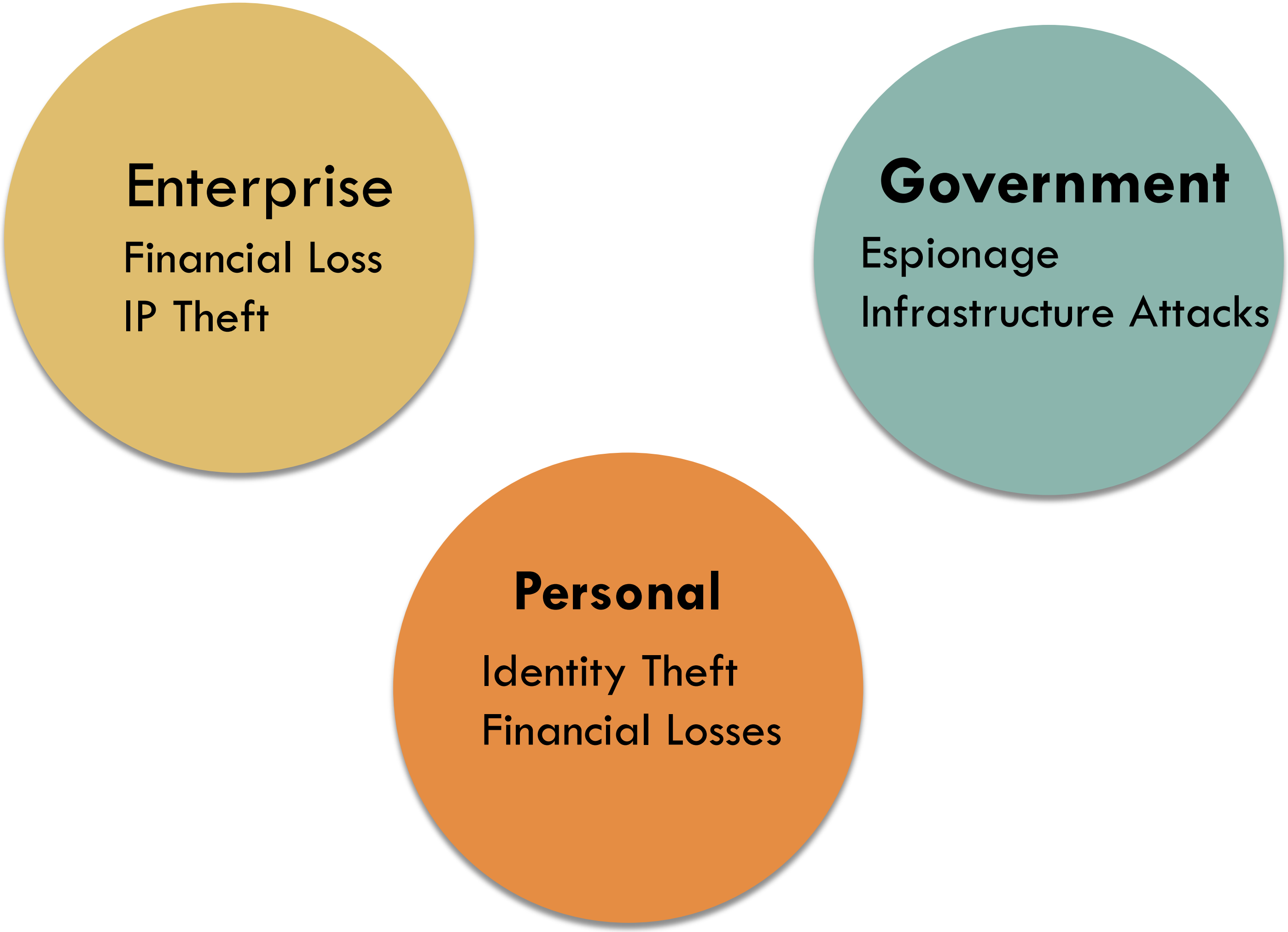
QUANTITATIVELY ANALYZING STEALTHY COMMUNICATIONS CHANNELS

By Patrick Butler, Kui Xu, Danfeng (Daphne) Yao
Computer Science @ Virginia Tech

Botnet Threats are Pervasive



Enterprise

Financial Loss

IP Theft

Government

Espionage

Infrastructure Attacks

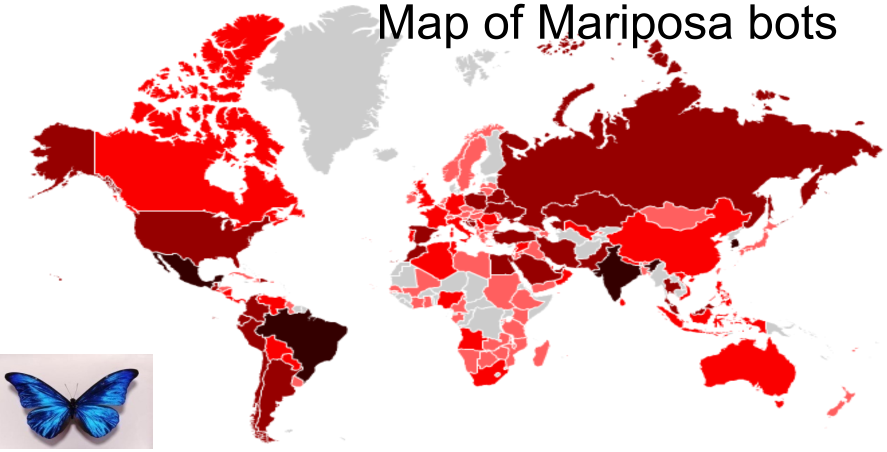
Personal

Identity Theft

Financial Losses

Botnets: Mariposa

- □ 12 Million IPs
- □ Data from 800k users
- □ Changes malware every 48 hours

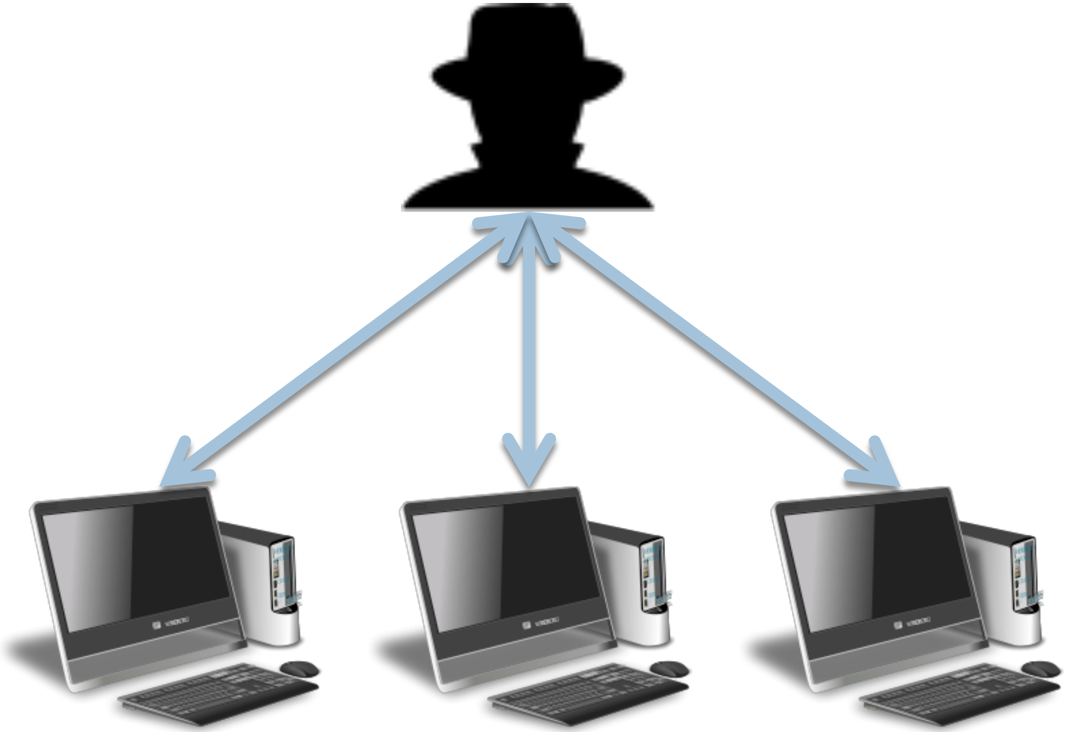


Map of Mariposa bots

How are they controlled?

Botnet Command and Control

- Current Channels for Command and Control
  - IRC
  - HTTP
  - E-Mail
  - Skype
  - Bluetooth
