Who wants to snoop on your Internet traffic?

Dr Richard Clayton

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

fipr

BCS, Hertfordshire
30th September 2009

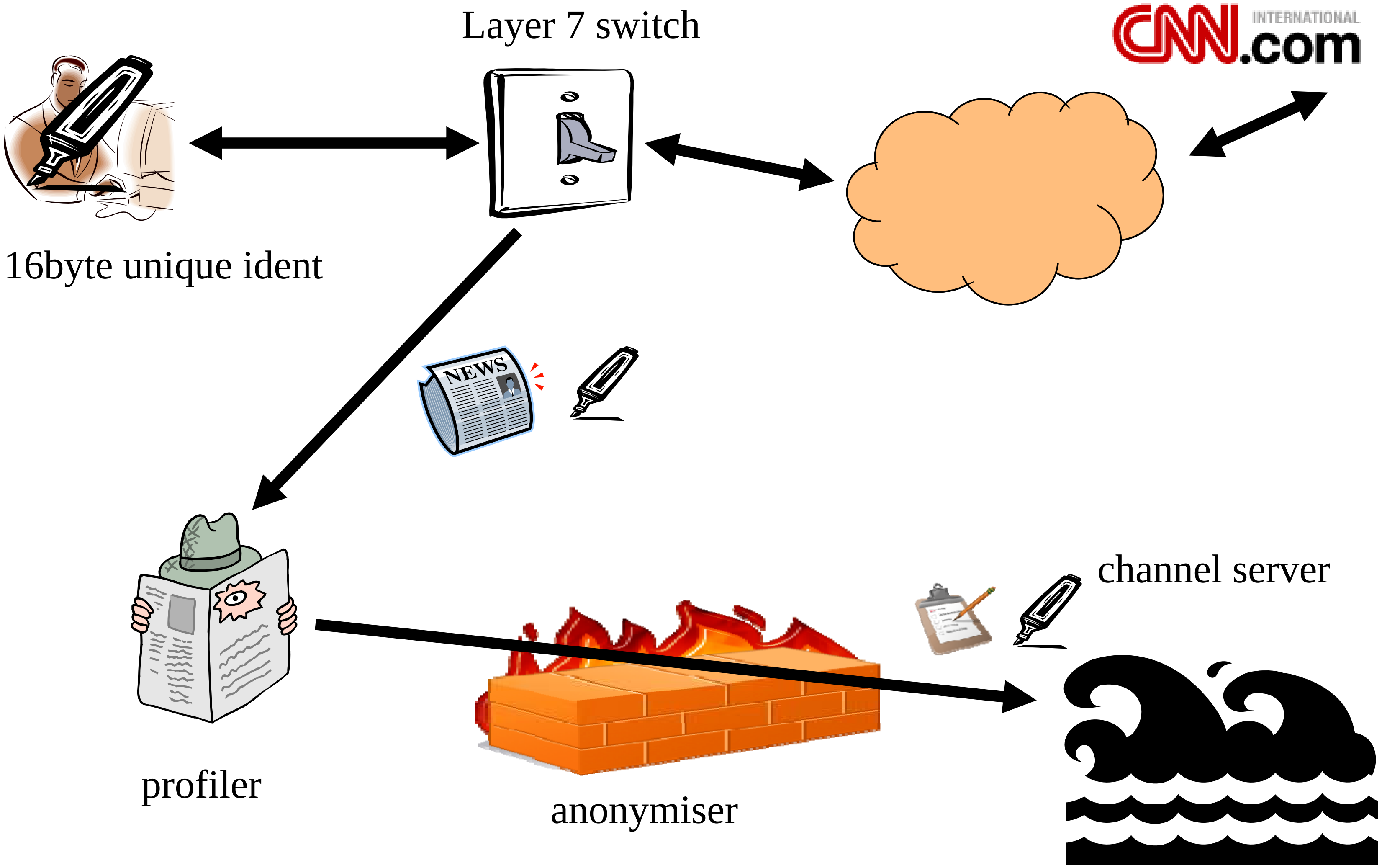
Overview

- Phorm
- Great Firewall of China (GFC)
- Peer-to-Peer (p2p)
- Internet Watch Foundation (IWF)
- Interception Modernisation Programme (IMP)
- ...and many more

