Yryskeldi Emilbek uulu
Dr. Patrick Shepherd
CSC 486: Network Dynamics
2 May, 2022

Final Project Report

Introduction

Network science is highly applicable within the domain of viral marketing. Under an effective marketing strategy, the business may select a small set of initial users based on trust among close social circles of friends or families so as to maximize the spread of influence in the social network. Under this strategy, the business may choose a pool of users who adopt their product based on specified criteria. These criteria often include the propagation probabilities, centrality metrics, or other node-level characteristics. Influence maximization algorithms (for example, linear threshold or independent cascade models) are the conventional models used to identify the most influential potential seed nodes.

However, researchers have proposed that in addition to this selective product adoption, people naturally share product recommendations with their friends, even without marketing intervention. Thus, we arrive at the idea of “self-activation”, whereas initial seed process will include not only the selected seed nodes (based on a particular algorithm) but also self-activated nodes that are selected randomly apart from the former seed nodes. This project explores the effect that the self-activation paradigm has on the outcome influence sets, and whether a particular type of initial seed process (random, quantitative, and set-based) responds best to the notion of self-activation.

Model/Data

For the input data, this project uses a unimodal, undirected, and unweighted historical Quaker social network dataset from an open-source software repository *Programming Historian.* The dataset encompasses relationships between seventeenth-century Quakers. The data has been excerpted from a dataset employed for the Six Degrees of Francis Bacon project, originally compiled by John Ladd, Jessica Otis, Christopher N. Warren, and Scott Weingart for their Programming Historian tutorial about NetworkX (a Python package for working with network data used in this project). The relationships between seventeenth-century Quakers lend themselves well to social network analysis because, as the authors of the Programming Historian tutorial suggest, "scholars have long linked Quakers’ growth and endurance to the effectiveness of their networks". Within the dataset, the 96 total nodes represent Quakers and the 162 total edges represent computationally inferred relationships between the Quakers based on the Oxford Dictionary of National Biography and confirmed and expanded through the crowdsourced contributions of scholars and students.

Using the Networkx Python package, we can visualize the Quaker social network in many different ways using circular (Figure 1), spiral (Figure 2), or the Kamada-Kawai (Figure 3) layouts.

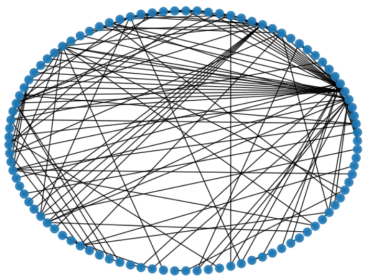

Figure 1. Circular visualization of the Quaker social network using Networkx

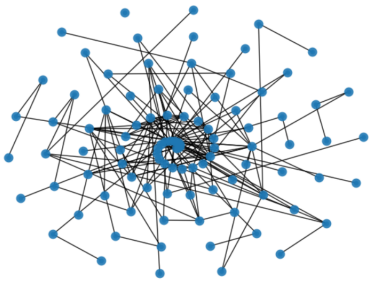

Figure 2. Spiral visualization of the Quaker social network using Networkx

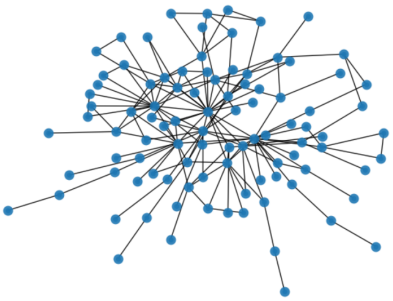

Figure 3. Kamada-Kawai visualization of the Quaker social network using Networkx

These visualizations show that the Quaker social network resembles a structure of a scale-free graph characterized by the presence of large hubs highly interlinked with other nodes.

In order to test the notion of self-activation within the influence maximization paradigm and its effect on the total influence set, we first conduct simulations and explore how the type of algorithm and the number of initial seed nodes all influence the total activation set. We use the conventional Independent Cascade model as well as the self-activation model to run our simulations. Activation set refers to the total number of nodes activated directly or indirectly by the seed nodes selected by one of the three algorithms: random, quantitative, and set-based. Random algorithm (Figure 4) selects the seed nodes based on randomly choosing nodes within a graph without considering node-level metrics. Quantitative algorithm (Figure 5) selects the seed nodes based on the sizes of their activation sets - the higher the potential activation set of a particular node, the higher is the chance that this node will be included in the seed node set. The set-based algorithm (Figure 6) selects seed nodes based on how many additional unique nodes each particular node adds to the total activation set. We use the range of even numbers between 2 and 20 for choosing the size of initial seed node set. We also use the 25% propagation probability for edges.