  - DNS?



Our Contributions: DNS C&C

- Formalize a DNS C&C protocol
  - Tunneling
  - Codewords
- How does a hacker hide illegitimate traffic?
  - Piggybacking
  - Exponentially Distributed Query Strategy
- Give a formal definition of perfect stealth in covert channels
- Define a method to generate domain name flux

DNS Communication

- Tunneling
  - Upstream: Encode data as a query
  - Downstream: Encode response as answer
  - Bidirectional, but client must continually poll
  - Arbitrary messages

- Codeword
  - Use common hostnames to signify particular command
  - Uni-directional

Blackhat’s Setup

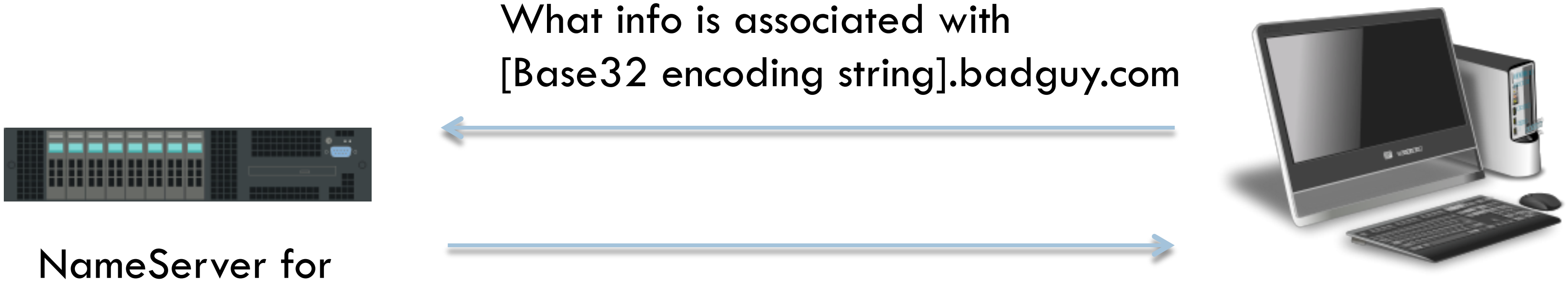
- □ Create a malicious nameserver for badguy.com

