Characterizing Intelligence Gathering and Control on an Edge Network

MARTIN ARLITT
HP Labs, Palo Alto, CA, USA
University of Calgary, Calgary, AB, Canada
NIKLAS CARLSSON
Linköping University, Linköping, Sweden
PHILLIPA GILL
University of Toronto, Toronto, ON, Canada
ANIKET MAHANTI
University of Calgary, Calgary, AB, Canada
CAREY WILLIAMSON
University of Calgary, Calgary, AB, Canada

There is a continuous struggle for control of resources at every organization that is connected to the Internet. The local organization wishes to use its resources to achieve strategic goals. Some external entities seek direct control of these resources, to use for purposes such as spamming or launching denial-of-service attacks. Other external entities seek indirect control of assets (e.g., users, finances), but provide services in exchange for them.

Using a year-long trace from an edge network, we examine what various external organizations know about one organization. We compare the types of information exposed by or to external organizations using either active (*reconnaissance*) or passive (*surveillance*) techniques. We also explore the direct and indirect control external entities have on local IT resources.

Categories and Subject Descriptors: C.2.0 [**Computer-Communications Networks**]: General

General Terms: Measurement

Additional Key Words and Phrases: Workload Characterization

Authors' address: M. Arlitt, HP Labs, Palo Alto, CA, USA, martin.arlitt@hp.com; N. Carlsson, Linköping University, Linköping, Sweden, niklas.carlsson@liu.se; P. Gill, University of Toronto, Toronto, ON, Canada, phillipa@cs.utoronto.ca; A. Mahanti, University of Calgary, Calgary, AB, Canada, mahantia@ucalgary.ca; C. Williamson, University of Calgary, Calgary, AB, Canada, carey@cpsc.ucalgary.ca

©ACM, (2011). This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version will be published in ACM Transactions on Internet Technology.

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
© 20YY ACM 0000-0000/20YY/0000-0001 $5.00

1. INTRODUCTION

Many organizations rely on the Internet and other IT resources to achieve their goals. However, there is a continuous struggle for control of resources at each and every organization connected to the Internet. In addition to management teams that allocate resources to achieve internal goals, there typically are many other external organizations (and/or individuals) interested in gaining control of the organization's resources. As part of this struggle, the organization's technical team faces numerous challenges, one of which is responding to security risks that could prevent the IT infrastructure from functioning as intended.

There are a variety of external entities that are interested in the local organization. Some seek direct control of the local organization's assets (e.g., computers, finances). Others seek indirect control of an organization's assets, but provide services in exchange. Both groups collect information in pursuit of these goals, typically via *reconnaissance* (i.e., active measurements like scanning) of the local organization or *surveillance* (i.e., observation through passive measurements) of the local organization's use of Internet services.

Using a year-long trace of network activity from a large university, we examine how much information is leaked to external organizations, how the information is leaked, and how much control they have within the target organization. The purpose of our characterization study is to improve the understanding of these issues, so that proper solutions can be developed.

We used the following five questions to guide our work:

—*Who are we dealing with?*

—*What do they know about the local organization?*

—*How did they obtain that information?*

—*Which intelligence gathering technique is the most effective?*

—*What control do they have over local resources?*

The primary contribution of our work is the characterization of a year in the life of an edge network. We quantify the extent of information that various external organizations learn about the IT infrastructure of an edge network. On this topic, we believe we are the first to compare the information gained by two different *intelligence gathering* techniques. We quantify the control (direct or indirect) that external entities have on local IT resources. Our results show that many external entities have extensive, up-to-date information on the edge network. While some of the "leaks" could be prevented, others will be more difficult to eliminate. Instead, edge network operators should stay informed of what these external entities learn, so that problems can be quickly remediated.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our research methodology. Section 4 presents summary characteristics for our data sets. Section 5 investigates the participating organizations in the observed traffic. Section 6 examines what selected external organizations know about the local edge network. Section 7 considers which local resources these external organizations control. Section 8 summarizes our results and future directions.

2. RELATED WORK

Internet security has received significant attention from the research community, because of its critical importance. We briefly consider research in four areas: *Surveillance*, *Reconnaissance*, *Botnet Characterization*, and *Intrusion Detection* (i.e., "real-time intelligence" for computer networks).

Surveillance: Surveillance is an intelligence gathering technique that obtains information through passive observation of activity. This method is commonly used on the Web, for example by behavioral tracking sites (e.g., `doubleclick.net`) or Web analytics sites (e.g., `google-analytics.com`) [Krishnamurthy and Wills 2009]. Krishnamurthy and Wills [2009] examine how third parties are able to track user actions across many Web sites, through the use of "hooks" like cookies and Javascript. Although our work focuses on knowledge obtained on organizations rather than users, their work clearly demonstrates how organizations can improve their understanding of targets of interest.

Reconnaissance: Reconnaissance comprises efforts to actively gather information of interest. In the computer systems and networks domain, *scanning* is an active measurement technique used to gather information about resources at a target organization. There is significantly more research on this topic than on surveillance. For example, Yegneswaran *et al.* [2003] provided an extensive characterization of four months of third-party scanning activity, by examining firewall logs from 1,600 organizations. Pang *et al.* [2004] monitored unused IP address space to characterize the sources and intent of "Internet background radiation". Jin *et al.* [2007; 2007] examined third-party activity on *gray space* (unassigned IP address space) to identify and categorize scanners. Allman *et al.* [2007] conducted a longitudinal study of third-party scanning, investigating the onset of scanning, scanning frequency, and scanned services over a 12-year period.

A key difference from these works is that we focus on understanding what the third party learns about the targeted environment. This is important as it helps prioritize the actions that the technical team needs to take in response to the scan. Also, these works focus on *reconnaissance*, while our work also considers information an external party could learn through *surveillance*. We do, however, leverage the heuristics defined by Allman *et al.* [2007]. Other related works to understanding and/or addressing *reconnaissance* traffic include scan identification techniques [Gates et al. 2006; Xu et al. 2008; Jung et al. 2004; Allman et al. 2007], visualization techniques [Muelder et al. 2005; Jin et al. 2009; Yin et al. 2004], and worm detection/mitigation techniques [Jung et al. 2007; Zou et al. 2005; Sommers et al. 2004; Weaver et al. 2004]. Our work complements these.

Botnets: Identifying resources under the control of external organizations is challenging, as the controlling party may try to conceal this fact. On the Internet, botnets (sets of compromised hosts) are a commodity desired by certain organizations. A considerable number of researchers have characterized botnets [Barford and Blodgett 2007; Barford and Yegneswaran 2006; Zhuang et al. 2008; Li et al. 2009], or developed techniques for identifying them [Collins et al. 2007; Karasaridis et al. 2007]. We leverage the observation that botnets are often used to send spam to identify hosts that are (potentially) under the control of an external organization.

Intrusion Detection: Network Intrusion Detection Systems (NIDS) are used

to monitor network activity and alert network administrators to potentially important events. A challenge is to accurately identify prioritized, actionable events from the large volumes of network activity. Shankar and Paxson [2003] proposed a technique called Active Mapping that helps reduce the number of false alarms. Katti *et al.* [2005] demonstrated the value of multiple collaborating NIDS to combat "common enemies". Duffield *et al.* [2009] used machine learning with flow signatures to detect malicious or unwanted traffic. Our work complements such studies, as by understanding what information is being leaked, we assist in assessing the significance of different events. This could be leveraged to help reduce the number of false alarms, by eliminating (or lowering the priority of) events that do not reveal sensitive information.

3. METHODOLOGY

3.1 Data Collection

We use three types of measurements collected from the University of Calgary's 400 Mbps full-duplex link to the Internet. Two of the data sets span a full year and the third covers nine months. All the measurements were collected simultaneously using a SunFire server with four quad-core CPUs, 32 GB memory, and 1 TB disk space. One of the monitor's gigabit Ethernet NICs receives a mirror of all the university's Internet traffic. The monitor rotates and compresses the logs for each data set described in Section 3.2 on a daily basis. The compressed logs are periodically moved to a secure archive for long-term storage.

We take the issues of privacy and security very seriously. To protect user privacy, we limit the types of data we record, restrict access to the recorded data, and do not conduct analyses to try and identify individual users. Regarding security, we share *actionable information* with the campus IT staff.

3.2 Data Sets

While recording full-packet traces to disk could make a lot of interesting and useful information available to us, it would be difficult to sustain indefinitely and would also pose significant privacy concerns. Therefore, we determined what data we could gather continuously (without ever recording full-packet traces to disk) that would enable us to answer the research questions at hand. We determined we needed three complementary types of data sets in our work: connection-level records, HTTP transaction records, and frame-level summaries. We next describe each of these data sets.

Connection: The data set that we study most extensively is a collection of connection summaries. We use the `conn` feature of the open-source Intrusion Detection System *Bro*[1] to collect these summaries. Each connection summary contains information such as the source and destination IP addresses and port numbers, the number of bytes transferred in each direction, and the "state" of the connection. A detailed description of the connection summaries is provided in the online Bro documentation.[2] This data set was collected from April 1, 2008 to March 31, 2009.

[1] `http://www.bro-ids.org/`
[2] `http://tinyurl.com/bro-conns`

HTTP: To supplement the connection data, we gathered summaries of Web transactions. We implemented this as a script in a separate Bro process. This script records information such as the URL, the `User-Agent`, and the presence (or absence) of certain HTTP headers. To preserve user privacy, we do not record the local IP address involved in the transaction, nor do we record any cookies. This data set was collected between July 1, 2008 and March 31, 2009.

