Chapter 0: Apply Graph Theory on Big Data

Big data can be defined as various descriptions by academic and industrial fields. In statistics, big data is focused on analysis of large data set. Also big data is about collecting, storing, and preprocessing of huge and complex data in computer science. On the other side, big data is managed for making and implementing good decisions. Generally big data has three features: volume, variety, and velocity.

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. Interested reader may read the article: J. Manyika, et al., Big data – The next frontier for innovation, competition, and productivity, *McKinsey Global Institute* (2010).

Techniques for analyzing big data:

There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyze datasets. Here, we provide a list of some categories of techniques applicable across a range of industries (in alphabetical order):

A/B/N testing; Ensemble learning; Genetic algorithms; Machine learning; Natural language processing; Neural networks; Network analysis; Optimization; Supervised learning; Simulation; Time series analysis; Unsupervised learning; Visualization.

Following we shall show two real cases:

0.1 Patent Big Data Analysis

Material cited from: S. Jun, Patent Big Data Analysis using Graph Theory, *Advanced Science and Technology Letters*, **85** (Information Technology and Computer Science 2015), 25-28.

Patent documents are considered as big data, because patent data also have three characteristics of big data.

The collected patent documents are transformed into transaction-thing matrix (TTM) for big data analysis, because most analytical tools of big data such as statistics as well as graph theory need to structured data type consisted of row and column. The row and column of our TTM are transaction and thing, respectively. Using the TTM, we get visualization of TTM by graph theory.

Case study: We searched 100 patents related to graphic user interface (GUI) technology from KIPRIS (Korea Intellectual Property Rights Information Service).

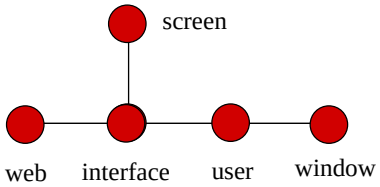
The following table shows the TTM of 10 transactions (patents) of 100 patents:

interface	screen	user	web	window
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
0	0	0	0	0
0	0	0	0	0
1	0	0	0	0
1	0	1	0	0
0	0	0	1	0
0	1	0	0	0
0	0	1	1	1

Using a statistical software R on that 100 transactions, we get the following correlation coefficient matrix of TTM

	interface	screen	user	web	window
interface	1.00000	0.12312	0.62770	0.03964	-0.03430
screen	0.12312	1.00000	-0.03342	0.03254	-0.02272
user	0.62770	-0.03342	1.00000	0.02437	-0.03703
web	0.03964	0.03254	0.02437	1.00000	-0.03357
window	-0.03430	-0.02272	-0.03703	-0.03357	1.00000

From the above data (considering the large number at each row besides the diagonal) we visualize the relationship by the graph below:



Using this result, and we can perform Research and Development planning for new technology development related find vacant or emerging areas of target technology. Finally we develop new technology development related to GUI. Here we only use five keywords as vertices of graph structure. We may use more keywords for patent analysis and visualization and get more diverse results for Research and Development planning in patent management.

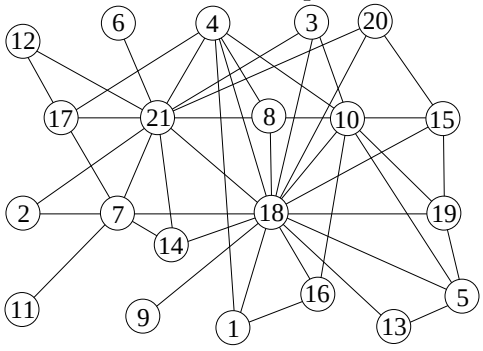
0.2 Truss in Massive Networks

Material cited from: J. Wang and J. Cheng, Truss Decomposition in Massive Networks, *Proceedings of the VLDB Endowment*, **5(9)** (The 38th International Conference on Very Large Data Bases, August 27th - 31st 2012, Istanbul, Turkey), 812-823.

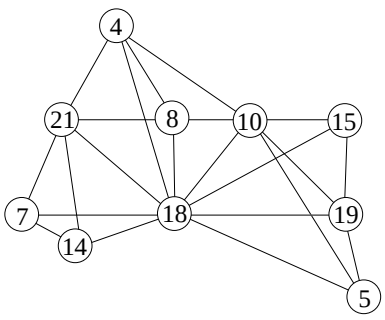
Social network: Many companies have recently discovered the value of data from social media. Telecom companies have found that some information from social networks is useful in predicting customer churn. They discovered that customers who know others who have stopped using a certain telecom are more likely to do so themselves, so these likely-to-churn customers are then targeted for retention programs. Other consumer-facing companies have found that they can learn about customers' attitudes, buying trends, and taste from sentiments expressed online, allowing them to make timely changes in their marketing, and going forward changes in their product planning.

Given a graph G, the *k-truss* of G is the largest subgraph of G in which every edge is contained in at least $(k-2)$ triangles within the subgraph.

In a social network, a triangle implies a strong tie among three friends, or two friends having a common friend. Intuitively, we may consider this as in a social network, the more common friends two people have, the stronger their connection it implies. So researchers study the k-truss of a social network.



(a) A manager-relationship graph G.



(b) The 4-truss of G (no 5-truss exists).