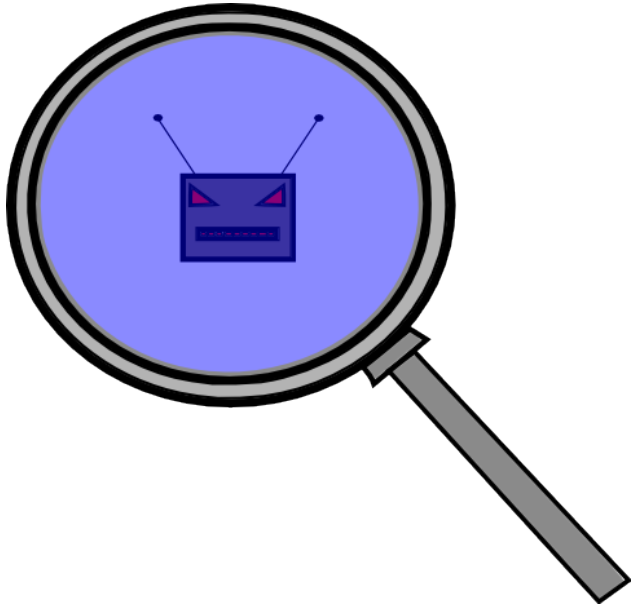
BotMagnifier: Locating Spambots on the Internet



Gianluca Stringhini
Thorsten Holz
Brett Stone-Gross
Christopher Kruegel
Giovanni Vigna

USENIX Security Symposium

August 12, 2011

Spam is a big problem


```
Received: (from spam@localhost)
        by [redacted] (8.14.1/8.13.8) id p6QL0aOt008828
        for spam-remail; Tue, 26 Jul 2011 14:00:36 -0700
X-Envelope-From: <decriesa36@eoriginal.com>
Delivered-To: mlmapdar@[redacted]
Received: from NXDOMAIN (HELO [84.240.215.82]) (84.240.215.82)
    by [redacted] (qpsmtpd/0.43rc1) with ESMTP; Tue, 26 Jul 2011 14:00:36 -0700
Received: from  (192.168.1.71) by eoriginal.com (84.240.215.82) with Microsoft SMTP Server id 8.0.685.24; Wed, 27 Jul 2011 03:00:35 +06
Message-ID: <4E2F2AE3.408090@eoriginal.com>
Date: Wed, 27 Jul 2011 03:00:35 +0600
From: "Monty Ward" <decriesa36@eoriginal.com>
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.9) Gecko/20101112 Thunderbird/3.1.4
MIME-Version: 1.0
To: <mlmapdar@[redacted]>
Subject: Telex to help defeat web censors
Content-Type: text/html; charset=UTF-8
Content-Transfer-Encoding: 7bit

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
  <head>

    <meta http-equiv="content-type" content="text/html; charset=UTF-8">
  </head>
  <body bgcolor="#ffffff" text="#000000">
    <p align="center">
                <img  src="http://swpills.ru/Az2yBmfipqsxfqBjDrlrzoAzmE/38864.jpg" border="0">
</p>
<p align="center">cheap viagra</p>
<p align="center"><a href="http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE">http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE</p>
  </body>
```

Spam is sneaky

THE COMPUTER SECURITY GROUP AT UCSB

```
Received: (from spam@localhost)
        by [redacted] (8.14.1/8.13.8) id p6QL0aOt008828
        for spam-remail; Tue, 26 Jul 2011 14:00:36 -0700
X-Envelope-From: <decriesa36@eoriginal.com>
Delivered-To: mlmapdar@[redacted]
Received: from NXDOMAIN (HELO [84.240.215.82]) (84.240.215.82)
    by [redacted] (qpsmtpd/0.43rc1) with ESMTP; Tue, 26 Jul 2011 14:00:36 -0700
Received: from  (192.168.1.71) by eoriginal.com (84.240.215.82) with Microsoft SMTP Server id 8.0.685.24; Wed, 27 Jul 2011 03:00:35 +06
Message-ID: <4E2F2AE3.408090@eoriginal.com>
Date: Wed, 27 Jul 2011 03:00:35 +0600
From: "Monty Ward" <decriesa36@eoriginal.com>
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.9) Gecko/20101112 Thunderbird/3.1.4
MIME-Version: 1.0
To: <mlmapdar@[redacted]>
Subject: Telex to help defeat web censors
Content-Type: text/html; charset=UTF-8
Content-Transfer-Encoding: 7bit

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
  <head>

    <meta http-equiv="content-type" content="text/html; charset=UTF-8">
  </head>
  <body bgcolor="#ffffff" text="#000000">
    <p align="center">
                <img  src="http://swpills.ru/Az2yBmfipqsxfqBjDrlrzoAzmE/38864.jpg" border="0">
</p>
<p align="center">cheap viagra</p>
<p align="center"><a href="http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE">http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE</p>
  </body>
```

Spam is sneaky


```
Received: (from spam@localhost)
        by [redacted] (8.14.1/8.13.8) id p6QL0aOt008828
        for spam-remail; Tue, 26 Jul 2011 14:00:36 -0700
X-Envelope-From: <decriesa36@eoriginal.com>
Delivered-To: mlmapdar@[redacted]
Received: from NXDOMAIN (HELO [84.240.215.82]) (84.240.215.82)
    by [redacted] (qpsmtpd/0.43rc1) with ESMTP; Tue, 26 Jul 2011 14:00:36 -0700
Received: from  (192.168.1.71) by eoriginal.com (84.240.215.82) with Microsoft SMTP Server id 8.0.685.24; Wed, 27 Jul 2011 03:00:35 +06
Message-ID: <4E2F2AE3.408090@eoriginal.com>
Date: Wed, 27 Jul 2011 03:00:35 +0600
From: "Monty Ward" <decriesa36@eoriginal.com>
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.9) Gecko/20101112 Thunderbird/3.1.4
MIME-Version: 1.0
To: <mlmapdar@[redacted]>
Subject: Telex to help defeat web censors
Content-Type: text/html; charset=UTF-8
Content-Transfer-Encoding: 7bit

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
  <head>

    <meta http-equiv="content-type" content="text/html; charset=UTF-8">
  </head>
  <body bgcolor="#ffffff" text="#000000">
    <p align="center">
                <img  src="http://swpills.ru/Az2yBmfipqsxfqBjDrlrzoAzmE/38864.jpg" border="0">
</p>
<p align="center">cheap viagra</p>
<p align="center"><a href="http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE">http://swpills.ru/?ELtAz2yBmfipqsxfqBjDrlrzoAzmE</p>
  </body>
```