Frame: For validation purposes, we used frame-level summaries that count the number of frames and bytes transferred in each direction broken down by network and transport-layer protocols. The functionality (implemented in C) is kept as simple as possible to minimize the overhead it places on the monitor. The data set is considered the "ground truth" (particularly for the amount of data transferred) and is used to validate other results. The counts were recorded for every one minute interval for the same one-year period as the connection data.

3.3 Scalability Challenges

While the one-year duration of our traces allows us to examine long-term behaviors, coping with the large volume of data is a challenge. To address this challenge, we apply best practices, such as developing analyses on small subsets of the data [Paxson 2004]. The bulk of our analyses were run on a server with four single-core AMD Opteron processors and 8 GB of memory. Each analysis used two processors, one to decompress the data and stream it to the analysis program on another processor. Many of the analyses use the same parser (written in C) to extract the fields of interest from each connection summary. Specialized functions are added as needed to perform specific analyses of interest.

Developing efficient and scalable analyses is important to us, as *real-time* intelligence is our long-term goal. We made a key design choice to focus initially on the activity of distinct /24 prefixes (i.e., the first three octets of an IPv4 address). As a result, most of our analyses use an array of 2^{24} data structures, one for each possible /24 prefix. The contents of the data structure vary by analysis. For example, since a /24 network can have at most 2^8 hosts, we use a bit vector of length 32 bytes (256 bits) to record the unique IPv4 addresses seen per prefix. Other variables keep a running count of the number of hosts and flows seen for the prefix. This approach provides a reasonable balance between state and time overheads. For example, the individual analyses we conducted on the year-long "connection" data set required 16–25 hours to complete.

3.4 Supplementary Data Sets

We also use several secondary sources to supplement our data sets. In particular, we needed a mapping between external organizations and their corresponding /24 prefixes. This information is available from the Regional Internet Registries (RIRs). We queried the RIRs (obeying rate limits) for the organization identifier (OrgID) for the most popular /24 prefixes observed (based on number of connections), to determine the external organizations. Unfortunately, some organizations have multiple OrgIDs, for reasons such as acquisition or internal policy [Krishnamurthy and Wills 2009]. We attempted several methods to discover the set of OrgIDs affiliated with an organization: using the organization lookup feature available in some RIRs; extracting the domain of the contact email for an OrgID and grouping OrgIDs with

the same contact domain; and exploiting regular patterns in the OrgID (e.g., organization name followed by a number).

4. SUMMARY CHARACTERISTICS

This section examines some high-level characteristics of the data sets and discusses several limitations. In subsequent sections, we examine specific characteristics in more detail.

Table I provides information from the Frame data set. Over one trillion frames and nearly 700 TB of data were transmitted over the network, averaging about 2 TB per day. Approximately one half of the frames were *inbound*; i.e., sent by computers on the Internet to destinations within the university network. Slightly more data (57%) was observed for inbound frames than for *outbound* frames (43%). This indicates (not surprisingly) that the university consumes more data from the Internet than it provides to others.

Table I. Information from Frame Data Set.

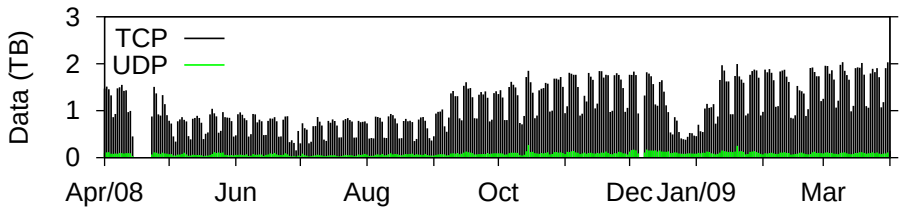
Description	Value
Total Frames Observed	1,173.0 billion
Total Data Observed	695.9 TB
Inbound Frames	50.2%
Inbound Data	56.6%
Outbound Frames	49.7%
Outbound Data	43.4%

Using the Frame data set, Figure 1 shows time series plots of the volume of data transferred on a daily basis into or out of the university. As noted earlier, a slightly greater volume of data is consumed; this can be seen in the higher daily volume of data in the inbound direction, which peaks near 2 TB per day in March 2009. As one might expect, the daily volume is higher on weekdays than on weekends. The daily volume reflects the university's annual calendar; volumes are highest when classes are in session[3], and lower otherwise. This characteristic affects the rate at which external entities learn information about the organization (as will be seen later). TCP is the dominant transport-layer protocol used on this network. For both the inbound and outbound directions, TCP transfers about 80% of all frames and 90% of the data bytes. UDP accounts for almost all of the remaining traffic.

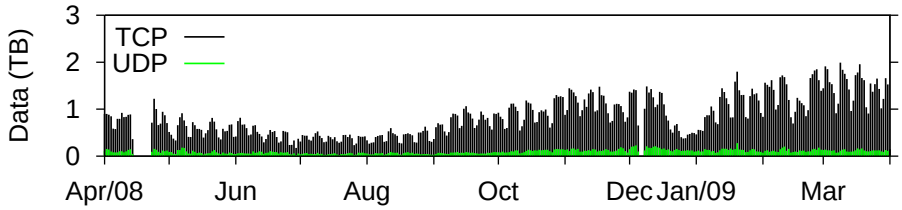
There are two periods of missing data in Figure 1. From April 14–21, 2008, our monitor was offline due to a hardware problem. On December 6th and 7th, 2008 (a weekend), no packets were forwarded to our monitor due to scheduled maintenance on the campus network infrastructure. Otherwise, our data set provides a complete picture of a year in the life of a campus network.

Table II summarizes the "Connections" data set. The year-long data set contains more than 39 billion connection summaries. Inbound connections are initiated by an external host and destined to a host on campus. Outbound connections are initiated

[3]Regular sessions span from September-December and January-April, with a five-day workweek from Monday through Friday, and a one-week study break in the middle of February.



(a) Inbound Direction



(b) Outbound Direction

Fig. 1. Data Volume by Transport Protocol.

by a host on campus and destined to a host on the Internet. The connections are roughly split evenly across these two categories (48% versus 52%).

Table II. Summary of Connections Data Set.

Description	Value
Data Set Size	850 GB (compressed)
Data Set Size	3.3 TB (uncompressed)
Connection Summaries	39.3 billion
Total Data Observed	1,620.3 TB
Inbound Connections	48.1%
Outbound Connections	51.9%

One problem with the connection summaries is that the total data volume reported is incorrect. This is due to the large percentage of TCP connections that are terminated with a RST (reset) packet rather than a FIN packet. The RST packet may not contain a valid sequence number, which can result in incorrect calculation of the number of bytes transferred. This reset behavior has existed on the Internet for some time [Arlitt and Williamson 2005]. Weaver *et al.* [2009] provide a thorough discussion of reset TCP connections. While we ignore blatantly incorrect size information (e.g., MB or GB transfers in less than one millisecond), the total data transfer estimate is still more than double the data transfer reported by the low-level data set. Thus in this work we focus primarily on flow volumes rather than data volumes to identify types of activity. For future studies we plan to alter the data collection to address this particular issue.

Table III provides a breakdown of the connection states for both the inbound and outbound directions. Following the work of Allman *et al.* [2007], we group

the connections according to their state (as determined by the Bro `conn` script).[4] "Good" connections account for just under half of all connections. About 40% of the inbound connections are considered "Bad", while 30% of outbound connections also fall in this category. The remaining connections are labeled as "Unknown". We use information on the mix of good, bad, and unknown connections in subsequent sections.

Table III. Summary of Connection States.

Type	State	Inbound (%)	Outbound (%)
Good	SF	43.42	39.99
	RSTO	2.63	4.55
	RSTR	0.43	1.81
	Total	46.48	46.35
Bad	S0	35.71	27.13
	RSTOS0	3.54	1.10
	REJ	0.99	2.40
	Total	40.24	30.63
Unknown		13.28	23.02

5. WHO ARE WE DEALING WITH?

5.1 Aggregating by Network Prefix

As an initial step towards answering our first question regarding *who* might be gathering intelligence about the university, we examine the diversity of external IPv4 addresses observed in the connection summaries. In addition, we also counted the distinct /24 prefixes observed. The motivation for considering the prefixes is to gain insight into the organizations involved in the communications. While IPv4 addresses can be allocated in blocks other than /24, we felt that /24 represented a reasonable starting point.[5] Once we understand the popular prefixes, we can determine the actual address blocks they belong to, and focus on the corresponding organizations.

Table IV shows the number of distinct IPv4 addresses and /24 prefixes observed in our connection summary data set. Considering only the inbound connections,

[4]The Bro `conn` script classifies connections into one of thirteen states (see `http://tinyurl.com/bro-conns`). The states shown in Table III corresponds to: normal establishment and termination (SF); connection established, originator aborted (i.e., a RST was sent by originator) (RSTO); established, responder aborted (RSTR); connection attempt seen, no reply (S0); originator sent a SYN followed by a RST, a SYN-ACK was not seen from the responder (RSTOS0); connection attempt rejected (REJ). For completeness, the remaining connections, which we do not further classify, but leave as "unknown", are: connection established and close attempt seen only from originator (S2); connection established and close attempt seen only from responder (S3); responder sent a SYN ACK followed by a RST, a SYN not seen from the (purported) originator (RSTRH); originator sent a SYN followed by a FIN, a SYN-ACK not seen from the responder (SH); responder sent a SYN ACK followed by a FIN, a SYN not seen from the originator (SHR); no SYN seen, just midstream traffic (OTH). For a detailed discussion of each of these states, see Allman *et al.* [2007].

