CSCI4190 Course Project

Social Network Analysis

-Task 3: Simulate cascading behaviors-
-in networks-

Chung Tsz Ting 1155110208

Content Page

1. Abstract

This project is an analysis of the simulation of cascading behaviors in networks. This is based on the human behavior of referencing decisions that are done by other people. To study the behavior of information cascade in social networks, the dataset from the Stanford Network Analysis Platform (SNAP) project of Stanford University is used as the network in which each of the nodes simulates the decision of a person. Moreover, the analysis is conducted through the tool of SNAP.py developed by Stanford University. From the result of the analysis, the influence of payoff value, the proportion of initial adopters as well as the total degree of initial adopters to the virality of cascading are highlighted. Meanwhile, clusters are also shown to be a natural obstacle to cascade.

2. Objective

Social network is a social structure from social actors and relationships between them in which the former is represented by nodes while the later is represented by dyadic ties. In the meantime, information cascade is a common practice in our daily life which poses influence in different aspects of our life, like social networks, political positions, and business decisions. For example, Kaschesky and Riedl propose a model for the analysis of the political opinions' formation on the Internet and highlight the sentiment and idea adoption from the structure and information flow across the network[1]. Cascading in networks is the phenomenon of individuals being influenced by their particular network neighbour, this is also the study target of this project. Because of the significant influence of information cascade, this project examines how the cascading procedure is ongoing in a network and simulating the procedure for a new innovation be introduced and cascaded in a market from the formation of the social force of conformity. In the meantime, the effect of different parameters (for example, threshold calculated from payoff, the portion of initial adopters as well as the way of selecting initial adopters) affecting the virality of cascading is also examined. Finally, the relation between cascading and clusters will also be examined.

3. Dataset

This project uses the Slashdot Social Network Dataset[2] to simulate the cascading between individuals. This dataset is an online social network dataset in which edges represent the interaction between individuals. The dataset is available in the Stanford Large Network Dataset Collection obtained in February 2009, and the detail statistics are as below,

Dataset statistics	
Nodes	82168
Edges	948464
Nodes in largest WCC	82168 (1.000)
Edges in largest WCC	948464 (1.000)
Nodes in largest SCC	71307 (0.868)
Edges in largest SCC	912381 (0.962)
Average clustering coefficient	0.0603
Number of triangles	602592
Fraction of closed triangles	0.008168
Diameter (longest shortest path)	11
90-percentile effective diameter	4.7

From the statistic, we can see the dataset is a giant component and the majority of nodes are connected. After knowing the basic information of the dataset, the degree distribution of the network is investigated.

