

A Scalable Hybrid Framework for Continuous Influencer Intelligence

Nidhi Gupta¹, Dr. Sohit Agarwal²

¹Research Scholar, Department of Computer Engineering and Information Technology, Suresh Gyan Vihar University, Jaipur, Rajasthan, India

²Principal, ASTER School of Computer Science & Professor, Department of Computer Engineering and Information Technology, Suresh Gyan Vihar University, Jaipur, Rajasthan, India

Abstract—Finding a small number of influential users who are able to disseminate information to potentially the largest number of users has been known as the problem of Influence Maximization (IM) in social network analysis. While the standard solution to this problem is indeed correct, it involves many cycles of simulation of how information propagates through the social network which makes it infeasible for the network with more than a small number of nodes. The common and fast method to tackle this problem is to rank users according to their out-degree, and therefore, to select seed users according to this ranking. At first glance, this method is both efficient and convenient, but it discards important information, such as the degree of influence of users, the engagement of their followers, or the Uniqueness of their role in the community. This paper presents and tests the so-called Champion Hybrid Framework which consists of seven different scores combined into the Hybrid Score that is then used for the seed user's selection while avoiding overlaps in the influence of the selected users.

With its 288 connections and samples from more than 2000 users, the method of Hybrid has led to the greatest number of individuals reached in percentage terms (3.57%) after overcoming out-degree ranking by almost 2% and PageRank by more than 120%. Additionally, Gradient Boosting was able to predict Hybrid Score using only five basic predictors, which makes it possible to carry out this assessment easily and low-cost in the future by using only relatively low-cost features. The paper presents several relevant examples and describes the process of the work.

Index Terms—Influence maximization, social network analysis, independent cascade, linear threshold, PageRank, seed selection, hybrid scoring, gradient boosting

I. INTRODUCTION

In any social media, few accounts receive most of the engagement. For corporations, health institutions and platforms that want their message to be heard by many people, the immediate question is: which few users to approach first? This is known as the Influence Maximization (IM) problem. The aim is to find a small seed set of users so that as soon as they disseminate information, it will reach the maximum number of people through the network.

The standard technique to solving the IM problem provides strong mathematical guarantees. It repeatedly runs simulations to determine how much in terms of reach every potential user can contribute and builds the seed set in a greedy manner (Li et al., 2023). However, this approach has a significant limitation, namely computational efficiency. Since the process must be done for every candidate at each selection stage, the cost of computing grows quickly as the network scales beyond a few thousand users. Therefore, many real-life systems take advantage of computationally efficient alternatives such as ranking users by their degrees, PageRank score, and role as intermediaries in the network. Available evidence suggests that the methods in question are associated with trade-offs between speed, scalability, and accuracy, while no technique seems to be good enough in every type of network (Fan et al., 2025).

This situation poses a challenge to practitioners looking for efficient influencer rankings that are regularly updated. Computation-intensive simulation techniques lead to good results, but the processes are very costly. Alternatively, faster techniques fail to consider many details. The Champion Hybrid

Framework presented in this paper aims to overcome this issue. The framework is based on the principle that instead of relying on a single measure of influence, seven influencing signals are calculated and then combined into an overall score; furthermore, a special selection method is applied to minimize overlapping substances in the finals score.

Table I. Key Terms at a Glance

Term	Plain-language meaning
Seed set	The minor initial set of users who have been selected to receive messages first.
Out-degree	The amount of other users that an account has followed or pointed at by that account outwardly (its outgoing links).
PageRank	An importance score indicating general significance given the users that link to it rather than the link count.
Betweenness	The frequency by which a user is present on the shortest path connecting two other users (the bridge score).
Independent Cascade (IC)	A community propagation model where each newly active user tries once to infect each of their neighbors.
Linear Threshold (LT)	A community propagation model, whereby a user becomes active only after considerable number of neighbors have become active.
Diffusion coverage	The proportion of the entire network reached from its seed set.

