

Statistical Approaches to Mining Multivariate Data Streams

Eric Vance, Duke University

Department of Statistical Science



David Banks, Duke University

Tamraparni Dasu, AT&T Labs - Research

**JSM July 31, 2007
Salt Lake City, Utah**



Data Streams

- Huge amounts of complex data
- Rapid rate of accumulation
- One-time access to raw data



Change Detection

- Problem: Change detection in complicated data streams
- Three criteria
 - Nonparametric
 - Fast
 - Statistical guarantees



E-Commerce Server Data

- **Data description**
 - One server in a network of servers
 - Data polled every 5 minutes
 - 27 variables
 - 5 week time period
- **Quality issues**
 - Many variables unimportant or unchanging
 - Missing data



E-Commerce Server Data

- **Variable Elimination**

- Several variables remain constant (Total Swap)
- Predictable and non-informative (Used SysDisk Space)
- Correlated (CPU Used%, CPU User%)

- **6 Variables Selected**

- CPU Used%
- Number of Procs
- Number of Threads
- Ping Latency
- Used Swap
- $\text{Log Packets} = \log(\text{In Packets} + \text{Out Packets})$



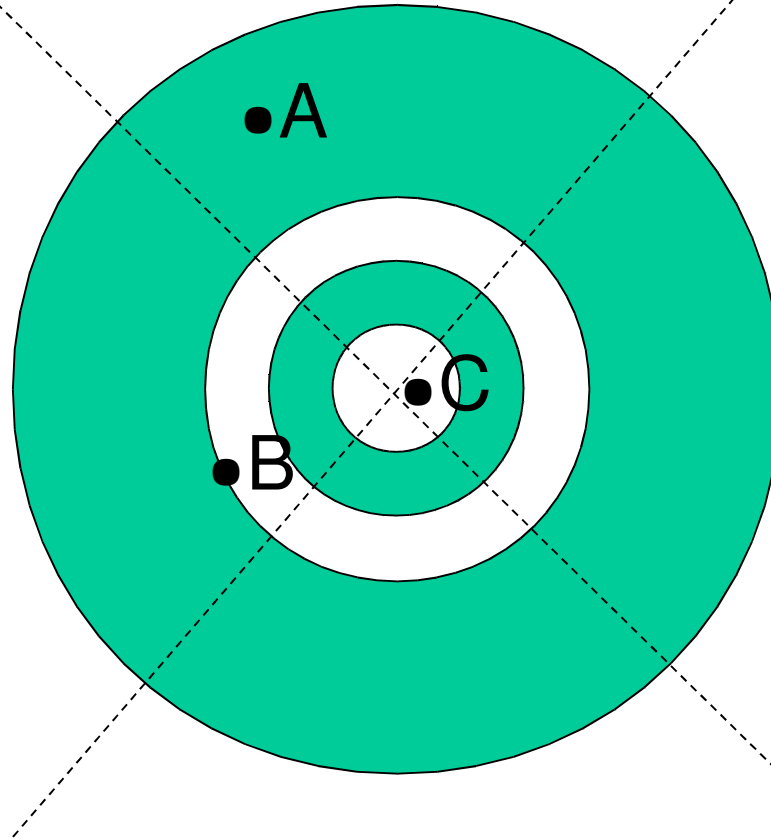
Our Approach

- Partitioning scheme (Multivariate Histogram)
 - Data depth: Rank each point in relation to Mahalanobis **distance** from center of data
 - Data Pyramid: Determine which **direction** in the data is most extreme
- Profile based comparison
 - Identify changes in profiles over time (week to week)



Partitioning Example in 2D

- 5 “center-outward” depth layers
- 4 pyramids



- A: depth 4 pyramid +y
- B: depth 3 pyramid -x
- C: depth 1 pyramid +x



Identify Depth and Direction

- Compute center of comparison Data Sphere \bar{X}
- Calculate Mahalanobis distance for each point x_i

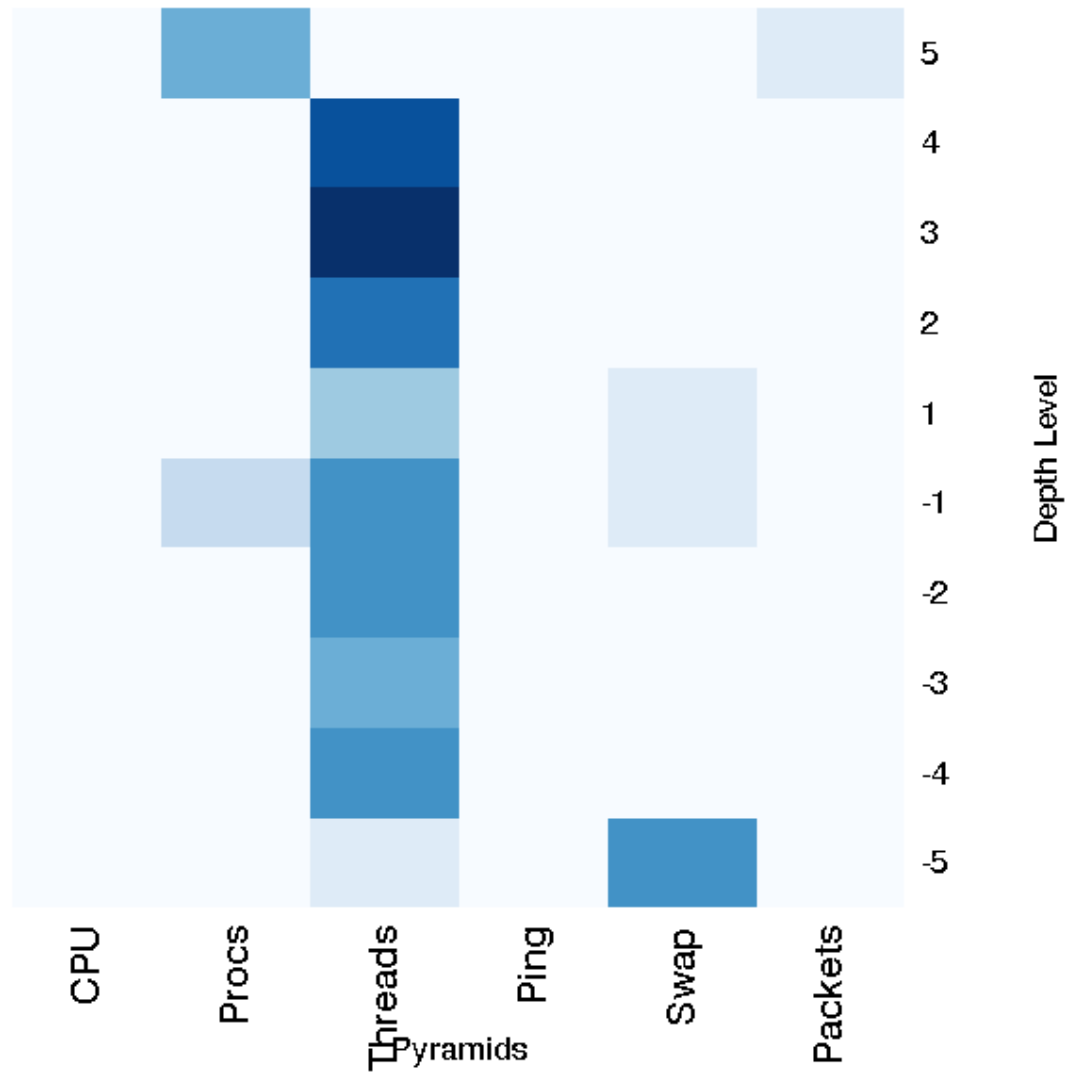
$$M_i = \sqrt{(x_i - \bar{X}) S^{-1} (x_i - \bar{X})}$$

- In which of the 5 quantiles of depth is M_i
- Determine direction of greatest variation for x_i