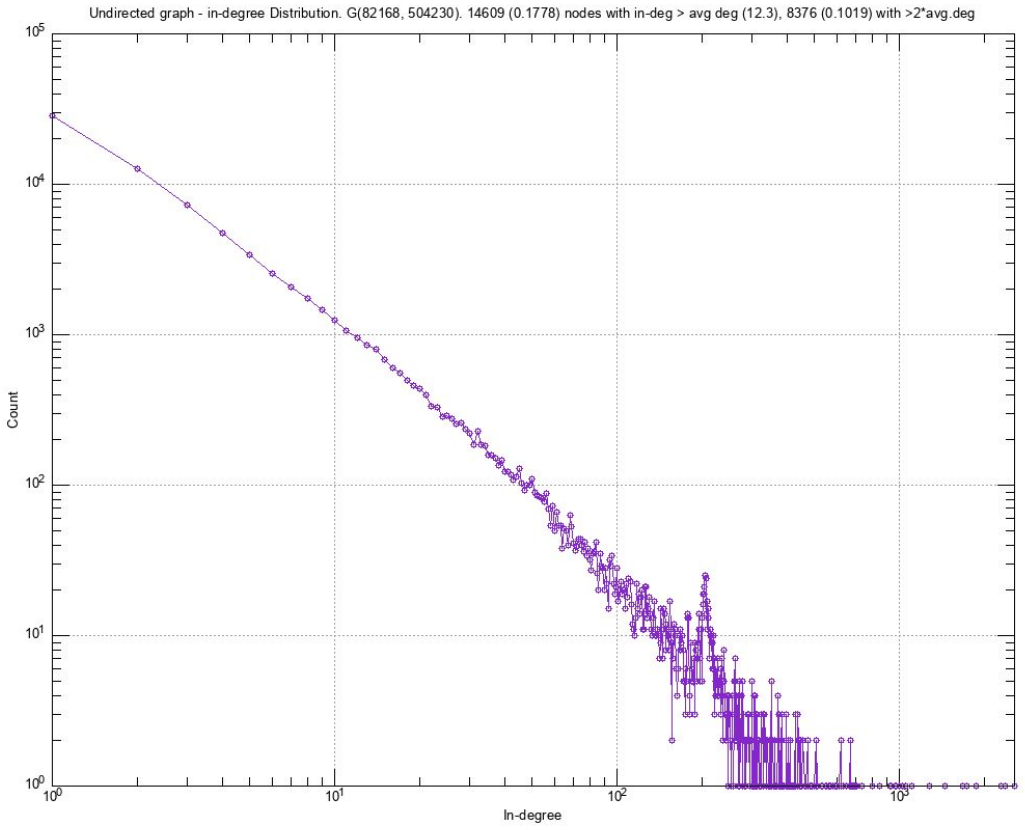

Fig.1. Degree Frequency Distribution and Summary

Maximum degree	2552
Minimum degree	1
Average degree	12.273
Edges in largest WCC	948464 (1.000)

This graph is generated from the PlotInDegDistr function from snap for the type of undirected graph. From the result, it can be seen that the degree and the number of nodes with the degree are inversely proportional.

4. Methodology

4.1. Tools for analysis

4.1.1.SNAP

SNAP is a general Social Network Analysis(SNA) tool and graph mining library to examine giant networks. This project adopted snap.py which is a python interface for snap and conducted in MacOS. The requirement to reproduce this task result in MacOS requires brew installation of python 3 and snap installation from pip.

4.1.2. Matplotlib

To give visualization on the result of the virality of cascading as well as the cluster density analysis, Matplotlib is adopted. It is a python library that provides interactive visualization. The code for visualization as well as code for cascading is separated into two files for the convenience of reproducing the task result in different aspects.

4.1.3. Numpy

Numpy is a python package for scientific computing which is used for assisting graph plotting in this project.

4.2. Cascading Procedure

All the nodes in the network belong to class ‘0’ at the initial start. To select initial adopters for the cascading in the network, different approaches described in detail below are used to initialize a certain proportion of nodes as class ‘1’ which is the new innovation to be marketed over the network. For every iteration of cascading, all nodes in the network will be checked. If the fraction of the neighbour of the node in adopting the new product is greater than a threshold q defined below, the node will

change from class ‘0’ to class ‘1’. When the class distribution (i.e. the ratio between nodes in adopting different classes) remains unchanged after an iteration, the iteration will come to an end.

4.3. Analysis Cases

This project considers the information cascading performed in an undirected network that every node (i.e. individual) can influence each other.

4.3.1. Random Two-Classes Initial Adopters

	A	B
A	a,a	-c,-c
B	-c,-c	b,b

In real-life marketing, when a new product A is introduced into a market with existing product B, value a is assumed as the payoff of people in adopting product A for a person while the neighbor of that person is also adopting product A. Vice versa, the payoff for that person adopting product B while his/her neighbor is also adopting B is b and payoff for the person in adopting product different than his/her neighbor is -c. Thus, we will have the following equations.

$$pda - (1-p)dc \geq (1-p)db - pdc$$
$$pa + pb + 2pc \geq b + c$$
$$p \geq \frac{b+c}{a+b+2c} = q$$
$$p \geq \frac{b}{a+b} = q \text{ when } c = 0$$

where p is the fraction of the neighbour of the node in adopting the new product and d is the number of neighbour of the node

For a different combination of values of variables a, b, and c, there are different q values, i.e. [0, 0.05, 01, 0.15 … 0.95, 1] . Therefore, a range of q values from 0 to 1 have been examined in the analysis of the effect on cascading corresponding to different a,b, and c, the result is shown in the below graph. In this case of analysis, we consider the marketing approach as randomly selecting the initial adopters for a range of proportion from 0 to 1, i.e. [0, 5%, 10%, 15% … 95%, 100%]. The higher value of the new product and the larger penalty for people adopting products different from their neighbor, the lower q value.

