Some thoughts on Application Identification and Classification

Andrew Moore
Computer Laboratory
University of Cambridge
andrew.moore@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Roadmap

- Why do networking characterization?
- How to do network characterization
  (and network monitoring...)
- What makes network characterization hard?
- What can we do with network characterization?
- A method for improving network characterization
- Network characterization futures

Why Identification?

(some examples from today’s papers)

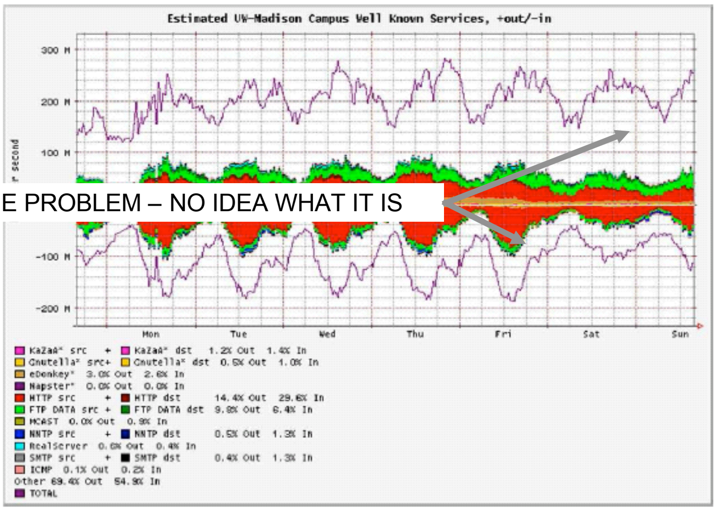
- identifying new applications
    - p2p, botnets, new applications - good and bad
- traffic patterns (traffic analysis)
- identifying better features
- classify and characterize new apps
- smart-networking - application specific routing

Characterise to protect

- Signatures into virus detectors
    - Brad Karp’s Autograph
    - Christian Kreibich’s HoneyComb
- Bad host detection *that guy is port scanning*
    - he is probably a bad guy,
    - a good guy identifying bad machines, (oops)
    - some new application (double oops)

Understanding

traffic for a large university - not Cambridge



THIS IS THE PROBLEM – NO IDEA WHAT IT IS

Traffic Distribution of the network of the University of Wisconsin for the week 7-13 Sept. 2003. Courtesy of wwstats.net.wisc.edu

other

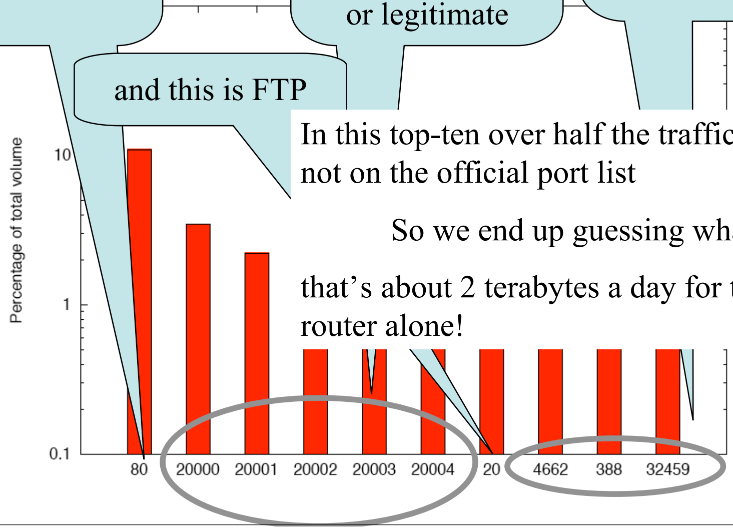
e ISPs r

Port numbers seem helpful, this is web

But these top 5 are either: keyboard loggers or viruses or legitimate

And these three are peer-2-peer and perhaps another virus

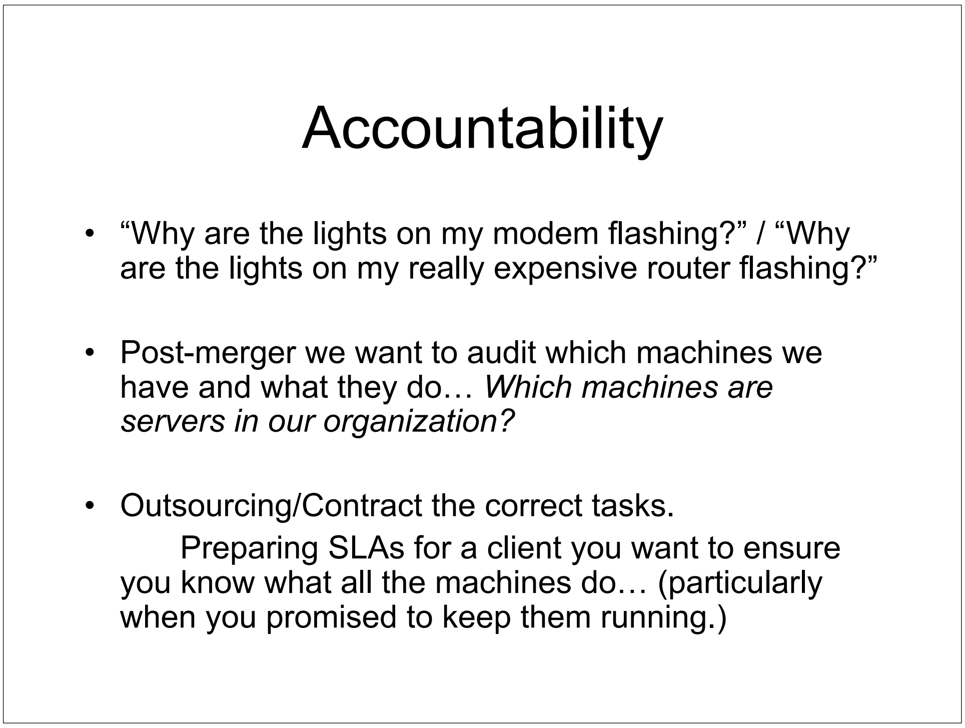
and this is FTP



In this top-ten over half the traffic is not on the official port list

So we end up guessing what it is

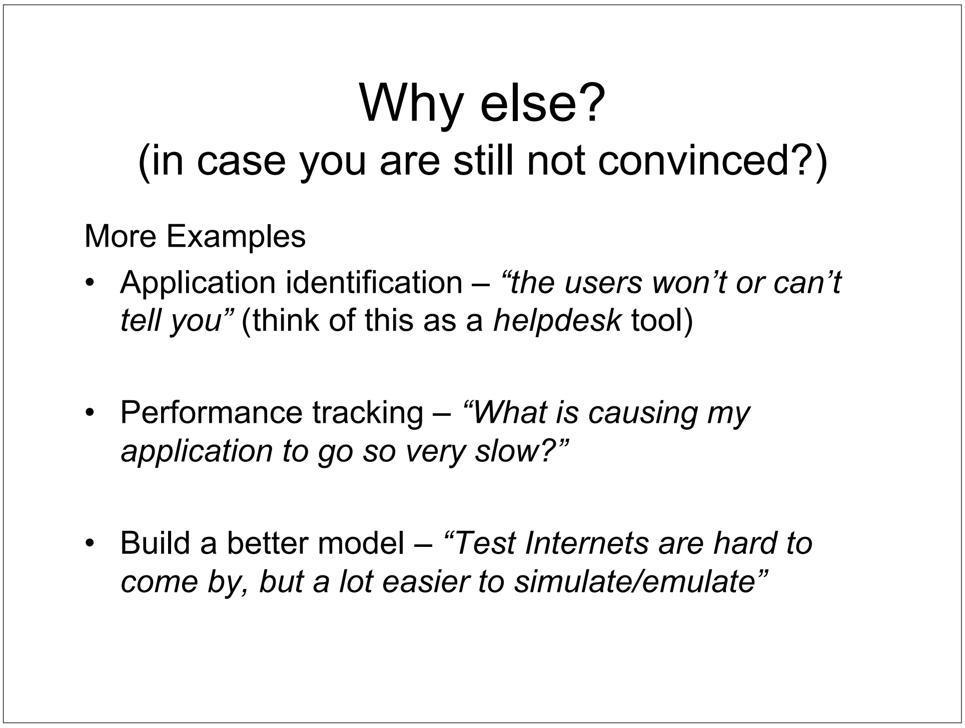
that's about 2 terabytes a day for this router alone!



Accountability

- "Why are the lights on my modem flashing?" / "Why are the lights on my really expensive router flashing?"
- Post-merger we want to audit which machines we have and what they do… *Which machines are servers in our organization?*
- Outsourcing/Contract the correct tasks.

  Preparing SLAs for a client you want to ensure you know what all the machines do… (particularly when you promised to keep them running.)



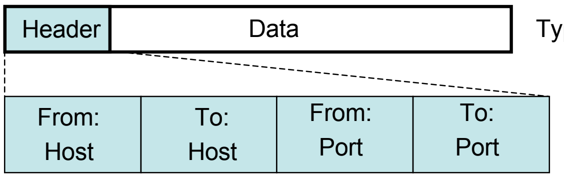
Why else?
(in case you are still not convinced?)

More Examples

- Application identification – *"the users won't or can't tell you"* (think of this as a *helpdesk* tool)
- Performance tracking – *"What is causing my application to go so very slow?"*
- Build a better model – *"Test Internets are hard to come by, but a lot easier to simulate/emulate"*

How do people do this now?

Use packet headers (addresses)



Typical Internet packet

Extract of the header

- Use the port number
- Maybe in concert with the host info
  - *that host is a web server*
  - *this host is a NAT gateway*

Why is this a problem?

For one particular traffic sample...