Tracking Spambots is important

Botnets are responsible for 85% of worldwide spam

- ISPs and organizations can clean up their networks
- Existing blacklists (DNSBL) can be improved
- Mitigation efforts can be directed to the most aggressive botnets

Tracking Spambots is challenging

- The IP addresses of infected machines change frequently
- It is easy to recruit "new members" into a botnet

e

An approach is to set up *spam traps*. However, a few problems arise:

- Only a subset of the bots will send emails to the spam trap addresses
- Some botnets target only users located in certain countries

Basic Insight

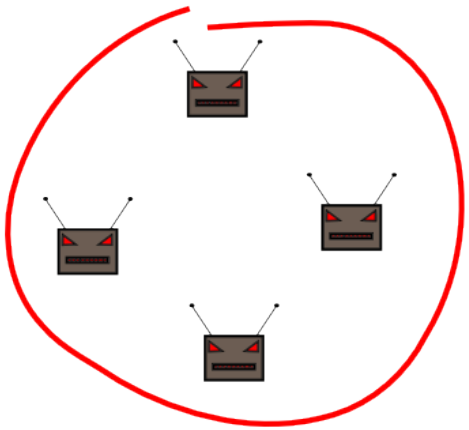
Bots that belong to the same botnet share similarities

As a result, they will follow a similar behavior when sending spam

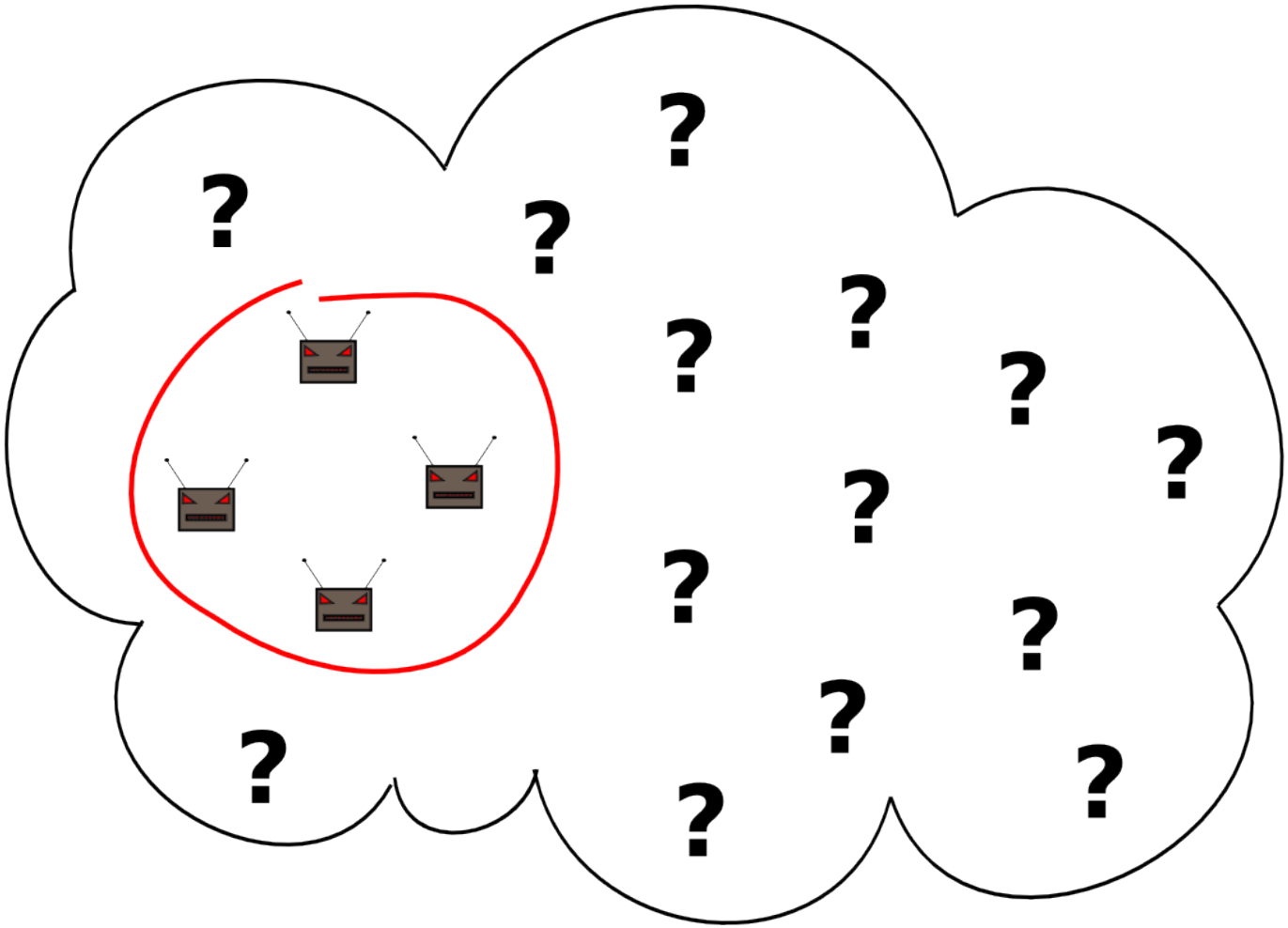
Commoditized botnets could appear as multiple botnets

By observing a portion of a botnet, it is possible to identify more bots that belong to it

