Preliminary Analysis Report of the Network Dataset

Tian Guo, Konstantin Kutzkov, Mohamed Ahmed, Jean-Paul Calbimonte, Karl Aberer

1 Data set

All observations in the following are based on the udp dataset for the first half of April. Unless otherwise specified, the reported measurement is jitter_up.

2 General approach

The measurements conducted by the probes, for example, the individual UDP-jitter tests, may be described in terms of the attributes of the probe (`ranNode, apNode, hubType, headlineSpeed, location`, and etc), the test (`week day, hour, minute, target`, and etc), and in fact anything else that we can monitor (the features). With this, if the distribution of a variable we are interesting in monitoring, say the "up-stream UDP jitter", is dependent on a particular combination of feature values (or combinations of features), we can attempt to characterization this dependency, for example in terms of the correlation between a particular observation and some (set of) feature value(s).

In this way, we can attempt to identify and explain the performance measurements collected, such as for the jitter, in terms of the features we observe. For example, we may identify that the up-stream UDP jitter measured by probes on DSL-lines in London to targets in New York is normally high/unstable in the evenings on weekdays, and should not be interpreted in the same way as say a measurements from the morning. Conversely then, anomalies here are observations that cannot be adequately explained by expected outcome of the feature combination that generates them.

Our goal is therefore to build models that are able to explain measurements at the network level, in terms of the features we are able to monitor. something on flexibility, space, multi modal data ...

To motivate the importance of considering the range in the values of features, and the combinations of features, we report below a small study that shows we are able to better characterize the distribution of the variable of interest (UDP jitter-up), if we explicitly account for the range of individual features and the impact of considering features jointly.

The distribution for individual features In Figure 1 we plot the mean and the median value for the 60 different values of the feature 'minute'. We consider the 'mean' because it is very sensitive to outliers and just a few of them can dramatically change the mean value. On the other hand, the median is robust towards outliers. From Figure 1, we see that the median appears to be stable across different values with a value of approximately $400\mu s$. In contrast, due to a few outliers, the mean values are much larger and also quite skewed. In Figure 2 we plot the mean and median values for the feature on a log scale in order to highlight how much the mean values are different from the median values due to outliers.

In Figure 3 we plot the mean and median distribution for different apNode values and in Figure 4 the values for different ranNode values. Again, we observe that the median values are stable but the mean values differ significanlty. Furthermore, we observe that the distribution of the mean, median are not the same accross features. For example, comparing the distrinution of the UDP-jU for the features headlineSpeed (Figure 5) and hubType (Figure 6), we observe that the mean and median values differ significantly.

The distribution for pairs of features In this section we present evidence that the jitter distribution for different feature combinations differ significantly. For example, the median jitter value the the subset of the data where 'minute=1' and 'apNode=7' will be considerably different from the median jitter value

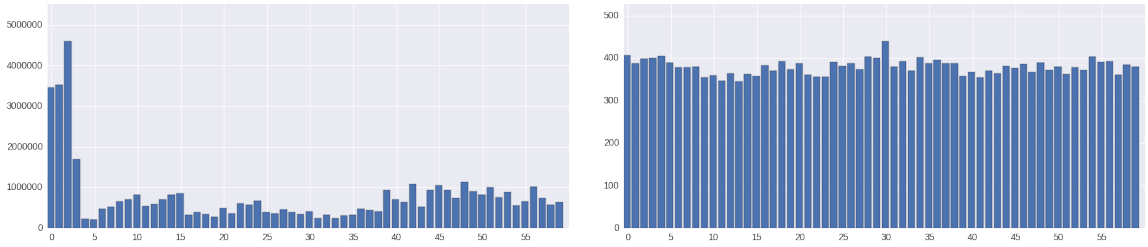

Figure 1: The mean(left) and median(right) for different minutes.