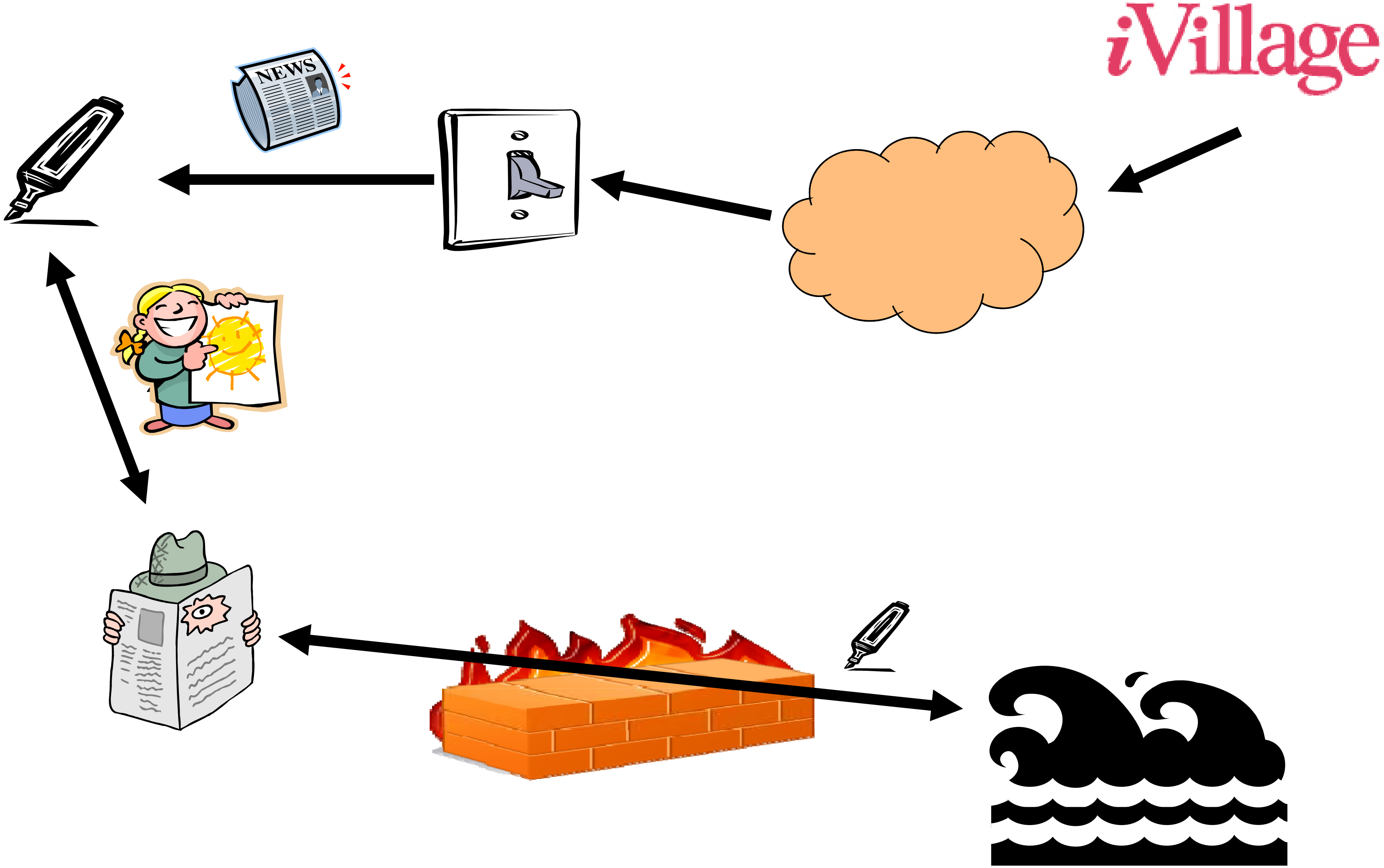
Behavioural advertising

- Advertising is big bu$ine$$!
- Basic Google model is “put ads on relevant pages”
- Alternative approach is “show ads that are relevant to people who happen to visit”
  - DoubleClick tracks visits to participating sites by cookies (returned to DoubleClick)
  - Phorm proposed to inspect HTML on *(almost) all* visited pages to deduce nature of content, then serves relevant advert if you visit a participating site
- Advertisers want to know what you do, not who you are
  - they break people down into categories
  - ABC1, “empty nesters”, lots of fancy new names….
- So they can live with anonymity