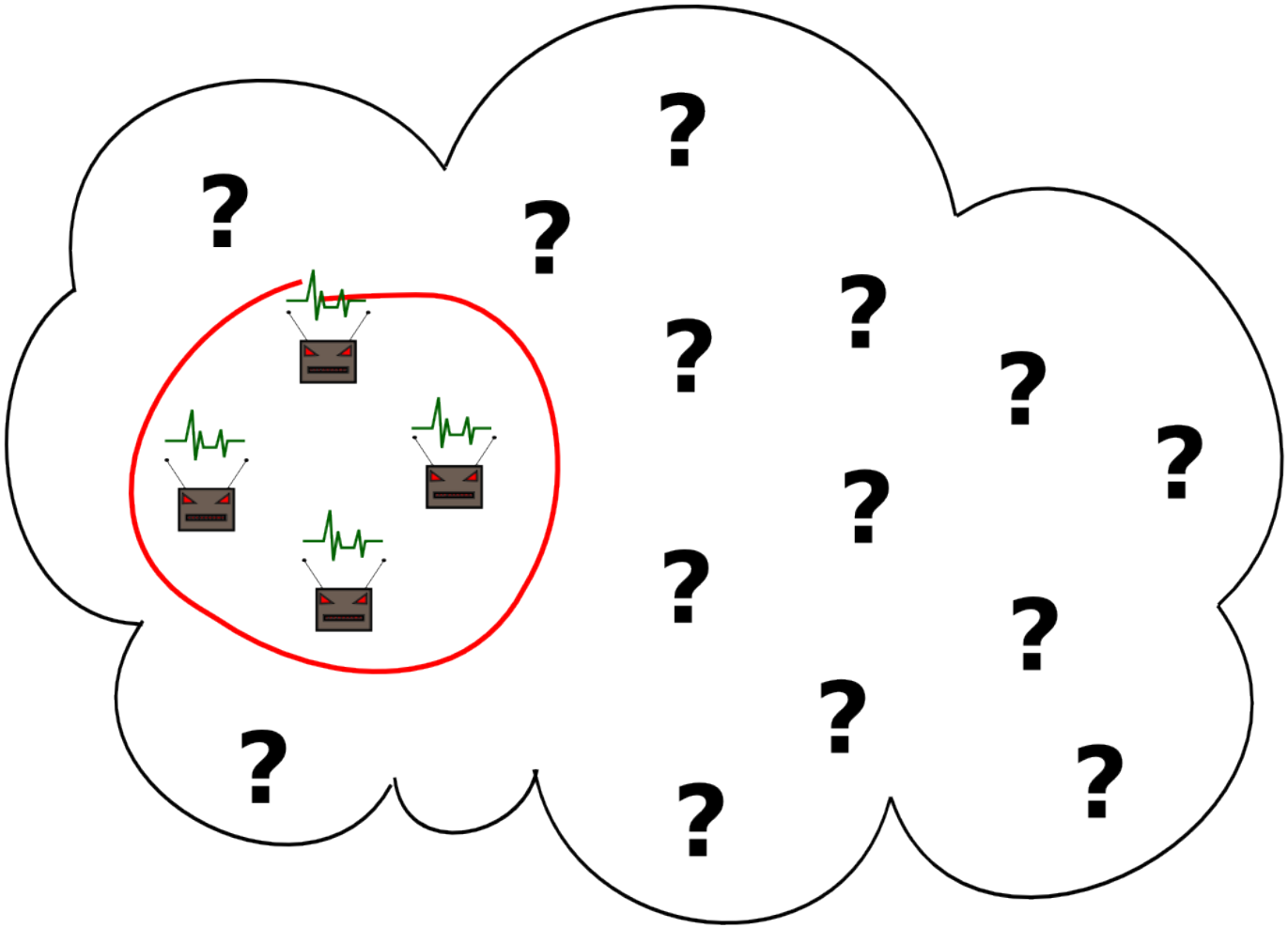
Our Approach



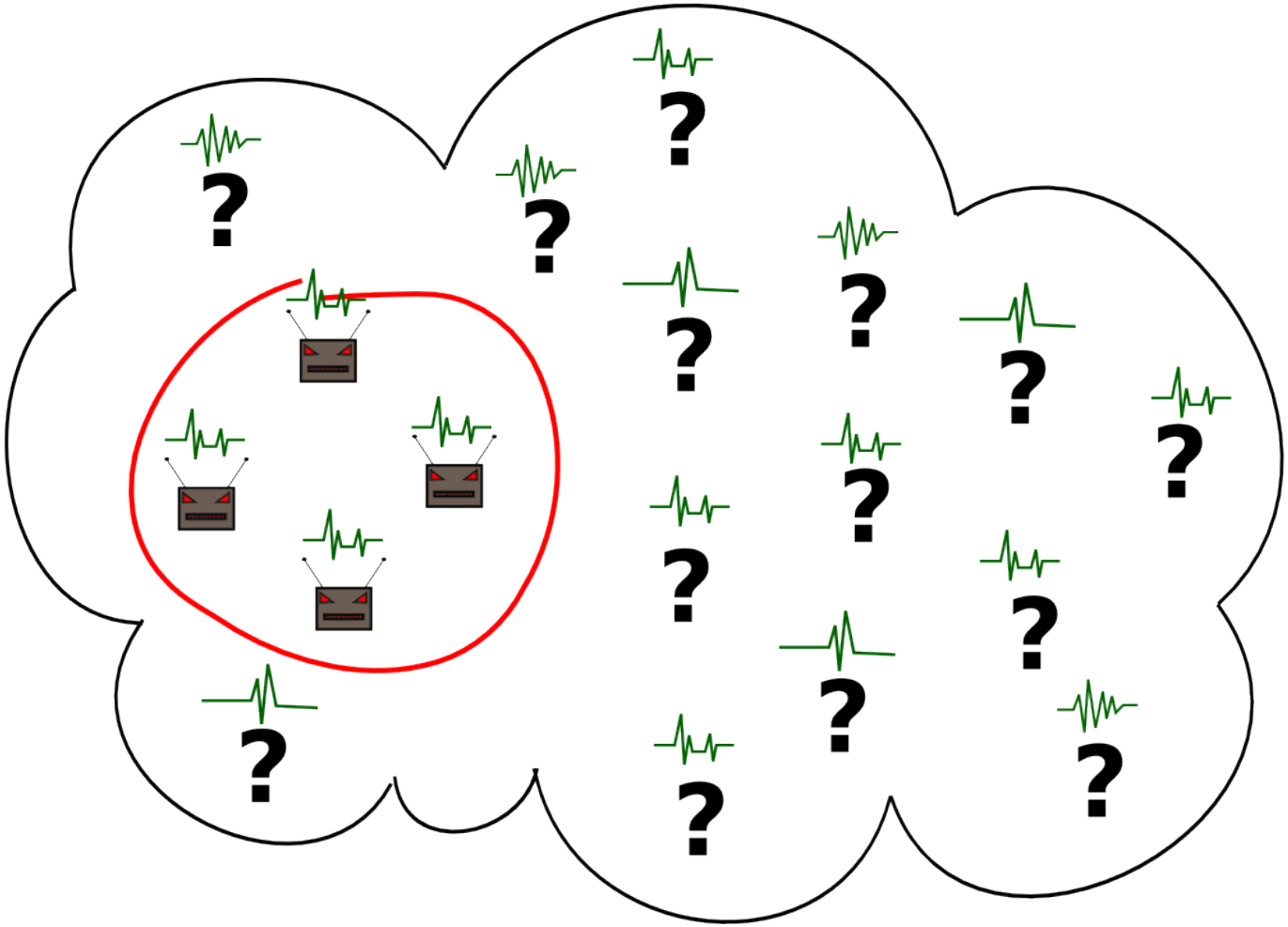
Our Approach



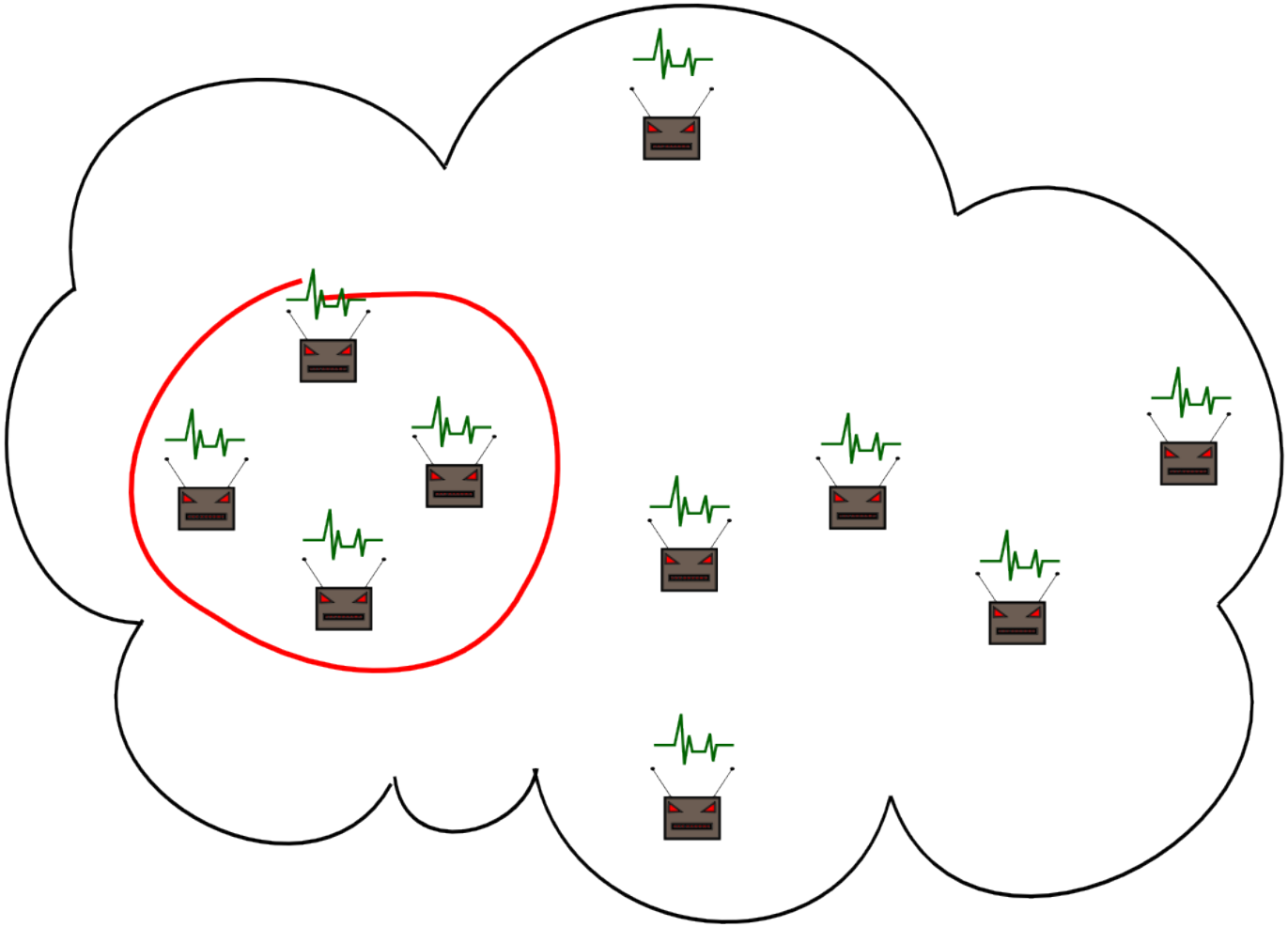
Our Approach



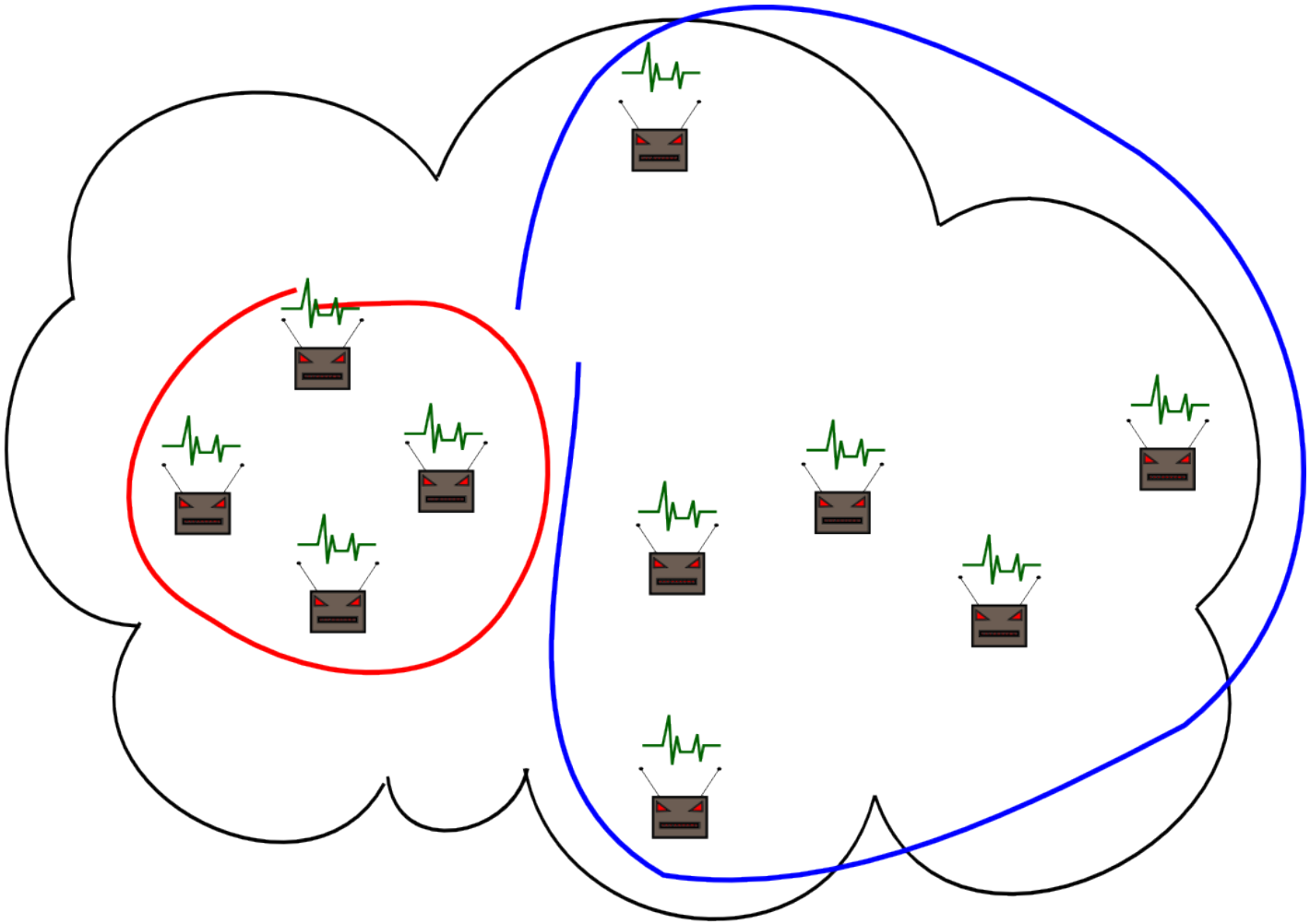
Our Approach



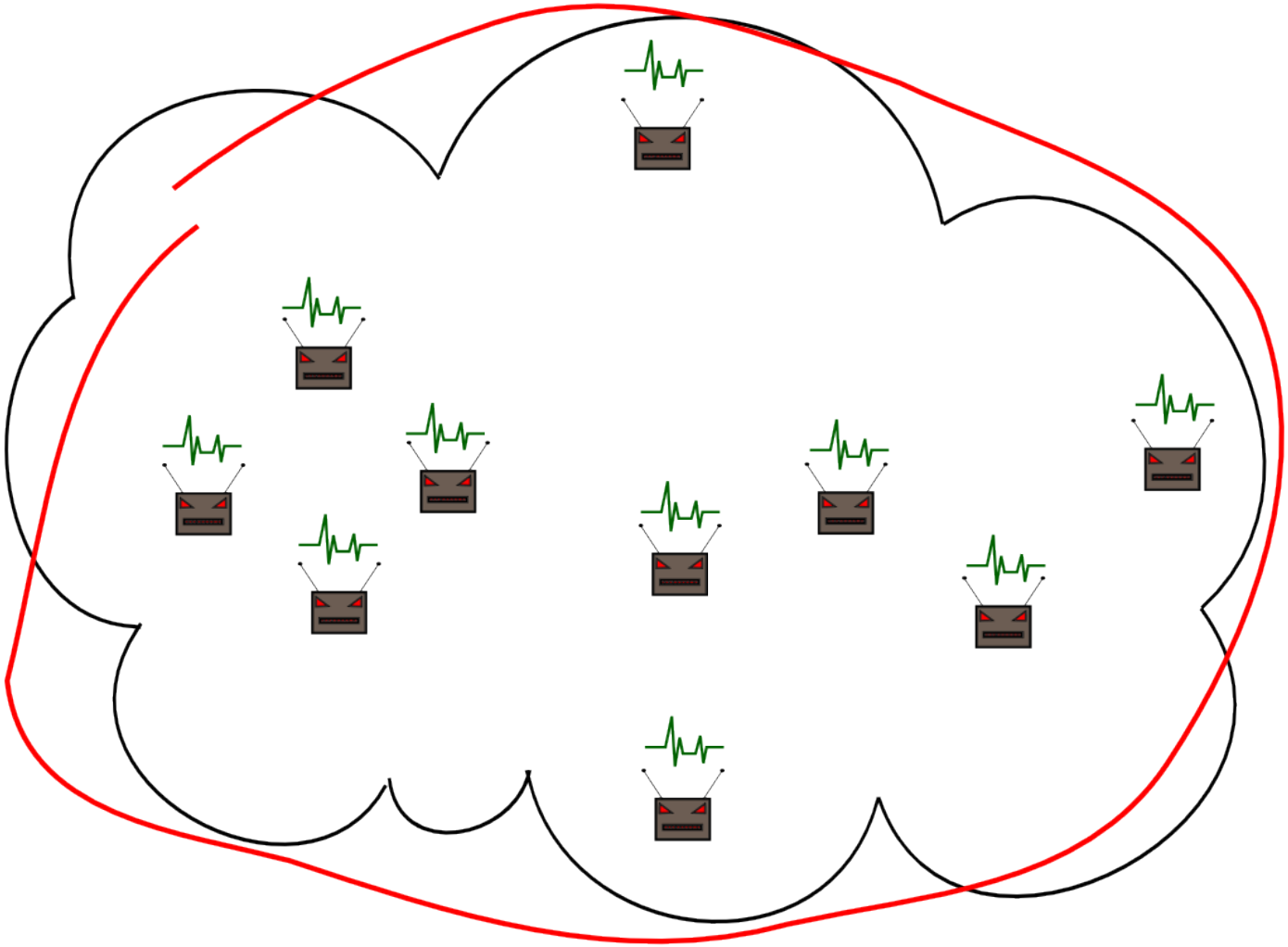
Our Approach



Our Approach



Our Approach



Input Datasets

How can we achieve this?

Our approach takes two datasets as input:

- The IP addresses of known spamming bots, grouped by spam campaign (**seed pools**)
- A log of email transactions carried out on the Internet, both legitimate and malicious (**transaction log**)

Our System

We implemented our approach in a tool, called BotMagnifier

We used a large spam trap to populate seed pools

We used the logs of a Spamhaus mirror as transaction log

