Influence Maximization in Networks

Raj Patel (201901306),* Vishv Joshi (201901453),† and Ashray Kothari (201901457)‡
Dhirubhai Ambani Institute of Information & Communication Technology,
Gandhinagar, Gujarat 382007, India
SC-435, Introduction to Complex Networks

Analysis of a network is extremely important to know how a network is constituted and in what manner influence spreads in the network. Social Networks hold importance in the eyes of marketing and advertisement agencies in order to reach out to the maximum number of people using the least possible budget. We have analyzed this with the help of *Game of Thrones* Network as the story has a lot of people interacting in various communities, both within and between. The spread of information in such a network may provide insights into how information spreads in a network.

I. INTRODUCTION

Think about the following imaginary event as an inspiring example. A tiny business creates a fascinating application for an online social network with the intention of marketing it on the same network. Because of its limited resources, it can only choose a few initial network users to use it. The company hopes that these early adopters will adore the app and begin persuading their friends to use it, and then their friends will persuade their friends' friends, and so on, leading to the application being used by a sizable portion of the social network. Finding influential people in a social network is a challenge since it involves choosing the first members who will eventually impact the greatest number of others in the network. Many businesses and people interested in promoting their goods, services, and original ideas would be interested in this topic, also known as influence maximisation.

To analyse this, we look at a network built around the Volumes of "*A Song of Ice and Fire*", which were turned into the hugely successful web series "*Game of Thrones*". The paper correlates various centrality measures and spreading behavior to identify the most influential characters in the network.

II. DATASET AND NETWORK PREPARATION

The Game of Thrones Network dataset is generated by taking all the volumes of the "Song of Ice and Fire" saga. The characters who appear within 15 words of one another are considered to be interacting with one another and are hence connected. The number of interactions corresponds to the edge weight in the graph. The graph has 796 nodes and 2823 edges.

For our report, we shall focus on a sub-network consisting of only the key characters. Only those nodes are considered in the sub-network whose weight are greater than 100. Now our graph has been reduced to 124 nodes and 985 edges. Since the network just depicts how the characters interact, it generates an undirected graph.

Source	Target	Weight
Eddard-Stark	Robert-Baratheon	334
Jon-Snow	Samwell-Tarly	228
Joffrey-Baratheon	Sansa-Stark	222
...	...	...

TABLE I: Data representation of the Network

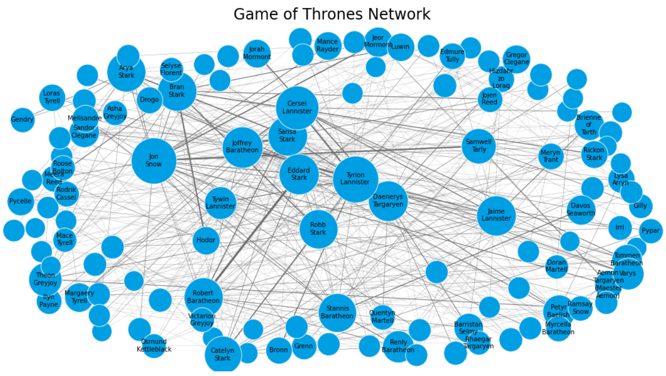

FIG. 1: Network Visualization

# Nodes	124
# Edges	985
Avg. Path Length	2.237
Avg. Clustering Coeff.	0.634
Diameter	4
Connected Components	1
Density	0.129
Transitivity	0.43

TABLE II: Network Statistics

*Electronic address: 201901306@daiict.ac.in
†Electronic address: 201901453@daiict.ac.in
‡Electronic address: 201901457@daiict.ac.in

III. NETWORK ANALYSIS

A. Power-Law

To get insights into the structure of the network and thereby compare it with other well-known networks, degree distribution is a simple, yet powerful, metric. The plots in Fig (2) and Fig (3) shows the degree-distribution. We can see from Fig (3) that the network follows power law distribution with α value equal to 2.708.

