Degree Assortativity in Influence Maximization of Partially Observable Networks

STAT 175: Statistics and Data Science of Networks

Derek Chang, Ben Elliott, Lyla Kiratiwudhikul, Caleb Saul

May 3, 2023

1 Introduction

Understanding how information disseminates in today's world has never been more important: individuals, companies, organizations, and governments face limited resources and increasing competition for audience attention in spreading their messages and encouraging action. The influence maximization problem has a wide range of applications: spanning product marketing strategy, humanitarian crowdfunding, participatory data collection, and spreading awareness about health resources. Decision-makers must not only consider how their messages will be disseminated within their direct sphere of influence—existing customers, social media followers, or current patients — but also how it will spread beyond what is known—friends of customers, followers of followers, or coworkers of patients.

At the heart of influence maximization are networks. Modeling how individuals are connected with and influence one another gives insights about whom to target with initial messaging or intervention. One popular approach is the independent cascade model: a set of nodes are initially activated, and then each activated node activates its neighbors independently with some probability p. The challenge of selecting the seed set to disseminate a message as widely as possible is known as seed selection for influence maximization.

The majority of this literature analyzes fully known networks, but in many real-world applications, knowledge of a network is limited: only pieces of a network are observable. Studying the independent cascade model over undirected graphs, Eckles et al. (2022) develops algorithms for selecting a seed set with the network bounded by how much is observed. The paper also analyzes two approaches to acquire network information: influence sampling—observing the influence spread of random nodes—and edge queries, which reveal the identity of neighboring nodes in an adjacency matrix. Stein et al. (2017) proposes four heuristic algorithms for selecting an initial seed set: random selection, random selection with neighbor activation, degree-based selection, and state-of-the-art with limited visibility. The authors find that heuristic algorithms which target nodes at the boundary between the observed and unobserved portions of a graph outperform the state-of-the-art by up to 38%. However, they note that the relatively good performance of these degree heuristics may depend on certain characteristics of the network such as degree assortativity.

In this work, we examine the effects of degree assortativity on the average spread using degree-based heuristics.[1] To do so, we alter connections in the NetHept to generate new networks with various assortativities, and follow the experimental setup as in Stein et al. (2017). We also repeat the experiment on other datasets: Gnutella, Stanford, Amazon, Wikipedia, and Email EU. The results show that applying degree-based heuristics to lower assortativity graphs results in higher spread. For networks with positive assortativity, applying degree-based heuristics for seed selection across all visibilities and datasets results in activation of less than 1% of overall graph; on the other hand, for networks with negative assortativity, we see activation of greater than 1% and up to 20%. These results suggest that understanding network characteristics is essential in applying a seed selection algorithm to obtain optimal performance.

2 Methodology

The influence maximization challenge is to effectively disseminate a message as far as possible within a network, given that it is only partially observable and there is a limited budget. To do so, we must select the most influential members within a network to distribute the message far beyond the boundaries of the known network.

The problem setup is as follows: consider a directed graph $G(V, E, w)$ where we have a set of nodes V, directed edges $E \subseteq V \times V$, and a weight function $w : E \to [0, 1]$ in which $w(i, j)$ represents the probability of node i influencing its neighbor j. We hope to select a subset of the network $S \subseteq V$ that maximizes our spread through the network, which we assume to follow an independent cascade model. In this model, all nodes $v \in S$ are activated, and each activated node subsequently activates its neighbors independently with probability $w(i, j)$. Specifically, we use a weighted cascade model where the activation probability of an edge (i, j) depends on the in-degree of j: $w(i, j) = \frac{1}{\sum_{k \in V} (k,j) \in \mathcal{E}}$. Intuitively, this means that nodes are more likely to be activated (or receive the information we hope to spread) if they have fewer neighbours that are connected to them.

However, the assumption of having complete knowledge of the full network is unrealistic. Thus, we focus only in the case of partially observable networks. A useful concept here is the *organization-partitioned* network $O \subseteq G$, a subset of the complete network to which decision-makers have complete information. Then, decision makers also see the neighbours of all the nodes in O which are not in O, denoted here as *boundary nodes*. We assume that only members of O can be selected as seeds, so our objective is to find a set of seed nodes $S \subseteq O$ that maximizes information spread across the underlying network G.

