Identifying Network Traffic Activity Via Flow Sizes

REDJACK

REDJACK

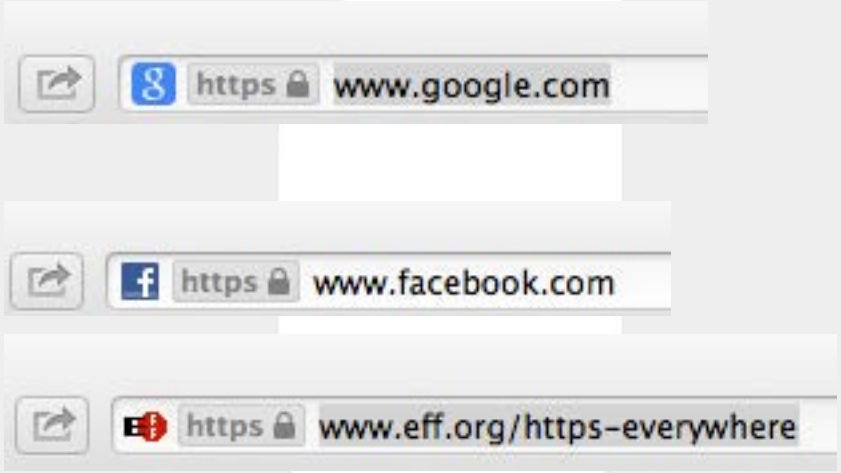
Overview

- Motivation – identifying activity via payload
- Theory behind the idea
- Measuring NetFlow
- Measuring DNS traffic captures
- Implications and future work

REDJACK

Motivation

- Users don't have the common decency to send plaintext all over the place anymore
- HTTPS prevalence
- OTR encryption for IM
- SSL for email



REDJACK

This Expands on Previous Work

- 2007 Paper on BitTorrent detection that focused on multiple behaviors – fumbling, file transfers, &c
- Now doing in depth study of control messages to see what we can find
  - Advantage – this time, have payload
- Questions:
  - Size of control messages
  - Distribution of control messages
  - Frequency of combinations?

REDJACK

Identifying Protocols Via Flow Sizes

- Hypothesis: traffic consists of three families of data
  - "Chatter"
    - Short (< MTU) , roughly symmetric packets of variable size
    - SSH, Telnet, IRC, ICQ, AIM
  - Transfer
    - MTU packets, met by payload-zero packets
    - FTP, Mail, HTTP
  - Control
    - < MTU packets, fixed sizes "fill in the blank" templates
    - All protocols

