Anomaly Detection Score White Paper

Initial feature set:

total\_rcvd\_value: total value received for all history

total\_sent\_value: total value sent for all history

total\_balance: total balance over all history (avg\_sent\_value-avg\_rcvd\_value)

total\_rcvd\_count: total number of TXs received for all history

total\_sent\_count: total number of TXs sent for all history

avg\_rcvd\_value\_per\_block: historical avg value received per block

avg\_sent\_value\_per\_block: historical avg value sent per block

avg\_rcvd\_count\_per\_block: historical avg # TXs received per block

avg\_sent\_count\_per\_block: historical avg # TXs sent per block

max\_rcvd\_value\_per\_block: historical max value received per block

max\_sent\_value\_per\_block: historical max value sent per block

max\_rcvd\_count\_per\_block: historical max # TXs received per block

max\_sent\_count\_per\_block: historical max # TXs sent per block

avg\_sent\_count\_per\_sec: historical avg # TXs sent per second

avg\_rcvd\_count\_per\_sec: historical avg # TXs received per second

avg\_rcvd\_value\_per\_sec: historical avg value received per second

avg\_sent\_value\_per\_sec: historical avg value sent per second

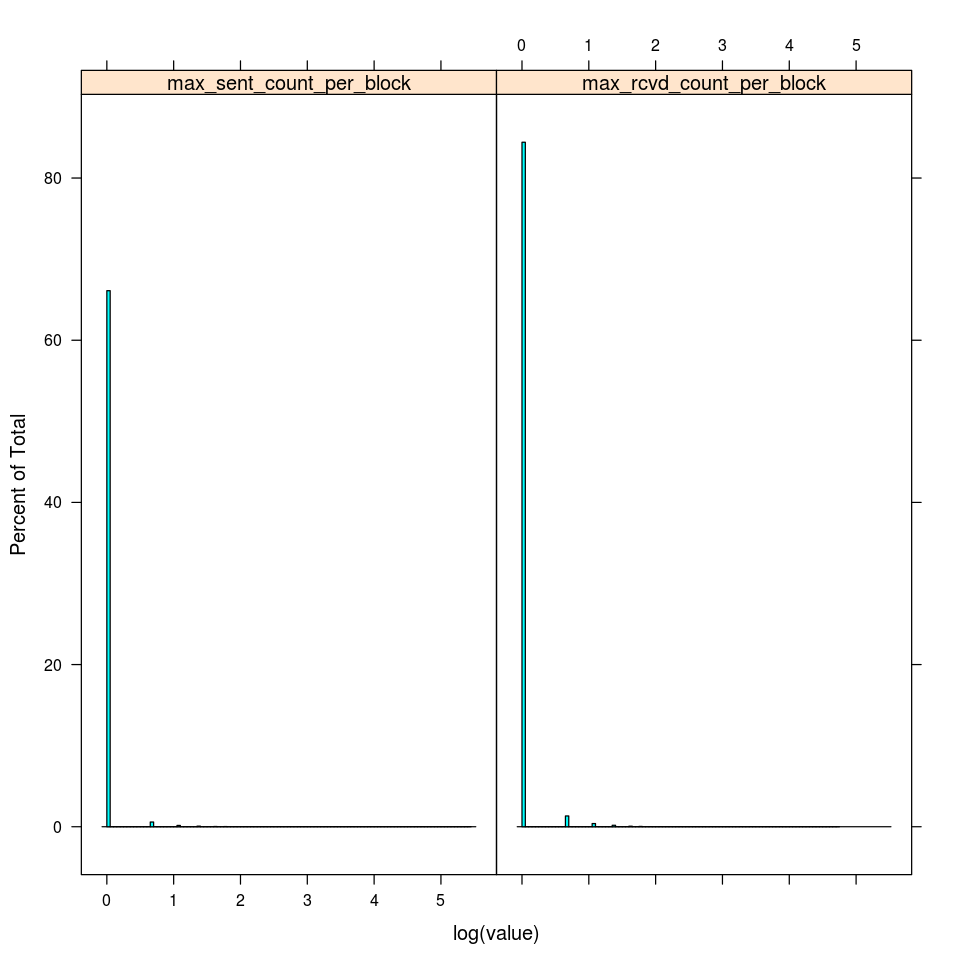
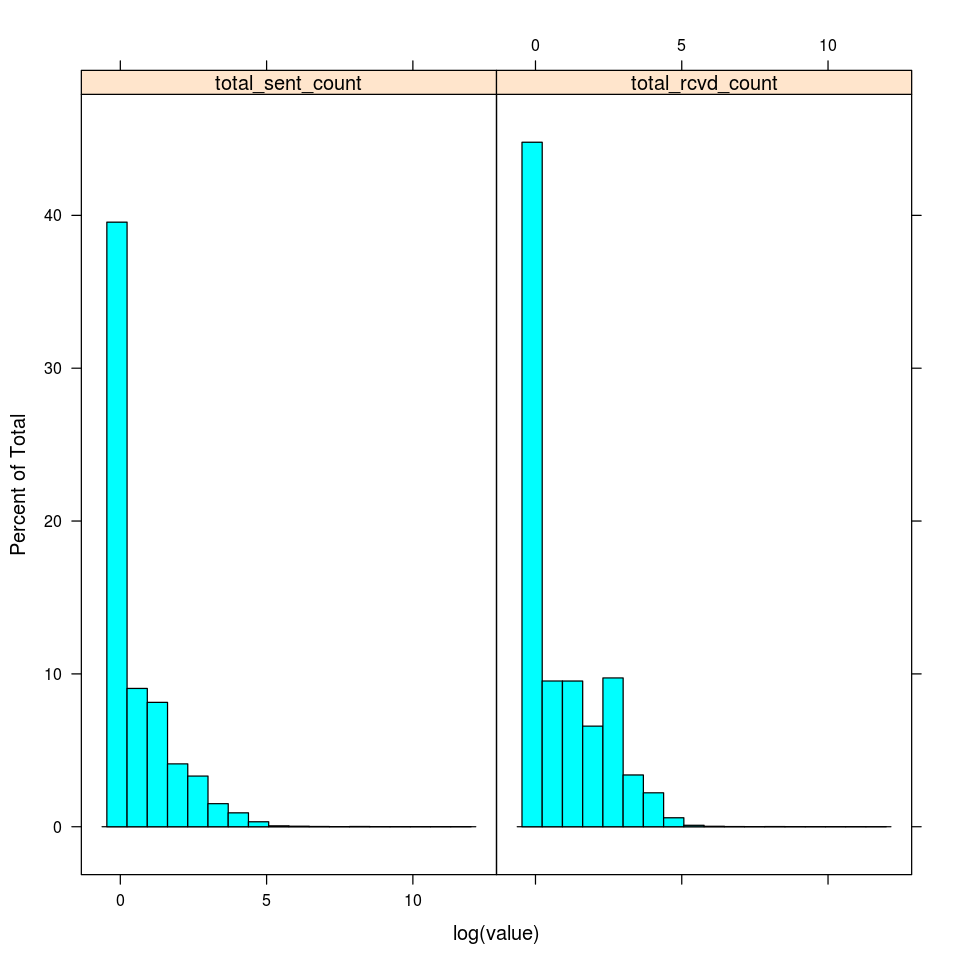
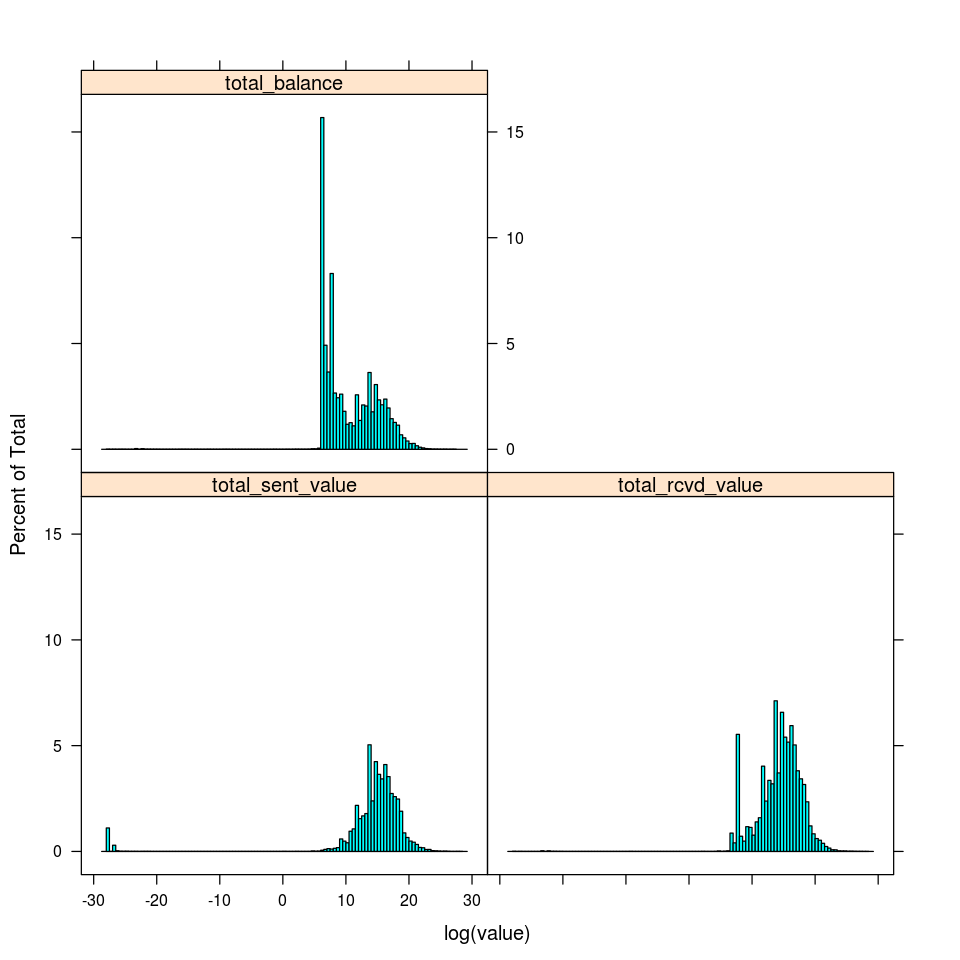
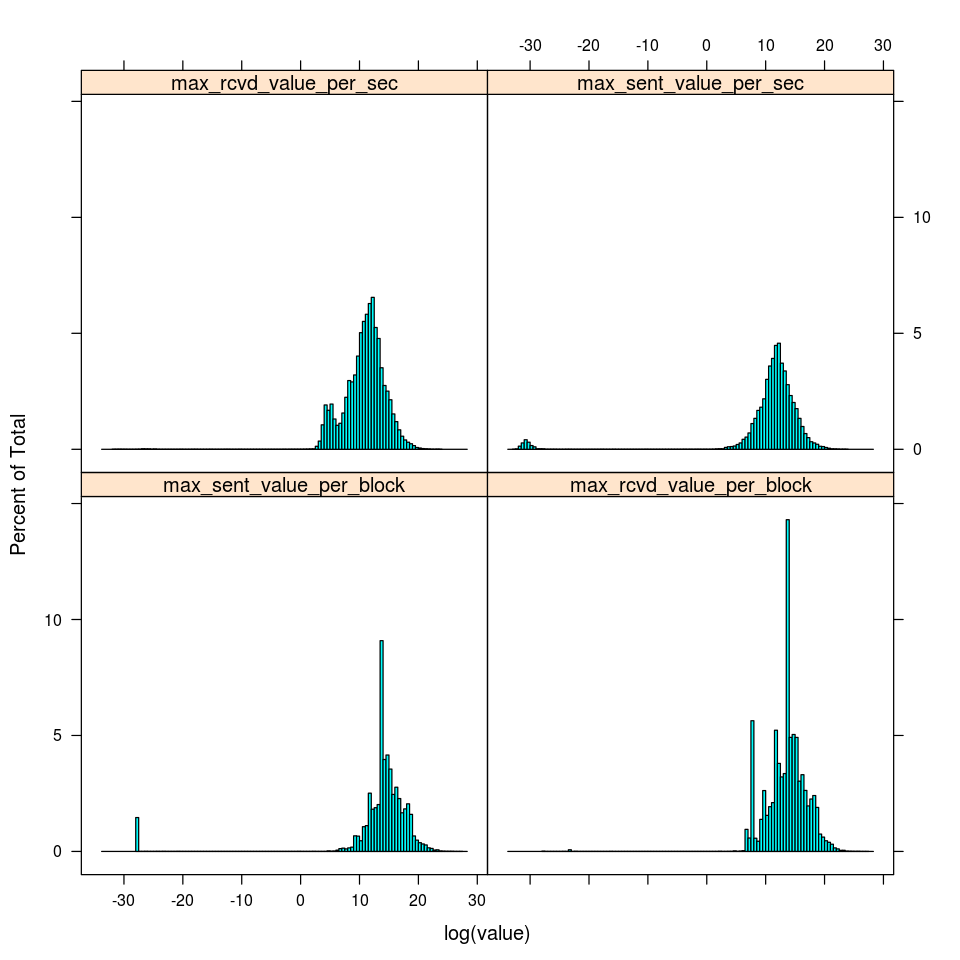
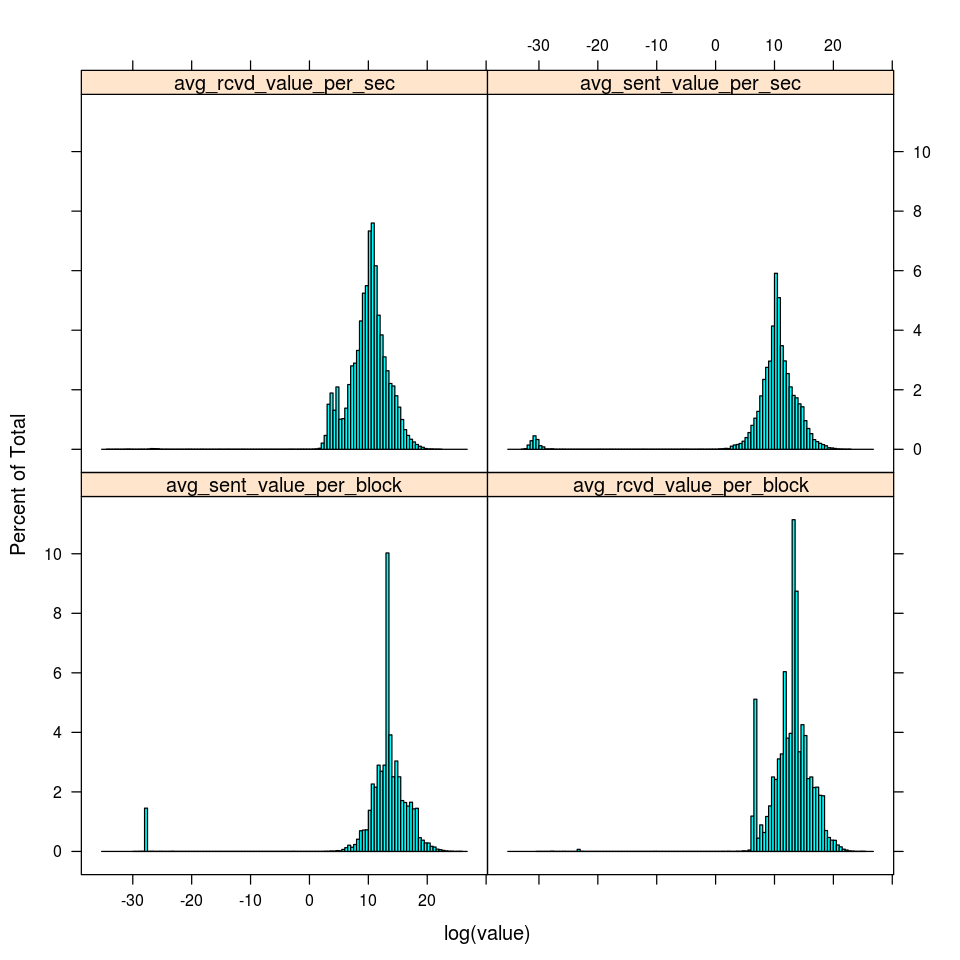
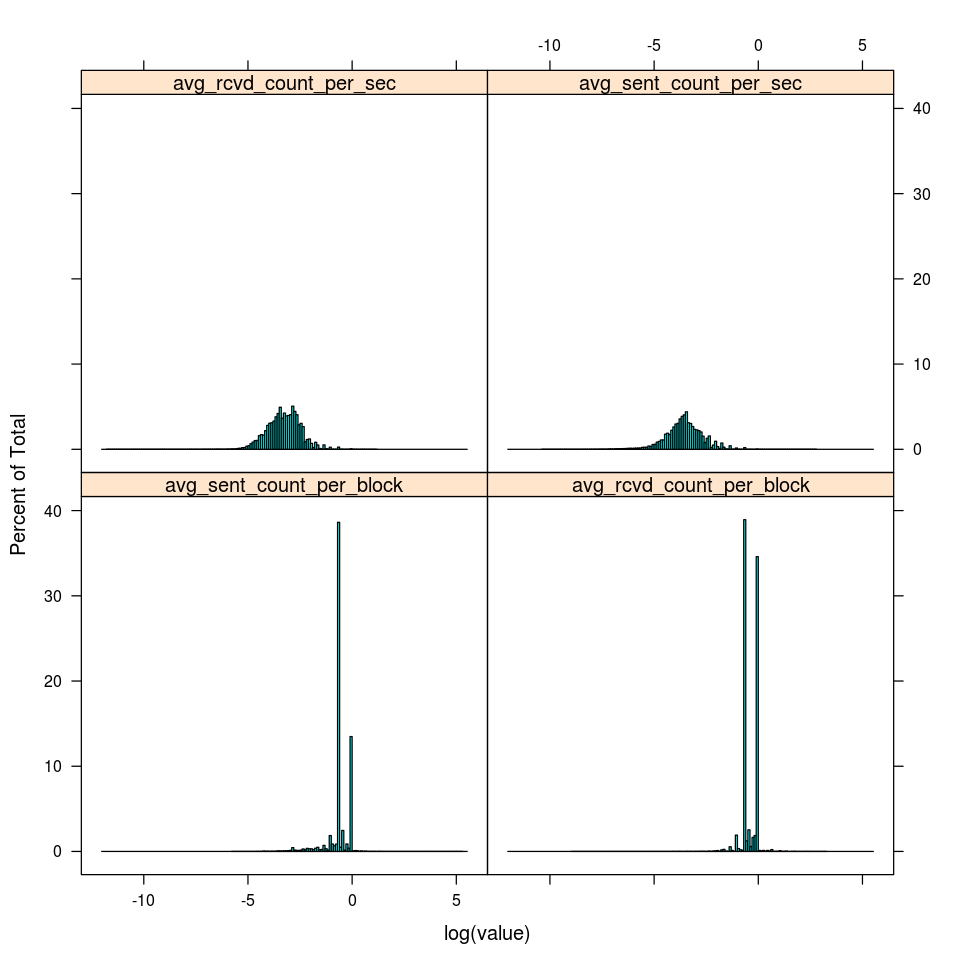
max\_sent\_count\_per\_sec: historical max # TXs sent per second

