Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 10, 2014

C. Huitema
Microsoft Corporation

November 06, 2013

Passive Traffic Analysis Threats and Defense
draft-huitema-perpass-trafficanalysis-00.txt

Abstract

Traffic analysis is used by various entities to derive "meta data" about Internet communications, such as who communicates with whom or what, and when. We analyze how meta-data can be extracted by monitoring IP headers, DNS traffic, and clear-text headers of commonly used protocols. We then propose a series of actions that would make traffic analysis more difficult.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The massive monitoring attacks that we know about seem to fall into three categories: listening to the content of communications in

transit, accessing content of documents and past exchanges at a server, and analyzing traffic to find patterns of communications and deduce social exchanges.

Other efforts address the "listening on conversations" attack, and how to prevent them with more or better encryption. There are some good ideas for reducing the risk of accessing contents on server, such as storing encrypted contents on servers, or enabling distributed services so that users can chose server locations that they find more acceptable. Enabling encryption will also reduce the capability to extract information from the e-mail or http headers. This draft focuses on a different set of threats, the monitoring and analysis of Internet Protocol headers to extract "metadata" such as the structure of social graphs or the timing of social events.

This draft proceeds by analyzing first the information that the monitoring entities desire to acquire and that privacy advocates would like to protect. These monitoring tools are expected to work for both IPv4 [RFC0791] and IPv6 [RFC2460]. We present then the mechanism of IP header monitoring, and discuss the critical problem of associating IP addresses to user identities. We then review a series of mechanisms that might be used to mitigate IP header monitoring.

2. Passive Analysis Targets

Questioned about revelation that his secret services were monitoring all the phone calls of the populace, a famous leader defended himself by saying that no, we don't listen to your phone calls, we merely gather "meta data." It turns out that meta data such as who called what telephone number and at what time is actually very valuable.

The first target of traffic analysis is the graph of connectivity within a given population. If we known that two phone numbers frequently call each other, we can infer that there is a relation between the owners of these numbers. For example, if investigative services discover a pattern of calls between an old general and some young lady, they can infer the existence of some inappropriate relation, and eventually force the general to relinquish his leadership position. Similarly, if we find a pattern of frequent calls between a small set of telephone numbers, we can infer the existence of some tight-knit network. Further analysis can then lead to the evaluation that these are just the members of the same family or the same sports team, or on the contrary it can find that these are political opponents organizing themselves, or maybe in rare cases some members of an underground criminal organization.

The graph of connectivity may sometimes take very simple forms. For example, visiting the web site of a banned organization may be sufficient to get flagged as a dissident by some autocratic regimes.

The second target of traffic analysis is the discovery of traffic surges, or the opposite, sudden absence of traffic indicating that a particular group has gone silent. If the monitoring of traffic reveals increased activity between a particular group, secondary analysis can be used to obtain more information on the activities of the group. That secondary analysis will be able to find the difference between a family preparing a birthday event, a sports team training for a particular competition, a group of activists planning a political protest, and maybe in rare cases a group of criminals planning some nefarious act.

Traffic can operate across multiple media. Analysis of phone calls reveals patterns between phone numbers, but similar analysis can be applied to IP addresses. Traffic analysis becomes much more valuable if the IP address can be associated with a personal email address or with a personal phone number. This correlation is also a target of traffic analysis.

Traffic analysis may also reveal the targets location. If the same user appears to connect to the Internet from a succession of IP addresses at different locations, the monitoring services can deduce the itinerary of that user.

For the defenders, the targets of traffic analysis become as many assets to be protected. In the following analysis, we will focus on ways to thwart discovery of the graph of connectivity, timing of activity, and correlation between identifiers.

3. Analysis of IP headers

Internet traffic can be monitored by tapping Internet links, or by installing monitoring tools in Internet routers. Of course, a single link or a single router only provides access to a fraction of the global Internet traffic. However, monitoring a number of high capacity links or monitoring a set of routers placed at strategic locations provides access to a good sampling of Internet traffic.

Tools like Cisco's NetFlow [RFC3954] allow administrators to acquire statistics about "sequence of packets with some common properties that pass through a network device." The most common set of properties is the "five tuple" of source and destination addresses, protocol type, and source and destination ports. These statistics are commonly used for network engineering, but could certainly be used for other purposes.

Let's assume for a moment that IP addresses can be correlated to specific services or specific users. Analysis of the sequences of packets will quickly reveal which users use what services, and also which users engage in peer-to-peer connection with other users. Analysis of traffic variations over time can be used to detect increased activity by particular users, or in the case of peer-to-peer connections increased activity within groups of users.

4. Linking IP addresses to user identities

In Section 3, we have assumed that IP addresses can be correlated with specific user identities. This can be done in various ways.