- Each query to the Spamhaus mirror corresponds to an email
- We show how BotMagnifier also works when using other datasets as transaction logs

Our System

BotMagnifier is executed periodically

It takes as input a set of seed pools

At the end of each observation period, it outputs:

- The IP addresses of the bots in the magnified pools
- The name of the botnet that carried out each campaign

Phase I: Building Seed Pools

Set of IP addresses that participated in a specific spam campaign

Built using the data of a spam trap set up by a large US ISP

$$\approx 1\text{M messages / day}$$

We consider messages with similar subject lines as part of the same campaign

Design decisions:

- Minimum seed pool size: 1,000 IP addresses
- Observation period: 1 day

Phase II: Characterizing Bot Behavior

THE COMPUTER SECURITY GROUP AT UCSB

For each seed pool:

- We query the transaction log to find all the events that are associated with the IP addresses in it
- We analyze the set of destinations targeted and build a **target set**

Problem

The target sets of two botnets might have substantial overlaps

We extract the set of destinations that are **unique** to each seed pool (**characterizing set**)

Phase III: Bot Magnification

Goal: find the IP addresses of previously-unknown bots

BotMagnifier considers an IP address x as behaving similarly to the bots in a seed pool if:

- x sent emails to at least N destinations in the target set
- x never sent an email to a destination outside the target set
- x has contacted at least one destination in the characterizing set

How large should N be?

