Know Your Enemy:

Know Your Friends

(... and how you share data with them)

David Dittrich Chuck Costarella Tom Holt Jeff Nathan David Pisano
Adam Pridgen Lukas Rist Christian Seifert Camilo Viecco David Watson

January 7, 2013

Abstract

This is the abstract...

1 Introduction

Why share data?

1.1 Benefits of Sharing Data

Encouraging Cross-Disciplinary Research

More Efficient Use of Resources

Gaining New Insights Correlation of data.... Data mining... Visualization...

2 Issues with protecting data vs. code or publications

NOTE: ***Christian Seifert is looking into this section.***

The free software movement has been around for many years. There is a wide range of software from operating systems like Linux, to utilities and applications, to complete software suites. There are many types of open source licenses in use today (expand...)

2.1 Copyright does not cover data

2.2 Databases are not protected exactly like copyright

2.3 Existing mechanisms and their limits/applicability

2.4 Differences between international legal juridictions

Describe differences between privacy laws and data use restrictions across countries, including the penalties involved.
Wombat (`http://www.wombat-project.eu`) as an example of an EU FP7 funded project that includes techinical and proceedural/policy discussion for data sharing

2.4.1 Database right

http://en.wikipedia.org/wiki/Database_right

2.4.2 Database Directive

http://en.wikipedia.org/wiki/Directive_on_the_legal_protection_of_databases

3 Issues with data sharing in a confederated environment

3.1 Activities to be covered

NOTE: ***Adam Pridgen is looking into this section.***

Similar to documents and source code, there are numerous activities associated with data that must be considered when defining rights transfer.

3.1.1 Produce

Producing security data...

3.1.2 Consume

Security data has value in a number of ways from basic research through operational detection and response activities.

3.1.3 Replicate/Store

In order to make data available to a consumer, a producer must chose some mechanism for allowing access to that data. This can be a real-time feed, read-only access to stored files, or even access to a database management system (DBM) for remotely querying for portions of the data. The simplest means of sharing data is by replication of the data (i.e., copying it) and then storing that copy locally for use.

3.1.4 Perform or display

Performance or display of data is using that data in some presentation method such as a graph, a visualization, a real-time display, etc. Who is able to see the performance or display is controlled by attendance at a conference, making the display available on a web page, or creating a video that can be accessed on demand.

3.1.5 Modify

Prior to using data in a performance or display, it may be beneficial to modify the data in some way. This could be anonymizing elements of the data, or substituting one name for another. Producers of data who are known to the public may not want the data they provided to be modified in ways that reflect badly on them (e.g., substituting names or words in a way that could result in legal liability or negative publicity to the data producer.)

3.1.6 Distribute/redistribute

Once a consumer has acquired a copy of some data, replicating it for their own use, it is a very short step to further replication and potential redistribution to additional parties. This the where a large part of the conflict arises in data sharing.

3.2 Impediments to data sharing

NOTE: ***Tom Holt is looking into this section.***

There are several factors that limit the amount and type of sharing of computer security data. Many of these factors result from a lack of accepted ethical guidelines and their enforcement within the computer security field [2], allowing actors who lack integrity to take advantage of trust for their own personal gain.

3.2.1 Limitations and difficulties in anonymization

NOTE: ***David Watson will check with Sebastien, et al re: anonymization tools related to this section.***

Anonymization of data reduces the precision of data elements such that sensitive aspects like personal identifiers, sources of the data, or the existence of vulnerabilities that can be exploited, are rendered useless. The problem of anonymization, from a research perspective, is that the reduction in fidelity of the data can also render is useless for research. There are many cases where anonymization has been un-done and people or systems believed to be "hidden" in the data have been exposed [1, 3, 4].
As a result of these limitations, data sharing agreements may focus on tightly controlling the redistribution and/or performance and display rights.

3.2.2 Fear of competitive disadvantage from sharing

In a highly competitive field, such as the financial industry, knowledge about vulnerabilities in a competitor's services can be used to embarass that competitor and steal their customers. This risk alone may be viewed as greater than the risk associated with being compromised, hindering effective response or law enforcement involvement.

3.2.3 Fear of liability

There are many ways in which releasing information about how others are using the internet (either for good or bad) can incur liability on the part of the party sharing such data.

False accusation from false-positive reporting Data about computer intrusions and computer misuse often includes data that is associated with legitimate, innocent activity as well. Systems that are designed to identify and quarantine "malicious software" based on signatures often identify features that are shared by legitimate software. In both of these cases, someone whose analysis of the data yields a false positive determination of malintent may be called to answer for their determination by an innocent party who was incorrectly identified. For example, AV software often identifies factors (such as the use of compression, or *packing*) that in and of itself is not proof of hostile intent, even though the vast majority of malware may use a given packer. Or communications in a chat channel being monitored by a botnet researcher may include innocent third parties along with botnet controllers, and those innocent parties may have no knowledge of, nor involvement, in the malicious botnet activity.

Exposure of unsanctioned activities

3.2.4 Loss of control

Transitive trust when data is re-used

Authentication/Authorization/Access controls

Are discretionary access controls + trust good enough?

4 Comparison of data sharing agreements

NOTE: ***David Watson is looking into this section.***

4.1 GDH 1 and 2

4.2 Honeeebox

4.3 hpfeeds

4.4 Other security industry agreements...

- Agreement1
- Agreement2

4.5 Summary

A table that illustrates the elements described above is found in Table 1.

Cloak Mask	User `ezk`				User `joe`						Meaning for files J1–E10
	J1	J2	J3	J4	E5	E6	E7	E8	E9	E10	
+000											Show files to owners only
+007			A				A	A			Show files to owners and others
+070		A	A		A		A	A			Show files to owners and group members

Table 1: Here is a complex table that spans two columns. It shows how also to straddle the table cells.

5 Suggestions for trusted confederated data sharing

5.1 Metadata tagging to identify ownership and allowed rights

5.1.1 Traffic Light Protocol

5.2 Encryption in storage/transit

5.3 A workable data sharing agreement for multi-level trust?

5.4 A workable means of trusted sharing in hpfeeds

6 Conclusions

These are our conclusions...

References

[1] Mark Allman and Vern Paxson. Issues and etiquette concerning use of shared measurement data. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 135–140, 2007.

[2] D. Dittrich, M. Bailey, and S. Dietrich. Building An Active Computer Security Ethics Community. *Security Privacy, IEEE*, 9(4):32–40, July/August 2011.
[3] Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57:1701–2010, August 2009.
[4] Michael Zimmer. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, December 2010.