Categorizing Computing Assets According to Communication Patterns

Dieter Gantenbein[1], Luca Deri[2]

[1]IBM Zurich Research Laboratory, 8803 Rueschlikon, Switzerland
dga@zurich.ibm.com, http://www.zurich.ibm.com/~dga/

[2]NETikos S.p.A., Via Matteucci 34/b, 56124, Pisa, Italy
deri@ntop.org, http://luca.ntop.org/

Abstract. In today's dynamic information society, organizations critically depend on the underlying computing infrastructure. Tracking computing devices as assets and their usage helps in the provision and maintenance of an efficient, optimized service. A precise understanding of the operational infrastructure and its users also plays a key role during the negotiation of outsourcing contracts and for planning mergers and acquisitions. Building an accurate inventory of computing assets is especially difficult in unknown heterogeneous systems and networking environments without prior device instrumentation. User mobility and mobile, not-always-signed-on, computing devices add to the challenge. We propose to complement basic network-based discovery techniques with the combined log information from network and application servers to compute an aggregate picture of assets, and to categorize their usage with data-mining techniques according to detected communication patterns.

Keywords. Network and System Management, Inventory and Asset Management, IT Audit, Due Diligence, Data Mining, and OLAP.

1 Computing Infrastructure as a Critical Business Resource

Modern e-business environments[1] tightly link the customer and supplier systems with the internal computing infrastructure. Hence the performance of the end-to-end business processes becomes critically dependent on the availability of the underlying computing infrastructure. From an economic perspective, the efficient cost-effective and resource-optimized provision of the required services is an argument in many organizations to justify the tight grip on the deployed computing assets and their usage [16].

Classical methods for asset and inventory management quickly reach their limit in today's dynamic environments: Periodic physical inventories ("wall-to-wall") have the clear advantage of identifying the actual location of the devices but require costly human visits ("sneaker net") and can detect neither mobile, currently out-of-office

[1] An e-Business definition can be found at http://searchebusiness.techtarget.com/sDefinition/0,,sid19_gci212026,00.html

equipment nor the existence and use of contained logical assets. Financial asset tracking, while being an accepted process in its own right, cannot detect additional equipment brought into or remotely accessing the resources of an organization. Periodic self-assessment questionnaires to be filled out by individual end users or their cost-center managers are another and often complementary approach. Apart from the human effort they require and the inaccurate incomplete data that results, most forms pose questions the answer of which could be easily read out of the infrastructure itself.

Well-managed computing infrastructures typically equip servers and end-user devices with software daemons for the tracking of resources and the system as well as for application performance monitoring [29][30]. There are many situations, however, in which this cannot be assumed and used. In many organizations, there are a fair number of devices that are brought in ad-hoc and are not instrumented accordingly, for which instrumentation is not available, or on which instrumentation has been disabled. After a merger/acquisition, for example, we can hardly assume to find an encompassing management environment in place across the entire holding organization. However, a good understanding of the provided infrastructure and its users is essential, actually already prior to the acquisition or while negotiating an outsourcing contract. Such investigations to gather the data required allowing more accurate service cost predictions and reducing the risk of unforeseen contractual responsibilities are often called Due Diligence or Joint Verification.

In this paper, we argue that it is no longer sufficient to keep a static inventory of information technology (IT) assets, but that the online tracking and categorization of resource usage based on observed communication patterns provide a much richer information base and enables faster and more accurate decisions in today's evolving e-business environment. For example, some typical workstation hardware could well be used in server role. While the maintenance of higher security and availability levels is essential for servers, this increases the IT cost. Hence early server-role detection is key to operational planning and financial compensation. In general, detailed asset configuration and usage records enable integrated Infrastructure Resource Management (IRM) processes, with value propositions ranging from network management, over help desk services, asset tracking and reporting, software distribution and license metering, service contract and warranty information, to the support of leasing, procurement, acquisition, and migration services.

Section 1 of this paper surveyed the business rationale for detailed assets inventory and monitoring in a networked world. Section 2 reviews current network-based discovery techniques, while Section 3 analyses information found in common network and application log files. Sections 4 and 5 then propose how to warehouse and compute aggregated activities to prepare for the data mining to categorize assets and users. Section 6 concludes with a validation of the log analysis techniques on a small campus network.

2 Network-Based Asset Discovery and Tracking Techniques

An increasingly popular inventory method is to collect information using the network itself. Network-based inventories can significantly reduce the time and cost of an IT audit and can also make regularly scheduled inventories feasible, providing more up-to-date information [2]. Building an accurate inventory of computing assets in an unknown systems and networking environment is a challenging task. Redundant sources of information may be desirable to build reliable algorithms and tools addressing largely heterogeneous environments. Accounting records, physical inventory baselines, end-system configurations, traffic monitors, networking services, and historic network and application server logs all represent valid sources of information (see Figure 1). The ultimate goal is to integrate all incoming records into a consistent model of what is out there and how it is being used [16].

