Results

There was are two major objectives/goals for this major research project. The first was to develop a Bayesian network model that is able to predict fraudulent credit card transactions with minimal errors and minimal false positive predictions. The second objective was to parallelize the Bayesian network inferencing process, so that the parallel version has a run time that is faster than the non-parallel version.

In order to validate the performance of the Bayesian network inferencing, I used a combination of a confusion matrix, Receiver Operating Curve (ROC), and the Area under the curve (AUC). Since the data set is highly class imbalanced, (i.e. the majority of the classes in the dataset are non-fraudulent), the ROC curve is an appropriate validation metric. A ROC graph provides information regarding the number of fraudulent transactions correctly classified (true positive rate) and the number of genuine transactions that were incorrectly classified (false positive rate). As you can see in Table 1, the Bayesian network was able to correctly classify fraudulent transactions, while producing zero false positives. As you can see from Figure 3 the model has a AUC value of 0.5968, which is greater than the baseline value of 0.50. These performance measures indicate that Bayesian networks can be used as efficient models to detect fraudulent credit card transactions.

Table 1: Confusion Matrix, Actual Classes Vs Predicted Classes

Actual \ Predicted	Non-Fraudulent	Fraudulent
Non-Fraudulent	34745	0
Fraudulent	50	12

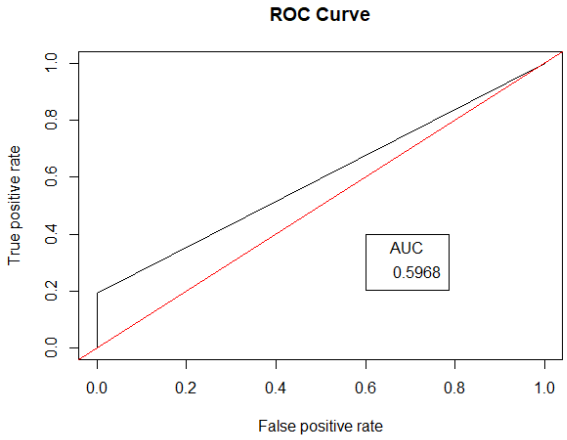

Figure 3: ROC Curve for The Bayesian Network

In order to compare the performance of the parallelized inferencing model to the non parallelized version, I computed the Karp-Flatt metric for various iteration values. The Karp-Flatt metric is a measure to evaluate the efficiency of the algorithm parallelization. The closer the value to zero the more efficient the algorithm parallelization is. Below is the formula for to calculate the Karp-Flatt Metric. Where n represents the number of machines and **Ψ** represents the actual speed up percentage.

$$e = \frac{\frac{1}{\psi} - \frac{1}{n}}{1 - \frac{1}{n}}$$

As you can see from Table 2, as the number of iterations increases the Karp-Flatt metric decreases. Indicating that the algorithm parallelization is more efficient as you increase the number of iterations. At the low iteration values the non-parallel version has a run time that is faster than the parallel version. This is the case because in the parallel version of the algorithm there is a time cost associated to communication between computing nodes. At the low iteration values this communication cost is larger than the benefit of parallelizing the algorithm. However,

as you can see from Figure 4, the benefit of parallelization surpasses the cost of communication when the iterations values become greater than or equal to approximately 12500.

Table 2: Karp-Flatt Metric for Various Iteration Values

Iterations	Non-Parallel Time (sec)	Parallel Time (sec)	Speed Up (Ψ)	Karp-Flatt Metric (n=3)
300	374	472	0.79	1.39
1000	447	481	0.93	1.11
10000	612	667	0.92	1.13
15000	842	788	1.07	0.90
21000	1027	983	1.04	0.94
30000	1252	1113	1.12	0.83
45000	1768	1248	1.42	0.56
60000	2268	1572	1.44	0.54
100000	3690	2135	1.73	0.37

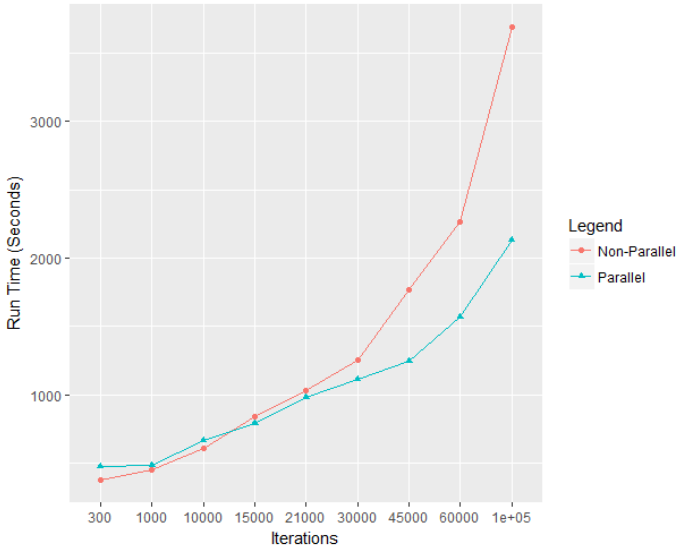

Figure 4: Time Comparison Between The Parallel Inference Model And The Non-Parallel Inference Model