Improved Hunt Seeding with Specific Anomaly Scoring

Brenden Bishop

January 8, 2019

New presentation who dis?

New presentation who dis?

- My formal training was in quantitative psychology and statistics at The Ohio State University, graduated 2017

New presentation who dis?

- My formal training was in quantitative psychology and statistics at The Ohio State University, graduated 2017
- Started at Columbus Collaboratory, working on a variety of projects, quite a bit of prototyping

Columbus Collaboratory

New presentation who dis?

- My formal training was in quantitative psychology and statistics at The Ohio State University, graduated 2017
- Started at Columbus Collaboratory, working on a variety of projects, quite a bit of prototyping

Columbus Collaboratory

- Love cyber projects because, by and large, one can actually measure all the stuff required to answer the question

Hunting

Hunting

- Hunting has become an integral component of mature cyber security operations

Hunting

- Hunting has become an integral component of mature cyber security operations
- Network defenders spend a portion of their time *hunting* for vulnerabilities, misconfigurations, or previously unnoticed security events

Hunting

- Hunting has become an integral component of mature cyber security operations
- Network defenders spend a portion of their time *hunting* for vulnerabilities, misconfigurations, or previously unnoticed security events
- The practice has evolved beyond grepping randomly through logs

Hunting

- Hunting has become an integral component of mature cyber security operations
- Network defenders spend a portion of their time *hunting* for vulnerabilities, misconfigurations, or previously unnoticed security events
- The practice has evolved beyond grepping randomly through logs
- Hunts can now be seeded using ML/AI/Statistical models, leading to a directed search rather than a random walk

Sounds simple enough, but...

Sounds simple enough, but...

Challenges

Frequent challenges when finding anomalies:

Challenges

Frequent challenges when finding anomalies:

1. "Find anything strange on the network" is not sufficiently specific

Challenges

Frequent challenges when finding anomalies:

1. "Find anything strange on the network" is not sufficiently specific (neither is "Find any lateral movement.")

Challenges

Frequent challenges when finding anomalies:

1. "Find anything strange on the network" is not sufficiently specific (neither is "Find any lateral movement.")
   - Statistics requires problem identification, consideration of available variables, and understanding how observations arise

Challenges

Frequent challenges when finding anomalies:

1. "Find anything strange on the network" is not sufficiently specific (neither is "Find any lateral movement.")
   - Statistics requires problem identification, consideration of available variables, and understanding how observations arise
2. Cyber and statistics/data science folks can talk past one another

Challenges

Frequent challenges when finding anomalies:

1. "Find anything strange on the network" is not sufficiently specific (neither is "Find any lateral movement.")
   - Statistics requires problem identification, consideration of available variables, and understanding how observations arise
2. Cyber and statistics/data science folks can talk past one another
3. Unsupervised learning is prone to a high false alarm rate; Machine Learning/Artificial Intelligence/Automated-Inference are not immune

Addressing challenges

Addressing challenges

1. Scope problems appropriately (e.g. Find strange outbound connections to cloud storage.)

Addressing challenges

1. Scope problems appropriately (e.g. Find strange outbound connections to cloud storage.)
2. Cyber and statistics/AI/ML experts must iterate collaboratively; interdisciplinary teams are optimal for innovation

Addressing challenges

1. Scope problems appropriately (e.g. Find strange outbound connections to cloud storage.)
2. Cyber and statistics/AI/ML experts must iterate collaboratively; interdisciplinary teams are optimal for innovation
3. Turn big data into managable data, and, where possible, turn unsupervised problems into supervised. Collect data and validate models

Addressing challenges

1. Scope problems appropriately (e.g. Find strange outbound connections to cloud storage.)
2. Cyber and statistics/AI/ML experts must iterate collaboratively; interdisciplinary teams are optimal for innovation
3. Turn big data into managable data, and, where possible, turn unsupervised problems into supervised. Collect data and validate models (practice security as a science)

Addressing challenges

1. Scope problems appropriately (e.g. Find strange outbound connections to cloud storage.)
2. Cyber and statistics/AI/ML experts must iterate collaboratively; interdisciplinary teams are optimal for innovation
3. Turn big data into managable data, and, where possible, turn unsupervised problems into supervised. Collect data and validate models (practice security as a science)

- The remainder of the talk essentially focuses on item three

Good news everyone

Good news everyone

- Cyber security data is particularly well suited to statistical inference

Good news everyone

- Cyber security data is particularly well suited to statistical inference
  - Logs are typically a census of network activity, we have the population