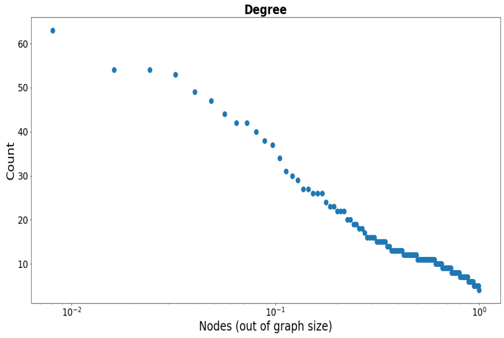

FIG. 2: Degree Distribution

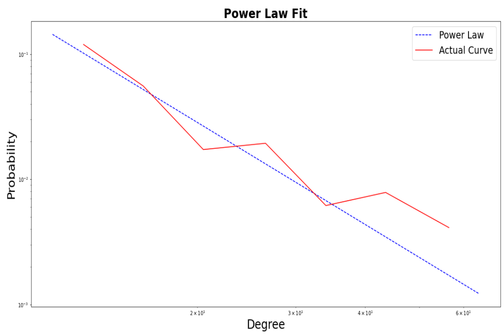

FIG. 3: Power-Law Fit

B. Centrality Measures

In network analysis, centrality is an measure of a node's relative significance and hence to rank them, based on some criteria, inside a graph. Since finding the most prominent users in a social network is ultimately about measuring relevance and social networks are intrinsically tied to graph theory and network analysis, centrality metrics have attracted the greatest focus.

Degree Centrality helps us understand how well nodes, here people, are connected in the graph. It's the most basic centrality measure. With this one may observe that most of the people do not have much role in the network. Similar to how the novels depict him as a member of the richest and maybe most powerful family in the show, Tyrion Lannister is the most connected individual in the graph. This can be accurately inferred from the plots of Fig. (4) and Fig. (5). As degree centrality looks at number of connections, it can't indicate accurate influence of the nodes.

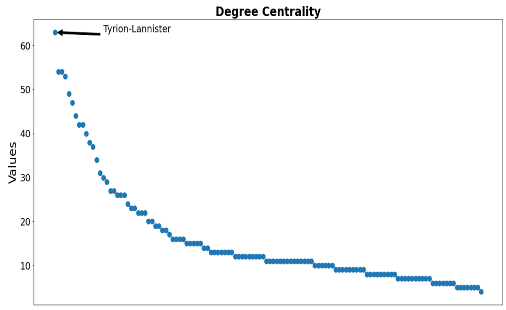

FIG. 4: Degree Centrality

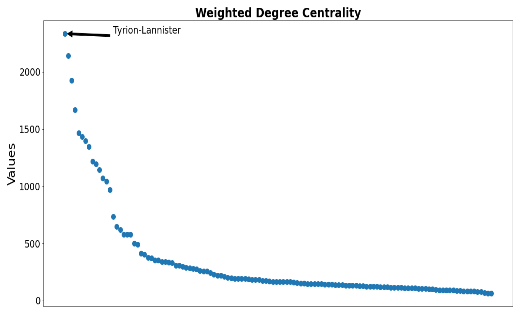

FIG. 5: Weighted Degree Centrality

Eigenvector centrality is based on the principle that a node having connections to other nodes with a high score in turn gets a contribution in raising its own score. And it makes sense as having 100 popular connections is far more impactful than having 100 unpopular connections. Thus, Tyrion Lannister has the highest centrality here as shown in Fig. (6).

PageRank Centrality aids in assessing a node's significance in a graph. Tyrion Lannister is the most interconnected figure, yet he is not the most powerful on his own. Because of his own special traits and abilities, the plot in Fig. (7) demonstrates that Jon Snow is the most significant character.

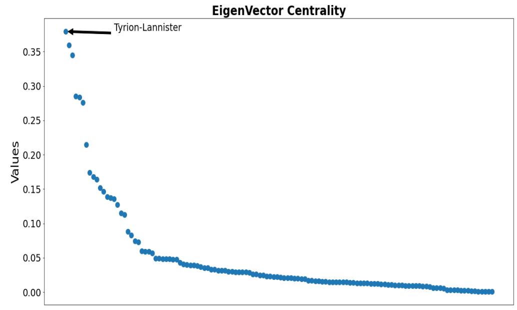

FIG. 6: Eigenvector Centrality

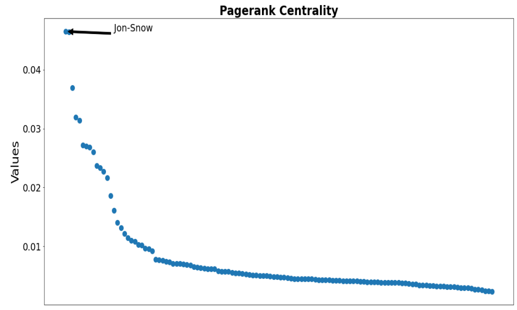

FIG. 7: PageRank Centrality

It's quite often seen that an influential node having few connections propagates information efficiently. So to quantify how a node, as an intermediary, controls the interaction of other nodes in a social network, Betweenness Centrality is a crucial measure. In the plot of Fig (8), the highest value belongs to Robert Baratheon. This makes sense because his death kickstarts the whole story and is practically the reason for nearly all aspects of the story.