4.3.2. Top N degree Two-Classes Initial Adopters

Similar to the above case, different q values and proportions of initial adopters are examined for the effect of cascading while nodes (i.e. individuals) are not randomly selected this time. The degree of all nodes in the network is examined and the nodes with top N degrees are taken as the initial adopters, i.e. the n% of nodes with the highest degree.

4.3.3. Clustering

To investigate the relationship between clusters and complete cascading, two theories below are investigated for different portions of initial adopters as well as different threshold q.

i. If the remaining network after cascading contains a cluster of density greater than $1 - q$, then the set of initial adopters will not cause a complete cascade.
ii. Moreover, whenever a set of initial adopters does not cause a complete cascade with threshold q, the remaining network must contain a cluster of density greater than $1 - q$.

The former theory shows that clusters are obstacles to cascades while the latter shows that clusters are the only obstacles to cascading.

To investigate theories, this project examines the later theory by treating the cascading network as a cluster and the remaining network as another cluster. Because of the theory that the union of two clusters of density p is also a cluster of density p, even if the remaining network contains 2 separate clusters, we can still put the union of them under examination of the above theory. Then, the cluster density is calculated for the later cluster (i.e. remaining network) for different portions of initial adopters as well as q value and checked whether its value is greater than 1-q if there is no complete cascading.