(Codeword or Tunneling)

Or

- □ Be able to seed a known DNS entry with information

(Codeword Only)



What info is associated with
[Base32 encoding string].badguy.com

NameServer for
badguy.com

CNAME/TXT [Base32 encoded string]

Bot

Codewords

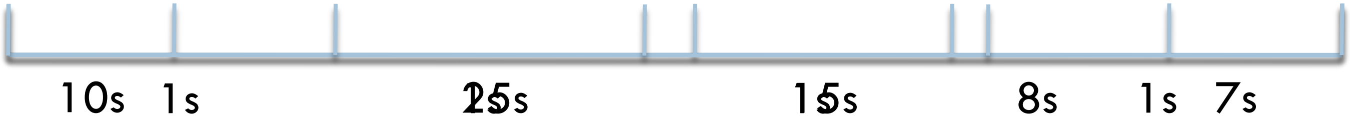
- Look up www.subdomain.domain.com
  - If address resolves to 127.0.0.1: Do Nothing
  - Else attack address
- Look up ftp.subdomain.domain.com
  - If address resolves to 127.0.0.1: Do Nothing
  - Else report status to port 2314 and download updates

Both methods allow communication between bot and controller

How do we detect codewords if they look like normal domain names?

Temporal Detection

- □ Random processes do not show uniform intervals
- □ Poisson Process: For given interval of time the probability of an event occurring is fixed.



WWBHD?

- We propose to model a normal rate and try to replicate it or hide behind it
  - Examples Include:
  - CNN.com λ =39/hour / 50 hosts
  - Google.com λ =131.5/hour / 50 hostss
- We present the Piggyback query strategy:
  1. Wait for a valid DNS request
  2. Attach a message as part of a legitimate request or send alongside a legitimate request

Experiments

- We evaluate quantitative techniques for distinguishing stealthy C&C traffic from legitimate DNS traffic
  - Packet contents, the contents of each packet are different if they are encoded data vs. valid domain
  - Timing, extra packets change the intervals between packets

