Assignment 7

In assignment 5 you will use association rules and clustering to assess fraudulent behavior in simulated Medicare data.

Download the assignment data from Dropbox:
https://www.dropbox.com/s/t9kz2xeai6bw7in/assignment_5_data.zip?dl=0

Turn in a single word document to Blackboard with your ***brief but accurate*** answers.

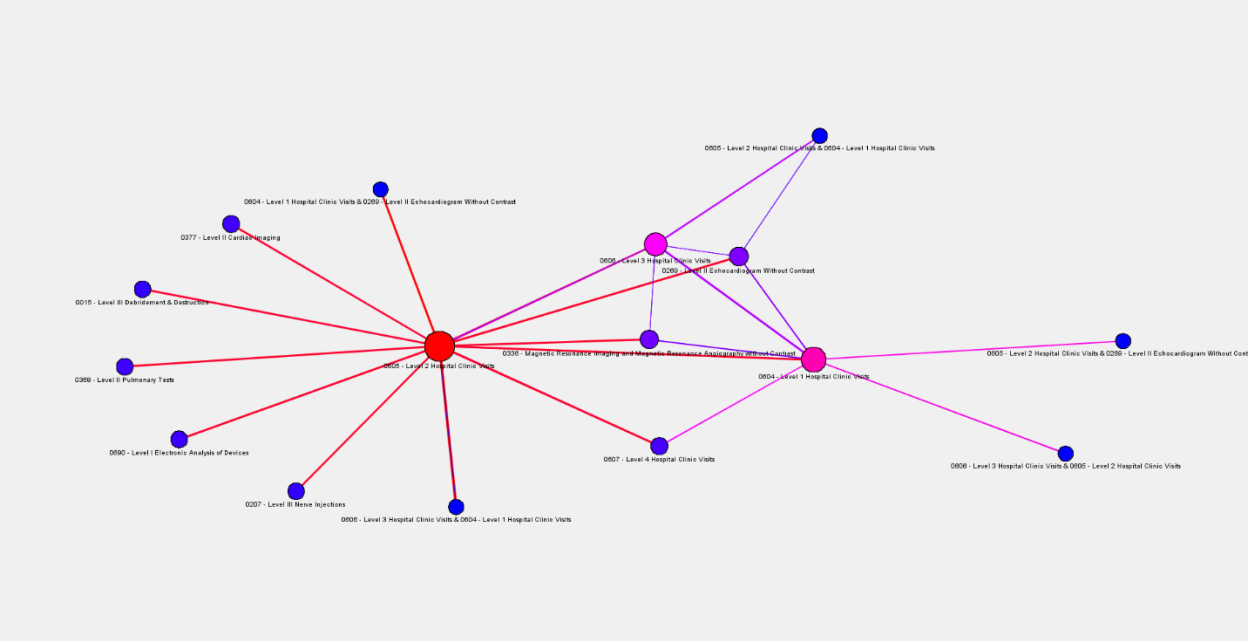
1.) Find fraudulent behavior patterns using association rules.

The transaction data set contains the medical procedures of a sample of patients from the larger, general population. While it may contain some fraudulent behavior, it presumably contains what could be called normal patient behavior.

The transaction_review data set contains the medical procedures of 5000 patients who have been manually labeled as involved in Medicare fraud.

The proc_key_map data set contains the English names of medical procedures. Please report results for question number one using these English procedure names.

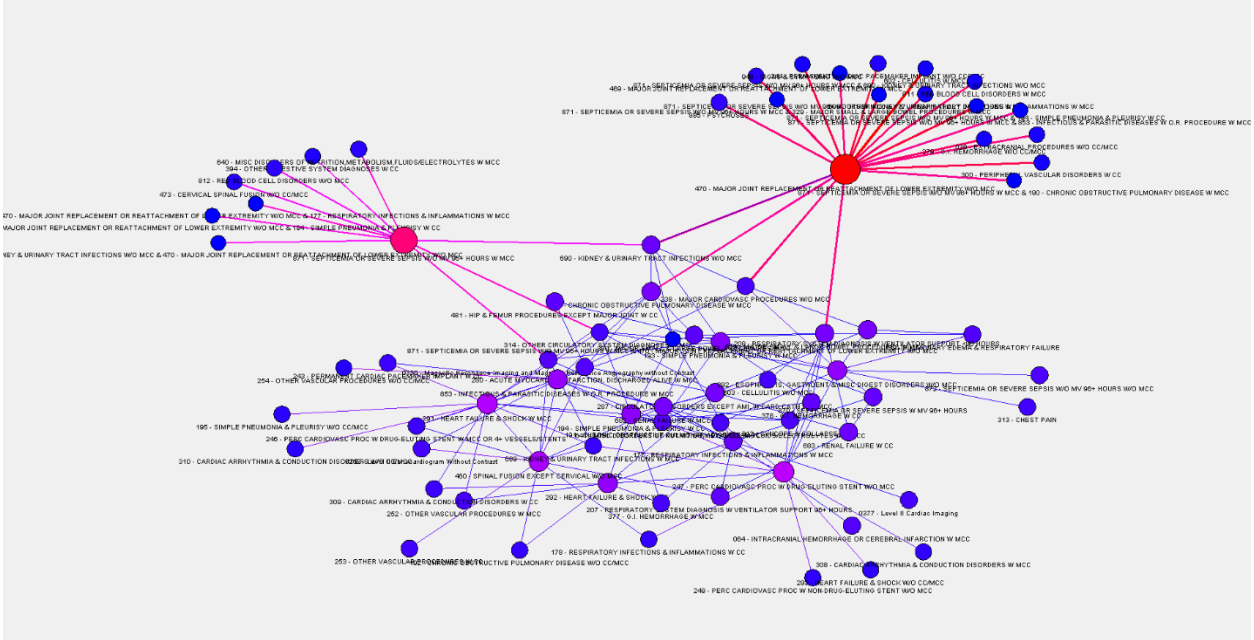
(2 pts.) *Using the SAS Enterprise Miner Association node on all defaults, create a link graph of normal patient behavior using the transaction data set and briefly describe the highest lift rules.*



