Network Science **12** (1): 65–87, 2024
doi:10.1017/nws.2023.23

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Audience selection for maximizing social influence

Balázs R. Sziklai[1,2] and Balázs Lengyel[1,3,4]

[1]Institute of Economics, HUN-REN Centre for Economic and Regional Studies, Budapest, Hungary, [2]Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest, Budapest, Hungary, [3]Corvinus Institute for Advanced Studies, Corvinus University of Budapest, Budapest, Hungary, and [4]Institute of Data Analytics and Information Systems, Corvinus University of Budapest, Budapest, Hungary
Corresponding author: Balázs R. Sziklai; Email: sziklai.balazs@krtk.hun-ren.hu

Abstract

Viral marketing campaigns target primarily those individuals who are central in social networks and hence have social influence. Marketing events, however, may attract diverse audience. Despite the importance of event marketing, the influence of heterogeneous target groups is not well understood yet. In this paper, we define the Audience Selection (AS) problem in which different sets of agents need to be evaluated and compared based on their social influence. A typical application of Audience selection is choosing locations for a series of marketing events. The Audience selection problem is different from the well-known Influence Maximization (IM) problem in two aspects. Firstly, it deals with sets rather than nodes. Secondly, the sets are diverse, composed by a mixture of influential and ordinary agents. Thus, Audience selection needs to assess the contribution of ordinary agents too, while IM only aims to find top spreaders. We provide a systemic test for ranking influence measures in the Audience Selection problem based on node sampling and on a novel statistical method, the Sum of Ranking Differences. Using a Linear Threshold diffusion model on two online social networks, we evaluate eight network measures of social influence. We demonstrate that the statistical assessment of these influence measures is remarkably different in the Audience Selection problem, when low-ranked individuals are present, from the IM problem, when we focus on the algorithm's top choices exclusively.

Keywords: Influence maximization; sum of ranking differences; innovation spreading; audience selection; linear threshold diffusion model

1. Introduction

Suppose an NGO would like to launch a drug prevention campaign in schools or a politician competes for votes in towns. Due to limited resources, only few high schools can be visited in the prevention campaign or few speeches can be held before elections. How should they select a handful set of audiences to achieve the widest reach through word-of-mouth?

Targeting individuals in a marketing campaign to trigger a cascade of influence has been in the spotlight of research in the past two decades. In their seminal paper, Kempe et al. (2003) formulate the *Influence Maximization problem* (IM) as follows: Given a diffusion model D, a network G and a positive integer k, find the most influential k nodes in G whose activation results in the largest spread in the network.

However, it is less understood how the IM method can be applied in traditional marketing settings, like public event organization. Marketing campaigns typically include group targeting in the form of commercials or public events but the message of these potentially spreads further in the social networks of the initial audience (Shi et al., 2019; Yan et al., 2017; Saleem et al., 2017).

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Motivated by the frequent combination of target group selection and viral marketing, which is difficult to address in the original IM setting, we formulate the *Audience Selection* (AS) problem. Given a diffusion model D, a network $G = (V, E)$ and an activation mechanism $f{:}2^V \to 2^V$ assess the collection of possible node sets (audiences) $S_i \subset V,\ i = 1, \ldots, n$ according to the spread that their activation generates in order to find $S_{\hat{i}}$ of maximum spread. We pose no restriction on the subsets, they might be of varying sizes and some might even overlap. At the beginning of the campaign, one S_i is targeted. The activation mechanism denotes the individuals $f(S_i) \subseteq S_i$ who were successfully persuaded to participate in the campaign. The diffusion model is applied on $f(S_i)$, and the influence spread is determined.

Note that f and D are different in nature. The activation mechanism is the result of a campaign event that consists of various marketing elements (speeches, ads, or distributed product samples), while the diffusion mechanism imitates the personal interactions between users, commonly referred to as word-of-mouth.

In the AS problem, we do not seek to construct the best seed set, but to compare available seed sets. Each S_i set represents an audience that could be reached through a campaign involving several public events. There are two inherent difficulties involved. Firstly, the number of sets that need to be evaluated can be astronomical. Secondly, the seed sets are relatively large and contain a wide variety of people—not only opinion leaders, but ordinary folk too.

Although ordinary individuals do not have much influence on their own, they can be pivotal collectively. This stands in contrast to influencers, who in turn can be powerful even if there are only a few of them. The roles of both groups are illustrated by Wang et al. (2019) who found that complex news stories spread much more effectively via "ordinary people." Influencers only play an important role in disseminating simple messages.

In order to evaluate the potential of the possible audiences we need to understand what ordinary agents bring to the table. Influence in networks can be proxied by the centrality of the individuals, but certain centrality indices might be better in ranking ordinary agents than others. Put it differently, an influence measure that is successful in identifying the best spreaders on the whole node set might be less efficient in classifying agents of average or low influence.

In this paper, we devise a statistical test to uncover the real ranking of influence maximization proxies under the AS model. The situation is somewhat similar to polling. When a survey agency wants to predict the outcome of an election it takes a random, representative sample from the population. Our aim is to observe the performance of the measures on "average" nodes; thus, we will take samples from the node set, then compare the obtained results using a novel comparative test method: Sum of Ranking Differences (SRD or Héberger test).

SRD ranks competing solutions based on a reference point. It originates from chemistry where various properties of a substance measured in different laboratories (or by various methods) need to be compared (Héberger, 2010; Héberger and Kollár-Hunek, 2011). SRD is rapidly gaining popularity in various fields of applied science such as analytical chemistry (Andrić, 2018), pharmacology (Ristovski et al., 2018), decision-making (Lourenço and Lebensztajn, 2018), machine learning (Moorthy et al., 2017), political science (Sziklai and Héberger, 2020), and even sports (West, 2018).