Threshold Computation

N should be greater for campaigns targeting a larger number of destinations

$$N = k \cdot |T(p_i)|, 0 < k \leq 1$$

where $|T(p_i)|$ is the size of the target set, and k is a parameter

Precision vs. Recall analysis on ten campaigns for which we had ground truth (coming from Cutwail C&C servers)

$$k = k_b + \frac{\alpha}{|T(p_i)|} \rightarrow k_b = 8 \cdot 10^{-4}, \alpha = 10$$

Phase IV: Spam Attribution

We want to "label" spam campaigns based on the botnet that carried them out

Running Malware Samples

We match the subject lines observed in the wild with the ones of the bots we ran

Botnet Clustering

- IP overlap
- Destination distance
- Bot distance

Validation of the Approach

To validate our approach, we studied *Cutwail*, for which we had direct data about the IP addresses of the infected machines

The C&C servers we analyzed accounted for approximately 30% of the botnet

We ran the validation experiment for the period between July 28 and August 16, 2010

For each of the 18 days:

- We selected a subset of the IP addresses referenced by the C&C servers
- With the help of the spam trap, we identified the campaigns carried out
- We generated the seed and magnified pools

BotMagnifier identified 144,317 IP addresses as bots. Of these, 33,550 were actually listed in the C&C databases ($\approx$ 23%).

Overview of Tracking Results

We ran our system between September 28, 2010 and February 5, 2011

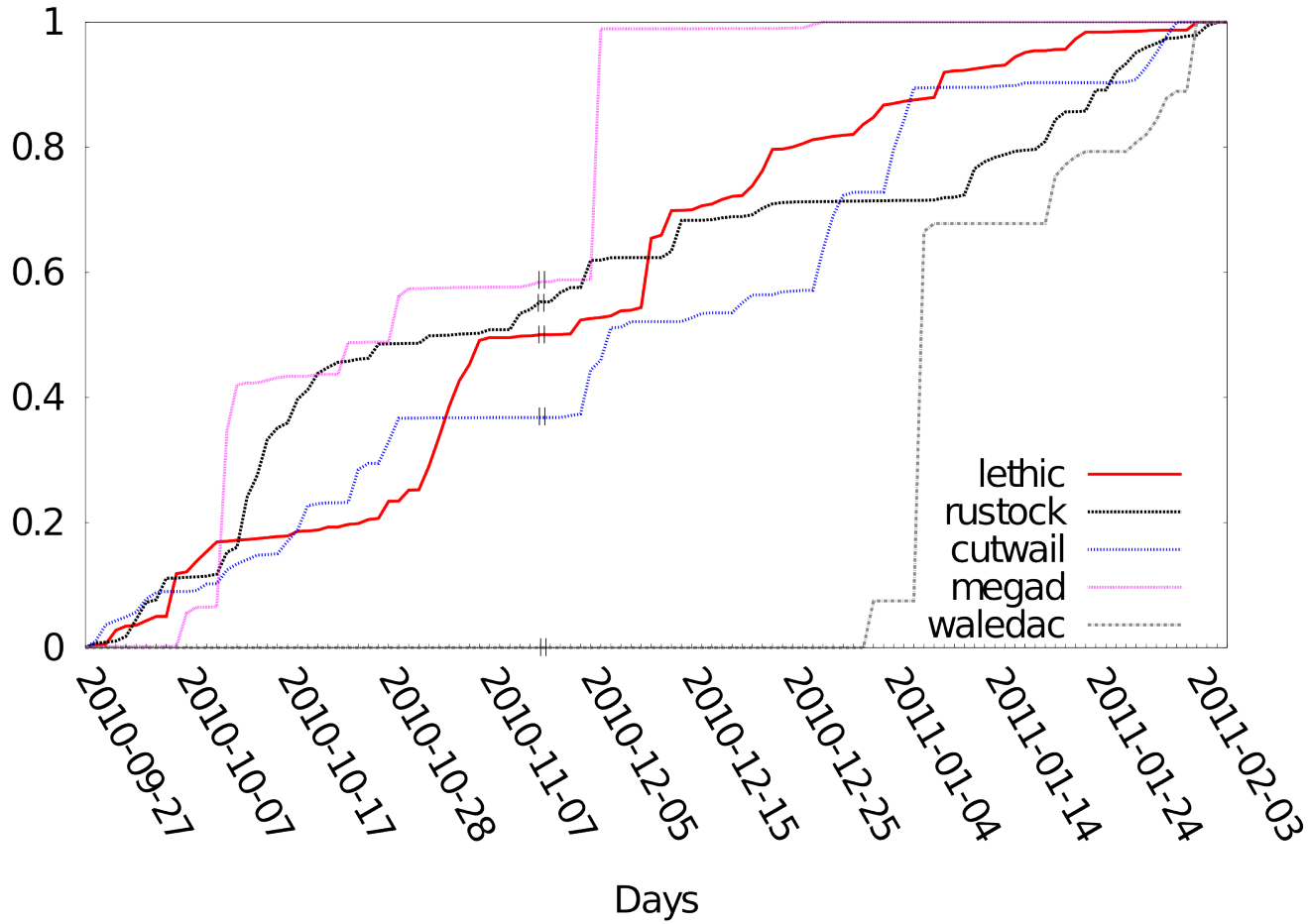
BotMagnifier tracked 2,031,110 bot IP addresses

Of these, 925,978 belonged to magnified pools, while the others belonged to seed pools