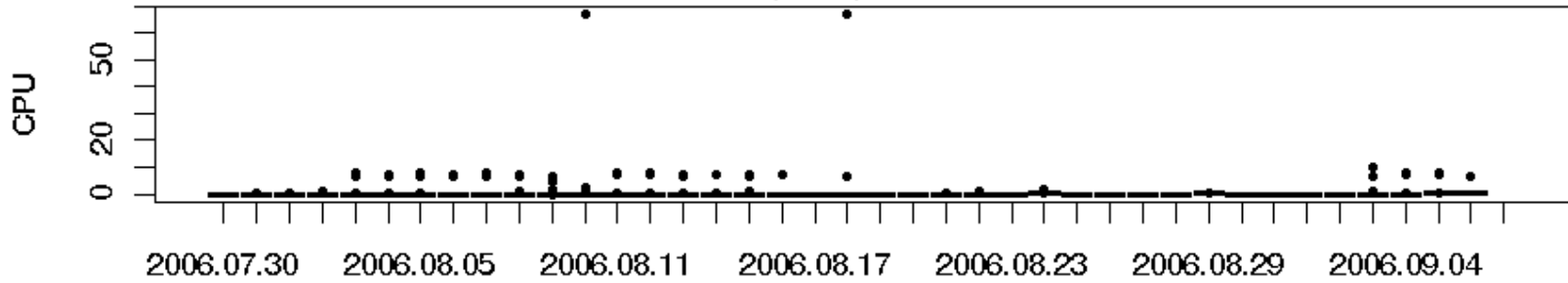
Data Partition in 6 Dimensions

Week1 v. Week1: Trim=0

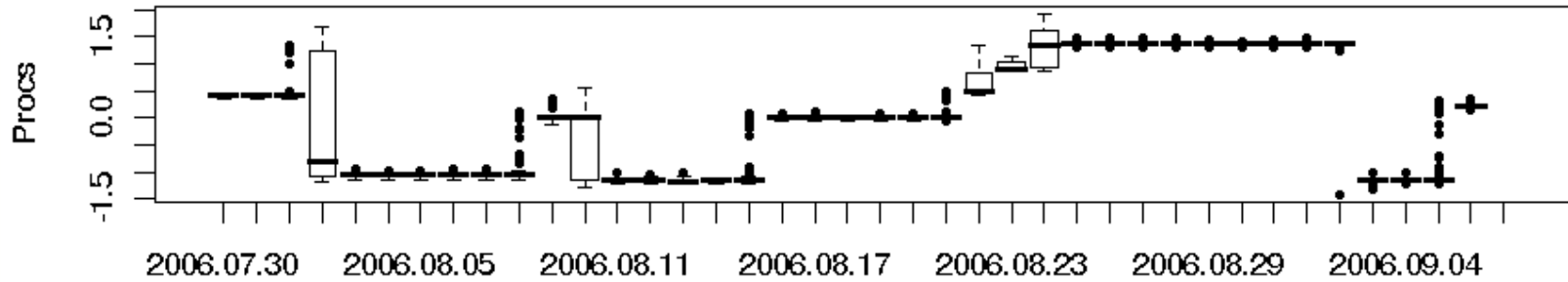


Daily Boxplots

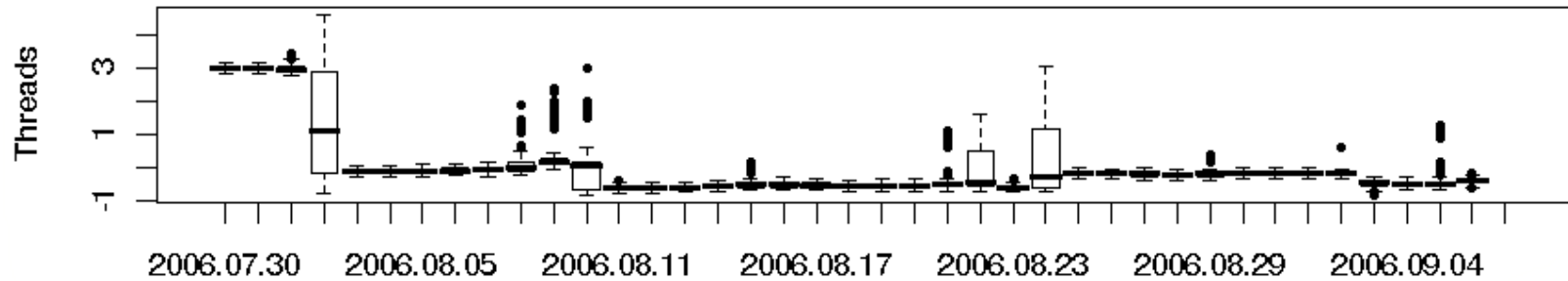
CPU



Procs



Threads

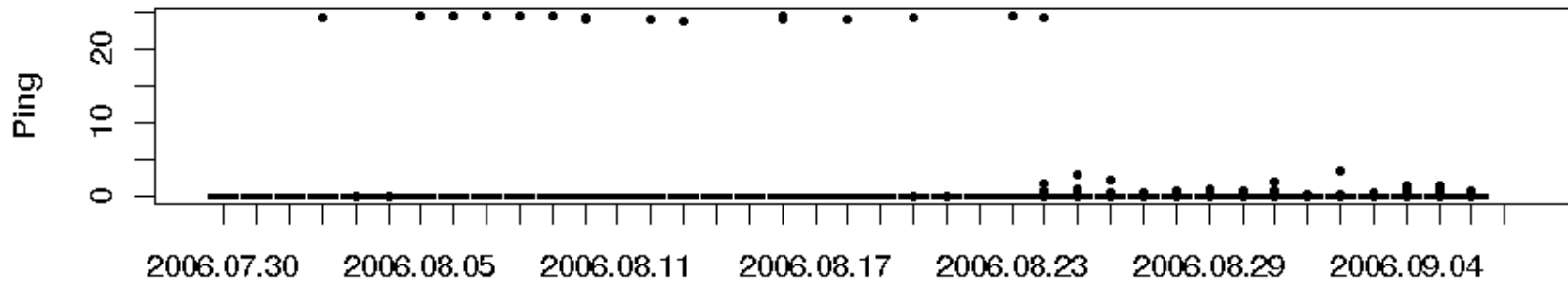