Phorm design #1



Phorm design #2



Distilled pages

attack
CleanFeed
format
inquiry
ISP
legal
packets
paper
PDF
system

content
document
event
partners
Phorm
PIA
privacy
School
system
Thinking

advertising
consumers
leading
OIX
online
Phorm
technology
Virgin
websites
Webwise

This is what all those rubbish search engines
used to do before Google came along!

Channel server

- Channel server is also told about URLs
- Hence channel server is also told of search terms (Google &c keep them within the URLs) and these are then mined
- Channel server only learns UID not IP address
    - hence some “anonymity” properties
- Channel server matches top 10 words against advert “channel”
- Then records just the UID and time against channel
- When user visits a partner site, sees best (££) advert for matches like “has visited 3+ travel sites in the past week”
- Phorm promise rules about channel specifications that would prevent identification of individuals

Dance of the cookies

#1 faked cnn.com response redirects user to “webwise.net”

```
GET cnn.com/index.htm
307 webwise.net/bind?cnn.com/index.html
```

#2 16byte UID allocated and sent in webwise.net cookie

```
GET webwise.net/bind?cnn.com/index.htm
307 webwise.net/bind-2?cnn.com/index.html
```

#3 check that the user is returning webwise cookies

```
GET webwise.net/bind-2?cnn.com/index.html
307 cnn.com/magic?cnn.com/index.html&UID
```

#4 faked cnn.com responds with a faked cnn.com cookie

```
GET cnn.com/magic?cnn.com/index.html&UID
307 cnn.com/index.html
```

#5 now permit access to real cnn.com, since cookie holds UID

```
GET cnn.com/index.html
```

Opting-out

- User can opt-out with a webwise.net cookie
- User can effectively opt-out by refusing to return webwise.net cookies (or cnn.com cookies)
- But note that deleting all cookies will set you back to default state (and opt-out is forgotten)
- User will have significant problems if they set webwise.net to resolve to 127.0.0.1
  - System supposed to disable itself if lack of browsing progress
- ISPs looking at network level opt-outs
  - presumably RADIUS setting to select IP pool
  - some hints that this turned out to be complicated...

Opt-in versus opt-out

- May be processing "sensitive personal data" (religion, trade union, medical etc)

  `<h1>Union advice for vicars living with AIDS</h1>`

  - DPA requires an informed opt-in for this
- Information Commissioner says that Privacy and Electronic Communications Regulations requires an opt-in
- But it's illegal wiretapping so opt in/out irrelevant:
  - RIP 2000 requires permission from *both* ends of communication
  - RIP s16 shows Phorm keywords do infringe
  - whatever user says, permission for data TO servers not given
  - whatever user says, permission for data FROM servers not given
  - whatever user says, permission from THIRD PARTIES not given
    - think "email" or "web forum"

Privacy

- Privacy and Data Protection are not the same!
- Data Protection just mechanistic approach to controlling corporations with mainframes
  - and UK has minimal watered down variant
  - to a first approximation, anonymity fixes everything
- Privacy relates to controlled disclosure of information that matters TO YOU
  - your privacy is violated even if you are anonymous
- ANALOGY: Suppose the Post Office opened all your letters, so you can get a better class of junk mail

Great Firewall of China

- Chinese firewall shuts connections if it spots specific keywords passing by
  - for example GET /?falun HTTP/1.0
- Keywords spotted as they pass by in both directions (dealing with requests & results)
- CAUTION: parts of Chinese system DO use other blocking methods, the academic network isn't currently using the scheme & other protocols are blocked at the application level!
- Shutting of connections is done by sending TCP reset packets
- If you ignore these packets your connection is unhindered!