Our test method is devised as follows. Let us suppose that the diffusion model D and the network G are fixed. We take samples from the node set of G. For the sake of simplicity, we assume that the activation function is the identity function, $f(S) = S$; thus, every user in the sample is activated. Although this is unrealistic in real campaigns, it is equivalent to the case of taking a bigger sample and activating only a fraction of the users. Since, in this paper, we will work with random samples, there is no difference between the two approaches.

We run the diffusion model with these samples as input and observe their performance, the latter will serve as a reference ranking for SRD. Then, we calculate the influence maximization proxies on the whole network. We aggregate the proxy values for each node in the sample sets. This induces a ranking among the samples: different proxies will prefer different sample sets. Finally, we compare the ranking of the sample sets made by the proxies with the reference ranking.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

The measured distances (SRD scores) show which proxies are the closest to the reference, that is, the most effective in predicting the spreading potential of the different audiences.

We contribute to the literature in two ways. First, we introduce the Audience Selection problem and show that it is inherently different from the Influence Maximization problem. Second, we apply a new statistical test, Sum of Ranking Differences, which enables us to compare network centrality measures. We test the efficacy of different network centralities in a social network environment under both the AS and IM frameworks.

2. Literature and research problem

Social networks became important elements of marketing campaigns. Targeting individuals based on their position in the network and their capacity to influence others is an essential marketing tool and an active area of network science (Aral, 2011; Watts and Dodds, 2007). However, marketing strategies often focus on communities and not on individuals. For example, health interventions target specific groups and by changing their behavior also aim to have a wider impact that can spread in the network (Gold et al., 2011). Similarly, a frequently used technique of brand management is the identification of communities of customers that engage with the brand and can also influence others (Gensler et al., 2013).

Here we argue that the market segmentation and influence maximization techniques should be combined in marketing campaigns run on social networks. However, such an endeavor requires a better understanding of what groups of individuals should be selected for targeting such that social influence is maximized. To demonstrate this research niche, we shortly review grouping techniques and then introduce the research niche in group targeting.

2.1 Grouping for targeting

Identifying the groups of influential individuals is a widely used technique in social network targeting. One line of literature explores how network communities, groups of nodes that are densely connected within and loosely connected across groups, interconnect and influence each other. Yan et al. (2017) introduce a statistical method to reveal the influence relationship between communities, based on which they propose a propagation model that can dynamically calculate the scope of influence spread of seed groups. Rahimkhani et al. (2015) identify the community structures of the input graph, select the most influential communities, then choose a number of representative nodes based on the result. Weng et al. (2014) predict which memes will be successful in spreading using the network and the community structure. Hajdu et al. (2020) study passenger travel behavior to reveal passenger groups that travel together on a given day and model epidemic spreading risk among passengers.

Co-location of individuals is also used by many to group the users for targeting. Li et al. (2014) examine the challenge of providing new companies (*e.g.*, restaurants) with marketing services by locating their potential customers in a region. Similarly, Song et al. (2016) investigate a departmental sales event where the store's location plays a significant role in the campaign. In this scenario, individuals residing far away from the store are unlikely to be affected by the promotional efforts. Both papers aim to find k users who can maximize the number of influenced users within a specific geographical region.

An alternative spatial approach is presented by Cai et al. (2016) who propose a two-layered graph framework, where one layer represents the social network and the other layer connects users who are geographically close. In this framework, particular events (*e.g.,* exhibitions, advertisements, or sports meetings) take place in the physical world and influence multiple users located near the event's position. These influenced users then have the potential to propagate the event further by sharing it with their friends in online social networks or with their neighbors in the physical world. Similarly, Shi et al. (2019) consider influence maximization from an online-offline

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

interactive setting. They propose the location-driven Influence Maximization problem, which aims to find the optimal offline deployment of locations and durations to hold events subject to a budget, so as to maximize the online influence spread. Saleem et al. (2017) use geo-tagged activity data and track how users are navigating between locations and based on this information they select the most influential locations.

2.2 Audience Selection: a niche for methodological development

Targeting groups of individuals is fundamentally different from targeting influential individuals. Here, the task is to select an audience of the message that can be composed of highly influential persons and of less influential ones.

We have found only one paper, that explicitly addresses the problem of group targeting. Eftekhar et al. (2013) consider a billboard campaign that targets groups. The advertiser is allowed to select k locations (that is, k groups) for the campaign. The individuals in the groups are activated with a certain probability. This so-called Fine-Grained Group Diffusion Model (FGD) has a similar mathematical framework as ours, although our formulation, that we will introduce in Section 3, is more general. For instance, Audience Selection can model when either three campaign events can be organized on the East coast or four on the West coast, but it is impossible to formulate this as an FGD.

The statistical assessment of the quality of group targeting is even scarcer. The comparison of centrality measures is accomplished by Perra and Fortunato (2008) who analyze the rankings of nodes of real graphs for different diffusion algorithms. Their focus is on spectral measures such as PageRank and HITS and they adopt Kendall's τ index to calculate their pairwise correlation. However, they do not validate their results by ground truth of diffusion. On the other hand, the robustness of centrality measures has been studied, see Martin and Niemeyer (2019) and the references therein. Here we propose the SRD ranking method, introduced in detail in Section 4, that can assess the quality of Audience Selection and thus can provide new insights for group targeting. SRD scores and distribution can be generated by rSRD, an R package written by Staudacher et al. (2023) and downloadable from the Comprehensive R Archive Network. For the reliability of the employed cross-validation technique, see Sziklai et al. (2022).

3. Audience Selection vs Influence Maximization

Since the proposed problem of Audience Selection is closely related to Influence Maximization, it is worth to explore the differences between the two approaches. Let us consider a small example in which grouping is based on co-location of individuals. Figure 1 depicts a social network with additional geographic information. Colors signify different locations, say towns. Each town has three inhabitants (agents). Red agents are all connected to all the orange and blue agents. The top green agent is connected to all the orange agents and two blue ones, while the bottom green is connected to all the blue agents and two orange. Finally, the middle green agent is connected to all the pink agents.

