urge to collect information that might be 'interesting later'.



Trick question revisited

We know everything about John Doe

- He is an easy target for social engineering.

Jane Doe 'does not exist'

- She is an easy target for impersonation.

The same goes for organisations.

Remember the dilemma?



Thus...in conclusion

- As long as there is a demand for it, companies dealing in individual's information will continue to flourish.
- The way that these companies handle information is driven by the average netizen..and
- The new generation of Internet users are not privacy aware.
- As this behaviour becomes more acceptable, more invasive technology will appear (“ don't worry child – it's just a Google Bot “)



Thus...in conclusion

The mouse is mightier than the pen

- Security experts tend to focus on technology itself, ignoring the application and surroundings of it's use.
- The web 2.0 contains great tech (?secure?) but little is known about the security implications when the tech is actually used.
- Real criminals don't write buffer overflows – they follow the route of least resistance.
- Mainstream criminals tend to lag behind. We knew about phishing attacks back in 95.
- What will be on their minds in 2010?
- I am guessing it would be something close to this...