5. Result

5.1. Random Two-Classes Initial Adopters

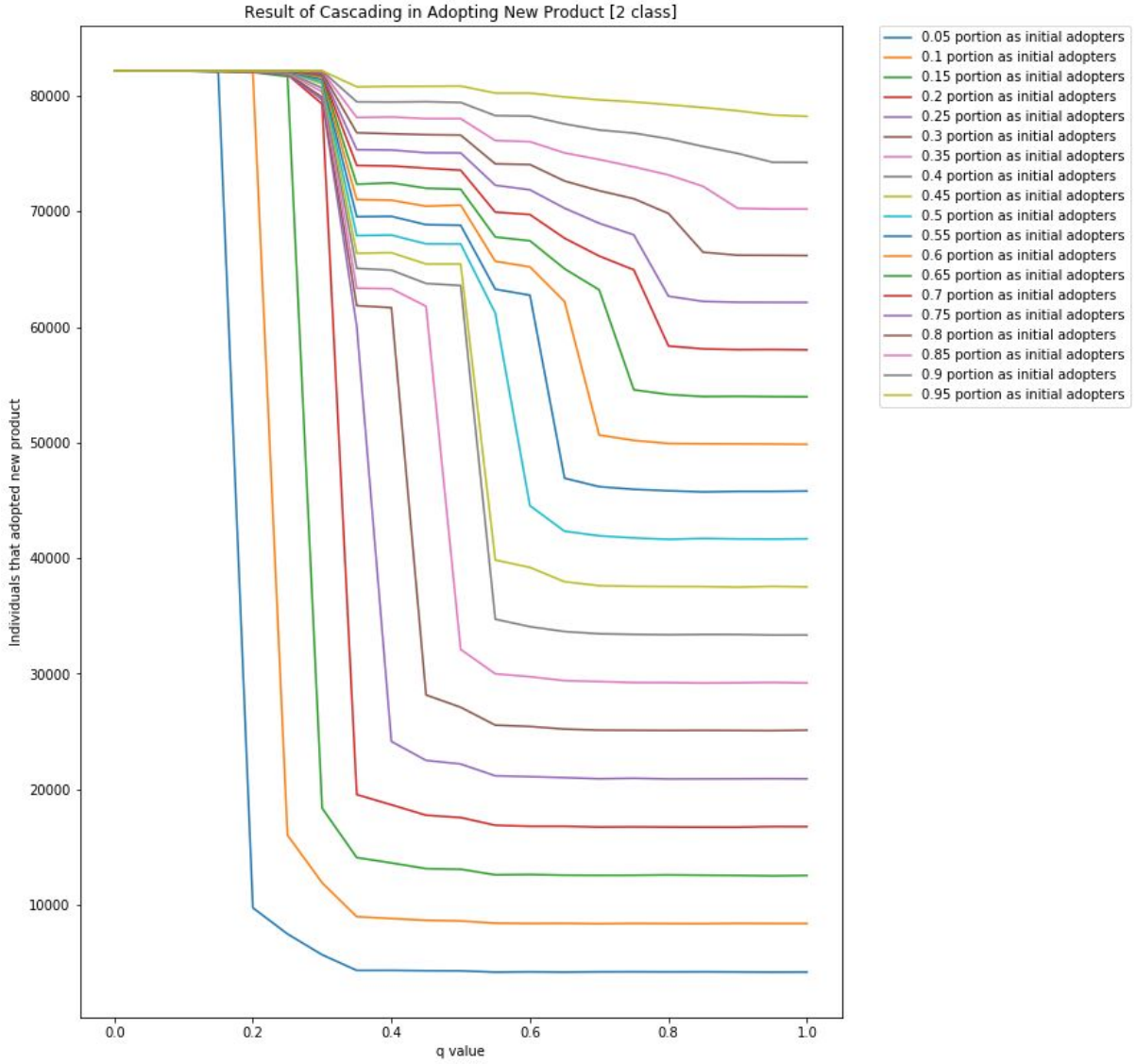

Fig.2. Overview of Result for Case1

From the above plot, the relation of q value and proportion of initial adopters with the effect of cascading is shown. It is observed that both parameters affect whether a complete cascade can be achieved. For a higher proportion of initial adopters and a lower q value, the cascading can cover more nodes and become more complete. However, in reality, businesses could not easily enhance the number of initial adopters. Instead, the market share may even be the target of their marketing campaign. Therefore, focus on the q value, it is observed that most cases can attain complete cascading when q value is smaller or equal to 0.2 for the proportion of initial adopters higher than 10%.

Moreover, the choice of initial adopters actually influences the completeness of cascading, for random selection, it can form a complete cascade for some run and cannot form for some run. All the graphs' results are the result from a specific run.