Date

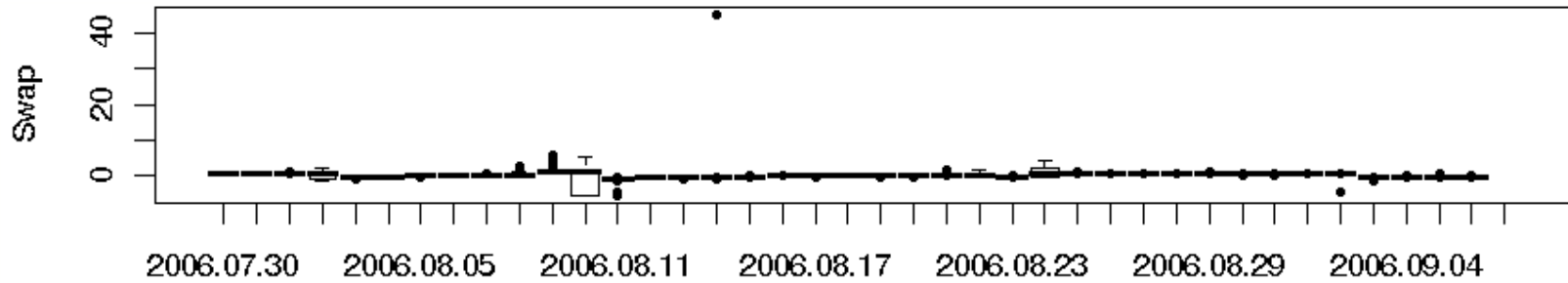


Daily Boxplots

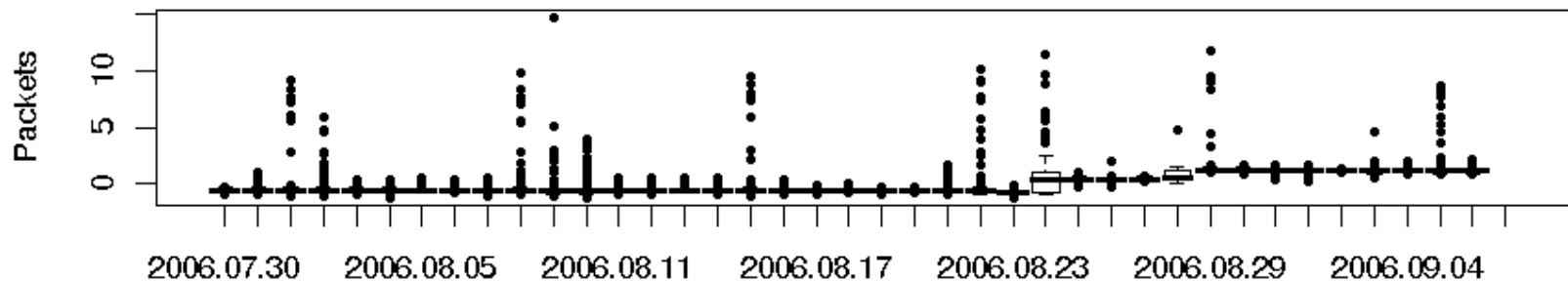
Ping



Swap



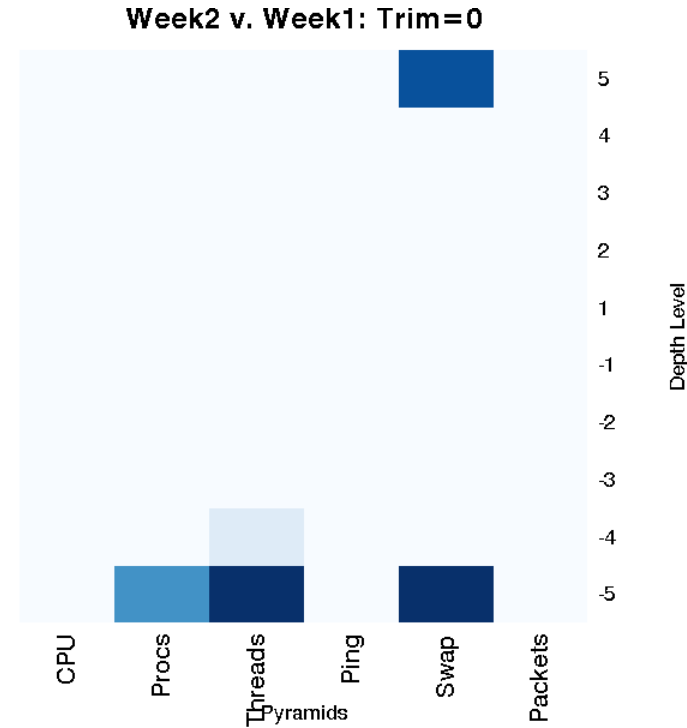
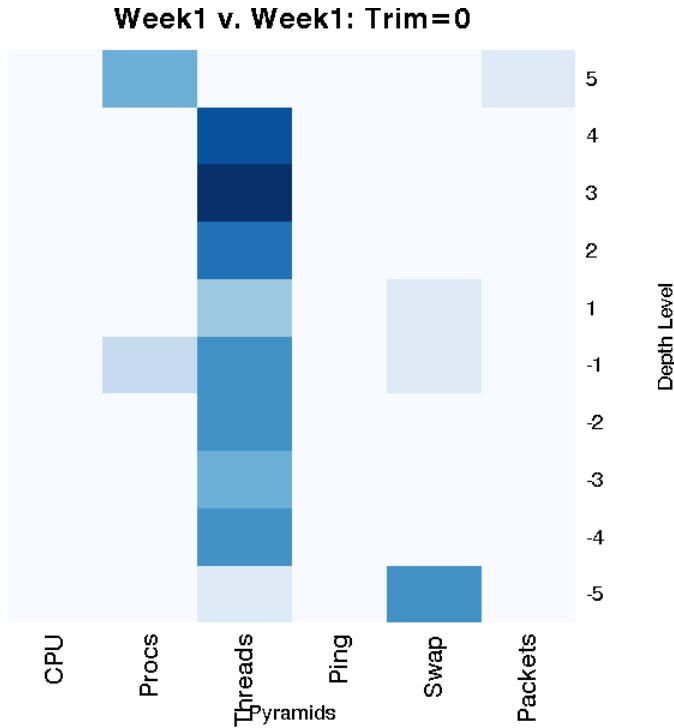
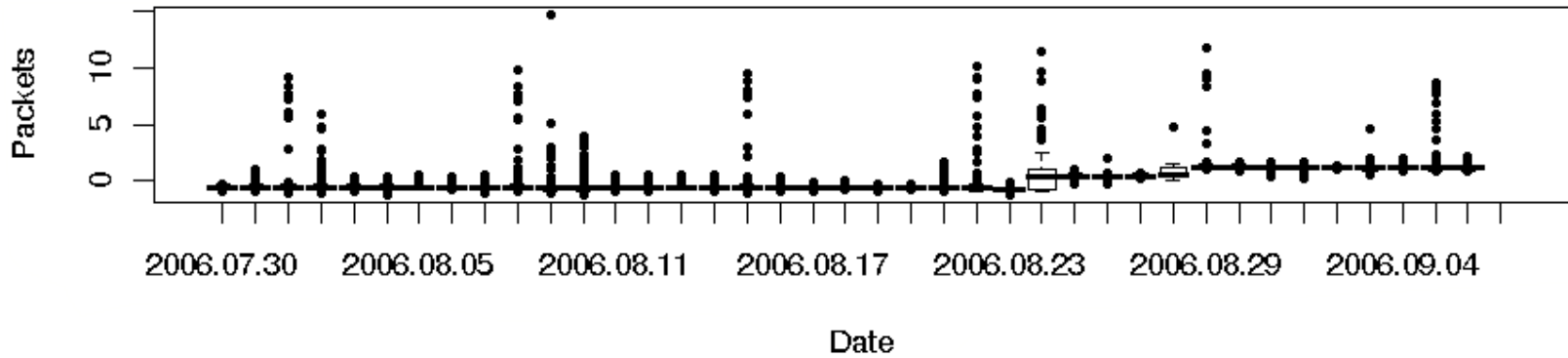
Packets