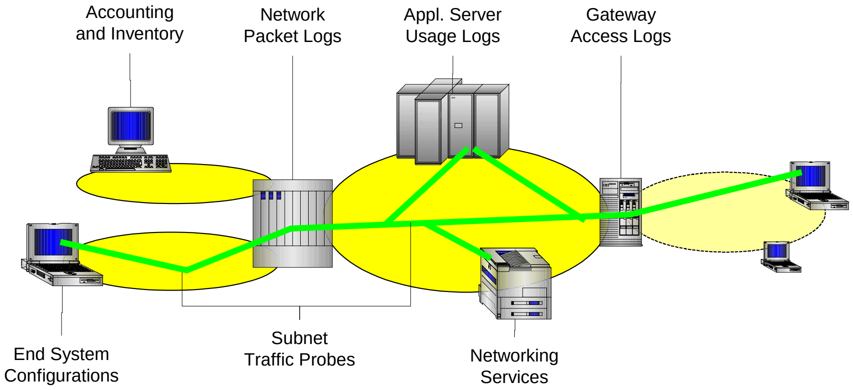

Fig. 1. Networks, Communication Paths, and Sources of Information

Network-based asset discovery and tracking techniques can be classified into “online” methods (to determine the actual state of end-systems, network and services) and “historic” log information processing (to analyze recorded network and services usage traces). Although online monitoring may also keep historic information, it cannot see into the past, i.e. into events that happened prior to its start. If the time period available for asset discovery is limited, i.e. too short to see all sporadically connecting devices, or if it is difficult to obtain sufficiently broad and deep administrative access to the overall infrastructure, it is attractive to reconstruct the global picture of computing assets and their usage from historic and redundant log information [10].

Before focusing on the analysis of log information in the subsequent chapters, the following section surveys currently known techniques for online discovery. As described in [27], there is no single technique that can perform an exhaustive network discovery, as every network has some peculiarity (e.g. some network parts are protected by firewalls, whereas others are accessible only from a few selected hosts) and also because not all networks run the same set of network protocols. In addition, network interface cards for mobile computers can easily be shared (plug-n-play) and swapped, and more and more portables are already equipped with wireless network interfaces, making these devices difficult to track and identify as they can move and change link and network addresses during their lifetime.

Network-based online discovery techniques can be further classified into methods that (i) passively listen and map the network, (ii) actively talk and walk the network and services, and (iii) interact and footprint individual hosts and target application services. The following sections highlight and position various techniques used by the authors to explore an unknown networking environment.

2.1 Passive Network Mapping

Passive network mapping enables the discovery and identification of network assets in a purely passive fashion, i.e. without generating any kind of traffic that stimulates target machines in order to discover their presence.

Network Packet Sniffing

Packet sniffing consists of capturing packets that are received by one or more network adapters, and does not interfere with normal network operations as the packet capture application (network probe) generates no traffic whatsoever. As modern networks make intensive use of switches for filtering out unnecessary traffic, a probe can see only traffic directed to the host in which the probe is running and broadcast/multicast traffic. The network administrator's reaction to switched networks is to adopt techniques such as ARP poisoning (MAC layer address-resolution protocol interaction to enable "man in the middle" attacks) and port mirroring (a network debugging feature assigning a port from which to copy all frames, and a port to which to send those frames) to avoid duplicating probes on each sub-network to be monitored. Packet capture is location dependent; hence the probe should be placed where the traffic actually flows, which can pose quite a challenge. The probe needs to have decoders for each of the protocols the network administrator is interested in. As network traffic can be quite bursty, probes must be fast enough to handle all traffic in quasi real-time and to avoid loosing track of the ongoing traffic sessions. Applications that belong to this category include [24] and [9].

Subscription to Network and Syslogs

As most of the network devices and services store activity reports in log files which can easily get enabled for incremental remote forwarding via syslog, it is quite common to subscribe to such log file events for tracking network activities. In particular, log file data can be very useful in the case of dial-in modem servers, corporate VPN gateways, mobile users, and WAP gateways. Some drawbacks of using logs are that their format is usually fixed and not customizable by network administrators, that it is necessary to periodically read the logs as they can wrap and hence overwrite historical information, and that access typically requires administrator privileges.

2.2 Active Network Mapping

There are several different techniques that can be employed for actively mapping network assets. They all share the principle that the network needs to be exhaustively explored from a starting point using a repetitive algorithm that walks the entire network up to an endpoint or until the entire IP address range has been exhausted.

SNMP Walking of Network Topology

Starting point: the current default route of the host that performs the mapping. Recursive propagation algorithm: using SNMP [1] contact all adjacent routers, learn all their current interfaces, and read their ARP table for learning all local hosts. Applications that belong to this category include [21]. Specific MIBs can also provide hardware-related configuration information, e.g. allow the algorithm to drill down to racks, frames, and individual plugs in wiring closets. Termination condition: recurs until network closure or until a limited access authorization (SNMP community string) blocks walking. The technique is potentially hazardous for networks as SNMP traffic is not suitable for large networks and can interfere with normal operations. For security reasons, network administrators may deny remote SNMP GET operations. Moreover, SNMP can be either disabled or misconfigured.

Network-Wide IP Ping Sweeps

Starting point: the host that performs the mapping. Contact all hosts of the network being explored, e.g. using ICMP ECHO (a.k.a. ping). Terminate when all addresses have been tried or when a firewall/router blocks the traffic. Evaluation: End-to-end method that works in many situations where SNMP router walking is blocked. NAT and firewall devices block inbound IP ping sweeps, whereas unmanaged IP addresses can still talk to most servers, even in other network segments. Hence the starting point for ping sweeps should be selected carefully. Ping typically works as long as end systems have an inherent business need to communicate. The generation of ICMP traffic may interfere with normal operations. ICMP ECHO can be (partially) disabled for security reasons. Other techniques [22] [31] may produce better results in terms of efficiency and accuracy.

DNS Network Domain Name-Space Walking

Starting point: the local DNS server. Algorithm: walk the DNS space by performing a zone transfer in order to know all known hosts and DNS servers. Recurs until network closure or until a DNS forbids the zone transfer. Evaluation: technique can produce misleading data as some DNS servers may be out of synchronization with the actual network state. Typically provides good information about stable network services and applications; naming conventions may allow further conclusions on intended main host usage (DNS, Notes, Mail Exchange, etc.). DNS walking is useless in non-IP networks of course, and fails on networks in which names have not been configured. Its results need to be carefully analyzed in particular when dynamic address protocols (e.g. BOOTP and DHCP) are in use.