Stein et al. (2017) introduces four heuristic methods for maximizing this information spread. First, and most trivially, is the random seed selection method, in which we choose observable nodes uniformly at random. The second approach is random selection with neighbor activation, which selects random nodes and, for each node, activates one of its neighbors. This leverages the friendship paradox: on average, a node's neighbors have more neighbors than that node.

[1] The code along with the experiment's results can be accessed at `https://github.com/lylakirati/influence_maximization/`

These first two methods do not take advantage of any information we have on the observable part of a network, so they are typically not great candidates for selecting seeds in the influence dissemination process. To remedy this, Stein et al. (2017) proposes the (weighted) degree-based seed selection method, in which nodes with the highest out-degrees are selected as seeds. However, under this selection scheme, it is likely that we would select the most central nodes in the observable part and might not promote message propagation in the unobserved part of the underlying network G. Thus, Stein et al. introduces a weighted degree-based version which adds additional weight w to the boundary nodes (i.e. if a boundary node has degree of k, we would rank it in our seed selection as if it had degree of $k + w$). This promotes the selection of nodes at the edges of the observed network, increasing the probability that we activate nodes outside of O.

Finally, Stein et al. (2017) benchmarks their degree-based heuristic against the state-of-the-art influence maximization algorithms such as TIM (Tang et al., 2014) and IMM (Tang et al., 2015), which identify optimal seed nodes based on the concept of *reverse reachable sets*. The reverse reachable set for a node v in G is the set of nodes in G' that have a directed path to v, where G' is the network after removing each edge $e \in E$ with some probability $1 - p(e)$. IMM generates θ many of these sets, having estimated θ based on the input network G and the number of initial seed nodes k. Once IMM has obtained these reverse reachable sets, the optimal selection of the k seed nodes are those which cover the most reverse reachable sets generated.

To situate our findings in the literature, we give an overview of the experiments in Stein et al. (2017). The study is focuses on the NetHept dataset—a collaboration network within the high energy physics theory community between 1991 and 2003. The study generates an organization-partitioned network from a set of observable nodes O by selecting a set of initial nodes uniformly at random, and then iteratively adding a randomly selected neighbor of the nodes in O that is not already in O until reaching the desired number of nodes. If the set O is not yet of the desired size and, at any point, there are no more neighbouring candidates to add, then some other node $v \in V$ is selected uniformly at random and added to O. After obtaining O, they select seeds using the four approaches mentioned above and simulate information dissemination using the weighted independent cascade model. Using these generated, partially observed networks, Stein et al. finds that in settings where network visibility is low (e.g. less than 5%), the weighted degree-based seed selection tends to perform best. However, when more of the network is observed, the IMM method starts to outperform in terms of average spread; yet the algorithm runs much slower than the degree-based. Also, when more of the network is observed, the weighted degree method with larger weights performs worse than those with smaller weights. And as expected, the two naive, random methods perform poorly.

3 Simulation

The results obtained by Stein et al. (2017) indicate that, using the NetHept dataset, simple heuristics perform best in settings where both network visibility and the size of the initial seed set are very low. In settings where only 1%-5% of the network is observable and fewer than twenty initial seeds are chosen, seed selection based on weighted degrees consistently achieves better results than state-of-the-art algorithms and stochastic seed selection processes. Yet, without rigorous testing and a broad application of these methods across multiple datasets, it is uncertain to conclude

whether these findings are due to the power of degree-based heuristics or a confounding property of the underlying NetHept network structure.

One characteristic of the NetHept dataset worth investigating is its degree assortativity, defined as the preference for a network's nodes to attach to others that are similar in degrees, which the authors note may have influenced the success of their proposed heuristics. Similar to most citation networks, the NetHept network exhibits high degree assortativity; high degree nodes are more likely to be connected to other high degree nodes than low degree nodes and vice versa. While significant research has been conducted to establish a distinction between influential nodes and high-degree nodes (Cha et al., 2010), possible links between these two properties may still affect influence dissemination results.

To explore the impact of assortativity on the performance of Stein et al. (2017)'s proposed degree-based heuristics, we run two experiments, one working directly with manipulations of the NetHept dataset and the other leveraging applications to other datasets for comparison. Together these simulations shed light on how changes in degree assortativity affect the spread of influence on a graph network when initial seed selection is based on weighted degree. The design and results of each experiment is described below.

