Failures in a Hybrid Content Blocking System

Richard Clayton

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

PET Workshop,
Dubrovnik, Croatia

31st May 2005

Summary

- Content blocking system taxonomy
- The novel hybrid design of CleanFeed
- Simple evasion of CleanFeed
- Attacking the system itself
- Detecting the IWF & CleanFeed
- The oracle attack to locate blocked content
- Conclusions

Content blocking methods

- Blackhole routeing of IP addresses
  - Edelman identified significant overblocking
- Use web proxy to filter if URL match
  - Expensive, at a time when web proxy caches are going out of fashion
- DNS poisoning (do not provide IP address)
  - Dornseif found that often done incompetently

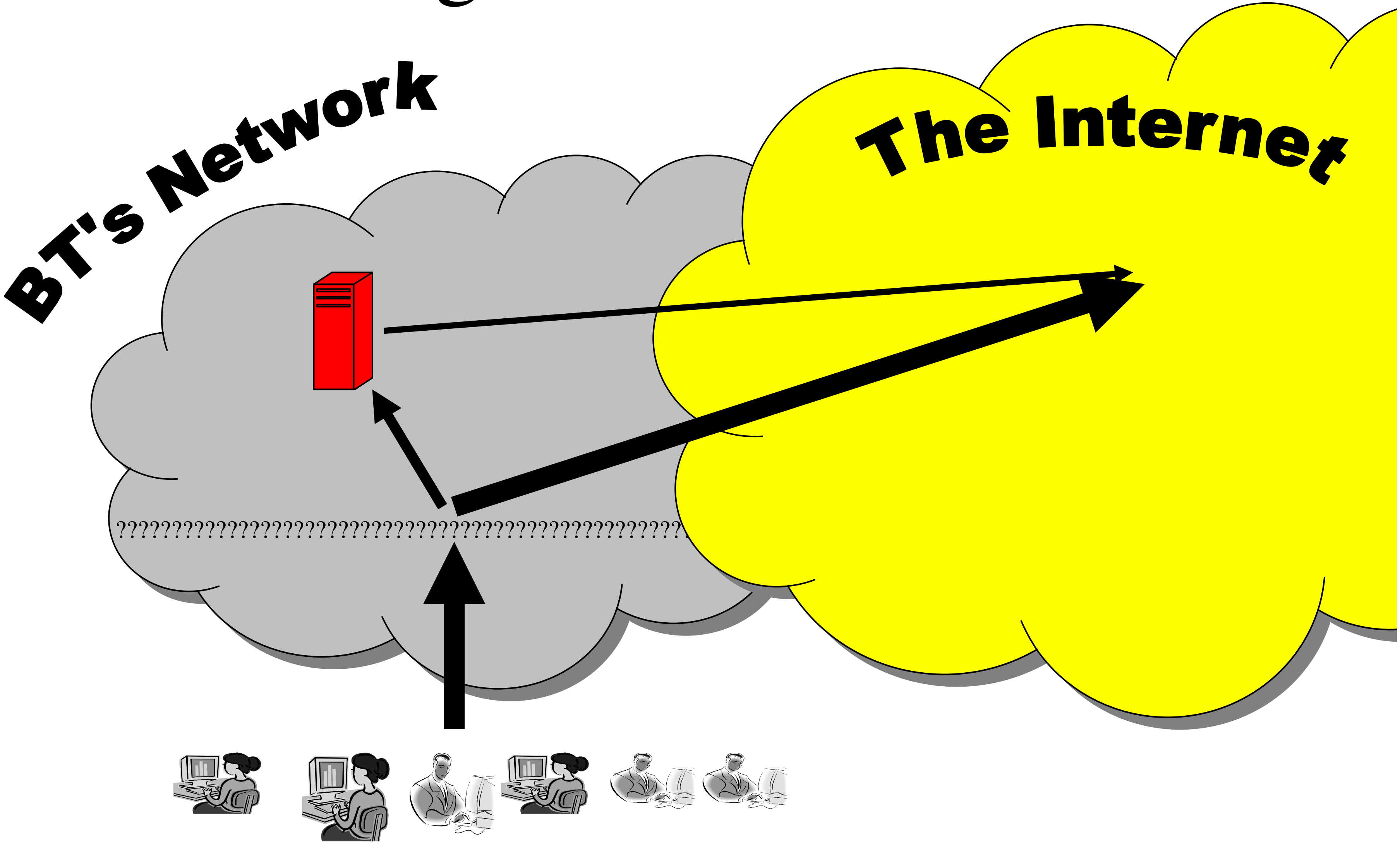
The IWF

- Internet Watch Foundation
- Set up 1996 to address problem of child pornography on Usenet
- Operate a consumer “hot-line” for reports
- Now mainly concerned with websites
- Have a database of sites not yet removed
- Database could underpin a blocking system

Design of 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

Design of CleanFeed



So it’s an elegant design…

… are there any problems with it ?

YES!

Fragility

- Evading either stage evades the system
- Moving IP address or port evades stage #1
- Using unusual escape forms evades stage #2
  - www.example.com/%%37%37ebpage.html
- Lots more ways of breaking the system
    - By user acting alone
    - By content provider
    - By both acting together
- Limited benefits from hybrid design

Can attack the system

- Redirect extra traffic
  - add specious IP addresses into DNS lookup so that high bandwidth sites are sent to stage #2
- Block valid traffic

  google cache: **`66.102.9.104/search=?q=cache:FF9etc`**

  'etc venues': **`195.224.53.128/directions/parkstreet`**
- NB: more efficient when sure is the IWF