II. RELATED WORK

According to recent surveys, Influence Maximization (IM) techniques are divided into three types: structural, diffusion-based, and learning-based methods. There has been an increase in the number of hybrid methods that utilize multiple signals to identify influential users (Li et al., 2023; Fan et al., 2025). Effectiveness of signals through studies show that embracing the hybrid model of degree, PageRank, and

betweenness is more rewarding compared to using individual signaling systems, especially in dependence of network structure (Tang et al., 2023; Khatri et al., 2023). The research strives to include unique community structures and higher levels of networking relations for the process of seed selection that ensures spreading in different regions of the structure, and minimizing the concentration of seeds within one community only (Zhang et al., 2024; Zhao et al., 2025). This can be proved by different multimodal fusion studies, which state that integration of unique heterogeneous characteristics creates better representation rather than application of single characteristics only (Kansal, 2026). Another trend is based on the use of Gradient Boosting ideas for relevant prediction of influence scores through application of combination between structural and behavioral characteristics (Subramani et al., 2023). The framework suggested here will build on the results of the studies mentioned above by gathering seven different signals that exert an effect and merging them into a single score; the effectiveness of Gradient Boosting in predicting this score with high precision and scalability will be examined afterward.

III. METHODOLOGY

The framework follows a total of eight steps, which include loading and cleaning user data, creating a network of connections, determining the useful attributes for every user, putting those attributes together to form one Hybrid Score, choosing a comprehensive seed set based on the score attained, evaluating the seed set using two different spreading simulation approaches, developing a model which uses low-cost attributes to predict the score, and creating the articles with the charts that will be shown in this report.

A. Preparing The Data and The Network

User data is processed first; slow and expensive work with big datasets is avoided by filtering them in favor of the most promising profiles according to some simple performance indicators (such as followers' count, engagement level, and growth dynamics). To prevent the distortions made by some outrageous values (where one account has millions of followers), numerical columns are normalized using the 1st and 99th percentiles as boundaries. Finally, the network is

formed from who-follows-whom records and the weights in this connection emphasize how probable it is for the given user to disseminate information (and increase this weight a bit later in our study as the power of the engagement itself).

B. Measuring Each User

For each user, the framework assesses: the number of followers and users they follow; the same values, but weighted; PageRank, which rewards users who are connected with important users; and betweenness centrality, which denotes the users who serve as a bridge between the isolated components of the network. These indicators are integrated with the basic values of the user's followers and his engagement rate to obtain one set of indicators for every user.

C. The Champion Hybrid Score

The framework combines seven normalized signals into one Hybrid Score, assigning a fixed percentage to each signal as follows:

Out-Degree (40%): This is the most significant part of the score and indicates how many accounts are reachable with the target account.

Out-Strength (25%): This indicates the same information as Out-Degree signal but adds the weight in accordance with the likelihood that certain connection will lead to passing the information.

PageRank (15%): This is given to accounts that have numerous connections with connection accounts.

In-Degree (8%): This indicates the number of the followers of the account.

Reach (5%): This adds a bonus to the score due to the good number of followers the account has.

Engagement (4%): The score indicates how much the audience interacts with the account.

Betweenness (3%): This give points to users who have connections with various communities.

Written as a formula: $H(v) = 0.40 \cdot \text{OutDeg} + 0.25 \cdot \text{OutStr} + 0.15 \cdot \text{PR} + 0.08 \cdot \text{InDeg} + 0.05 \cdot \text{Reach} + 0.04 \cdot \text{Eng} + 0.03 \cdot \text{Betw}$. Almost 65% of the total weight

is attributed to raw connectivity (out-degree and out-strength) because raw connectivity is known to be an important predictor of reach by itself. The remaining signals serve as tiebreakers and adjustments and are responsible for identifying quite a few kinds of accounts, such as high reach but low connection profiles like that of a celebrity from the foregoing connections.

D. Choosing a Diverse Seed Set

Selecting the top 20 users based on Hybrid Score could turn out to be ineffective because the users with highest Hybrid Scores would probably know each other and significantly overlap in their audiences. Hence instead of taking the users with the highest scores, the approach taken here is to develop the seed set of users gradually. Every time the leading candidates who have not been chosen earlier are evaluated, and the user with the best combination of high Hybrid Score and low overlap with the already picked up users is chosen based on the similarity of their two-hop networks.