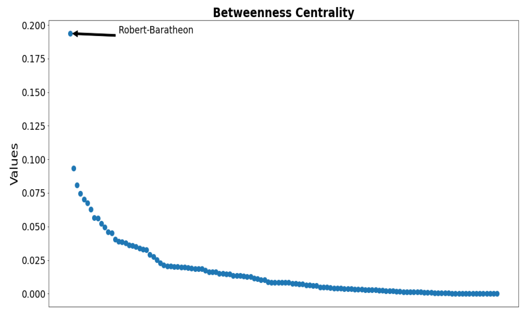

FIG. 8: Betweenness Centrality

C. Community Detection

We take a look at communities formed by the Louvain algorithm and Clauset-Newmann-Moore(CNM) algorithm. The CNM algorithm (10) gives out 4 communities which can be classified as:

- The Starks and the Baratheons who are against the Lannisters
- Arya Stark's associates, Lannisters, and the Tyrells
- Daenerys Targaryen and her associates
- The Night's Watch

The Louvain algorithm (9) gives 6 communities:

- The Lannisters and their associates
- Bran Stark's and Robb Stark's associates
- The Night's Watch
- Stannis Baratheon and Renly Baratheon and their associates
- Daenerys Targaryen and her associates
- Arya Stark and her associates

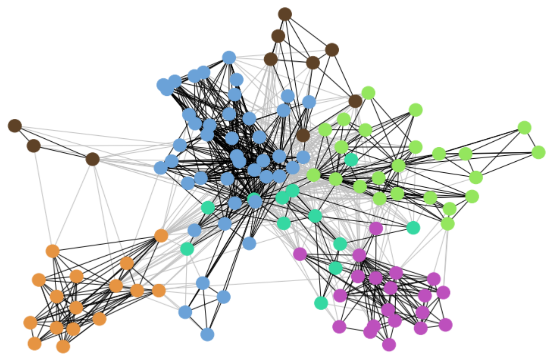

FIG. 9: Louvain Community

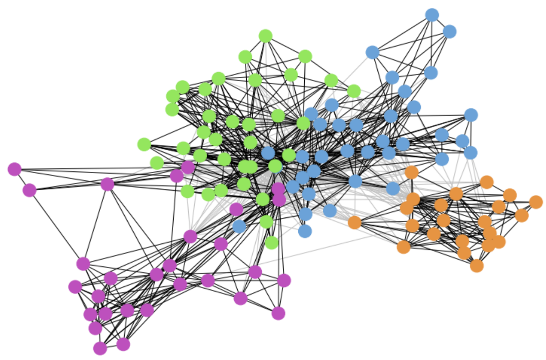

FIG. 10: CNM Community

In Fig (9) and Fig (10), the lighter edges represent inter-community edges, and darker edges represent intra-community edges.
The modularity for the CNM algorithm is 0.39 and the Louvain algorithm is 0.477. This leads us to the conclusion that the Lovain method is more effective at identifying communities in our network. Even real circumstances that follow the GOT narrative reflect this as communities detected are quite accurate.

IV. INFLUENCE MAXIMIZATION

The influence maximization issue outlines a marketing scenario, where the marketer's objective is to choose a small number of network nodes (the seeding set) in a way that will allow the impact to expand spontaneously to as many nodes as feasible. For instance, think about inviting a select group of influencers to a prominent product launch event so they can tell their network about it. Such a study is **NP-Hard** Problem.

A. Information Flow

We look at two models for information flow. This can also resemble the spread of infection or diseases in the manner that information passes from one person to another upon contact or in other words, adjacent nodes in a graph. The spread of an idea can also be modeled from these models.

1. **Linear Threshold Model (LTM)**
   This model follows the accumulative nature of influence from peers. There is a certain threshold that determines whether a node is *active* or not. If the threshold is crossed, the node becomes *active* and it never deactivates. A node i is said to be active if

   $$\sum_{\text{active j}} w_{ji} \geq \theta$$

   where w_{ji} represents weights from node j to node i and θ is the threshold.

   Let's say that at the beginning of the process, Samwell Tarly and Bran Stark, two of Jon Snow's closest friends, are aware of an extremely significant piece of information about his life. One can see how this information flows using LTM through this animation.