The highest lift rules describe consistent medical behavior. A patient fills ill, go to the doctor's office (clinic) or the hospital and then has another routine procedure.

(2 pts.) *Using the SAS Enterprise Miner Association node, set minimum confidence level = 5% and support percentage = 1%. Create a link graph of fraudulent patient behavior using the transaction_review data set and briefly describe the highest lift rules.*

Assignment 7



The highest lift rules describe inconsistent patient behavior, in particular having multiple major, unrelated procedures, such as MRIs (w/o previous screenings), treatment for renal failure, and treatment for parasites.

(1 pt.) *How could these findings be used to detect fraud in the future?*

Fraud can be characterized from this data as patients who seek multiple major treatments that are seemingly unrelated.

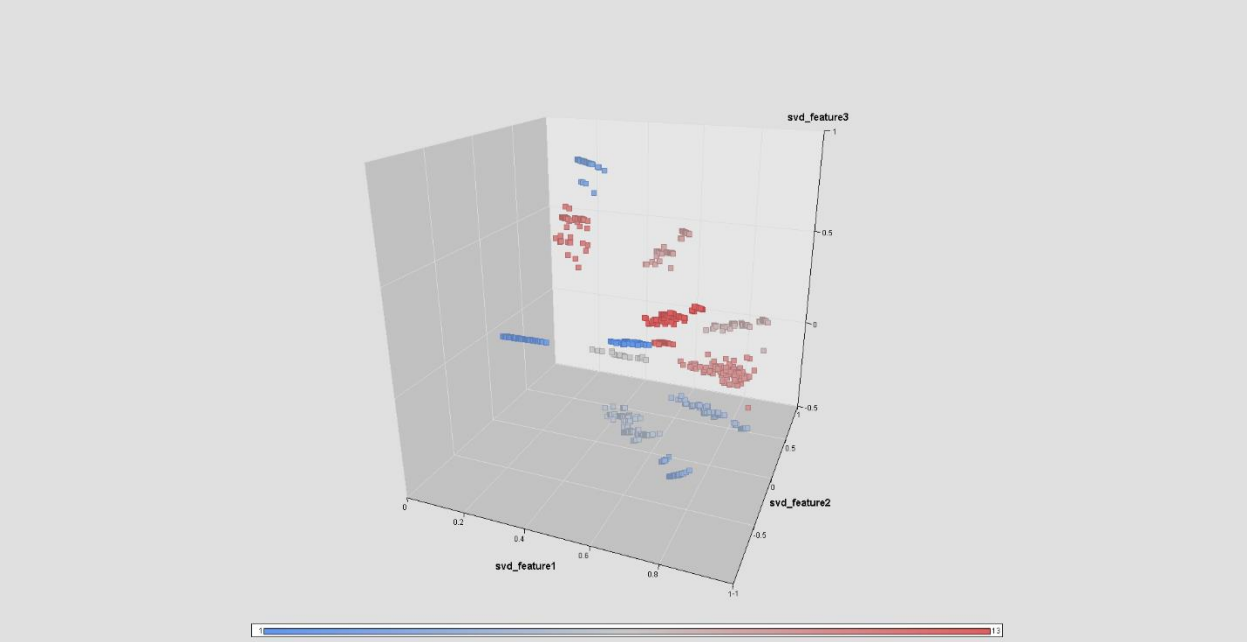
2.) Profile fraudulent patient populations using clustering.

The patient_history data set contains basic demographic information about the normal population sample and SVD features extracted from their procedures.