DHCP Lease Information

Starting point: the local DHCP service or administrative access to the corresponding server. There is no standardized access across products. Microsoft’s Win/NT

Resource Kit contains utilities to find DHCP servers and list clients. Resulting data contains the currently assigned IP addresses with associated MAC address as key to client information. The value of this technique lies in particular in the tracking of devices connected only sporadically with the network.

Windows and Novell Network-Domains, LDAP, and Active Directory

Starting points: the local directory services of LDAP and the corresponding Microsoft and Novell application networking. This technique nicely complements DNS walking on IP-based networks. On networks in which dynamic DNS is implemented results can partially overlap owing to misconfigurations; directories tend to have richer data [18].

2.3 Host and Service Mapping

As hosts get discovered, the procedure typically continues by drilling down on each active host. We may want to employ specialized tools to learn about the currently running operating system (OS) and the services the host provides. One of the principles of network mapping is that the more we know about a host the more we can find out. The easiest way to learn the OS is to parse the logon banners a server returns when opening TCP/IP ports. Unfortunately, not all hosts offer such services. However, a large class of workstations provides further Windows-specific data. If both approaches fail, the ultimate resort is to use advanced techniques to directly analyze the TCP/IP stack.

TCP/IP Stack Analysis and OS Detection

The standardized TCP/IP protocols allow a certain degree of local-system freedom. Such local choices may impact the system tuning and application performance. Hence different stack implementations feature slightly differing behavior, especially when challenged with peculiar protocol situations such as bogus flags, out-of-scope packets, or windowing information. The current internet-shared database contains fingerprints for more than 500 IP implementations. The most popular tool is [22]. The results of stack-based OS analysis must be carefully interpreted as they are based on heuristics that can produce vague or even wrong results. Moreover, care must be exercised when interacting with some - typically old and badly maintained - hosts as the odd requests against the stack may crash the host currently being mapped.

UDP/TCP Port Scans

Port scanning is the technique that tries to communicate with remote ports, and map the TCP/IP services available from a host. Access can be tried by using either a small list of well-known ports (such as TELNET, FTP, HTTP and POP3), the entire range of named services, or by scanning the entire 64K-large port range. In addition, access can stop when granted (early close) or continue (full open) until the banners are displayed. The former can be an alternative to ping in network environments that block ICMP packets. There are also various tools available to security analysts that further scan for possible security vulnerabilities [23]. To reduce the impact on end-systems and visibility in intrusion-detection systems, “stealth” modes are commonly available to sequence randomly across IP addresses and ports. Unfortunately port scan is a potentially hostile activity, hence it needs to be used carefully and only after that the local network administrators have been informed. There is a growing list of personal tools that are able to detect port scans[2].

Remote Windows Fingerprinting
For Windows systems, there are specialized scanners that connect to the remote interfaces of Windows systems management. In particular, and especially with proper credentials, this yields a wealth of information on the hardware, networking, and software configuration of the remote host. An example is WinFingerprint [33].

In the literature and in our experience, all the techniques described above produce good results although none is accurate by itself. In fact, to improve mapping accuracy it is necessary to combine various techniques and filter out results using advanced data-mining techniques. This is the approach we use for mapping network hosts and services. The following chapter explains which techniques enabled the authors to extend the map to mobile computers and learn more about asset usage in general.

3 Processing Logs to Track Assets and Their Usage

The ability to access information virtually anywhere - at any time - is transforming the way we live and work. Pervasive computing encompasses the dramatically expanding sphere of computers from small gadgets to computers embedded within and intrinsically part of larger devices. Such devices are characterized as not always powered on, and may not always be connected to a network. Not each pervasive device is a wireless terminal. Some connected devices are not IP networked. Pervasive devices typically have a shorter life cycle than classical workstations do [15]. Together, these effects make asset management hard.

Roaming individuals may remotely access corporate resources from corporate and third-party devices, using various network access points. While this renders the borderline of an organization fuzzier, the tracking of corporate assets and resource usage remains a business-critical issue. The network discovery methods described, which typically poll a tightly connected network and systems environment, tend to fall short in what they observe in this case. It may not be feasible to subscribe to network access servers and events in real time. As in the case of tracking shy deer that visit the watering holes only at night, we may have to live with after-the-fact analysis. Even if the device has already disappeared, there is still a wealth of traces recorded at always-on servers. Analyzing server logs actually allows computers to be tracked via communication traces over extended periods in much shorter time.

Consider, for example, polling access-specific gateways, IP address lease records, processing packet/session logs of firewalls and VPN servers, and analyzing application-specific logs from Intranet web, mail, and other subsystems. Figure 2 shows a model of the data contained in log files from the following network services: SOCKS network proxy gateways, DNS name servers, HTTP web servers, POP/IMAP email access and SMTP mail submission servers. Despite the many differences

[2] Scanlogd http://www.openwall.com/scanlogd/, and
Norton Personal Firewall http://www.symantec.com/sabu/nis/npf/

among the various protocols and log entry formats, there are some obvious commonalities with respect to referencing external objects such as hosts, ports, and users (depicted as self-contained entities in the center of the picture).