•Using a port-based method we could not identify 30% of the traffic **at all**

Why? Many ports are not "designated", have unofficial uses or an ambiguous designation

32343: Err no-idea

4662: that would be eMule, but it isn't in any "official" list

•Of the 70% we could identify with port-based schemes a further 29% was **incorrectly** identified

Why? Official port lists don't tell the whole tale

"If I wrap my new application up to look like HTTP it will get through the firewall"

80: HTTP is that a server or a proxy or a VPN or a ...?

Ports as poor practice

- Ports are still used as some sort of definitive classifier
- Commonly by studies examining the effectiveness of new methods

  (using traffic without "ground-truth")
- BUT

  ground-truth error >> evaluation accuracy

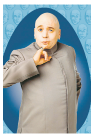
What is an application anyway?

- port 80?
- http on port 80?
- html on http on port 80?
- web page on html on http on port 80?
- So what about gmail?
  - email or web (browser) traffic?
  - What about when my MUA gets the email via the webmail interface?

Email

- MTA vs MUA
- Spam vs Ham

- Commercial vs Domestic
- Decent vs Wicked



Speaking of evil… phishing

- US: $200 million/year
- UK: £30 million/year

(*a nice little earner* - D. Trotter)

- Rock-phish example:
  - Compromised machines run as a proxy
  - Domains do not infringe trademarks
  - Distinctive URL style
    - http://session9999.bank.com.lof80.info/signon
  - Some usage of fast-flux since Feb'07
    (resolving 5+ IP addresses at once)
    limits impact of take-down orders

facts'n'figures stolen from slides by Richard Clayton

Going phishing?

(rock-phish example)

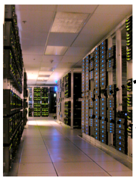
Here is what you will need….

Target

rate of increased availability: 1/minute

(Barnum, P.T. *various*)

Zombie army



Safe, secure, *legitimate* data-center hosting

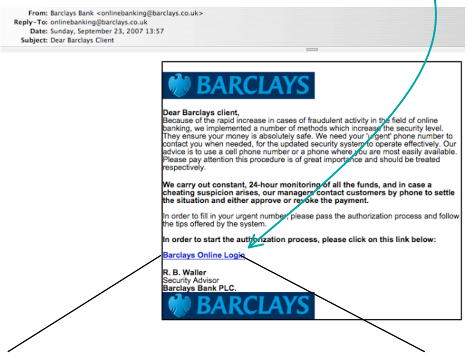
Evil clone-bank

(or just the back-end)

DNS server

(under your control)

Something wrong with my account? well I better click on this embedded link



From: Barclays Bank <onlinebanking@barclays.co.uk>
Reply-To: onlinebanking@barclays.co.uk
Date: Sunday, September 23, 2007 13:57
Subject: Dear Barclays Client

Dear Barclays client,
Because of the rapid increase in cases of fraudulent activity in the field of online banking, we implemented a number of methods which increase the security level. They ensure your money is absolutely safe. We need your 'urgent' phone number to contact you when needed, for the updated security system to operate effectively. Our advice is to use a cell phone number or a phone where you are most easily available. Please pay attention this procedure is of great importance and should be treated respectively.

We carry out constant, 24-hour monitoring of all the funds, and in case a cheating suspicion arises, our managers contact customers by phone to settle the situation and either approve or revoke the payment.

In order to fill in your urgent number, please pass the authorization process and follow the tips offered by the system.

In order to start the authorization process, please click on this link below:

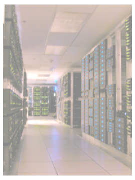
Barclays Online Login

R. B. Waller
Security Advisor
Barclays Bank PLC.

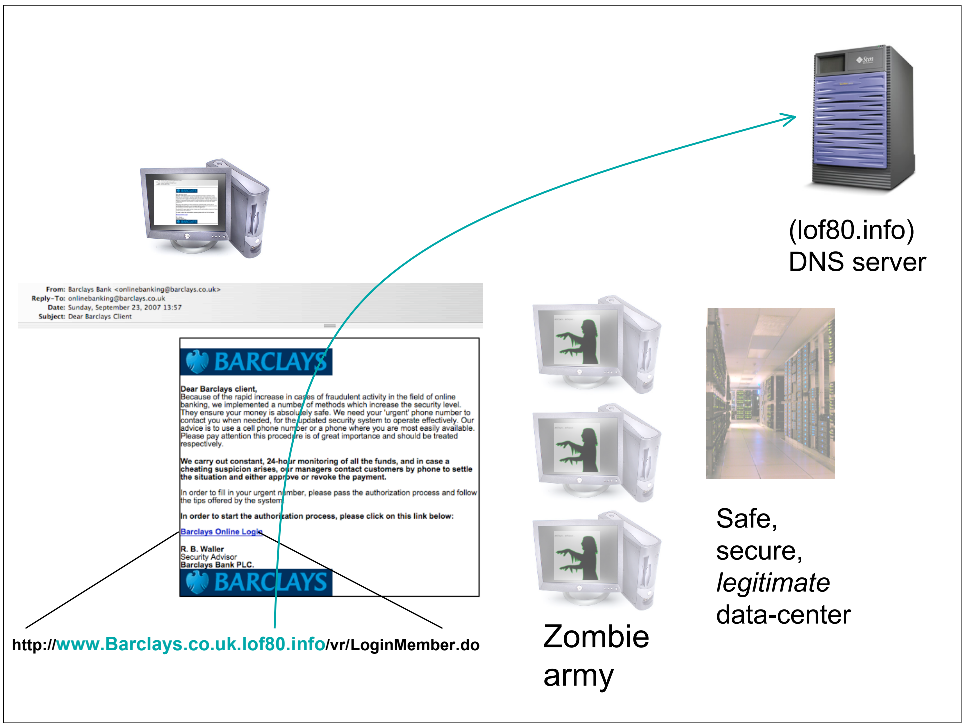
<http://www.Barclays.co.uk.lof80.info/vr/LoginMember.do>

(lof80.info)
DNS server

Zombie army



Safe, secure, *legitimate* data-center



From: Barclays Bank <onlinebanking@barclays.co.uk>
Reply-To: onlinebanking@barclays.co.uk
Date: Sunday, September 23, 2007 13:57
Subject: Dear Barclays Client

BARCLAYS

Dear Barclays client,
Because of the rapid increase in cases of fraudulent activity in the field of online banking, we implemented a number of methods which increase the security level. They ensure your money is absolutely safe. We need your 'urgent' phone number to contact you when needed, for the updated security system to operate effectively. Our advice is to use a cell phone number or a phone where you are most easily available. Please pay attention this procedure is of great importance and should be treated respectively.

We carry out constant, 24-hour monitoring of all the funds, and in case a cheating suspicion arises, our managers contact customers by phone to settle the situation and either approve or revoke the payment.

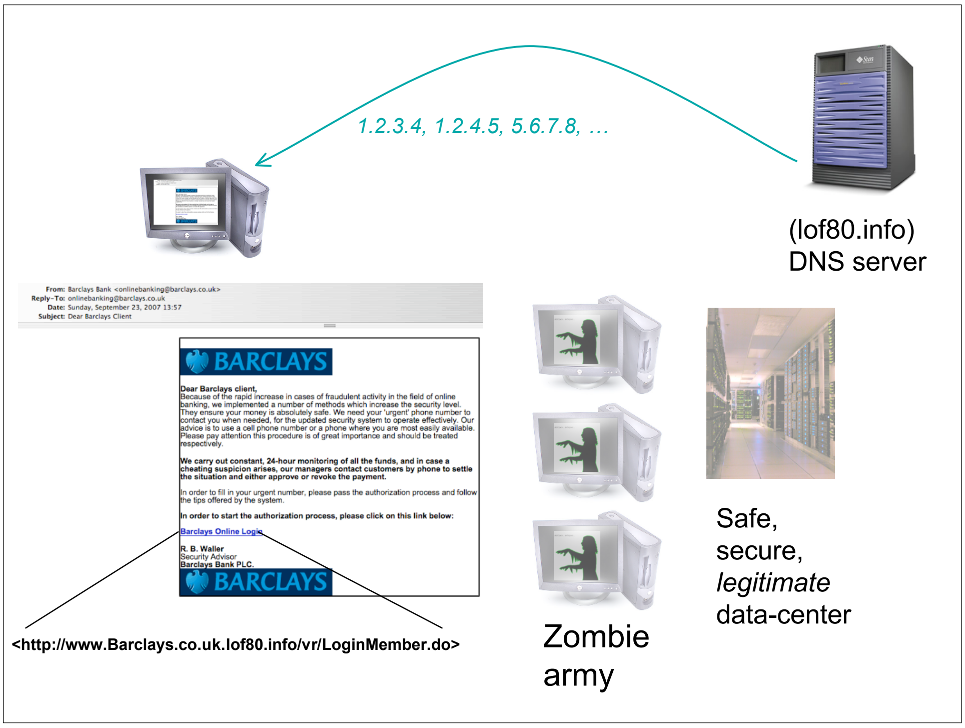
In order to fill in your urgent number, please pass the authorization process and follow the tips offered by the system.

In order to start the authorization process, please click on this link below:

Barclays Online Login

R. B. Waller
Security Advisor
Barclays Bank PLC.

BARCLAYS



From: Barclays Bank <onlinebanking@barclays.co.uk>
Reply-To: onlinebanking@barclays.co.uk
Date: Sunday, September 23, 2007 13:57
Subject: Dear Barclays Client

BARCLAYS

Dear Barclays client,
Because of the rapid increase in cases of fraudulent activity in the field of online banking, we implemented a number of methods which increase the security level. They ensure your money is absolutely safe. We need your 'urgent' phone number to contact you when needed, for the updated security system to operate effectively. Our advice is to use a cell phone number or a phone where you are most easily available. Please pay attention this procedure is of great importance and should be treated respectively.