In the Audience Selection problem, we have to find the most suitable locations for a series of marketing events. Agents corresponding to the same location can be reached simultaneously. Let us assume that the linear threshold diffusion model (a standard choice in the literature, see details in Section 5) adequately represents influence spreading in this network. In our example, one event is organized and every user is activated at the selected location. Which town should be chosen?

Formulating the question like this, we are asking which town provides the best audience. In contrast, the Influence Maximization problem tries to find the k most pivotal agents in a diffusion.

Let us try to answer our questions from the perspective of Influence Maximization. Since each town accommodates three agents, let us fix $k = 3$. Kempe et al. (2003) proposed a greedy algorithm

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

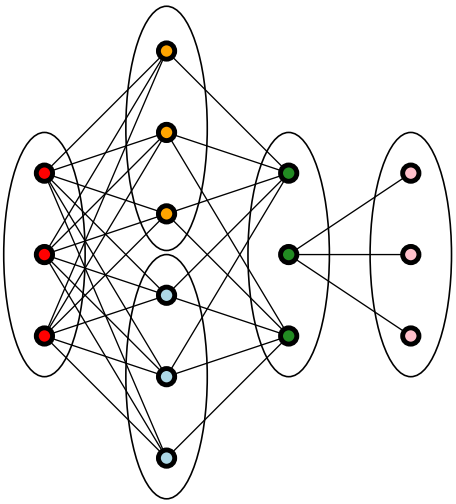

Figure 1. Sketch of a social network with five geographical locations (red, orange, blue, green, and pink) each accommodating three agents.

to approximate the best set of k agents. Since then, many clever heuristics have been invented to improve either the approximation or the running time of the greedy algorithm. However, for such a small example we do not need sophisticated techniques. We can try each combination of agent triplets in a diffusion simulation. It turns out that choosing any two of the red agents, and the middle green agent is the best: On average, they can reach (activate) 78.6% of the nodes.

Since the majority of the most influential agents are red, we are inclined to choose the red town. However, that would be a mistake. Green agents can activate on average 74.0% of all the nodes, while red agents only score 63.6%. Thus, the green town outperforms the red one by a hefty 10%.

Let us approach the question from another angle. Influence Maximization techniques often suffer from computational limitations. It is difficult to even approximate a suitable k set for a huge network. The usual workaround is to use proxy measures, that is, network centralities that were designed to find influential spreaders. For instance, we may calculate the average PageRank value of each agent in a town (for definition and discussion of network centralities, see Section 6). Table 1 contains the PageRank values as well as the Harmonic centrality of each node. It also lists the average centrality values for each town.

PageRank correctly predicts the potential of the towns, in the sense, that larger PageRank values correspond to greater spreading capabilities. Harmonic centrality is far less effective, as it ranks the town with the best audience as the second worst.

From this small example, we can deduce a couple of interesting points.

- Influence Maximization techniques are ineffective in comparing sets of nodes, because the most influential agents might be spread over different sets.
- Although some sets contain less influential agents (top and bottom green agents are less influential than any of the red ones), as a group they might be more effective than any other.
- Aggregating network centralities at a set level can be a good predictor of the sets' spreading potential.
- Some centrality measures are more effective than others. Their performance might also vary depending on the underlying graph structure and the applied diffusion model.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table 1. The PageRank values and Harmonic centrality of the social network depicted in Figure 1. Agents are referred as x_i, where x denotes the first letter of the towns' color (**r**ed, **o**range, **b**lue, **g**reen, **p**ink,) and i denotes the agents location in the town (**t**op, **m**iddle, **b**ottom). Bold numbers represent the maximum value among towns

Agent/ town	PageRank ($\alpha = 0.8$)	Harmonic centrality	Agent/ town	PageRank ($\alpha = 0.8$)	Harmonic centrality
r_t	0.0770	0.571429	g_m	0.1259	0.214286
r_m	0.0770	0.571429	g_b	0.0660	0.52381
r_b	0.0770	0.571429	p_t	0.0469	0.142857
o_t	0.0547	0.488095	p_m	0.0469	0.142857
o_m	0.0653	0.535714	p_b	0.0469	0.142857
o_b	0.0653	0.535714	Red town (avg)	0.0770	**0.5714**
b_t	0.0653	0.535714	Orange town (avg)	0.0617	0.5198
b_m	0.0653	0.535714	Blue town (avg)	0.0617	0.5198
b_b	0.0547	0.488095	Green town (avg)	**0.0859**	0.4206
g_t	0.0660	0.52381	Pink town (avg)	0.0469	0.1428

Based on these observations, we can formulate a hypothesis. We expect to observe a discrepancy between the rankings of the best Influence Maximization proxies and the best algorithms for the Audience Selection problem. In the next section, we will describe the ranking frameworks and present our hypothesis formally. To avoid inferring too much from a small example, we test our hypothesis on two real-life social networks.

Note that the best way to uncover the potential of an audience is by simulation. However, this might be prohibitive for various reasons. Running diffusion simulations are costly, especially since we have to do it many times—often on a very large network too—to get a reliable estimate of the set's spreading potential.

But the real difficulty comes from the cardinality of the sets rather than the running time of one simulation. If a marketing campaign consists of a series of events, each tied to a different location, the number of sets we need to assess can grow excessively. For instance, if we have to pick five towns among 100, that already means more than 75 million combinations that each need an evaluation by simulation. Calculating influence proxy measures, that is, network centralities that were specifically designed with the underlying problem in mind, is a much cheaper and sometimes the only feasible solution. In the example at hand, the analyst would select the five towns with the highest average node centrality.

4. Methodology

In this section—following a brief overview of the data and the SRD method—we introduce our hypothesis and the proposed testing framework.

4.1 Data

For demonstration, we use two real-life social networks. iWiW was the most widely used social network in Hungary before the era of Facebook (Lengyel et al., 2020). Over its life cycle, it had more than 3 million subscribers (nearly one-third of the population of Hungary) who engaged

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

in over 300 million friendship ties. For computational reasons, we use a 10% sample of the node set of this network chosen uniformly at random but stratified by towns and network community structure. Pokec is a Slovakian dating and chatting website. Similarly, we took a random sample of the users of Pokec.