Example packet trace

```
cam(54190) → china(http)[SYN]
china(http)→ cam(54190) [SYN, ACK] TTL=39
cam(54190) → china(http)[ACK]
cam(54190) → china(http) GET /?falun HTTP/1.0<crlf><crlf>
china(http)→ cam(54190) [RST] TTL=47, seq=1, ack=1
china(http)→ cam(54190) [RST] TTL=47, seq=1461, ack=1
china(http)→ cam(54190) [RST] TTL=47, seq=4381, ack=1
china(http)→ cam(54190) HTTP/1.1 200 OK (text/html)<crlf>..
cam(54190) → china(http)[RST] TTL=64, seq=25, ack zeroed
china(http)→ cam(54190) . . . more of the web page
cam(54190) → china(http)[RST] TTL=64, seq=25, ack zeroed
china(http)→ cam(54190) [RST] TTL=47, seq=2921, ack=25
```

Blocking peer-to-peer traffic

- Rights holders are concerned about file sharing of copyrighted material (they believe it is costing them money)
- Music industry is joining in to the networks to determine the identity of peers, but they find this slow and expensive
- So they'd like to see technical measures taken by the ISPs
- ISPs concerned about traffic implications of widespread use of file sharing protocols
- ISPs actually control the networks, so their concerns have been addressed for some time:
  - traffic shaping (slowing things down)
  - traffic blocking (stopping it altogether)

Peer-to-peer traffic detection #1

- Once upon a time you could tell what traffic was by looking at the port number (25: email, 80: http, 53: dns, 6699: napster)
- Firewalls stopped this being so useful (80 & 53 go through)
- So initially you could categorise peer-to-peer by port numbers
- But, once these ports began to be blocked software evolved to use many different port numbers (and/or just port 80!)
- ISPs then deployed “deep packet inspection” kit to look for telltale signs of p2p protocols:

```
69 74 54 6f 72 72 65 6e 74 20 70 72 6f 74 6f 63 6f 6c 65 78
80 2c 01 03 01 00 03 00 00 00 20 00 00 34
```

- Protocols started to use encryption
  - albeit compatibility may leave handshakes in plain text

Peer-to-peer traffic detection #2

- Peer-to-peer traffic is (fairly) distinctive:
    - connections to multiple peers, patterns of traffic in and out
- Trend towards using heuristics to decide what is p2p
- This is fine for ISP (who wants to reduce usage)
- Useless for rights holders – since there are important non-infringing uses of file sharing technology
    - standard examples: World of Warcraft patches; Linux distros
- Music industry was much enamoured of “Audible Magic”
    - Picked apart p2p protocol to extract payload
    - Identified payload by signal processing and checking dictionary
- Encryption makes Audible Magic a non-starter today
    - Belgium (Scarlet) case drags on; Irish (eircom) settled out of court

