Modeling the Active and Idle Durations of Network Hosts

Soumyo Moitra
smoitra@cert.org
FloCon 2015

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JAN 2015		**00-00-2015 to 00-00-2015**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Modeling the Active and Idle Durations of Network Hosts	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Software Engineering Institute,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
FloCon 2015, Portland, OR, January 12-15, 2015.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

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

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

Bhattacharya, R. N. and Waymire, E. C. (2009) Stochastic Processes with Applications. SIAM.

Brostrom, G. (2012) Event History Analysis with R. CRC Press.

Crovella, M. and Krishnamurthy, B. (2006) Internet Measurement. John Wiley & Sons.

Hayden, L. (2010) IT Security Metrics. McGraw Hill.

Lawless, J. F. (2002) Statistical Models and methods for Lifetime Data. Wiley.

Maindonald, J. and Braun, W. J. (2010) Data Analysis and Graphics using R: An Example-Based Approach. Cambridge University Press.

Mills, M. (2011) Introducing Survival and Event History Analysis. Sage.

Rausland, M. and Heyland, A. (2003) System Reliability Theory. Wiley.

Ross,S. (2014) Applied Probability Models. Academic Press.

Snyder, D. L. and Miller, M. I. (2011) Random Point Processes in Time and Space. Springer.

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)

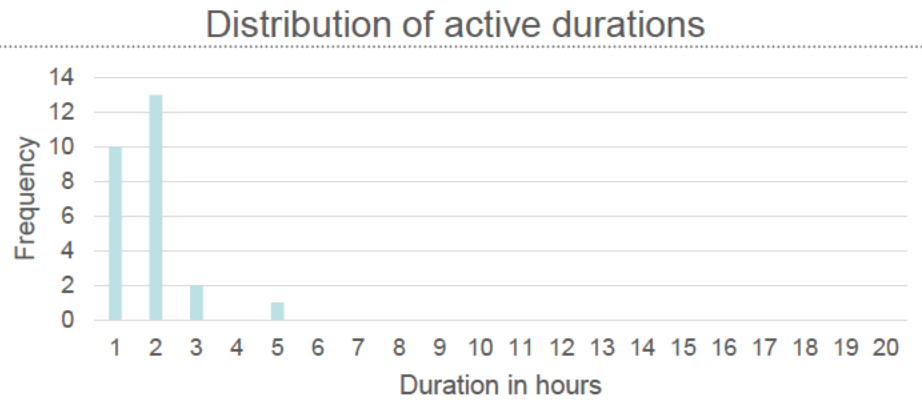
Compute the durations of active and idle times

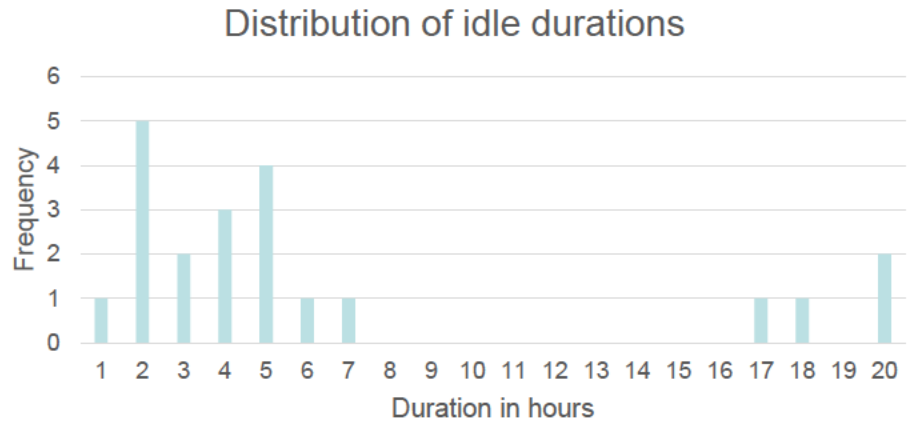
Plot the frequency distributions

Durations from Flows (Hypothetical)

Flows from rwcount	Conversion to 1/0	T	U
123	1		
456	1		
789	1	3	
0	0		
0	0		2
234	1		
90	1	2	
0	0		
0	0		
0	0		
0	0		4
55	1	1	
0	0		1
99	1		

Results


Distribution of active durations
Frequency
0
2
4
6
8
10
12
14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Duration in hours


Distribution of idle durations
Frequency
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Duration in hours

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?