The iWiW graph sample contains 271,913 users and 2,712,587 friendship ties, while the Pokec is somewhat scarcer featuring 277,695 nodes and 2,122,778 edges. The iWiW dataset is accessible upon request from the Databank of the HUN-REN Centre of Economic and Regional Studies,[1] while the Pokec dataset is downloadable from the Stanford Large Network Dataset Collection.[2]

4.2 SRD

SRD is a statistical test method that compares solutions via a reference. The input of an SRD analysis is an $n \times m$ matrix where the first $m-1$ columns represent the methods we would like to compare, while the rows stand for the different input instances of the solutions. In our context, columns are centrality measures (influence proxies) and the rows are the samples we have taken from the node set (potential audiences). The last column of the matrix has a special role. It contains the benchmark values, called *references*, which form the basis of comparison. From the input matrix, we compose a *ranking matrix* by replacing each value in a column—in order of magnitude—by its rank. Ties are resolved by fractional ranking, *i.e.,* by giving each tied value the arithmetic average of their ranks. SRD values are obtained by computing the Manhattan distances (or ℓ_1-norm) between the column ranks and the reference ranking. In our paper, the reference values are given externally (they are the average spread of the sample sets in the simulation), but in some applications, the reference values are extracted from the first $m-1$ columns, this process is called *data fusion* (Willett, 2013). Depending on the type of data, this can be done by a number of ways (taking the average, median, or minimum/maximum). Since we have an external benchmark, we do not delve into the intricacies of data fusion. A more detailed explanation with a numerical example can be found in Sziklai and Héberger (2020).

SRD values are normalized by the maximum possible distance between two rankings of size n. In this way, SRD values corresponding to problems of different sizes can be compared. A small SRD score indicates that the solution is close to the reference (ranks the rows almost or entirely in the same order as the latter). The differences in SRD values induce a ranking between the solutions. SRD calculation is followed by two validation steps.

i. The permutation test (sometimes called randomization test, denoted by CRRN = comparison of ranks with random numbers) shows whether the rankings are comparable with a ranking taken at random. SRD values follow a discrete distribution that depends on the number of rows and the presence of ties. If n exceeds 13 and there are no ties, the distribution can be approximated with the normal distribution. By convention, we accept those solutions that are below 0.05, that is, below the 5% significance threshold. Between 5% and 95% solutions are not distinguishable from random ranking, while above 95% the solution seems to rank the objects in a reverse order (with 5% significance).

ii. The second validation option is called cross-validation and assigns uncertainties to the SRD values. Cross-validation enables us to statistically test whether two solutions are essentially the same or not, that is, whether the induced column rankings come from the same distribution. We assume that the rows are independent in the sense that removing some rows from the matrix does not change the values in the remaining cells (in our setting this is trivially true). Cross-validation proceeds by taking ℓ samples from the rows and computing the SRD values for each subset. Different cross-validation techniques (Wilcoxon, Dietterich, Alpaydin) sample the rows in different ways (Sziklai et al., 2022). Here we opted for 8-fold cross-validation coupled with the Wilcoxon matched pair signed rank test (henceforward Wilcoxon test). For small row sizes ($n \leq 7$) leave-one-out cross-validation

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table 2. SRD computation. nSRD stands for normalized SRD

	Input data					Ranking matrix		
	Solution 1	Solution 2	Ref.			Solution 1	Solution 2	Ref.
Input1	0.37	0.65	0.49	⟶	Input1	1	3	2
Input2	0.51	0.14	0.34		Input2	2	1	1
Input3	0.82	0.88	1	⟶	Input3	3	5	5
Input4	0.93	0.65	0.84		Input4	5	3	3.5
Input5	0.88	0.65	0.84	⟶	Input5	4	3	3.5
					SRD	6	2	
					nSRD	0.5	$0.1\dot{6}$	

is applied, each row is left out once. For larger row sizes (as in this paper), a reasonable choice is leaving out $\lceil n/\ell \rceil$ rows uniformly and randomly in each fold. The obtained SRD values are then compared with the Wilcoxon test with the usual significance level of 5% for each pair of solutions.

4.2.1 Example

Table 2 features a toy example. Suppose we would like to compare two solutions along five input instances for which we have reference values. The input table shows how the two solutions perform according to the different data. The ranking matrix converts these values to ranks. Note that there is a three-way tie for Solution 2. Since these are the 2nd, 3rd, and 4th largest values in that column, they all get the average rank of 3. Similarly, in the reference column the 3rd and 4th largest value coincide, thus they get an average rank of 3.5.

We compute the SRD scores by calculating the distances between the ranking vectors of the solution vector and the reference ranking vector in ℓ_1-norm. For the first solution, this is calculated as follows:

$$6 = |1-2| + |2-1| + |3-5| + |5-3.5| + |4-3.5|.$$

The maximum distance between two rankings that rank objects from 1 to 5 is 12. Thus, we divided the SRD scores by 12 to obtain the normalized SRD values. In the permutation test, we compare these values to the 5% thresholds of the discrete distribution that SRD follows. For $n = 5$, we accept those solutions (with 5% significance) which have a normalized score of 0.25 or less. This value comes from the discrete SRD distribution which depends on the number of rows and the presence of ties. Compared to the reference Solution 1 seems to rank the properties randomly, while Solution 2 passes the test.

4.3 Testing framework for the Audience Selection problem

In the Audience Selection problem, we are presented with several possible audiences (node sets of the underlying social network) and we need to predict that, upon activation, which are spreading influence better. Spreading potential can be proxied by network centrality. Our aim is to find out which centrality has better prediction power. Now we describe a step by step a comparison method that enables us to evaluate network centralities with statistical significance.