[5]Of the roughly 700,000 /24 prefixes we resolved, only 1.82% were from address blocks with larger prefixes.

we observe nearly 300 million unique IPv4 addresses, with 3.2 million distinct /24 prefixes. This is roughly 7% of all possible (but not necessarily routable) IPv4 addresses, and 19% of the possible /24 prefixes.

For the outbound connections, the number of destination IP addresses is slightly larger at 324 million, but the diversity in terms of /24 prefixes increases substantially to 10.6 million, or 63% of the possible /24 prefixes. Throughout the measurement period, there is a lot of overlap in the external addresses observed from inbound and outbound connections. When all connections are considered together, 392 million distinct IPv4 addresses and 10.6 million unique /24 prefixes are seen (9.1% and 63.4% of the possibilities, respectively).

Table IV. Summary of External IPv4 Addresses.

Description	Unique IPv4 Addresses	Unique /24 Prefixes
Inbound Connections	298,367,959	3,152,020
Outbound Connections	324,174,907	10,577,626
All Connections	391,875,529	10,637,400

The left-hand side of Table V shows how the /24 prefixes were "discovered"; i.e., which protocol and state the connection had when a distinct /24 prefix was first encountered in the trace. For inbound connections, about two-thirds of the /24 prefixes are encountered for "good" connections (i.e., connections that exchanged data). UDP connections account for half of the discoveries. "Bad" connections (i.e., those typically associated with scanning, such as S0 or ICMP echo) discovered about 25% of the distinct /24 prefixes.

For outbound connections, the breakdown is quite different. "Good" connections discovered less than 10% of the 10.5M unique /24 prefixes. 90% of the discoveries were with "bad" connections; mostly using TCP (57%), but a lot with UDP (20%) or ICMP echo (13%).

Table V. Breakdown of External IPv4 /24 Prefix and External IPv4 Address Discovery.

Type	Protocol	External IPv4 /24 Prefixes		External IPv4 Addresses	
		Inbound (%)	Outbound (%)	Inbound (%)	Outbound (%)
Good (Successful)	TCP	18.4	2.9	10.9	4.6
	UDP	49.6	6.0	59.2	22.8
	Total	68.0	8.9	70.1	27.4
Bad (Exploratory)	TCP	6.5	57.1	5.1	20.0
	UDP	17.0	20.0	20.2	43.4
	ICMP	1.3	12.5	0.2	5.4
	Total	24.8	89.6	25.5	68.8
Unknown (Other)	TCP	2.2	0.8	1.6	1.9
	UDP	0.4	0.0	0.0	0.0
	ICMP	4.6	0.7	2.7	2.0
	Total	7.2	1.5	4.4	3.8

The right-hand side of Table V shows the breakdown of the discovery of IPv4 addresses. A notable difference from the discovery of /24 prefixes is "good" connections discover a greater percentage of the addresses in both the inbound and outbound directions. This indicates that the "bad" connections are often quite distributed across the IPv4 address space, rather than concentrated within small portions (e.g., a /24 prefix).

Since TCP is utilized for a much larger fraction of all connections than UDP, it is somewhat surprising to see UDP account for such a large portion of the /24 prefix and IPv4 address discoveries. One reason for this is the use of UDP for control communication by some popular P2P applications (e.g., eMule).

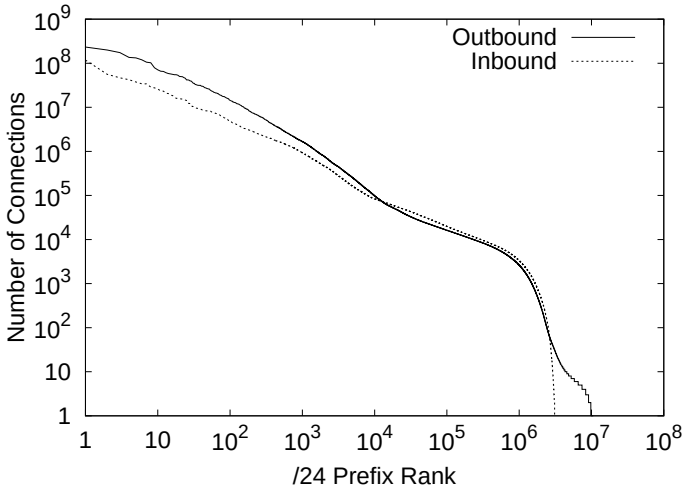


Fig. 2. Frequency of /24 Prefixes.

Figure 2 shows the frequency distribution of the inbound connections across the unique /24 prefixes. The distribution is clearly non-uniform, with some prefixes responsible for significantly more inbound connections than others. On this log-log plot, the graph appears roughly linear through the first one million prefixes. The distribution drops off sharply for the remaining two million prefixes. Many of the one million highest ranked prefixes involved "good" connections, while the "bad" connections are more commonly associated with prefixes with fewer connections.

Figure 2 also shows how the 20.4 billion outbound connections were distributed across the 10.6 million distinct /24 prefixes observed on outbound connections. The frequency distribution for outbound connections is more skewed than in the inbound direction, as local users use popular Internet-based services. The frequency distributions are quite similar from prefixes 10,000 through one million. The tails of the distributions differ noticeably. The outbound connections are addressed to an additional 7.4 million distinct /24 prefixes. Figure 2 shows that most of these prefixes receive only a handful of connections from the university. Again, many of

the connections associated with prefixes in this "unpopular" region were classified as "bad" connections.

The results in this section demonstrate one of the first challenges in determining who we are dealing with - the local organization communicates with a large and diverse group. With many connections, it is also unclear if communication with an organization actually occurred. In the subsequent sections, we take additional steps towards answering our guiding questions.

5.2 Identifying Organizations

Using the statistics by /24 prefix, we match each prefix to an OrgID, as described in Section 3. To reduce the overhead on the RIRs, we only issued queries for the most popular prefixes, namely those cumulatively responsible for 75% of all connections.

Table VI lists the ten prefixes responsible for the most inbound connections. The table also lists the percentage of connections that are classified as "bad" and the number of hosts (out of at most 256 in the /24 subnet) that are contacting machines on campus. The top /24 prefix (CHINANET-AH) belongs to China Telecom; it generated 118 million incoming connections (over 300,000 per day). The states of these connections are almost exclusively "bad", using the definitions of Allman *et al.* [2007]. The next four most popular /24 prefixes belong to ISPs in four different countries: RCMS (RoadRunner) in the USA, ISPSYSTEM in Russia, VE-DEMA (Desca.com) in Venezuela, and LUNA-DSL (luna.nl) in the Netherlands. Each of these organizations has a number of users interested in selected services at the university (e.g., Web), even if they are not aware they are visiting the university (e.g., for P2P). These /24 prefixes are an example of how rankings can change as data is aggregated; even though these OrgIDs are in the top 10 when ranked by a single /24 prefix with the most inbound connections, when we aggregate all of the /24 prefixes associated with an OrgID, these organizations drop from the top of the list, behind larger organizations.

Three of the top ten /24 prefixes belong to popular Internet companies (Google, Microsoft, and Yahoo!). These inbound connections are primarily the activity of the crawlers associated with each company's search engine. Two interesting observations about the activity of these prefixes are that they generate a moderate number of "bad" connections (7–19%), and that they are distributed across more hosts (25-84% coverage of 256 possible hosts) than any of the other prefixes (less than 14% host coverage).

The remaining two /24 prefixes in the top ten belong to local companies. Telus (TACE) provides commercial and residential communications services, including Internet access. FCL-13 is a small local company. The high percentage of "bad" connections from the Telus prefix come from 13 residential computers scanning the university network.

While Table VI provides high-level statistics for the top ten prefixes for incoming connections, we also computed other measures for these connections, including byte counts, port counts, and fan-out ratios. These metrics provide additional dimensions to characterize the different external organizations. Rather than claiming that some measures better characterize the differences between organizations than others, we used Principal Component Analysis (PCA) together with clustering to

Table VI. Top 10 /24 Prefixes for Inbound Connections.

Rank	Inbound				
	OrgID	Country	Connections	% Bad	Hosts
1	CHINANET-AH	China	118,036,837	99.995	35
2	RCMS	USA	56,077,430	0.106	24
3	ISPSYSTEM	Russia	46,649,041	15.347	9
4	VE-DEMA	Venezuela	42,282,187	0.345	28
5	LUNA-DSL	Netherlands	37,429,823	0.004	33
6	GOGL	USA	34,135,193	7.299	216
7	MSFT	USA	34,017,285	18.976	200
8	TACE	Canada	28,781,112	74.452	13
9	YHOO	USA	27,589,471	4.051	65
10	FCL-13	Canada	26,016,295	10.511	6

Table VII. Top 10 /24 Prefixes for Outbound Connections.