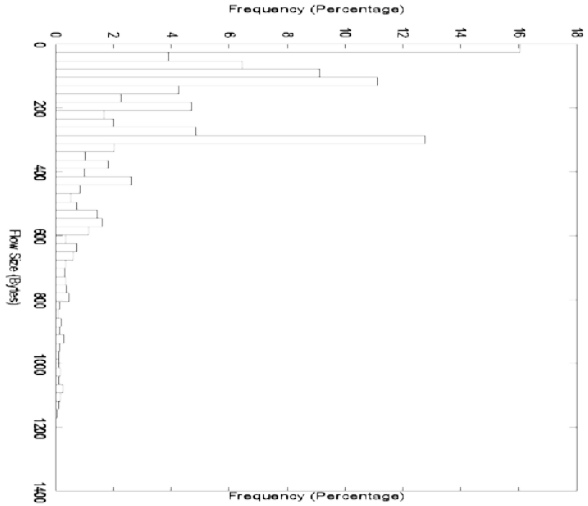
REDJACK

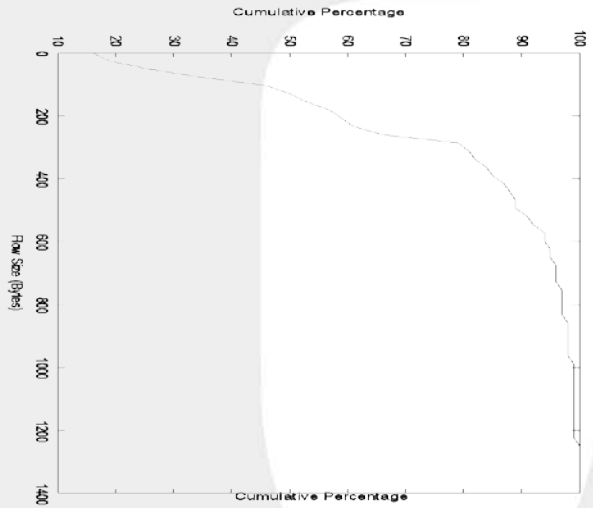
Differentiate Via Control Message Sizes

Histogram

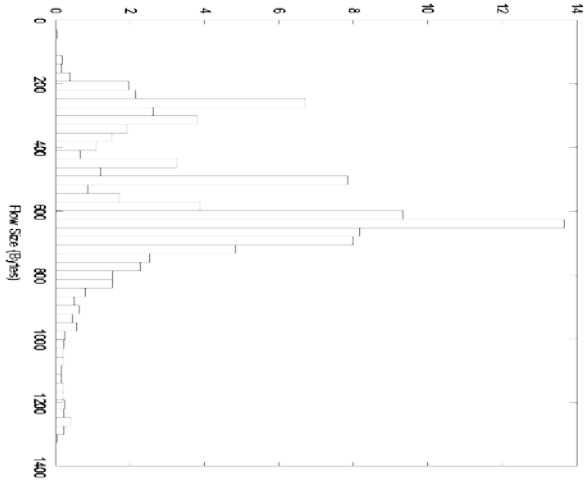
CDF

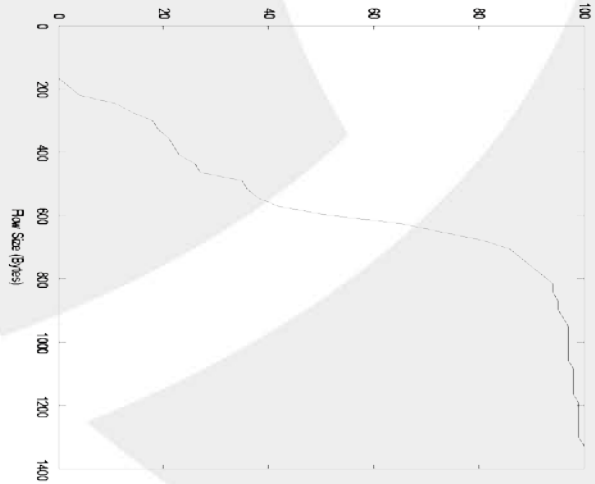
SMTP





HTTP





REDJACK

Done Some of This Already

- 2007 paper on p2p identification showed that you could find BitTorrent by looking for specific behaviors
  - Control packet sizes were one particular behavior
- However…
  - What are the actual packets?
  - What are the sizes
- Didn't have ground truth in previous work
  - Now have access to it via DNS records