1. We determine the influence measures (centralities) for each node of the network.
2. We take n samples of size q from the node set.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

3. For each set, we calculate its average centrality according to each measure. Depending on the data, alternative statistical aggregates, *e.g.,* the sum (for unequal set sizes) or the median (in the presence of outliers), can be applied.
4. The activation function is used on the sets. In this paper, every user in the set is activated.
5. We run a Monte Carlo simulation multiple times with diffusion model D for each of the sample sets as seeds and observe their performance, that is, how many nodes do they manage to activate on average.
6. We compile the ranking matrix from the values obtained in step 3 and step 4. These correspond to the solution columns and the reference column, respectively.
7. We compute the SRD values and validate our results.

This framework enables us to test our hypothesis. The SRD analysis results in a ranking that tells us which centrality measures perform well in the Audience Selection problem. In parallel, we also run a diffusion simulation on the centralities' top choices, which gives us the ranking of the centralities in the Influence Maximization problem. Our hypothesis is that different measures prevail in one setting than the other.

Hypothesis 1.*The ranking of centralities in the Audience Selection problem significantly differs from a ranking obtained in the Influence Maximization problem.*

For validation, we test our hypothesis on two datasets, the iWiW and the Pokec networks. If the set size, q, is chosen to be too big, then the whole graph will be activated during the diffusion simulation, if it is set too small, then hardly anybody will be infected outside the original sets. Keeping this in mind, the hypothesis should be true for a wide range of set size values. Thus, we set $q = 500$ for iWiW, but opted for $q = 2000$ for the Pokec graph. These correspond to 0.183% and 0.720% of the total number of nodes, respectively. Note that iWiW and Pokec have approximately the same number of nodes, although Pokec is somewhat less dense.

Increasing the number of sets, n, is costly as we have to run simulations for each set, but it also helps to create a clear ranking over the proxy measures. Under a high n parameter, even small differences in performance will be statistically significant. Considering the distribution of SRD, we opted for $n = 24$. SRD already follows approximately normal distribution if $n \geq 13$ provided there are no ties in the rankings. As n becomes larger than 13, the distribution becomes sufficiently dense to detect small differences. In the presence of a few ties, the distribution is even denser, although it moves away slightly from normality. Thus, we use the empirical distribution, provided by the rSRD package, to determine p-values.

5. Diffusion models

A central observation of diffusion theory is that the line that separates a failed cascade from a successful one is very thin. Granovetter (1978) was the first to offer an explanation for this phenomenon. He assumed that each agent has a threshold value, and an agent becomes an adopter only if the amount of influence it receives exceeds the threshold. He uses an illuminating example of how a peaceful protest becomes a riot. A violent action triggers other agents who in turn also become agitated and a domino effect ensues. However, much depends on the threshold distribution. Despite the similarities, network diffusion can substantially differ depending on the type of contagion we deal with. A basic difference stems from the possible adopting behaviors. In most models, agents can adopt two or three distinct states: *susceptible*, *infected,* and *recovered*. Consequently, the literature distinguishes between SI, SIS, and SIR types of models.

In the conventional Influence Maximization setup, a seed set of users are fixed as adopters (or are infected), then the results of diffusion are observed. In the SI model, agents who become adopters stay so until the end of simulation (Carnes et al., 2007; Chen et al., 2009; Kim et al., 2013).

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

A real-life example of such a diffusion can be a profound innovation that eventually conquers the whole network (*e.g.,* mobile phones, internet). In SIS models, agents can change from susceptible to adopter and then back to susceptible again (Kitsak et al., 2010; Barbillon et al., 2020). For example, an innovation or rumor that can die out behaves this way. Another example would be of a service that agents can unsubscribe from and subscribe again if they want. In SIR models, agents can get "cured" and switch from adopter to recovered status (Yang et al., 2007; Gong and Kang, 2018; Bucur and Holme, 2020). Recovered agents enjoy temporary or lasting immunity to addictive misbehavior (*e.g.,* to online gaming). Other models allow users to self-activate themselves through spontaneous user adoption (Sun et al., 2020).

In this paper, we feature a classical SI diffusion model: the Linear Threshold (LT) framework. We represent our network G with a graph (V, E), where V is the set of nodes and E is the set of arcs (directed edges) between the nodes. In LT, each node, $\mathbf{v}$ chooses a threshold value $t_{\mathbf{v}}$ uniform at random from the interval $[0, 1]$. Similarly, each edge e is assigned a weight w_e from $[0, 1]$, such that, for each node the sum of weights of the entering arcs add up to 1, formally

$$\sum_{\{e\in E|e=(\mathbf{u},\mathbf{v}),\mathbf{u}\in V\}} w_e = 1. \qquad \text{for all } \mathbf{v} \in V.$$

One round of simulation for a sample set S_i took the following steps.

1. Generate random node thresholds and arc weights for each $\mathbf{v} \in V$ and $e \in E$.
2. Activate each node in V that belongs to S_i.
3. Mark each leaving arc of the active nodes.
4. For each $\mathbf{u} \in V \setminus S_i$ sum up the weights of the marked entering arcs. If the sum exceeds, the node's threshold activate the node.
5. If there were new activations go back to step 3.
6. Output the spread, *i.e.,* the percentage of active nodes.

We ran 5000 simulations and took the average of spread, that is, the expected percentage of individuals that the campaign will reach if S_i is chosen as initial spreaders.

In this setup, we can think of node thresholds as a measure of innovativeness or risk attitude of the users, while edge weights represent how much influence they bear on each other. Randomization of the parameters is useful in two ways. First, it is difficult to measure peer influence and individual thresholds. There are much more data available, since Kempe et al. (2003) introduced this diffusion model, so there is room for potential improvements in this aspect. For instance, a strong argument can be made that the parameters for innovators and early adopters should be chosen from different intervals, as they play a key role in the diffusion (Sziklai and Lengyel, 2022). Nevertheless, randomization is also essential for obtaining multiple observations needed for statistical inference. Perturbing parameter values is a standard technique to achieve this.