5.2. Top N degree Two-Classes Initial Adopters

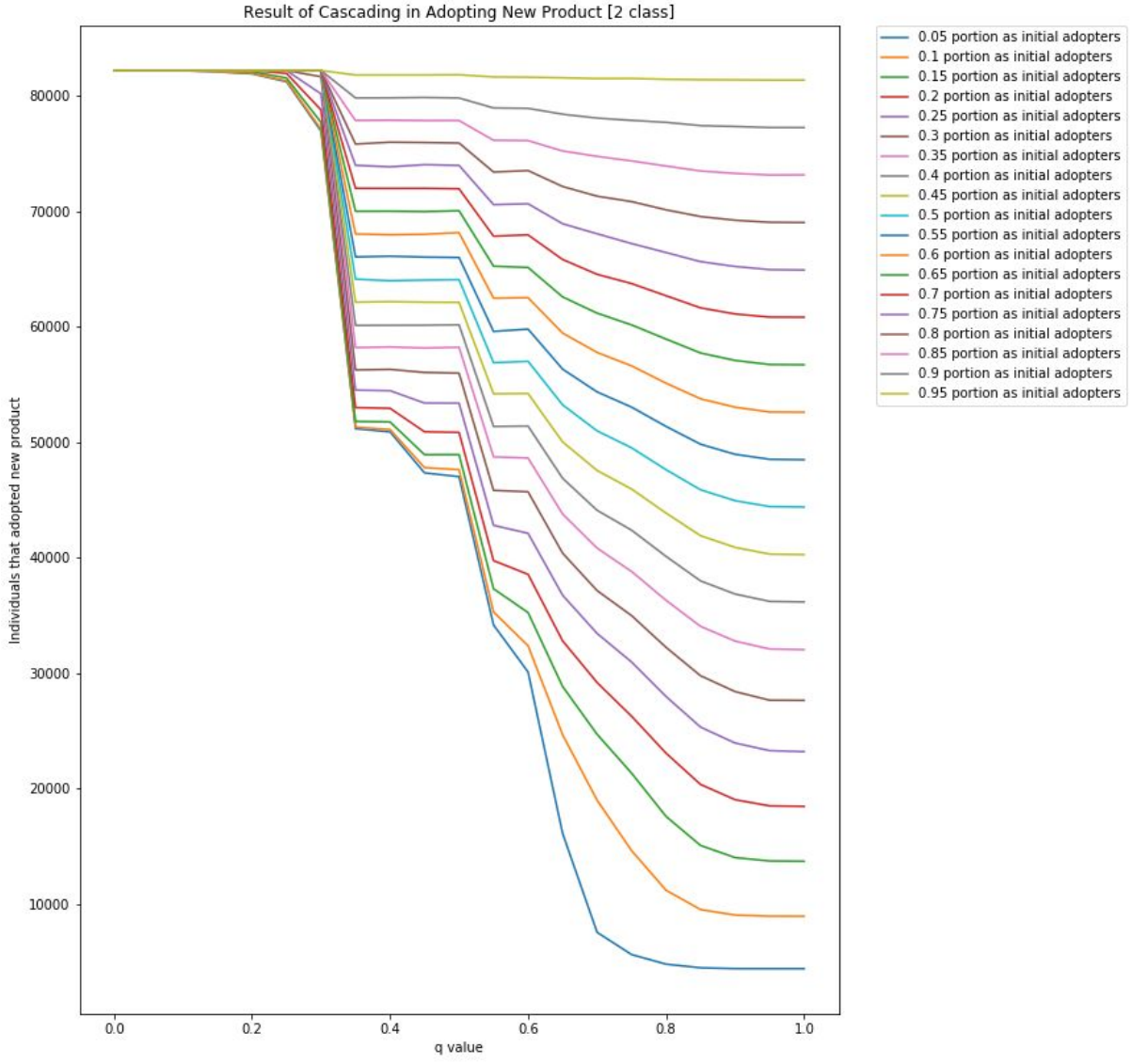

Fig.3. Overview of Result for Case2

This case chooses the top degree nodes as initial adopters and it is observed that the cascade virality is significantly higher than the previous implementation. Moreover, observing from a larger q value to a smaller one, the virality of cascading converges faster to attain a complete cascade than the previous one.