Date

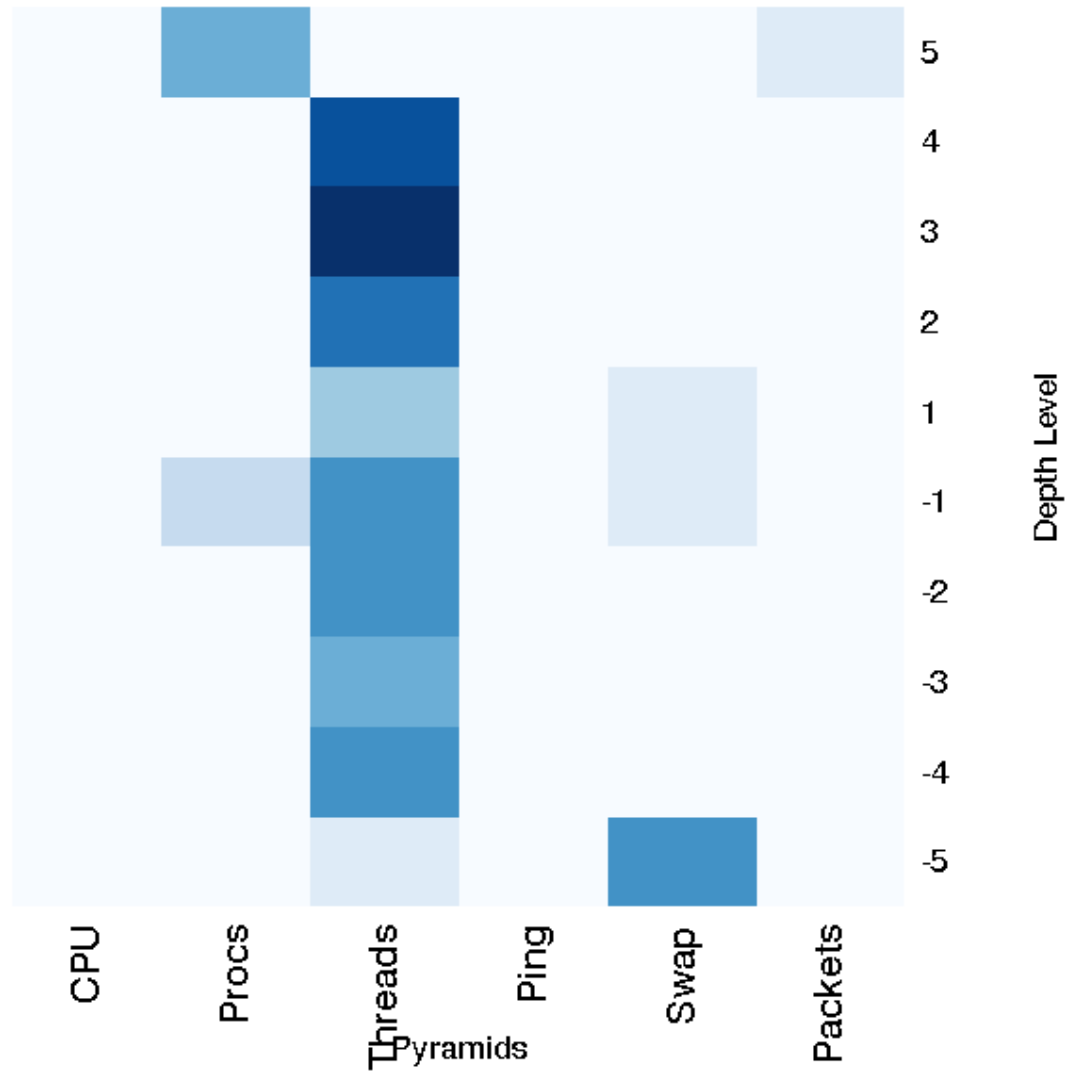


Partitioning Into Bins



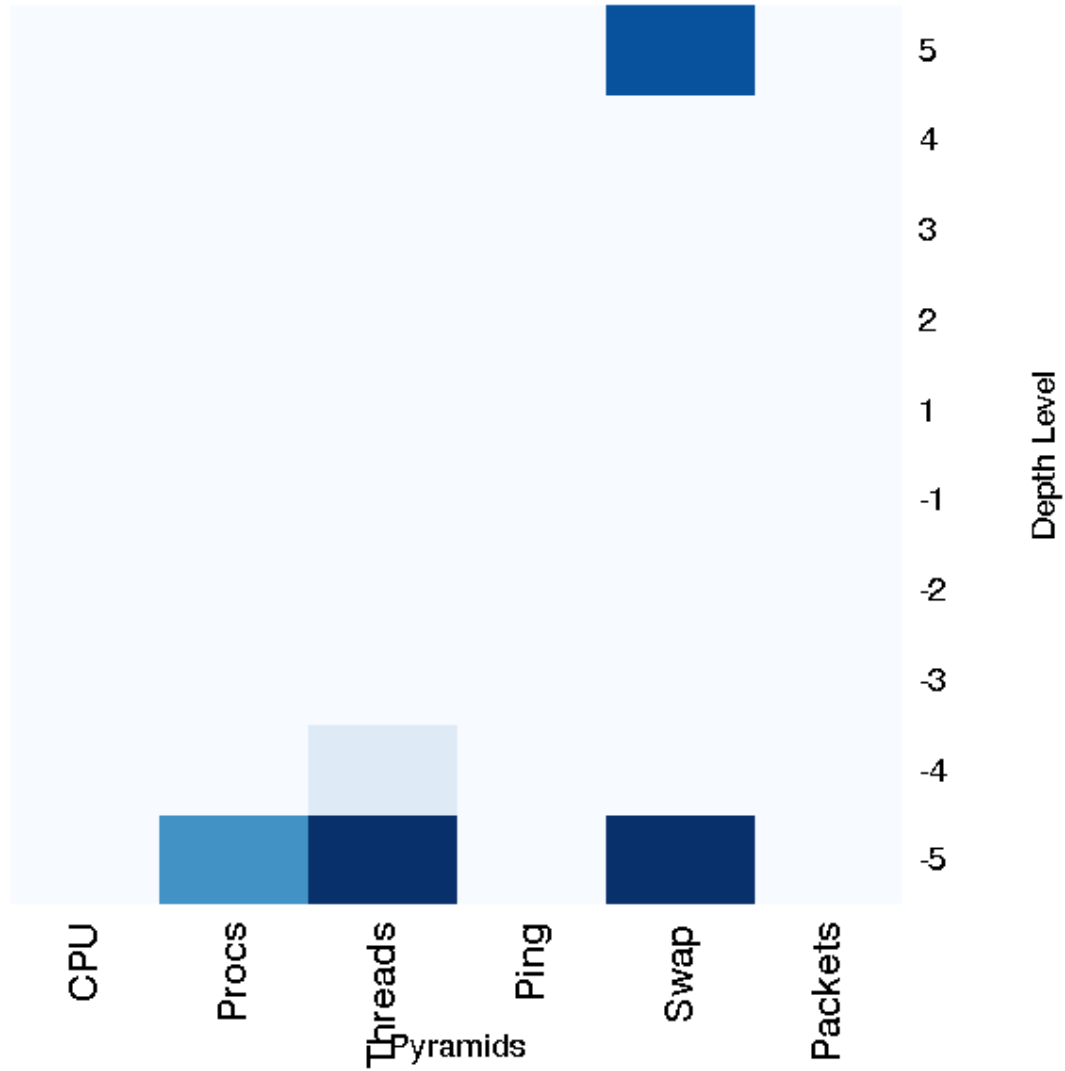
Week 1 Results

Week1 v. Week1: Trim=0



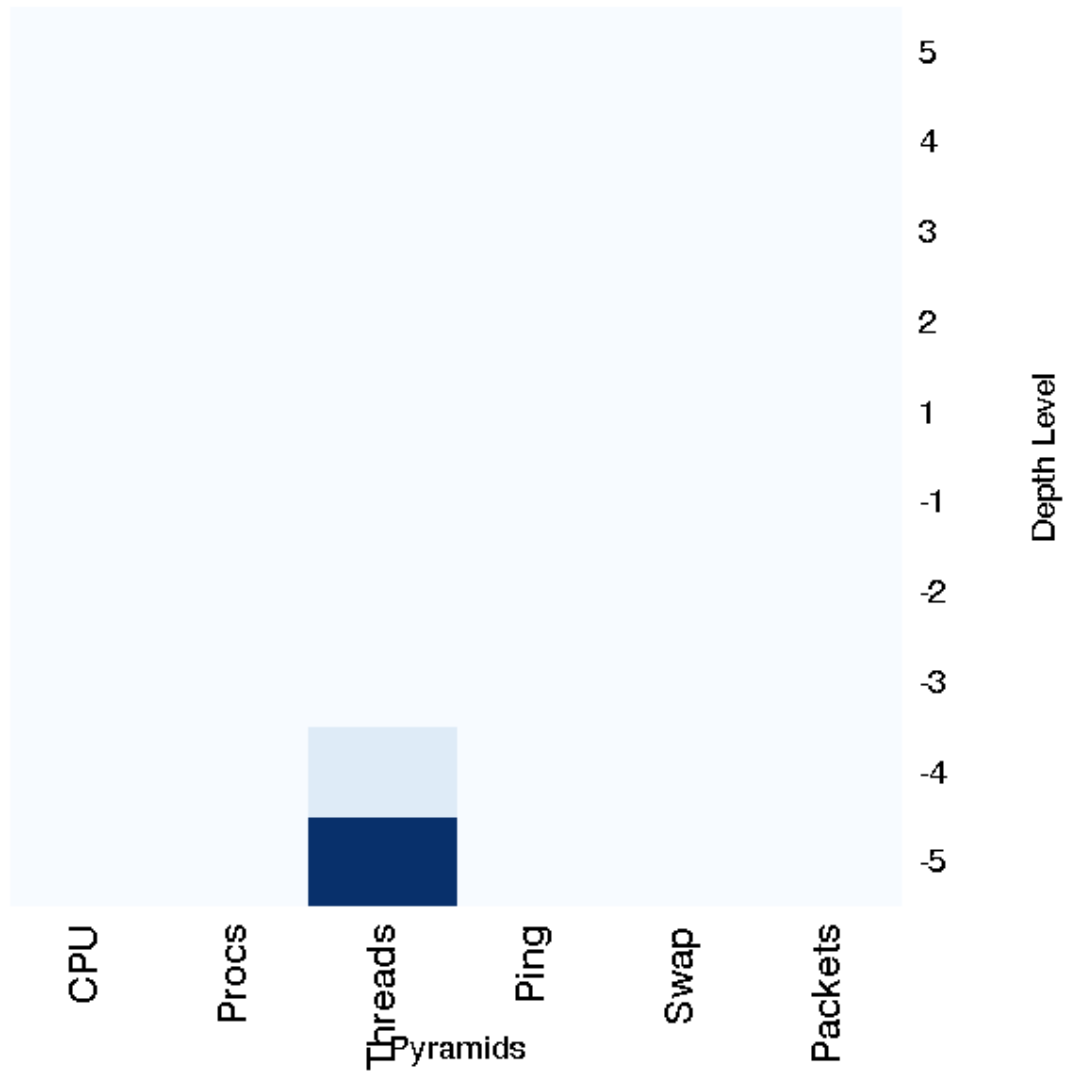
Week 2 Results

Week2 v. Week1: Trim=0



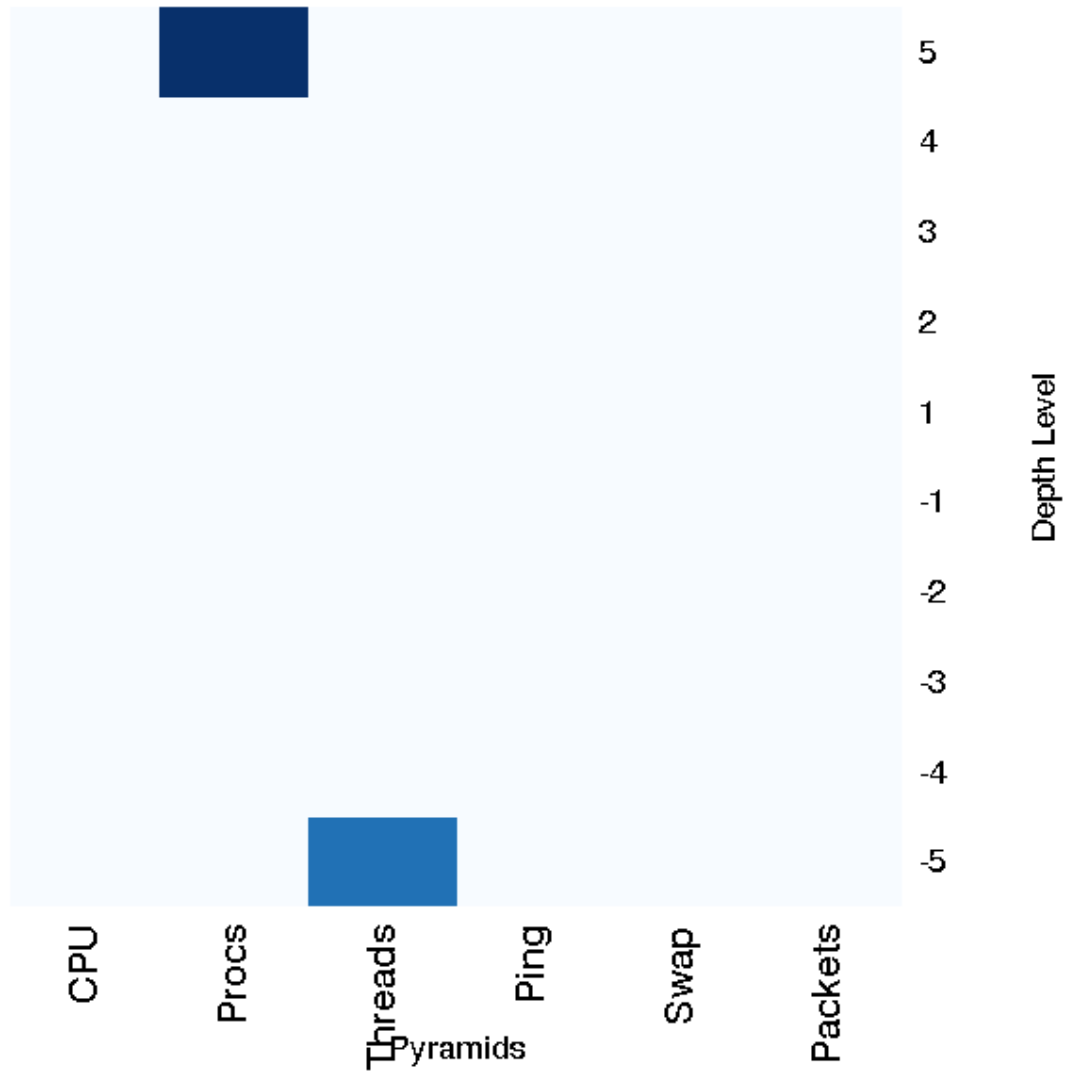
Week 3 Results

Week3 v. Week1: Trim=0



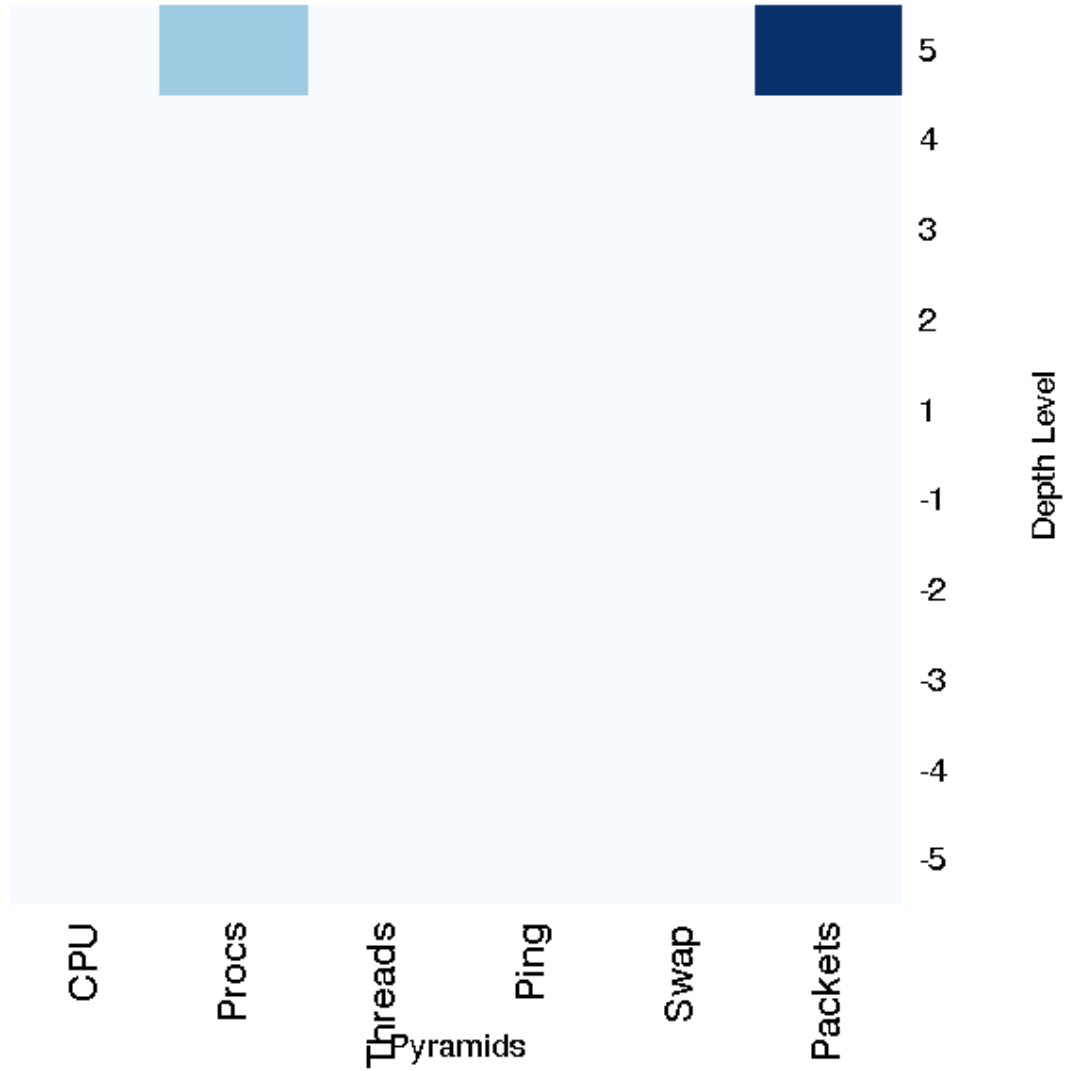
Week 4 Results

Week4 v. Week1: Trim=0



Week 5 Results

