James Legg *j.legg.17@ucl.ac.uk*
Kevin Mulder *k.mulder.17@ucl.ac.uk*
Patrick Roddy *patrick.roddy.17@ucl.ac.uk*

Group Project, Spring Term 2018

Intrusion Detection from Network Logs

THE DATA

Two 1TB datasets from the Los Alamos National Laboratory, namely network traffic events and host security events collected from their enterprise network over 90 days, as described by and to be found at:
M. J. M. Turcotte, A. D. Kent, and C. Hash, "Unified Host and Network Data Set", ArXiv e-prints, Aug.2017.
https://csr.lanl.gov/data/2017.html
Our mission, which we chose to accept, was to find intrusions to the network, which are undoubtedly there somewhere!

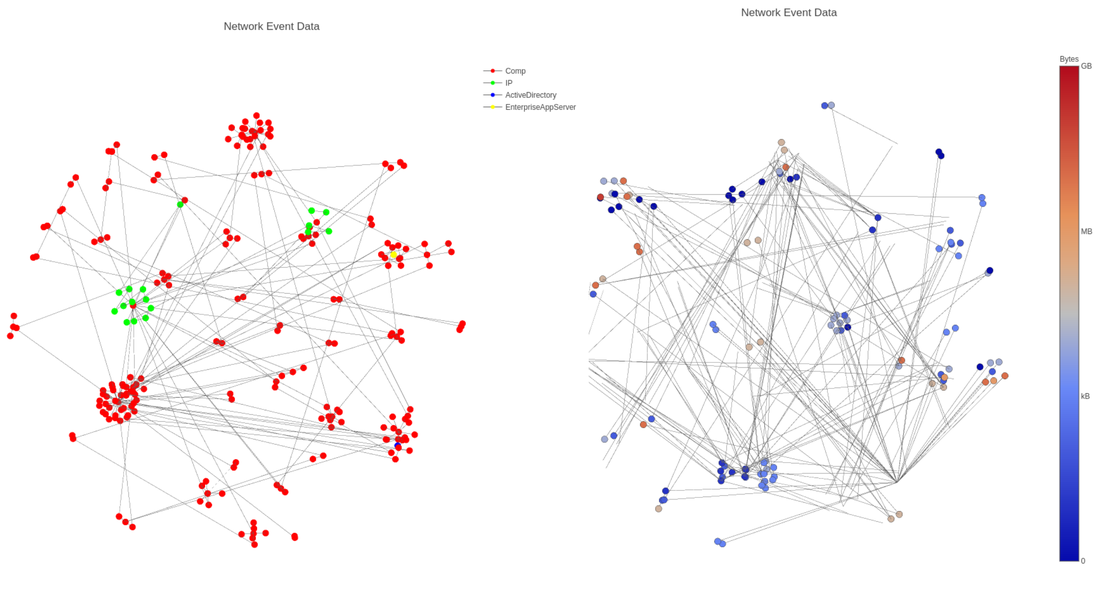
REDUCTION 1TB → 300GB → 25GB → 5GB

Terabyte data sets are unwieldy. A first problem was to get space allocated on server where we could all access them. Python allows rapid script development but takes days to process this much data. To cope we reduced the size of the datasets by the processes of coding the data (repeated string values and long integers to small integers) and successive time binning. The latter risked losing the correlation information between the many fields describing each network event; however, there remained much to be discovered.

INSPECTION

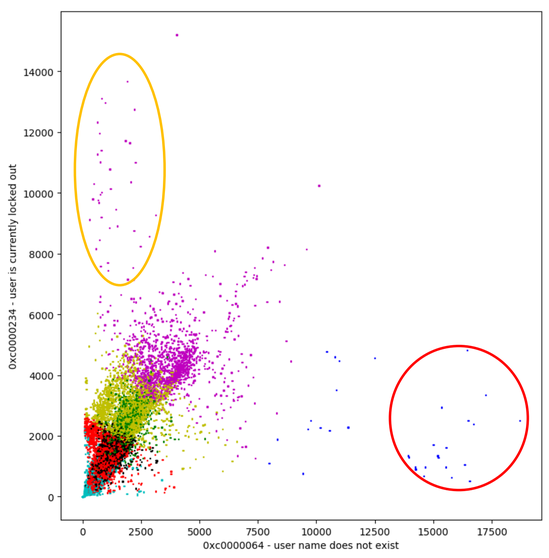
GRAPHING

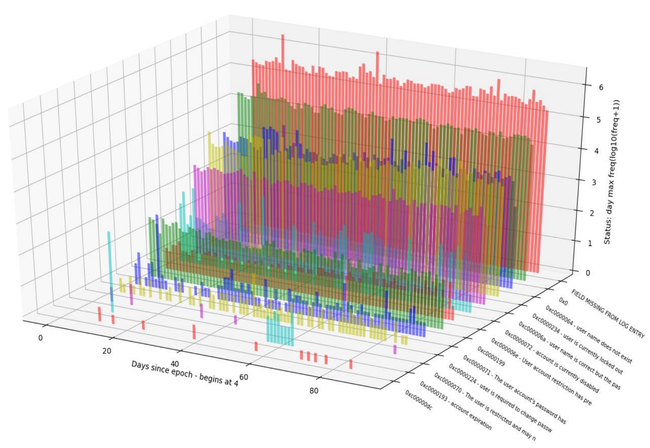
A network graph was created from the network flow data in order to visualise the network. The network graphs are drawn in either: two or three dimensions; using the Kamada-Kawai (KK) or Fruchterman-Reingold (FR) force-directed layout algorithms; and with or without lines connecting nodes. The source node is taken as the parent node, and the unique destination nodes as the children nodes. This visualisation can help to identify nodes which are potentially malicious – namely they are singly connected on the outskirts of the network graph.