Measurements

- Entropy

$$Entropy = \sum_{i=1}^{k} p_i \log_2 p_i$$

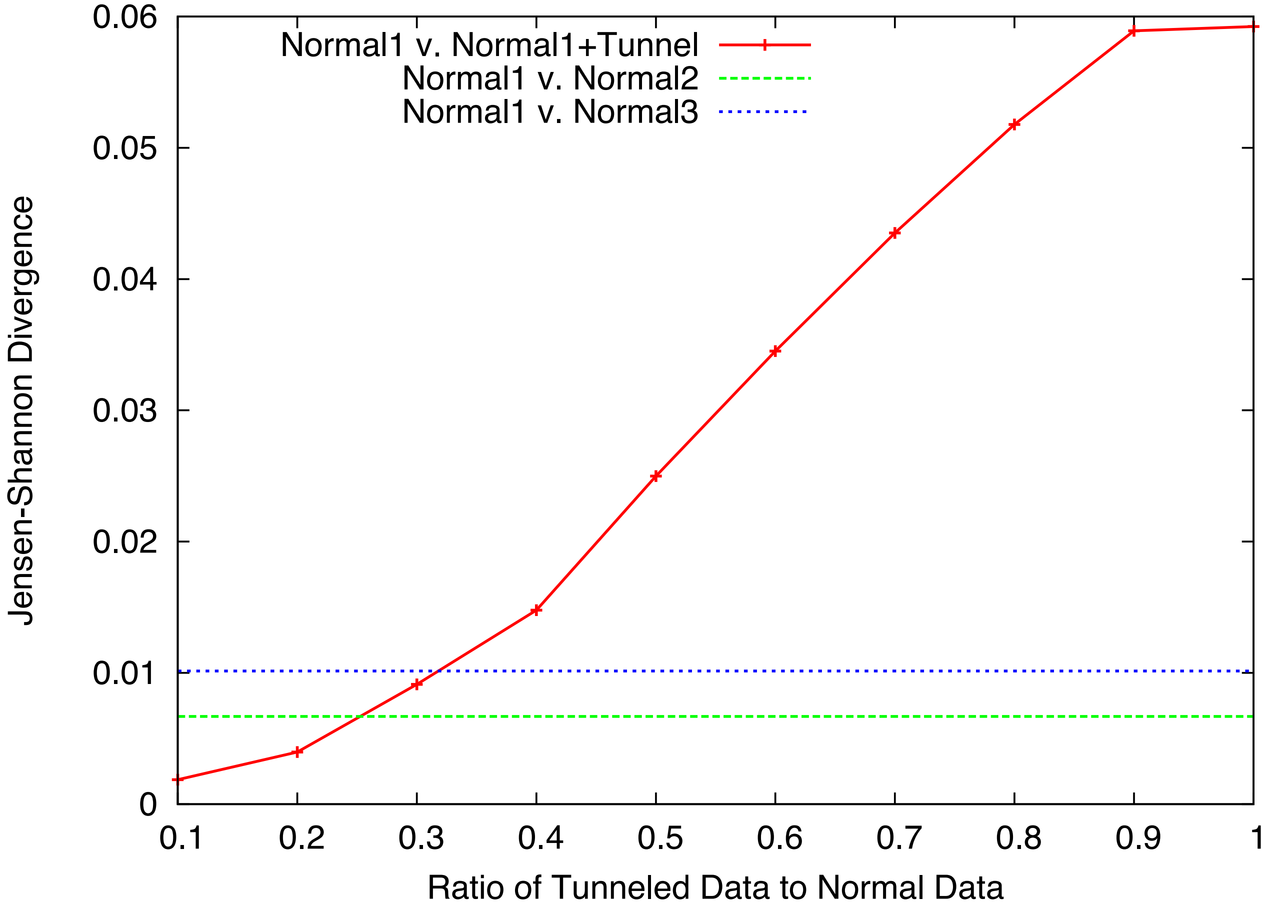
- Jensen-Shannon Divergence

$$M = \frac{1}{2}(P + Q) \qquad (2)$$

$$D_{KL}(P, Q) = \sum_{i=0}^{n} p_i \log \frac{p_i}{q_i} \qquad (3)$$

$$D_{JS} = \frac{1}{2}(D_{KL}(P, M) + D_{KL}(Q, M)) \qquad (4)$$

Packet Measurements



Differences can be measured between infested(red) data when the data contains >40% tunneled data

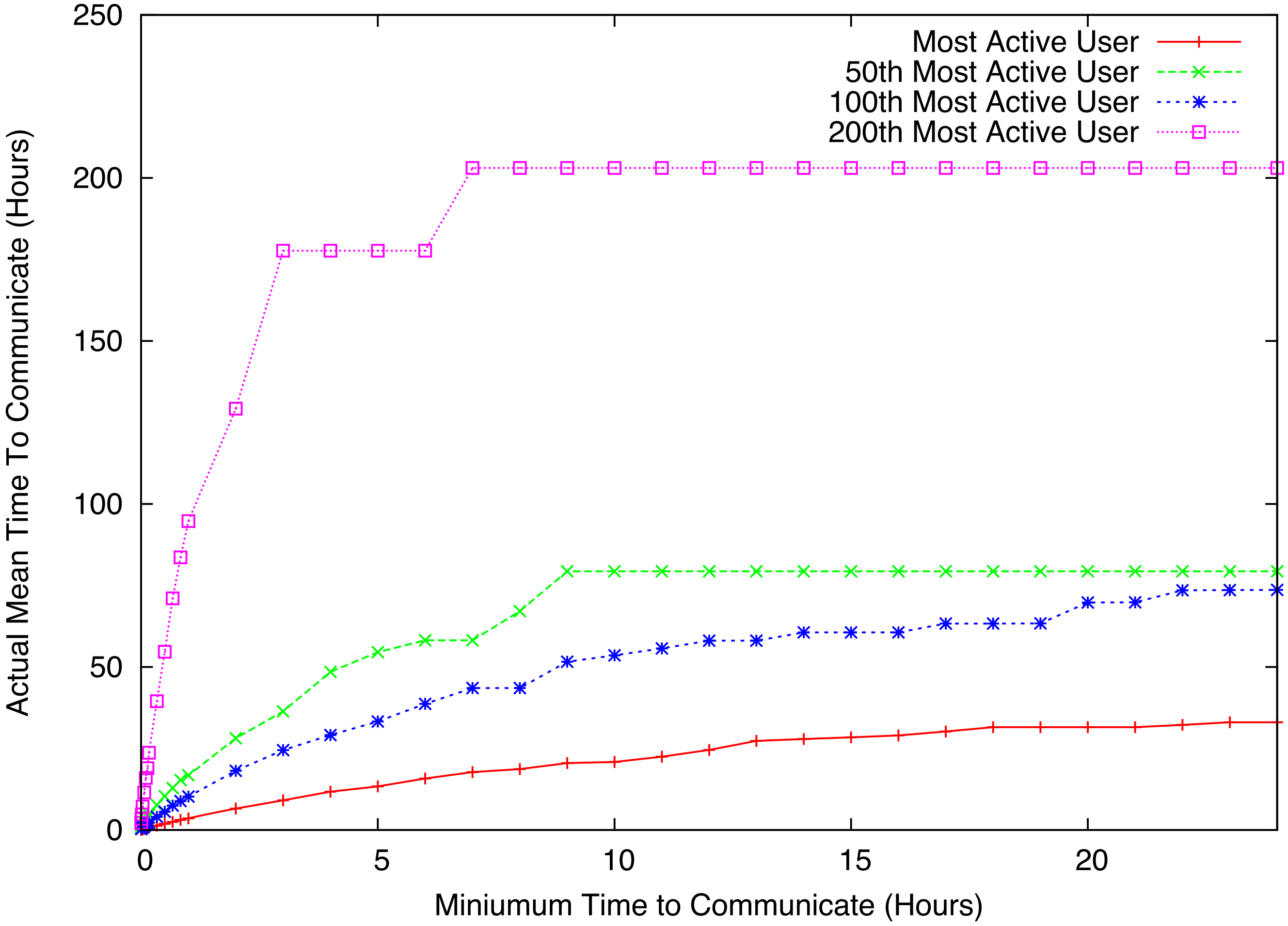
Time To Communicate

- Time-to-communicate (TTC) is defined as the time interval between two network connections (DNS queries in our setting)