Week5 v. Week1: Trim=0



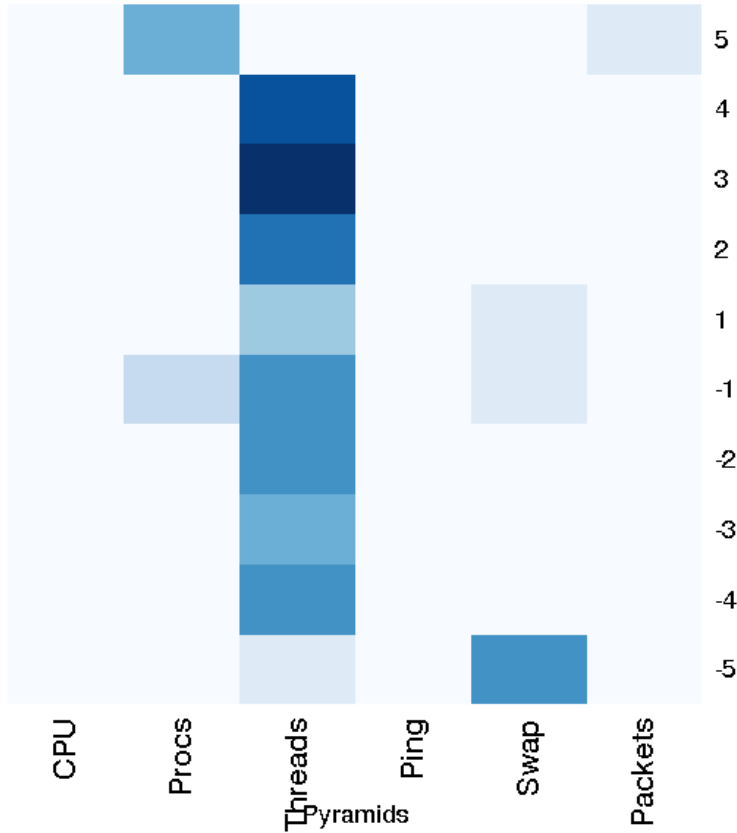
Trimmed Mean and Covariance

- Applying partitioning method using a trimmed mean and trimmed covariance matrix
 - Use 90% most central data points
 - Recompute center
 - Recompute Mahalanobis distances using new center and new covariance matrix
- Bins are more uniformly filled
 - More points appear closer to trimmed center
 - More variables become “most extreme”

