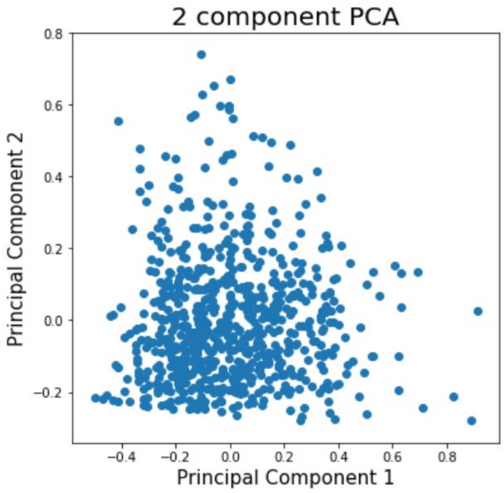
Outlier Detection using Random walks

Objective : Unsupervised learning model to detect Outliers in the Vehicle Insurance business. Detect Outliers among **708** Automotive dealers in the US based on their repairs charges associated with the vehicle insurance business. The analysis and study is done on behalf of the auto manufacturer who supplies the parts and pays the repair costs to the dealers, in order to minimize insurance payout cost to the manufacturer



Exploratory data analysis and Feature engineering : We look at the data of contracts from 2016 - 2019 (4 year contracts) and perform feature engineering to identify key features that identifies poor performance amongst the dealers.
The Key variables identified for analyzing the performance of each dealer are :

- ❖ **Loss Ratio** :
    - ➢ Total Repair cost against an ESB Contract/Total premium under the contract
- ❖ **Average parts repaired per contract** :
    - ➢ Average number of parts repaired under a contract by a dealer for a vehicle
- ❖ **Repair cost ratio(6 months)** :
    - ➢ The total repair cost incurred in the last 6 months of the contract period vs the total repair cost for the entire period of the contract
- ❖ **Claim to contract ratio** :
    - ➢ Number of claims against a contract.
- ❖ **Concentration of part repair :**
    - ➢ Score computed as concentration of repair on a part by a dealer in comparison with the repairs of other dealers, taking into account the average number of repairs warranted by the part.
- ❖ **Overcharged part cost :**
    - ➢ Score computed as the net cost of repairs charged above the average repair cost associated with a part across all parts repaired by the dealer across all contracts.

Stochastic modelling for outlier detection: **As can be seen from the plot of data above after applying PCA to reduce into 2 dimensional space, the majority of the dealers are concentrated on a central cluster.** The data is modelled as a stochastic graph where the data points associated with each dealer forms a node and edges weights are determined by the **inverse of euclidean distance between two nodes**.
The model is envisioned as a random walk on the graph network and the aim of the model is to determine the stationary distribution associated with each state(which is the probability of visiting a node).
Nodes in high dense regions will have higher probabilities of being visited.

The **similarity** between two nodes is determined by **the number of shared neighbours**. Consider objects, p1 and p2. Suppose p1 has a set of neighbors {p3, p4, p5, p7} and p2 has a set of neighbors {p3, p4, p6, p7} . The set of neighbors shared by both p1 and p2 is {p3, p4, p7}. In this algorithm, we take the cardinality of this set as the similarity measure.The transition probability matrix 'S' is defined as :

$$S[i,j] = \frac{Sim[i,j]}{\sum_{k=1}^{n} Sim[i,k]}$$

The **Markov chain** thus formed is clearly **irreducible**(connected). We make the chain **aperiodic** by adding a small damping factor **d**'. This scenario can be viewed as a random walk on a Markov chain where the walker visits one of the adjacent states with probability (**1-'d')** or requests another random state with probability 'd'. Outliers will have a fewer number of nodes than the normal objects in general, and this further helps to isolate outliers. In fact we can see that the similarity measure used here is the number of high-similarity shared neighbors. The **stationary distribution** associated with the irreducible, aperiodic Markov chain is then determined .

Nodes which are in the dense regions will receive votes from neighbours which are themselves highly connected and will therefore have a higher chance of being 'visited' in the random walk model
A low probability indicates a high chance for a dealer to be an Outlier and all the dealers are ranked according to their probability of being an outlier. The method is similar to the page rank algorithm.

Results : The dealers are ranked according to their probabilities of being an outlier. The analysis is validated by performing a **standard Z-score computation** on each of the features to verify if the identified .
From a Total of 708 dealers investigated across all states in the US, the net potential revenue gain if the performance is standardized for the top ten outliers is 191 k USD