- A bot master sets the Minimum TTC (MTTC) this affects the bot's Actual TTC (ATTC)

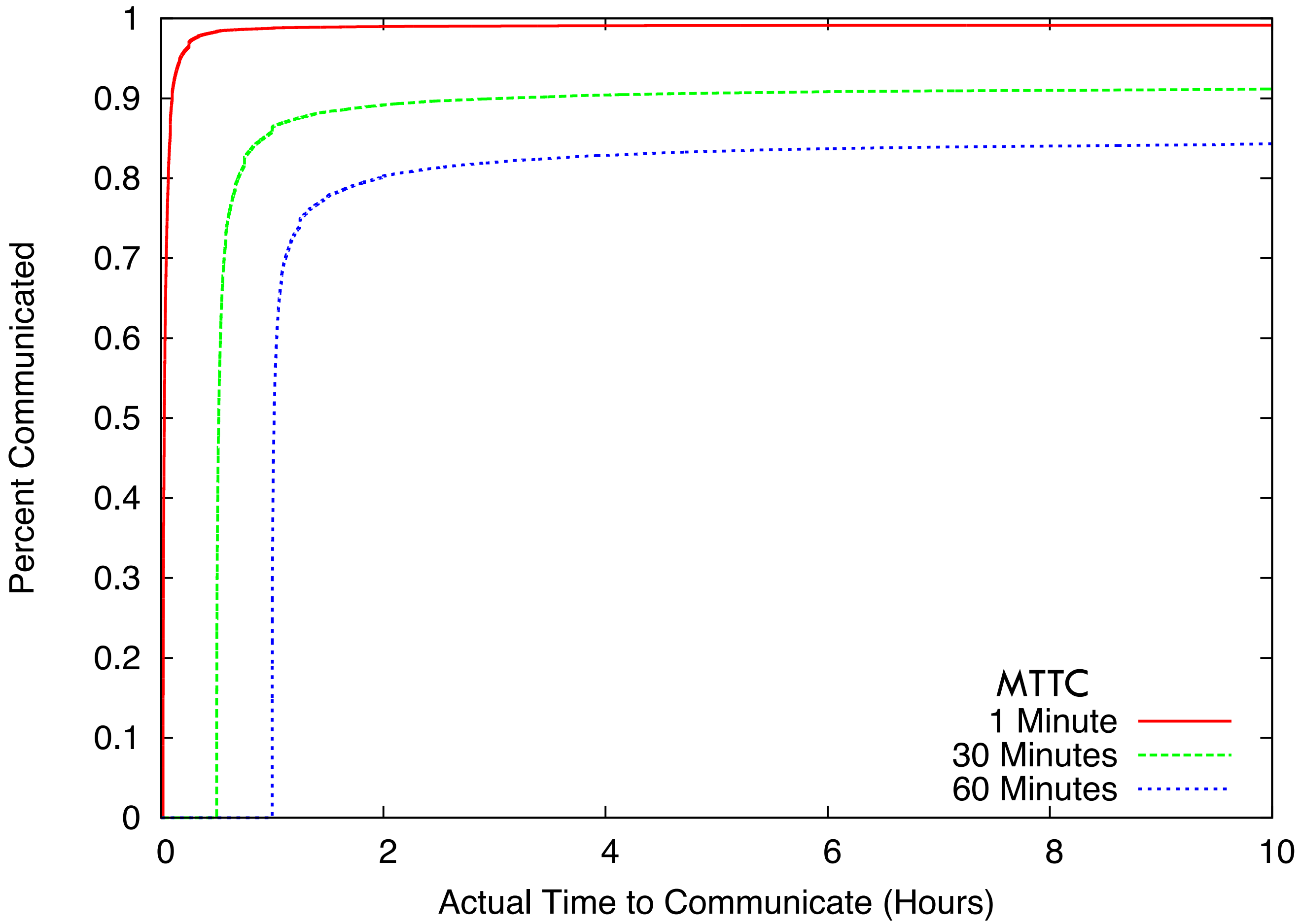
- Smaller TTC means more frequent communication