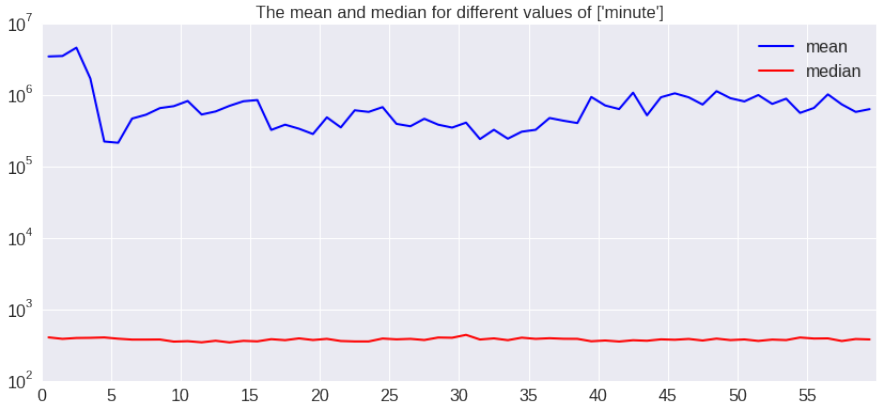

Figure 2: The median and mean jitter for different 'minute' values.

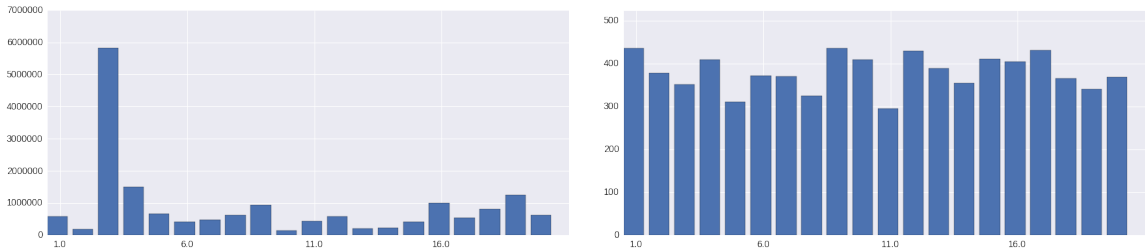

Figure 3: The mean(left) and median(right) for different apNode values.

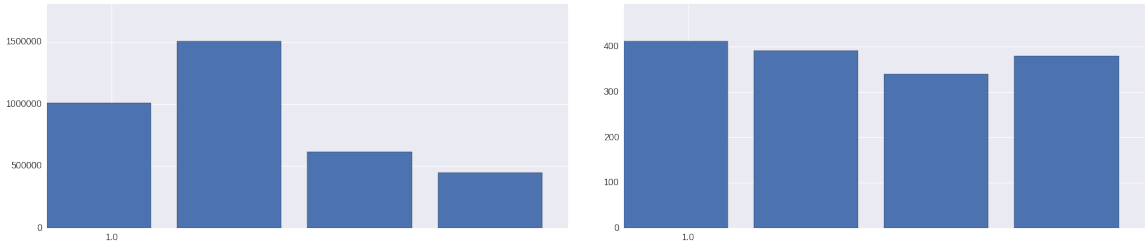

Figure 4: The mean(left) and median(right) for different ranNode values.

of the subset of the data where 'minute=53' and 'apNode=17'. We will also show that such combinations are in a sense unique and we won't be always able to predict the median jitter value of a given combination from its components. For example, the median jitter value for the subset of the data where 'minute=53' and

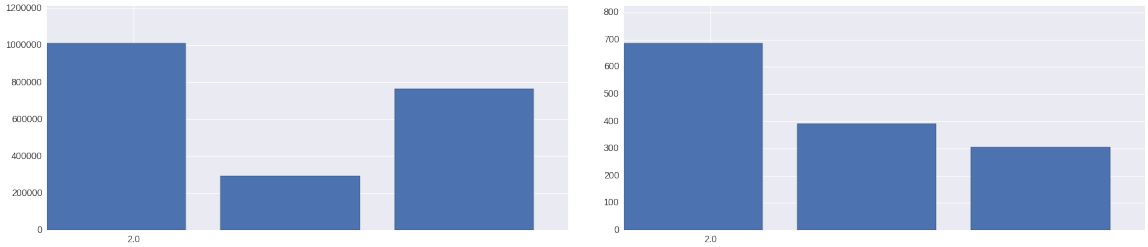

Figure 5: The mean(left) and median(right) jitter for different headline speed values.