3.1 Degree Assortativity and the Effects on Spread: NetHept Dataset

3.1.1 Experimental Design

By altering the assortativity of the NetHept dataset and examining how influence spreads, we can gain insights on how assortativity impacts the performance of the degree-based heuristics. The original NetHept dataset has a degree assortativity of 0.3161, indicating that high degree nodes tend to have neighbors of high degree. Throughout our experiment, we generate manipulations of the dataset with varying levels of degree assortativity by randomly altering connections between nodes. We first sample k edges (u, v_1) from the original network. Then, for each edge (u, v_1), we remove the edge from the graph and replace it with a new edge (u, v_2) where $v_2 \in V - v_1$ is a randomly chosen node in the graph. In this way, we encourage random connections between nodes, reducing the tendency of similar degree nodes to be linked and, thus, decreasing assortativity. We do this for multiple k values equivalent to 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the total number of edges in NetHept. As a result, we create 12 variations of the original NetHept network with corresponding degree assortativities ranging from 0.11 to 0.315.

In this experiment, we focus specifically on evaluating the weighted degree-based seed selection heuristic, the best performing heuristic found by Stein et al. (2017) given limited visibility. Below we visualize algorithm performance when attaching a weight of three to boundary nodes, though similar results can be observed with different weights. To remain consistent with Stein et al., we generate partially observed subgraphs, use five initial seed nodes, and simulate spreading based on the weighted cascade model in the same manner. On each assortativity variation of the NetHept dataset, we generate eleven partially observed networks with visibilities ranging from 1% to 21%; and, for each observed graph, we repeat the described experiment for five times to obtain an average spread.

3.1.2 Results

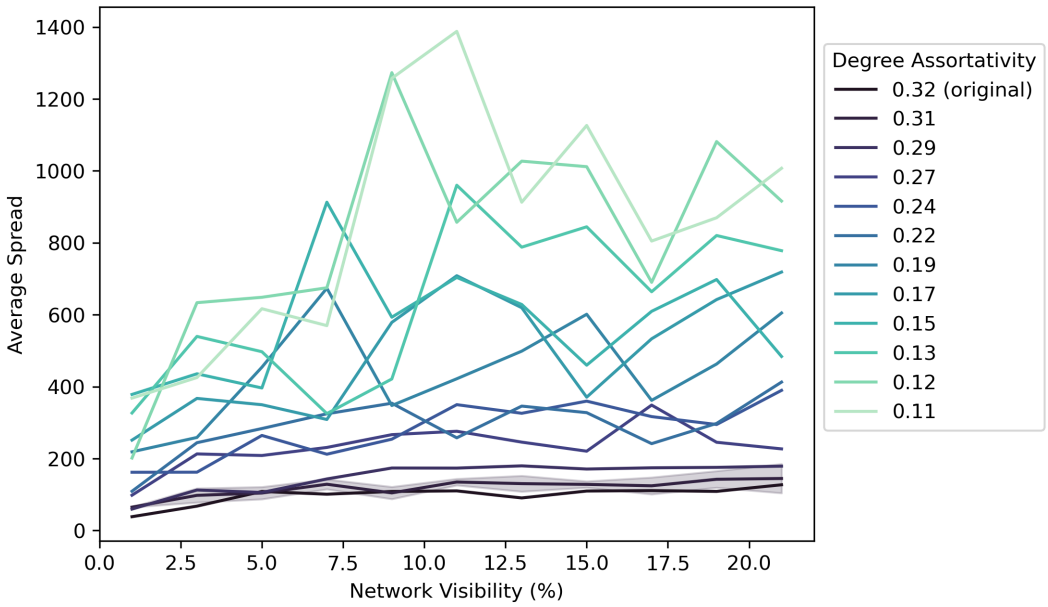

Figure 1: Influence spread on modifications of the NetHept dataset with varying assortativity

Figure 1 indicates one evident trend: across all levels of network visibility, as the assortativity of the NetHept dataset decreases, the average spread increases. With higher levels of assortativity, matching or near the original network's value, average spread never exceeds 200, aligning with results from Stein et al. (2017). However, with reduced levels of assortativity, seeding with the weighted degree-based heuristic performs significantly better, consistently reaching spread levels above 500. The greatest such difference comes when we run the algorithm to initialize seeds on a network with 10% visibility. The original dataset, with degree assortativity of 0.32, leads to an average spread of 110.0 nodes, while the modified dataset, with assortativity of 0.11, leads to an average spread of 1387.8 nodes; this mere 66% decrease in assortativity gives rise to an 1162% increase in average spread.

