SOCIAL NETWORK REPORT

ON KARATE DATASET USING INDEPENDENT CASCADE MODEL

MARCH, 2022

PREPARED BY

ALIPOURMOTLAGH MOHAMMADHADI

ARFAOUI WAJIH

GOPALAKRISHNAN HARIKRISHNAN

NAKIB AHMADOMAR

OMUVWIE ENITA

RANA INDERPREET

Contents

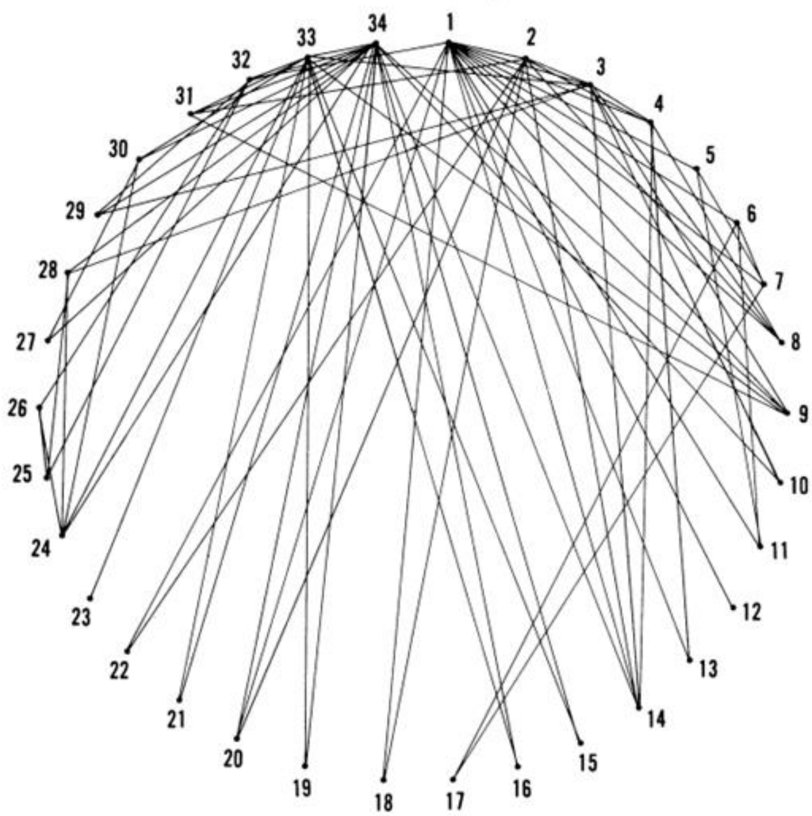
INTRODUCTION

We live in an increasingly digital age, where recent technologies and social media are providing us with new channels of communication and great convenience to share information and express opinions. With the proliferation of social media, its influence on business, politics, and society become evident and significant as it spreads massive volumes of information over its different platforms. The study of how innovations, behaviors, and diseases spread through social networks has a long history in the social sciences. Many models of information and influence diffusion have been proposed for a wide variety of applications, such as viral marketing, cascading behavior & prediction and information spreading. In this project, we chose an **independent cascade diffusion model** to maximize the spread of the influence of Zachary karate club network. We will also walk through the methodology & implementation of the model, our model parameters and statuses, simulation execution and finally the outcomes of the model.

DATASET DESCRIPTION

The Dataset that we chose is about the social network between members of a university karate club, led by President John A. and karate instructor Mr. Hi (pseudonyms).

Social Network Model of Relationships in the Karate Club



This above figure illustrates the graphic representation of the social relationships among the individuals in the karate club. Each time a line is drawn between two points, it represents an interaction outside of karate classes, workouts, and club meetings. Such a line is referred to as an edge.

The edge weights are the number of common activities the club members took part of. Zachary studied conflict and fission in this network, as the karate club was split into two separate clubs, after long disputes between two factions of the club, one led by John A., the other by Mr. Hi.

The data format is an undirected igraph graph object. Vertex no. 1 is Mr. Hi, vertex no. 34 corresponds to John A.

Graph attributes: 'name', 'Citation', 'Author'.

Vertex attributes: 'name', 'Faction', 'color' is the same as 'Faction', 'label' are short labels for plotting.

Edge attribute: 'weight'.