REDJACK

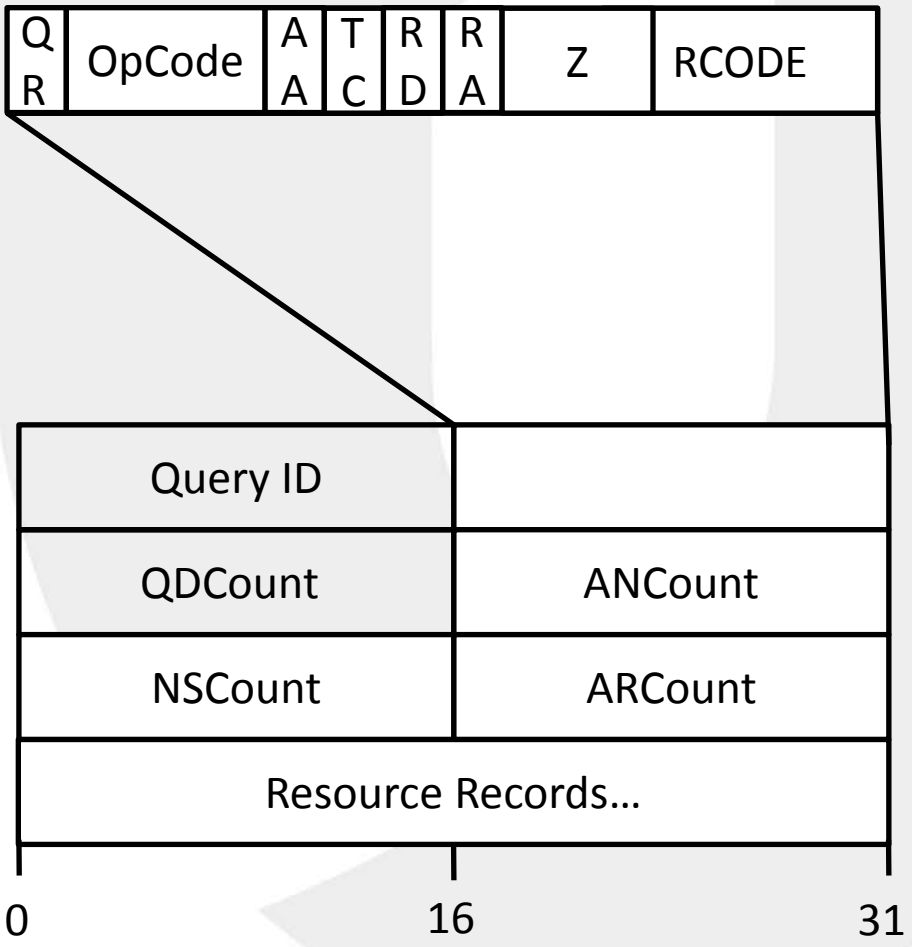
DNS Analysis

- Using DNS data, we can compare the exact messages sent against packet sizes
- See what messages produce what packet sizes
- Determine if we can predict messages via sizes
- Can't predict *content*, but we can guess what the user was looking for

REDJACK

The DNS Datagram

- State is maintained by Query ID
- Other flags set various info – authoritative, recursive, &c
- Response is sent in one or more RR's (resource records)



REDJACK

Resource Records and DNS Information

- DNS handles a *lot* of information
  - Name lookup
  - Name *ownership*
  - Authentication
  - Redirection
  - Email

REDJACK

Ripping Apart DNS Message Contents

- A DNS message contains 1 or more RR's (resource records)
  - Different RR's serve different purposes
  - Each RR has a different format, although most contain at least one variable length domain name
- Multiple different RR's may be sent to comprise a single message
- There's no requirement that the RR's actually be related to the original query, they may be annotative information
- ~40 RR's currently defined, including a couple of optional ones
- Responses are rarely just one message

Multiple Records Will Appear Simultaneously

	A	AAAA	CNAME	MX	NS	OPT	SOA	TXT
A	**99.33**	100.00	52.56	98.15	99.33	99.30	99.59	50.00
AAAA	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00
CNAME	0.69	0.00	**1.30**	0.00	1.30	1.36	0.00	0.00
MX	1.88	0.00	0.00	**1.90**	1.90	0.11	69.18	50.00
NS	100.00	100.00	100.00	100.00	**100.00**	100.00	100.00	100.00
OPT	49.56	0.00	52.03	2.94	49.57	**49.57**	0.41	50.00
SOA	2.65	0.00	0.00	96.29	2.65	0.02	**2.65**	50.00
TXT	0.00	0.00	0.00	0.05	0.00	0.00	0.04	**0.00**

- Table provides P(record of row type|record of column type); blue columns are P(record of row type)
- Some records (NS,A) are common
- Some (SOA) have a strong dependency P(SOA|MX)=96%
- Records will show up in group (5,10 NS records common)

REDJACK

What Are These Messages?

