Attribute Requirements for Handling
Suspicious Illicit Downloading of Intellectual Property

Matt Bishop
Dept. of Computer Science
University of California at Davis
1 Shields Ave.
Davis, CA 95616-8562

Abstract

This document describes the attribution requirements using the example of responding to a "takedown letter" from a hypothetical company.

1 Introduction

"Attribution" is the association of data with an entity (person, process, file [2]. For example, authentication is a mechanism for attributing an identity to an entity, and is thus an example of attribution. In order to demonstrate the importance of attribution beyond its use in attribute-based access control (ABAC), itself a foundation of GENI, we considered five scenarios [1]:

1. Illicit downloading of music
2. DDoS attacks on GENI
3. Collecting and sharing data
4. Elections
5. Telephone-to-tweet service, as during the Arab Spring

As discussed, we selected the first scenario for examination because that describes a problem that the GENI Project Office, and member institutions, are encountering.

We emphasize that the use of an attribution infrastructure is at this point not possible because no such infrastructure exists. This document, and this approach, are intended as an example of how attribution might be used to assist in the investigation.

Our scenario is as follows. Some organization (for our purposes, the RIAA[1] claims that pirated music has been downloaded over a GENI resource. They serve an order to determine the culprit on the GENI Project Office, which in turn forwards the order to the university's information technology unit (ITU) department. The ITU then must identify possible perpetrators.

The source from which the intellectual property was downloaded may or may not be known. In the takedown requests, it usually is *not* given. This may be because the download was detected in such a way that the origin of the material could not be identified, or is suspect. Throughout this note, we assume it is not known. Were it known, an additional takedown order would be directed against the origin. We do note that the GENI node identified in the request may itself be a server; but that changes nothing that follows.

2 Overview of Approach

Takedown requests typically contain the following information:

[1] The Recording Industry Association of America and music downloads are used as exemplars here. The "takedown" requests may come from the Motion Picture Association of America, Random House, or indeed any holder of protected information. The material being requested to be removed may be music, movies, books, or any other protected material. The material may be protected by laws such as copyright, trade secret, or other laws.

- Name and contact information of complainant (for example, address, phone number, and email)
- Service provider (this may simply be a contact email such as "abuse@infringe.edu")
- Title of the song
- Timestamp of the infringement
- IP address of the host involved
- Type of infringement, for example BitTorrent
- Name of the file downloaded
- Size of the file downloaded
- Hash of the download (for example, of the Torrent)

As the download was performed on one or more GENI slices, the university must use the IP address to identify the slice(s) involved. From this, it then determines the creator of the slice(s), and from that it can determine the specific GENI project involved, which leads to the principal investigator. Note that this may (and probably will) require the university to work with the GMOC or some other GENI entity.

In more detail, the steps are as follows:

1. *Determine the slice(s) involved in the request.* Given the IP address, the slice(s) using that IP address at the time of download (infringement) must be identified. For expository purposes, we assume only one slice is using the IP address.
2. *Determine the GENI project using the slice.* This may (and probably will) require obtaining information from the aggregate operator for the campus; that entity in turn may need to work with the GENI Meta-Operations Center (GMOC) to obtain more detailed information about the entities involved in the project using the slice. Note that the experimenters may come from different universities, so the other academic institutions may need to be notified.
3. *Determine the responsible party for the GENI project.* This will require the assistance of the GMOC.
4. *Notify the responsible party of the takedown request.* At this point, internal procedures need to be followed, with the responsible party assisting to identify the downloader, or contradict the claim.

The attribution elements involved in this process can be derived from the above. The next section identifies those elements (and points out where the specific elements depend upon internal procedures). It also discusses the types of attribution required for each element.

3 Attribution Requirements

This section discusses the attribution elements required to provide the information leading to the identification of the possible downloaders, and the degree of assurance required for each element. We assume, for our purposes, that the attribution elements are accurately recorded by the recording mechanism, and that existing GENI infrastructure is augmented to collect this information. The problem of gathering this information and protecting it is akin to protecting the integrity of provenance data [3–5], and depends very much on the systems involved in GENI, and the GENI protection mechaisms. We leave this for future work. Given this, it is *critical* to note that these attributes should be seen as starting points for any investigation, and not proof *per se*.

An *attribute* consists of two components [2] that identify the *name* of the attribute and its *value*, and may be written as a pair $(name, value)$. Each attribution also has a type; for our purposes, the relevant types are:

- Perfect attribution, in which all the actors and systems are known to everyone with the requisite level of assurance; and
- Perfect selective attribution, in which all the actors and systems are known to a select group with the requisite level of assurance.