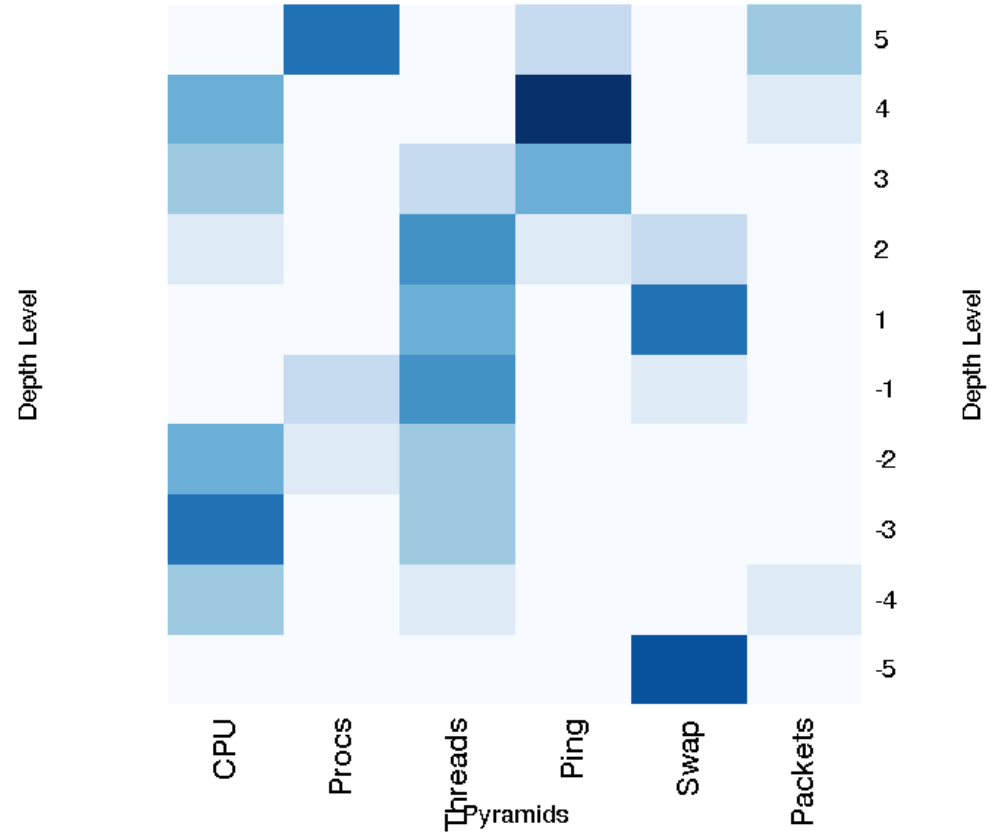


Trimmed Comparison: Week 1

Week1 v. Week1: Trim=0

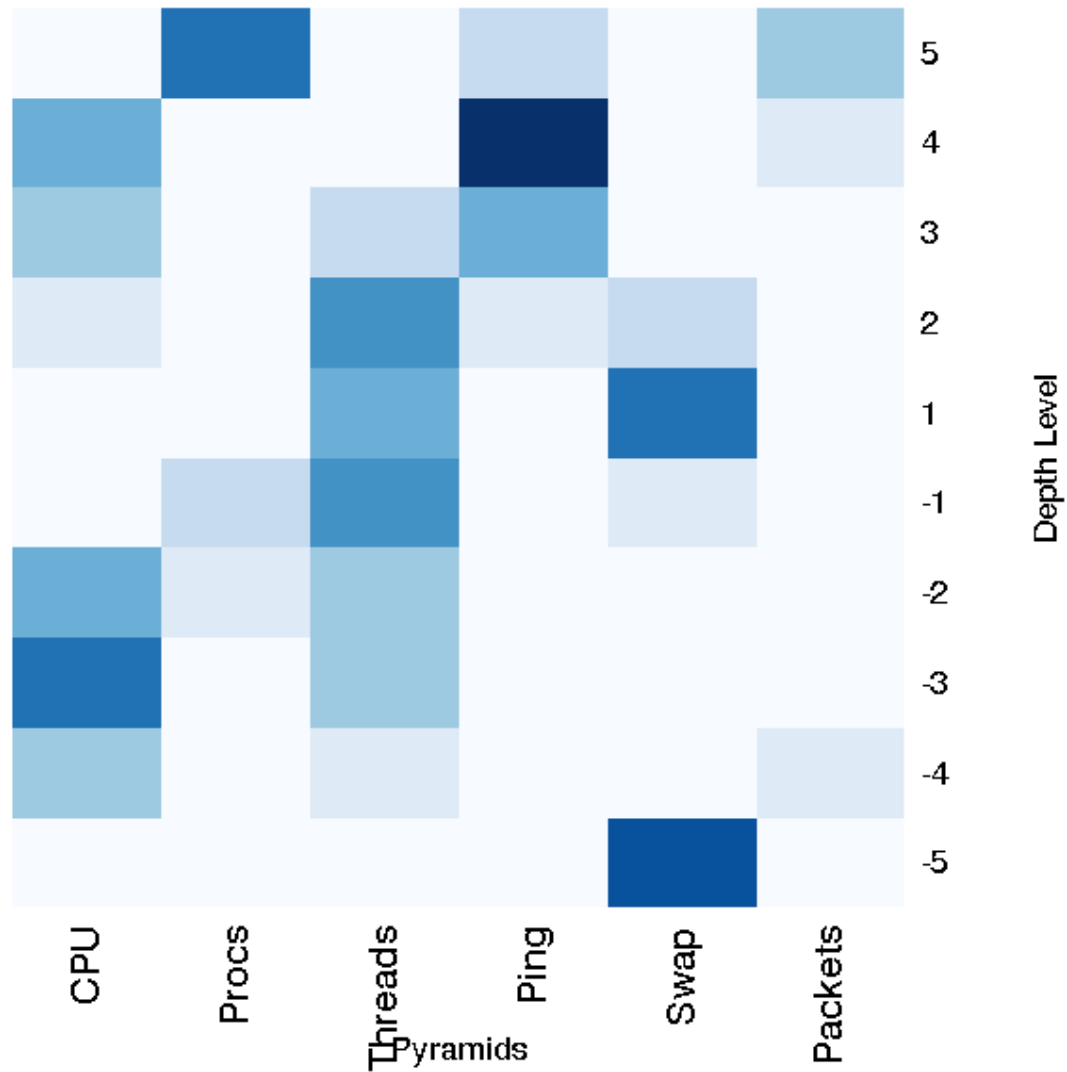


Week1 v. Week1: Trim=0.05



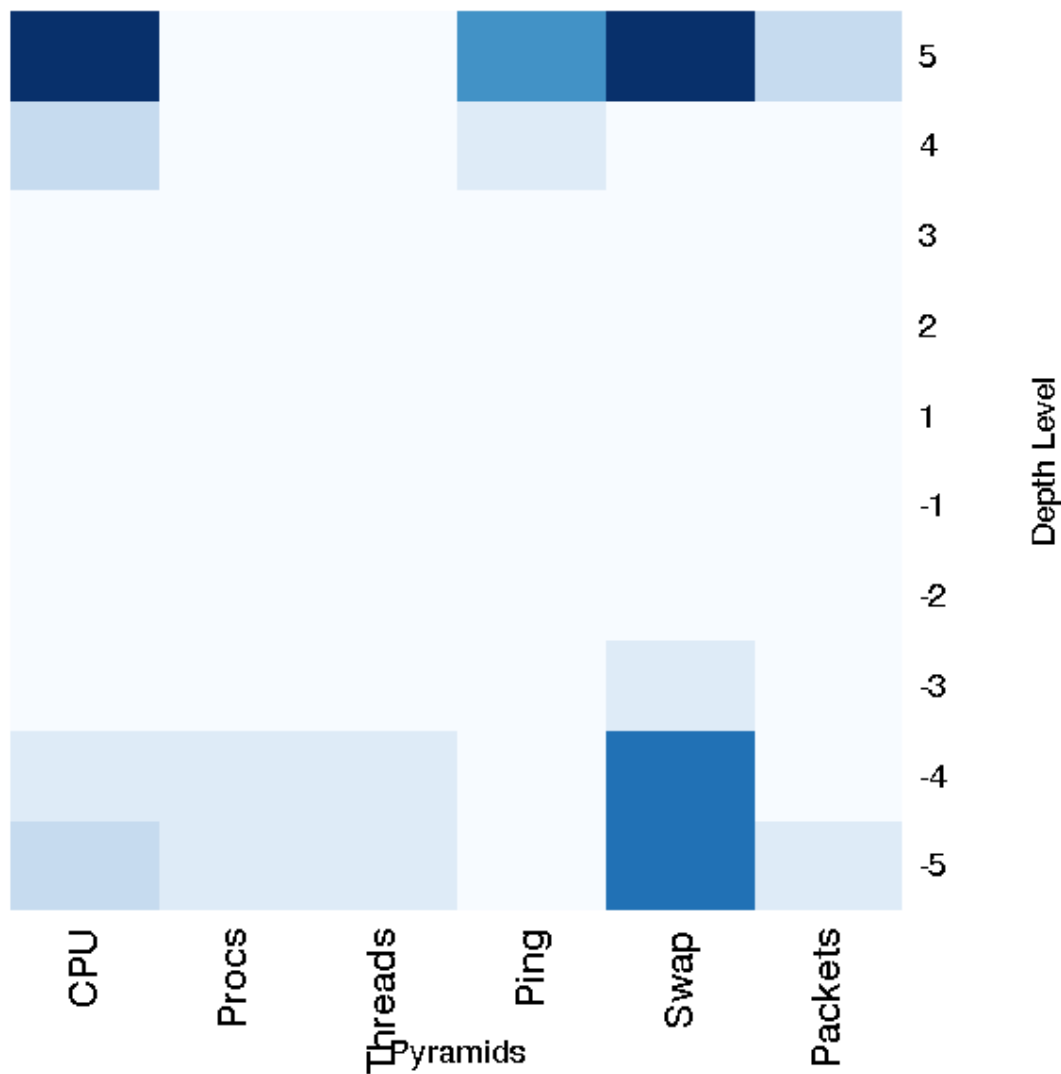
Trimmed Results: Week 1

Week1 v. Week1: Trim=0.05



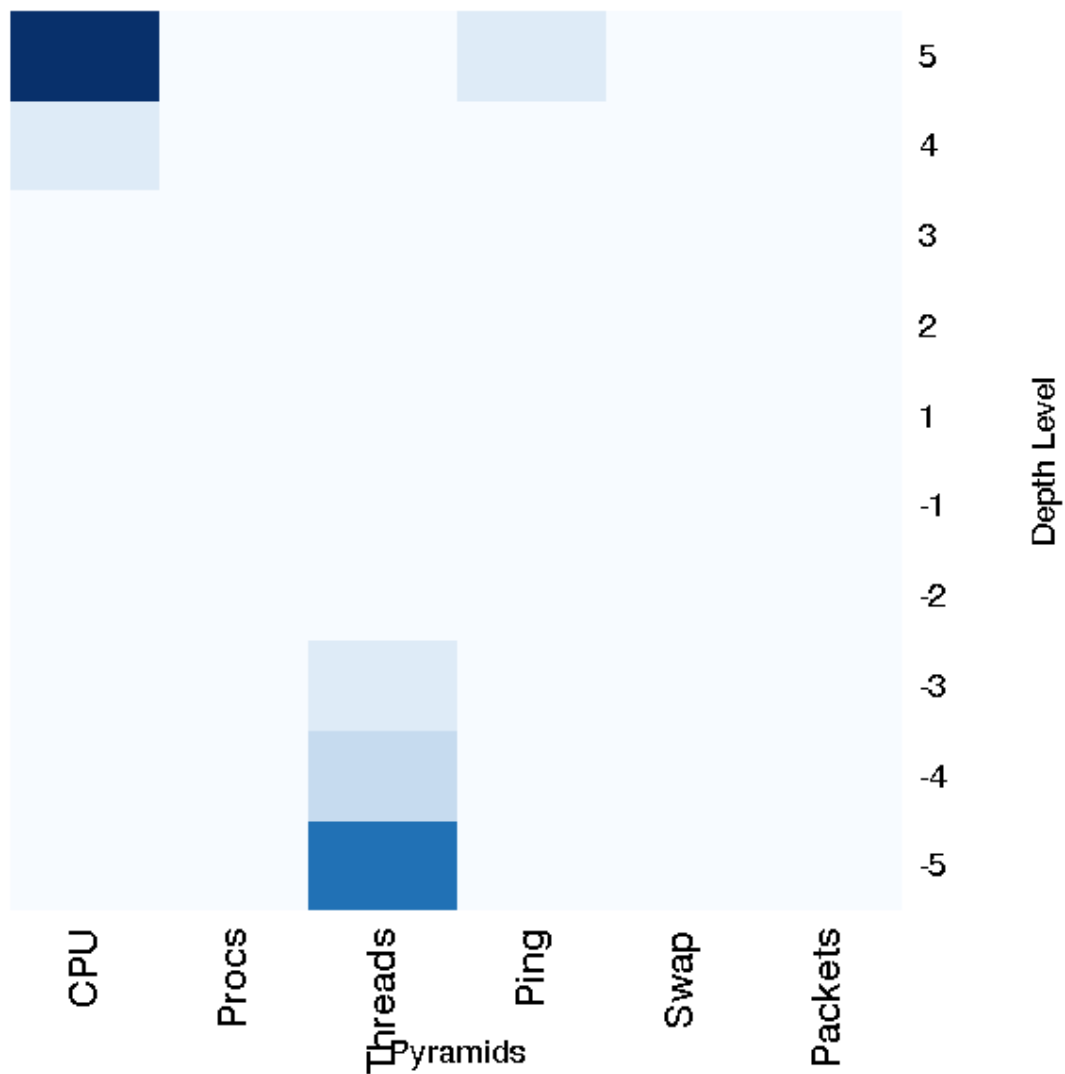