❖ Comparison

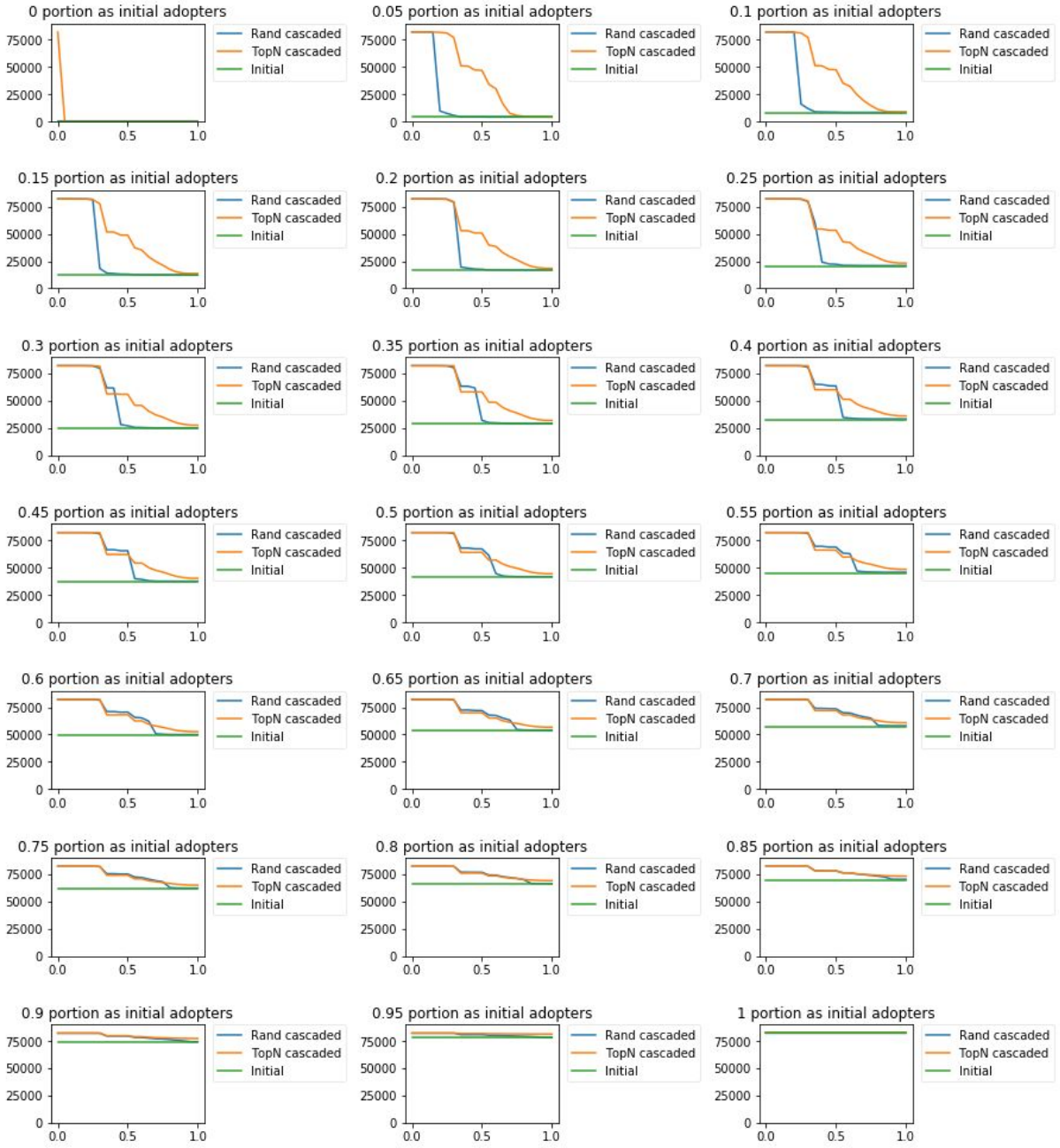

Fig.4. Separate Plot for Case1 and Case2

After taking an overview of the result, we take each of the cases of initial adopters' proportion separating and compare them to the distribution of product adoption and its distribution after cascading for the above two implementation cases (i.e. random selection and top N degree selection). For these plots, the difference between the two implementations as well as the between initial and after cascading is more significant for a smaller proportion of initialization of adopters. Starting for 60% of initial

adoption, the difference between random initialization and top N degree selection becomes negligible.

