

Endpoint Identification Using System Logs

Stephen Melvin

Abstract

Typical logging of public Internet activity involves recording only the source IP address associated with a transaction or access. Unfortunately this is usually insufficient to identify a unique endpoint associated with the activity, mainly due to the common existence of one or more NAT gateways between the originating network and the Internet and the lack of specific logging at the originating end. Here we address the information needed to solve this problem, where it needs to come from, and various scenarios for how it can be managed. Specifically, in order to uniquely identify a specific endpoint, three related pieces of information need to be recorded: 1. the MAC address associated with the local IP address at the originating network (known by a DHCP daemon); 2. the association between the local IP address and port numbers at each NAT gateway (known by a NAT daemon); and 3. the source port number of the public access (known by the remote server). A related problem of MAC address validation, allowing the association of a specific hardware device with a MAC address in use, is discussed as well.

Introduction

It is a usual practice for companies providing access to the Internet and for companies providing content and services on the Internet to generate logs of access and activity. Some examples of how logs are used are for debugging and troubleshooting, detection and monitoring of abuse, statistical analysis, demographic analysis, report generation and other general business purposes. A web server activity log will typically contain information regarding the access, but not the actual content of the access itself. For example, a web server log generally records the originating IP address, the name of the document that was requested, the number of bytes that were transferred to the client machine as well as other information about the user's computer if available. It is common to record in an access log file a record of each access. See for example the Common Log Format of the Apache Web Server [1].

In cases in which a user is directly connected to the Internet using a public IP address, the originating IP address is sufficient to identify the machine at which the request originated. This identification may require the cooperation of multiple parties responsible for IP address blocks, but it is generally possible due to the nature of how IP addresses are globally administered. However in most scenarios the IP

address of the access is not sufficient. Typically an IP address in use by a user lies on a local network and is an unregistered or un-routable address that can be used within an enterprise but cannot be used on the public Internet. Un-routable addresses are addresses that have been set aside in the ranges 10.0.0.0 to 10.255.255.255, 172.16.0.0 to 172.31.255.255 and 192.168.0.0 to 192.168.255.255. IP addresses in this range may be freely used within a private network as they are guaranteed to be unused and unusable on the public Internet. NAT Gateways are used to convert packets coming from un-routable IP addresses into packets with addresses valid on the public Internet. This scheme is utilized to allow many machines to be used on an internal network without tying up public IP addresses, which are global resources. The operation of NAT Gateways on the Internet is well known and in wide use today. One current implementation is the natd daemon of Unix implementations [2].

Frequently internal networks allocate IP addresses using a protocol known as DHCP. This requires the use of a DHCP server attached to the local network. Briefly, the DHCP protocol involves the allocation of IP address upon request by machines on the local network. This operation is known as a "lease" and generally has an expiration time associated with it. The DHCP protocol generally requires periodic communication between a user computer and a DHCP server in order for the user to continue to be allowed to use the IP address to which it has been granted.

Many machines may exist on a local network, and there may be multiple NAT Gateways. This means that a request for a document on the public Internet originating from a browser on a user's machine may be translated multiple times before it reaches the web server that is hosting the document.

Endpoint Activity

Figure 1 illustrates an example of activity logging in which a unique endpoint can be identified. The User Computer is connected to a Local Network which is connected to a NAT Gateway and a DHCP Server. In most scenarios, the NAT Gateway and the DHCP Server will be implemented on the same physical machine and there will only be one network connection from that machine to the Local Network. The NAT Gateway is coupled to the public Internet, which is in turn coupled to a remote Web Server. An Access Log receives information from the Web Server, the NAT Gateway and the DHCP Server. By combining information from all

three sources as described in more detail below, activity logs can be generated that uniquely associate the User Computer with activity on the Web Server.