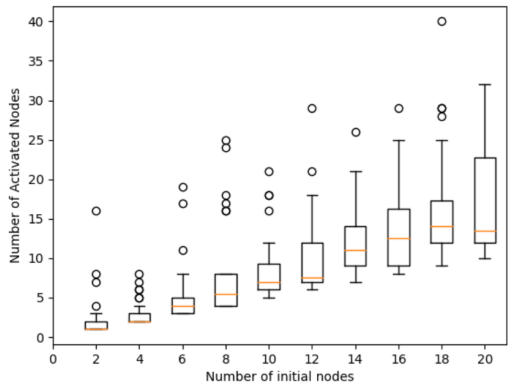

Figure 4. Random algorithm combined with self-activation and its effect on the activation set

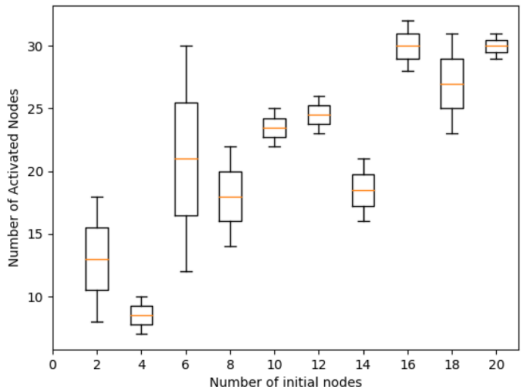

Figure 5. Quantitative algorithm combined with self-activation and its effect on the activation set

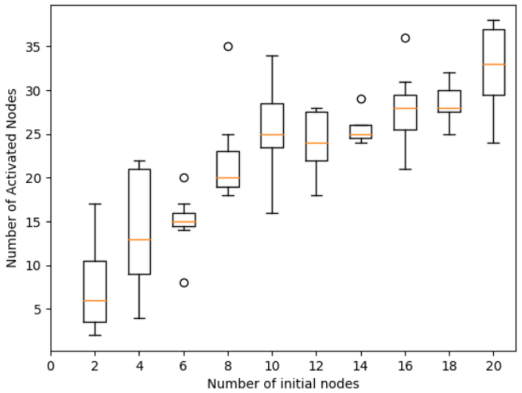

Figure 6. Set-based algorithm combined with self-activation and its effect on the activation set

Based on the above visualizations, it seems that the set-based algorithm performs best with self-activation when the number of initial nodes is higher, while the quantitative algorithm performs best with self-activation when the number of initial nodes is lower.

Results

We further explore the difference between influence sets of the three algorithms by running the Single-Factor Analysis of Variance (ANOVA) statistical test to determine whether there are any statistically significant differences between the means of the three algorithms and their activation sets (keeping the self-activation modification within the three algorithms). We run the ANOVA test because we have the three categorical groups in our independent variable (representing each algorithm) and the continuous dependent variable (each algorithm's activation set). We validate the assumptions that 1) our observations are independent (no relationship between the observations in each algorithm's performance or between the algorithms themselves); 2) there are no significant outliers in our observations; 3) our dependent variable is approximately normally distributed for each category of the independent variable; and 4) there is a homogeneity of variances.

Within the following statistical tests, we test the null hypothesis that the means of the activation sets of the three algorithms are the same. If they are statistically different, the results will indicate strong evidence to reject the null hypothesis. In the dataset, we activate 12 seed nodes, 10 of which are selected by one of the three algorithms and 2 of which are randomly self-activated (Figure 7). We run 40 total simulations to obtain the sample.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Random selection plus self activation	40	771	19.275	33.4865385
Algorithm A plus self activation	40	1276	31.9	14.1948718
Algorithm B plus self activation	40	1287	32.175	19.2762821

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4345.01667	2	2172.50833	97.3379574	1.2801E-25	3.0737629
Within Groups	2611.35	117	22.3192308			
Total	6956.36667	119				

Figure 7. One-way ANOVA test for the differences between means of the three algorithms' activation sets

The p-value corresponding to the F-statistic of the one-way ANOVA test is lower than 0.05, suggesting that the one or more algorithms' activation sets are significantly different (however, the ANOVA test doesn't yet tell us which pair of the algorithms are significantly different). We follow up this finding with the post-hoc Tukey HSD test (Figure 8). This post-hoc tests would likely identify which of the pairs of algorithms' activation sets are significantly differerent from each other. The Tukey HSD test will help us identify which specific groups's means (compared with each other) are different. The test compares all possible pairs of means.

