Detection and Analysis of Scans on Very Large Networks

FloCon 2004: Modeling Techniques Panel
July 21, 2004

Marc Kellner
Carrie Gates

CERT® Centers
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Sponsored by the U.S. Department of Defense

Needs Motivating this Approach

A comprehensive, integrated view of scanning activity across the network(s) of interest is needed to support situational awareness.

A historical record of network activity is needed to detect extremely low-intensity scans.

A historical record of identified scans is needed to study the evolving characteristics of Internet scanning.

Network defenders need support to

- **help identify higher-risk scans**
- **identify hosts at greater risk of compromise**
- **detect internal sources of scans**

Project Components and Initial Focus

The major thrusts of our effort are

- **scan detection**
- **scan database**
- **analysis of scans**

Our initial focus in scan detection is on

- **inbound traffic**
- **single source scans**
- **using TCP protocol**

As the scan database is populated, we will be able to commence analysis of the scans against the network(s) of interest.

Distinguishing Characteristics

Unlike most scan detection approaches, ours

- **is retrospective (not real-time)**
- **is based on flows (e.g., Cisco's NetFlow data)**
- **is multi-dimensional and extensible**
- **provides probability of traffic containing a scan**
- **supports long-term analysis of scanning activity**

Overview of Scan Detection Steps

Sort flow records by {source IP address (SIP), start time (stime)}

Identify the events (essentially clusters of traffic) for each SIP

Analyze each event (independently), for each SIP, to assess the probability it contains scanning activity

(Future) Combine traffic – not initially identified as scans – across time periods; analyze this combined traffic, for each SIP, to assess the probability it contains scanning activity

Scan Indicators

We compute several scan indicators for each event. These indicators are used to compute the probability that an event contains scanning activity.

Indicators are computed for

- **individual class C (/24) sub-nets (nets)**
  - **net coverage, net run length**
- **individual destination addresses (hosts)**
  - **low port coverage, low port run length**
  - **high port coverage, high port run length**
- **the event overall (event)**
  - **sub-net run length**
  - **flag combinations**
  - **packets per flow**
  - **unusual ports**

Applying Logistic Regression

Finally, we use the ten overall scan indicator values to determine how likely it is that an event contains scanning activity.

Let $I_1, \ldots, I_{10}$ represent the ten overall scan indicator values for any given event.

A "simple" logistic regression model using these variables predicts the probability (P) that this event contains scanning activity as

$$P = e^z / (1 + e^z) \quad \text{where}$$

$$z = ß_0 + (ß_1 * I_1) + (ß_2 * I_2) + \ldots + (ß_{10} * I_{10})$$

However, in order to apply this model we need to find the $ß$ values.

Model Estimation and Validation (1 of 2)

From a dataset of 155,827 events (reflecting 56M TCP flow records) we drew two samples:

- **for estimation of the model (120 events)**
- **for validation of the model (200 events)**

Each of the 320 sample events were independently classified by two experts as containing scanning activity or not. This was accomplished by examining the flow records without use of any scan detection tool or the scan indicator values.

These classifications were used as the "gold standard" against which the model was developed and validated.

Model Estimation and Validation (2 of 2)

Estimation sample:

- **drawn using a stratified sampling approach**
- **provided to a standard logistic regression program to estimate the β values for the model**
- **results: correctly classified all 120 sample events**

Validation sample:

- **drawn as a purely random sample from the dataset**
- **the logistic regression model was then used to classify each sample event**
- **results: correctly classified 197 of the 200 sample events**
  - **two false negatives (1.0% error rate)**
  - **one false positive (0.5% error rate)**

Results from the Sample Dataset

Running the scan detection system on the full dataset of 155,827 events yielded the following:

- **90,999 (58.40%) had prob. = 0.9 of containing scans**
- **64,288 (41.25%) had prob. = 0.1 of containing scans**
- **the remaining 540 events (0.35%) fell somewhere in between (i.e., > 0.1 and < 0.9)**

Using the customary probability of 0.5 as the threshold for a scan, led to classifying 91,381 (58.6%) of the total events as scans.

As points of reference, these events

- **contain 18,642,671 (33.1%) of the 56,344,051 TCP flow records**
- **came from 90,490 (17.7%) of the 511,602 unique source addresses sending TCP flows**

Scan Database

We have developed a database to record summary information about all detected scans. This will support the detection of distributed source scans and repetitive scans, as well as general analysis.

Vital information recorded in the database includes

- **scan source, start time, and end time**
- **all targeted destinations (i.e., {DIP, dport} pairs)**
- **size of the scan (in number of bytes, packets, flows, unique destinations, unique dips, unique dports, /24 subnets, and duration)**
- **indicator values from the scan detection program**
- **type of scan**
- **etc.**

Analysis of Scans

Based on information accumulated in the scan database

Determine appropriate and informative scan metrics and characteristics

Report results such as top scan sources, top scan targets (inside protected network), top ports scanned, average scan characteristics (e.g., intensity, scan rate, duration, number of different addresses scanned, number of different ports scanned, scan flow characteristics), how frequently an average target is scanned, repetitive scan sources, etc.

Planned Operational Capabilities

Phase 1: Provide capability to run scan detection and populate scan database on a routine basis

Phase 2: Provide basic access to scan database using a selection of pre-defined (but parameterized) queries

Phase 3: Provide scan information for lower tiers

- **identify scans that included any of the addresses of interest to that lower-tier organization**
- **facilitate investigating the scan from raw flow data**

Phase 4: Highlight hosts potentially compromised during a scan

Concluding Summary

The major thrusts of our effort are

- **scan detection**
- **scan database**
- **analysis of scans**

Unlike most scan detection approaches, ours

- **is retrospective (not real-time)**
- **is based on flows (e.g., Cisco's NetFlow data)**
- **is multi-dimensional and extensible**
- **provides probability of traffic containing a scan**
- **supports long-term analysis of scanning activity**

Scan database has been designed and implemented

Analysis capabilities will be provided for network defenders and will support scanning research goals