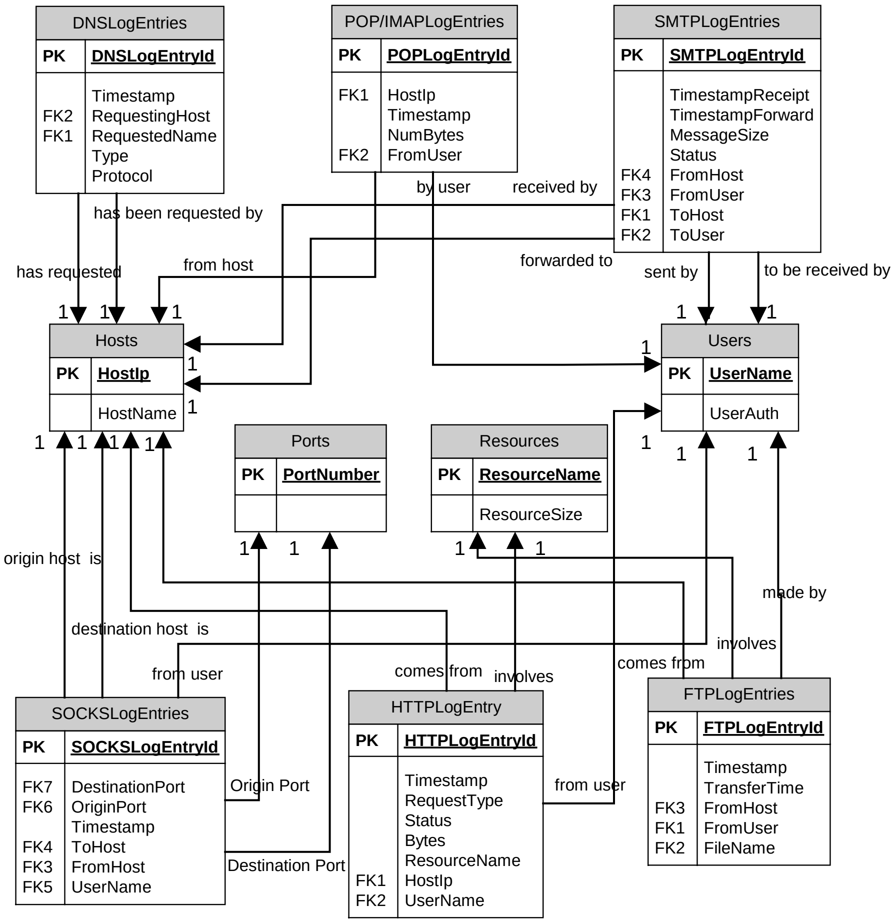

Fig. 2. Entity-Relational Modeling of Data Contained in Network Services Log Files

In summary, the growing complexity and dynamics of a typical IT environment are counterbalanced by the wealth of log information all around. While physical inventories are costly and can hardly face the dynamics of the growing number and diversity of end-user devices, substituting online methods is aligned with the trend to lean administration. This does not prevent - indeed, may require - tight control of the use of the organization’s resources. We anticipate that in the long run the gathering of usage information from the various network and application subsystems is the most cost-effective asset and usage management scheme. Being the best method for pervasive devices, it may actually become standard usage also for the management of other, more static hosts, always-on workstations and servers. In the following, we focus on investigating this approach.

4 Warehousing Asset Usage Records

Again, our proposed overall approach is to complement basic network-based discovery with the combined log information from network and application servers, and then to compute an aggregate picture of assets and categorize their usage with data-mining techniques. Figure 3 depicts the stages for warehousing the log information in a relational database.

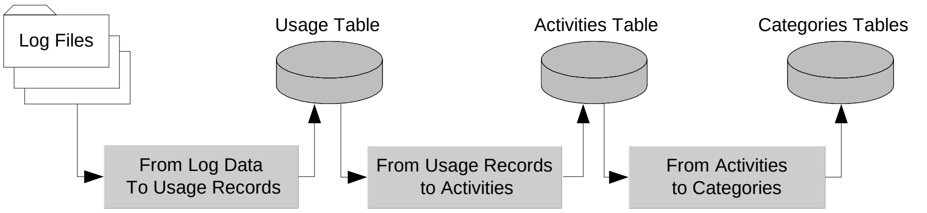

Fig. 3. Parsing, Cleaning, Coding, and Aggregating Log Information

Log files from selected network and application servers are first normalized into a denser data representation which we call usage records. This step also allows abstracting from the specific application server products and log-file formats. In a second stage (described in Section 5) related usage records from multiple network and application protocols and observation points are aggregated into a server-independent perception of activities. Usage and activity records are then input to analytic processing and data mining.

In our study we focused on log information related to services that are most widely used by network users today. We processed HTTP web and SMTP/POP/IMAP mail application logs, and analyzed TCPDUMP network packet logs. We will show how we can benefit from this redundancy of data belonging to several servers that manage different protocols. For example, if we see an HTTP activity originating from a particular host and would like to associate this with a user, we can search for email activities that originated from the same host in the same interval.

4.1 HTTP Logs and Usage Records

The W3 Consortium has standardized HTTP log file formats; in addition there is the basic Common Log File Format and the Extended Common Log File Format, which adds Referer and UserAgent information fields [32]. Fortunately, most web server products follow these formats. Deviations can be absorbed with a regular-expression-based parsing front end. The following shows an example of the original web server log file entries for accessing the root web page (index.html) on host 131.114.4.XX:

```
131.114.4.XX - - [25/Aug/2001:22:54:16 +0200] "GET / HTTP/1.0" 200 4234 "-"
"Mozilla/4.71 [en] (WinNT; I)"
131.114.4.XX - - [25/Aug/2001:22:54:16 +0200] "GET /images/header.gif HTTP/1.0" 200
9342 "http://www.di.unipi.it/" "Mozilla/4.71 [en] (WinNT; I)"
```

The corresponding consolidated single usage record stored into the database looks as follows:

