Modeling the Active and Idle Durations of Network Hosts

Soumyo Moitra
smoitra@cert.org
FloCon 2015
### Modeling the Active and Idle Durations of Network Hosts

#### Abstract
This material is based upon work funded and supported by SEI Line Funding under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

This material has been approved for public release and unlimited distribution.

The Government of the United States has a royalty-free government-purpose license to use, duplicate, or disclose the work, in whole or in part and in any manner, and to have or permit others to do so, for government purposes pursuant to the copyright license under the clause at 252.227-7013 and 252.227-7013 Alternate I.

This material was prepared for the exclusive use of FloCon 2015 attendees and may not be used for any other purpose without the written consent of permission@sei.cmu.edu.

DM-0001657
Introduction

Important to understand network behavior of hosts

Durations active and idle by host type

Patterns important for Situational Awareness

Baselining to detect anomalies

Decide whether a host should be in the inventory
Objectives of the Analysis

Distributions of the durations of active and idle times

Insights into different behaviors

Two metrics:

Probability of a host being active after a period of idleness

Conditional probability of a host becoming active within a time horizon
Given it has been idle for some time
Methodology

Flow data from the public domain
(http://tools.netsa.cert.org/silk/referencedata.html)

SiLK (CERT/SEI) and Unix Tools

Spreadsheets

Focus on web servers initially

Methodology applicable to all types of hosts
References

Analysis

Time series of network flows – out traffic

Time window = 23 hours

Time scale (bin size) = 1 hour

Convert volumes to a 0/1 series (1 => active)

Compute the durations of active and idle times

Plot the frequency distributions
## Durations from Flows (Hypothetical)

<table>
<thead>
<tr>
<th>Flows from rwcount</th>
<th>Conversion to 1/0</th>
<th>I</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>456</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>789</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Distribution of active durations

Distribution of idle durations
Discussion

Active durations
  Very compact (low variation – narrower than Poisson)
  Mean = 1.8
  Weibull?

Idle durations
  Long tail or two populations
  Issues with estimating the metrics
  Censoring/Truncation problems

Future Work
  Need much longer time series
  Need to estimate the metrics with more data sets
  Correct for biases
  Compare across different host types
  Effects of varying the time scales, time windows and time horizons
Thank you!

Questions/comments?