Copyright © 2016 Splunk Inc.

Dynamic Population Discovery for Lateral Movement Detection

Rod Sotto & Joseph Zadeh
Splunk UBA Team

splunk>

INTRODUCTION

$Whoami…

- Joseph Zadeh

Senior Data Scientist with Splunk User Behavioral Analytics.

- Rod Soto

Senior Security Researcher with Splunk User Behavioral Analytics.

Outline

- What is lateral movement? – Rod
- Process, Objective & Tools used for lateral move - Rod
- Why is lateral movement detection important? - Rod
- What are behavioral profiles? What does it apply to? - Joe
- How do we establish assets (current available technologies) – Rod - Joe
- The importance of asset management – Rod - Joe
- Combination of assets classes to establish behavioral profiles and risk - Joe
- How can we automate discovery and data aggregation of DMZ assets - Joe
- Proof of Concept / Example – Joe - Rod
- Explanation of data science tools and techniques used for analysis – Joe
- How can this be applied to layer 4/7 data or PCAP data - Joe
- Conclusion - Rod - Joe

What is Lateral Movement?

- Lateral movement are series of actions conducted after a successful exploitation attack or infiltration in a organization's network that seeks to further reconnaissance and expand reach of attacker by gaining knowledge of internal network assets and accessing them.

Lateral Movement process Example

Objectives

- Main objectives of lateral move are:

  - Gain further knowledge of internal network assets.

  - Expand access into other systems

Ultimate goal is to get “Crown Jewels” which may be AD Domain admin credentials or access to valuable, sensitive information.

Tools

- Some of the tools used for lateral movement include:

  - Keyloggers, ARP spoofing, PwDump, Mimikatz

  - PsExec, WMI, PowerShell, Metasploit, sc, at, wmic, reg, winrs

  - RDP, SSH, VNC

  - Exploits (PTH/PTT), BruteForce tools

Why is lateral movement detection important?

- Let's talk about the concept of DWELL, or Detection Deficit (VZN)
- Rapid detection of lateral movement can reduce, contain and prevent further impact of a breach
- Detection of lateral move enables SOC/SECOPS/IR/DR teams to act in a more efficient manner
- Increases cost/deters attackers and would be external attackers, as well as insiders

How do we establish assets (current available technologies)

- *"In information security, computer security and network security an Asset is any data,device, or other component of the environment that supports information-related activities. Assets generally include hardware (e.g. servers and switches), software (e.g. mission critical applications and support systems) and confidential information.[1][2] Assets should be protected from illicit access, use, disclosure, alteration, destruction, and/or theft, resulting in loss to the organization."* *Wikipedia
- Anything that is network enabled inside the perimeter should be consider an asset, most common assets are: Network servers, Routers, Switches, Databases, Application Servers, Workstations, Printers, IoTs.

The Importance of Asset Management

- No asset management = No risk analysis
- Unmonitored unsupervised assets are likely to be targeted and exploited by attackers.
- Lack of OS/Application version/patch level increases risk of compromise
- Enables access management to resources inside the perimeter
- Enables SECOPS/IR/DS to identify and assess resources in case of incident

Use the right tools for the job

PRACTICAL ML FOR SECURITY

Adversarial Drift

- Current status quo, is driven by adversaries developing and introducing changes in their TTPs, bypassing all current detection technologies.

Spammers

Organised crime

Script kiddies/crackers

Hacktivists/pressure groups

Disgruntled employees

Foreign government agencies

The competition

Domestic government agencies

Advesarial Models

- Machine Learning Looses Effectiveness the more complex the adversary

Advesarial Models

Non-Automatable Actions: Hybrid Human/Computer Analysis

Automatable Actions: Good for ML

splunk> listen to your data®

Learning = Compression?

- There is a duality between learning and compression

Primary Key	Time	UserID	Count
Row 1	…	…	…
Row 2	…	…	…
Row 3	…	…	…
…	…	…	…
Row N	…	…	…

Input Data Total Size = 1 GB

Learned output is a set of "coefficients"

Total Output Size = 1K

Learning = Compression?