We carry out constant, 24-hour monitoring of all the funds, and in case a cheating suspicion arises, our managers contact customers by phone to settle the situation and either approve or revoke the payment.

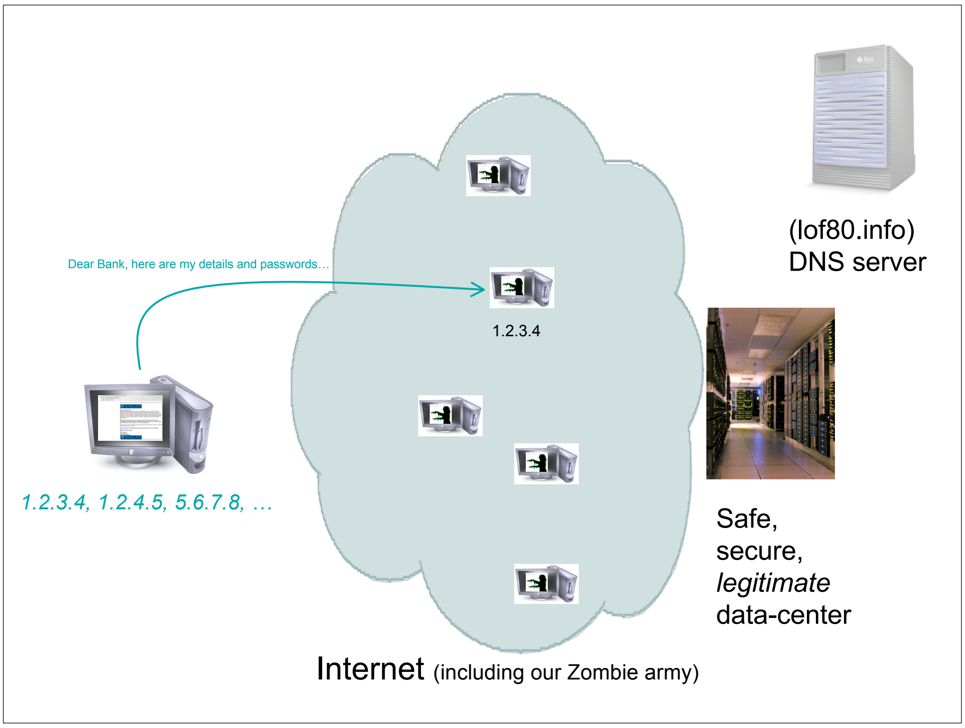
In order to fill in your urgent number, please pass the authorization process and follow the tips offered by the system.

In order to start the authorization process, please click on this link below:

Barclays Online Login

R. B. Waller
Security Advisor
Barclays Bank PLC.

BARCLAYS



(lof80.info)
DNS server

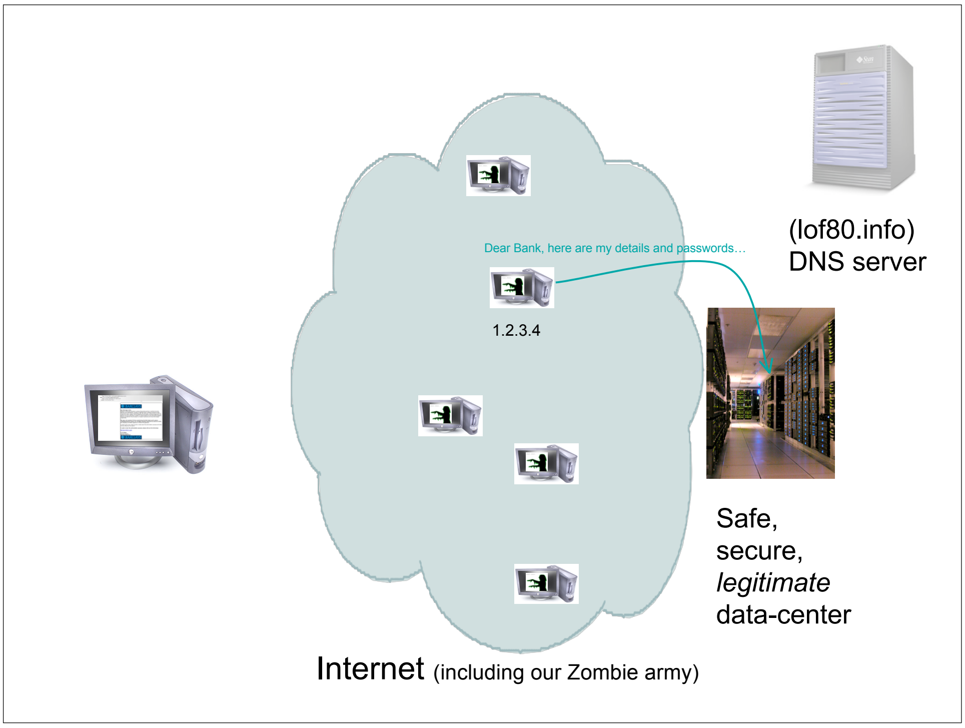
Dear Bank, here are my details and passwords…

1.2.3.4

1.2.3.4, 1.2.4.5, 5.6.7.8, …

Safe,
secure,
legitimate
data-center

Internet (including our Zombie army)



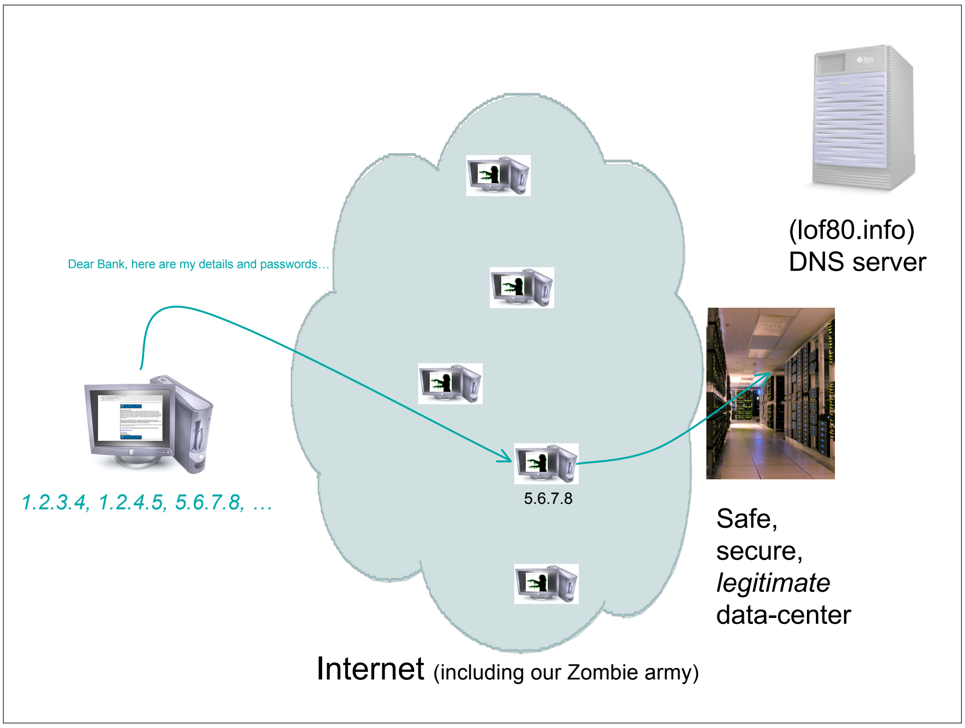
(lof80.info)
DNS server

Dear Bank, here are my details and passwords…

1.2.3.4

Safe,
secure,
legitimate
data-center

Internet (including our Zombie army)



(lof80.info)
DNS server

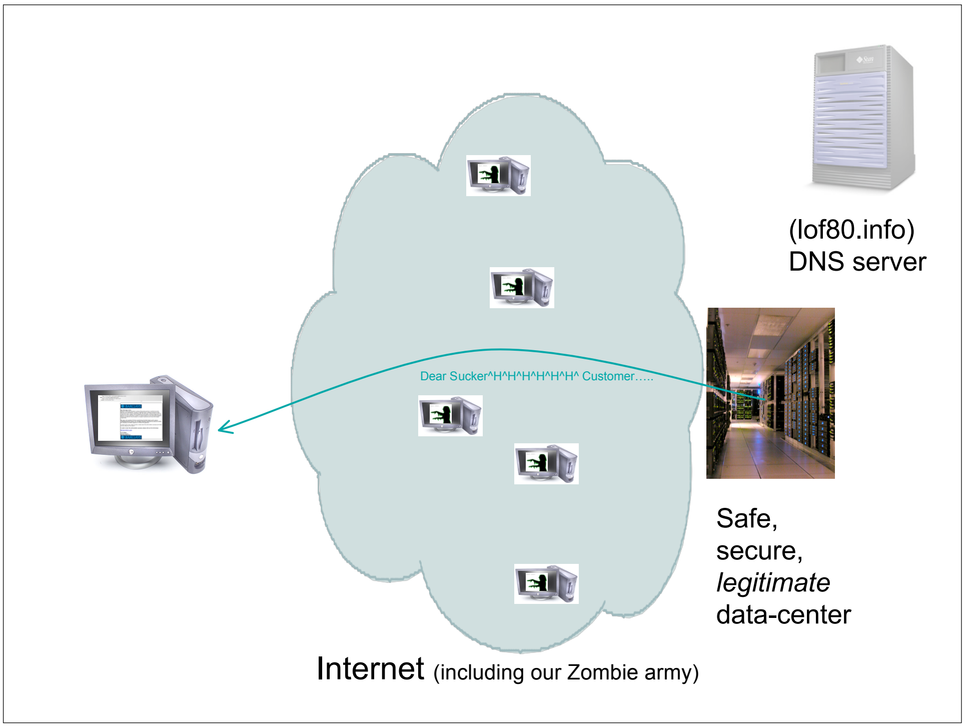
Dear Bank, here are my details and passwords…