Interestingly, for any given variant of the NetHept network, as the visibility increases, we do not see a monotonic increase in average spread. While the general trend is increasing, we often see several dips in performance followed by sharp pickups. These shifts are likely because heuristic performance depends not only on the amount of network visibility but also which particular subgraph of the network is observed. With the computational power to run more than five simulations on each modified dataset before averaging, it is probable that these curves would begin to increase more smoothly, as they do in the Stein et al. (2017) results, where a thousand simulations were performed. Noticeably, the variability is even more pronounced for modified sets with lower assortativity. This signals that in graphs with lower assortativity, the performance of degree-based heuristics for influence maximization is more sensitive to which part of the network is fully observable than in graphs with higher assortativity.

3.2 Degree-Based Heuristic Performance Across Datasets

3.2.1 Experimental Design

Exploring influence spread across multiple datasets provides another opportunity to see how assortativity, and potentially other network properties, may affect heuristic performance. In addition to the original NetHept dataset, we also apply the algorithm to the five following directed graphs: the Gnutella Peer to Peer network, the Stanford Web network, the Amazon Product network, the Wikipedia Vote network, and the European Email network, all publicly available through Stanford's Large Network Dataset Collection. The varying characteristics of these real-world datasets make them a good choice for a more complete analysis of the degree-based seed selection algorithm. Key descriptive statistics for each graph are presented in Table 1 below. More detailed information on the networks is available at Stanford's Large Network Dataset Collection.

Dataset	**Assortativity**	**Nodes**	**Edges**	**Diameter**	**Avg. Clustering Coefficient**
NetHept	0.3161	15233	58891	31	0.2621
Gnutella	−0.0037	10876	39994	9	0.0062
Stanford	0.04576	281903	2312497	674	0.5976
Amazon	0.1027	262111	1234877	32	0.4198
Wikipedia	−0.0189	7115	103689	7	0.1409
EU Emails	−0.0018	1005	25571	7	0.3994

Table 1: Descriptive statistics for NetHept and selected networks from Stanford's Large Network Dataset Collection

As in our simulations with the modified NetHept dataset, here we focus on evaluating the performance of the weighted degree-based seed selection heuristic with a weight of three. We again generate our partially observed graphs, use five initial seed nodes for the algorithm, simulate information spreading based on the weighted cascade model following Stein et al. (2017). Instead of raw spread numbers, we report the average spread as a percentage of the total nodes in the given network, allowing for algorithm performance comparison across datasets of different sizes. We repeat this process on each dataset, reporting results for network visibility levels of 1%, 5% and 11%, reflective of sparsely observed networks.

3.2.2 Results

Dataset	Assortativity	Average Spread (%)
NetHept	0.3161	0.32
Gnutella	−0.0037	4.74
Stanford	0.04576	0.16
Amazon	0.1027	0.0027
Wikipedia	−0.0189	1.86
EU Emails	−0.0018	19.44

Table 2: Average Spread in Various Networks with **1% Network Visibility**

Dataset	Assortativity	Average Spread (%)
NetHept	0.3161	0.59
Gnutella	−0.0037	5.28
Stanford	0.04576	0.13
Amazon	0.1027	0.0033
Wikipedia	−0.0189	1.75
EU Emails	−0.0018	18.33

Table 3: Average Spread in Various Networks with **5% Network Visibility**

Dataset	Assortativity	Average Spread (%)
NetHept	0.3161	0.80
Gnutella	−0.0037	8.81
Stanford	0.04576	0.18
Amazon	0.1027	0.0034
Wikipedia	−0.0189	1.83
EU Emails	−0.0018	20.36

Table 4: Average Spread in Various Networks with **11% Network Visibility**

Tables 2, 3, and 4 display the results of our simulations. First, observe a general trend in how assortativity affects performance—applying degree-based heuristics on networks that exhibit degree assortativity (i.e. positive assortativity coefficient) consistently results in less average spread than applying them to networks exhibiting degree disassortativity (i.e. negative assortativity coefficient). In graphs with positive degree assortativity, less than 1% of the overall network is activated. On the other hand, graphs with negative degree assortativity achieve influence spreads greater than 1% of the overall network, with a maximum average spread over 20%. Note that these negative assortativities are all also very close to zero, so lack of assortativity, rather than disassortativity, could be driving this change. While less precise, this does align well with our findings from experimenting with the NetHept dataset where we also find that lower assortativity levels lead to better

performance of the algorithm.