E. Testing the Seed Set

A high score alone does not imply that a seed set will effectively disseminate the information in real life. Thus, there is a need to evaluate each seed set based on 150 repeated simulations done in two independent diffusion processes, Independent Cascade, where every new user has only one attempt to influence each of his/her connections, and Linear Threshold, where the user adopts the information when his/her connections have already adopted it. The evaluation of seed sets in two processes guarantees that results are not biased by the assumption of a single information diffusion process.

F. Predicting the Score Cheaply

In the end, to test if the Hybrid Score can be calculated in an effective way without the need to perform costly network measures again and again, two models, Random

Forest and Gradient Boosting, are used to predict the Hybrid Score based on the five features that can be extracted easily, such as followers, engagement rate, in-degree, out-degree, and PageRank. This approach is like the work of other researchers who attempted to build models predicting influencer scores based on minimal features (Subramani et al., 2023).

IV. EXPERIMENTAL SETUP

For the study, a directed network comprising of 2,000 users and 288 connections has been adopted. The significant parameters were set which included a size of the seed set of 20, a candidate pool that contained the best 500 users who scored well, a mild penalty for diversity of 0.02, a propagation boost of 1.8, a maximum of 10 hops, and 150 simulation runs have occurred in relation to every seed set taking into consideration every spread model. The whole process, which consists of data loading, scoring, seed selection, total runs of 600 simulation runs, and model training took about 5.4 seconds and thus confirms that the method is efficient enough for the regular simulation process.

V. RESULTS

A. How Far Did Each Method Spread?

Table II. Mean diffusion coverage across 150 simulations per method, averaged over IC and LT

Table II. shows the average share of the network reached

Method	IC Coverage	LT Coverage	Avg. Coverage
Hybrid	0.0336	0.0378	0.0357
OutDegree	0.0327	0.0373	0.0350
PageRank	0.0161	0.0163	0.0162
Random	0.0103	0.0104	0.0103

Hybrid had the highest results under both spreading schemes, exceedingly only the basic out-degree ranking by 2%, and PageRank by about 120% and random selection by more than 240%. This small difference can be explained by the scoring system, since out degree and its weighted version constitute 65% of the total score and therefore both methods produce highly overlapping seed sets. The remaining five signals then provide a small but rather steady improvement than the fundamentally different rankings. The lower performance of the PageRank method can be explained mainly by the network's sparsity, having only 288 connections among 2,000 users; as a result, there is less than one outgoing link per user on average. Because of this lack of depth in the network, PageRank is biased toward users who are linked to highly ranked users, rather than users able to reach many users directly.

B. Who Got Chosen, and How Well Can We Predict the Score?

Table 2 indicates the first-ranked Hybrid seeds. The highest-ranking seed is user 64231, featuring a total number of outgoing connections equal to 70 that is the maximum number of outgoing connections in the whole network. However, the number of its followers is only equal to 739. Thus, this indicates that the structure can choose between far-reaching connections and the number of followers. There are several seeds at the lower end of the table having only 2-3 outgoing connections and thousands of followers, e.g., user 26210 with 18,200 followers. This shows that the composite method is able to keep popular accounts with the number of outgoing connections much lower than for all other seed accounts. It is worth mentioning that the average score of the chosen accounts is equal to 0.41 that is much higher than the regular average score of any user equal to 0.124.

Table III. Top 8 of 20 selected Champion Hybrid seeds, ranked by Hybrid Score.

Rank	User ID	Hybrid Score	Out-Deg.	Followers
1	64231	0.8692	70	739
2	26210	0.5446	26	18,200
3	66715	0.4943	7	2,700
4	64125	0.4757	4	1,400
5	65449	0.4643	5	1,100
6	64612	0.4250	3	745
7	39395	0.4154	3	1,600
8	26300	0.4119	3	1,500

When it comes to predictions, the Gradient Boosting model performed much better than Random Forest in predicting the Hybrid Score using just five basic features. It achieved a testing R^2 of 0.9913, which equates to the model explaining around 99% of the variability in scores while having a very low RMSE value of 0.0044.