2. **Independent Cascading Model (ICM)**
   This model does not take into account the action of other nodes while considering a particular node. Every active node has an independent chance of activating the neighboring nodes. Contrary to the linear approach of the threshold model, this is a probabilistic approach.
   This animation demonstrates how ICM is used in the prior information flow scenario.

Model parameters have a large role in these dynamics, which can cause the spreading process to take diverse paths.

B. Budgeting

We execute a brute-force simulation as the first investigation into budgeting and choosing the seeding set for Influence Maximization, taking certain top values as seeds for different centralities. For example, for a seeding set of size 5, we take the top 5 respective values for each centrality. The seeding set size increases by one after each iteration until the whole network is covered.

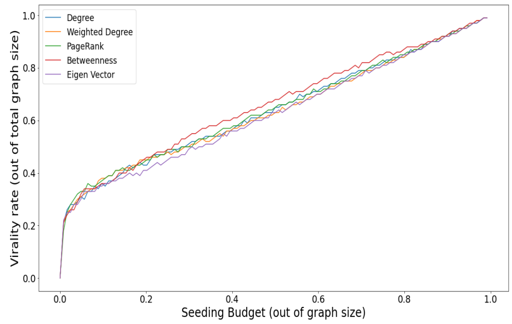

FIG. 11: Virality Plot for Linear Threshold Model

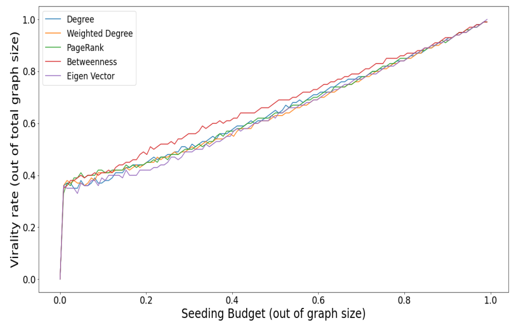

FIG. 12: Virality Plot for Independent Cascading Model

We observe that even after taking a small seeding set, we can get nearly 20% coverage using LTM and 40% coverage using ICM. Also, of all centralities, the Betweenness Centrality works better for most of the seeding set sizes in both models.

Then, using a brute-force search strategy, we identify the ideal seeding "couple" (budget = 2) by considering all feasible pairs, regardless of their centrality ranks.

- Using Independent Cascade Model,
  The best-performing nodes here are John Snow and Arya Stark. They cover a little over 41% of the network. This marks an interesting observation that if two nodes of non-intersecting domains are taken then influence spread is the maximum. This can be confirmed by the fact that the storylines of Jon Snow and Arya Stark largely operate independently of each other and hence they can spread influence to a wider reach. Another Observation is that it is not necessary that maximum influence will spread if and only if we take the top 2 important people in the network as Arya Stark is not even in the top 10 pagerank values.
- Using Linear Threshold Model,
  The maximum area coverage is done by Renly Baratheon and Catelyn Stark. They can cover over 40% of the network. Again, we see that neither Renly Baratheon nor Catelyn Stark have the top pagerank values. But they still manage to do well because they belong to disjoint groups in the story. Another pair that comes close is the Renly Baratheon and Jon Snow pair which also covers 40% of the network. It is again observed that both characters fall under different storylines.

The results of the two studies mentioned above lead us to the conclusion that top characters do not necessarily disseminate information to the greatest number of network nodes in our network. Even if some characters are less significant since they come from other groups, they could be able to disseminate knowledge more widely.

V. CONCLUSION

Network analysis is a sophisticated and helpful tool for a wide range of fields, particularly in the domain of social networks. Marketing influence maximization, fraud detection, and recommender systems are examples where network analysis is used. For analyzing network datasets, there are numerous tools and approaches but they must be chosen carefully, considering particular traits of the problem and the network.

[1] Goldenberg, Dmitri. “Social network analysis: From graph theory to applications with python.” PyCon 2019 — 3rd Israeli National Python Conference, Israel, 2019. arXiv preprint arXiv:2102.10014 (2021).

[2] Chen, Wei, Yajun Wang, and Siyu Yang. ”Efficient influence maximization in social networks.” Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009.

[3] Sela, Alon, et al. ”Active viral marketing: Incorporating continuous active seeding efforts into the diffusion model.” Expert Systems with Applications 107 (2018): 45-60.

[4] Shakarian, Paulo, et al. ”The independent cascade and linear threshold models.” Diffusion in Social Networks. Springer, Cham, 2015. 35-48.

[5] [Dataset]