ID	Record Type	StartTime	EndTime	Initiating User	Initiating Host	Target User	Target Host	Global Ref	Local Ref	Description	Data Pkts	Data Vol
1	Http Session	2001-08-25 22:54:16	2001-08-25 22:55:20		131.114.4.XX		www.di.unipi.it	/	/ /images/header.gif …	Mozilla/4.71 [en] (WinNT; I)	7	45701

4.2 SMTP Logs and Usage Records

Unfortunately in this case there is no standard log file format across email server products [26] [19], although they all relate to RFC821 [28]. Parsing of specific log file formats can be simplified with regular-expression-based frontends. The following shows an example from Sendmail Version 8 for the case of user Rossi@di.unipi.it sending an email to local user Verdi@di.unipi.it and to remote user Bianchi@informatik.uni-freiburg.de.

```
Jun 18 09:26:37 apis sendmail[30933]: JAA14975: from=<rossi@di.unipi.it>, size=1038,
class=0, pri=61038, nrcpts=2, msgid=<005101c0f7ee$54e36640$5b027283@kdd>,
proto=SMTP, relay=pc-rossi [131.114.2.91]
Jun 18 09:27:06 apis sendmail[30934]: JAA14975: to=<verdi@di.unipi.it>,
ctladdr=<rossi@di.unipi.it> (15124/110), delay=00:00:29, xdelay=00:00:00, mailer=local,
stat=Sent
Jun 18 09:27:06 apis sendmail[30934]: JAA14975: to=<bianchi@informatik.uni-freiburg.de>,
ctladdr=<rossi@di.unipi.it> (15124/110), delay=00:00:29, xdelay=00:00:28, mailer=esmtp,
relay=mailgateway1.uni-freiburg.de. [132.230.1.211], stat=Sent (OK id=15BxcV-0003Xy-00)
```

The corresponding usage records stored into the database look as follows:

ID	Record Type	StartTime	EndTime	Initiating User	Initiating Host	Target User	Target Host	Global Ref	Local Ref	Data Pkts	Data Vol
1	MsgSending	2001-06-18 09:26:37		Rossi @di.unipi.it	pc-rossi.di.unipi.it			Msg1 @di.unipi.it	JAA14975	1	1038
2	LclForwarding	2001-06-18 09:27:06	2001-06-18 09:27:06			Verdi @di.unipi.it			JAA14975	1	
3	RmtForwarding	2001-06-18 09:27:06	2001-06-18 09:27:xx			Bianchi @informatik.uni-freiburg.de	mailgateway1.uni-freiburg.de		JAA14975	1	

Note that we chose not to compress the data into a single record in order to maintain easy access to the individual destinations in the usage table.

4.3 POP Logs and Usage Records

Post Office Protocol Version 3 (POP) is an Internet protocol that allows a client to download email from a server to his or her Inbox on the local host, where messages are then managed. This is the most common protocol, and works well for computers that are unable to maintain a continuous connection to a server. Internet Mail Access Protocol Version 4 (IMAP) enables a client to access email on a server rather than downloading it. The following is an example where user Rossi starts a POP session from his host pc-rossi.di.unipi.it in order to download all emails that arrived since the last poll:

```
Jun 18 09:26:49 apis ipop3d[733352]: pop3 service init from 131.114.2.91
Jun 18 09:26:50 apis ipop3d[733352]: Auth user=Rossi host=pc-rossi.di.unipi.it [131.114.2.91]
nmsgs=32/32
Jun 18 09:26:51 apis ipop3d[733352]: Logout user=Rossi host=pc-rossi.di.unipi.it
[131.114.2.91] nmsgs=27 ndele=5
```

The corresponding usage record stored into the database looks as follows:

ID	Record Type	StartTime	EndTime	Initiating User	Initiating Host	Target User	Target Host	Global Ref	Local Ref	Data Packets	Data Volume
1	Pop Session	2001-06-18 09:26:49	2001-06-18 09:26:51	Rossi@di.unipi.it	Pc-rossi.di.unipi.it		popserver.di.unipi.it			27	32

4.4 TCPDUMP Packet Logs and Usage Records

Network packet logs definitely provide the largest volume of data with an immediate need for compression. Aggregating all entries referring to the same TCP connection into a single usage record was not sufficient (tcpdump compression tool www.tcptrace.org), especially for HTTP V1 clients that keep multiple connections to the same server open at the same time. Therefore, we decided to aggregate all data of all simultaneous connections between two hosts into a single usage record. Each such connection is considered closed when the time elapsed since receipt of the last packet is greater than a fixed gap time period. After an idle gap, further packets between the same pair of hosts result in the creation of a new usage record. We obtained best results with gap values of approx. 1 min. The following example shows a situation in which host 131.114.2.1XY has a POP session with server 217.58.130.18, followed by another host 131.114.4.1ZZ having an HTTP session with a web server on 212.48.9.22 using two simultaneous connections on ports 2099 and 2100, respectively:

```
999598392.171337 > 131.114.2.1XY.45316 > 217.58.130.18.pop3: S
3331168056:3331168056(0) win 5840 <mss 1460,sackOK,timestamp 364005685 0,nop,wscale
0> (DF)
…
999598421.515854 > 131.114.2.1XY.45320 > 217.58.130.18.pop3: S
```

```
3369225773:3369225773(0) win 5840 <mss 1460,sackOK,timestamp 364008620 0,nop,wscale
0> (DF)
…
999597426.543181 > 131.114.4.1ZZ.2099 > 212.48.9.22.www: S 2586282406:2586282406(0)
win 16384 <mss 1460,nop,nop,sackOK> (DF)
…
999597471.802370 > 131.114.4.1ZZ.2099 > 212.48.9.22.www: R 2586282731:2586282731(0)
win 0 (DF)
```

The corresponding usage record stored into the database looks as follows:

ID	Record Type	StartTime	EndTime	Initiating User	Initiating Host	Target User	Target Host	Global Ref	Local Ref	Data Packets	Data Volume
1222	Pop Session	2001-09-04 12:13:12	2001-09-04 12:13:21		131.114.2.1XY		217.58.130.18			24	236
213	Http Session	2001-09-04 11:57:06	2001-09-04 11:57:51		131.114.4.1ZZ		212.48.9.22			21	1732

In addition to the columns shown, each usage record also contains Tmin, Tmax, Taverage and Tstddev fields that statistically describe the time relationships between the constituent log file entries of the usage record.