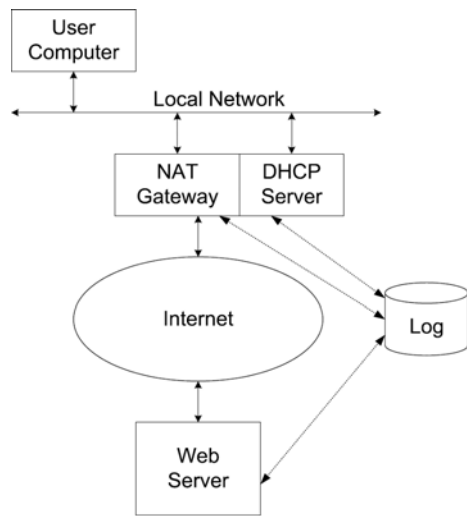


Figure 1

The Access Log is shown in Figure 1 as a single unit for illustrative purposes. The storage of activity data can be distributed across multiple machines and the physical location or locations of the log storage can vary. The Web Server, the NAT Gateway and the DHCP Server could locally store activity information and then periodically transfer it to a central location, or alternatively the activity information could be transmitted immediately to a central repository. In still other scenarios, the activity information could never be stored together in one physical location but could be maintained separately and controlled by separate entities. As long as the requisite information is recorded in some fashion, there are many alternatives to how, when and where the information is combined to create the association between the User Computer and the Internet activity.

This system relies on each User Computer having a unique identifier, for example a MAC address. Every computer having an Ethernet interface in principle has a globally unique MAC address, which is a 48-bit address associated with the Ethernet interface and used as the source address for Ethernet frames transmitted from that interface. The MAC address is created by the manufacturer at the time the interface is created and is part of the link layer encapsulation of an IP packet [3]. Alternative identifiers could be used to uniquely identify a particular computer. For example, some central processing units (CPUs) have unique processor IDs that are created by the microprocessor manufacturer and are globally unique and cannot be changed by the user.

In order to associate the MAC address used by the User Computer with activity that occurs on the Web Server, it is first necessary to record the association between the MAC address used by the User Computer and an IP address

allocated by the DHCP server. Additionally, it is necessary to record the association between an internal and external IP address that is created by the NAT Gateway. Figure 2 illustrates communication and logging events to accomplish these goals.

The User Computer exchanges messages with the DHCP/NAT Gateway, which in turn is coupled to the Remote Server. Figure 2 illustrates the types of information that are logged in order to associate the User Computer with remote activity. When the User Computer is first connected to a local network on which the DHCP/NAT Gateway is also connected, it communicates with the DHCP/NAT Gateway in order to get an IP address to use. The DHCP protocol is typically used to perform this function, although there are alternative dynamic IP address allocation protocols that can be used. When a dynamic IP address is allocated to the User Computer, this is known as a “lease” and will typically last for a defined period of time at which point it needs to be renewed through further exchange of messages.

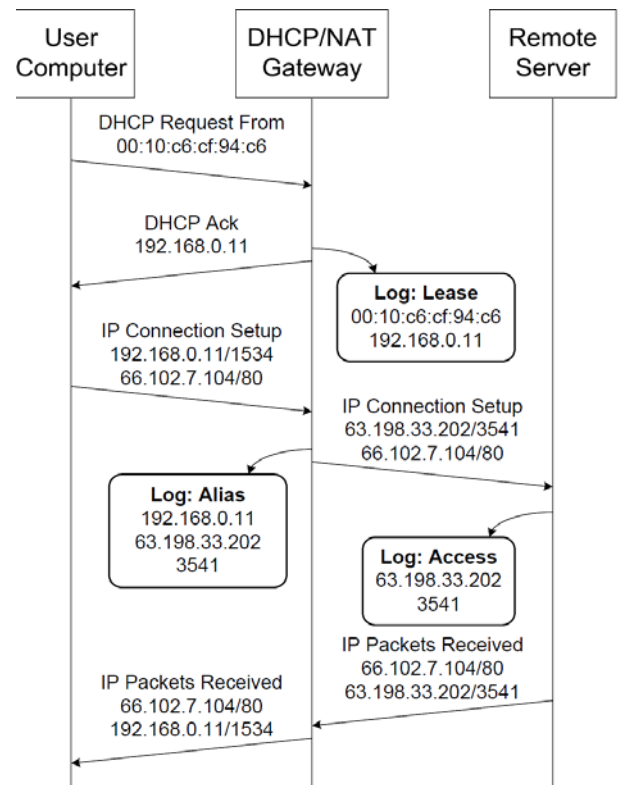


Figure 2

Figure 2 illustrates a simplified exchange of DHCP messages between the User Computer and the DHCP/NAT Gateway for illustrative purposes. In the example of Figure 2, first the User Computer, using the MAC address 00:10:c6:cf:94:c6 requests an IP address from the DHCP/NAT Gateway. Next, the DHCP/NAT Gateway allocates dynamic IP address 192.168.0.11 to the User Computer and sends an acknowledgement message to the User Computer with this

information. At this point, the lease of IP address 192.168.0.11 to MAC address 00:10:c6:cf:94:c6 is recorded.