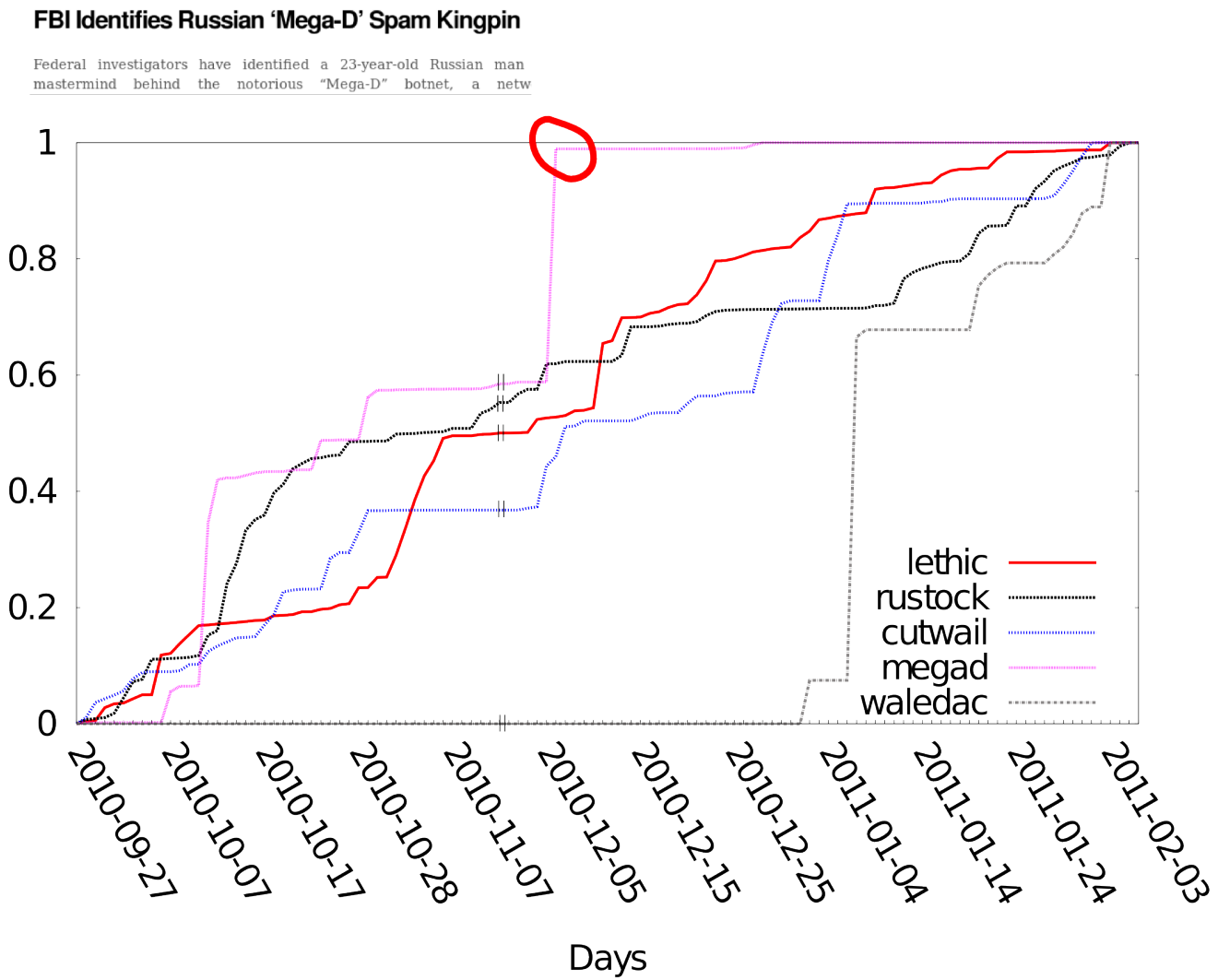
1.6% estimated false positives

Botnet	**Total # of IP addresses**	**# of "static" IP addresses**
Lethic	887,852	117,335
Rustock	676,905	104,460
Cutwail	319,355	34,132
MegaD	68,117	3,055
Waledac	36,058	3,450