max\_rcvd\_count\_per\_sec: historical max # TXs received per second

max\_rcvd\_value\_per\_sec: historical max value received per second

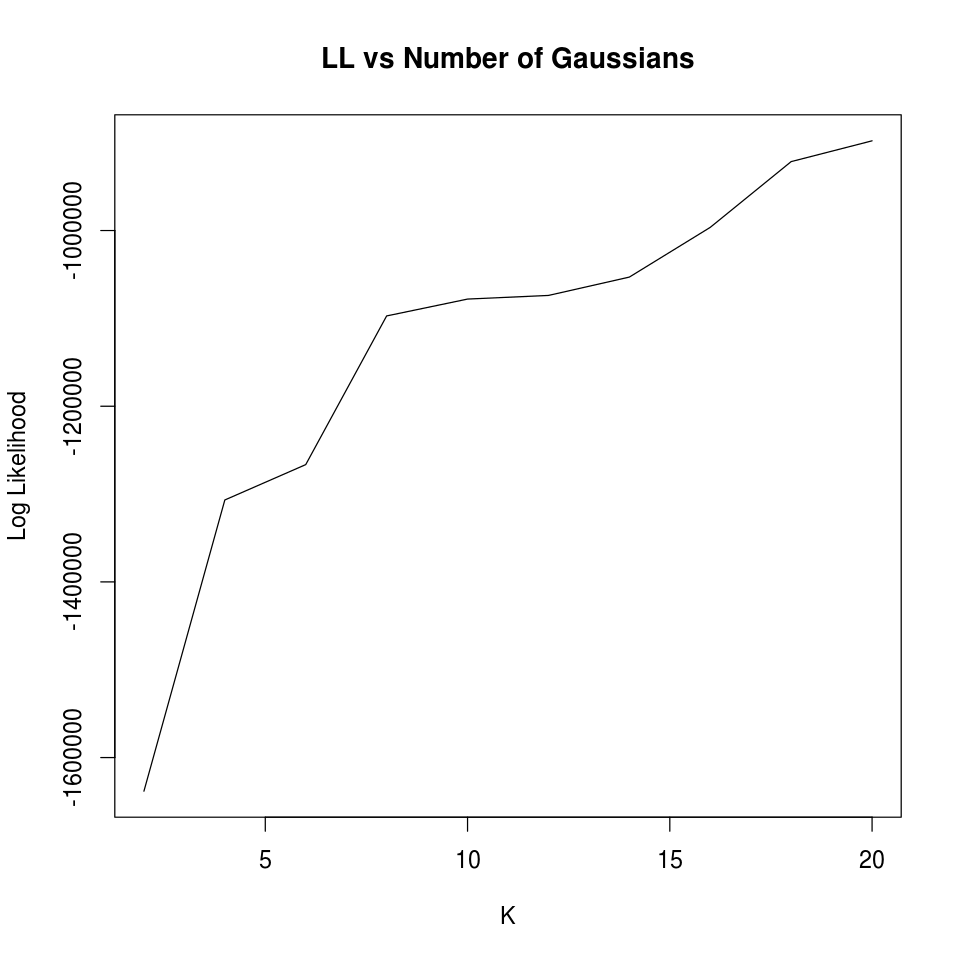
max\_sent\_value\_per\_sec: historical max value sent per second

I analyzed all of the above variables for collinearity by looking at their correlation matrix. Out of 441 possible relationships, I did not find any that were correlated over 90%.



We can see that in general, the variables involving averages/maxima over blocks are much further from normal than the variables involving averages. Therefore I decided to exclude all aggregate variables over blocks in favor of aggregations per second.

Finally, I ran a GMM for k={2,4,6,8,10,12,14,16,18,20} and observed the log likelihood. I found the following curve for the log likelihood:



There appears to be a significant “elbow” in this curve at K=8; therefore I will pick K=8 as the number of independent gaussians for this model.