- A – IPv4 Address, 32 bit integer
- AAAA – IPv6 Address, 128 bit integer
- CNAME – Canonical Name, domain name string
- MX – Mail record, 16 bit preference value + domain name string
- NS – Nameserver name, domain name string
- OPT – Option record, variable option length
- SOA – 2 domain names and 128 bits of integers
- TXT – Variable length text

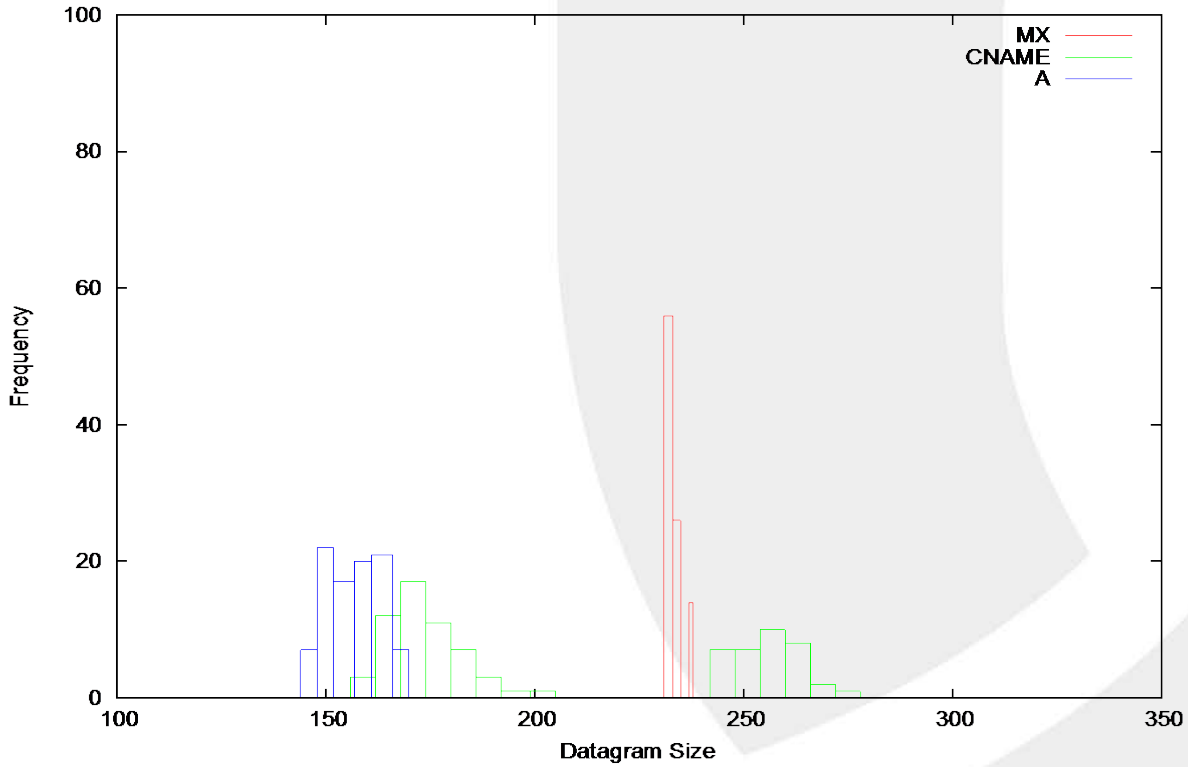
REDJACK

What Do We Do With DNS?

- Really, three major queries
  - Queries returning MX – Mail lookups
  - Queries returning CNAME – looking up aliases (CDN's love this)
  - Queries returning A on its own – simple lookups
- We can split out these queries and calculate frequencies for each one

REDJACK

Resulting In This



REDJACK

Observations

- Simple A records (least baggage) are smallest
- CNAME records broken into two groups
  - Differentiation is by NS records
  - 5 NS responses – smaller group
  - 10 NS responses – larger group
- MX is a very narrow spike (231-238 bytes)
  - Actual MX record is just a domain name, the rest of the offset is due to the SOA record

REDJACK

Conclusions

- Control messages in protocols can be used to differentiate the types of messages sent
  - We can use this information to differentiate protocols
  - Can use it to identify specific behaviors within protocols
- Variance in domain names is not significant enough to cause ‘overlap’ in messages
- Where can we go with this?
  - Facebook? Graph API? REST interfaces?
  - Markov Models?