- Example of Linear Regression in R

```
> mtcars
                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

Learning Machine: R Linear Model

Learning = Compression?

- Train a model to predict mpg as a function of car weight, number of cylinders and displacement

```
> mtcars
                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

Learning Machine: R Linear Model

splunk> listen to your data®

Learning = Compression?

- Train a model to predict mpg as a function of car weight, number of cylinders and displacement

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1

Learning Machine: R Linear Model

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

splunk> listen to your data®

Learning = Compression?

- The overall input data is reduced in a “compressed form” to use in future predictions

```
> mtcars
                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

Learning Machine: R Linear Model

```
> (41.107678, -3.635677, 0.007473,-1.784944)
```

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

splunk> listen to your data®

Learning = Compression?

- This process is extremely brittle in terms of modeling a changing signal or an adversary that changes patterns over time

```
> mtcars
                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

Learning Machine: R Linear Model

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
wt          -3.635677   1.040138  -3.495  0.00160 **
disp         0.007473   0.011845   0.631  0.53322
cyl         -1.784944   0.607110  -2.940  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.595 on 28 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8147
F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11

> y = c_1 x_1 + c_2 x_2 + c_3 x_3 + b
```

Learning = Compression?

- The simple linear model gives us output that separates the **Signal** from the **Noise** (this is not always possible with a model)

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1

Learning Machine: R Linear Model

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
wt          -3.635677   1.040138  -3.495  0.00160 **
disp         0.007473   0.011845   0.631  0.53322
cyl         -1.784944   0.607110  -2.940  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.595 on 28 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8147
F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11

> y = c_1 x_1 + c_2 x_2 + c_3 x_3 + b
```

splunk> listen to your data®

Learning = Compression?

- Real example of random forest trained on C2 traffic

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
europartsplus.org	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

Learning Machine: MLLib Random Forest

```
Random Forest:
TreeEnsembleModel classifier with 6 trees

 Tree 0:
   If (feature 7 <= 0.0)
    If (feature 11 <= 0.0)
     If (feature 80 <= -51.45518112)
      Predict: 1.0
     Else (feature 80 > -51.45518112)
      If (feature 76 <= 7.0)
       If (feature 60 <= -48.02338409)
        If (feature 81 <= 0.0)
         If (feature 84 <= 2.0)
          Predict: 1.0
         Else (feature 84 > 2.0)
          Predict: 0.0
        Else (feature 81 > 0.0)
         Predict: 0.0
       Else (feature 60 > -48.02338409)
```

Learning = Compression?

- We really “learn” a function we can call in batch or real time

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
europartsplus.org	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

```
import org.apache.spark.mllib.tree.model.RandomForestModel
```

```
val rfModel: RandomForestModel = RandomForest.trainClassifier(trainData.union(cvData),
  numClasses, categoricalFeaturesInfo, loopNumTrees, featureSubsetStrategy, loopImpurity, loopDepth, loopBins)
```

```
//decorate the (domain,feature vector) pair with predicted output of the model
//use pair RRD's to recover the full data we need for anomaly generation by a inner join
val predictedOutput = vectorizeFeatures.map(x => (x._1, randomForestModel.predict(x._2)))
```

When is a model ready?

A model stabilization algorithm:

splunk> listen to your data

“But all too often we forget the first rule of battle - the battlefield – the attacker can escape everything it cannot escape the terrain – choose the terrain, use the terrain – we win” Sun Tzu

SECURITY ANALYTICS FOR DEFENSE

High Level Objectives

- Asset Class Discovery
  - Identify all things acting like device type “X”
- Identify key services/assets in the DMZ
- Identify human / non human by device
- Anomalies on rare paths
  - U->S
  - S->U
  - U->U (LAN to LAN)
  - S->S (DMZ to LAN)
- Identity Resolution Impossible Mappings

Modeling Methodology

- Step 1: Identity Resolution
- Step 2: Topology Discovery
- Step 3: Behavioral Profiles
- Step 4: Client/Server Relationship Discovery
- Step 5: Monitor for changes in asset relationship graph