Piggyback in the Real World



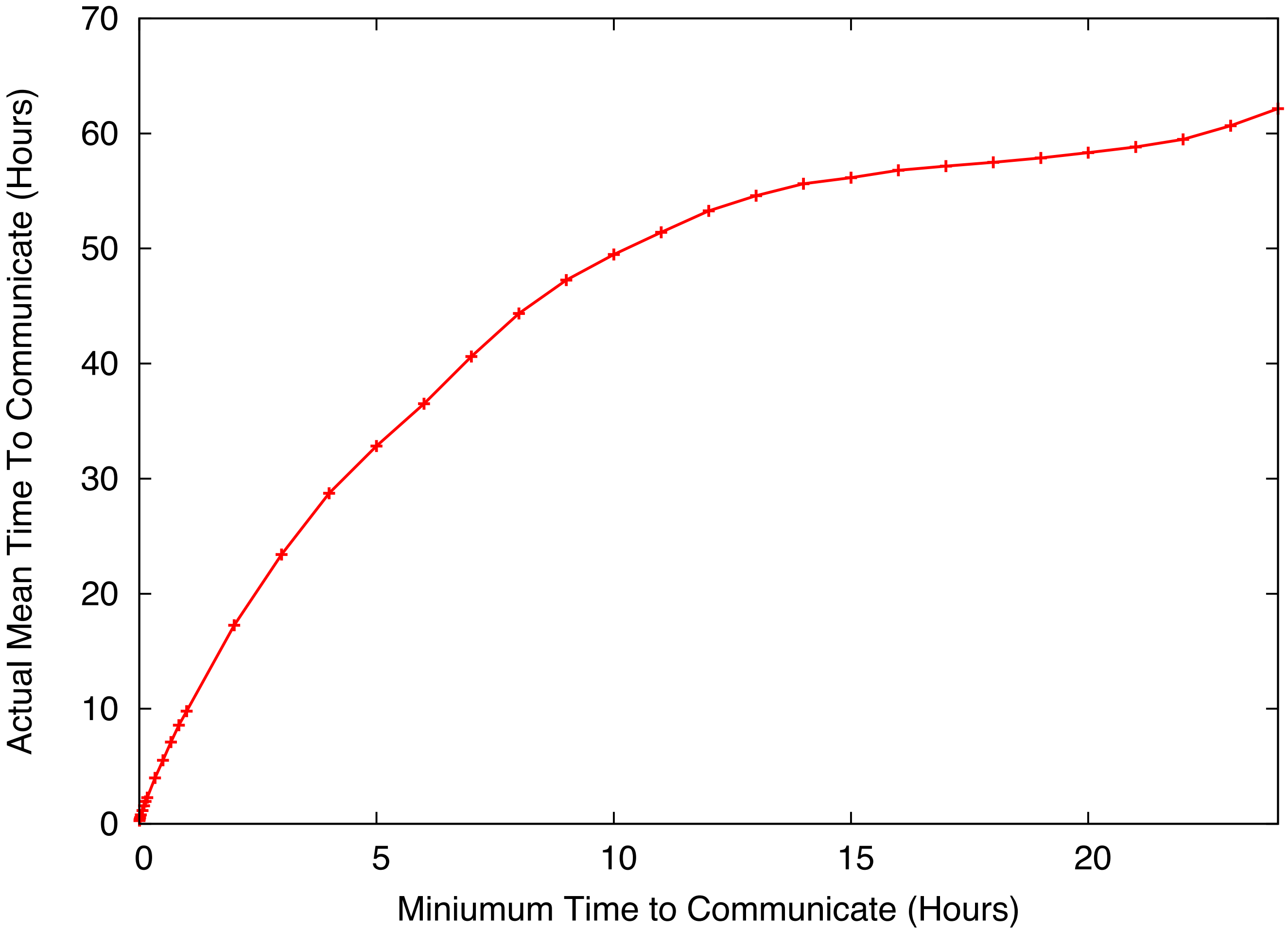
MTTC ~ ATTC for the most active users, and degrades as a function of usage frequency.

Piggyback in the Real World



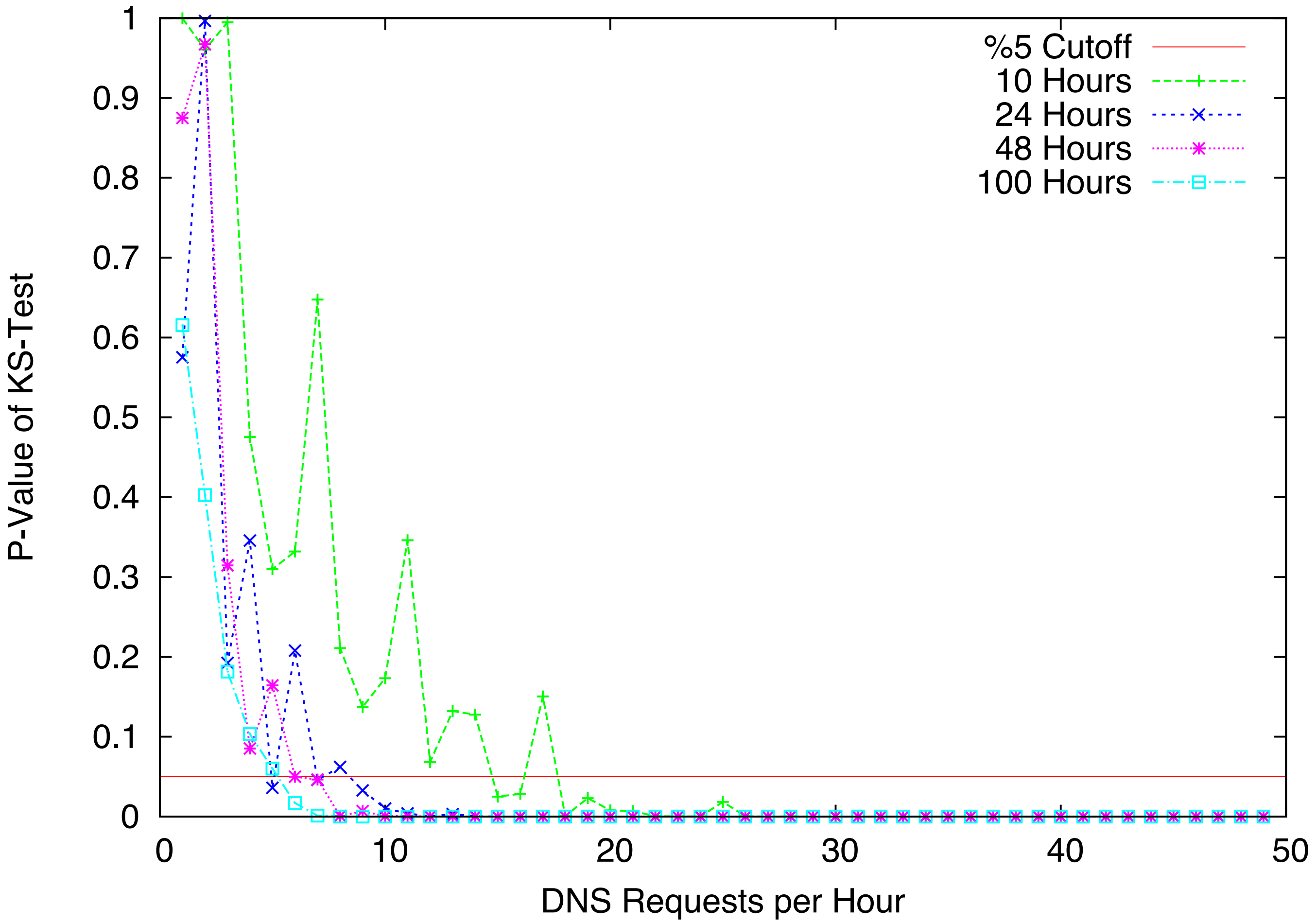
80 % of the machines communicated within 2 hours ATTC with an MTTC of 1 hr