Good news everyone

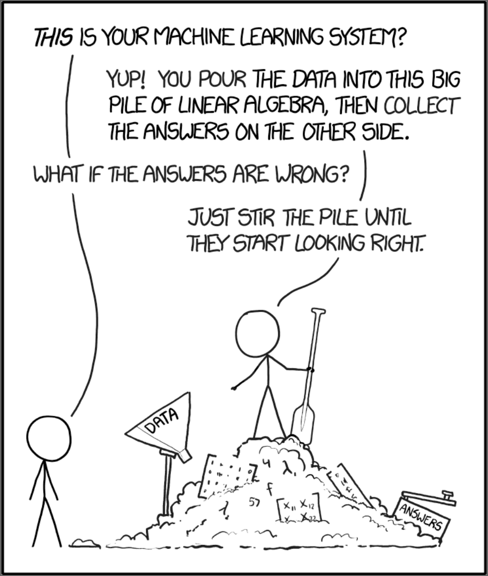
- Cyber security data is particularly well suited to statistical inference
  - Logs are typically a census of network activity, we have the population
- Probability measures offer single-number summaries of all available information; anomalies are events with low probability

Good news everyone

- Cyber security data is particularly well suited to statistical inference
    - Logs are typically a census of network activity, we have the population
- Probability measures offer single-number summaries of all available information; anomalies are events with low probability
- Building an anomaly scoring model is tantamount to estimating a probability distribution

Good news everyone

- Cyber security data is particularly well suited to statistical inference
  - Logs are typically a census of network activity, we have the population
- Probability measures offer single-number summaries of all available information; anomalies are events with low probability
- Building an anomaly scoring model is tantamount to estimating a probability distribution
- Models can be validated during the course of regular hunting


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Some fundamentals

1. Network activity can be quantified (e.g. time, bytes sent, bytes received, protocol, connection type)

Some fundamentals

1. Network activity can be quantified (e.g. time, bytes sent, bytes received, protocol, connection type)
2. Quantified information can be stored in a numeric matrix with each row representing a single multivariate observation

Some fundamentals

1. Network activity can be quantified (e.g. time, bytes sent, bytes received, protocol, connection type)
2. Quantified information can be stored in a numeric matrix with each row representing a single multivariate observation
3. The observations are realizations from some probability distribution

Some fundamentals

1. Network activity can be quantified (e.g. time, bytes sent, bytes received, protocol, connection type)
2. Quantified information can be stored in a numeric matrix with each row representing a single multivariate observation
3. The observations are realizations from some probability distribution
4. Anomalies are aberrant rows, from low-density regions

Estimation

- Statisticians have been improving density estimation for around a century

Estimation

- Statisticians have been improving density estimation for around a century
- Kernel density estimators allow nonparametric estimation of any p dimensional probability distribution

Estimation

- Statisticians have been improving density estimation for around a century
- Kernel density estimators allow nonparametric estimation of any p dimensional probability distribution
- Though in practice, whenever p is larger than about 5 estimation can become quite burdensome

Estimation

- Statisticians have been improving density estimation for around a century
- Kernel density estimators allow nonparametric estimation of any p dimensional probability distribution
- Though in practice, whenever p is larger than about 5 estimation can become quite burdensome
- One promising approach that circumvents this effective dimensionality constraint is the use of vine copulas

Vine copulas in a nut shell

Vine copulas in a nut shell

- Copulas can partition multivariate densities into the product of their marginals and a component which captures all dependencies

Vine copulas in a nut shell

- Copulas can partition multivariate densities into the product of their marginals and a component which captures all dependencies
- Vine copulas split the dependency portion into p(p-1)/2 bivariate copula densities, decoupling convergence speed and dimension

Vine copulas in a nut shell

- Copulas can partition multivariate densities into the product of their marginals and a component which captures all dependencies
- Vine copulas split the dependency portion into p(p-1)/2 bivariate copula densities, decoupling convergence speed and dimension
- tl;dr One can estimate complicated multivariate distributions fairly accurately and quickly

Scoring

- Possessing an estimate of a distribution allows for the evaluation of the estimated density for novel values

Scoring

- Possessing an estimate of a distribution allows for the evaluation of the estimated density for novel values
- One can assign a probability to each record log and sort low probability events to the top

Scoring

- Possessing an estimate of a distribution allows for the evaluation of the estimated density for novel values
- One can assign a probability to each record log and sort low probability events to the top
- The most rare events can be given to a hunter, beginning iterative evaluation of the model

Raw data

- We'll use a subset of publicly available data from Kent [2015]

Raw data

- We'll use a subset of publicly available data from Kent [2015]
- The full data represents 58 consecutive days of events from Los Almos National Laboratory corporate, internal network (csr.lanl.gov/data/cyber1/)
- Data is de-identified, even the time variable