Tools like reverse DNS lookup can be used to retrieve the DNS names of servers. In fact, since the addresses of servers tend to be quite stable and since servers are relatively less numerous than users, we can expect that large scale monitoring services maintain databases of servers' IP addresses to facilitate such retrieval. On the other hand, the reverse lookup of users addresses is less informative. For example, a lookup of the address currently used by my home network returns a name of the form "c-xxx-xxx-xxx-xxx.hsd1.wa.comcast.net" in which the symbols "xxx-xxx-xxx-xxx" correspond to the IP address used by my home network. This particular type of reverse DNS lookup does not reveal much interesting information.

Traditionally, the police has relied on Internet Service Providers (ISP) to provide identification on a case by case basis of the "owner" of a specific IP address. This is a reasonably expedient process for police investigations, but large scale monitoring requires something more efficient. If the monitoring service can secure the cooperation of the ISP, they may obtain the link between identity and address through some automated update process. We may expect that some ISP will not willingly cooperate with large scale monitoring of their customers, in which case the monitoring entities have to rely on other methods.

Even if the ISP does not cooperate, identity can often be obtained by analyzing the traffic. We will discuss in the next section how SMTP and HTTP can leak information that links the IP address to the identity of the user.

4.1. Monitoring POP3, IMAP or SIP clients for identifying users of IP addresses

POP3 [RFC1939] and IMAP [RFC3501] are used to retrieve mail from mail servers, while a variant of SMTP [RFC5321] is used to submit messages through mail servers. The IMAP connections originate from the client, and typically start with an authentication exchange in which the client proves its identity by answering a password challenge.

If the protocol is executed in clear text, monitoring services can "tap" the links to the mail server, retrieve the user name provided by the client, and associate it with the IP address used to establish the connection.

The same attack can be executed against the SIP protocol, [RFC3261] if the connection between the SIP UA and the SIP server operates in clear text.

There are many instant messaging services operating over the Internet using proprietary protocols. If any of these proprietary protocols includes clear-text transmission of the user identity, it can be tapped to provide an association between the user identity and the IP address.

4.2. Retrieving IP addresses from mail headers

The SMTP protocol specification [RFC5321] requires that each successive SMTP relay adds a "Received" header to the mail headers. The purpose of these headers is to enable audit of mail transmission, and perhaps to distinguish between regular mail and spam. Here is an extract from the headers of a message recently received from the "perpass" mailing list:

```
Received: from xxx-xxx-xxx-xxx.zone13.example.org (HELO ?192.168.1.100?)
 (xxx.xxx.xxx.xxx)
 by lvpsyyy-yyy-yyy-yyy.example.net with ESMTPSA
 (DHE-RSA-AES256-SHA encrypted, authenticated);
 27 Oct 2013 21:47:14 +0100
Message-ID: <526D7BD2.7070908@example.org>
Date: Sun, 27 Oct 2013 20:47:14 +0000
From: Some One <some.one@example.org>
```

This is the first "Received" header attached to the message by the first SMTP relay. For privacy reason, the field values have been anonymized. We learn here that the message was submitted by "Some One" on October 27, from a host behind a NAT (192.168.1.100) that used the IP address "xxx.xxx.xxx.xxx." The information remained in the message, and is accessible by all recipients of the "perpass" mailing list, or indeed by any monitoring service that sees at least one copy of the message.

For monitoring services, such information is just plain candy. Monitor enough e-mail traffic and you can regularly update the mapping between IP addresses and individuals. Even if the SMTP traffic was encrypted, the monitoring service could still register to receive a copy of public mailing lists like "perpass," and then log the header fields.

Similar information is available in the SIP headers [RFC3261].

4.3. Tracking address use with web cookies

Many web sites only encrypt a small fraction of their transactions. A popular pattern was to use HTTPS for the login information, and then use a "cookie" to associate following clear-text transactions with the user's identity. Cookies are also used by various advertisement services to quickly identify the users and serve them with "personalized" advertisements. Such cookies are particularly useful if the advertisement services wants to keep tracking the user across multiple sessions that may use different IP addresses.

As cookies are sent in clear text, a monitoring service can build a database that associates cookies to IP addresses. If the IP address is already identified, the cookie can be linked to the user identify. After that, if the same cookie appears on a new IP address, the new IP address can be immediately associated with the pre-determined identity.

4.4. Tracking address use with network graphs

There have been many publicly reported instances in which the police managed to find the owner of a "disposable" cell phone. In theory this is hard, because there is no direct registration of the owner's identity. But in practice, the identity can be inferred through analysis of network graphs.

Suppose that the new owner of the cell phone uses it carelessly to call his mother, his brother, his boss and his preferred restaurant. Mother, brother, boss and restaurant are part of the "network graph" already collected by pervasive monitoring, and in fact constitute an almost unique signature of this particular individual. A quick database search and voila, the cell phone is identified.