6. Influence Maximization proxies

In this section, we give a brief overview of the measures we employed in our analysis.

Among the classical centrality measures, we included **Degree** and **Harmonic Centrality**. The former is a self-explanatory benchmark, and the latter is a distance-based measure proposed by Marchiori and Latora (2000). Harmonic centrality of a node, $\mathbf{v}$ is the sum of the reciprocal of distances between $\mathbf{v}$ and every other node in the network. For disconnected node pairs, the distance is infinite; thus, the reciprocal is defined as zero. A peripheral node lies far away from most of the nodes. Thus, the reciprocal of the distances will be small which yields a small centrality value.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

PageRank, introduced by Page et al. (1999), is a spectral measure and a close relative of *Eigenvector centrality* (Bonacich, 1972). Eigenvector centrality may lead to misleading valuations if the underlying graph is not strongly connected. PageRank overcomes this difficulty by (i) connecting sink nodes (*i.e.,* nodes with no leaving arc) with every other node through a link and (ii) redistributing some value uniformly among the nodes. The latter is parameterized by the so-called damping factor, $\alpha \in (0, 1)$. The method is best described as a stochastic process. Suppose we start a random walk from an arbitrary node of the network. If anytime we hit a sink node, we restart the walk by choosing a node uniform at random from the node set. After each step, we have a $(1 - \alpha)$ probability to teleport to a random node. The probability that we occupy node **v** as the number of steps tends to infinity is the PageRank value of node **v**. The idea was to model random surfing on the World Wide Web. PageRank is a core element of Google's search engine, but the algorithm is used in a wide variety of applications. The damping value is most commonly chosen from the interval (0.7, 0.9), here we opted for $\alpha = 0.8$. For an axiomatic characterization of PageRank, see Was and Skibski (2023).

Generalized Degree Discount (GDD) introduced by Wang et al. (2016) is a suggested improvement on *Degree Discount* (Chen et al., 2009) which was developed specifically for the *independent cascade* model. The latter is an SI diffusion model where each active node has a single chance to infect its neighbors, transmission occurring with the probability specified by the arc weights. Discount Degree constructs a spreader group of size q starting from the empty set and adding nodes one by one using a simple heuristic. It primarily looks at the degree of the nodes but also takes into account how many of their neighbors are spreaders. GDD takes this idea one step further and also considers how many of the neighbors' neighbors are spreaders. The spreading parameter of the algorithm was chosen to be 0.05.

***k*-core**, sometimes referred to as, *k*-shell exposes the onion-like structure of the network (Seidman, 1983; Kitsak et al., 2010). First, it successively removes nodes with only one neighbor from the graph. These are assigned a *k*-core value of 1. Then it removes nodes with two or less neighbors and labels them with a *k*-core value of 2. The process is continued in the same manner until every node is classified. In this way, every node of a path or a star graph is assigned a *k*-core value of 1, while nodes of a cycle will have a *k*-core value of 2.

Linear Threshold Centrality (LTC) was, as the name suggests, designed for the Linear Threshold model (Riquelme et al., 2018). Given a network, G with node thresholds and arc weights, LTC of a node **v** represents the fraction of nodes that **v** and its neighbors would infect under the Linear Threshold model. In the simulation, we derived the node and arc weights that is needed for the LTC calculation from a simple heuristics: each arc's weight is defined as 1 and node thresholds were defined as 0.7 times the node degree. Note that we ran LTC and GDD under various parameters and chose the one which performed the best during the CRRN test (see Fig. A1 in the Appendix).

Suri and Narahari (2008) define a cooperative game on graphs and derive node centrality by computing the Shapley value. Nodes are considered players who cooperate to compose the best *k*-element set, for some given *k*. The Shapley value is the expected marginal contribution of each player when players are added to the set one by one and each order of the players is equally likely. Marginal contribution of a node **u** is just the difference between the value of the node set with and without **u**. Depending on how we define the value of a node set, we can derive different cooperative games. Here we use the G1 game variant proposed by Michalak et al. (2013) who also gave a polynomial-time algorithm to compute the corresponding **Shapley(G1)-value**. In G1, the characteristic function value of a node set C, is the number of nodes in C plus the number of distinct neighbors of C.

Top candidate (TC) algorithm originates from the field of social choice. It is a group identification method designed to find experts on recommendation networks (Sziklai, 2018, 2021). The algorithm takes a graph as an input and outputs a list of experts. The size of the output can be adjusted with a parameter, $\alpha \in [0, 1]$. The smallest set ($\alpha = 0$) corresponds to the elite, while

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

the largest set ($\alpha = 1$) is composed by anyone who could conceivably considered as expert. The algorithm works as follows. In the beginning, each node is considered as an expert and asked to nominate α fraction of its most popular neighbors. Popularity is measured by the number of recommendations the node received, that is, in the number of entering arcs. The nodes that receive no nominations are surely not experts; hence, they are removed from the expert set and their nominations are withdrawn. This in turn might leave some other nodes without nominations. The removal of experts continues until the set stabilizes. The α parameter is what makes TC suitable as a centrality index for strongly connected graphs. We run the algorithm multiple times, incrementing the α parameter by 0.01. Each node is assigned the largest $1 - \alpha$ value for which the node first becomes expert. This can be considered as a measure of exclusivity. For strongly connected graphs, every node becomes an expert at $\alpha = 1$ the latest, but those nodes will have an exclusivity value of 0. Note that social networks consisting of one component become strongly connected when the undirected link between the users (the "friendship") is converted into two opposing directed edges.

This is by no means a comprehensive list, there are other proxy and centrality measures, *e.g.,* HITS (Kleinberg, 1999), LeaderRank (Lü et al., 2011), DegreeDistance (Sheikhahmadi et al., 2015), IRIE (Jung et al., 2012), GroupPR (Liu et al., 2014), Attachment centrality (Skibski et al., 2019) for more see the survey of Li et al. (2018). Announcing a clear winner falls outside the scope of this manuscript. A real ranking analysis should always consider the diffusion model and a network characteristics carefully, as different algorithms will thrive in different environments.

