APPENDIX

Tuning System Parameters

After the choice of the grouping algorithm has been made, the next task is to tune system parameters in order to maximize performance. It is seen that some of these parameters could be generalized across user profiles, while the others revolve around the user traffic profile.

The following are the important system parameters in consideration:

1. Window Size (N);
2. Number of bands (K); and
3. Error Percent (t).

The optimal values for each of the above parameters were empirically determined, based on the following factors:

- Maximum Performance – lesser number of false alarms.
- Light-weight detection and quicker response – decrease response time.

Experiments on N and K

The objective of the experiment is to determine the ideal values of {N, K, t} at which the performance will be maximal. It is obvious that N and K are related to each other – K refers to the amount of compression allowed on the total number of events in the event space, which in turn, is determined by the value of N , for good performance. A higher value of N will result in more probabilities, which may (or may not) call for an increase/decrease in the value of K in order to accommodate performance requirements. But it is observed that although N and K are related to each other, the value of the two tuple {N, K} need not be tied to a particular user traffic profile, and can be generalized across user sites. However, N and K should be chosen in a way that the relation between N and K is still preserved across all user sites. On the other hand, t by definition is specific to the user traffic considered. Hence, t may have to be varied from site to site, and involves trial-and-error.

The performance is mentioned in terms of False Negatives (FN) or True Positives (TP). The parameter considered for measurement is the True Positive Rate (TPR), the ratio of

True Positive windows to the total number of Attack windows. The approach taken to finding the right {N, K, t} is as follows:

- The value of t is kept at a constant, and the system performance for different {N, K} combinations is observed to derive the relation between N and K.
- The value of t is chosen in a way that it exceeds the maximum value of t for all datasets available for experimentation. This is based on the fact that the system performance saturates at a maximum, beyond a value of t for any dataset. It is seen that t = 1% is a suitable value for the experiment. Precise values of t will be determined for every dataset separately.
- The varying values of N and K are taken from the following discrete values: N' = {50, 100, 150, 200} and K' = {5, 10, 15, 20, 30, 40}.
- A performance matrix, which describes the performance (TPR) of the system at each possible {N, K} combination from N' and K' is arrived at. The resultant matrix is a 6*4 matrix.
- The matrix is analyzed to determine {N, K} values for which performance is maximum. Bounds are determined on the values for which the system performance is guaranteed to get maximized. This will also bring to light the relation between N and K, if it is quantifiable.
- Based on the above observation, choose good values of N, K, and determine the exact value of t for every data set.

Experiment

The experiment is the same as the general experiment approach – for a chosen value of {N, K}, train with normal traffic and pass the training traffic into the testing phase in order to determine the number of False Positives. Train with normal traffic and pass the attack traffic into the testing phase, in order to determine the number of False Negatives. Determine TPR for each {N, K} and populate a cell of the performance matrix.

Results and Interpretations

The experiment was conducted (with t = 1%) for the above said values of K for two data sets – DARPA data set and the AUCK-II data set. The performance matrices for the two data sets are indicated in Figures A1 and A2.

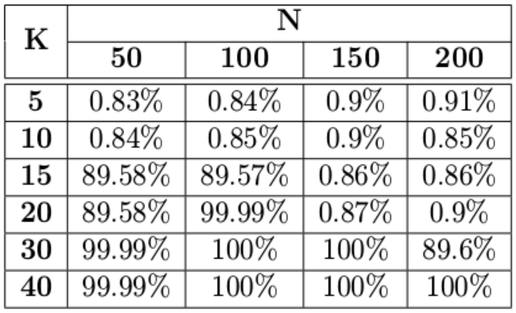
The following are the important observations:

• The performance matrix can be analyzed in two ways – Column wise analysis is done when N is fixed and a suitable K has to be arrived at, and row wise analysis to be done when the right N is chosen for fixed K. The former is more useful.

• The matrices show an important fact discussed above – for a given N , performance saturates after a particular value of K is reached, and addition of a few more bands does not improve the performance any more.

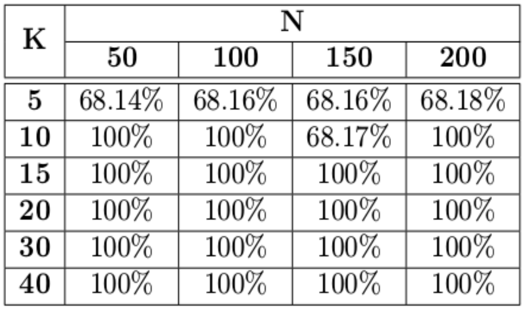
• The well tagged DARPA data set appears to have ideal properties – for even higher values of N , performance starts converging from small values of K. In essence, it means that the data set contains very few event occurrences for all N , which could be captured with fewer bands.