The 'Faction' vertex attribute gives the faction memberships of the actors. After the split of the club, club members chose their new clubs based on their factions, except actor no. 9, who was in John A.'s faction but chose Mr. Hi's club.

PROPOSED APPROACH - INDEPENDENT CASCADE MODEL

Independent cascade model is the type of epidemic model which assumes that a node adopts an innovation when one of its neighbors adopts it.

In the Independent Cascade model, every edge $(i, j) \in E$ is associated with a propagation probability

$$\boldsymbol{p}: (\boldsymbol{V} \times \boldsymbol{V}) \rightarrow (\boldsymbol{0}, \boldsymbol{1}]$$

Where $\boldsymbol{p_{i,j}}$ represents the probability that node $\boldsymbol{j}$ is influenced by node $\boldsymbol{i}$ through the edge (i, j) at step $\boldsymbol{t}$ when node i is activated at step $\boldsymbol{t - 1}$.

Thus, we denote an Independent Cascade model by a triple $\boldsymbol{GIC} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{p})$

The Independent Cascade model for influence maximization (IM) problem is characterized by a triple (G, K, w),

Where:

$\boldsymbol{G}$ is a given directed graph

$\boldsymbol{K} \leq \boldsymbol{L}$ is the cardinality of source nodes

$\boldsymbol{w}: \boldsymbol{E} \rightarrow [\boldsymbol{0}, \boldsymbol{1}]$ is a probability weight function mapping each edge $\boldsymbol{e} \in \boldsymbol{E}$ to a real number $\boldsymbol{w}(\boldsymbol{e}) \in [\boldsymbol{0}, \boldsymbol{1}]$.

The model needs to choose a set of $\boldsymbol{K}$ source nodes $\boldsymbol{S} \subseteq \boldsymbol{V}$ based on (G, K, w). Then a random binary weight function w, which encodes the diffusion process under the IC model, is obtained by independently sampling a Bernoulli random variable $\boldsymbol{w}(\boldsymbol{e}) \sim \boldsymbol{Bern}\,(\boldsymbol{w}(\boldsymbol{e}))$ for each edge $\boldsymbol{e} \in \boldsymbol{E}$.

The model's objective is to maximize the expected number of the influenced nodes:

$$\max(S) \colon |S| = Kf(S, w)$$

$f(S, w) = E(w[f(S, w)])$ is the expected number of influenced nodes when the source node set is $\boldsymbol{S}$ and w is sampled according to $\boldsymbol{w}$.

METHODOLOGY

One of the most heavily researched modelling techniques for influence propagation is the independent cascade model. It is a probabilistic model that was proposed by Jacob Goldenberg (Jacob Goldenberg) and others when they studied marketing models.

The model's fundamental assumption is that the success of a node u's attempt to activate its neighbour v is a random event with probability $\boldsymbol{p}(\boldsymbol{u}, \boldsymbol{v})$. And the likelihood that a node in an inactive state will be activated by a neighbour node that has recently entered the active state is independent of the activity of neighbours who have previously attempted to activate the node.

Furthermore, the model assumes that any node u in the network has only one chance to try to activate its neighbour node $\boldsymbol{v}$, regardless of if it succeeds or not, in the future, even if u itself remains active. If it regains influence, it is referred to as a node in the network without influence.

This is the model algorithm or methodology:

- At time $\boldsymbol{t}$, the newly activated node u exerts an influence on its adjacent node $\boldsymbol{v}$ with a success probability of $\boldsymbol{p}(\boldsymbol{u}, \boldsymbol{v})$. If node v has multiple neighbors who are all newly activated nodes, these nodes will attempt to activate node $\boldsymbol{v}$ in any order.
- If node $\boldsymbol{v}$ is successfully activated, node v becomes active at $\boldsymbol{t} + \mathbf{1}$, affecting its adjacent inactive nodes; otherwise, node v's state does not change at $\boldsymbol{t} + \mathbf{1}$.
- This process is repeated indefinitely until there are no influential active nodes left in the network, at which point the propagation process terminates.