1.2.3.4, 1.2.4.5, 5.6.7.8, …

5.6.7.8

Safe,
secure,
legitimate
data-center

Internet (including our Zombie army)

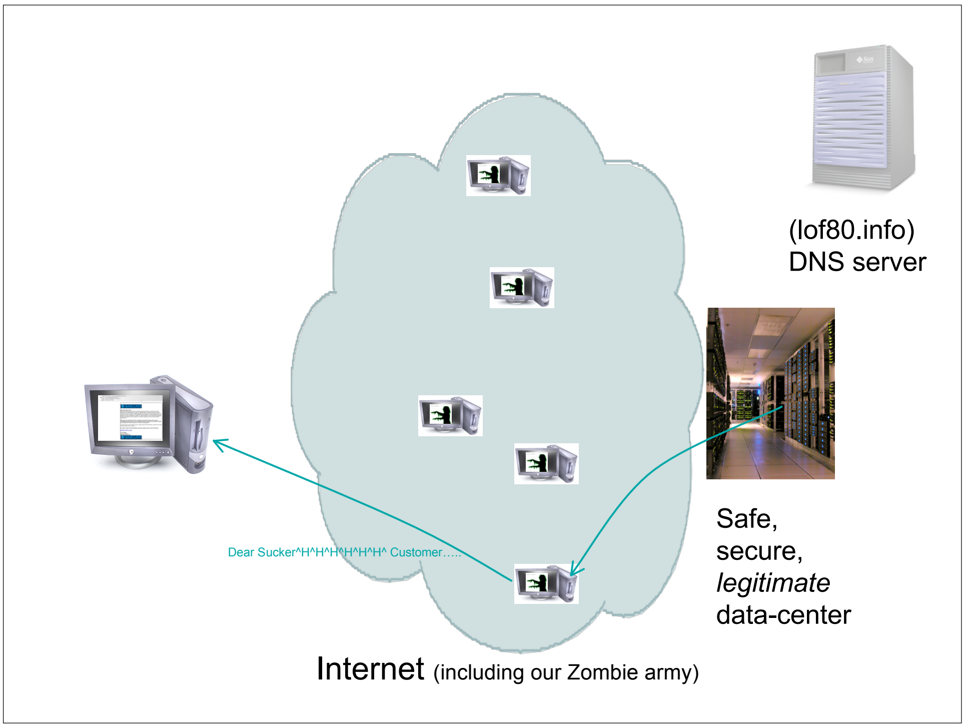


(lof80.info)
DNS server

Dear Sucker^H^H^H^H^H^H^ Customer…..

Safe,
secure,
legitimate
data-center

Internet (including our Zombie army)

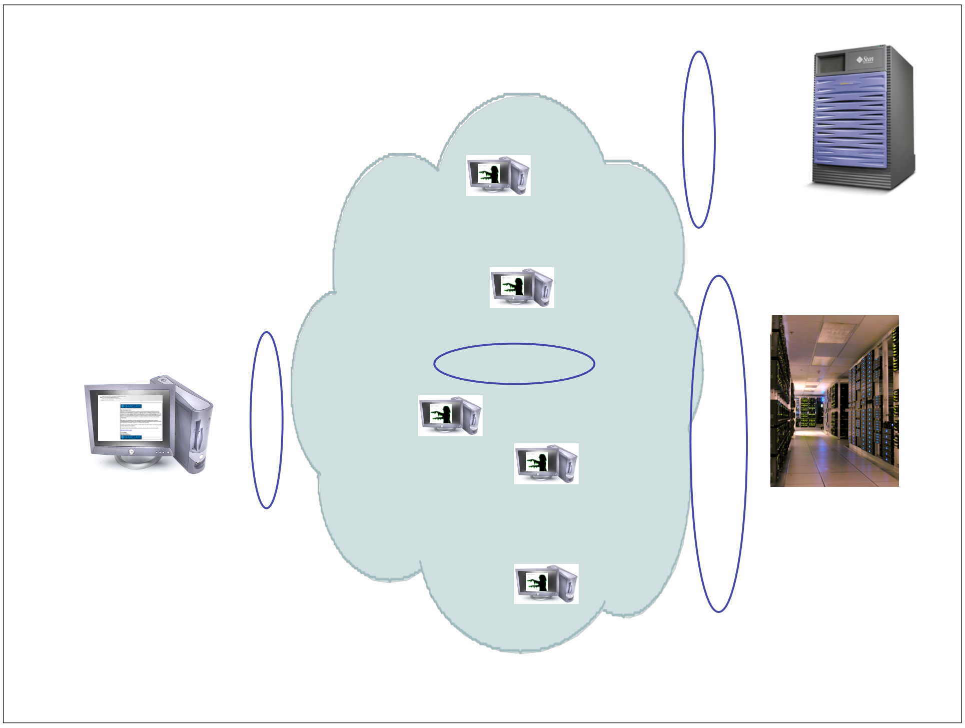


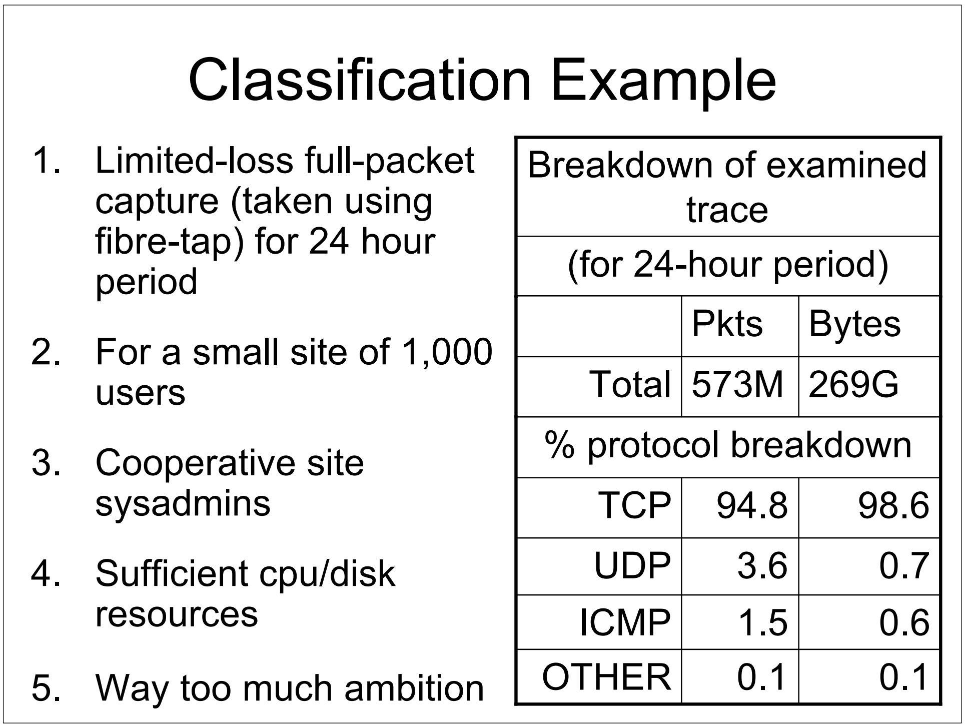
(lof80.info)
DNS server

Safe,
secure,
legitimate
data-center

Dear Sucker^H^H^H^H^H^H^ Customer…..

Internet (including our Zombie army)

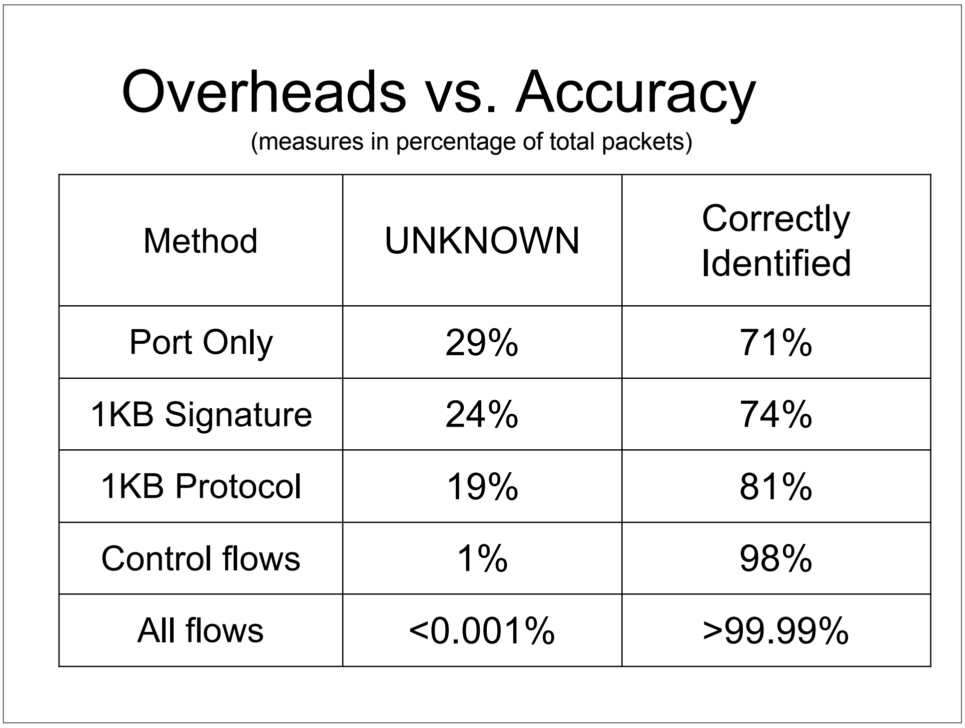




Classification Example

1. Limited-loss full-packet capture (taken using fibre-tap) for 24 hour period
2. For a small site of 1,000 users
3. Cooperative site sysadmins
4. Sufficient cpu/disk resources
5. Way too much ambition