Asset/Service Dynamic Discovery

Asset/Service Dynamic Discovery

Semi Aggregates

MAC	R2L	L2L	L2R	R2R	Total Cnt
00:AA:00:BB:AA	0%	90%	10%	0%	

MAC/IP	Application Distribution (Dynamic)		Fixed Apps		Port/Protocol Distribution	
	Dynamic Counts		Windows Update	Java	Dynamic Counts	
00:AA:00:BB:AA			5%	20%		

Device (MAC/IP)	Domain Distribution (layer 7)		IP Distribution (layer 4)		Num of Unique External/ Internal Devices	Weekly Access Heatmap
	Dynamic Counts		Dynamic Counts			
00:AA:00:BB:AA						

IP	Fixed Event IDs (possibly can just use Meis model here)	
10.10.0.1		

MAC	Identity Resolution Checks		
	IP	Cnt	NodeDegree
00:AA:00:BB:AA	10.10.0.1		

IP	Identity Resolution Checks
10.10.0.1	00:AA:00:BB:AA

MAC	Time resolution roll ups
00:AA:00:BB:AA	

Device	Top 10 Apps		Top 10 domains	
	Dynamic Counts			
10.10.0.1				

Seeing the Analytic In Action

Seeing the Analytic In Action

Seeing the Analytic In Action

- Once identity resolution/learning process is complete we create new anomalies based on new paths/actions that are rare for a particular population profiel

GARBAGE IN GARBAGE OUT

STEP1: IDENTITY RESOLUTION

Identity Resolution

- Many possible ways to attack the identity resolution problem with enterprise solutions but this usually has complexity
- Smaller scale shops should leverage work already done here - SIEM is a good example a tool that normalizes lots of these scenarios
- Advanced Pattern – Inventory Based Trust : Usenix 2016 "BeyondCorp: Design to Deployment at Google"

SECURITY

BeyondCorp: Design to Deployment at Google

Figure 2: Device Inventory Service

ID Resolution WORKFLOW

splunk> listen to your data®

ETL Online Mode: Raw Individual Streams

Incremental load: Prioritizing updates to state table in real time

1. Assign priority to data streams for automated ETL of daily/weekly/ incremental updates
2. Update Active ID Table before any other workflow task begins

ETL Online Mode: Raw Individual Streams

DHCP State Table

IP_Address	Hostname	MAC	LastLease_Timestamp
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T13:00:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T14:00:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T22:30:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-11T09:00:00
10.100.1.23	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-11T14:00:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T10:00:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T14:30:00
192.168.1.65	scott.hr.acme.com	00:50:a6:d2:21:01	2014-03-10T14:30:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T17:30:00
192.168.1.65	scott.hr.acme.com	1b:31:a5:1d:b0:11	2014-03-11T14:50:00
10.13.11.221	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-12T14:30:00

1. Drop all tuples not containing Event ID = 673, EventID = 4663
2. ID data extractor for keeping only key data points necessary for AD State table

ID Res Event ID Filter

AD State Table

AD_IP	Username	FQDN	Event_Time
10.10.50.25	dave	dave.eng.acme	2014-03-10T13:00:00
10.10.50.25	dave	dave.eng.acme	2014-03-10T14:00:00
10.10.50.25	dave	dave.eng.acme	2014-03-11T09:00:00
10.100.1.23	dave@acme.com		2014-03-11T14:00:00
10.5.12.2	scott	scott.hr.acme	2014-03-10T10:00:00
192.168.1.65	scott@acme.com		2014-03-10T14:30:00
10.13.11.221	scott		2014-03-12T14:30:00
192.168.1.65	scott	scott.hr.acme	2014-03-11T14:50:00

ETL Online Mode: Real Time Active State Table

IP_Address	Hostname	MAC	LastLease_Timestamp
10.10.50.25	steve.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T13:00:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T14:00:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-10T22:30:00
10.10.50.25	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-11T09:00:00
10.100.1.23	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-11T14:00:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T10:00:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T14:30:00
192.168.1.65	scott.hr.acme.com	00:50:a6:d2:21:01	2014-03-10T14:30:00
10.5.12.2	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-10T17:30:00
192.168.1.65	scott.hr.acme.com	1b:31:a5:1d:b0:11	2014-03-11T14:50:00
10.13.11.221	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-12T14:30:00