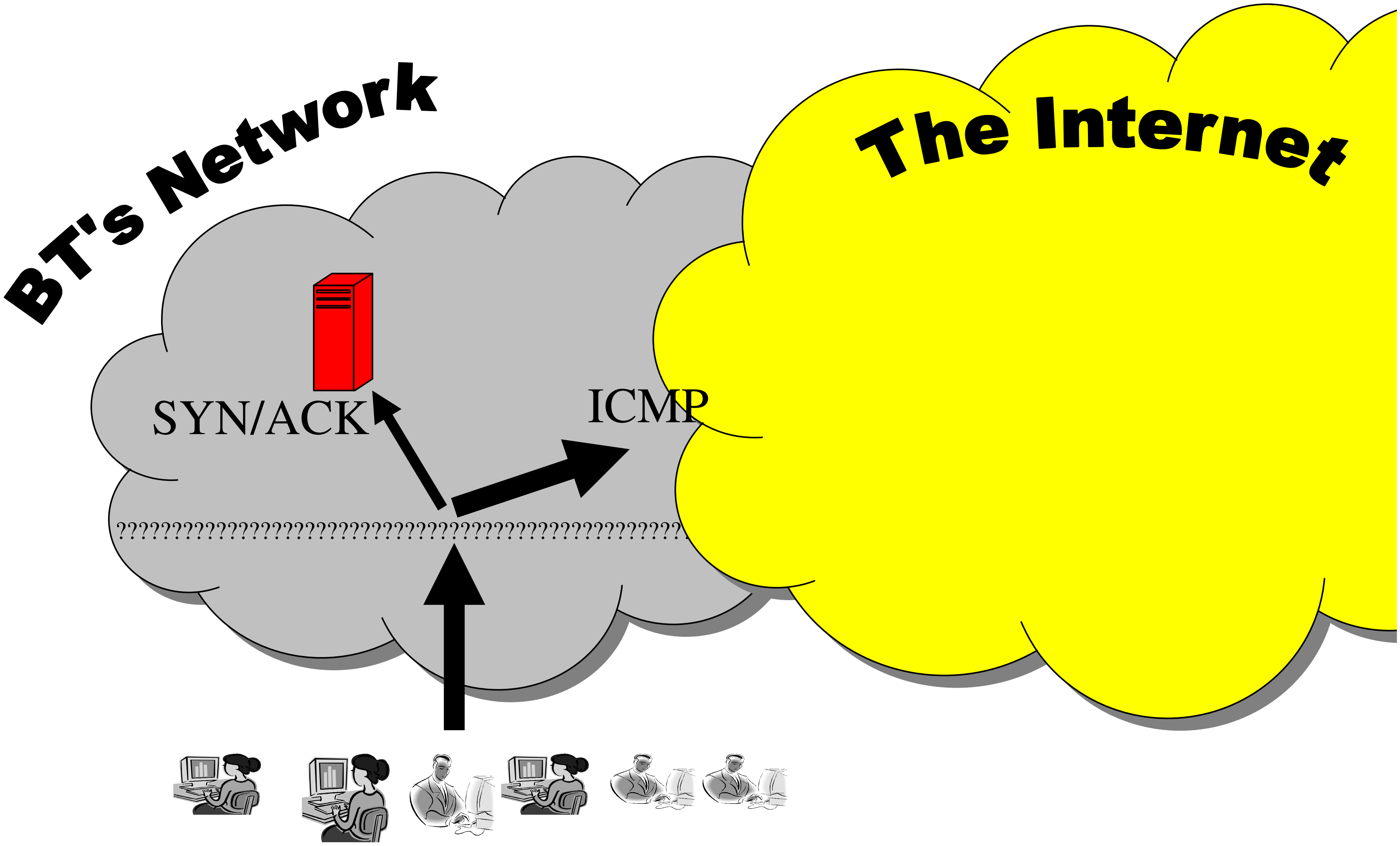
Detecting IWF accesses

- Content providers can self-report
  - provides valuable info about timing etc
  - NB: recognising CleanFeed also relevant
- IWF have a fixed /26 network
  - need anonymising systems (caches, Tor, JAP..)
- Detect multiple accesses for same identifier
  - first AS is (outraged) consumer, second IWF, third the police or other investigators

The oracle attack

- Detect the redirection by the first stage by seeing that traffic reaches the second
- Send `tcp/80` packets with TTL set to 8, see what then comes back:
  - ICMP time exceeded means no redirect
  - RST (or SYN ACK) means redirect to proxy
- Then use a suitable database to get domain names, eg: `whois.webhosting.info`

The oracle attack



Oracle attack results I

```
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
17:54:28  Scan: To [~~~.~~~.191.43] : [166.49.168.9],  ICMP
17:54:28  Scan: To [~~~.~~~.191.44] : [166.49.168.5],  ICMP
17:54:28  Scan: To [~~~.~~~.191.45] : [166.49.168.9],  ICMP
17:54:28  Scan: To [~~~.~~~.191.46] : [166.49.168.13], ICMP
17:54:28  Scan: To [~~~.~~~.191.47] : [166.49.168.9],  ICMP
17:54:28  Scan: To [~~~.~~~.191.48] : [166.49.168.9],  ICMP