The same approach can be applied to IP addresses. Users do a lot of repeat visits to web sites, mail servers, game servers, instant messaging servers, etc. These visits tend to follow time patterns. It is easy to imagine that if a particular pattern was seen from address "A" one day, and the same pattern from address "B" the next day, then A and B point to the same user, whose computer just got a

new address. At that point, the user may be identified only as a "case number," but the real identity can be filled as soon as email monitoring is successful, or sip monitoring, or maybe some ISP cooperation.

4.5. Static IPv6 interface identifiers

The IP Version 6 Addressing Architecture [RFC4291] suggests that "for all unicast addresses, except those that start with the binary value 000, Interface IDs... may have universal scope when derived from a universal token (e.g., IEEE 802 48-bit MAC or IEEE EUI-64 identifiers [EUI64])."

When implementors follow this recommendation, the IID part of the IPv6 address becomes a globally unique 64 bit number. Even if the IPv6 host moves to a new location, the IID will remain constant. Monitoring services can use that property to correlate IPv6 addresses belonging to the same host, and thus to the same user.

4.6. Stuff we have not thought off yet

The previous sections listed a number of known ways to extract identities from IP addresses. This is by no means an exhaustive list. There are certainly other possibilities, for example monitoring of public Wi-Fi networks and tracking of association between MAC addresses and IP addresses, or monitoring of various authentication services.

5. Defenses against IP header monitoring

In the current state of the Internet, defense against monitoring is very hard. There are many ways to associate IP addresses with user identity. Tapping of big Internet pipes is bound to provide a trove of data. Retrieving social graphs and detecting surges of activity is well within the means of a well funded monitoring service. But this does not mean that the Internet engineering community should just give up. Even if we cannot stop this monitoring completely, we can certainly make it harder and less reliable.

The first version of this internet draft presents a list of potential defenses that have been mentioned in various discussions. This list is not exhaustive, and is also not prioritized. It is merely a recollection of a number of suggestions.

5.1. Client server encryption

The previous analysis shows that IP traffic analysis is facilitated by the discovery of relations between IP addresses and users.

Encryption of the client-server protocols will deprive monitoring of this source of information.

The analysis was conducted for mail protocols (POP3, IMAP, SMTP) and for SIP. Encrypting these protocols is of course a priority. But if we want to really mitigate the threat of disclosing identity to address mappings, we should encrypt any protocol that carries a description of the user identity.

Encryption may not always completely remove the possibility to monitor connections. Encrypted traffic may still contain patterns, such as for example VOIP connections sending one RTP audio message every 20 ms. With elaborate pattern analysis, monitoring entities may be able to find user-specific patterns in the encrypted messages. Studying cryptographic protection against such analysis is probably beyond the scope of this document.

5.2. Clean-up E-mail headers

The email service is by itself a rich target for traffic analysis. Analyzing who sent mail to whom and when provides a rich set of meta data, even when the messages' contents are encrypted. This would probably justify a draft focusing just on email trafic analysis and protection. Since this document focuses on IP header monitoring, we will just point here a tiny problem related to discovery of linkage between IP and email addresses.

The initial "Received" field of e-mail headers carries the IP address from which the e-mail was submitted. This is equivalent to broadcasting the mapping between that IP address and the user identity. We should seriously consider the tradeoff between privacy and auditability that this feature afford.

A reasonable tradeoff could be to not publish the IP address or the domain name of the initial submitter, and to start the "Received" list with the IP address of the mail server. We should however consider the case where the first server is a "home" server, whose public IP address is the same as that of the user. Ideally, we should not publish that either.

The same reasoning should apply to any protocol that publishes a trace of successive server addresses in its headers. At some point, auditability should give way to privacy.

5.3. Source address obfuscation

Jon Crowcroft suggested a nice idea a few years ago, although for a different reason: sourceless network architecture [SNA]. Send

packets with no source address, and you make the metadata much less useful. (Of course, if the packet is to get a reply, the source address needs to be encrypted in the payload.)

The idea is largely theoretical, and would require significant changes in a number of widely deployed protocols, including TCP.

5.4. Network address translation

Many home networks use "network address translation" (NAT) [RFC3022] to share a single IPv4 address between several computers, and possibly several users. NAT are also used in some enterprise networks, and in some Wi-Fi "hot spots." Some ISP have also begun to use NAT, providing "private" addresses to their subscribers.

NAT complicates the task of IP header monitoring, because a particular address may be shared between multiple users. If the address is only shared between few users, like the members of a family sharing a home network, monitoring services can probably use analysis techniques to retrieve the individual connections, and NAT may not be more than a speed bump. If the sharing pool is much larger, like all the subscribers to a medium size ISP, monitoring becomes significantly harder.