Rank	Outbound			
	OrgID	Country	Connections	Hosts
1	THEFA-3	USA	233,151,450	105
2	LLNW	USA	199,406,030	182
3	TACE	Canada	173,986,912	34
4	GOGL	USA	135,869,335	60
5	GOGL	USA	132,321,311	48
6	GOGL	USA	119,603,516	56
7	AKAMA-3	USA	105,659,863	82
8	C01342375	USA	102,145,390	21
9	GOGL	USA	78,934,636	53
10	EBSCO-1	USA	70,677,068	18

visualize the relationship between the top 50 prefixes observed on campus.[6]

Figure 3 shows the results of a cluster analysis of the output of a two-dimensional PCA of the top 50 /24 prefixes based on incoming connections. Here, each dimension consists of a linear combination of nine different measures and the weights are selected such that the two dimensions account for as much of the variability in our data as possible. This approach provides a visualization of how similar the different /24 prefixes are to each other.

Several distinct clusters are apparent in Figure 3. In particular, prefixes belonging to the popular service providers Google, Microsoft, and Yahoo! (labeled as "Providers") form an exclusive cluster in the top right corner. On the other end of the spectrum, ChinaNet (marked with a '1') and a cluster with three ill-behaved (and black-listed) subnets from Asia (in the lower left corner) are located relatively

[6]This part of the analysis was done using SPSS. For the clustering, we used hierarchical clustering, but only show example results with eight "clusters" (including some singletons).

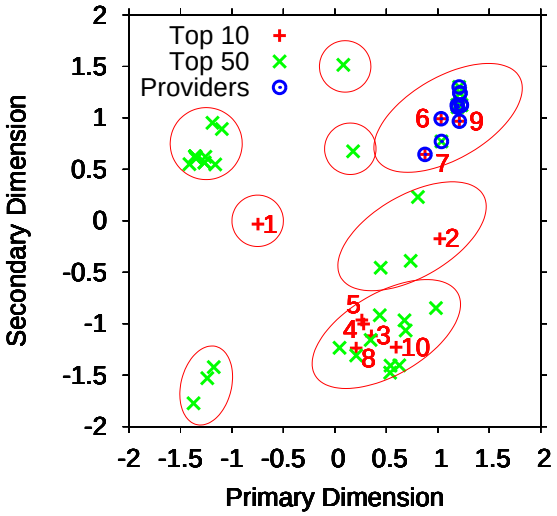


Fig. 3. Clustering /24 prefixes using incoming connections.

far to the left.[7]

We apply the same methodology to determine the top ten destination /24 prefixes for outbound connections. Table VII provides the results. Five of these prefixes belong to organizations that provide globally popular services. The most popular prefix belongs to Facebook (THEFA-3), while four others are assigned to Google (GOGL). Two additional /24 prefixes in this list are related to content delivery - Limelight Networks (LLNW) and Akamai (AKAMA-3). The third most popular prefix belongs to Telus (TACE), a local ISP.

5.3 Organizational View

Next, we aggregate the /24 prefix statistics by OrgID. As stated earlier, some organizations have more than one OrgID [Krishnamurthy and Wills 2009]. Aggregating OrgIDs into distinct organizations provides a clearer picture of external entities.

Table VIII shows the ten OrgIDs responsible for the most inbound and outbound connections. Inbound connections have a mix of residential network providers (SHAWC and TACE), search providers (GOGL, INKT (Yahoo!)), and several foreign providers (e.g., TurkTelekom, ChinaNet, Hinet, and TPNET.PL (NEOSTRADA)). For outbound traffic, the top OrgIDs tend to be popular Web-based service providers (GOGL, MSFT, THEFA-3, YHOO), content delivery providers (LLNW, AKAMAI), and providers of various other services (LVLT, VGRS, DOUBLE-3).

After completing the aforementioned analyses, we selected several organizations to represent the different types of external entities observed in the traces. For simplicity, we use three types: *first-party providers*, *third-party providers*, and *In-*

[7]When interpreting these results, we note that the primary component is dominated by measures that capture the ratio between good and bad connections, good and bad fanout, and the number of connections contacted using bad connections (with negative values being strongly correlated with subnets being more "bad"). The secondary component, on the other hand, is dominated by the number of hosts (with positive values being strongly correlated with larger number of hosts).

Table VIII. Top 10 OrgIDs, Inbound/Outbound.

Rank	Inbound		Outbound	
	OrgID	Conns (10^6)	OrgID	Conns (10^6)
1	SHAWC	2,558	GOGL	1,082
2	TACE	1,069	MSFT	456
3	TurkTelekom	392	THEFA-3	321
4	NEOSTRADA-ADSL	240	YHOO	286
5	GOGL	153	LLNW	280
6	CHINANET-AH	141	TACE	280
7	AR-APGO-LACNIC	127	LVLT	234
8	HINET-NET	104	VGRS	181
9	INKT	92	DOUBLE-3	180
10	CHINANET-GD	81	AKAMAI	163

ternet Service Providers (ISPs). The first two types were used by Krishnamurthy and Wills in their study of privacy diffusion on the Web [Krishnamurthy and Wills 2009]. First-party providers are those that provide services directly visited by users, such as search engines (e.g., `www.google.com`) and social networking sites (e.g., `www.facebook.com`). Third-party providers are indirectly visited by users when they visit first-party providers. Examples of third-party providers include Content Delivery Networks (CDNs) like Akamai or Limelight. ISPs provide Internet connectivity to users at other edge networks.

The main criteria we used to select external organizations to represent each type was *popularity* (based on the volume of connections). The non-uniform distribution of traffic across organizations presents an opportunity to learn what information about an edge network is leaked by examining only a few external organizations. As our results in this section will show, an organization's popularity strongly influences how much it knows about the local organization.

The set of first-party providers that we selected consists of four popular Web-based service providers (Facebook, Google, Microsoft, Yahoo!). We selected three third-party providers: a CDN (Limelight Networks), a global registry service (Verisign) and an Infrastructure-as-a-Service (IaaS) provider (Amazon). While Table VIII shows that Akamai is another popular CDN, we omit them from our set of third-party providers as our network monitor does not see traffic between university computers and the Akamai edge nodes located on the campus network. Even though Amazon is also a first-party provider via their online store, we believe their IaaS business is responsible for much of their on-campus traffic. Thus we consider them a third-party provider for our study.

In the set of ISPs, we included two local providers (Telus, Shaw) and four foreign providers that appear in the top ten list for inbound connections (ChinaNet, Hinet, TPNET.PL, and TurkTelekom). At the time of our study, these four foreign ISPs were all on the Composite Blocking List[8], indicating that other organizations have reported "bad" traffic originating from them.

Table IX provides summary statistics on the selected external organizations. Ta-

[8] `http://cbl.abuseat.org/domain.html`

ble IX reveals distinctive commonalities within groups, and differences between them. For example, the first-party providers used relatively few distinct prefixes (tens to hundreds), and popular first-party providers tend to have a higher proportion of outbound connections than inbound ones. The first-party providers that offer search services (Google, Microsoft, and Yahoo!) have a noticeably larger fraction of inbound connections than the social networking provider (Facebook). This is primarily due to the search services scanning for Web servers, retrieving documents for indexing, or gathering consistency information for cached documents from hundreds of different Web sites on campus. These three first-party providers also offer email service, and thus also have inbound connections to deliver messages to email servers on campus. They also have inbound connections to the campus DNS servers, to learn about the hosts using their services. For first-party providers like Google, Microsoft, and Yahoo!, this assortment of tasks can result in hundreds of thousands of inbound connections per day.

The third-party providers have characteristics similar to the first-party providers. Since the third-party providers do not provide Web search functionality, they have very few inbound connections (i.e., no crawling activity). It is interesting to note that Limelight has almost as much outbound traffic as Facebook, even though most Internet users are likely only familiar with the latter. Also, more /24 prefixes were seen for Amazon than either of the other third-party providers, as well as more than Google and Facebook. One possible explanation for this is the traffic is for Amazon's IaaS infrastructure, rather than its online store. This is one reason why we included Amazon as a third-party rather than first-party provider.

The ISPs exhibit quite different characteristics. For example, ISPs have hosts on thousands of /24 prefixes, rather than tens or hundreds. They are also responsible for proportionally more inbound connections than outbound connections. The local ISP Shaw seemingly has a disproportionately large fraction of the inbound connections (almost 15%). This is due to students, faculty, staff and other local people accessing university resources from their homes.

For the remainder of the paper, we focus on the organizations in Table IX.

Table IX. First-Party, Third-Party, and ISP Statistics.

Class	Organization	Prefix (/24)	Connections	
			In (%)	Out (%)
First-Party Providers	Google	263	0.56	7.44
	Microsoft	398	0.42	3.20
	Yahoo!	630	0.45	2.71
	Facebook	19	0.04	1.90
Third-Party Providers	Amazon	365	0.02	0.26
	Limelight	114	0.06	1.73
	Verisign	29	0.00	1.86
ISPs	ChinaNet	20,022	2.78	1.10
	TurkTelekom	13,309	2.24	0.09
	Hinet	15,197	0.84	0.58
	Shaw	7,046	14.94	0.06
	Telus	7,844	4.89	1.67
	TPNET.PL	6,907	1.40	0.15

6. WHAT DO OTHERS KNOW?

In this section, we search for answers to the next two guiding questions. We begin with a discussion of how others are gathering "intelligence" on the local organization in Section 6.1. We then describe how we filter Distributed Denial of Service (DDoS) incidents from the data, to more accurately interpret what others know. In Section 6.3, we explain what others (potentially) know about the local organization.