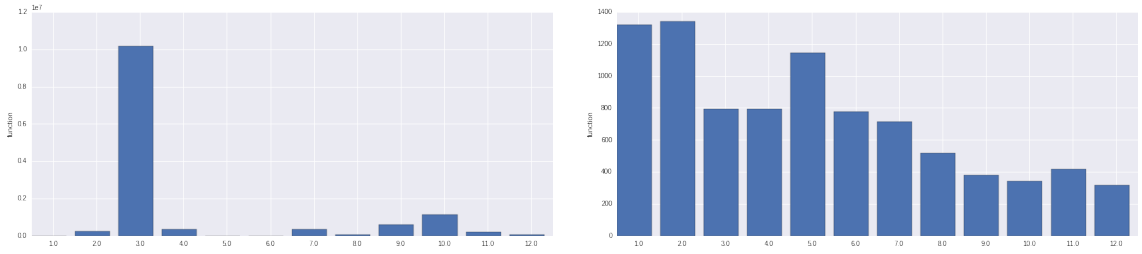

Figure 6: The mean(left) and median(right) for different hub types.

'apNode=17' cannot be predicted from the median values for the two subsets of the data where 'minute=53' or 'apNode=17'.

Let D be our dataset, $f_1, \ldots, f_k$ a set of feature values and $D[f_1, \ldots, f_k]$ the subset of the data with observation having the feature values $f_1, \ldots, f_k$. For example, $D[minute = 53, ranNode = 3]$ is the subset of the data where the test started in 53rd minute and the probe has ranNode=3. In Figure 7 we plot the median jitter values for the different combinations of the 'minute' and 'ranNode' features, represented by blue bars. For a given feature value f_i we denote by $j(D[f_i])$ the jitter median jitter value for f_i, e.g. $j(D['minute' = 12])$ is the median of the jitter values of all observations in D for which 'minute'=12. For each minute-ranNode combination (m_i, r_i) the green line gives the value $\min(j(D[m_i]), j(D[r_i]))$ and the red line gives $\max(j(D[m_i]), j(D[r_i]))$. For example, let m_i be $'minute' = 12$ and r_i be $'ranNode' = 3$. Let $j(D['minute' = 12]) = 600$ and $j(D['ranNode' = 3]) = 900$ and $j(D['minute' = 12,' ranNode' = 3]) = 1000$. Then the green line will have a value 600, the red line a value of 900 and the blue bar will be above both values at 1000. As evident from the figure, such cases occur and therefore the minimum and maximum values would be a bad predictor for the combination and thus probes that are represented by a unique feature value combination are likely to follow a different distribution.

We make similar observations for the jitter median values for the feature pairs 'hubType' × 'apNode' (Figure 8).

3 Anomaly detection for different probe properties

The above discussion suggests that what constitutes an anomaly should be feature dependent. Recall that D is the given dataset, $f_1, \ldots, f_k$ a set of feature values and $D[f_1, \ldots, f_k]$ is the subset of the data with observation having the feature values $f_1, \ldots, f_k$. For example, $D[minute = 53, apNode = 17]$ is the subset of the data where the test started in 53rd minute and the probe has apNode 17.) Consider the following definition of an anomaly: a jitter value for a set of feature values $f_1, f_2, \ldots, f_k$ is said to be anomalous if it is at least $c \cdot \text{median}(j(D[f_1, f_2, \ldots, f_k]))$ for some constant $c > 1$.