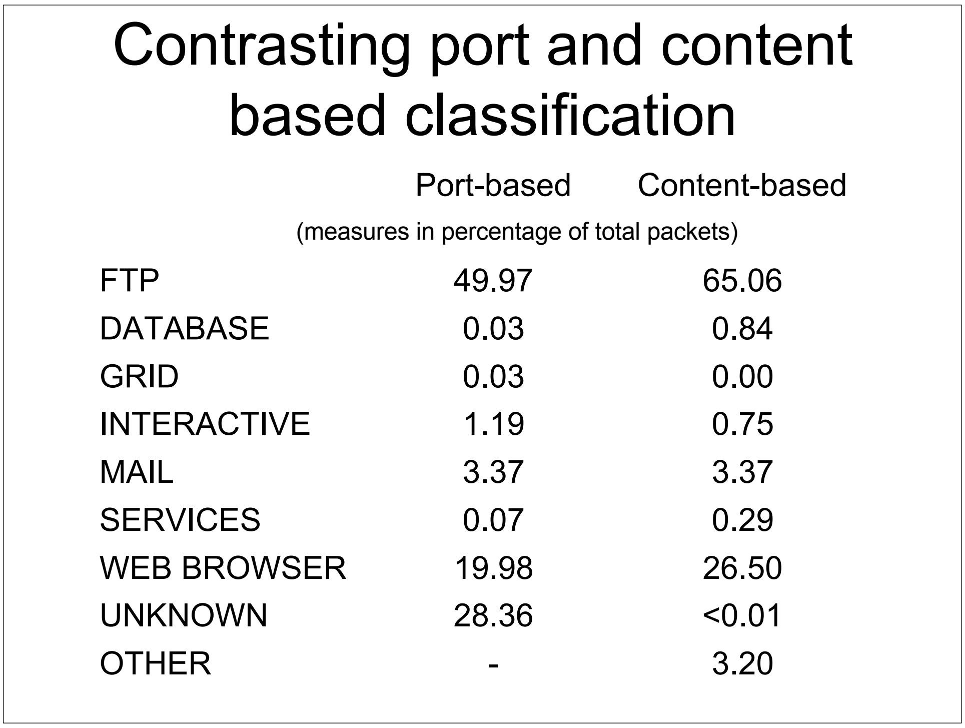
Breakdown of examined trace		
(for 24-hour period)		
	Pkts	Bytes
Total	573M	269G
% protocol breakdown		
TCP	94.8	98.6
UDP	3.6	0.7
ICMP	1.5	0.6
OTHER	0.1	0.1



Overheads vs. Accuracy

(measures in percentage of total packets)

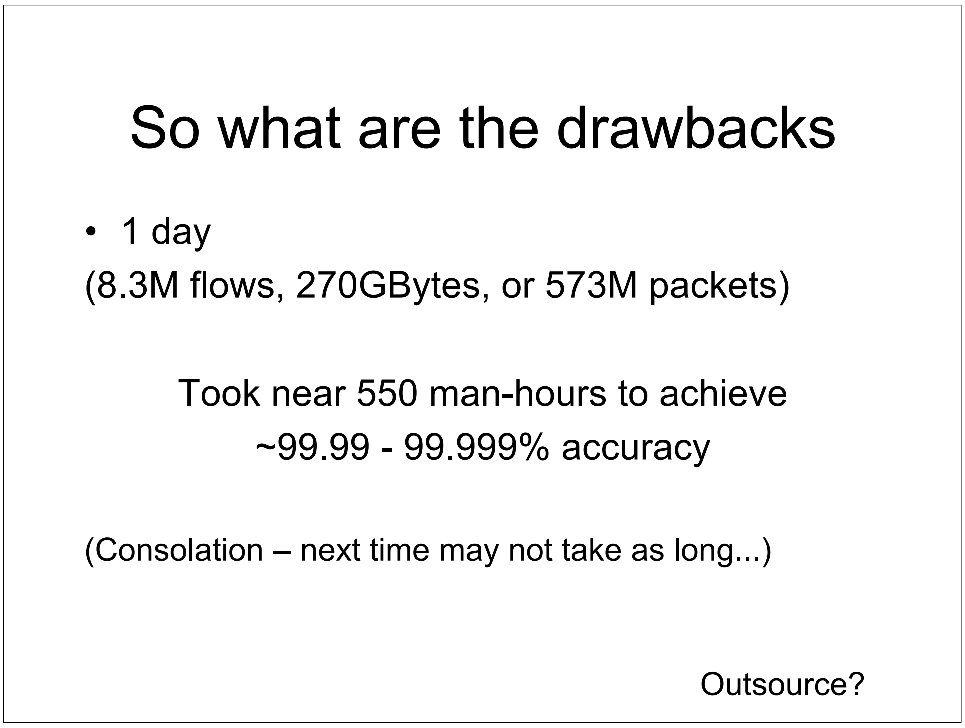
Method	UNKNOWN	Correctly Identified
Port Only	29%	71%
1KB Signature	24%	74%
1KB Protocol	19%	81%
Control flows	1%	98%
All flows	<0.001%	>99.99%



Contrasting port and content based classification

(measures in percentage of total packets)

	Port-based	Content-based
FTP	49.97	65.06
DATABASE	0.03	0.84
GRID	0.03	0.00
INTERACTIVE	1.19	0.75
MAIL	3.37	3.37
SERVICES	0.07	0.29
WEB BROWSER	19.98	26.50
UNKNOWN	28.36	<0.01
OTHER	-	3.20



So what are the drawbacks

- 1 day

(8.3M flows, 270GBytes, or 573M packets)

Took near 550 man-hours to achieve
~99.99 - 99.999% accuracy

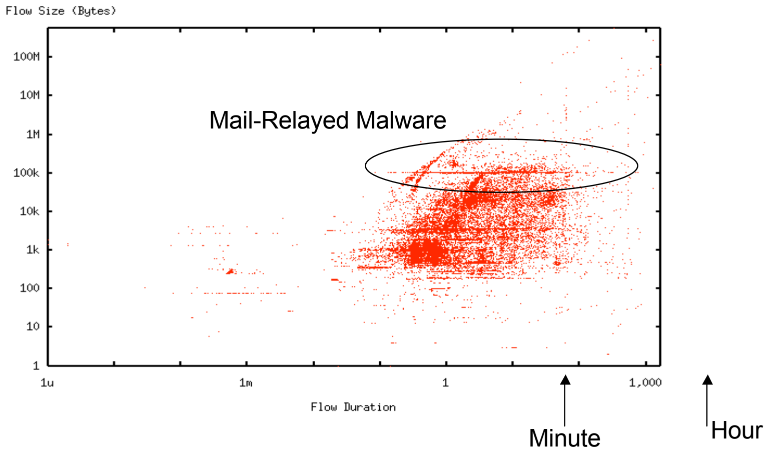
(Consolation – next time may not take as long...)

Outsource?

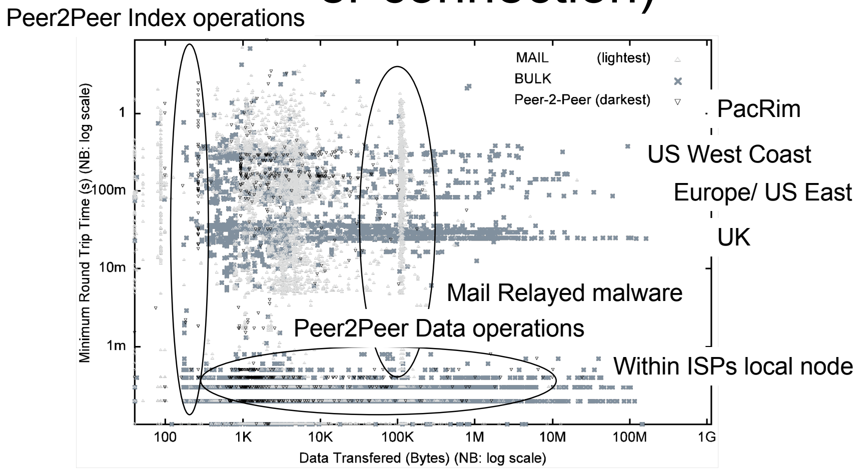
Errors?

- Encrypted Protocols
  - ssh: 831MBytes, (0.3 %)
    - Interactive sessions (Talk to the users)
- Covert channels
  - legitimate protocols carrying undesired traffic
- Unrecognized samples
  - too-small a sample to decode: e.g., one packet for a unique host for the 24 hour trace
    - Commonly from off-site
    - Residual background radiation (Pang *et al.* IMC04)

Flow size (Bytes) vs duration (s) (point per connection)



RTT vs. data transferred (point per connection)



Peer2Peer Index operations

PacRim

US West Coast

Europe/ US East

UK

Mail Relayed malware

Peer2Peer Data operations

Within ISPs local node

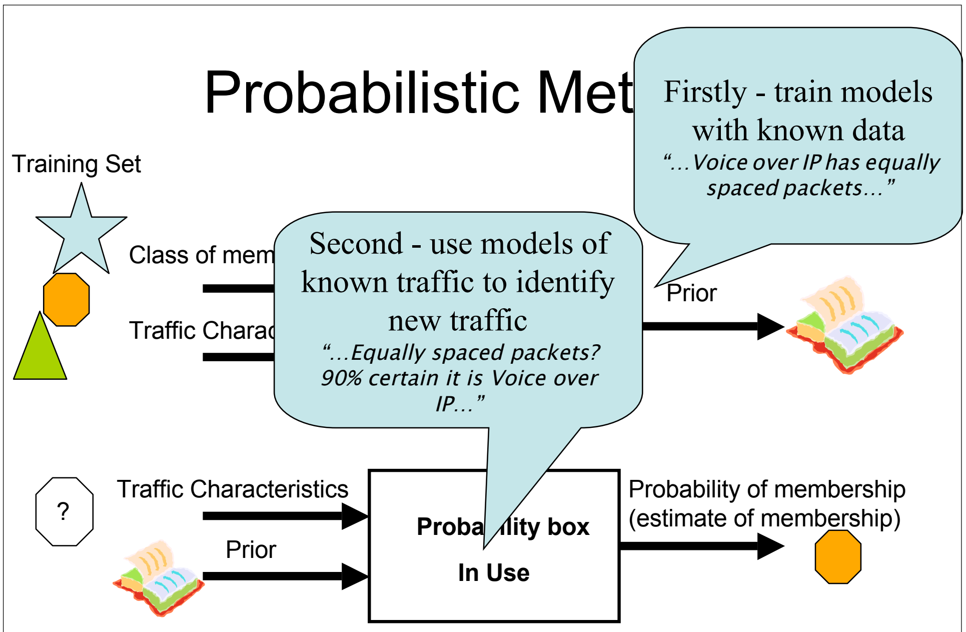
A further alternative?