Raw data

- We'll use a subset of publicly available data from Kent [2015]
- The full data represents 58 consecutive days of events from Los Almos National Laboratory corporate, internal network (csr.lanl.gov/data/cyber1/)
- Data is de-identified, even the time variable
- Say one is looking for anomalous, successful authentication events

```
1,C625$@DOM1,U147@DOM1,C625,C625,Negotiate,Batch,LogOn,Success
1,C653$@DOM1,SYSTEM@C653,C653,C653,Negotiate,Service,LogOn,Success
1,C660$@DOM1,SYSTEM@C660,C660,C660,Negotiate,Service,LogOn,Success
```

Wrangle data and analyze

Wrangle data and analyze

- Dummy code login-type and authentication-type factors, and engineer other desired features

Wrangle data and analyze

- Dummy code login-type and authentication-type factors, and engineer other desired features
- Wrangled data set is 13 dimensional binary

Wrangle data and analyze

- Dummy code login-type and authentication-type factors, and engineer other desired features
- Wrangled data set is 13 dimensional binary
- Employ a continuous convolution to allow for kernel density estimation

Wrangle data and analyze

- Dummy code login-type and authentication-type factors, and engineer other desired features
- Wrangled data set is 13 dimensional binary
- Employ a continuous convolution to allow for kernel density estimation
- Use the `kdevine` or `vinecopular` R libraries to estimate the density

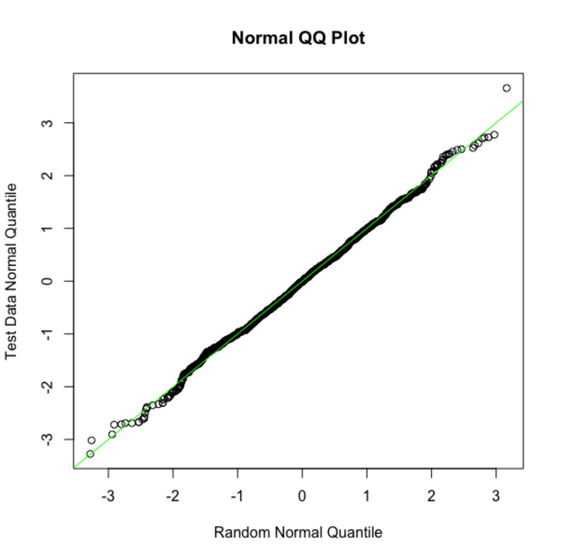
Just that easy

```r
vinedat <- dat[sample.int(nrow(dat), 10e3), -c(1:5)]
vinedatcc <- cctools::cont_conv(vinedat)
dest <- kdevine(vinedatcc, xmin = rep(-.5, 13),
                xmax = rep(1.5, 13), cores = 6)

scoreObs <- function(X){out <- cbind(X, dkdevine(X, dest))}

f <- sort(rep_len(1:2000, length.out = nrow(datcc)))
datcclist <- lapply(unique(f), function(i){datcc[f == i, ]})
outlist <- parallel::mclapply(datcclist, scoreObs, mc.cores = 5)
scored <- do.call("rbind", outlist)

results <- dat[, 1:5] %>%
  mutate(lpd = log(scored[, 14])) %>%
  arrange(lpd)
```


Normal QQ Plot
Test Data Normal Quantile
Random Normal Quantile
-3
-2
-1
0
1
2
3

Were you talking just now?

- With minimal investment, defenders can easily build probability models for any logs they want, not bound by existing tools

Were you talking just now?

- With minimal investment, defenders can easily build probability models for any logs they want, not bound by existing tools
- Models be generated on the fly, one-offs for a given hunt

Were you talking just now?

- With minimal investment, defenders can easily build probability models for any logs they want, not bound by existing tools
- Models be generated on the fly, one-offs for a given hunt
- Models can be refined/tuned as hunters check examine outputs and iterative development continues

Were you talking just now?

- With minimal investment, defenders can easily build probability models for any logs they want, not bound by existing tools
- Models be generated on the fly, one-offs for a given hunt
- Models can be refined/tuned as hunters check examine outputs and iterative development continues
- If at some point a model is found to have a satisfactory hit-rate, the anomalies are interesting, then one create an automatic detector

Thank you, kindly.

Columbus Collaboratory

Collaborative Security

K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

A. D. Kent. Comprehensive, Multi-Source Cyber-Security Events. Los Alamos National Laboratory, 2015.

T. Nagler. Kernel methods for vine copula estimation. 2014.

T. Nagler. A generic approach to nonparametric function estimation with mixed data. *Statistics & Probability Letters*, 137:326–330, 2018.

T. Nagler and C. Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.