6.1 How Others Gather Intelligence

As stated in Section 2, there are two general methods external organizations can use to gather information on a local organization: surveillance and reconnaissance.

Active measurements (e.g., scans) typically receive the most attention. We refer to this type of intelligence gathering as *reconnaissance*. Some reconnaissance may occur as part of normal operations; e.g., contacting a DNS server to determine the IP address of the target organization's SMTP server. Other reconnaissance may be more difficult to assess, as an entity may intentionally try to hide the fact that any sort of intelligence gathering is occurring. For example, an entity may use a botnet (i.e., a set of compromised hosts from other organizations) to make it difficult for the targeted organization to recognize the source or intent of the reconnaissance.

External organizations can also gather intelligence on a target through *surveillance*; i.e., by passively monitoring activity between the two organizations. In particular, external organizations can passively gather data by examining the target organization's use of the external organization's (Internet-based) services. This technique has received relatively little attention in the research community. Krishnamurthy and Wills [2009] recognized that some organizations use surveillance, and investigated the implications on user privacy. We consider what information about the IT infrastructure is leaked.

6.2 Filtering DDoS Incidents

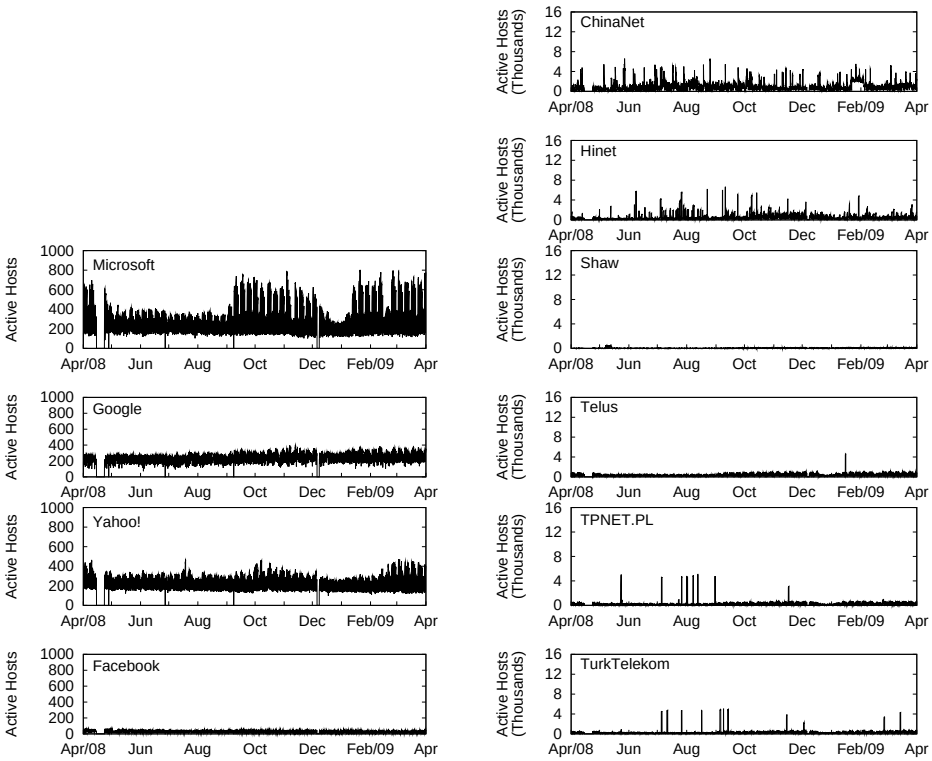
While quantifying what an external organization learns through surveillance, we realized that it is important to establish the *identity* of a host before it can be considered active. For TCP, this means observing a complete SYN handshake.

This requirement was not anticipated, since we knew the university's network used best practices like egress filtering to reduce Distributed Denial of Service (DDoS) attacks [Specht and Lee 2004] initiated from hosts on campus. However, while testing our *surveillance* analysis, we realized that some compromised machines on campus were launching DDoS attacks that only used source addresses within the university's assigned IPv4 address space. This resulted in all 2^{16} possible local IPv4 addresses being observed, even though many are not in use. To filter such traffic from the trace, we only consider (outbound) connections for which the identity of the source has been verified. This may result in fewer active hosts being discovered in our surveillance analysis, but it provides more realistic results than considering all observed connections.

6.3 What Intelligence is Gathered

6.3.1 *IT Infrastructure Reconnaissance.* Since networks are dynamic (e.g., many hosts are online only part of a day, week, or year), we decided to look at how many

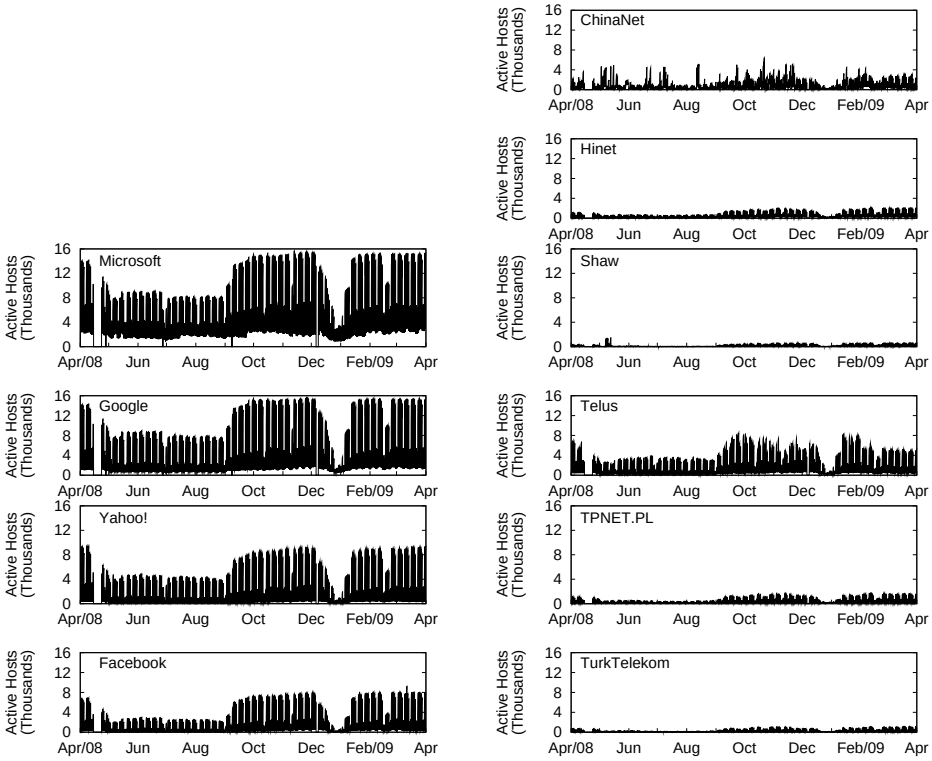
active (i.e., in use) hosts each external organization observes on a daily basis. While an organization could use longer-term information, our analysis helps indicate how often the information is "refreshed" by the external organization.



(a) First-Party Provider Reconnaissance (b) ISP Reconnaissance

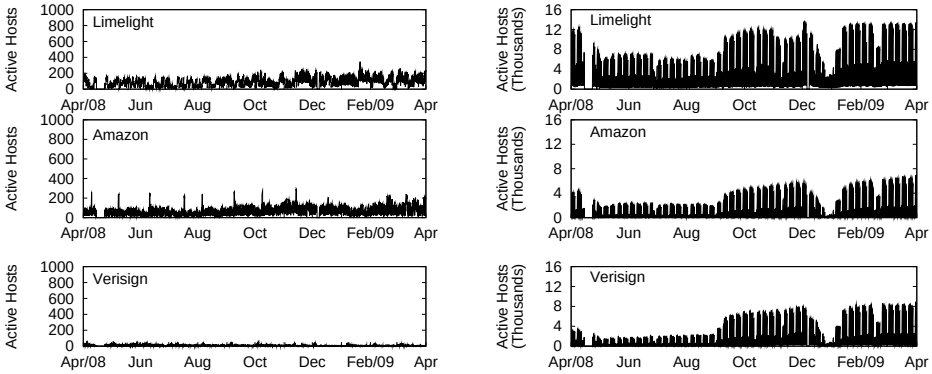
Fig. 4. First-Party Provider and ISP Knowledge of Active Hosts from Reconnaissance.

Figure 4(a) shows a time series plot illustrating the information the selected first-party providers obtain from reconnaissance. This graph reveals several interesting observations and insights. First, even with fewer connections than Google and slightly less than Yahoo! (Table IX), Microsoft is always aware of more active hosts than either Google or Yahoo!. Microsoft's knowledge varies a lot over the course of the year, reflecting the changes in the user population that correspond to changes in the academic calendar (and hence the number of computers in use). This is due to the prevalence of the Microsoft Windows operating systems on university computers. Google and Yahoo! obtain relatively constant information via reconnaissance throughout the year. Microsoft's peak discovery via reconnaissance is 803 active hosts, compared to 400 for Google and 473 for Yahoo!. The main reason for the difference is that Microsoft scans more "servers" running on student computers, resulting in the greater variation in active hosts over the course of the year.



(a) First-Party Provider Surveillance (b) ISP Surveillance

Fig. 5. First-Party Provider and ISP Knowledge of Active Hosts from Surveillance.



(a) Third-Party Provider Reconnaissance (b) Third-Party Surveillance

Fig. 6. Third-Party Provider Knowledge of Active Hosts.

