FRAUD DETECTION

Sarith Divakar M

Department of CSE

L.B.S College of Engineering

(Managed by LBS Centre for Science & Technology - A Govt.of Kerala Undertaking)

Online Session: 04-04-2020

FRAUD DETECTION

- Typical examples for which fraud detection is relevant are: credit card fraud, insurance claim fraud, money laundering, tax evasion, product warranty fraud, and click fraud.
- A first important challenge in fraud detection concerns the labeling of the transactions as fraudulent or not.
- A high suspicion does not mean absolute certainty, although this is often used to do the labeling.
- Supervised, unsupervised, and social network learning can be used for fraud detection.

Supervised learning

- Labeled data set with fraud transactions is available.
- A common problem here is the skewness of the data set because typically only a few transactions will be fraudulent.
- Hence, a decision tree already starts from a very pure root node (say, 99 percent nonfraudulent/1 percent fraudulent) and one may not beable to find any meaningful splits to further reduce the impurity.
- Similarly, other analytical techniques may have a tendency to simply predict the majority class by labeling each transaction as nonfraudulent.

Common schemes: over and undersampling

- oversampling, the fraudulent transactions in the training data set (not the test data set!) are replicated to increase their importance
- In undersampling, nonfraudulent transactions are removed from the training data set (not test data set!)
- Both procedures are useful to help the analytical technique in fi nding a discriminating pattern between fraudulent and nonfraudulent transactions.

Adjust the predictions

$$p(C_i|x) = \frac{\frac{p(C_i)}{p_t(C_i)} p_t(C_i|x)}{\sum_{j=1}^{m} \frac{p(C_j)}{p_t(C_j)} p_t(C_j|x)}$$

whereby C_i represents the target class (e.g., C_1 is fraudulent and C_2 is nonfraudulent), $p_t(C_i|x)$ represents the probability estimated on the over- or undersampled training data set, $p_t(C_i)$ is the prior probability of class C_i on the over- or undersampled training data set, and $p(C_i)$ represents the original priors (e.g., 99/1 percent). The denominator is introduced to make sure that the probabilities sum to one for all classes.

Unsupervised learning

- Used to detect clusters of outlying transactions
- SOM and look for cells containing only a few observations that might potentially indicate anomalies requiring further inspection and attention.

Social network analysis

- Exploiting relational information provides some interesting insights in criminal patterns and activities.
- Legitimate nodes only sparsely connect to each other, fraudulent nodes are characterized by a dense structure, with many links between all the members.
- *spider constructions: F*raudulent constructions look like a dense web in which all nodes are closely connected to each other.