In what follows, we assume perfect selective attribution, in which those involved with GENI are given accurate information with the desired (presumably very high) level of assurance. But those not involved with GENI have no right to the information because making the attributes generally visible may compromise the confidentiality of some of the experiments.

We also write attributes informally. For example, associated with each IP address is an attribute; then the attribute for the host with IP address 10.0.0.1 is, precisely, ("IP address", 10.0.0.1). Suppose this host has a slice with slice identifier S17. The attribute associated with the slice is ("slice identifier", S17). Thus, the association of the attributes is:

$$((\text{``IP address''}, 10.0.0.1), (\text{``slice identifier''}, \text{S17}))$$

We simplify this notation by assuming the attribute names are given implicitly, and so would write this association as

$$(10.0.0.1, \text{S17})$$

But the reader should be aware of the need to have an underlying database that makes the above associations among attributes.

We consider each of the above steps separately.

3.1 Determine the slice(s) involved in the request

Here, we must go from the given network transfer data to the specific slice or slices with properties that could have performed the transfer being complained of. The network properties we have are the timestamp of the infringement, the IP address of the destination, the protocol involved in the infringement (which the letter terms "type of infringement"; see above), and the size of the file downloaded. The name of the file downloaded, and its hash, are properties of the *file*, not of the connection, and it may not be clear what parts of the transfer belong to a file. So we consider them separately; for now, we ignore them.

Given the IP address, we can determine which slices either transit or terminate at the suspect IP address. It is important to note that one could perform the transfer from an intermediate point, because a "slice" may involve general-purpose computers as infrastructure. The idea is that such a computer could copy the contents of the transfer for later "analysis" (playback).It may therefore enable users or administrators on that system with access to that slice to initiate the transfer.

In order to determine the slice involved, one must have the slice IDs that transit or terminate at that IP address. Thus, for each slice, we need the following:

$$(\text{IP address}, \{\text{sliceID}_1, \ldots, \text{sliceID}_n\})$$

Note that how this is actually stored, and where, is not relevant to our discussion. Indeed, it may be stored somewhere (at the GMOC, for example) as

$$(\text{sliceID}, \{\text{IP}_1, \ldots, \text{IP}_{m_{\text{sliceID}}}\})$$

What is critical is that one be able to obtain this information.

This may lead to one slice ID, in which case we would go to the next step. But if the IP address has multiple slices, we next try to narrow down the slices that could be involved using the traffic parameters of time, type, and size.

The timestamp is the time of the suspect download. Thus, the data was transferred beginning at that time. So, examine the amount of traffic over each slice that transited or terminated at the identified IP address. This means that each packet coming in and out of the suspect address must be tied in some way to the slice that produces it. Note that some packets may not be tied to any slice; we can ignore these, as the download is using GENI resources. Were the download not associated with a slice, it would not be associated with GENI.

Thus, each packet needs the following attributes:

$$(\text{packetID}, (\text{i/o}, \text{time}, \text{protocol}, \text{payload size}))$$

Here, "i/o" is a flag indicating whether the packet is entering the host ("incoming") or leaving ("outgoing"), "time" is the time the packet enters or leaves, "protocol" is the type of protocol (BitTorrent, HTTP, etc.), and "payload size" is the size of the *data* portion of the packet. Note we use a sequence rather than a set, so that we can implicitly include the names of the attributes (by the ordering of the attributes value).

When a slice is identified as potentially suspicious, the packet attributes associated with that slice and the protocol identified in the takedown request are gathered for all hosts associated with the slice, beginning with the time associated with the takedown request. A simple sum will probably eliminate many slices as not carrying enough data to have downloaded the content being complained about.

We now have a set of slices $\{\text{sliceID}_1, \ldots, \text{sliceID}_k\}$ that could be involved in the download.

3.2 Determine the GENI project using the slice

The next step is to find the principal investigators of the project. As each GENI slice is associated with a project, the slice must have an attribute of the associated project, as well as an attribute of the project's associated principal investigators. Thus, for each slice, we need the following:

$$(\text{sliceID}, \text{projectID})$$

From that, we can go into the set of project attributes:

$$(\text{projectID}, \text{P.I. information})$$

The "P.I. information" may involve additional attributes, depending on how that information is stored.

Next, we examine who was running an experiment on the slices at the time. In what follows, we consider only a single slice. The steps will be followed for each slice.