• From the results, it can be seen that when K is upper bound by N*5, the system performance peaks for almost all values of N and K in all three data sets.



K	N			
	50	100	150	200
5	0.83%	0.84%	0.9%	0.91%
10	0.84%	0.85%	0.9%	0.85%
15	89.58%	89.57%	0.86%	0.86%
20	89.58%	99.99%	0.87%	0.9%
30	99.99%	100%	100%	89.6%
40	99.99%	100%	100%	100%

Figure A.1: Performance Matrix for AUCK-II



K	N			
	50	100	150	200
5	68.14%	68.16%	68.16%	68.18%
10	100%	100%	68.17%	100%
15	100%	100%	100%	100%
20	100%	100%	100%	100%
30	100%	100%	100%	100%
40	100%	100%	100%	100%

Figure A.2: Performance Matrix for DARPA

Conclusion

Hence, it is hypothesized that K should be at least N*5 for the system to achieve good performance levels. The absence of constant values for K for any N is expected to create a few space overheads in hardware implementation (of detection), since N is expected to vary across multiple user sites, depending on their traffic profiles. This can be worked around by implementing detection assuming a reasonably large value of N , and choosing to keep N 5 bands for model probabilities. This space can be dynamically handled for all values of N lesser than the chosen value. This may amount to wastage of reasonably small space, which can be easily handled by modern hardware.

Experiments on Error Percent 't'

The objective of this experiment is to determine precise values of t for different data sets. The previous experiment hypothesized the relation between N and K, and experiments on t preserve this relation. For the purpose of these experiments, N is assumed to be 100, and K, according to our hypothesis is chosen to be 20. With these values fixed, experiment is conducted by varying the value of t until a saturation point in performance, in terms of True Positives is reached. Since it has already been explained that increasing values of t will increase false positives, we are concerned more about the attack traffic and how efficient the system is with respect to classifying attack windows as attack, with a given t.

Experiment

The following procedure was used to conduct the experiment:

1. Training phase: Train with normal traffic.
2. Testing phase: Pass as input the attack traffic, and estimate the True Positives for every t.
3. Repeat the experiment for different values of t, and find out the value of t at which the number of True Positives converges to a maximum. This value is the actual value for the error percent t.

Results and Interpretations

The results of the experiment are graphically described in Figure A3 and A4. The value of t differs from site to site. The number of bands K determines the level of accuracy of modeling, and should be carefully chosen during system design. N and K are related to each other.

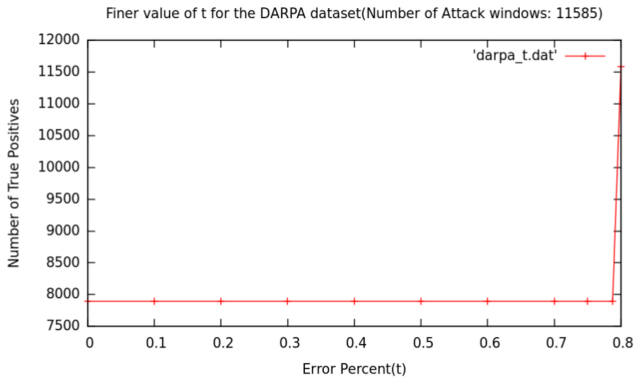


Figure A.3 Experiments to determine t (DARPA)

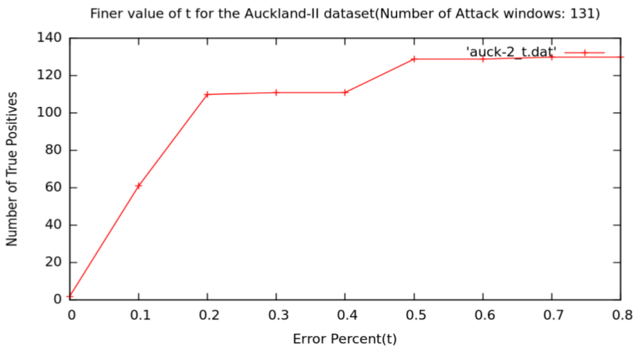


Figure A.4 Experiments to determine t (AUCK-II)