Although the presence or lack of assortativity appears to affect results, it is difficult to see a direct relationship between the strength of degree assortativity and performance. Across all visibility levels, graphs with similar assortativity values sometimes exhibit drastically different average influence spreads when using the degree-based heuristics for initial seed selection. This observation points to the possibility that other graph characteristics likely also have a significant impact on the success of the algorithm; we keep this in mind as we consider directions for future work.

Generally, we can also observe that for the majority of the datasets explored, across different visibilities, on average, less than 5% (and often even less than 1%) of the nodes become activated. Not only does this well reflect the difficulty of influence maximization in general, but it may also be an indicator that five initial seeds (as used in our trials) can not sufficiently initialize an influence spread in these graphs, or even that degree-based heurisitcs are not powerful enough to spur significant activation in these settings. Both of these avenues should be explored in future work.

4 Discussion

This paper both summarizes the heuristic-based influence maximization approaches proposed by Stein et al. (2017) and presents preliminary results on how network assortativity impacts the performance of these heuristic. In examining the efficacy of weighted degree-based seeding on variants of the NetHept and several additional networks, we discover a clear trend: influence tends to spread better in networks with lower assortativity. This is a surprising result; Stein et al. (2017) hypothesizes that one of the reasons that degree-based heuristics perform well on the NetHept dataset is that this network has a high level of assortativity. The relative success of degree-based heuristics on networks with lower assortativity implies a distinction between high-degree nodes and high-influence nodes, since, in networks with low assortativity, high degree nodes are not likely to be connected to other high degree nodes. Despite this compelling result, our experiments still only marks a small first step in this exploration.

Additionally, Stein et al. (2017) finds that the ability of degree-based heuristic methods to promote spread through a network is sensitive to the choice of boundary weights. Specifically, the paper notes that in settings with lower network visibility, increasing boundary weights leads to increased spread, since, intuitively, larger boundary weights increases the probability that we move to a node in the network that we have not yet seen. However, Stein et al. also notes that in settings where network visibility is high, these boundary nodes become less important. Indeed, experimentally the authors see that as network visibility increases, degree-based heuristics with larger boundary weights tend to underperform those with smaller boundary weights. It would be interesting extension to introduce a weighting scheme that is a function of network visibility, in which boundary weights decrease with network visibility. For the purposes of our experiments, we choose to isolate network assortativity as the variable of interest.

In future work, we hope to study this setting in greater detail, expanding our analysis past these simulations. Given more greater computational capacity, we would like to increase the size of each

experiment, running the algorithm more than five times for each graph-visibility pair. This would not only lead to more accurate averages, but also allow us to report valid confidence intervals for the spread values. We also hope to investigate more thoroughly the impact of other graph properties on heuristic performance. It is unlikely that degree assortativity is the only graph characteristic affecting the spread of influence; the effects of changes in attributes such as triangle density, clustering coefficient, node-to-edge ratio, and more could provide further insight on when degree-based heuristics perform best.

Moreover, we hope to apply state-of-the-art models for influence maximization to the datasets used for heuristic analysis above. While our work centers on comparing degree-based heuristic performance across networks with different assortativities, it would be interesting to see how well state-of-the-art models perform in these same settings. Lastly, we hope to study the impact of using different sampling methods and initial seed set sizes on heuristic performance.

References

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm).* Retrieved from `http://scholar.google.de/scholar.bib?q=info:rqhbqWEH79kJ:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=10`

Eckles, D., Esfandiari, H., Mossel, E., & Rahimian, M. A. (2022, July). Seeding with costly network information. *Operations Research*, *70*(4), 2318–2348. doi: 10.1287/opre.2022.2290

Stein, S., Eshghi, S., Maghsudi, S., Tassiulas, L., Bellamy, R. K. E., & Jennings, N. R. (2017). Heuristic algorithms for influence maximization in partially observable social networks. In *SocInf@IJCAI.*

Tang, Y., Shi, Y., & Xiao, X. (2015). Influence maximization in near-linear time: A martingale approach. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*.

Tang, Y., Xiao, X., & Shi, Y. (2014). *Influence maximization: Near-optimal time complexity meets practical efficiency.*