The establishment of a “lease” represents the grant of an IP address to a particular machine identified by an Ethernet address. In this case the IP address granted is an internal, un-routable address, which can be used on a local network but cannot be used on the public Internet. DHCP servers can typically be configured to allocate either internal or external IP addresses, and can allocate from a pool of IP addresses, or can be configured to associate particular IP addresses with particular MAC addresses.

An example of software that performs the DHCP functionality is the dhcpd daemon commonly used in UNIX based systems. [4] The dhcpd daemon can be configured to listen on certain interfaces and to respond to broadcast messages from machines requesting IP addresses. Some implementations of dhcpd can be configured to automatically log the granting of leases and the expiration of leases. In the present system, the dhcpd daemon is configured to generate this information, and/or is modified to transmit this information to another host, immediately or periodically.

The next sequence illustrated in Figure 2 relates to network address translation (NAT). Because the User Computer is utilizing an un-routable IP address, this address needs to be translated to an external IP address before packets can be sent over the public Internet. This is the job of the NAT gateway. The establishment of an association between an internal IP address and port number to an external IP address and port number is sometimes known as an “alias link.” Because there may be many internal machines communicating with the same remote host, it may be necessary for the NAT gateway to change the port number from the one utilized by the User Computer. Because TCP connections are uniquely identified by source and destination IP address and source and destination port numbers, multiple connections from the same IP address can be established to the same destination port number as long as the source port number is different for each connection.

In the example illustrated in Figure 2, the User Computer sends a packet to set up a connection to a remote Web server at IP address 66.102.7.104, port 80. The source IP address for the User Computer is 192.168.0.11 and the source port number is 1534. Upon receiving the packet from the User Computer, the DHCP/NAT Gateway establishes an alias link, rewrites the outgoing packet and sends it to the Internet. The source address for the outgoing packet is replaced with the source address for the DHCP/NAT Gateway, which in this example is 63.198.33.202. Figure 2 illustrates that the DHCP/NAT Gateway associated port 3541 with the User Computer source port 1541. At this point the alias of source

IP address 192.168.0.11 to external IP address 63.198.33.202, port 3541 is recorded.

An example of software that performs NAT functionality is the natd daemon. In some implementations natd relies on a library known as libalias which performs the function of maintaining a table or database of IP number and port number associations. The libalias library code adds and deletes alias links as needed. In the present system, the libalias library is modified to log certain alias links to a file and/or to transmit this information to another host, immediately or periodically.

The next sequence illustrated in Figure 2 is the receipt of the packet by the Remote Server and the return of a packet to the DHCP/NAT Gateway, which subsequently returns a packet to the User Computer. The Remote Server could be a Web Server, an Email Server or any other server on the Internet for which activity logging is desired. In the example shown in Figure 2, the Remote Server, which is at IP address 66.102.7.104 receives a packet to port 80 from source IP address 63.198.33.202, port 3541. The Remote Server logs the access, *including the source port number* (which is not typically recorded in current implementations). When the source IP address and source port number are correlated with the alias link information and with the DHCP lease information, it is possible to associate a particular computer with activity that occurs on a remote server.

In certain systems, it may not be necessary to record the IP to MAC address association at the time the lease is generated by the DHCP Server. Instead, this information may potentially be generated at the same time the alias information is generated. This is because the packet that is received by the DHCP/NAT Gateway may contain the source Ethernet address of the User Computer. In this case, the DHCP/NAT Gateway can just look at the source Ethernet address and record this as the MAC address associated with the source IP address that is also in the packet. In this case, the information contained in the “Lease” information block and the information contained in “Alias” information block are combined into a single entry created at the same time by the DHCP/NAT Gateway. However this simplification is not always possible because in some systems, the source Ethernet address of the packet received by the DHCP/NAT Gateway is not the original source Ethernet address of the User Computer. This could be the case if there are intervening routers between the User Computer and the DHCP/NAT Gateway. There may also be situations where multiple NAT gateways are employed between the computer originating a packet and the machine that is ultimately responsible for delivering that packet to the Internet.