Tukey HSD results

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inferfence
A vs B	16.9014	0.0010053	** p<0.01
A vs C	17.2695	0.0010053	** p<0.01
B vs C	0.3681	0.8999947	insignificant

Figure 8. Post-hoc Tukey HSD test to identify which algorithms' activation sets are significantly different (A stands for the random algorithm, B stands for the quantitative algorithm, and C stands for the set-based algorithm)

The Tukey HSD test results show that not all algorithms' activation sets are statistically different: the only algorithm that brought statistically different results from the rest is the random algorithm, which indicates strong evidence to reject the null hypothesis that the means of random algorithm's influence set and other algorithms' influence set are different. The quantitative and set-based algorithms are not statistically different, which means that we failed to reject the hypothesis that the the means of their activation sets are statistically different from each other.

This project then proceeds to conducting a test to identify whether the means of the activation sets of the two non-random algorithms (quantitative and set-based) before and after introducing the self-activation model are statistically different. To test the null hypothesis that the means before and after introducing the self-activation model are the same, we employ the Paired Two-Sample for Means T-test (Figure 9 and 10). We emphasize the two-tail p-value metric to test our hypothesis since we're not yet concerned about the direction of the difference. In both influence sets, the graph had 12 total unique initial seed nodes both before and after introducing self-activation; in the group with no self-activation, all initial seed nodes were selected with the specified algorithm; in the group with self-activation, 10 seed nodes were selected using the specified algorithm, and 2 nodes were self-activated. We run 40 total simulations to obtain the sample.

T-test: algorithm a no self activation vs algorithm a with self-activation

t-Test: Paired Two Sample for Means

	Algorithm A - no self activation	*Algorithm A plus self activation*
Mean	31.1	30.95
Variance	17.06666667	24.2025641
Observations	40	40
Pearson Correlation	0.020438328	
Hypothesized Mean Difference	0	
df	39	
t Stat	0.149184582	
P(T<=t) one-tail	0.441088387	
t Critical one-tail	1.684875122	
P(T<=t) two-tail	0.882176773	
t Critical two-tail	2.02269092	

Figure 9. Two-Sample for Means T-test for differences between quantitative algorithm's performances before and after introducing self-activation

T-test: algorithm b no self activation vs algorithm b with self-activation

t-Test: Paired Two Sample for Means

	Algorithm B - no self activation	*Algorithm B plus self activation*
Mean	33.525	31.05
Variance	20.51217949	17.84358974
Observations	40	40
Pearson Correlation	-0.26275756	
Hypothesized Mean Difference	0	
df	39	
t Stat	2.249776535	
P(T<=t) one-tail	0.015090405	
t Critical one-tail	1.684875122	
P(T<=t) two-tail	0.03018081	
t Critical two-tail	2.02269092	

Figure 10. Two-Sample for Means T-test for differences between set-based algorithm's performances before and after introducing self-activation

T-test results show that there is no statistically significant difference between the means of activation sets of the quantitative algorithm before and after introducing self-activation; the activation sets of the set-based algorithm before and after self-activation, however, are statistically different. Thus, we obtained strong evidence to reject the null hypothesis that the means of the set-based algorithm's activation sets before and after introducing self-activation are the same. The set-based algorithm seems to respond negatively to the self-activation model, while the quantitative algorithm doesn't hold statistically significant influence from self-activation.

Conclusion

Influence maximization models hold significant business value because they can reduce marketing intervention costs. Traditional influence maximization models, however, fail to acknowledge the real-world property of social network: some seed nodes may self-activate and affect the activation set. This project explored this idea further and compared the differences in output between the three node-finding algorithms - random, quantitative, and set-based - with the self-activation engrained into the Independent Cascade model. We found that the algorithms that consider factors like average influence set size (quantitative) and unique influence sets for each node (set-based) produce statistically different results as compared to the randomness-based algorithm. We also then explored the idea of self-activation further and conducted statistical tests to determine if self-activation brings statistically significant differences in activation sets. We found that the introduction of the self-activation model produces statistically significant differences in activation sets for the set-based node-finding algorithm, but not for the quantitative algorithm. These findings suggest that if the business influence maximization model employs the quantitative algorithm to find influential nodes, the total activation set won't undergo statistically significant decrease in the activation set size if some of the nodes are self-activated. In contrast, if the business influence maximization employs the set-based algorithm, costs are much more significant and more attention must be brought to proactive marketing intervention.

Resources:

https://github.com/melaniewalsh/sample-social-network-datasets/tree/master/sample-datasets/quakers - Dataset

http://oak.ucc.nau.edu/rh232/courses/EPS525/Handouts/Understanding%20the%20One-way%20ANOVA.pdf - One-Way ANOVA

https://researchbasics.education.uconn.edu/t-test/ Two-Sample for Means T-Test

https://arxiv.org/abs/1906.02296 - Self-Activation within Influence Maximization Models (Spontaneous User Adoption)