# Featurization

Featurization is one of the most important techniques to account for social network effects. Hence, I'll discuss it in more detail here. Featurization refers to mapping neighbor and network characteristics into features and combining them with the local variables for predictive modeling. This process involves three types of features. First, we can construct features that summarize the behavior of the dependent variable of the neighbors. We can also add features that aggregate the independent variables of the neighbors. Finally, we can also include network characteristics of the node itself.

Here you can see an example of featurization in a fraud detection context. Every customer is characterized by local variables such as age, recency, and so on. We also included features that summarize the dependent variable behavior of the neighbors. Hence, we only consider the fraud behavior of the neighbors. You might notice that we included the features that Lu and Getoor introduced, as discussed earlier.

Here you can see another example of featurization in a churn-prediction setting. The local variables are age, average duration, average revenue, and promotions received. In this case, the features summarize the independent variables of the neighbors, such as average age of friends, average duration of friends, and so on. The dependent variable of the neighbors (churn behavior) is disregarded.

As mentioned before, the third type of feature summarizes the network characteristics. The example network shown here illustrates this type of feature. Assume that we work in a fraud-detection setting. The black nodes represent the fraudsters and the white nodes are the legitimate customers.

Here you can see the network metrics summarized for each of the nodes A to J. Let's briefly discuss these. The degree represents the number of connections. It can be decomposed into the fraud degree and the legit degree, which represent the number of connections to fraudulent and legitimate customers, respectively. A triangle is a group of three nodes that are all connected to each other. A fraud triangle is a triangle where both connecting nodes are fraudulent. A legit triangle is a triangle where both connecting nodes are legitimate. A semi-fraud triangle is a triangle where one connecting node is fraudulent and the other one legitimate. The geodesic path is the shortest path to a fraudulent node. The number of 1-hop paths indicates how many fraudulent nodes we encounter in all possible 1-hop steps from the node. The number of 2-hop paths indicates how many fraudulent nodes we encounter for all possible 2-hop steps from the node. For node I, the 2-hop path number equals 6, which represents the six 2-hop paths I-K-I, I-A-I, I-N-I, I-G-I, I-G-D, and I-G-F. Each of these contains one fraudulent node. The number of 3-hop paths indicates how many fraudulent nodes we encounter for all possible 3-hop steps from the node. Finally, we can also include the closeness and betweenness values. Here is the rest of the table, which shows the metrics for nodes K to T.

To conclude, when you are doing featurization, it is important to include as many features as possible. Ideally, it is recommended to add all three types of features to the data: features representing the dependent variable characteristics of the neighbors, features representing the independent variable characteristics of the neighbors, and features representing network characteristics. Obviously, this will substantially increase the data set in terms of the number of inputs or predictors. Hence, input selection procedures should be used to select the most predictive inputs. The predictive model itself can then be constructed using any of the techniques that were discussed before, such as logistic regression, neural networks, SVMs, random forests, and others. As always, the choice will be made by considering the characteristics of the business problem.