5 Computing Aggregated Activities to Prepare for Data Mining

As described above, log file entries were first consolidated into denser usage records. As a next step, usage records originating from multiple protocols and observation points are aggregated into a server-independent perception of activities. We continue with our example of sending an email to a local and a remote user. Usage record 1 computed from the network packets between pc-rossi and mailserver gets combined with usage records 2-4 derived from the mail server log.

UsageRecords

ID	Category	Server	Source	Record Type	StartTime	EndTime	Initiating User	Initiating Hostname	Target User	Target Hostname	Global Ref	Local Ref	Data Pkts	Data Vol
1	Net Log	net-gw.di.unipi.it	/etc/logs/net.log	NetSmtp Packets	2001-06-18 09:26:36			pc-rossi.di.unipi.it		mailserver.di.unipi.it			xx	1038
2	SMTP Log	mailserver.di.unipi.itt	/etc/logs/May20.txt	Message Sending	2001-06-18 09:26:37		Rossi@di.unipi.it	pc-rossi.di.unipi.it			Msg1@di.unipi.it	JAA14975	1	1038
3	SMTP Log	mailserver.di.unipi.it	/etc/logs/May20.txt	LocalMsg Forwarding	2001-06-18 09:27:06	2001-06-18 09:27:06			Verdi@di.unipi.it			JAA14975	1	
4	SMTP Log	mailserver.di.unipi.it	/etc/logs/May20.txt	RemoteMsg Forwarding	2001-06-18 09:27:06	2001-06-18 09:27:07			Bianchi@informatik.uni-freiburg.de	mailgateway1.uni-freiburg.de		JAA14975	1	

The result is just one activity record. Note that further usage records from other network observation points and mail servers would just confirm this single activity record with global perspective. The relationship between activities and their constituent usage records is maintained in a separate table. Auxiliary tables are also used to index the hosts and users involved in activities.

Activities

ID	Category	Server	Source	Record Type	StartTime	EndTime	Initiating User	Initiating Hostname	Target User	Target Hostname	Global Ref	Local Ref	Data Packets	Data Volume
5	activity			Email Sending	2001-06-18 09:26:36	2001-06-18 09:27:07	Rossi@di.unipi.it	pc-rossi.di.unipi.it			Msg1@di.unipi.it		xx+3	1038

Activity UsageRecords

ID	ActivityID	UsageRecordID
1	5	1
2	5	2
3	5	3
4	5	4

Activity Hosts

ID	ActivityID	HostName
1	5	pc-rossi.di.unipi.it
2	5	mailserver.di.unipi.it
3	5	mailgateway1.uni-freiburg.de

Activity Users

ID	ActivityID	UserName
1	5	Rossi@di.unipi.it
2	5	Verdi@di.unipi.it
3	5	Bianchi@informatik.uni-freiburg.de

With these aggregation algorithms, continuous web surfing originating at a particular host results in a single activity record. A new activity is created after an inactivity period of 60 min. An activity ends with the last end time of the constituent usage records. Concurrent web surfing and email processing result in separate activities of different types. Sending and checking/receiving email also result in separate activities of different types. When not exceeding the 1-hour inactivity period, sending multiple emails results in a single activity. A background daemon frequently checking for arriving email also results in a single activity.

6 Process Validation on a University Network

To validate the process of combining log information from network and application servers to compute an aggregate picture of computers and users according to detected communication patterns, we tested it with the Computer Science departmental staff network of the University of Pisa. This network of 512 possible IP addresses is rather heterogeneous and mostly unprotected, with the exception of a few systems used for accounting purposes that are shielded by a packet filter firewall. There exist a total of approx. 300 hosts (50 servers and 250 workstations). Under a confidentiality agreement we were able to get access to a full 7-day week of real traffic logs from the departmental web and mail servers and the gateway to the rest of the university networks. The validation playground is depicted in Figure 4.

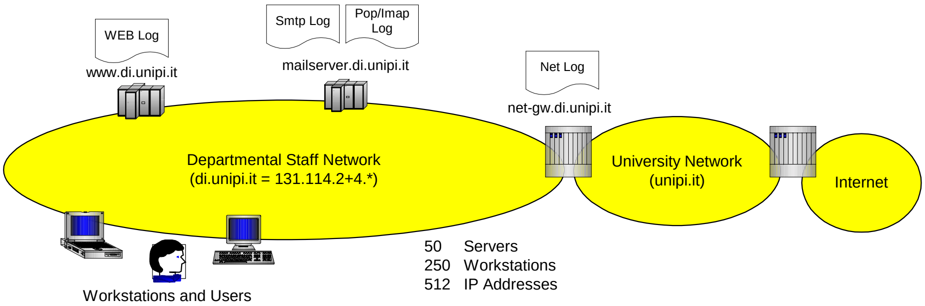

Fig. 4. Actual Network and Log-File Sources at University of Pisa

A significant problem tackled by the authors is the large amount of data to parse. The first week in September in the setup described corresponds to a total of 13 GB uncompressed raw TCPDUMP, 46 MB HTTP, and 8+7 MB POP/SMTP log files. The warehousing of usage records by means of a small Java program took several hours, parsing approx. 100,000,000 log entries, directly filtering out 30% as not-studied protocols, and creating 335,000 usage records. The usage records distribute as follows over the protocols studied: 64% HTTP, 19% SMTP, 14% POP, 3% IMAP. The distribution of the usage-record log source: 68% Net, 10% Web, 22% Mail. In other words, most usage records correspond to Internet web surfing logged by the network gateway.