AD_IP	Username	FQDN	Event_Time
10.10.50.25	dave	dave.eng.acme	2014-03-10T13:00:00
10.10.50.25	dave	dave.eng.acme	2014-03-10T14:00:00
10.10.50.25	dave	dave.eng.acme	2014-03-11T09:00:00
10.100.1.23	dave@acme.com		2014-03-11T14:00:00
10.5.12.2	scott	scott.hr.acme	2014-03-10T10:00:00
192.168.1.65	scott@acme.com		2014-03-10T14:30:00
10.13.11.221	scott		2014-03-12T14:30:00
192.168.1.65	scot	scott.hr.acme	2014-03-11T14:50:00

Real Time Identity Table

IP	DHCP.hostname	DHCP.MAC	DHCP_Lasteventtime	AD_username	AD_FQDN	AD_Lasteventtime
10.100.1.23	dave.eng.acme.com	58:5c:35:c3:6e:a4	2014-03-11T14:00:00	dave@acme.com	dave.eng.acme.com	2014-03-11T14:00:00
10.13.11.221	scott.hr.acme.com	12:3a:74:b2:6a:22	2014-03-12T14:30:00	scott	scot.hr.acme.com	2014-03-12T14:30:00
10.131.1.4	admin	NULL	NULL	domain_admin	acme.com	2014-03-12T23:00:00

Primary Key

splunk> listen to your data®

Learning the Local Layers

STEP 2: TOPOLOGY DISCOVERY

Map the lower layers of the OSI model passively

- Infer key properties
  - DMZ blocks (often times we find new segments this way)
  - LAN only blocks
  - VLAN behavior (Student VLAN, ADMIN VLAN, STAFF VLAN)
- Keep in mind we loose visibility into switched traffic flows (layer 2 is hard to see at scale)

Basic Features Built in this Step

- Graph Features
  - Source/Destination behavior
    - How many hosts talk to this IP? (In Degree)
    - How many hosts are talked to by this IP? (Out Degree)
- Layer 2/Layer 3 Features
  - IP Subnet Behavior
    - # LAN to LAN conversations (non routable IP flows)
    - # LAN to WAN conversations (non routable address to routable routable address)

Graph Features: Example

- Example: Frequency of serving as Destination Device vs Source Device.

Table: Device "10.131.16.51"

	History	Current-day
#Times as Destination	38	372
#Times as Source	83	50

  - Hypothesis: new observed data distribution = historical distribution
  - Chi-square statistics= 160.68 (df=2-1=1)
  - p-Value< 2.2e-16. Conclude that distribution of pairs (destIP, srcIP) based on history vs. new day are different.

STEP 3: BEHAVIOR BASED PROFILING

Asset Fingerprints

- Goal is to use machine learning, vanilla/fuzzy correlation to discover some common asset classes (ML term sometimes is class labels)

Desktop

Laptop

MS Server

*nix Server

IOT

Biomedical Devices

Energy Meters

What are behavioral profiles? What does it apply to?

- Windows 2008/2003 Server Profile
  - Flow Characteristics:
    - Byte distribution ratios are asymmetric
  - Application Layer Characteristics
    - SMB, Netbios MS-Update
    - Number of unique domains per day

- Windows 2008/2003 End Device Profile
  - Flow Characteristics:
  - Application Layer Characteristics:
    - Facebook Chat, social media, twitter,
    - Non uniform browsing patterns

- *nix Server/End Device Device Profile
  - Flow Characteristics:
  - Application Layer Characteristic
    - Software updates for distros (ubuntu, rhel)

Representing a Profile Over Time

Comparing with entity history:

Comparing a Profile to a Group

tes Sent Out to External in Total & Compared with Group
embers.

Application Fingerprints

- Instead of using raw counts in the histogram, we use total bytes per application as weights.
- Hypothesis: new observed distribution of bytes out by application = historical distribution of bytes out by application
- An example of contingency table for device "172.31.6.70":

	History	Current-Day
dns	7194071	1313328
apple-push-notifications	48317	3
apple-maps	152980	19324
good-for-enterprise	553771	18460
icloud-base	1483250	2944784
web-browsing	40815	35102
ssl	306477	380370
itunes-base	133537	49260
facetime	0	621
gmail-base	0	64179
ntp	0	78058
icloud-mail	0	1064740
constant-contact	0	8028

- Chi-Square Test statistics= 5846500
- Degree of freedom=12
- P-value$\leq$ 2.2e-16.