CLUSTERING

Scatter plots, coloured using the K-means clustering algorithm, were produced for selected pairs of field values. This one is for the "user locked out" and "user name does not exist" values of the log on status field. The yellow circle is for times of very frequent log on attempts to locked out accounts. Red is for times of very frequent log on attempts with the wrong password. Both are suspicious. Note that these are unusual events compared to the norm, but compare opposite, which are rarer still.

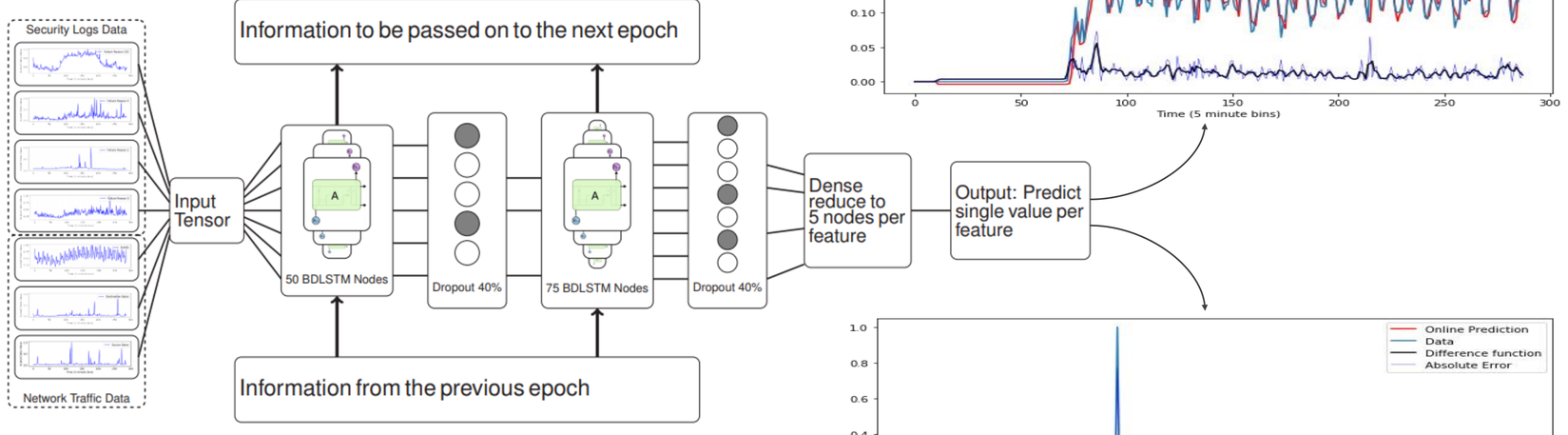




RARE EVENTS

Unless a network is hopelessly compromised, malicious events will be rare. The host security events were histogrammed, producing one series per field value against time. This plot is of the histograms for the various possible values the log on status field. Rare events on these and plots for other fields were identified visually and the original log records for them were then retrieved. Our project sponsors, a network security company, considered these and reported that the events included: various kinds of unauthorised log on attempts, poor practice such as automatic scripts with hard coded credentials and possibly even a simulated hacking process. (Asking an expert to label events is known as "consulting the oracle").

PREDICTION



RNN

A BiDirectional Long Short Term Memory (BDLSTM), a type of RNN, were used to analyse offline and real-time data. This enables more thorough network monitoring as well as continual improvement of the BDLSTM itself. An LSTM differs from a basic RNN due to its internal memory which can be updated for every epoch, which enables long term pattern recognition.

Using a BDLSTM offers automated real-time monitoring plus pattern recognition between features, which would be easy to miss by human or algorithmic monitoring.

The network was implemented in Keras, with a Tensorflow backend. Running on the network activity as a feature a loss of 2.87 × 10^{-4} (MAE) and a validation loss of 5.81 × 10^{-4} (MAE) is reported. Due to the scaling of the data, a custom percentage error was implemented and an average predictive power of 92.80% was reported.

CONCLUSION

We may not have labelled all the bad actors, but the project demonstrated that, even with a hefty reduction applied to the data and slow processing tools, significant events of concern to security professionals can be identified through traffic flow visualisation and observation of rare events and isolated clusters. Furthermore, bulk flow of traffic was successfully predicted with a neural network, opening up the possibility of real time monitoring for anomalous flow e.g. bulk data theft.

UCL CDT DIS

nccgroup
PROJECT SPONSOR