5.5. IPv6 privacy addresses

It is ironic to notice that as IPv6 improves "address transparency" by removing the need for address translation, it also makes monitoring significantly easier than when using NAT. But the Privacy Extensions for Stateless Address Autoconfiguration in IPv6 [RFC4941] allow users to configure temporary IPv6 addresses out of a global prefix. Privacy addresses are meant to be used for a short time, typically no more than a day, and are specifically designed to render monitoring based on IPv6 addresses harder.

Privacy extensions only affect the least significant 64 bits of the IPv6 address. The most significant 64 bits remain unaffected. The 64 bit prefix is typically allocated to a small network, e.g., a single household or a Wi-Fi hot spot. It has pretty much the same identifying power as an IPv4 address. If the network is small in size, the use of privacy addresses, just like the use of NAT, will be a mere speed bump for IP header monitoring.

5.6. Frequent address renumbering

In the days of modem networking, a computer would receive a new IPv4 address each time it connected to the Internet. Always on broadband connections may or may not provide the subscribers with permanent stable addresses. Some users pay extra for the convenience of a stable address. Of course, stable addresses greatly facilitate IP header monitoring.

In contrast, we could imagine that the broadband modem is re-provisioned at regular interval with a new IPv4 address, or with a new IPv6 address prefix. Some convenience will be lost, and TCP connections active before the renumbering will have to be reestablished. However, the renumbering will significantly complicate the task of IP header monitoring.

5.7. Multihoming

Multihoming is the practice of using multiple connections simultaneously. If done well, multihoming will split the graph of connectivity in interesting ways. Packets will travel over different routes, IP addresses will be different. Multihoming could make IP header monitoring harder.

5.8. Virtual Private Networks

Virtual private networks (VPN) allow users to set up a "tunnel" across the Internet to a "virtual" connection point, and effectively provide a form of multihoming. Since the connections are virtual, VPN could also provide a form of frequent address renumbering. As such, VPN can provide some resistance against IP address monitoring.

VPN's require careful configuration and setup to prevent leakage of identifying information. Tech that purports to secure or privatize your communication but that actually leaks - or worse, can be coerced into revealing your traffic, is worse than no tech at all.

5.9. Web proxies

Sending HTTP requests through web proxy is a way to hide the actual IP source of the request, and as such a way to complicate monitoring.

Much like VPN, web proxies are a two edged sword. If the proxy is compromised, the true origin of the traffic can be retrieved. Moreover, the proxy could become an observation point to monitor the web traffic.

If the monitoring services can observe the traffic coming in and out of the proxies, they can use correlation methods to match incoming and outgoing flows. This is obvious in the trivial case where there is just 1 user of the proxy. The monitoring will reveal that a message to the proxy from address A was quickly followed by a message from the proxy to address B, and the monitoring service will infer a connection from address A to address B. Even if the number of proxy users increase, the monitoring services may still be able to use timing information and correlate input and output messages.

5.10. Onion routing and shuffle nets

Services like Tor provide an obvious form of resistance against IP header monitoring.

5.11. And there is more

There are certainly more potential defenses, which will emerge during the discussion of this draft.

6. Recommendations

The following recommendations are an attempt to summarize the threat and mitigation analysis in the previous sections:

- Use encryption. In particular, never send a user identity in clear text.

- Ask "submission" SMTP server to obfuscate the IP address of the user, and not place it in mail headers.

- Not completely written yet...

7. Security Considerations

This draft does not introduce new protocols. It does present a series of attacks on existing protocols, and proposes an assorted set of mitigations.

8. IANA Considerations

This draft does not require any IANA action.

9. Acknowledgments

The inspiration for this draft came from discussions in the Perpass mailing list. Some of the text was contributed in messages to the list by Dave Nix, Brian Trammel and Brian Carpenter. Stephen Farrell provided guidance and useful suggestions. Thanks to Linus Nordberg for checkproofing the draft.

10. References

10.1. Normative References

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

10.2. Informative References

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC1939] Myers, J. and M. Rose, "Post Office Protocol - Version 3", STD 53, RFC 1939, May 1996.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC3022] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC3501] Crispin, M., "INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1", RFC 3501, March 2003.

[RFC3954] Claise, B., "Cisco Systems NetFlow Services Export Version 9", RFC 3954, October 2004.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC4941] Narten, T., Draves, R., and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6", RFC 4941, September 2007.

[RFC5321] Klensin, J., "Simple Mail Transfer Protocol", RFC 5321, October 2008.

[SNA] Crowcroft, J. and M. Bagnulo, "SNA: Sourceless Network Architecture", June 2008, <http://www.cl.cam.ac.uk/~jac22/talks/sna.ppt>.

Author's Address

Christian Huitema
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
U.S.A.

Email: huitema@huitema.net