Layer 7 Info

- Not always possible to build these kind of statistics without higher layer application data or PCAPs

- Example for Device "10.131.16.5" on Day 14 with its own history:

Table: Raw Counts

	History	Current-Day
ms-ds-smb	7268360	7093270
ldap	1204522	711268
msrpc	11019549	9019549
kerberos	287060	227019
active-directory	231385	49233
dns	15840	17003
ntp	3244	1360

- Left: Chi-square statistics=228050 with p-value<2.2e-16;

splunk> listen to your data®

ML/Stat Workflow Engine

STEP 4: CLIENT/SERVER RELATION DISCOVERY

Its all in the Bytes

- Depending on what type of visibility you have you can leverage certain levels of granularity
  - Flows (Netflow v 7) you get number of packets per flow very important
  - PCAPS best case scenario but hard to log/process at scale for large environments
  - Higher layers might get a loss of signal

- Example: Bytes Sent Out in Total.

- Hypothesis: new observed data distribution = historical distribution
- Kolmogorov Smirnov statistics=Max Deviance=0.15385
- p-Value = 0.9996

Histogram of Byte Distribution

- In terms of "total bytes out to external devices per application", h
different device "172.31.6.70" on day 11 as compared with its owr
history?

Group Based Comparisons

Vertical Time Series Comparion

- In terms of "total bytes out to external", how different is device "172.31.11.151" as compared with other devices in the same group?

“At this point all the hard work is done”…

STEP 5: MONITOR FOR CHANGES IN ASSET RELATIONSHIP GRAPH

Mining For Relationship Anomalies

- Anomalies on rare paths
  - U->S
  - S->U !!
  - U->U (LAN to LAN)
  - S->S (DMZ to LAN)!!

Beyond the Indicator

Seeing the Analytic In Action

Seeing the Analytic In Action

Seeing the Analytic In Action

- Once identity resolution/learning process is complete we create new anomalies based on new paths/actions that are rare for a particular population profiel

Conclusion - Rod - Joe

- New approaches in machine learning and data science can help improve lateral movement detection.
- Establish behavioral patterns based on data driven approaches can provide tools for detecting and predicting unusual, high risk and malicious behavior patterns in users and use of assets.

- We have been successful catching webshells and new kinds of in memory malware using the rare path approach
  - U->S
  - S->U !!
  - U->U (LAN to LAN)
  - S->S (DMZ to LAN)!!

Q&A

- Thank you

- Rod Soto @rodsoto
- Joseph Zadeh @josephzadeh

APPENDIX

Model Value vs. Total Cost of Validation and Impact Risk

Model Value vs. Total Cost of Validation and Impact Risk

Model Value vs. Total Cost of Validation and Impact Risk

Key to ML: Label Your Analysis

- This is how the algorithms will "learn" from human expertise and help support a common security workflow

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa	Outcome
yyfaimjmocdu.com	144	6.05	1	1	0	0	Malicious
jjeyd2u37an30.com	6192	5.05	0	1	0	0	Malicious
cdn4s.steelhousemedia.com	107	3	0	0	0	0	Benign
log.tagcade.com	111	2	0	1	0	0	Benign
go.vidprocess.com	170	2	0	0	0	0	Benign
statse.webtrendslive.com	310	2	0	1	0	0	Benign
cdn4s.steelhousemedia.com	107	1	0	0	0	0	Benign
log.tagcade.com	111	1	0	1	0	0	Benign

Human Expertise is manually encoded into a format computers understand: Sometimes this process is called Labeling or "Truth-ing" the data