One example of multiple NAT gateways is a wireless network, which employs a wireless hub, which itself is on a local network that is connected to the public Internet through a

NAT gateway. [5] In this case, packets from a User Computer go through two NAT gateways before reaching the public Internet. A first un-routable address is used on the wireless local network; these packets are translated into packets utilizing a second un-routable address and sent between the wireless hub and a NAT gateway on the wired local network. Finally, the wired NAT gateway translates the packets from the second un-routable address to an external IP address for use on the public Internet. Traceability back to the User Computer requires that the association between the computer and the first un-routable IP address be recorded, that the alias link between the first and second un-routable addresses be recorded and that the alias link between the second un-routable address and the external IP address be recorded.

Figure 1 illustrates a single log repository for illustrative purposes. The repository may be distributed and the correlation of the multiple pieces of information necessary to establish the identity of activity need not be actually performed until needed. For example, since the activity known to the Web Server is under the control of the entity operating the Web site or sites associated with the Web Server, it may be stored separately from the other information. Similarly, the access information known to the NAT gateways and the DHCP servers are typically under the control of the entity who provides access of the computer to the Internet, which may be a different entity from that operating Web Server.

In some cases, it may be sufficient that the information necessary to correlate a specific computer with specific Internet activity is available if and when necessary. Thus, the actual correlation is not performed unless required. For example, it may be the case that the entity providing access of a computer to the Internet protects the alias link and IP lease information unless required to provide it by a Court or law enforcement official, or dictated by an internal investigation. In some cases the entity providing access of a computer to the Internet may be required to preserve the alias link and IP lease information, either by laws governing the entity in whatever jurisdiction they operate, or by contract dictated by the Internet service provider they connect through.

MAC Address Validation

One issue that can arise when logging MAC address to IP address associations, such as through a DHCP lease or other address allocation mechanism, is the validity of the MAC address or other identifying information that is utilized by the User Computer. Some Ethernet interfaces can be re-programmed by a sophisticated user to set the MAC address to an arbitrary value not set by the manufacturer. This facility would allow a computer to masquerade as an arbitrary MAC address, which in some cases would defeat the purpose of uniquely identifying the machine that is connected. The same

is true of any ID number used to identify a computer if the number can be selected arbitrarily by the user. One way to address this issue is to require identifying information to be validated.

In some cases of public access to the Internet, user authentication takes place at the application level where users must type in user names and passwords. In such a case, it can be relatively simple to associate MAC addresses in use and/or allocated IP addresses with individual users. In this case, the related user account can be logged along with the other access information, allowing for possible later association to an individual. In this case, it may not be necessary to validate the MAC address in use, since the user is being identified through other means. In cases where there is no explicit user identification, or where it is important to further validate the access information, identification validation can be performed.

Figure 3 illustrates MAC address registration and validation. The MAC Address Registrar is responsible for receiving a MAC address and producing a signed version of the MAC address. The MAC Address Validator is responsible for receiving an encrypted and signed MAC address and validating the MAC address to generate a validation status. The registration/validation process is based on the use of public key cryptography. Public key cryptography is based on a matched pair of keys, one used to encode information and one used to decode information. By keeping one of the matched keys private and making the other public, the functions of authentication and encryption can be realized.

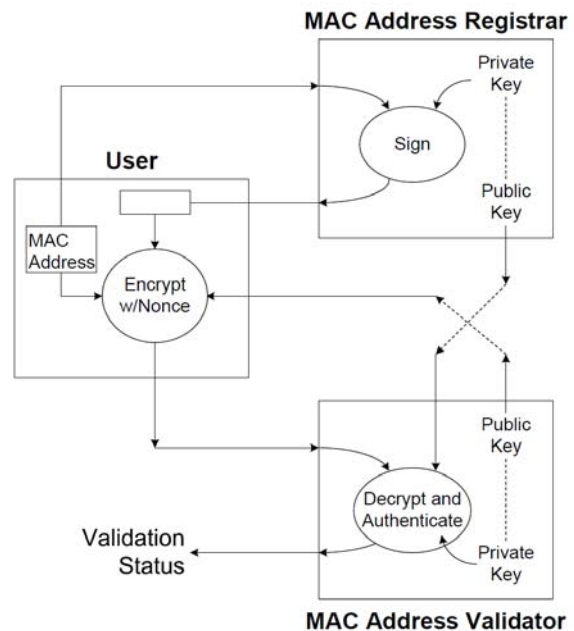


Figure 3

The MAC Address Registrar receives a MAC address and signs it and produces a signed MAC address. The Sign