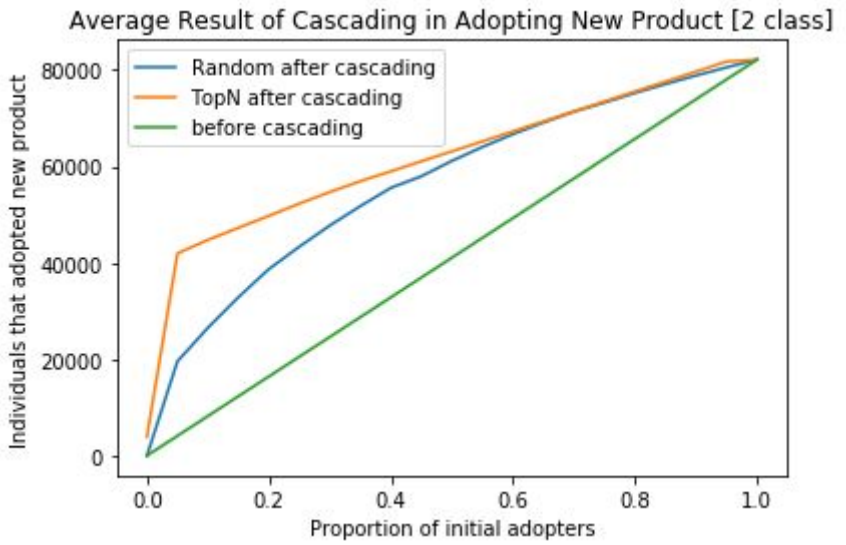

Fig.5. Adopting key nodes vs cascading for Case1 and Case2

For a real marketing campaign by the behavior of information cascading, it is not realistic to have a high portion of initial adopters, therefore it is important to reach a higher virality of cascading for a small proportion of initial adopters to make the marketing campaign economical and effective. For the result shown in Figure 4, it can be observed that the using TopN selection approach can give a significantly higher virality of cascading than using a random selection approach, the virality difference ratio is almost 2:1 when the proportion of initial adopters is 5%. Thus finding people with a high degree of connection can indeed give a significant market cascading effect.

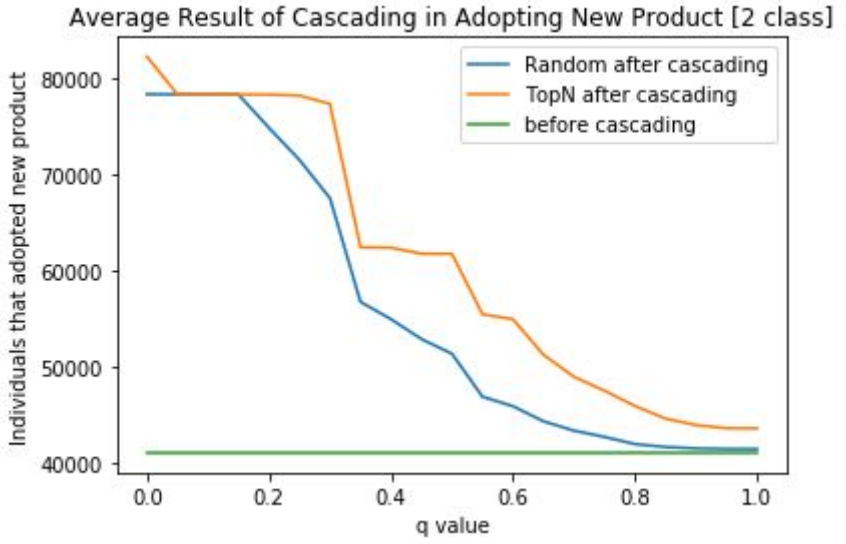

Fig.6. Payoff (threshold) vs cascading for Case1 and Case2

From the relation between the number of nodes in adopting new products after cascading and the q value, it is observed that the random initialization of new product adopters has substantially less influence in changing their neighbour adoption of new products than the TopN approach.

5.3. Multi-Classes Initial Adopters

If there exist two products in the market, according to the previous result, the distribution of the 2 existing classes should be concentrated within clusters. Thus, for 3-class cases(the introduction of the new product is the 3rd one in the market), we use the result of the previous result.