Piggyback in the Real World



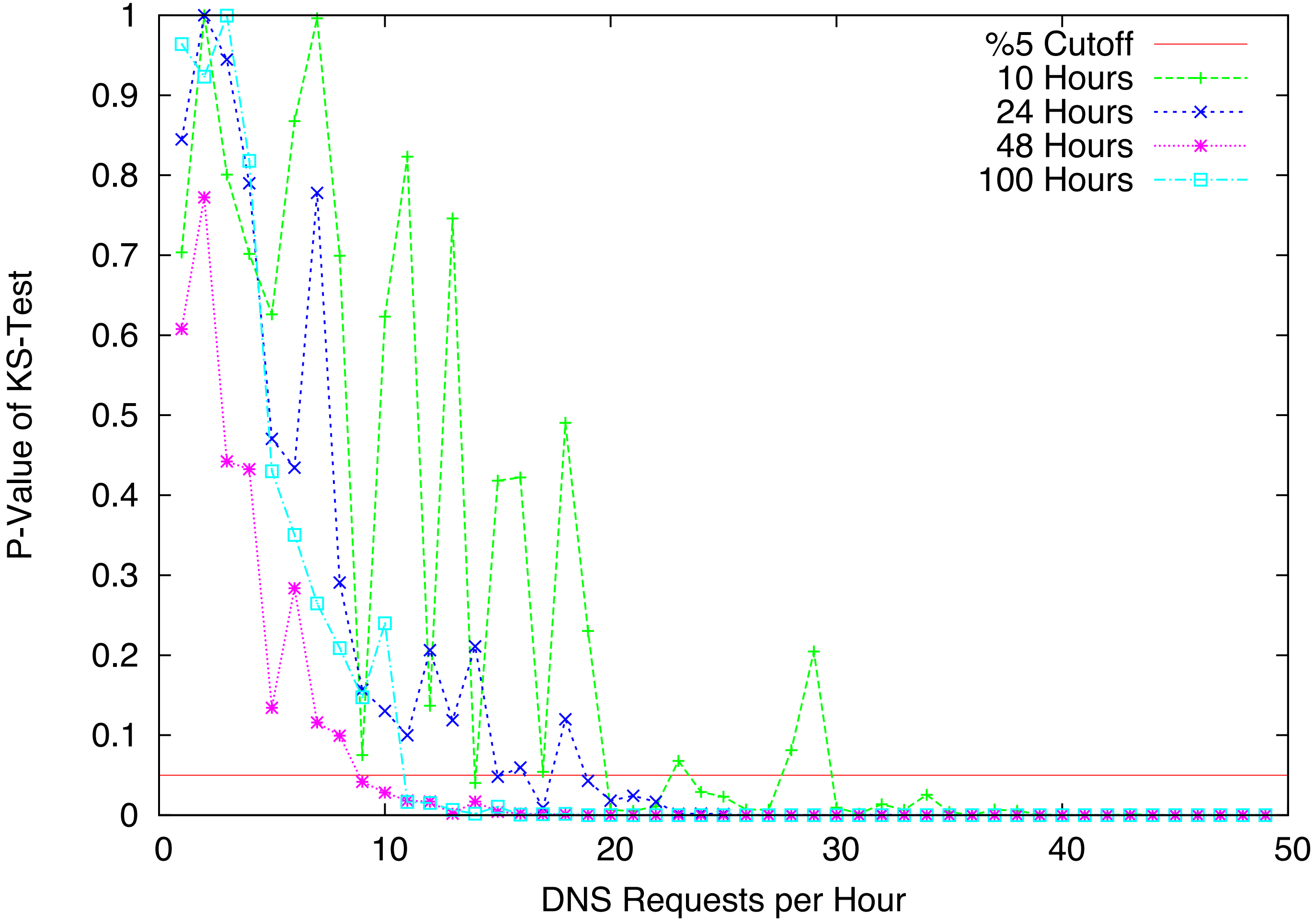
A MTTC of 5 hours will results with a mean host ATTC of 24 hours