This is significant because it permits frequent updates of the ranking with cheap simple account data, as opposed to having to compute more expensive network-wide metrics such as betweenness every time the ranking is updated. This is in line with what several researchers found out earlier: that gradient-boosting models can reconstruct influencer rankings from elementary input features (Subramani et al., 2023).

VI. DISCUSSION AND LIMITATIONS

Several honest caveats are important to mention. Firstly, the network for the tests is very sparse (less than one outgoing connection per user). As a result, any topology-sensitive technique cannot yield favorable results than degree ranking, since measurements are very different even in the two networks used for tests (Tang et al., 2023). Secondly, the values for the seven weights for the Hybrid Score (0.40, 0.25, 0.15, and so on) were initially chosen

manually. Previous research shows that an automatic search for an optimum weight combination can help adapt Hybrid Score to a network's topology (Khatri et al., 2023), which worth doing. Thirdly, both simulations for spreading used simpler fixed probability instead of one calibrated by the actual historical data, therefore, the numbers from the research have to be read as a good comparison of the techniques without a possibility to use them for predicting the actual outreach of a promotional campaign.

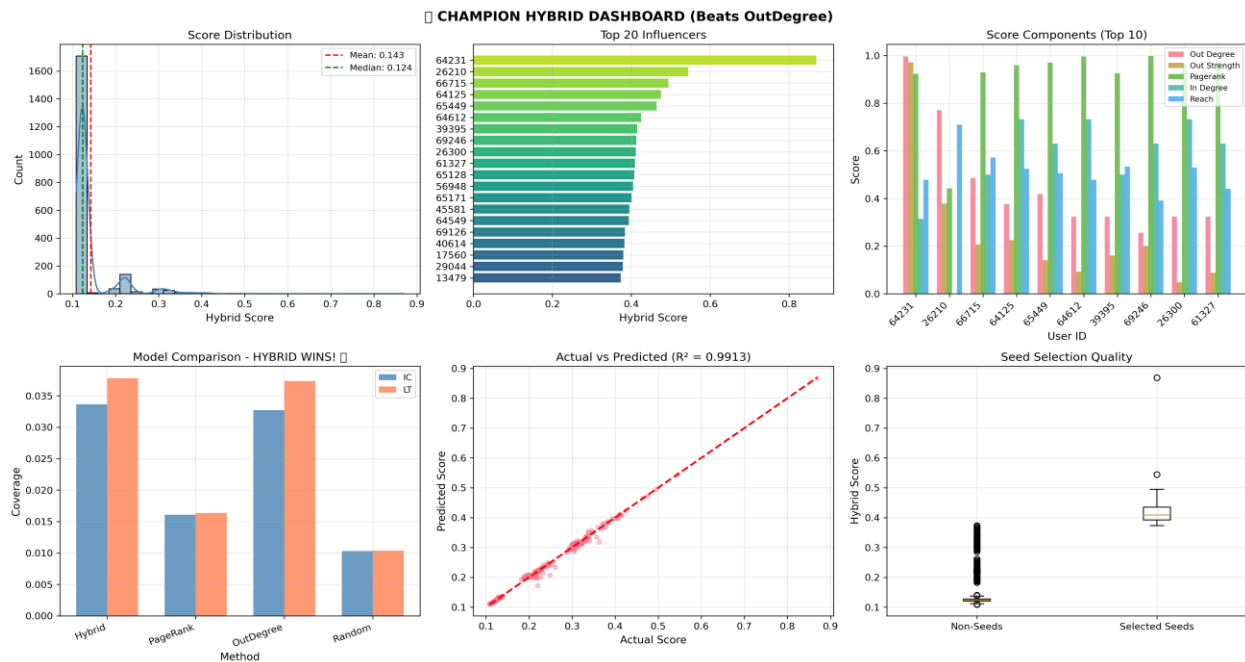


Fig I. Champion Hybrid Dashboard: score distribution; top-20 influencers; per-user score components; diffusion-coverage comparison; predicted-vs-actual score ($R^2 = 0.991$); and seed-selection quality.