5.4. Clustering

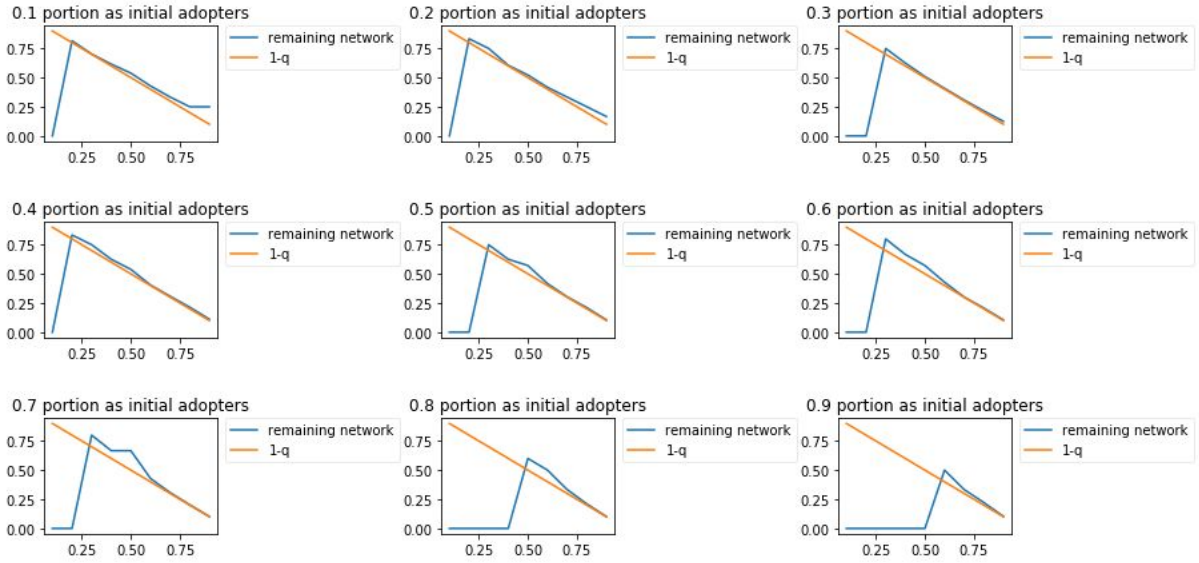

Fig.7. Relation Between Cluster and Cascading

To reduce the run time on finding the cluster density, this task only evaluates 9 values of q ranging from 0.1 to 0.9 for the same 9 proportion of initial adopters. The result is shown as above with every plot representing a specific proportion of initial adopters. From the result, it is observed that the cluster density of the remaining network after cascading is higher than the value of 1-q while the cascading is not complete. On the contrary, the value is lower than 1-q when there is a complete cascading. This observation is aligned with the theory (ii) for the relation between cluster and virality of cascading. Clusters are shown to be an obstacle for cascading.

6. Limitation

Because of the network size, the tools recommended in tutorials like NodeXL and SocNetV cannot be used to visualize and analyze the network features as the graph cannot be loaded into the software.

7. Discussion

Further investigation can be conducted for multiple products existing in the market. For example, the table of payoff for three classes cascading (i.e. new introduction of innovation to a market with existing 2 products) will be as below,

	A	B	C
A	a,a	-c,-c	-c,-c
B	-c,-c	b,b	-c,-c
C	-c,-c	-c,-c	e,e

For a multi-class case, there are more number of inequalities as the condition for the calculation of threshold q.

$$pda - (1-p)dc \geq (1-p)db - pdc$$
$$and$$
$$pda - (1-p)dc \geq (1-p)de - pdc$$

This project is only studying the case for monopoly without consideration other kinds of market structures like oligopoly, monopolistic competition, monopsony and perfect competition. Further study can be conducted on it following the above threshold calculation rule. For the initialization of market product adoption distribution, it can be the cascading result from lower class cascading (i.e. the result for 2 class cascading can be used as the initialization of the market of a 3 class cascading simulation).

8. Conclusion

From the analysis, the lower the q threshold and the higher the initial adopters' population, high the virality of cascading. To enhance the virality of cascading, the TopN degree selection approach as the selection of initial adopters outperform random selection. Moreover, it is also found that the cascading is hindered by clusters through the analysis of cluster density.

9. Reference

[1] M. Kaschesky and R. Riedl, “Tracing Opinion-Formation on Political Issues on the Internet: A Model and Methodology for Qualitative Analysis and Results,” 2011 44th Hawaii International Conference on System Sciences, 2011.

[2] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1) 29--123, 2009.