In Figure 9 we plot the jitter values for $D[\text{minute} = 37, \text{apNode} = 7]$ (left) and $D[\text{minute} = 53, \text{apNode} = 17]$ (right). The median is given by a yellow line. For the first dataset it is below 300 and for the second, above 1000. As evident from the plots, for $c = 4$ this results in very different values for the outliers: in the first set a jitter value above 1,200 is considered to be an outlier and in the second, a value of 3000 is not an

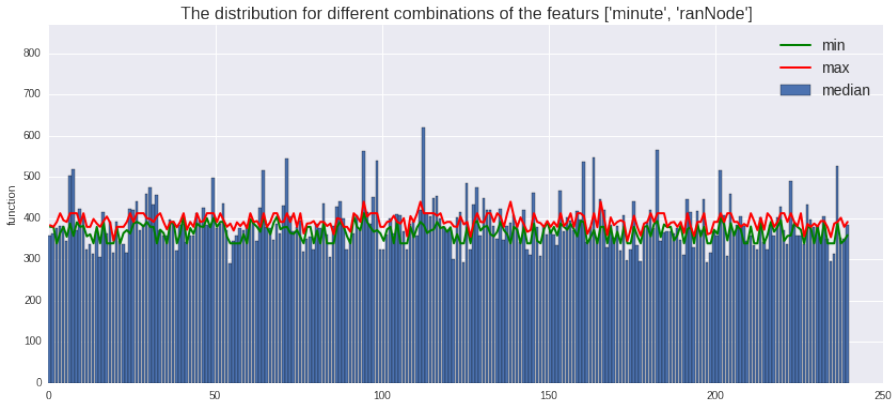

Figure 7: The median jitter for different 'minute' × 'ranNode' value combinations.

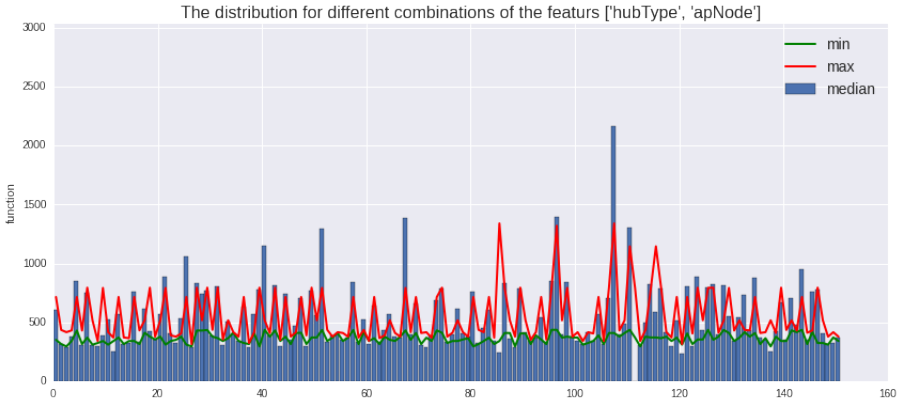

Figure 8: The median jitter for different 'hubType' × 'apNode' value combinations.

outlier.

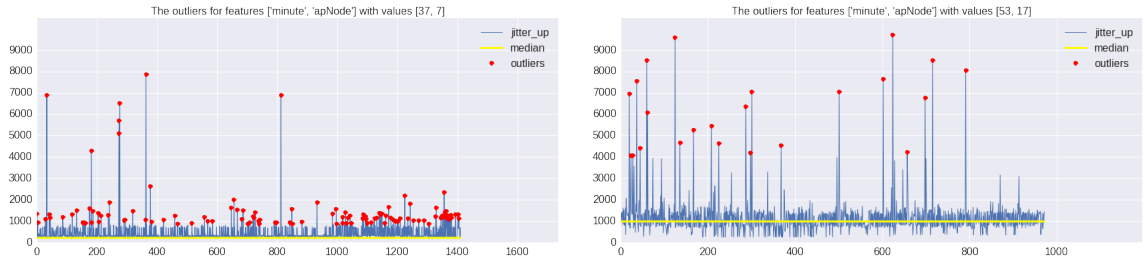

Figure 9: The anomalous data points for minute '37' and apNode '7' (left) and minute '53' and apNode '17'(right).

In Figure 10 we make similar observation for two different feature values of the features ranNode and hubType.