- We could encode in software the manual process

  work in progress - but maybe not robust

- Could we use a probabilistic method – a Bayes method?

Probabilistic Met



Firstly - train models with known data

"…Voice over IP has equally spaced packets…"

Second - use models of known traffic to identify new traffic

"…Equally spaced packets? 90% certain it is Voice over IP…"

What is Bayes theory anyway?

100 years of theory in 100 seconds

- P(H|D) = P(H)P(D|H) / P(D)
- H the Hypothesis
- P(H) – the "Prior" probability
- Observe data D

Hypothesis "Bayes is dead"

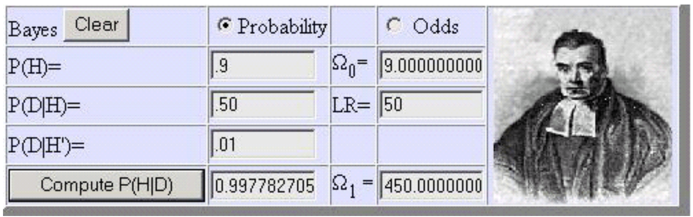
- P(H) .9 (given that outfit)

thanks to Derek McAuley for the pictures

Bayes II – make an observation

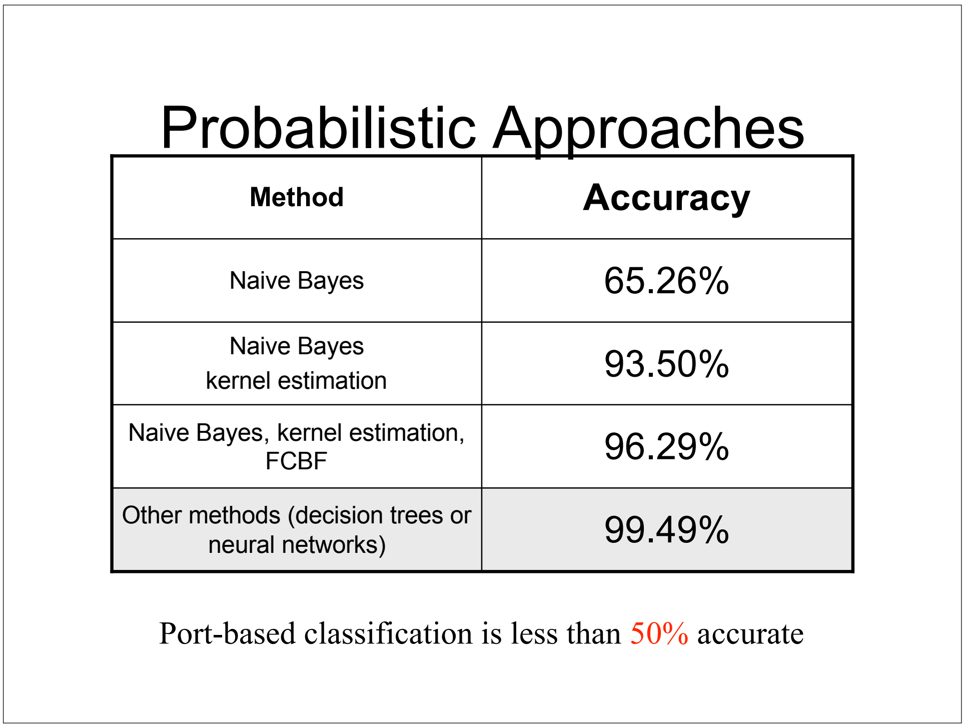
Bayes III – reach a conclusion

- P(H), say .9 Hypothesis “Bayes is Dead”
- P(D|H), say .5 Pr(dead given a grave)
- P(D|H’), say .01 Pr(not dead given a grave)
- P(D) hence .451



Bayes Clear	● Probability		○ Odds
P(H)=	.9	Ω_0=	9.000000000
P(D|H)=	.50	LR=	50
P(D|H')=	.01		
Compute P(H|D)	0.997782705	Ω_1 =	450.0000000

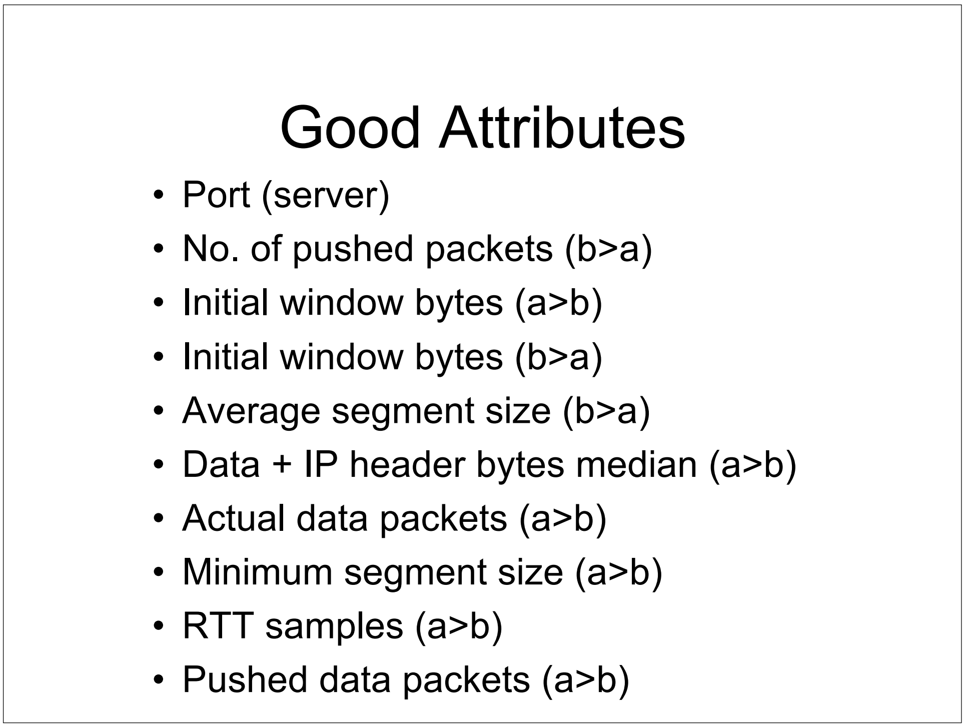
- Posterior P(H|D) is .99778..
  - Okay, so he is dead (probably)



Probabilistic Approaches

Method	Accuracy
Naive Bayes	65.26%
Naive Bayes kernel estimation	93.50%
Naive Bayes, kernel estimation, FCBF	96.29%
Other methods (decision trees or neural networks)	99.49%

Port-based classification is less than 50% accurate



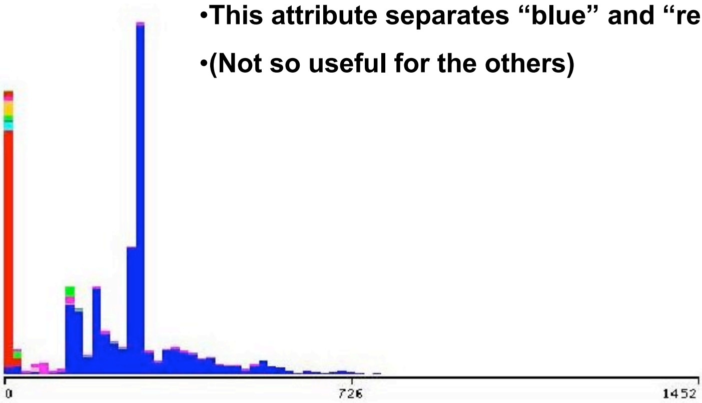
Good Attributes

- Port (server)
- No. of pushed packets (b>a)
- Initial window bytes (a>b)
- Initial window bytes (b>a)
- Average segment size (b>a)
- Data + IP header bytes median (a>b)
- Actual data packets (a>b)
- Minimum segment size (a>b)
- RTT samples (a>b)
- Pushed data packets (a>b)

Example attribute

Colours represent classes

- **This attribute separates “blue” and “red” well**
- **(Not so useful for the others)**

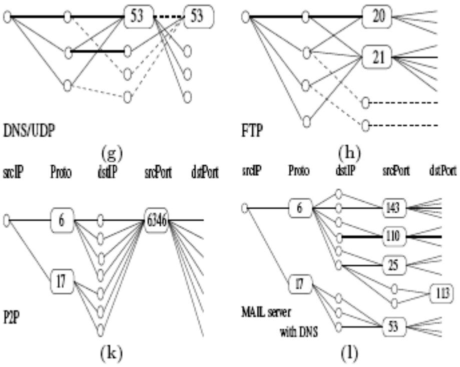


Other features

A simple number is not the only feature

- A graph shape (e.g., histogram) is a feature
- A set of activities over time and space is a feature

