Anonymizing Network Flow Data

Timothy J. Shimeall (tjs@cert.org)

January 2007

FloCon 2008

Overview

The balance of anonymization

Subnet-preserving

Subnet-collapsing

Host-preserving

Host-collapsing

Ports & Other issues

Conclusion

The Balance of Anonymization

Flow itself preserves some privacy by aggregation and eliding content.

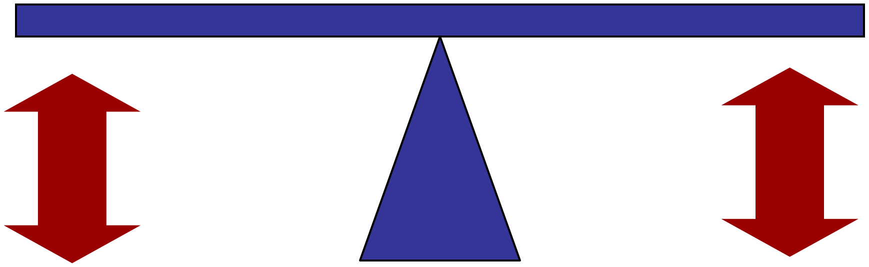
Anonymization is to aid in preserving the privacy of organizations represented in the data

- Data owner
- Partner or Customer
- Incidental
- Attacker

The more you anonymize the data, the less analyses can be done with it.

Need to explore a range of options

Privacy Utility



Subnet Preserving

Preserve host identity while concealing network.

How:

- Prepare list of networks
- Assign random substitution for network prefix
- Mask and replace prefix on each address
- Associative array works well for substitutions

Balance:

- Enables analysis down to host identity, but not organization identity
- Can be reversed by outside knowledge (server suffixes)

~~248.204~~.5.3

010.005.5.3

Subnet Collapsing

Conceal network structure and host identity, but preserve commonality of network

How:

- Reduce all address to the network
- Prepare random substitution for network
- Replace address with network substitutions

Balance:

- Allows network-level behavior analysis
- Might be reversed by organizations with lots of contact with data source

248.204~~.5.3~~

~~248.204~~.0.0

010.005.0.0

Host Preserving

Preserve host identity while concealing network commonality

How:

- Generate list of addresses
- Generate random substitution for each address
- Replace each occurrence with same substitution

Balance:

- Allows host-specific analysis
- Difficult to reverse

~~248.204.5.3~~

10.2.3.9

~~248.204.5.12~~

192.168.12.7

Host Randomizing

Do not preserve host or network identity (a.k.a., remove address content in any useful way)

How:

- Replace each occurrence of each address with random value
- Allow repetition of random values

Balance:

- Only permit analysis that does not involve address information
- Extremely difficult to reverse

~~248.204.5.3~~ ~~128.0.3.2~~

10.2.3.7 192.168.7.12

~~248.204.5.3~~ ~~0.5.4.1~~

192.168.17.37 10.2.3.7

Ports and Other Issues

There's more to anonymization of flow than addresses

- Network ports can be very revealing (OS fingerprinting)
- Timing information might be revealing
- TCP flags might be revealing (odd patterns)

Can anonymize this information:

- Ports: reduce to service, substitute; reduce to common/reserved/dynamic
- Timing: restart epoch; rescale timing; collapse interval
- TCP flags: reduce to function; remove OS-dependencies

Conclusion

Data sharing is difficult

Anonymization can be useful, but limiting

Anonymized does not mean private or irreversible