Second, first-party providers that offer Internet search services (e.g., Google, Microsoft, Yahoo!) consistently learn more about the local IT environment than those who do not offer such a service (e.g., Facebook). In normal operation, these latter first-party providers primarily discover DNS (and occasionally SMTP) servers, rather than many hosts. Google, Microsoft, and Yahoo! also contact DNS and SMTP servers in the course of providing various services.

Figure 4(b) shows the reconnaissance activities of the ISPs. This activity is quite different from the first-party providers. First, the ISPs are the source of much more aggressive reconnaissance, and thus learn of more active hosts (e.g., ChinaNet discovered a peak of 6,606 active hosts in one day). However, the reconnaissance activities of ISPs are much burstier than that of first or third-party providers. Hosts from ChinaNet issued on average more than 300,000 connection attempts per day; many of these did not discover anything. However, this does not mean that the targeted IP address is not in use, as hosts with personal firewalls may choose not to respond to certain connection attempts even if a port is open to select hosts.

The reconnaissance activity of the third-party providers (Figure 6(a)) like Limelight and Amazon is much like that of Facebook. That is, they discovered relatively few active hosts, and those that were discovered tended to be DNS or SMTP servers.

6.3.2 *IT Infrastructure Surveillance.* Figure 5(a) shows the number of active hosts that first-party providers are aware of through surveillance. An initial insight from this figure is that first-party providers know a lot more about the number of active hosts through surveillance than through reconnaissance. For example, Microsoft discovered up to 15,633 active hosts in a single day through surveillance, compared to 803 through reconnaissance. It is also important to note that the popularity of an organization matters; using surveillance, both Microsoft and Google learn about the same number of active hosts. Yahoo! appears to be less popular with users on campus, and thus knows about 9,531 active hosts on its peak day; Facebook is aware of a similar number, seeing a peak of 9,234 active hosts.

Figure 5(b) shows the number of active hosts that ISPs may know from surveillance. The numbers are substantially less than those of popular first-party providers. There is, however, more consistent behavior seen in Figure 5(b) than there was in Figure 4(b). For example, the outbound traffic to ChinaNet includes (legitimate) HTTP transactions to the Sina.com web portal, "a leading online media company ... for China and the global Chinese communities."[9] Similarly, Telus hosts numerous Web sites for local companies and organizations. The graph of Telus surveillance indicates that some of these are of interest to university users.

Figure 6(b) shows the knowledge of active hosts that third-party providers could gain via surveillance. The CDN Limelight discovered up to 13,902 active hosts per day, more than some popular First-Party Providers like Yahoo! and Facebook. This occurs because Limelight delivers content for multiple popular first-party providers. The graph of active hosts discovered via surveillance in Amazon traffic shows noticeable growth between September 2008 and April 2009, compared to other providers. This could be due to expansion of Amazon's IaaS business.

[9] `http://corp.sina.com.cn/eng/`

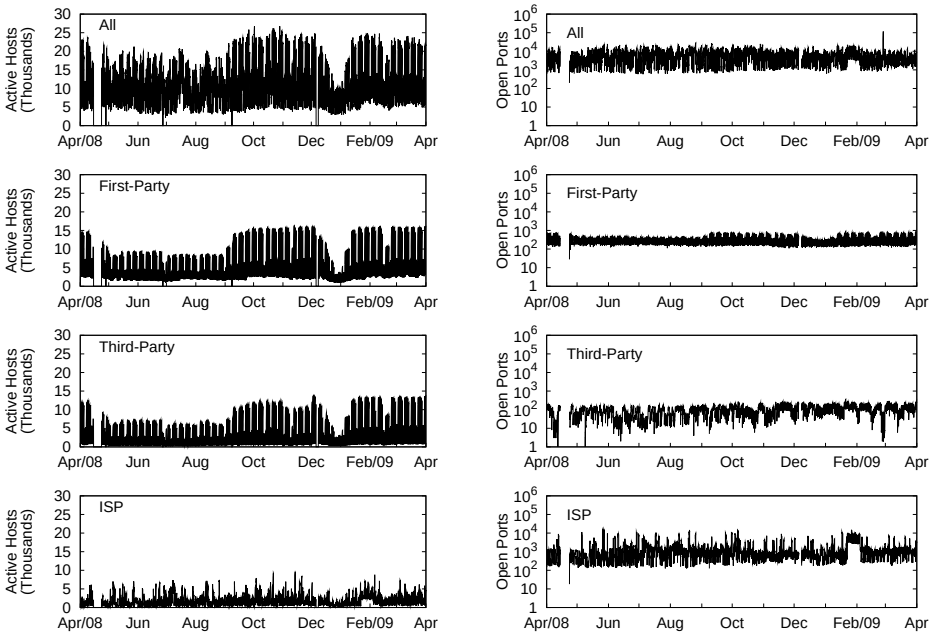
6.3.3 *Combined Knowledge of IT Infrastructure.* Figure 7(a) shows the knowledge that a first-party provider, a third-party provider, and an ISP have about the local organization. In this case, we compare the combined knowledge from reconnaissance and surveillance against the total number of local hosts engaged in TCP connections that successfully completed the SYN handshake (all). To keep the graph legible, we only include one organization from each category: Microsoft, Limelight and ChinaNet.

Figure 7(a) reveals several important insights. First, popular first-party providers consistently know more than providers in the other groups. Second, popular third-party providers can consistently know more than ISPs. This is particularly important to note. While popular Web sites have traditionally partnered with established companies (e.g., Limelight) to help scale their services, popular Web 2.0 sites like Facebook allow any third party to expose their services to a broad audience. This means that an edge network could leak information to an even broader audience in the future. Third, even the most popular first-party providers do not see all hosts on a given day (note the difference between the "All" and "First-Party" plots).

An open question is whether another (still unidentified) external organization knows more about this edge network than any of the entities considered in our study. To answer this, we examine the information that remains in the trace if we ignore the traffic involving our set of first-party providers, third-party providers and ISPs. The results of this analysis indicate that the remaining reconnaissance traffic discovers a maximum of 7,200 active hosts in a single day. Considering the remaining surveillance traffic, a peak of 19,981 active hosts are discovered in a single day. This means that the set of "other" (unidentified) organizations can at best know only slightly more about the active hosts on campus than the popular first-party providers like Google and Microsoft. Only a collaboration of multiple entities could compose a more complete picture of the active host behavior.

6.3.4 *Knowledge of Open Ports.* While we have shown that surveillance can provide an external organization with information about the active hosts on a remote network, it does not provide all information that might be of interest. For example, knowledge of specific open ports that are vulnerable to known problems could be exploited. Reconnaissance is a much more effective technique than surveillance for obtaining this information (because surveillance typically only sees a limited set of ports (e.g., HTTP, SMTP, DNS, etc.).

Figure 7(b) shows that ISPs see an order of magnitude more open ports than first or third-party providers. On a typical day, more than 10,000 open ports were discovered by various parties. The largest discovery occurred on February 26, when (in a 10 hour span) a single external host scanned 109 different protocol/port pairs on 45,848 addresses in the university's address space. About 100,000 open ports were discovered; we alerted IT to the most serious cases. This is an example of how knowing what others know about your organization can help improve the management of local IT resources. While it may be difficult to stop the leakage of the types of information we have discussed, proactively addressing issues could prevent others from exploiting them.



(a) Host Information (b) Port Information

Fig. 7. What External Organizations About Active Hosts and Open Ports.

6.3.5 *Discussion of Header-specific Surveillance.* In addition to greater knowledge of the IT infrastructure, surveillance may also provide the external organization with additional (richer) information. For example, HTTP headers can provide a lot of information about an organization's IT resources (and/or its individual users). The remainder of this section discusses the types of information leaked via HTTP headers. For security reasons, we chose not to release the details of this analysis; we only share some of the implications.

Operating systems and versions: Some external entities target specific weaknesses in operating systems, browsers, or Web-enabled software to gain control of machines. Using the `User-Agent` field, we confirmed that it can reveal the operating system on the user's machine (e.g., Windows), the version (e.g., Vista), and often patch levels. Clearly, this information could be used to compromise a machine. It also reveals which OS versions are most prevalent on the edge network, which could be used to determine which exploits the organization is most vulnerable to.

Applications and browsers: The `User-Agent` field also reveals the Web browser in use (e.g., Internet Explorer, Firefox) and the version. Aggregating this information over hosts indicates the adoption rates and other longitudinal trends. Such first-hand knowledge is much more valuable than (potentially) out-of-date public statistics (e.g., obtained from blogs). Self-monitoring of the `User-Agent` field could be used to resolve browser vulnerabilities, and to learn about other Web-enabled applications that could potentially provide a means of compromising the host.

Organizational interests: The `Referer` field provides the URL of the Web page a user traversed to reach the current page (and/or service provider). This

field can be used to infer information about the organization's interests, and the user's browsing habits.

To demonstrate this, we performed the following analysis. First, for each transaction in our HTTP data set, we extracted the *brand* of the visited Web site (e.g., the brand of `www.google.com`, `mail.google.com` and `www.google.co.uk` is `google`; in other words, we strip off the top level domain(s) information, and then use the next term as the brand).[10] We then ranked the brands by number of transactions. Next, we extracted the brands from the URLs contained in `Referer` fields of HTTP transactions that visited Google, Microsoft or Yahoo!. This set of extracted brands represents what these first-party providers learned (via surveillance) about traffic to other sites.