Despite these limitations, the framework fulfills its purpose: it operates in significantly less than ten seconds, consistently beats all accessible simple boundaries (out-degree), and remains interpretable because the score can be analyzed into seven parts and presented in comprehensible terms. Also, its methodical development permits the replacement of denser networks and addition of new behavior characteristics or intelligent weights without modifying the entire framework.

VII. CONCLUSION

The Champion Hybrid Framework is a combination of seven structural as well as behavioral signals that

produces a single interpretable score. Instead of using high-ranked users, it chooses a diverse group of influential users. The framework has been validated using two independent spreading simulations. In a network of 2,000 users, it achieved higher average network coverage than some methods like out-degree ranking and the PageRank and surpassed out-degree methods by 2 percent and PageRank by 120 percent. Notably, the method needed around five seconds to finish the selection process while being highly predictable ($R^2 = 0.991$).

Further research should focus on testing the method in large and dense real-life networks, learning the seven weights automatically instead of assigning them

manually, and calibrating the spreading simulations based on real information on past sharing practices.

REFERENCES

- [1] C. Fan, Z. Wang, J. Zhang, J. Zhao, and X. Meng, "Meta-heuristic algorithms for influence maximization: A survey," *Evolving Systems*, vol. 16, Art. no. 24, 2025, doi: 10.1007/s12530-024-09640-2.
- [2] M. Kansal and S. Agrawal, "Optimizing multimodal fusion for enhanced next-gen E-commerce," *AIP Conference Proceedings*, vol. 3426, no. 1, Art. no. 030045, 2026, doi: 10.1063/5.0327265.
- [3] I. Khatri, A. Choudhry, A. Rao, A. Tyagi, D. K. Vishwakarma, and M. Prasad, "Influence maximization in social networks using discretized Harris' Hawks Optimization algorithm," *Applied Soft Computing*, vol. 149, Art. no. 111037, 2023, doi: 10.1016/j.asoc.2023.111037.
- [4] H. Li, S. Yang, M. Xu, S. S. Bhowmick, and J. Cui, "Influence maximization in social networks: A survey," *arXiv*, arXiv:2309.04668, 2023, doi: 10.48550/arXiv.2309.04668.
- [5] J. T. Papadakis and H. Kondylakis, "Optimizing influence maximization in social networks via centrality-driven discrete particle swarm optimization (DPSO)," *Electronics*, vol. 15, no. 8, Art. no. 1730, 2026, doi: 10.3390/electronics15081730.
- [6] N. Subramani, S. Veerappampalayam Easwaramoorthy, P. Mohan, M. Subramanian, and V. Sambath, "A gradient boosted decision tree-based influencer prediction in social network analysis," *Big Data and Cognitive Computing*, vol. 7, no. 1, Art. no. 6, 2023, doi: 10.3390/bdcc7010006.
- [7] J. Tang, H. Zhu, J. Lan, S. Song, J. Qu, and Q. Du, "Identifying influential nodes for influence maximization problem in social networks using an improved discrete particle swarm optimization," *Social Network Analysis and Mining*, vol. 13, Art. no. 94, 2023, doi: 10.1007/s13278-023-01098-5.
- [8] C. Zhang, W. Cheng, F. Li, and X. Wang, "Hypergraph-based influence maximization in online social networks," *Mathematics*, vol. 12, no. 17, Art. no. 2769, 2024, doi: 10.3390/math12172769.
- [9] Z. Zhao, X. Liu, Y. Sun, N. Zhang, A. Hu, S. Wang, and Y. Tu, "Influence maximization based on bottom-up community merging," *Chaos, Solitons & Fractals*, vol. 193, Art. no. 116078, 2025, doi: 10.1016/j.chaos.2025.116078.