17:54:28  Scan: To [~~~.~~~.191.49] : [~~~.~~~.191.49], SYN/ACK
17:54:28  Scan: To [~~~.~~~.191.50] : [~~~.~~~.191.50], SYN/ACK
```

Oracle attack results II

```
~~~.~~~.191.40   lolitaportal.****
~~~.~~~.191.42   no websites recorded in the database
~~~.~~~.191.49   samayhamed.****
~~~.~~~.191.50   amateurs-world.****
                 anime-worlds.****
                 boys-top.****
                 cute-virgins.****
                 cyber-lolita.****
                 egoldeasy.****
                 elite-sex.****
             ... and 26 more sites with similar names
```

NB: missing names probably `.ru` or outdated database

NB: dodgy names on `.41 .43` … BUT no IWF "endorsement"

NB: It is illegal for me to check the ACTUAL contents

Conclusions

- Two stages; means two stages to fail
- You can use one stage to attack another
- Many (and deep) flaws come from relying on validity of content providers data
- Without attacks, assessing expense invalid
- Oracle attack shows risk of worse outcome when full public policy aspects considered

Failures in a Hybrid Content Blocking System

Richard Clayton

`http://www.cl.cam.ac.uk/~rnc1/`

UNIVERSITY OF CAMBRIDGE
Computer Laboratory