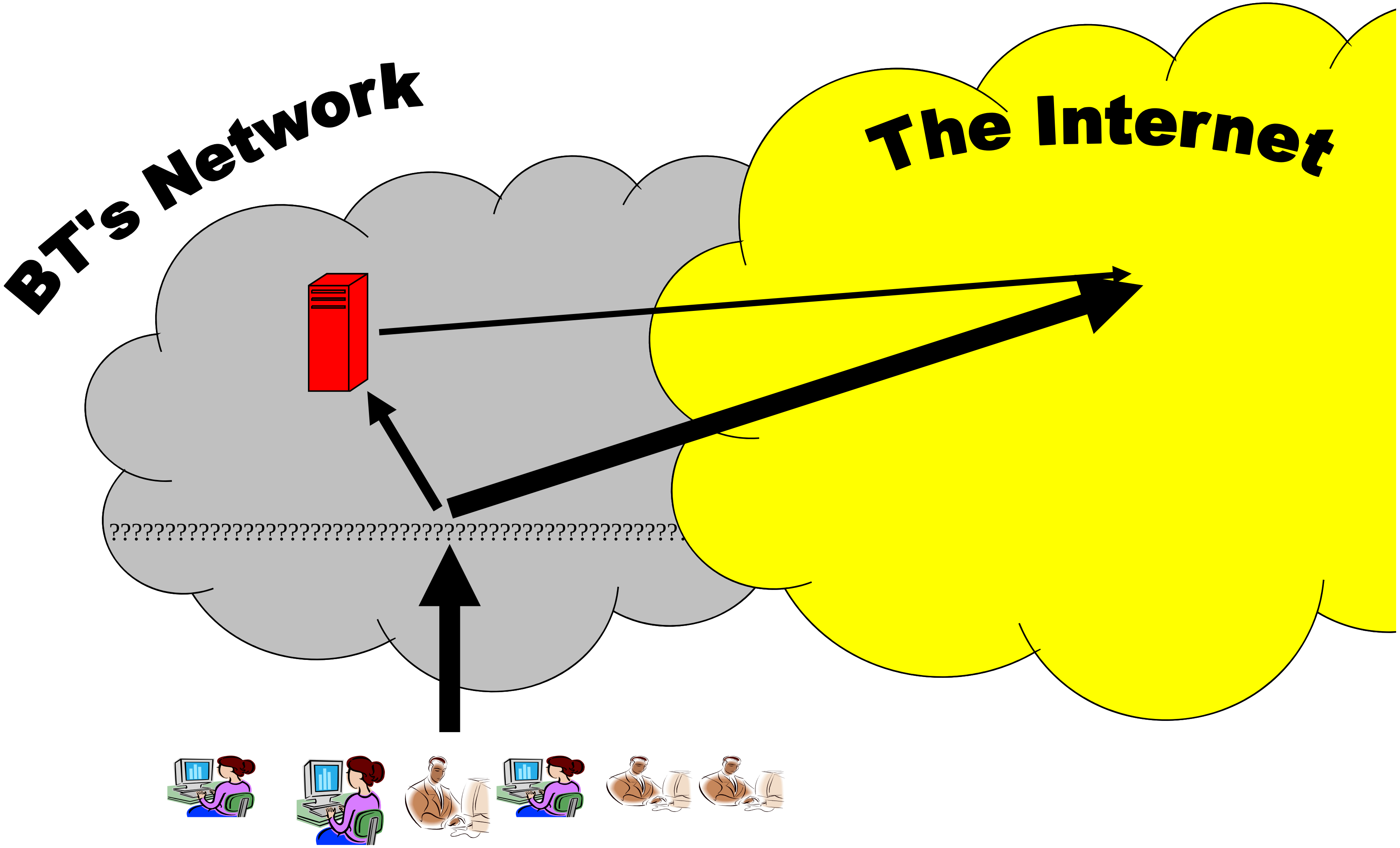
The IWF

- Internet Watch Foundation
- Set up in 1996 to address issue of child pornography on Usenet
  - phrases “child pornography” or “kiddy porn” seen to trivialise issue
  - politically correct term became “child abuse images” (CAI)
  - or rather more recently “child sexual abuse images”
- IWF operates a consumer “hot-line” for reports
- UK institution, but cooperates via INHOPE with other hotlines
- Funded by industry and also by EU (seen as leading light)
- Now mainly concerned with websites
- Has a database of sites not yet removed (for efficiency)
- Database now underpins various blocking systems

Taxonomy of blocking methods

- DNS poisoning
  - refuse to resolve the wicked domains
  - low cost, and highly scalable
  - overblocks (since all of geocities.com is affected)
- Blackhole routeing
  - refuse to carry the traffic to the wicked site
  - low cost, but limits to size of ACLs/routing-table
  - also overblocks, and struggles with “fast-flux” systems
- Proxy filtering
  - refuse to serve the wicked pages
  - high cost, and all traffic has to be inspected
- BT’s CleanFeed (2004)
  - combined custom iBGP routing with a proxy system