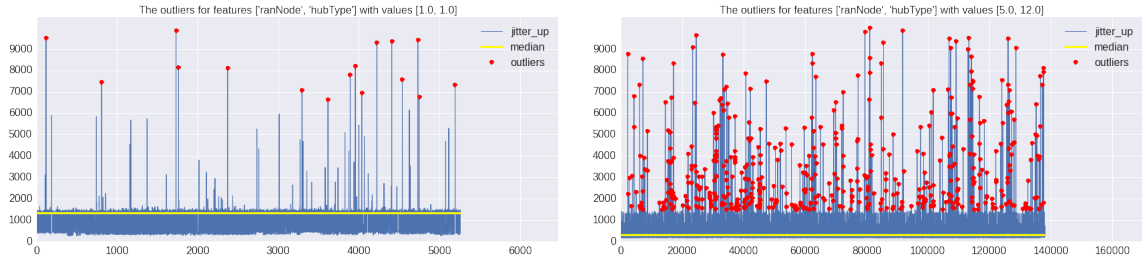

Figure 10: The anomalous data points for ranNode '1' and hubType '1' (left) and ranNode '5' and hubType '12'.

4 Interesting observations

After dividing the dataset into normal and anomalous data based on the BT labeling, we observe that certain feature values are much more likely to result in anomalies than others. In Figure 11 we plot the number of failures with a certain apNode value. It appears that almost all failures are observed in probes apNode=7 or apNode=11. Note that this distribution does not hold for the whole dataset, i.e., it is not the case that most observations come from probes with these apNode values. We observe a similar skewed distribution for the number of failures for different ranNode values in Figure 12. This distribution is again very different from the distribution of ranNode values over the whole dataset.

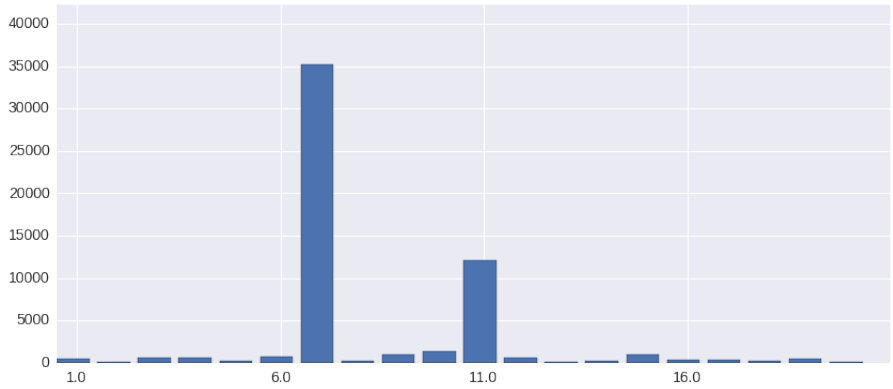

Figure 11: Failures per apNode.

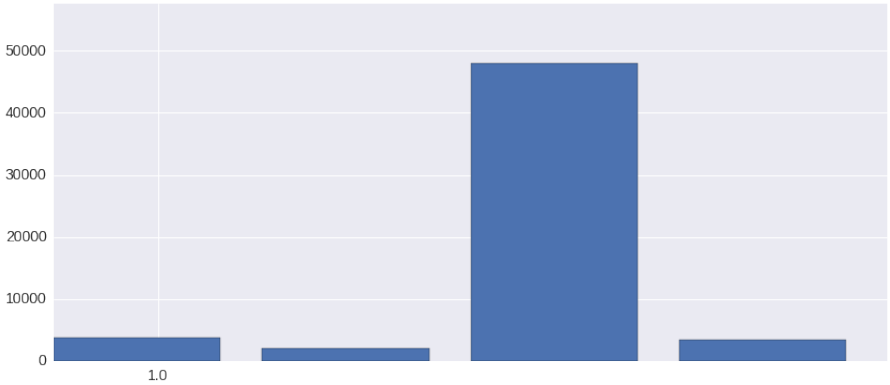

Figure 12: Failures per ranNode.