Figure 8 shows the results of this analysis. For each organization, a logarithmically binned weighted average of the rank is plotted on the x-axis. The actual rank (i.e., popularity of brands across all HTTP transactions) is shown on the y-axis. The results indicate that all three of these first-party providers have a reasonably accurate view of the brands used most by the university. They could also use other information in the `Referer` URL to enhance understanding of the target organization's interests.

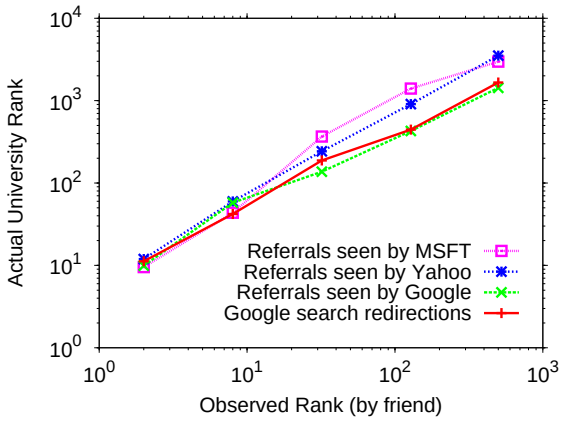


Fig. 8. Referral Information Leakage.

6.3.6 *Discussion of Web-services-based Information Leakage.* External organizations can also learn a lot about an edge network (and their users) through the use of their services. For example, in addition to the (private) information individual users may provide a social networking site, such as `Facebook`, we note that the combined leakage of many users may reveal much more about the organization (such as its structure, who is employed, who reports to whom, etc.). There are also many other services, such as email and other password protected services, that may reveal sensitive organizational information. This may become an even greater issue as organizations are outsourcing their services. As an example, we note that some

[10]Most host names follow this convention; however, exceptions do exist (e.g., `del.icio.us`).

organizations (including universities) have already started outsourcing their email and other collaboration services to Google.[11] From an organizational standpoint, adopting other (emerging) cost-effective services such as Amazon's IaaS may result in further information leaks.

Finally, as an example, we examined what Google learns from the search results it provides to local users (ignoring what Google may learn from the search query strings themselves). When a user clicks on a link in a page of search results, the user actually visits Google first and then gets redirected. Figure 8 also shows the results of this analysis. Although there are fewer search selections than `Referer` URLs for Google, the "Google search redirections" provide Google with a similar picture of the local organization's interests.

7. WHAT DO OTHERS CONTROL?

In this section, we tackle our final guiding question, and shed light on how much control external entities have over resources of the local organization. We focus on IP addresses (hosts) as the resource being controlled. As noted by Xie *et al.* [2009], "security rests on host accountability." Establishing the number of hosts directly controlled by external entities is a difficult endeavor, as they wish to hide the fact that they have compromised a machine. As such, our estimates should be considered lower bounds.

7.1 Direct Control of an Edge Network's Resources

Compromised machines that are controlled by an external entity may for example participate in a botnet that launches denial of service attacks, delivers spam, scans other computers for vulnerabilities, or attempts to exploit known vulnerabilities. As an initial step towards understanding how many machines are being directly controlled, we first look for significant spikes in the discovery of "new" (i.e., not previously seen in the data set) /24 prefixes observed on outbound connections.

Figure 9 shows the number of "new" prefixes observed daily. The largest spike occurs on the first day, since we have no prior knowledge of the "working set" of regularly visited external destinations. Most of the fluctuations in the graph reflect the general use of the network (e.g., higher activity during the main academic year).

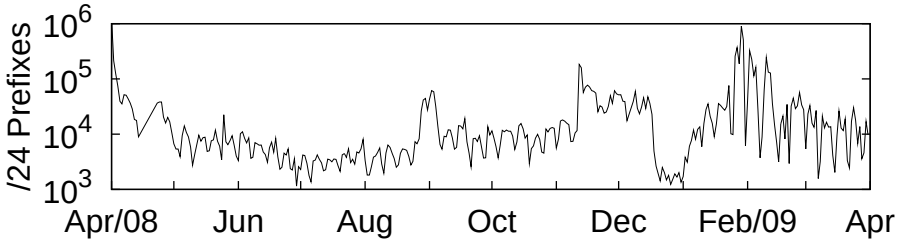


Fig. 9. New /24 Prefixes Observed Per Day.

[11] `http://www.google.com/a/help/int/en/edu/customers.html`

Figure 9 also shows several periods of increased discovery: a few days at the end of August, a sustained increase from mid-November to mid-December, and significantly increased activity from January 25 to February 12. We refer to this latter range as the "peak discovery" period. We now examine characteristics of it compared to the overall trace.

Table X provides some high-level characteristics of the discovery of /24 prefixes, for both the entire one-year trace as well as for the 18 day-long "peak period". Although the peak (discovery) period accounts for only 5% of the total trace duration, 35% of all distinct /24 prefixes were observed during that time. The bulk of the /24 prefixes were discovered by hosts on only a few local subnets. Hosts on the wireless subnet (WLAN) discovered about half of all distinct external prefixes. The student residence subnet is the next largest, discovering about one third of all unique /24 prefixes observed over the course of the year. The hosts on these two subnets are self-administered. The Student Union (SU) accounts for approximately 5% of all prefix discoveries. The computer science subnet (CPSC) is approximately the same size as the WLAN and Residence subnets. By comparison, the hosts on this subnet discover relatively few new /24 prefixes. Several notable differences between these groups are: (1) different user populations; and (2) CPSC machines are administered by IT staff. The peak period reveals that the discovery is done almost exclusively by hosts on the wireless subnet.

Table X. Discovery Characteristics

Characteristic	Label	Overall (%)	Peak (%)
Origin Subnet	WLAN	49.4	89.6
	Residences	32.4	7.9
	SU	4.8	1.7
	CPSC	0.9	0.0
	Others	12.5	0.8
Protocol	TCP	60.8	98.4
	UDP	26.0	1.3
	ICMP	13.2	0.3
Top Ports	445	36.3	92.4
	(ICMP Echo)	12.5	0.1
Top States	S0	76.0	98.9
	OTH	13.7	0.1
	SF	7.3	0.1
	REJ	0.1	0.2
	Others	2.9	0.8

If we consider the protocol in use when a distinct /24 prefix is discovered, we see a significant deviation from the overall distribution of packets or data traffic. Table X shows that while TCP is still used for a majority of the discoveries (60%), UDP (26%) and ICMP (13%)[12] are quite common as well. During the peak period, TCP was the dominant protocol used (98%), suggesting UDP and ICMP are more commonly used in long-lived discovery. The discovery during the peak period

[12] 95% of the ICMP traffic are Echo responses.

appears homogeneous; almost all of the discovery occurred on TCP port 445 (used for sharing services on Windows). Most of these connections end in an S0 state, indicating that no response was received from the destination. Our hypothesis is this activity was a worm (Conficker/Downadup) attempting to propagate. Relatively few local hosts were involved, and they aggressively attempted to contact hosts at other locations.

The bulk of the discoveries are done by a few hosts. Over the course of the year, the top two "discoverers" found a combined 17% of all observed distinct /24 prefixes. The top discoverer found 1.1 million /24 prefixes in two days during the "peak period". All of these were destined to TCP port 445. 99.5% received no response; almost all others received a REJ response (but indicating that the host is online). The second top discoverer was very active in November and December. This host used ICMP predominantly (96% of its discoveries), and discovered at much lower rates than the previously discussed host (tens of thousands per day rather than hundreds of thousands).

External entities who seek direct control of Internet hosts will actively search for opportunities to take control. Compromised hosts can then be used for numerous purposes, including DDoS attacks or sending spam [Staniford et al. 2002]. To search for potential bots on the edge network, we examine SMTP traffic - the exchange of email. To minimize false positives, we focus on successful SMTP transactions (`TCP` port 25) with known email servers. Over the year-long period, we observed successful transactions from about 3,000 local addresses to SMTP servers[13] at Google, Microsoft and Yahoo!. 745 local hosts successfully communicated with SMTP servers at Google, Microsoft or Yahoo!. 71% of these local addresses were from the WLAN and Residence subnets (i.e., self-administered machines), and are not authorized mail servers.

We performed a simple test to determine if the WLAN and Residence hosts sending SMTP messages to the selected first-party providers were trying to propagate spam. The Simple Mail Transfer Protocol specifies that client SMTP hosts should send a HELO (or EHLO) message to identify itself when the transmission channel is established [Postel 1982; Klensin 2001]. We used our network monitor to extract (in an online manner) the HELO/EHLO messages from several outbound SMTP connections from WLAN or Residence subnets. None of the observed HELO/EHLO messages contained the name or IP address of a host on the WLAN/Residence subnets. In fact, most of the observed messages included an unroutable IP address. Since the hosts were not revealing their actual identities to the SMTP servers, we conclude that they are most likely attempting to propagate spam.

