Clustering Clustering Which Clustering Should You Choose ??!!

(Comparison of Clustering Techniques on Network Datasets)

Anirudh Narasimhamurthy , Padmashree Teeka, Roshani Nagmote
CS 6140: Data Mining, University of Utah, Spring 2015

ABSTRACT

In this digital age of Facebook, Amazon, Twitter, Wikipedia, Pinterest, Yelp and other numerous online social networks , finding groups of individuals in other groups can be profitable for companies, intellectually satisfying for researchers or in case of a graduate student both of it, as it could potentially prove to be profitable for a project submission as well as one gets to understand better the concept of “clustering”

PROBLEM AND MOTIVATION

➢ Clustering is a fundamental part of explorative data mining and a very useful technique for statistical data analysis used in many fields.

➢ Finding the most effective clustering technique for grouping users in a large network.

➢ Network Datasets used : LiveJournal (Online social network), DBLP (Collaboration network), Gnutella (Internet peer to peer networks), Simulated Dataset

KEY IDEAS

➢ Deciding on which three clustering techniques to use for network datasets and compare their performances based on the cost and time parameters.

➢ Selecting hamming distance as the metric for measuring similarity.

➢ Implementing scalable algorithm for larger datasets.

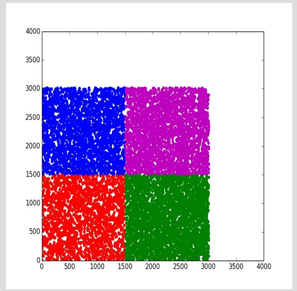
RESULTS

Dataset- GNutella (Internet peer to peer networks)
(Nodes: **6301** Edges: **20777** Number of Clusters: **15)**

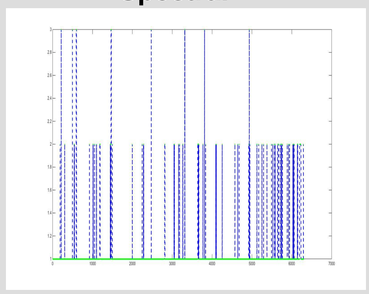
Clustering Algorithms	K-Means	Spectral	Hierarchical
Runtime	19s	2s	200s
Scalability	All datasets	All Datasets	Small Datasets
Suitability	General Purpose, Less number of clusters, Even cluster size, Flat Geometry	Graph Datasets, Few Clusters, Even cluster size, Non-Flat Geometry	Clusterable Datasets, Many clusters, Non Euclidean distances
Cost	7.3933*10^9	157.9451	-

Visualization of Cluster outputs

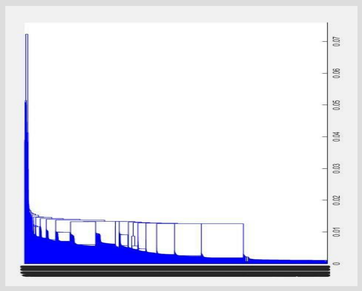
K-Means



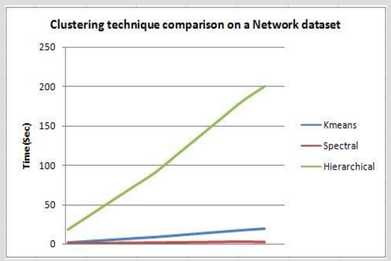
Spectral



Hierarchical



Comparison of clustering techniques using time

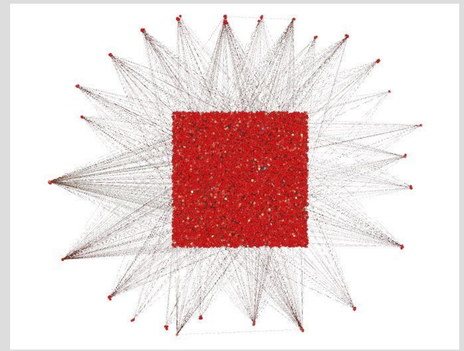


OBSERVATIONS

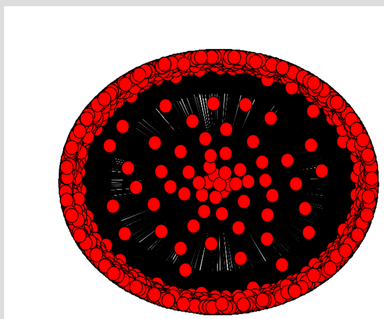
➢For K-means, cost decreased with increasing number of clusters.
➢For spectral, cost increased with increasing number of clusters.
➢**Spectral worked best for our network datasets**

INTERESTING FEATURE

Top k highly connected nodes/users in the given network dataset.



Neighbors of the most influential node/person in the given network.



This information can be potentially used for disseminating useful information across different user groups in a network.