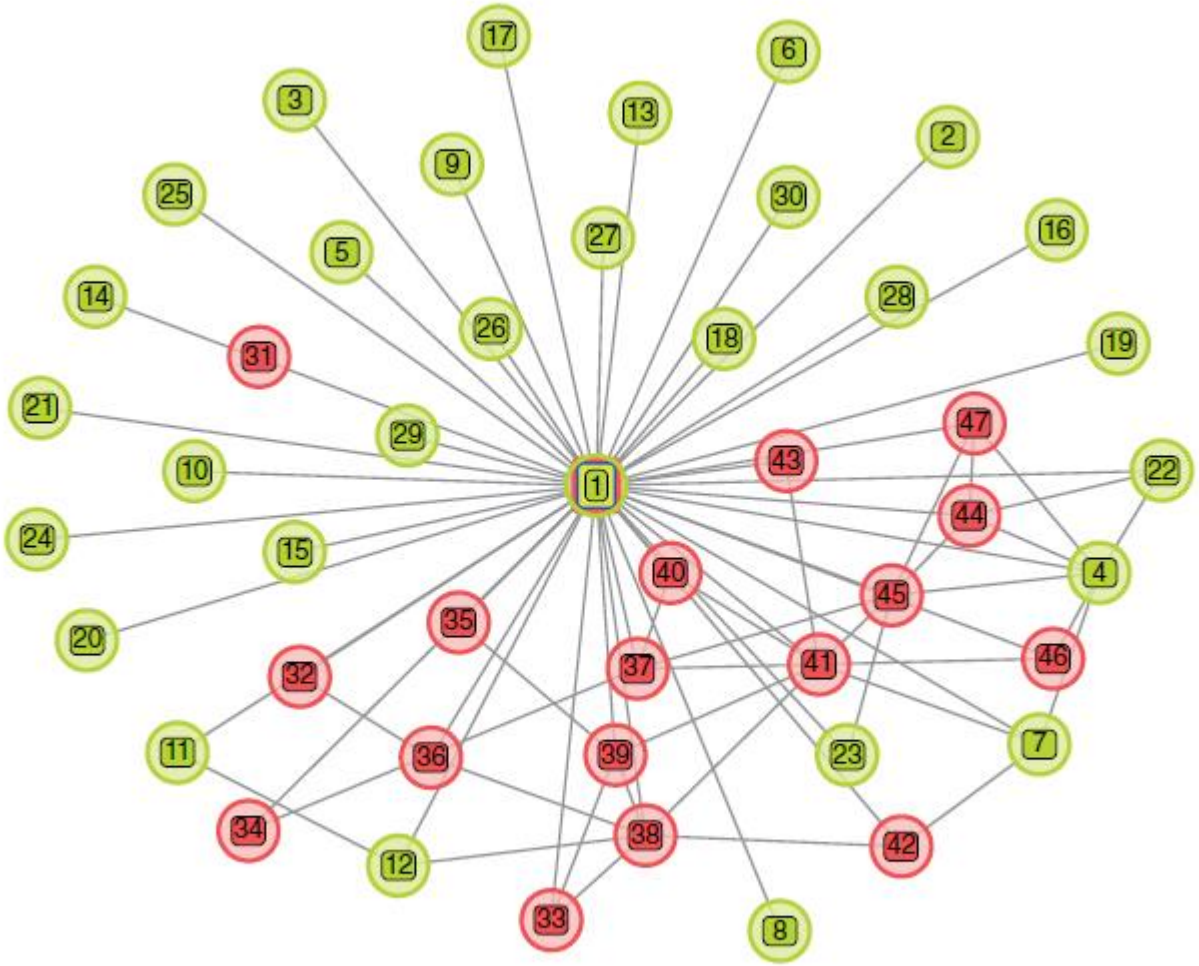

Figure 8.4 Fraud Network.
Light Gray Nodes Refer to Legitimate Individuals, While Dark Gray Nodes Represent Fraud

Network variables

Fraudulent degree

- In the network domain, the first-order degree refers to the number of immediate contacts a node has.
- The *n* -degree defines the number of nodes the surveyed node can reach in at most *n* hops.
- For the fraud domain, this means that the fraudulent first-order degree corresponds to counting the number of direct fraudulent neighbors

Triangles

- A triangle in a network is defi ned as a structure in which three nodes of the network are connected to each other.
- Especially triangles containing at least two fraudulent nodes are a good indicator of potential suspicious activities of the third node.
- Nodes that are involved in many suspicious triangles have a higher probability to commit fraud themselves.

Cliques

- A clique is an extension of a triangle
- clique as the maximal subset of the vertices in an undirected network such that every member of the set is connected by an edge to every other.
- While fraudulent triangles appear regularly in a network, fraudulent *k* -cliques (with $k > 3$) will appear less often.
- cliques are extremely precise indicators of future fraud.

Aggregated variables

- Although network variables as such can be very useful in detecting potential future fraud, these characteristics can also be converted in aggregated variables characterizing each node (e.g., total number of triangles/cliques, average degree weight, average triangle/clique weight).

Learning Techniques

- Using all the available attributes, standard learning techniques like logistic regression, random forests, and neural networks are able to estimate future fraud based on both network-related information and personal information