Lambda Architecture

- Architecture is described by three simple equations:

batch view = function(all data)

realtime view = function(realtime view, new data)

query = function(batch view, realtime view)

splunk> listen to your data®

Lambda Security

Lambda Security

Lambda Security

Lambda Security

splunk>
DHCP
IMS/IPAM
AD
FW
Proxy
VPN
Data Ingest
APACHE STORM
Real Time Identity Resolution
IP | DHCP.MAC | DHCP_Lasteventtime | AD_FQDN
10.100.1.23 | 58:5c:35:c3:6e:a4 | 2014-03-11T14:00:00 | joe.eng.acme.com
10.13.11.221 | 12:3a:74:b2:6a:22 | 2014-03-12T14:30:00 | ad.hr.acme.com
Distributed ETL
Username = select coallesce(user_name, hostname, IP) from Active_ID_Table where IP = '10.10.100.23)
Sequential Models and IOC's
Hybrid View (Batch + Real Time)
Large Scale Models and Non-Sequential IOC's
Spark
Data-Parallel
Table
Row
Result
GraphX
SELECT date, COUNT(*) FROM product GROUP BY date
Map
Reduce

Automated process to accelerate workflows like Splunk Query to retrieve PCAP for further analysis combined with automatic VT/heuristic correlations

ML + Sequencing the Security DNA

- We parallelize across many nodes (JVMs) and use both real time and batch computations

splunk> listen to your data

Time t_0

Command and Control (C2) traffic has been established between “Beachead” and command and control operator

Time t_0

Heartbeat traffic signals C2 operator that infected asset is up and ready for instructions

Time t_1

Obfuscated instructions get returned through an **Upstream** conversation embedded in PHP, .js, Flash, etc..

Commands obfuscated in this way can be through of as a hidden “Downstream Beacon”

Destination	Protocol	Length	Info
37.59.5.67	HTTP	260	POST /sears/sears.php HTTP/1.0
37.59.5.67	HTTP	260	POST /sears/sears.php HTTP/1.0
37.59.5.67	HTTP	260	POST /sears/sears.php HTTP/1.0

Embedded commands can signal infected asset to enumerate local information on the machine, attach to open network shares and perform lateral reconnaissance and privilege escalation throughout the compromised network

Time t_2

After targeted lateral movement and privilege enumeration all cases of targeted attacks eventually involve the compromise of the directory services roots servers (Usually AD Forest Roots) and exfiltration of key personnel information along with any

Time t_n

Edge weights can be encoded with key security features to increase overall model accuracy regardless of the underlying algorithms

BFS/DFS + Other classic graph search algorithms are a great examples of algorithms useful in detecting this graph signature

How can we automate discovery and data aggregation of DMZ assets - Joe

Proof of Concept / Example – Joe - Rod

Explanation of data science tools and techniques used for analysis – Joe

How can this be applied to layer 4/7 data or PCAP data - Joe

Copyright © 2016 Splunk Inc.

Time t_0
IP: 66.253.41.67

Command and Control (C2) traffic has been established between compromised hosts inside the corporate network and C2 servers

splunk>

Copyright © 2016 Splunk Inc.

Command and Control (C2) traffic has been established between compromised hosts inside the corporate network and C2 servers

splunk>

Copyright © 2016 Splunk Inc.

C2 Infrastructure changes locations of command and control server new communication path is established

splunk>

Copyright © 2016 Splunk Inc.

Time t_1
IP: 47.99.1.63

C2 Infrastructure changes locations of command and control server new communication path is established

splunk>

Copyright © 2016 Splunk Inc.

Time t_2
IP: 210.2.13.22

http://en.wikipedia.org/wiki/Fast_flux: Fast flux is a DNS technique used by botnets to hide phishing and malware delivery sites behind an ever-changing network of compromised hosts acting as proxies. It can also refer to the combination of peer-to-peer networking, distributed command and control, web-based load balancing and proxy redirection used to make malware networks more resistant to discovery and counter-measures. The Storm Worm is one of the recent malware variants to make use of this technique.

Copyright © 2016 Splunk Inc.

Time t_2
IP: 210.2.13.22

At each time step (typically a day or two) the C2 Infrastructure changes locations of command and control via this “Fluxing” behavior. A subset of these type of graph patterns is known as “Fast Fluxing”

splunk>

Copyright © 2016 Splunk Inc.

The constant mobility of command and control infrastructure will continue this IP/Domain fluxing movement until detected

Copyright © 2016 Splunk Inc.

The constant mobility of command and control infrastructure will continue this IP/Domain fluxing movement until detected