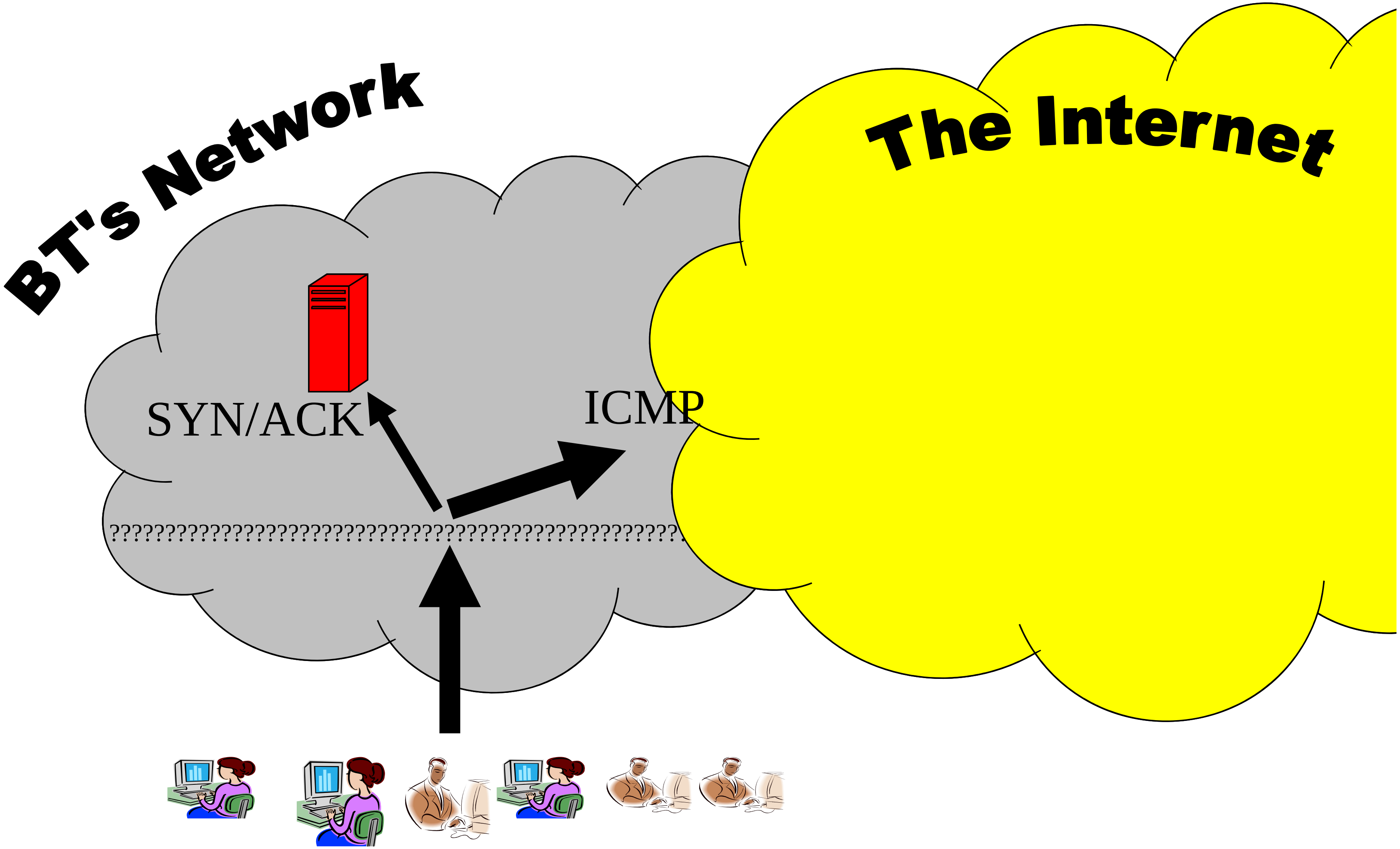
Design of CleanFeed



CleanFeed

- Part of BT “anti-child-abuse initiative”
  - two stage (hybrid) system, BT, June 2004
- First stage is IP address based
  - candidate traffic for blocking is redirected
- Second stage matches URLs
  - redirected traffic passes through a web proxy
- Best of both worlds?
  - highly accurate
  - but can be low cost because #2 is low volume
- BUT easy to avoid
  - use an external proxy or Tor
  - use HTTPS, or anything not on port 80
- AND can be reverse engineered
  - Raising public policy issue: does it do more harm than good?