Trimmed Results: Week 2

Week2 v. Week1: Trim=0.05



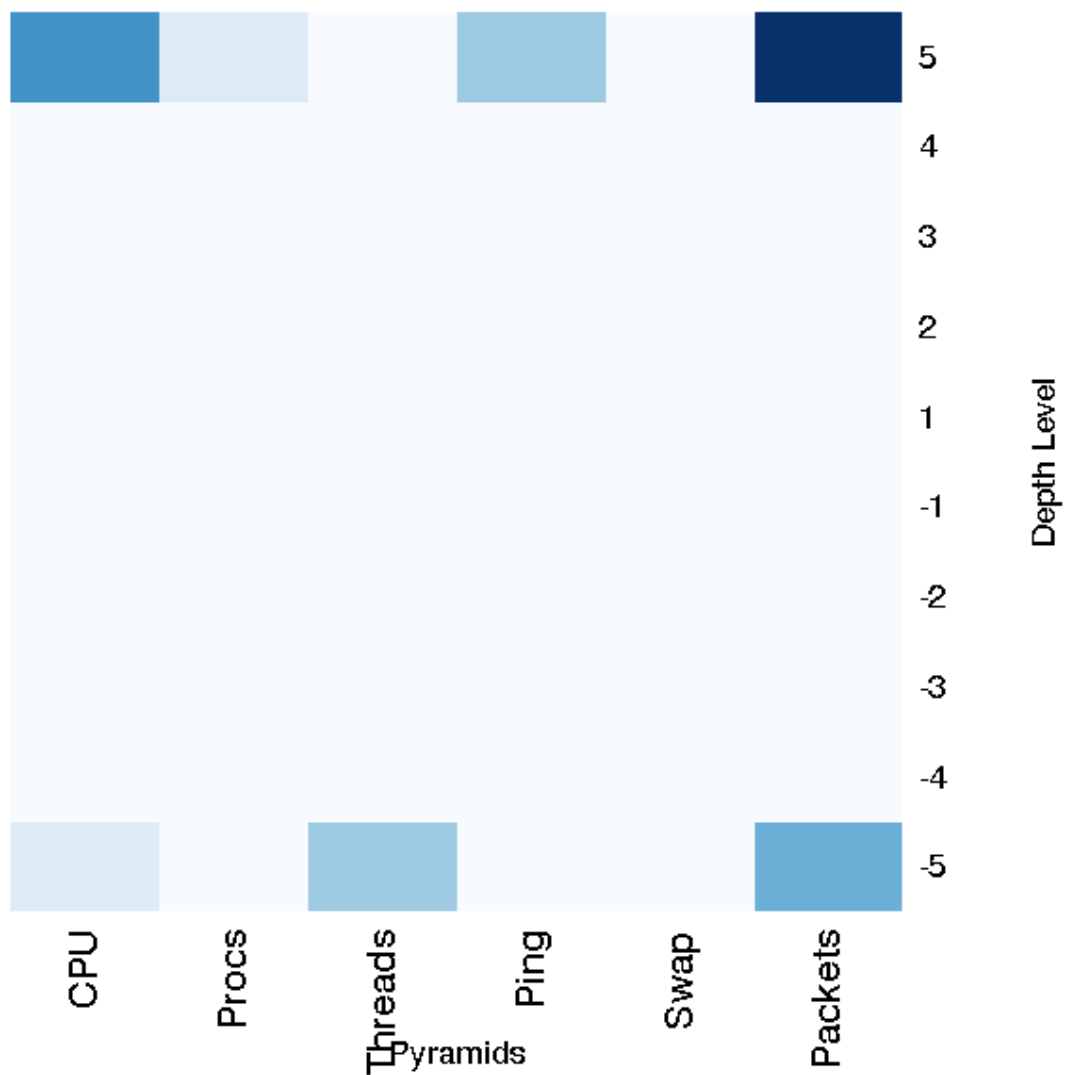
Trimmed Results: Week 3

Week3 v. Week1: Trim=0.05



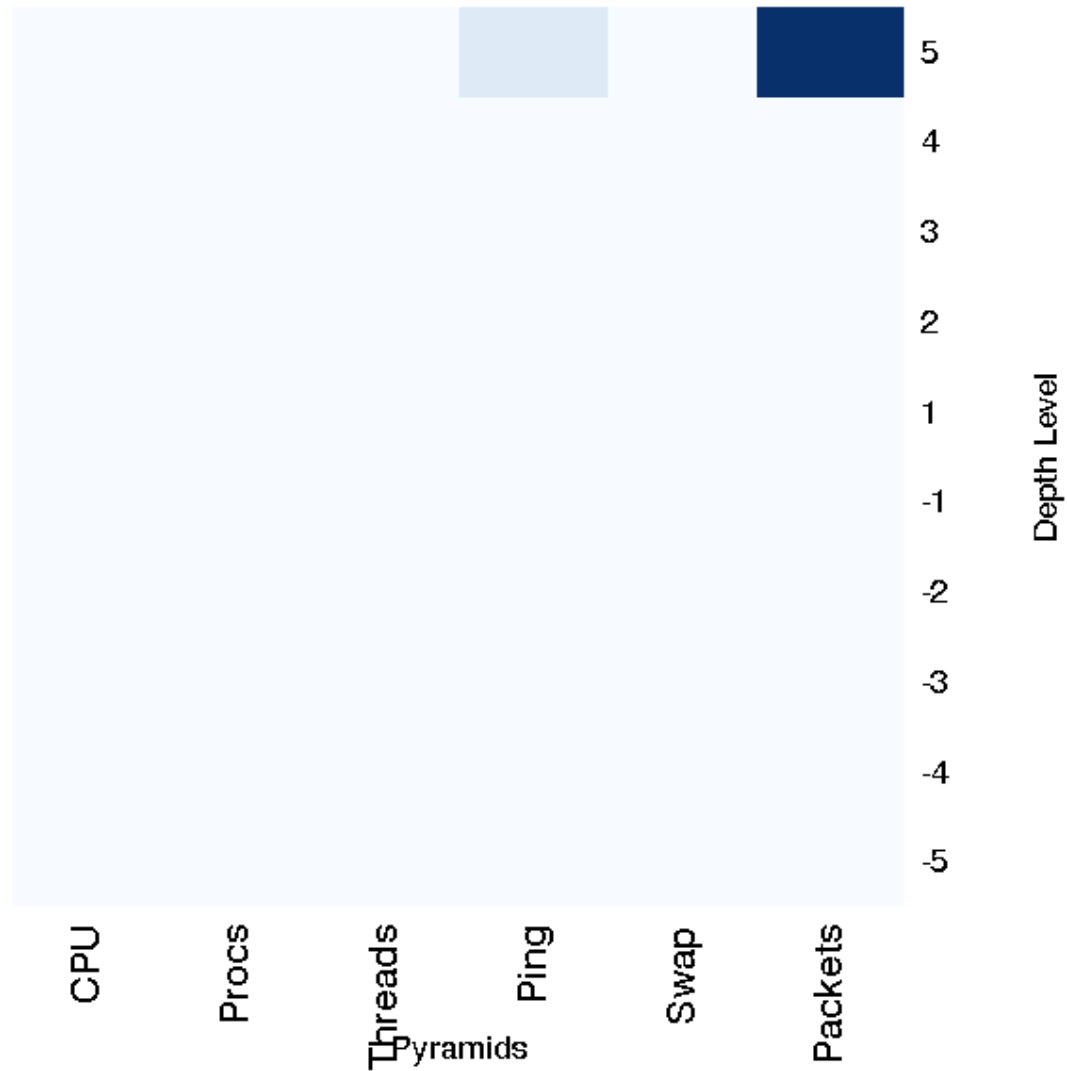
Trimmed Results: Week 4

Week4 v. Week1: Trim=0.05



Trimmed Results: Week 5

Week5 v. Week1: Trim=0.05



Multivariate Tests of Statistical Significance

- **Multinomial test**
 - Chi-squared test, G-test
- **Bootstrap and Resampling**
- **Bayesian methods**



Conclusion

- Quickly categorize data points non-parametrically
- Compare bins
- Identify which variables change most
- Questions or comments to
ervance @ stat.duke.edu