The following is the step for independent cascade modelling:

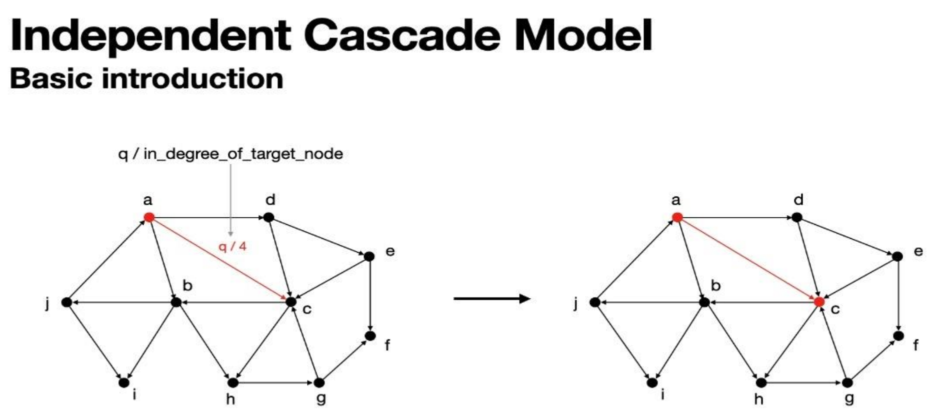
- The initial set of actived nodes which is setting the seed.
- An edge will be successfully cascaded with the probability 1/degree(karate), where 1 is a hyper parameter called threshold (which we always set to 0.2, 0.4, 0.6, 0.8, or 1.0).
- The propagation will be cascaded only once in each round, and the active nodes in the round will be set to actived node, which will no longer cascade to other nodes, even if the nodes did not cascade to every other node.
- Continue cascading until no nodes can activate others. The number of actived nodes is the final inclusion number at this period.

There are three node statuses:

- Inactive - these are common nodes.
- Active - seeds or nodes that were actived in the previous round; these nodes can now activate other nodes in the next round.
- Actived - nodes that have previously actived other nodes; these nodes are actived but cannot active other nodes.

Example for description explanation

The in-degree of node c in this graph is 4, so the propagation of the edge a->c is q/4. In the next round, c has a 50% chance of being actived by a.



This graph's independent cascade model dissemination process consists of the following.

Time step 0: A node is turned on.

Step 1: Node an attempt to activate node b with a probability of 0.5 and node c with a probability of 0.2. Assume that node b is activated effectively within this time step.

Time step 2: Node b attempts to activate node c with a probability of 0.3 and node d with a probability of 0.5. Assume that c- and d-nodes are successfully activated during this time step.

Time step 3: Node c attempts to activate e with a probability of 0.2, while node d attempts to activate e with a probability of 0.2. If all attempts during this time step have failed, no new nodes are activated, and propagation comes to a halt.

EXPLANATION OF GRAPHICAL APPROACH OF INDEPENDENT CASCADE MODEL

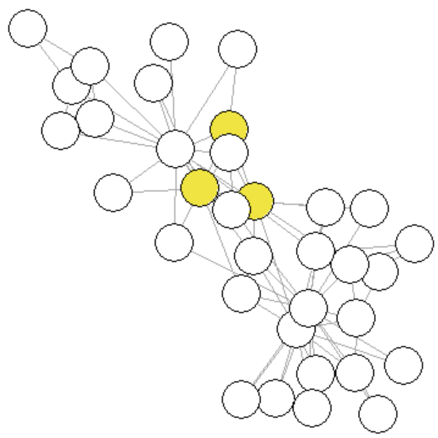


Figure 1: This graph illustrates the active nodes at the initial time

Our active nodes are shown in yellow, at the initial time, three nodes were activated. At the next time, these nodes have a chance of activating their neighbours. The figure below shows only three node was successfully activated. This step gets repeated until there are no more active nodes and the diffusion process stoped.