Exponential Query: CNN



Longer recording times allow detection at lower rates

Exponential Query: Google (high rate)



Higher legitimate traffic makes detection more difficult

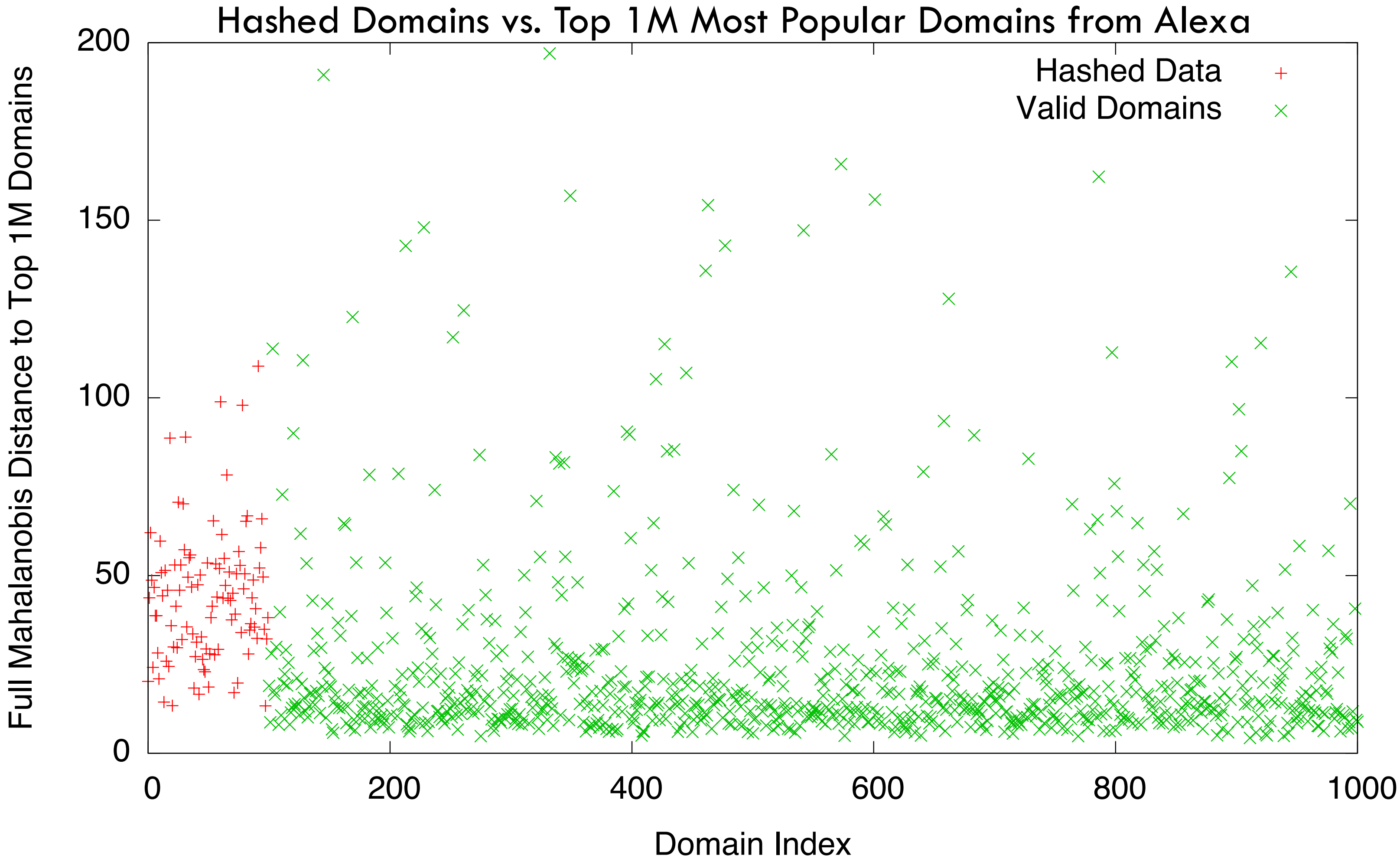
Domain Flux

- Bots and Controllers prevent blocking by generating short-lived domains
- Simple Method: $H(secret \,||\, counter)$
- Example:

$$H(\text{ACNS 2011} \,||\, 1234) = \text{d41d8cd98f00b20.com}$$

- But these do not look like real domains

Mahalanobis Distance



Hashed domains, generate a larger Mahalanobis distance

Related Works

- Karasaridis et al proposed the use of Kullback-Leibler distance to measure byte distribution of DNS packets
- R. Villamarin-Salomon and J. C. Brustoloni used DNS-based anomaly detection to detect botnets
- Stone-Gross et al observed domain flux in Torpig

Conclusions and Countermeasures

- Because almost all computers need domain-name resolution, it is impossible to block DNS traffic.
- For tunneled communications, probability distributions can be monitored to determine anomalies
- For codeword communications, monitor rate of communication for anomalies.

Take Home Message:

We demonstrate feasibility, effective, hard to detect.

Acknowledgements

- This work has been supported in part by NSF grant CAREER CNS-0953638 and CNS-0831186, and ARO grant STIR-450080.