7. Ranking analysis

The performance of the sample sets according to the various influence measures is displayed in Tables A1 (for iWiW) and A3 (for Pokec). The last column shows the percentage of nodes that the samples managed to activate on average in the diffusion simulation. The ranking matrix together with the SRD values is displayed in Tables A2 and A4. The tables can be found in the Appendix. The SRD score of a centrality is the distance between the ranking induced by the measure and the reference ranking, that is, the last column. In the CRRN test (Comparison of Ranks with Random Numbers), we compare the SRD scores with those of random rankings. Figures 2 and 3 show the result of the CRRN test. The SRD distribution is generated by observing the distances of random rankings from the reference. We accept a solution if the distance between the solution's ranking and the reference ranking falls left to the 5% significance thresholds.

For both networks, all of the methods pass the test, that is, they rank the sample sets more or less correctly. The performances vary, although degree and PageRank rank high, while *k*-core and Harmonic centrality rank low on both networks. The first places of degree should be taken with a grain of salt. A possible reason behind its exceptional performance might come from the fact that sample sets were created by random sampling. It remains to be seen how degree performs compared to other measures when sets are generated in a more realistic way, *e.g.,* based on location data.

Cross-validation reveals how the methods are grouped. Multiple SRD scores are generated by sampling the data and recalculating the rankings. We create 8-folds, by repeatedly removing three random rows (sample sets) and recalculating the rankings. Then, an 8-fold Wilcoxon test compares the median values. Figures 4 and 5 show the boxplots of the attained SRD values. On the iWiW network, there is no significant difference between Harmonic centrality, Shapley G1 and GDD. On the Pokec network, Linear Threshold Centrality and Shapley G1 are tied for the 4th place.

Our main goal was to show that Audience Selection and Influence Maximization are two different problems. So let us take a look at how the top choices of these measures perform in the simulation. In case of iWiW, we took the 500 highest-ranked agents for each measure. If the 500th and 501st agents tied with each other, we discarded agents with the same score one by

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

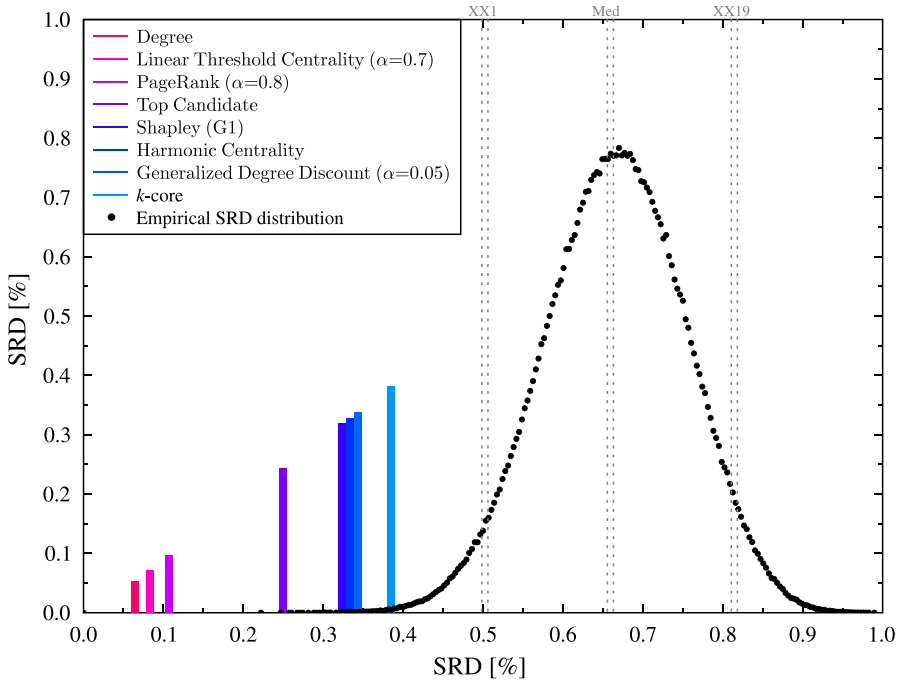

Figure 2. iWiW CRRN test. The order of the centrality measures in the legend (from top to bottom) follows the same order as the colored bars in the figure (from left to right). The bars' height is equal to their normalized SRD values. The black curve is a continuous approximation of the cumulative distribution function of the random SRD values. All (normalized) SRD values fall outside the 5% threshold (XX1: 5% threshold, Med: Median, XX19: 95% threshold).