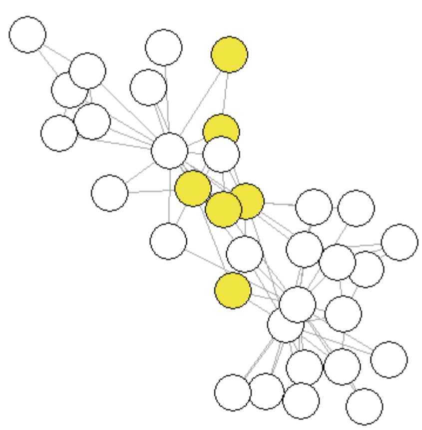
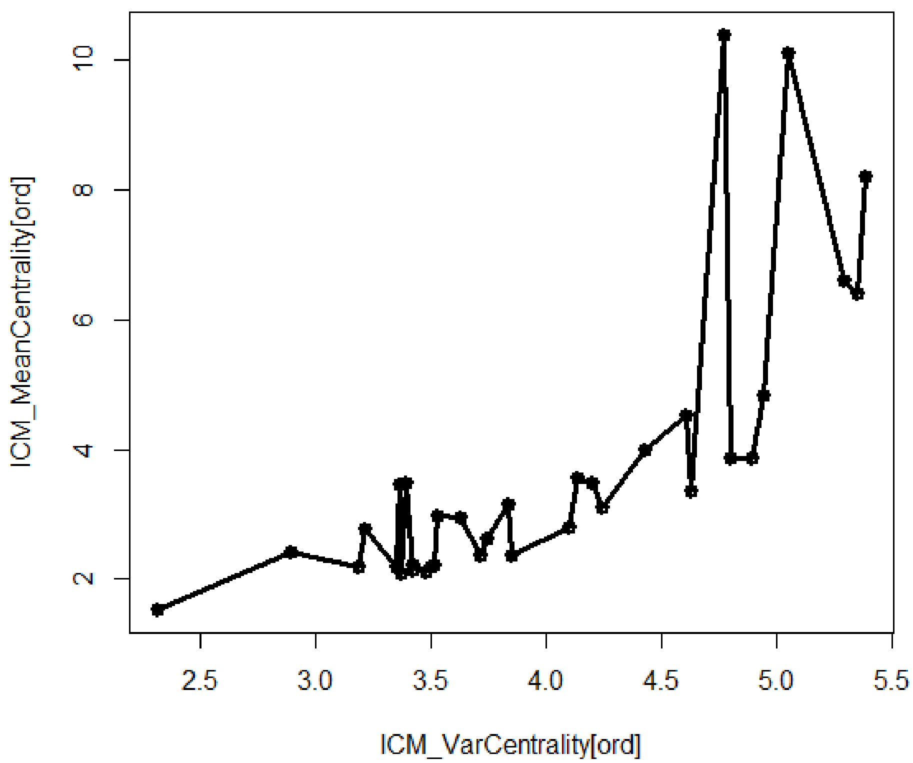


Figure 2: This shows the diffusion graph

BEST NODES

Based on the above explanations we ran the simulation with 1000 iterations and in each iteration we calculated the initial results of each node based on the average influence a node has we also calculated the influence and impact on the standard deviation. After running this loop we plot the result for each node based on the Average and Variance as follows:



We can see that the nodes number 27 and 31 will influence nearly 10 other nodes and since the variance of node number 27 is lower than 31 then we can conclude that node 27 is the optimal choice in this situation.

CONCLUSION

In this project, we have reviewed the data set using the Independent Cascade model and applied its algorithmic approach for achieving scalability. The sudden impact of social networks has provided a rich information base for trying to conduct relevant studies, allowing scientists to investigate information distribution processes and know the process of disseminating information using massive actual data, with incremental results. The distribution of information in social networks is mainly concerned with the fundamental network structure, the network's groups and the information distributed. The network structure-based research achievements primarily include the independent cascade model, linear threshold model, and extended model among others. The results of this project based on information characteristics primarily include multi-source information distribution model and data competition distribution model.

As we can see, the Cascade model helps us identify the most effective members of the network and select the optimal option according to the standard deviation and other factors such as cost and time. In the example above, we selected a small network and identified the effective members of this network. To do this, we used the two factors of mean impact and standard deviation and introduced the optimal options.

Finally, it can be said that to better identify the influential member, we should add as many different attributes as possible as long as they are involved in decision making in the network and expand the analysis based on them.

REFERENCES

GitHub - cbhua/model-independent-cascade: The implement of Independent Cascade Model, with different seeds selection algorithms. https://github.com/cbhua/model-independent-cascade

https://rdrr.io/cran/igraphdata/man/karate.html

http://www1.ind.ku.dk/complexLearning/zachary1977.pdf

https://appliednetsci.springeropen.com/articles/10.1007/s41109-018-0062-7

https://www.alessandro-giua.it/UNICA/PAPERS/JOUR/19access.pdf

https://misovalko.github.io/publications/wen2017online.pdf

独立级联模型. (2022). https://zhuanlan.zhihu.com/p/164608949