function utilizes a Private Key of MAC Address Registrar. The use of a private key accomplishes the function of authentication since one can verify using Public Key that the signed MAC address was produced by MAC Address Registrar. The mathematics of the matched key pairs makes it computationally infeasible to generate Private Key knowing only Public Key. Thus, it is impractical to generate a signed MAC address without access to Private Key. This means that Private Key should be maintained in confidence by MAC Address Registrar. There need not be a single MAC Address Registrar, there could be many. Indeed any entity responsible for granting access to the Internet may chose to maintain a separate MAC Address Registrar.

An Ethernet MAC address is 48-bits in length. The purpose of a MAC Address Registrar is to associate a MAC address with a known device, and potentially to verify the MAC address based on other criteria. This may be done, for example, by referring to the manufacturer and model of the hardware in use, by consulting a database of known MAC addressees, or by consulting a database of registered MAC addresses. Once the MAC address provided to the Registrar is verified, a signed version of the MAC address is generated. Because an arbitrary MAC address is usable to someone who can reprogram their Ethernet adaptor, any signed MAC address would be usable to someone wishing to bypass the MAC address registration process. This means that it is desirable for MAC Address Registrar to utilize enough bits in its signature so that it is impractical to guess signed MAC addresses even for arbitrary MAC addresses.

The User is responsible for delivering the MAC Address to the MAC Address Registrar and for saving the signed version of the MAC Address. Preferably the transmission of the signed MAC Address occurs over a secure channel. This is because if someone eavesdrops on this process, they could masquerade as the User by utilizing the MAC Address and signed MAC Address. A variety of techniques are possible to secure the transmission of signed MAC Address to the User. In some embodiments, this process may occur over a private network. The MAC Address Registrar may be operated by an equipment manufacturer, distributor or reseller and may register MAC Address before delivering it to a user. In other scenarios, an HTTP SSL connection is utilized to transfer a Signed MAC Address over an encrypted connection between a MAC Address Registrar and a User. Once the Signed MAC Address is delivered to the User, it is ideally stored in a manner inaccessible to unauthorized software running on the user's machine. This is needed to prevent malware running on the user's computer from retrieving the signed MAC address so that it could masquerade as a verified device. There are a variety of ways to accomplish this secure storage, including the use of passwords and additional encryption. Alternatively, the Signed MAC Address can be stored internal to an embedded microcontroller, such as on a smart card or within

an Ethernet adaptor. In this case, once the embedded system is programmed with the signed MAC address, the address cannot be retrieved even through an exhaustive analysis of software and storage on the user's computer.

The validation process depicted in Figure 3 begins with the use of a Public Key of MAC Address Validator delivered to a User. The use of public key encryption during the validation process guarantees that the Signed MAC Address is not intercepted by an eavesdropper. This would allow such an eavesdropper to masquerade as a User. In one scenario, the Public Key is delivered over a secure channel to User. This is desirable to avoid a Man-In-The-Middle attack, in which an intermediary intercepts the Public Key and replaces it with their own public key. In some cases, the Public Key is delivered to the User at the same time as the Signed MAC Address by the MAC Address Registrar. This may be convenient in situations where the MAC Address Registrar is operated by the same entity that operates the MAC Address Validator. In this case, the Public Key could be stored in the same manner as the Signed MAC Address, including on a smart card if such a facility is used. In another scenario, the Public Key is signed by a known Certificate Authority, the public key for which is previously known to the User. In this manner, the User can verify that the public key being input is indeed the public key for the MAC Address Validator. In order to protect the confidentiality of the Signed MAC Address, it is important to ensure that the User only encrypts it with keys from entities authorized to receive it.

In order to prevent a Replay attack, in which an eavesdropper listens to the transmission of an encrypted signed MAC address, it is useful to combine the signed MAC address with a number used once or "nonce." An example of nonce is a time stamp of sufficient length and granularity. Another possible implementation would be for the MAC Address Validator to generate a random number internally and send it to the User for combination with the signed MAC address. When the MAC Address Validator receives the encrypted and signed MAC address, decryption and authentication are performed using the Private Key and the Public Key, and a validation status is produced. The MAC Address Validator utilizes the Public Key of the MAC Address Registrar to authenticate the MAC Address. In one scenario, the delivery of the Public Key to the MAC Address Validator occurs on a secure channel, to prevent an attack in which a signed MAC address is faked according to keys not belonging to the MAC Address Registrar. In some cases, the MAC Address Registrar and the MAC Address Validator are co-located and operated by the same entity.