The oracle attack



The oracle attack on CleanFeed

- Detect the redirection by the first stage by seeing what traffic reaches the second
- Send **`tcp/80`** packets with TTL=8, see what then comes back:
  - ICMP time exceeded means no redirect
  - RST (or SYN ACK) means redirect to proxy

```
17:54:28  Scan: To [~~~.~~~.191.38] : [166.49.168.9],  ICMP
17:54:28  Scan: To [~~~.~~~.191.39] : [166.49.168.1],  ICMP
17:54:28  Scan: To [~~~.~~~.191.40] : [~~~.~~~.191.40], SYN/ACK
17:54:28  Scan: To [~~~.~~~.191.41] : [166.49.168.13], ICMP
17:54:28  Scan: To [~~~.~~~.191.42] : [~~~.~~~.191.42], SYN/ACK
```

- Then use a suitable database to get domain names
  - eg: **`whois.webhosting.info`**

```
~~~.~~~.191.40   lolitaportal.****
~~~.~~~.191.42   no websites recorded in the database
```

- Attack cannot be fixed, just detected
  - and it works against other 2-stage systems as well

Whitehall comprehension?

- Blocking considered “impossible” until BT deployed CleanFeed
- Parliament told: *“Recently, it has become technically feasible for ISPs to block home users’ access to websites irrespective of where in the world they are hosted”*
- In my view, doubtful that they actually understood the cost, fragility or ease of evasion of these blocking systems, let alone the reverse engineering of the blocking lists.
- Ministers want all (consumer?) broadband suppliers to filter
  - original target date of end of 2007 else “review our options”
- ISPA claimed 80% (more recently 95%) of consumers covered by systems that block illegal child images
  - methodology for count unclear (& not all ISPs filter all customers)

Most (all?) UK filtering is proxy based

- Comparison of URLs in proxy means no “overblocking”
- Proxying all web traffic very expensive (and other downsides)
- So select only traffic that might need filtering

1. DNS poisoning
   - resolve dubious domains to address of web proxy
   - low cost, and highly scalable – widely used in UK
   - assumes customers using the local DNS server!
2. custom iBGP
   - resolve dubious domains and route their /32 to web proxy
   - mechanism used by BT’s “cleanfeed” system
3. exotica (DPI, WCCPv2 etc)
   - can have scaling issues, so used mainly by smaller ISPs

Wikipedia

- Member of public reports Virgin Killer album cover to IWF
- IWF conclude it is an indecent image, and add URLs to blocklist
- List rolled out midday Friday December 5th 2008
- Large numbers of UK accesses to Wikipedia now proxied
  - this breaks Wikipedia security model!
- Mechanism rapidly identified, as is particular image
  - propriety of keeping image debated in May 2008
- Many instances of image located (some on Amazon US)
- On Monday 8th IWF considers Wikipedia “appeal” & rejects it
- On Tuesday 9th IWF board decide to remove URL from list
- Wikipedia blocked elsewhere for some time thereafter!

What was blocked?

- #1: Main page was blocked
  - http://en.wikipedia.org/…/virgin_killer
  - blocked entire text about The Scorpions album, not just the image
- #2: Image description page was blocked
  - http://en.wikipedia.org/…/Image:Virgin_Killer.jpg
  - this is also a text page (despite the URL!)
- Did not block ../Virgin_Killer (there are four duplicate URLs!)
- Some blocking systems were case sensitive, some were not
- Caused considerable confusion as to what blocking was in place
  - general lesson about this event and the archive.org event; most consumer reports were almost entirely inaccurate!
- Evidence that some ISPs did not block until Monday
  - possibly just slow, possibly because a high-traffic website

What is the IWF currently blocking ?