7.2 Indirect Control of an Edge Network's Resources

Based on the data we examined, we argue that popular first-party providers (indirectly) control more hosts than external entities who seek direct control. By providing interesting or useful functionalities, first-party providers can build a loyal audience. For example, thousands of machines on campus communicate with Mi-

[13] Pathak *et al.* [2009] observed that 90% of spam is destined to Yahoo! Mail, Google's Gmail, Microsoft's Hotmail, and Hinet. Thus we believe it is sufficient to consider SMTP connections destined to three of our first-party providers.

crosoft and Google every day. Many of these users use a particular service by habit, become personally *invested*, and are therefore more likely to be loyal to the service. These potentially large captive audiences represent a strong control point that can be leveraged for advertisement opportunities, marketing of new services, and even used to shape their perception and interests. As an example, Google recently made the following (hyperlinked) statement on their main search page: "Upgrade to Google Chrome, a faster way to browse the web. Install now." This represents a powerful position, one that the edge network operators should be mindful of.

Search engines are another powerful control point. Not only do they have the ability to redirect hosts to destinations of their own choosing, they can do so after they know the OS and browser of the user.

A third control point is online social networking platforms. First-party providers like Facebook realize this, and are opening their platforms to others, which may cause their captive audiences to become more available to third parties. For example, on April 27, 2009 Facebook announced the Open Stream API.[14] This allows any developer to access the stream of information within Facebook. Using this stream, developers can create new interfaces for it. While many positive things may come of this, one problem we foresee is the development of "impostor" sites that look and feel just like Facebook, created to compromise user computers, hijack user identities, etc.

Some external organizations take control to a higher level. For example, Google provides the Safe Browsing Extension for Firefox that limits the pages the user can visit.[15] While this extension (and other software and Web-based services) provides a helpful service to users and organizations, it demonstrates another opportunity that an external organization could establish a control point surreptitiously.

Systematic monitoring can help an organization understand the control external organizations have. For example, Figure 10 shows that Google's (indirect) "control" of resources on campus is increasing over time. We have also observed in Figure 6 that Amazon's knowledge of the local organization via surveillance is growing faster than others, perhaps due to third parties using their IaaS.

8. CONCLUSIONS

In this paper, we examined a year in the life of a campus network. We examined the communication patterns of three groups of external organizations: first-party providers, third-party providers, and ISPs. While reconnaissance activities like scanning receive a lot of attention in the research and operations communities, our results show that surveillance can also reveal a substantial amount of information about an edge network's infrastructure and usage. We demonstrated that popular first-party providers can obtain a much more accurate and up-to-date picture of an edge network through surveillance than an entity engaged in reconnaissance. We also showed that popular third-party providers can obtain a similarly complete assessment of an edge network, even though users do not intentionally direct traffic to them. We argue that by increasing an edge network's awareness of what external entities know about them, the security of the edge network could be improved. If

[14] `http://tinyurl.com/OpenStream`

[15] `http://www.google.com/tools/firefox/safebrowsing/`

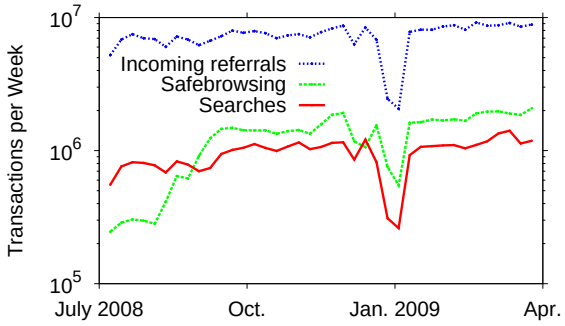


Fig. 10. Breakdown of Google Transactions.

this approach was widely adopted, it would benefit overall Internet security.

As future work, we plan to work towards obtaining "real-time", prioritized intelligence about the use of our network. Part of this plan will involve improving the performance and scalability of our analysis tools. There is a need for more scalable information services (e.g., mapping IP addresses to the organizations they are assigned to) that deliver more complete and consistent information to edge networks. Lastly, we believe that tools to facilitate sharing of selected information would also aid in improving overall Internet security. In particular, proper tools could enable edge networks to systematically help ISPs identify and shut down the sources of the types of undesirable traffic we reported.

Acknowledgments

This work was supported by the Informatics Circle of Research Excellence (iCORE) in the Province of Alberta and CENIIT at Linköping University. The authors thank the anonymous reviewers for their very constructive feedback. The authors also thank the contributors of the data sets; without their assistance this work would not have been possible.

REFERENCES

Allman, M., Paxson, V., and Terrell, J. 2007. A brief history of scanning. In *Proc. IMC*.

Arlitt, M. and Williamson, C. 2005. An analysis of tcp reset behaviour on the internet. *ACM SIGCOMM Computer Communication Review 35,* 1 (Jan.), 37–44.

Barford, P. and Blodgett, M. 2007. Toward botnet mesocosms. In *Proc. HotBots*.

Barford, P. and Yegneswaran, V. 2006. An inside look at botnets. In *Proc. Workshop on Malware Detection, Adv. in Information Security*.

Collins, M., Shimeall, T., Faber, S., Janies, J., Weaver, R., Shon, M., and Kadane, J. 2007. Using uncleanliness to predict future botnet addresses. In *Proc. IMC*.

Duffield, N., Haffner, P., Krishnamurthy, B., and Ringberg, H. 2009. Rule-based anomaly detection on IP flows. In *Proc. IEEE INFOCOM*.

Gates, C., McNutt, J., Kadane, J., and Kellner, M. 2006. Scan detection on very large networks using logistic regression modeling. In *Proc. ISCC*.

Jin, Y., Sharafuddin, E., and Zhang, Z. 2009. Unveiling core network-wide communication patterns through application traffic activity graph decomposition. In *Proc. ACM SIGMETRICS*.

Jin, Y., Simon, G., Xu, K., Zhang, Z., and Kumar, V. 2007. Gray's anatomy: Dissecting scanning activities using IP gray space analysis. In *Proc. SysML*.

Jin, Y., Zhang, Z., Xu, K., Cao, F., and Sahu, S. 2007. Identifying and tracking suspicious activities through gray space analysis. In *Proc. MineNet*.

Jung, J., Milito, R., and Paxson, V. 2007. On the adaptive real-time detection of fast-propagating network worms. In *Proc. DIMVA*.

Jung, J., Paxson, V., Berger, A., and Balakrishnan, H. 2004. Fast portscan detection using sequential hypothesis testing. In *Proc. SP*.

Karasaridis, A., Rexroad, B., and Hoeflin, D. 2007. Wide-scale botnet detection and characterization. In *Proc. HotBots*.

Katti, S., Krishnamurthy, B., and Katabi, D. 2005. Collaborating against common enemies. In *Proc. IMC*.

Klensin, J. 2001. Simple mail transfer protocol. RFC 2821.

Krishnamurthy, B. and Wills, C. 2009. Privacy diffusion on the web: A longitudinal perspective. In *Proc. WWW*.

Li, Z., Goyal, A., Chen, Y., and Paxson, V. 2009. Automating analysis of large-scale botnet probing events. In *Proc. ASIACCS*.

Muelder, C., Ma, K., and Bartoletti, T. 2005. A visualization methodology for characterization of network scans. In *Proc. VizSec*.

Padhak, A., Qian, F., Hu, C., Mao, M., and Ranjan, S. 2009. Botnet spam campaigns can be long lasting: Evidence, implications and analysis. In *Proc. ACM SIGMETRICS*.

Pang, R., Yegneswaran, V., Barford, P., Paxson, V., and Peterson, L. 2004. Characteristics of internet background radiation. In *Proc. IMC*.

Paxson, V. 2004. Strategies for sound internet measurement. In *Proc. IMC*.

Postel, J. 1982. Simple mail transfer protocol. RFC 821.

Shankar, U. and Paxson, V. 2003. Active mapping: Resisting NIDS evasion without altering traffic. In *Proc. SP*.

Sommers, J., Yegneswaran, V., and Barford, P. 2004. A framework for malicious workload generation. In *Proc. IMC*.

Specht, S. and Lee, R. 2004. Distributed denial of service: Taxonomies of attacks, tools and countermeasures. In *Proc. Parallel and Distributed Computing Systems*.

Staniford, S., Paxson, V., and Weaver, N. 2002. How to 0wn the internet in your spare time. In *Proc. USENIX Security Symposium*.

Weaver, N., Sommer, R., and Paxson, V. 2009. Detecting forged tcp reset packets. In *Proc. NDSS*.

Weaver, N., Staniford, S., and Paxson, V. 2004. Very fast containment of scanning worms. In *Proc. USENIX Security Symposium*.

Xie, Y., Yu, F., and Abadi, M. 2009. De-anonymizing the internet using unreliable ids. In *Proc. ACM SIGCOMM*.

Xu, K., Zhang, Z., and Bhattacharyya, S. 2008. Internet traffic behavior profiling for network security monitoring. *IEEE/ACM Transactions on Networking 16,* 6 (December), 1241–1252.

Yegneswaran, V., Barford, P., and Ullrich, J. 2003. Internet intrusions: Global characteristics and prevalence. In *Proc. SIGMETRICS*.

Yin, X., Yurcik, W., Treaster, M., Li, Y., and Lakkaraju, K. 2004. Visflowconnect: Netflow visualizations of link relationships for security situational awareness. In *Proc. VizSec*.

Zhuang, L., Dunagan, J., Simon, D., Daniel, R., Wang, H., and Tygar, J. 2008. Characterizing botnets from email spam records. In *Proc. LEET*.

Zou, C., Gong, W., Towsley, D., and Gao, L. 2005. The monitoring and early detection of internet worms. *IEEE/ACM Transactions on Networking 13,* 5 (October), 961–974.