The above description has been with regard to MAC addresses, but it equally applies to any form of identification that can be represented in digital form. In some cases, some or all of the functionality described in connection with the

User is built into a network interface card by the manufacturer and transparent to the user. For example, an Ethernet card could be pre-registered with a Signed MAC Address and the Public Key of the MAC Address Validator could be pre-installed. In order to validate, the Ethernet card merely encrypts the signed MAC address with a timestamp and makes it available to higher level software, which can then include this number during DHCP registration. In this case the validation of the MAC address is completely transparent to the user and would not affect implementations that do not rely on this feature. In some embodiments the encrypted and signed MAC address could be made part of the DHCP protocol, in which case the DHCP server could be modified to communicate with a MAC Address Validator before granting an IP address lease.

Alternatively, a different protocol could be used after an IP address lease but before packets are accepted by the NAT gateway. For example, an encrypted and signed MAC address could be sent to a machine on the local network on which it is installed, or the NAT gateway responsible for that network could accept the encrypted and signed MAC address and communicate with a MAC Address Validator before granting the opportunity to forward other packets. In other cases, a user may carry a portable smart card that can be used for authentication for use with any computer. In this case the actual MAC address used by the computer is not used for user authentication, but instead other identifying information that has been previously registered.

In some cases, the validation status generated by the MAC Address Validator is logged along with the lease information. A DHCP server, or other entity responsible to associating MAC addresses with IP addresses, could be modified to require additional information from a client computer and validate that the MAC address in use has been properly registered. Note that the NAT/DHCP Gateway need not know anything about the computer or have access to the registration information, but merely needs to know from the MAC Address Validator that the MAC address in use is valid. In this case, the Validation Status could be merely an affirmative result transmitted to a DHCP Server or NAT Gateway, which would then log an authorization code or an authentication string to prove that validation had been performed. In other cases, identification validation is done at the time an alias link is created and logged in connection with that information.

Conclusions

In this paper we have discussed how logging information can be generated sufficient to allow individual computers to be identified and associated with Internet activity. We also discussed how this information could be managed and

different scenarios in which the information is protected and controlled. We also presented a solution for how computer identification can be validated and how such a system can be designed to avoid eavesdropping, Man-in-the-middle and Replay attacks.

The idea behind endpoint identification is to guarantee that a reliable association can be made between activity on the Internet and a specific computer. This goal involves addressing abuse through accountability rather than protection, and thereby involves a fundamental tradeoff with privacy. While the anonymity of certain types of activity on the Internet is desirable and important, it is also desirable and important in many cases that those responsible for certain activity can be identified if necessary. The appropriate use of a carefully designed logging and validation system can appropriately balance these competing concerns. For example, information sufficient to identify endpoints can be maintained, while safeguards can be put in place to ensure that only in specific cases (such as a Court Order or Subpoena) would the information be made available. In another example, the information could be placed in the hands of an independent entity or organization, which would provide the information only under specific guidelines.

It is likely that the administrative and legal aspects of this problem are more challenging than the technical ones. The existence of multiple independent entities, under different legal jurisdictions, in different countries and with conflicting, or at least unaligned, interests greatly complicates the situation. Ultimately, however, the global community is better served by a system that can address abuse by allowing endpoints to be identified, as long as privacy can be adequately accommodated. Thus, there is a need for global cooperation to establish such mechanisms.

References

- [1] "Log Files," Apache HTTP Sever Documentation, The Apache Software Foundation, <http://www.apache.org>.
- [2] "Network Address Translation," FreeBSD Handbook, Section 31.9, The FreeBSD Project, <http://www.freebsd.org>.
- [3] *Link Layer*, W. Richard Stevens, TCP/IP Illustrated, Volume 1: The Protocols, Chapter 2, Addison-Wesley, 1994.
- [4] "Automatic Network Configuration," FreeBSD Handbook, Section 29.5, The FreeBSD Project, <http://www.freebsd.org>.
- [5] *Overview of 802.11 Networks*, 802.11 Wireless Networks: The Definitive Guide, Chapter 2, Matthew Gast, O'Reilly Media, Inc., 2005