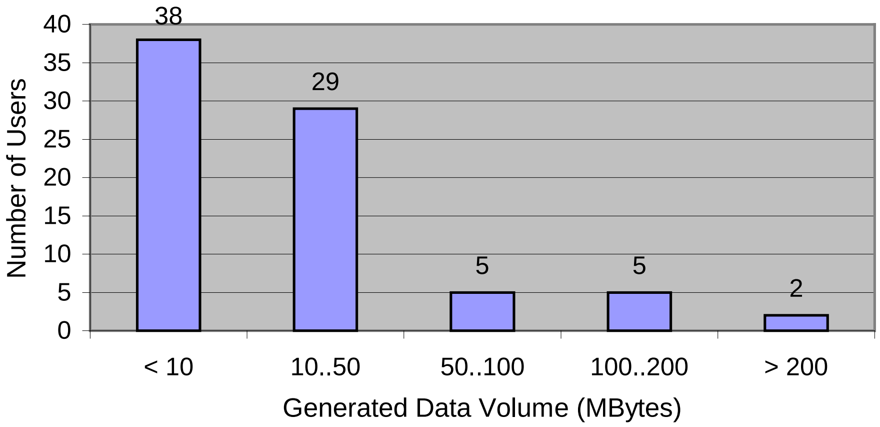

Fig. 5. Categorization of Users According to Generated Traffic

Finally, the usage records were aggregated into 6,700 activities, as consolidated input for the following data analysis.

Details of the validation work to test both the warehousing software and the concepts for categorizing hosts and users can be found in [10]. Some extracts from our findings are summarized here. The protocol used most in terms of usage records (with 1 min.

as usage gap) is definitively HTTP with 64%. As aggregated activities (with 60 min as activity gap) web surfing still represent 23%, whereas email sending and downloading activities are at 53% and 24%, respectively. 51% of the hosts analyzed are used by a single user, hence can be considered private workstations. Most of the traffic occurs during working hours (10AM – 6PM). The total number of activities increases with overall traffic, and the activity duration increases with the activity traffic. Most of the hosts that generate web traffic also send out emails. Most of the hosts generate less than 50 MB of traffic per day. Web activities usually last at least 10 min. 25 of the 79 identified users trigger email downloading explicitly; the others download emails periodically with a polling time above 10 min. For an example of a more advanced data exploration attempt, refer to Figure 5, which categorizes users with at least one personal host according to the total weekly web and email data volume generated by his/her hosts.

Eventually, we would like to derive conclusions such as: "Computer A is used by a secretarial person 5/7 days a week in the morning. Computer B probably is a student-lab workstation, shared by users X, Y and Z.” The lessons learned during the validation process are that (i) there is hope to achieve this – eventually – but (ii) it is of utmost importance to minutely parse, clean, code, and compress the original data sources, and (iii) there is no way around having a baseline sample population of users and computers to establish a data-mining model allowing OLAP predictions in newly discovered environments.

7 Conclusion

In today's dynamic information society, organizations critically depend on the underlying computing infrastructure. Tracking computing devices as assets and their usage helps in the provision and maintenance of an efficient, optimized service. Building an accurate inventory of computing assets is especially difficult in unknown heterogeneous systems and networking environments without prior device instrumentation. User mobility and mobile, not-always-signed-on, computing devices add to the challenge. We therefore propose to complement basic network-based online discovery techniques with the combined historic log information from network and application servers to compute an aggregate picture of assets, and to categorize their usage with data-mining techniques according to detected communication patterns.

In this paper we outlined the process of warehousing and analyzing network and application server logs to track assets and their usage. Given our initial validation, we hope to establish the potential of integrating the consolidated historic knowledge residing in access-specific gateways, firewalls, VPN servers, and network proxies, and the growing wealth of application-specific servers. We anticipate that in the long run the gathering of usage information from the various network and application subsystems will prove to be the most cost-effective asset and usage management scheme. Having already been established as the best method for pervasive devices, it may actually become standard usage for unknown heterogeneous environments and nicely complement the management of other, more static hosts, always-on workstations and servers.

8 Future Work

The authors are aware that this work is not complete as there are open problems and challenges in categorizing computing assets. Referring to the lessons learnt during the validation process, the value of the warehoused data increases with more careful cleaning and coding. A baseline population of assets and users is then essential as further input for OLAP processing to establish a data-mining model allowing later predictions in unknown environments. In particular, this work also needs to be extended in order to categorize accurately mobile and pervasive devices that can roam across different networks and be active only for a limited period of time.

Network intrusion-detection and customer-relationship management are established fields that benefit from OLAP techniques. We hope to promote the use of similar forms of data mining also as techniques for corporate asset management and to establish a flexible and dynamic management infrastructure for e-business services. Figure 6 proposes a chain of processing steps, starting with the classical network discoveries, adding log analysis for usage categorization, that may eventually allow questions about the cost, utility, and risk associated with individual assets to be answered on the one hand, and the computation of associated values on the other.