- Latest idea (NB: does not access the sites, since that’s illegal!)

```
for $hostname in (list of all valid hostnames)
   if (resolve(hostname) == cache-IP-address)
       print “hostname is blocked”
```

- List of hostnames comes from ISC “passive DNS” dataset
  - systems collecting anonymised copies of DNS responses
- *c* 120 million hostnames – 40 million are DNSBLs etc
- Further clean-up gives *c* 70 million hosts to check
- Takes about 2 days (and 22Gbytes) over home ADSL
- NB: does not identify URLs, merely hostnames

Current results (this is ongoing research)

- IWF list currently holds about 450 URLs (says a mole)
- 40% not yet identified by the methodology (too obscure?)
- 35% clearly (from hostname) intentionally wicked
- Remaining 25% are legitimate “free” hosting sites (etc)

100free.com, 2st.jp, 3dn.ru, 4shared.com, 50webs.com, adultdreamhost.com, adultshare.com, awardspace.biz, awardspace.info, bbs.zgsm.com, beam.to, boulay.be, byethost3.com, clan.su, club.telepolis.com, depositfiles.com, dump.ru, filehoster.ru, freeforum.tw, funkyimg.com, gayhomes.net, gratisweb.com, grou.ps, hotshare.net, i037.radikal.ru, image5.poco.cn, imagecross.com, imagevenue.com, imgsrc.ru, indexjunkie.com, ipicture.ru, letitbit.net, mail.su, megaupload.com, multipics.net, my1.ru, nakido.com, oo.lv, opendirviewer.com, pic.ipicture.ru, pic2us.com, picsbuddy.us, pornhome.com, pornspaces.com, pridesites.com, rapidshare.com, sapo.pt, sendspace.com, surge8.com, uploading.com, uppic.net, zshare.net

IWF removal process

- Bank phishing websites removed in 4 hours (when known about), 2 days (fast-flux systems), 10 days (not known about)
- Part time volunteers remove scam websites in 1-7 days
- Child Sexual Abuse Image sites: average lifetime ~ 4 **weeks**
- Only thing removed slower is fake pharmacy websites
  - and they are not tackled by any group we can locate
- We were amazed to uncover this, and consider it a scandal
- Main reason appears to be lack of prompt contact with hosters
  - IWF “not authorised” to contact foreign hosting providers
  - INHOPE rules mean local hotline must act, not the IWF
  - IWF not going after domain names, only the hosting
  - IWF (& INHOPE) confused as to whether aim is to remove content or to catch the criminals

Interception Modernisation Programme

- Spooks would like to get their hands on “traffic data”
- Although “content” is interception, “traffic data” is almost as valuable since it shows who communicates with who
- Idea is to use DPI equipment to snoop on all UK citizens
- Classic (1998) dumb question: “this IP address I’ve found in a web log, who’s it belong to? And what’s their Hotmail address?”
- BUT under IMP proposals, DPI can pick apart webmail HTML pages to identify Hotmail identities and email correspondents
- DPI can pick apart Second Life protocols to see which avatars you were near in (x, y, z) coordinates
- DPI can pick apart World of Warcraft protocols to determine which other characters you have been chatting to

Problems with IMP

- But current DPI can’t do all of these things at once
- Plus protocols evolve and what is of interest changes
- Hence the DPI kit has to be remotely reconfigured by GCHQ
- ISPs deeply unhappy about presence of snooping boxes that are not under their control (may be insecure)
- Spooks unlikely to want to say what protocols are currently being targeted except to people with clearances
- Hence day-to-day policing unlikely to benefit from scheme
- Cost looks enormous
- And if traffic is regularly encrypted system is useless
- Home Office have consulted, and are currently cogitating

And no time to mention...

- Criminals
  - “man in the browser” trojans snoop on your eBanking sessions
- Nation states
  - Greek Vodaphone scandal
  - Dutch wiretapping scandal
  - How do we know that the Chinese phone exchange isn’t relaying every conversation to Beijing?
- Partners
  - Internet romances now playing a big part in divorces
- Employers
  - Are you reading Facebook in the office?
  - Are you planning your holiday or doing your shopping?
  - Are you looking at porn?

Summary

- Advertisers want to know what you're interested in
- Some nations want to prevent you becoming interesting in particular topics: such as what really happened in Tiananmen Square back in 1989
- Some ISPs (and some ministers) want to ensure you don't view child sexual abuse images by accident – but they don't think that they're able to get the sites removed from the Internet
- Some spooks think that knowing everything about your Internet activity will make the world a safer place
- Some people think that privacy, and proportionality, matters!

Who wants to snoop on your Internet traffic?

http://www.lightbluetouchpaper.org

http://www.fipr.org

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

fipr