For example:



(g)

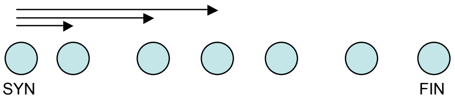
(h)

(k)

(l)

Netflow curiousness

- Netflow data is common & often held for long-term archive
- Sampled Netflow may reveal some flow structure - unintentional but useful…



Pick flows containing 2 packets and SYN flag

end time (last observation) - start time (first observation) = IAT

total bytes in flow = SYN packet + <other>

Result: some insight into the packet-by-packet size and timings

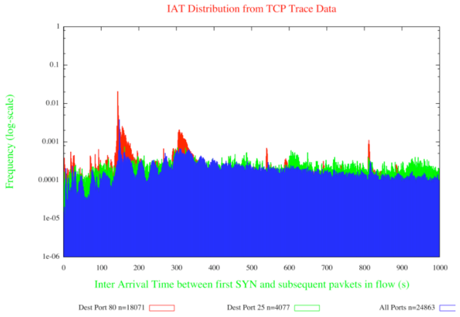
Notional Packet spacing (IAT)

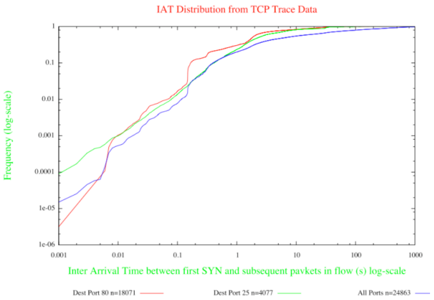
Downsides

- Need a lot of data
- Suffers all the disadvantages of sampling
- Encodes a lot of site/host/link information

Upsides

- May be a sufficiently useful change-detector
- Plentiful data-source
- Others have shown that packet-train sizes are a useful fingerprint





Why Characterize?

- **Identify**: “Hmmm, So *this is what an attack looks like*”
- **Understanding**: “*So what is my network doing anyway?”*
- **Accountability**: “*What has caused this enormous bill?”*
- **Application Enabler***:* Dynamic (application-specific) handling (e.g. routing) by end systems
- **Performance Tracking**: *“What is causing my application to go so very slow?”*
- **Application identification**: *“…telling helpdesk what the users won’t or can’t find out”*
- **Better Models**: Leading to better/more-realistic test traffic

How?

- Content classification - **Hard**.
  - But we are told us we don’t need flow reassembly for identification…. actually all he said was we could limit the traffic that required flow-reassembly
- Behavior classification
  - **requires some *ground truth***

    (which relied on content classification to begin with)

Where next?

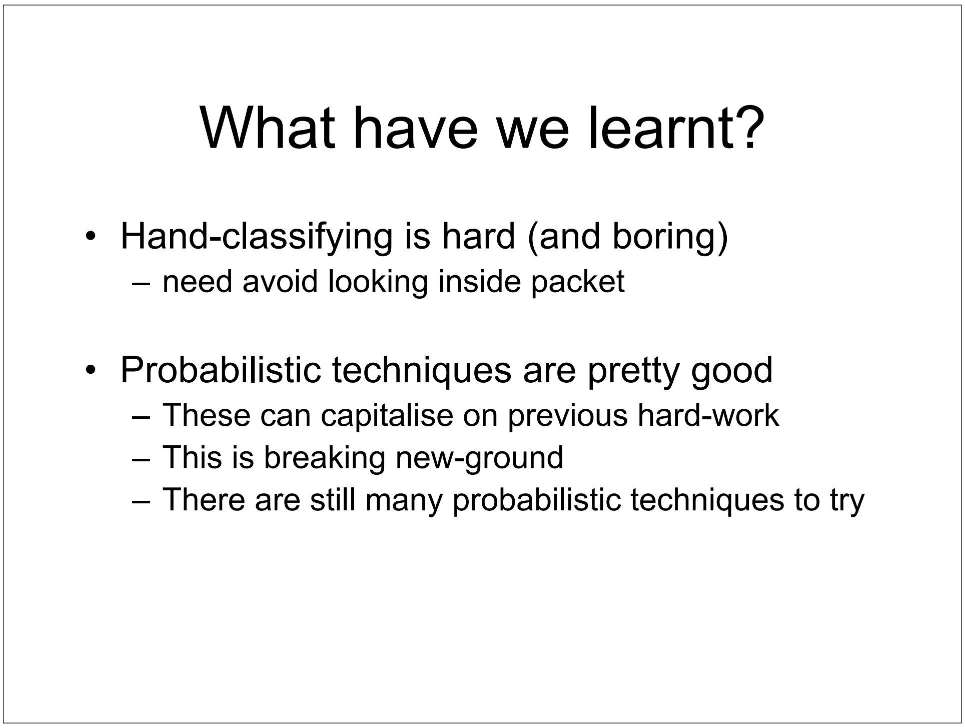
- Same Methods on New Data Sets
  - Same site on other days:
    - Assess Stationarity and Classification *Half-life*
  - Different sites on the same and more recent days:
    - Assess Classification Independence
- Other Methods

  e.g., Ones that do not assume flow independence
- Develop Better Attributes
- But most of all apply-better methods (or talk to others than can)

Domain Knowledge

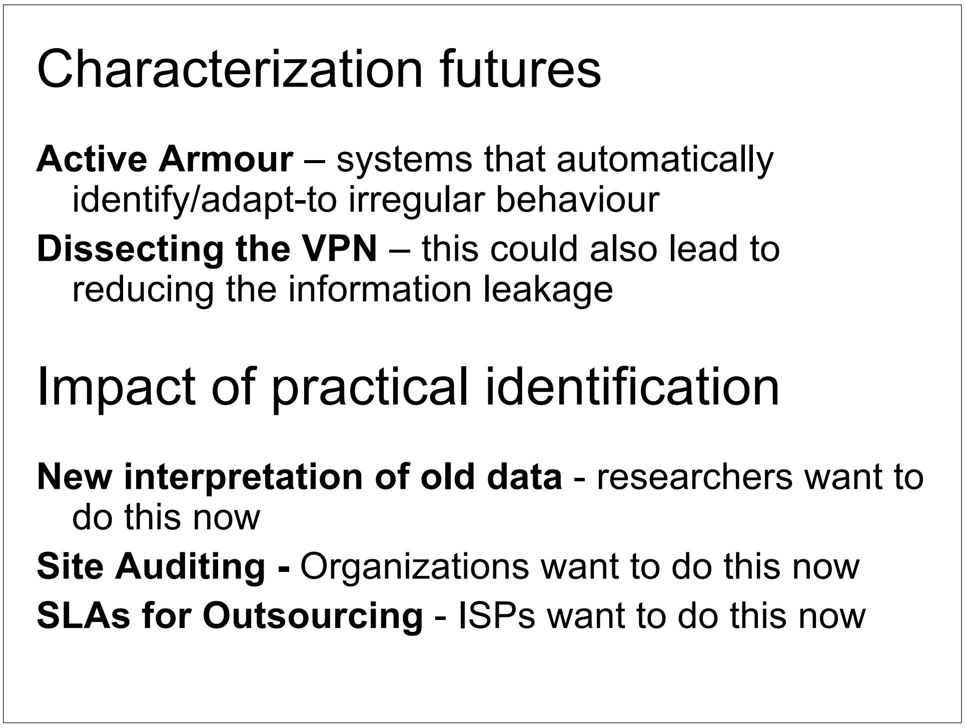
- Each of the motivations for “Why?” is a different domain of knowledge:
  - Hard to compare methods applied to different domains

  (Helping helpdesk may require significant site knowledge & historical knowledge)
  - Hard to compare data used in/by/for different methods (BLINC uses *flow-community* actions, mine is flow i.i.d)
- ML “headline”: These approaches encode domain knowledge



What have we learnt?

- Hand-classifying is hard (and boring)
  - need avoid looking inside packet
- Probabilistic techniques are pretty good
  - These can capitalise on previous hard-work
  - This is breaking new-ground
  - There are still many probabilistic techniques to try



Characterization futures

Active Armour – systems that automatically identify/adapt-to irregular behaviour

Dissecting the VPN – this could also lead to reducing the information leakage

Impact of practical identification

New interpretation of old data - researchers want to do this now

Site Auditing - Organizations want to do this now

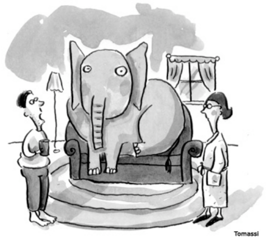
SLAs for Outsourcing - ISPs want to do this now

Elephants in the

Hallway/Driveway/Kitchen/Lounge(room)/Bathroom/Bedroom

- Limited engagement of/with the M-L community
  - Mea Cupla - I don’t read KDD output either
- Difficult-to-compare methodologies
- Difficult-to-compare datasets
- Lack of (annotated) Data
  - We don’t/**can’t** play nicely together
  - Privacy/Law

(Oops, I’m channeling kc claffy)



Classes as confusion

Network traffic Paper 1	Network traffic Paper 2	Network traffic Paper 3	Typical IDS paper
7 meta-classes (? classes)	11 meta-classes (40-50 classes)	11 meta-classes (40-50 classes)	2/3 meta-classes
domain, ftp-data, https, kazaa, realmedia, telnet, www	bulk(ftp), database, interactive, mail, services, www, p2p, attack, games, multimedia, unknown	web, p2p, data(ftp), network management, mail, news, chat/irc, streaming, gaming, nonpayload, unknown	Good, Bad, Ugly

How can I compare these methods?
I certainly can’t compare the output

Upshot - one persons great performance
is another persons rubbish performance

One day...