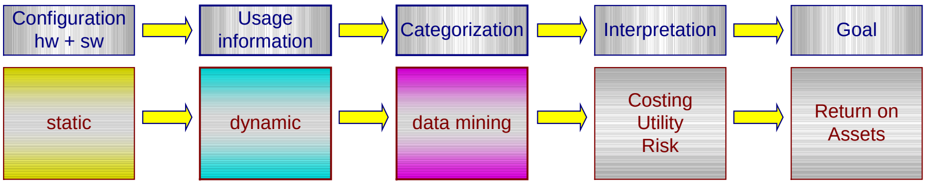

Fig. 6. Processing Stages for the Computation of Business Values

Additional ideas include the following: Associate a geographical location [20] with the categorized assets to facilitate the tasks of physically upgrading, replacing, and accounting. Enhance the proposed system with an accounting application that allows tracking the service usage, its users, and its availability. Study how to generate alarms (e.g. SNMP traps) when an asset modifies its behavior (e.g. if a computer that is known not to handle mail at some point routes emails, it means that something has changed or that a virus is running on the asset).

Acknowledgments

The proposed approach was investigated as part of a Diploma Thesis [10] and validated against real network traffic and application logs provided by the University of Pisa. In particular the authors acknowledge Prof. Maurizio Bonuccelli for his help in accessing this data, samples and statistical findings, which were also used in this paper. We would also like to acknowledge the dedication of Chris Giblin and Matthieu Verbert, who helped design and prototype most the discovery techniques described here.

References

1. J. Case et.al., Simple Network Management Protocol (SNMP), RFC 1157, 1990

2. Centennial, Network Inventory Audit, http://www.intertechnology.ca/soft-1.htm

3. R. Cooley, B. Mobasher, and J. Srivastava, Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, in Proc. KDEX'97, Newport Beach CA, 1997

4. M.S. Chen, J.S. Park, and P.S. Yu, Data Mining for path traversal patterns in a Web environment, in Proc. 16th Int. Conf. On Distributed Computing Systems, p. 385-392, 1996

5. Denmac Systems, Network Based Intrusion Detection: a Review of Technologies, http://www.denmac.com/, November 1999

6. L. Deri and S.Suin, Effective Traffic Measurement using ntop, IEEE Communications Magazine, May 2000, http://luca.ntop.org/ntop_IEEE.pdf.gz

7. L. Deri and S.Suin, Effective Traffic Measurement using ntop, IEEE Communications Magazine, May 2000, Monitoring Networks using Ntop, in Proc. 2001 IEEE/IFIP Int. Symp. On Integrated Network Management, Seattle, WA, 2001

8. L. Deri and S.Suin, Ntop: beyond Ping and Traceroute, DSOM, 1999, http://luca.ntop.org/ntop_DSOM99.pdf.gz

9. Ethereal free network protocol analyzer for Unix and Windows, http://www.ethereal.com/

10. M. Filoni, Computing assets categorization according to collected configuration and usage information, Diploma Thesis, University of Pisa, Italy, November 2001

11. D. Gantenbein, Network-based IT asset discovery and categorization, Presentation, University of Pisa, Italy, October 2000

12. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Academic Press, 2001, ISBN 1-55860-489-8

13. D. Heywood, Networking with Microsoft TCP/IP, 3rd edition, Microsoft Press, New Riders Publishing, ISBN: 0-7357-0014-1, 1998

14. J. Fielding et.al., Hypertext Transfer Protocol – HTTP 1.1

15. Pervasive Computing, IBM Systems Journal, Vol. 38, No. 4, 1999.

    http://www.research.ibm.com/journal/sj38-4.html

16. Intelligent Device Discovery (IDD), IBM Zurich Research Laboratory, Project http://www.zurich.ibm.com/csc/ibi/idd.html

17. V. Jacobson, C. Leres, and S. McCanne, *tcpdump*, Lawrence Berkeley National Labs, ftp://ftp.ee.lbl.gov/, 1989

18. A. G. Lowe-Norris, Windows 2000 Active Directory, O'Reilly, ISBN 3-89721-171-8, 2001

19. Microsoft Exchange: Tracking Log, http://www.microsoft.com/Exchange/en/55/help/default.asp?url=/Exchange/en/55/help/documents/server/XMT04027.HTM

20. Caida, NetGeo: the Internet Geographic Database, http://netgeo.caida.org/, 2000

21. NetView, Tivoli product, http://www.tivoli.com/products/index/netview/

22. NMAP Free Security Scanner, http://www.insecure.org/nmap/index.html

23. NSA Firewall Network Security Auditor, http://www.research.ibm.com/gsal/gsal-watson.html

24. L. Deri, Ntop: a Lightweight Open-Source Network IDS, 1998-2001, http://www.ntop.org/

25. J. Pitkow, In search of a reliable usage data on the www, in Sixth Int. Wold Wide Web Conf., pp. 451-463, Santa Clara CA, 1997

26. Sendmail mail service, http://www.sendmail.org

27. R. Siamwalla et al., Discovering Internet Topology, in Proc. IEEE INFOCOM ’99, 1999, http://www.cs.cornell.edu/skeshav/papers/discovery.pdf

28. Simple Mail Transfer Protocol, J. B. Postel, RFC 821, August 1982

29. Tivoli, Inventory management, http://www.tivoli.com/products/index/inventory/

30. Tivoli, Performance solutions, http://www.tivoli.com/products/solutions/availability/news.html

31. S. Branigan et al., What can you do with Traceroute?, *http://www.computer.org/internet/v5n5/index.htm*

32. Extended Common Log File Format, W3C working Draft, http://www.w3.org/TR/WD-logfile

33. Winfingerprint Windows Information Gathering Tool, http://winfingerprint.sourceforge.net/