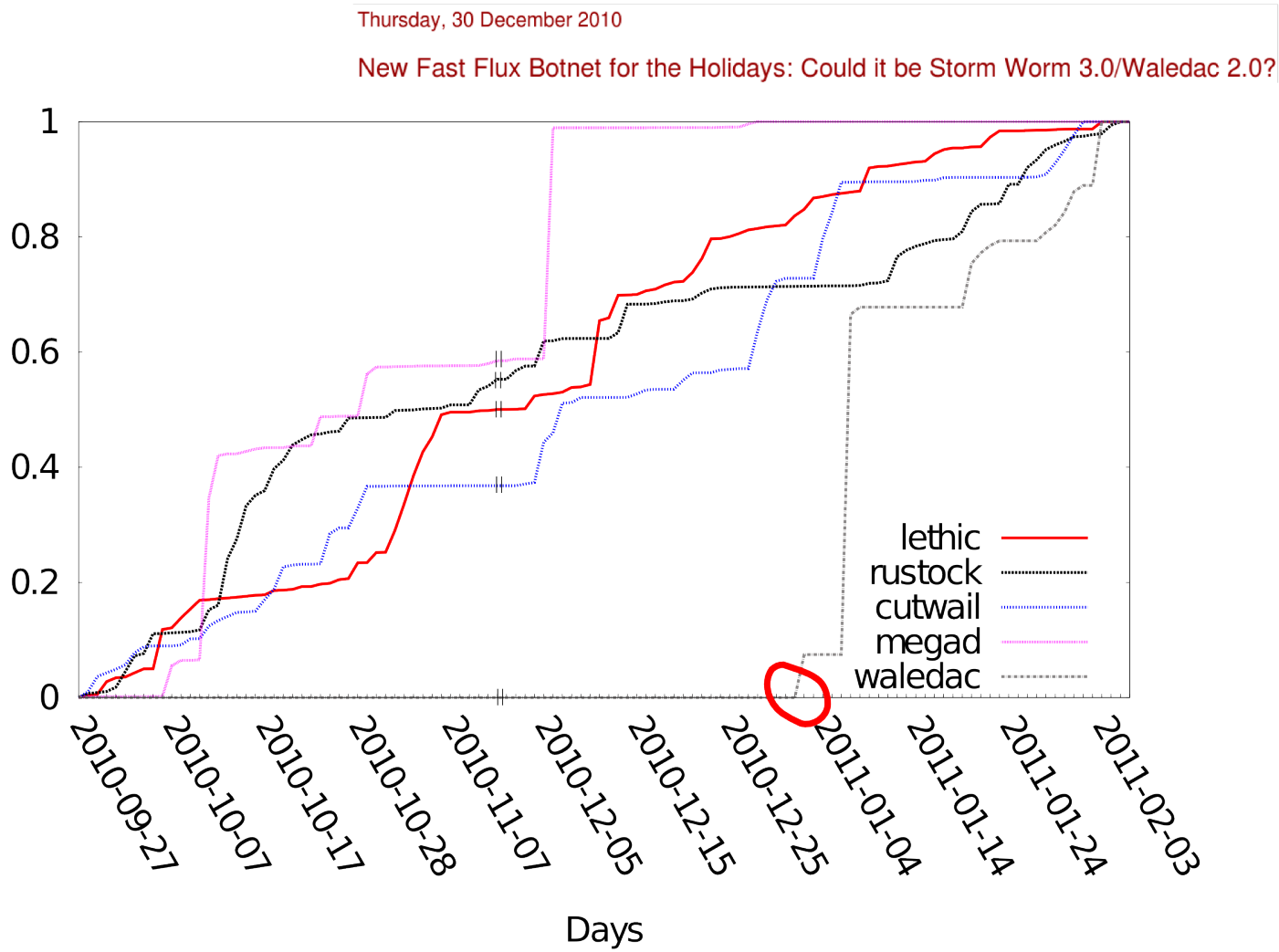
Overview of Tracking Results



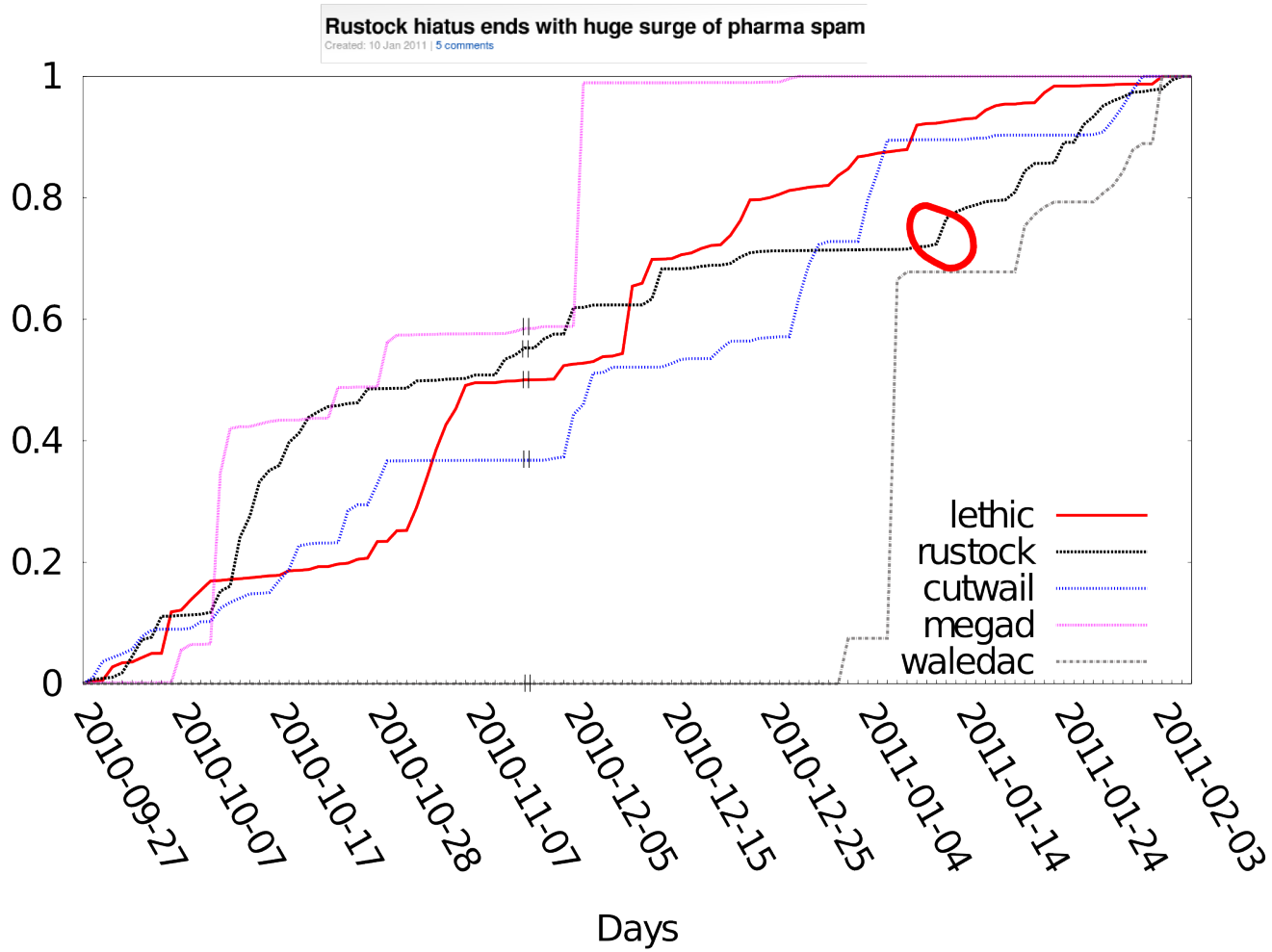
Overview of Tracking Results



Overview of Tracking Results



Overview of Tracking Results



Application of Results

Can BotMagnifier improve existing blacklists?

We analyzed the email logs from the UCSB CS mail server from November 30, 2010 to February 8, 2011

- If a mail got delivered, the IP address was not blacklisted at the time
- The spam ratios computed by *SpamAssassin* provide us with ground truth

28,563 emails were marked as spam, 10,284 IP addresses involved. 295 of them were detected by BotMagnifier, for a total of 1,225 emails ($\approx$ 4%)

We then looked for false positives. BotMagnifier wrongly identified 12 out of 209,013 IP addresses as bots.

Data Stream Independence

We show how BotMagnifier can be used on alternative datasets, too

We used the netflow logs from an ISP backbone routers
1.9M emails logged per day

We had to use new values for k_b and α

The experiment lasted from January 20, 2011 to January 28, 2011.

BotMagnifier identified 36,739 in magnified pools. This grew the seed pools by 38%.

Conclusions

We presented BotMagnifier, a tool for tracking and analyzing spamming botnets

We showed that our approach is able to accurately identify and track botnets

By using more comprehensive datasets, the magnification results would get better

Thanks!

email: gianluca@cs.ucsb.edu
twitter: @gianlucaSB