3.3 Determine the responsible party for the GENI project

The next question is to identify who in the project group was using the slice before or at the time of the suspect download. We need not worry about after the download, because we are interested only in who might have triggered it.

It is most likely that the download was triggered around the time of the timestamp on the takedown letter. However, it is also possible that someone set up a time-delayed process to use the slice to download the property.

Thus, we need to know who has accessed the slice and the timestamps of the process. So, associate with each user's access to a slice the following attributes: the identity of the user, the process(es) started, and the time each process started and ended:

$$((\text{userID}, \text{sliceID}), ((\text{processID}, \text{start time}, \text{end time}) \ldots))$$

Then the principal investigator can determine who was doing what at, or before, the time the download started.

3.4 Notify the responsible party of the takedown request

At this point, the principal investigator of the project running the slice has been identified, and all processes running on the slice at the time of the download (and before) have also been identified. Associated with these process identities are the user identities under which they ran. At this point, the principal investigator can simply begin contacting people to notify them of the request, or employ system-level forensics (for example, determining the owner of the file in question and seeing if that owner is one of the users in the "users" attribute of the slice.

3.5 Summary

From the above, the following attributes are required:

1. $(\text{IP address}, \{\text{sliceID}_1, \ldots, \text{sliceID}_n\})$
2. $(\text{sliceID}, \{\text{IP}_1, \ldots, \text{IP}_{m_{\text{sliceID}}}\})$
3. $(\text{packetID}, (\text{i/o}, \text{time}, \text{protocol}, \text{payload size}))$
4. $(\text{sliceID}, \text{projectID})$
5. $(\text{projectID}, \text{P.I. information})$
6. $((\text{userID}, \text{sliceID}), ((\text{processID}, \text{start time}, \text{end time}) \ldots))$

4 Conclusion

The need to obtain and track these attributes raises several issues that must be considered before any implementation and deployment is realized.

The first issue is the understanding that takedown requests do not arrive in "real time". Indeed, the slice used to download the offending material may be long gone when the letter arrives. In order to apply the above approach, one would need to keep a record of the attributes associated with the hosts, slices, packets, projects, and users even when the slice or project are no longer extant. This raises both storage issues and privacy issues. The former can be dealt with by applying some sort of a "time out" to purge old data. The latter cannot be easily handled.

The second issue is how to store the attributes. Centralized storage of some is quite simple; indeed, the GMOC stores information about projects and the resources allocated to them. But storing per-packet information at a central repository raises the privacy and storage issues, as well as performance ones: can a central site handle all the incoming attribute information? Local storage of some of the information, for example by making the packet information be stored by part of the underlying GENI control plane, may alleviate this problem globally—but it introduces the same problems on a per-host basis.

Finally comes the issue of performance. Currently, simply gathering these attributes and storing them will undoubtedly place a great burden on both the processing software and the storage systems of the hosts and GENI infrastructure. How to handle this, and indeed if it *can* be handled, is not known. It is an area ripe for future research.

On another note, Bishop *et al.* [2] point out that different research groups, institutions, and indeed governments have different degrees of attribution they will allow, ranging from no attribution to complete attribution. That these rules will affect GENI, especially as it spreads internationally, is clear. How it will affect the ability of GENI to capture this information, and how it will affect GENI's ability to use this information, is not at all clear.

References

[1] Matt Bishop, Mina Doroud, Jeffrey Hunker, and Carrie Gates. GENI attribution scenarios. Available at http://nob.cs.ucdavis.edu/attrib/scenarios.pdf, May 2012.

[2] Matt Bishop, Carrie Gates, and Jeffrey Hunker. The sisterhood of the traveling packets. NSPW '09, pages 1–12, New York, NY, USA, Sep 2009. ACM.

[3] Michael Factor, Ealan Henis, Dalit Naor, Simona Rabinovici-Cohen, Petra Reshef, Shahar Ronen, Giovanni Michetti, and Maria Guercio. Authenticity and provenance in long term digital preservation: Modeling and implementation in preservation aware storage. In *Proceedings of the First Workshop on the Theory and Practice of Provenance*, Feb. 2009.

[4] Carrie Gates and Matt Bishop. One of these records is not like the others. In *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance*, Berkeley, CA, USA, June 2011. USENIX Association.

[5] John Lyle and Andrew Martin. Trusted computing and provenance: Better together. In *Proceedings of the Second Workshop on the Theory and Practics of Provenance*, Feb. 2010.