Data Sharing: Lessons learned by the CERT/CC and the CERT/NetSA groups

Roman Danyliw <rdd@cert.org>

FloCon 2004: Data Sharing Panel

CERT® Network Situational Awareness Group
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890

The CERT Network Situational Awareness Group is part of the Software Engineering Institute. The Software Engineering Institute is sponsored by the U.S. Department of Defense.
Background

• CERT/CC has a long history of accepting incident reports, artifacts, and vulnerability information
  – Synthesizing this input into public analysis such as advisories and the coordination of patch releases

• CERT/SA has experience in analyzing operational data-sets of other organizations
  – Synthesizing these data-sets to form situational awareness, and new analytical approaches
Decomposing “Data Sharing”

• Data collection
  – Accepting data from outside your organization

• Data dissemination
  – Providing value-add back to data sources or constituency

An organization only involved in data collection is not “data sharing”
Concerns in Sharing

• Concerns for the data source
  – Is anything “sensitive” being released?
    – If so, what assurances do I have about my data?
  – Is there sufficient benefit to me in providing this information?

• Concerns for the data recipient
  – Is there any risk in accepting this information?
    – Does the data source know it is a data source?
    – Can others know that this data source is being used?
    – What responsibilities do I have with respect to handling/sharing this information with others?
  – Is there sufficient benefit to collecting this information?
Steps in the Sharing Process

1. 3rd Party
2. SEI
3. 3rd Party
4. SEI
(1) I am reporting data to CERT

- Sharing data is technologically hard and requires human intervention
  - Few tools provide native support for sharing
  - CERT does provide tools to extract, filter, and sanitize information
- What guarantees do I have for my data?
  - Once data is handed over, all guarantees are founded on trust – no practical technological solution
  - Accreditation of processes, technology, and facilities
(1) I am reporting data to CERT (cont’d)

• “My information is sensitive, I want to protect:”
  – Information revealed in packet payloads
    – Contents of email, clear-text authentication
  – Internal topology of the network
    – Size and the purpose of individual hosts
  – Laxness or lapses in security
    – Outbound attacks
    – Usage of certain services (e.g., P2P)
    – Indications of vulnerabilities

• Often raw data is not possible; only share summaries
(2) CERT is receiving my information

- Willingness to share does not always mean utility for the CERT
  - Impossible to mechanically parse free-form text reports
  - Organizational or obscure data formats (i.e., vendor X with tool Y version Z.zzz.z)

- Employ standard data use policies
  - For all automated data sharing, a formal MOU governs the exchange
  - Public, default data disclosure policy for all self-reported data

- Public knowledge of honey-pot addresses is problematic
(2) CERT is receiving my information

- Community specific constraints
  - Academic community
    - Cannot tie data back to students
    - IP address resolved to host names which contained a student’s name
  - Federal community
    - Cannot collect Personally Identifiable Information (PII)
      - Only present in the payload
  - Medical community
    - HIPPA prevents PII collection
      - Only present in the payload
(3) CERT is disseminating information

- Does not provide attribution
  - Sometimes obfuscates results to do peer comparison
- Coordinating pre-release information requires a substantial volume of encrypted email
  - Dedicated tool (srmail) to handle encryption/decryption among various standards (e.g., gpg, pgp, s/mime)
- How to control the use of data after it is made available?
  - Contractors and federal government “rights to use” on pre-release information
  - Data leak through a 3rd party
  - Reaction of some open-source vs. COTS vendors to a vulnerability
(3) CERT is disseminating information

- Who is the right audience?
  - Traditionally, advisories were for system administrators – now have summaries for management
  - How to reach home users?
(4) I am receiving CERT information

- Optimal format for receiving information:
  - Semantics: push vs. pull
  - Transport protocol: email, web, etc.
  - Machine parsable vs. human readable

- How timely is the information?
  - Incomplete information, but early notification
    - Incremental updates
  - Complete information, but late notification
Observations in Data Sharing

• Datasets based on more sites is not always better – a representative sample is key
  – *Defining representative is hard*

• The community needs to develop and adopt standards formats and protocols to exchange analytical results
  – *Adoption by the vendor community will be required*

• Centralization is not desirable; expertise to analyze data is rarely found in one place – build a community of analysts
  – *The politics of data sharing make this hard*