- Informed planning using **actual** application usage

- Self-defending household firewall, interface-card, and access-point

- Intelligent multiple-radio wireless usage

My thanks…

No (networking) researcher is an island

- **Dina Papagiannaki, Ian Pratt, Denis Zuev, and Richard Clayton among many others, along with a cast of thousands (of users)**

- **University of Cambridge and Intel**

WACI thanks:

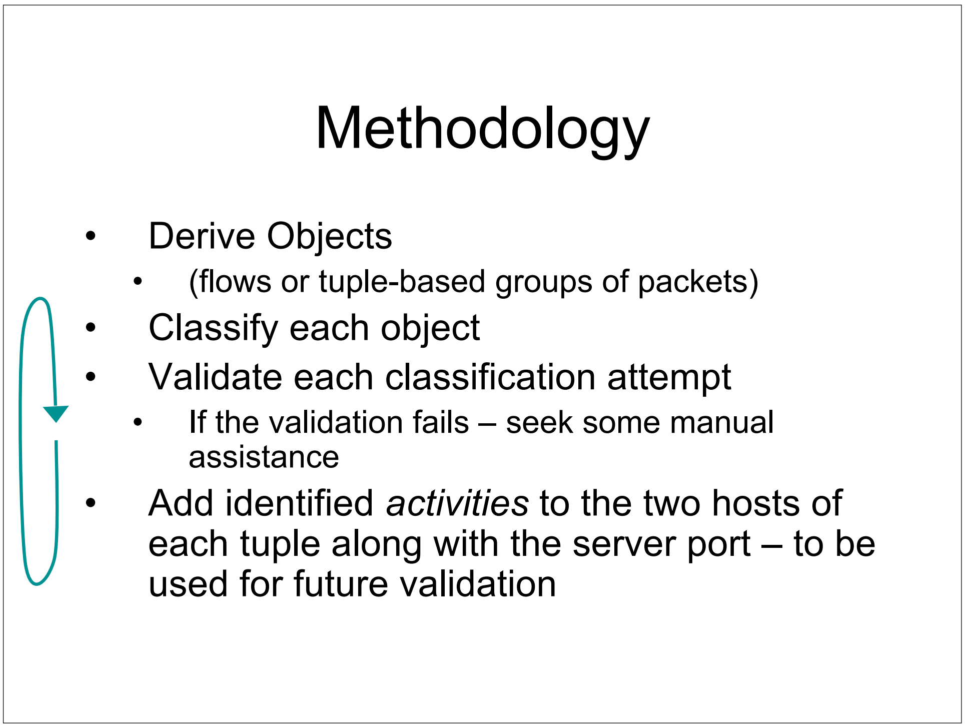
- **IRTF's Internet Measurement Research Group (Tim and Mark)**

- **BBN Technologies**

Question ?

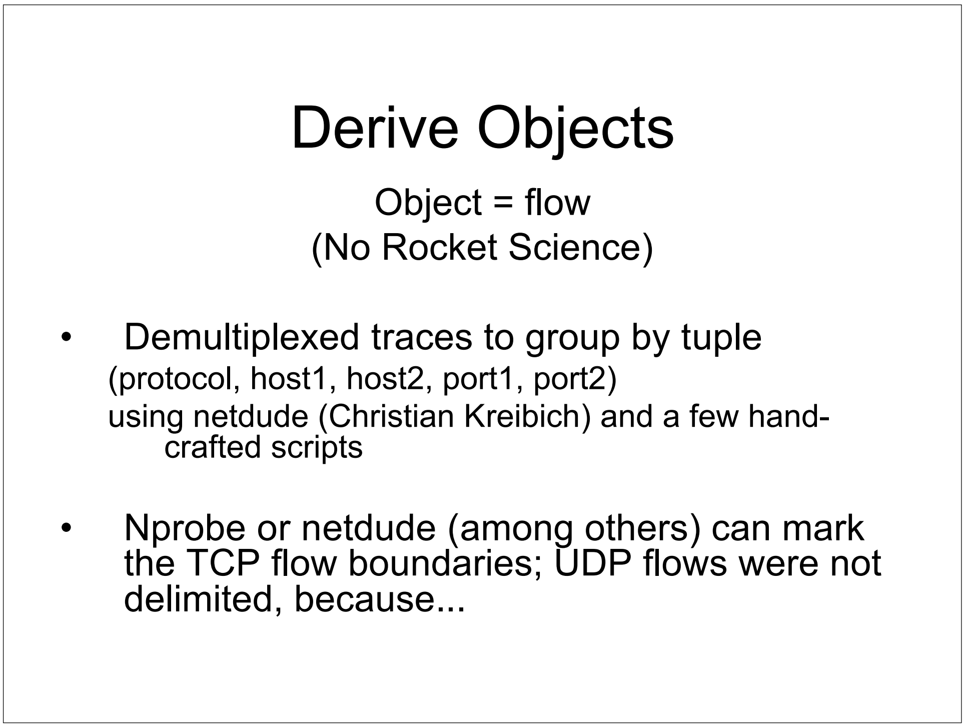
Our Approach

- Content-based classification
  - based upon full packet-capture
- Putting to one-side two issues:
  - privacy and practicality
- Need an identification of each application



Methodology

- Derive Objects
  - (flows or tuple-based groups of packets)
- Classify each object
- Validate each classification attempt
  - If the validation fails – seek some manual assistance
- Add identified *activities* to the two hosts of each tuple along with the server port – to be used for future validation



Derive Objects

Object = flow
(No Rocket Science)

- Demultiplexed traces to group by tuple

  (protocol, host1, host2, port1, port2)

  using netdude (Christian Kreibich) and a few hand-crafted scripts

- Nprobe or netdude (among others) can mark the TCP flow boundaries; UDP flows were not delimited, because...

Derive Objects - 2

- It quickly became clear that classifications for TCP flows and groups of UDP packets were (*surprisingly?*) stabile.
- Exceptions were not surprising:
  - P2P mixed in with HTTP
- Quantity was still pretty small
- UDP showed no such exception across any tuple

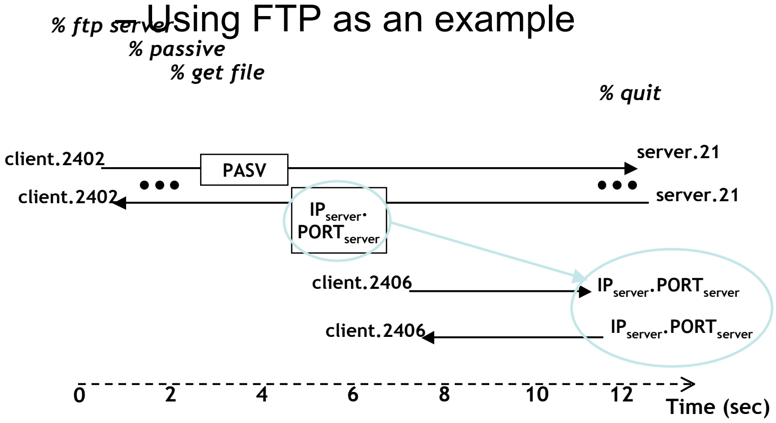
  (despite a laborious examination)

Traffic Identification Methods

- Flow-Behaviour
  - e.g., packets only travelling in one direction
- Recognisable contents strings
  - e.g., "*GET /.hash*" a P2P signature
- Protocol behaviour
  - e.g., "***MAIL…FROM…RCPT…DATA..***" a valid SMTP (mail) transfer

Traffic Identification Methods - II

- Control flow
  - Using FTP as an example



Traffic Identification Methods - III

- Format signatures:

  "Integer < 5, followed by string"

- Host behaviour

  Hosts have signatures too
  - DNS (names reveal purpose)
  - Routers transfer routing protocols, windows boxes (usually) do not

- Port (particularly server port)
  - the server port (identified as part of each object) formed the initial *seed* for classification – if the classification is known

Example – I

- H1,H2,P1,P2,TCP
- H2,P2 is a non-standard http server/port
  (identified previously)

  web client and web server (on non-standard port)?
- H1,P1 has not previously been active

  web client and web server (on non-standard port)?
- Parse TCP flow reveals a valid HTTP transaction

  web client / server verified

  H1 identified as HTTP client

Example – II

- H1,H2,P1,P2,TCP
- H2,P2 is a non-standard http server/port
  (identified previously)

  web client and web server (on non-standard port)?
- H1 previously identified as a windows box

  web client and web server (on non-standard port)?
- Parse TCP flow reveals an P2P signature

  web client / server rejected

  H2 identified as P2P server – revisit/revise H2 flows as required

Implementation

- A database containing an entry per-flow
  - known ports
  - signatures, etc.

    each added for a subsequent classification
- A database containing an entry per-host
  - based upon previously identified host traffic
  - clues from DNS (e.g. NAT boxes)

Processing Techniques

Increasing Complexity/Overheads

HP	**Header-Port-Based**	**25 = SMTP (mail) 80 = http (web)**
HF	**Packet-Header (Full)**	**Simplex flows** **Requests (but no acknowledgements)**
PS	**Packet Signature**	**Many malware signatures** **Offset(5) = 0xdeadbeef**
PP	**Packet Protocol**	**IDENT** **Integer < 5, followed by string**
1S	**Signature on 1st KByte**	**P2P** **GET = http://hash2546**
1P	**1st KByte Protocol**	**SMTP** **MAIL…FROM…RCPT…DATA..**
SP	**(Selected) Flow Protocol**	**FTP** **PASV <host>,<port>**
FP	**(Total) Flow Protocol**	**VNC** **Integer < 5, followed by string**
HH	**Host History**	**Port-Scanning**