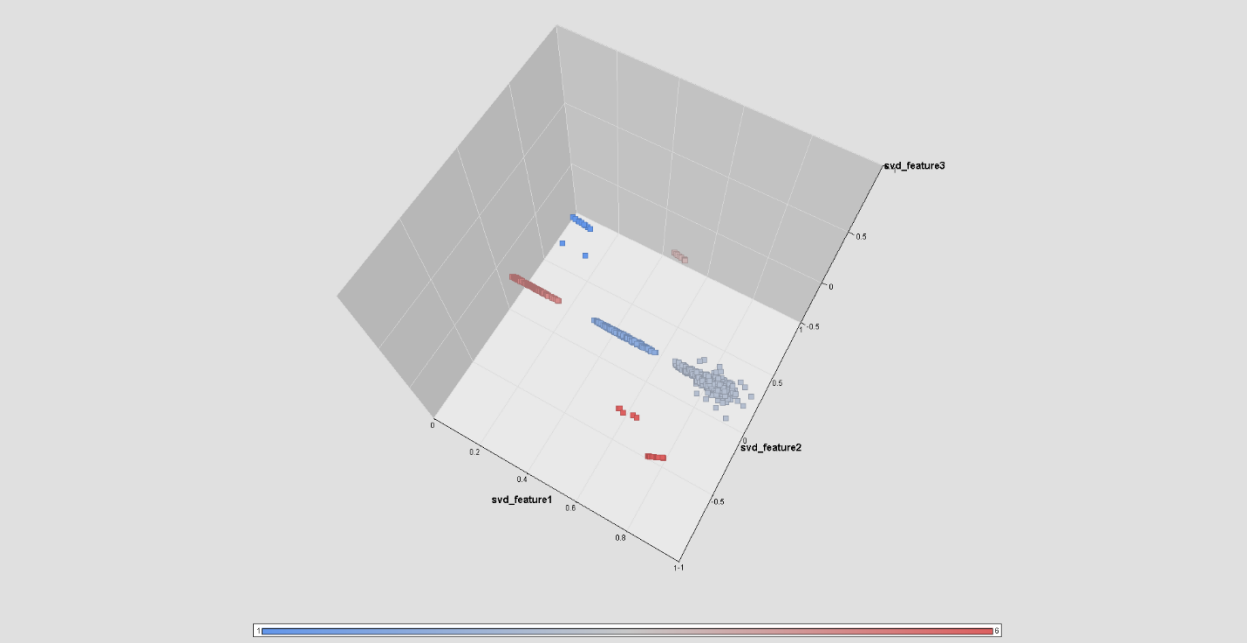
The patient_history_review data set contains basic demographic information about the fraudulent patient population and SVD features extracted from their procedures.

Since patient procedures can be used to detect fraud, build clusters in the SVD features already created from the transaction sets. Create 13 clusters in the patient_history set using the SAS Enterprise Miner Cluster node. Use only SVD_feature1, SVD_Feature2, and SVD_feature3 as clustering inputs. Aside from specifying 13 clusters, set standardization=none and seed initialization method=full replacement. Separately create 6 clusters in the patient_history_review set using the SAS Enterprise Miner Cluster node. Use only SVD_feature1, SVD_Feature2, and SVD_feature3 as clustering inputs. Aside from specifying 6 clusters, set standardization=none and seed initialization method= full replacement.

(1 pt.) *Explore the exported data from the cluster node. To confirm the clustering results in the patient_history data set, create a 3-D plot using SVD_feature1 as X, SVD_feature2 as Y, and SVD_feature3 as Z, and _SEGMENT_ as a color variable.*



(1 pt.) *Explore the exported data from the cluster node. To confirm the clustering results in the patient_history_review data set, create a 3-D plot using SVD_feature1 as X, SVD_feature2 as Y, and SVD_feature3 as Z, and _SEGMENT_ as a color variable.*



Use the following SAS code in a SAS Code node directly after the Cluster node(s).

```
proc freq
      data=&EM_IMPORT_DATA;
```

```
      table _SEGMENT_*age;
      table _SEGMENT_*gender;
run;
```

(1 pt.) *What can you say about the difference in* ***age*** *profiles between the patient_history and patient_history_review data sets?*

In the patient_history set age is remarkably homogenous across all clusters, while age varies a bit more across the patient_history_review set and one cluster there contains a noticeably higher percentage of young people.

(1 pt.) *What can you say about the difference in* ***gender*** *profiles between the patient_history and patient_history_review data sets?*

In the patient_history set gender is remarkably homogenous across all clusters, while gender varies a bit more across the patient_history_review set and one cluster there contains a noticeably higher percentage of females.

(1 pt.) *How could these findings be used to detect fraud in the future?*

As a group, fraudulent patients have fewer different clusters in the SVD feature space and the clusters are slightly less homogenous in terms of age and gender. The clusters tend to be very close to the SVD_feature2 axis as well. Any new clusters in that space that exhibited gender or age values that are noticeably different from the population distribution or that are very close to SVD_feature2 would be suspicious.

For an individual patient, they could be scored in both cluster spaces and see which cluster they fall nearest to. If they are closest to a cluster in the patient_history_review set then they would be suspicious.