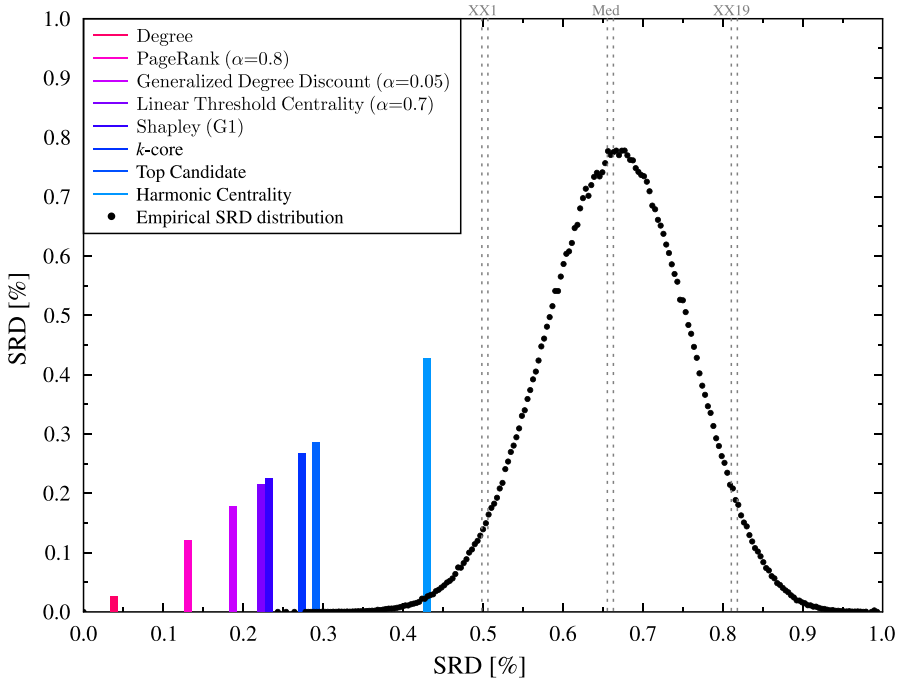

Figure 3. Pokec CRRN test. The order of the centrality measures in the legend (from top to bottom) follows the same order as the colored bars in the figure (from left to right). The bars' height is equal to their normalized SRD values. The black curve is a continuous approximation of the cumulative distribution function of the random SRD values. All (normalized) SRD values fall outside the 5% threshold (XX1: 5% threshold, Med: Median, XX19: 95% threshold).

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

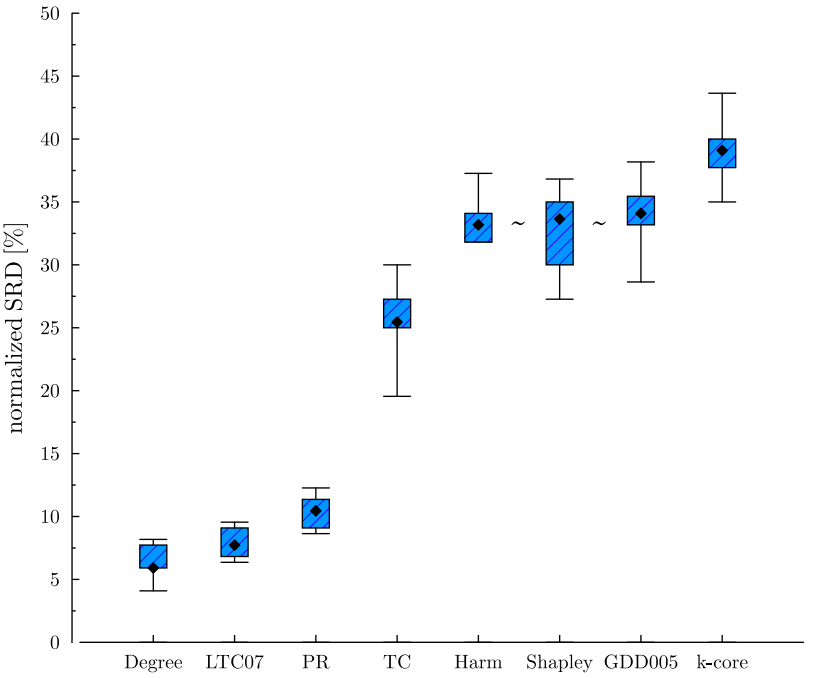

Figure 4. Cross-validation on iWiW data. The boxplot shows the median (black diamond), Q1/Q3 (blue box), and min/max values. The measures are ranked from left to right by the median values. The '∼' sign between two neighboring measures indicates that the Wilcoxon test found no significant difference between the rankings induced by the measures.

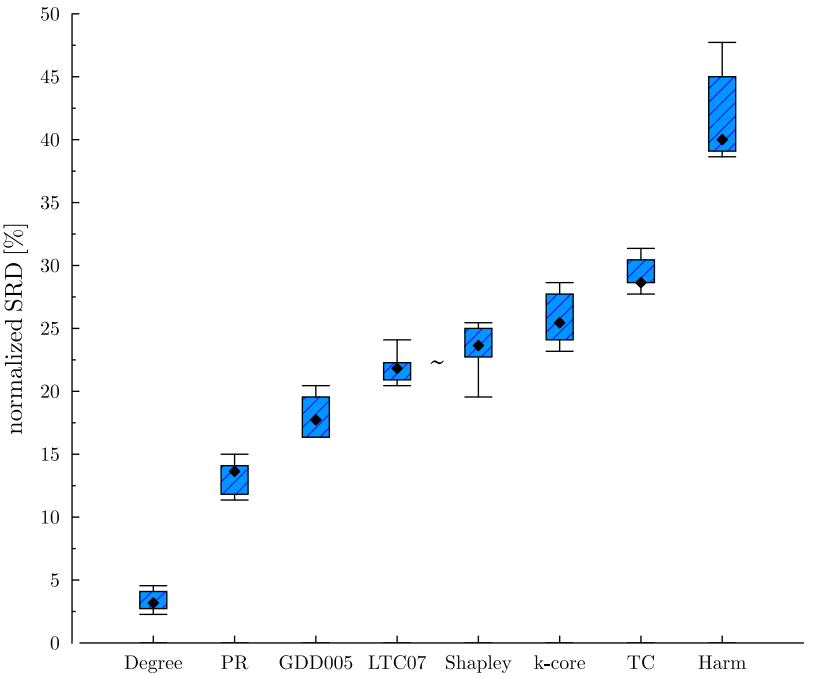

Figure 5. Cross-validation of Pokec data. The boxplot shows the median (black diamond), Q1/Q3 (blue box), and min/max values. The measures are ranked from left to right by the median values. The '∼' sign between two neighboring measures indicates that the Wilcoxon test found no significant difference between the rankings induced by the measures.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table 3. Comparing the rankings of centralities induced by the Audience Selection (AS) and Influence Maximization (IM) problem on iWiW. Ranks in brackets show the tied ranks according to the cross-validation

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC
Avg. spread (%)	21.124	16.243	21.013	21.178	4.559	20.970	20.207	17.770
Rank by IM	2	7	3	1	8	4	5	6
nSRD	0.066	0.333	0.108	0.344	0.385	0.083	0.330	0.250
Rank by AS	1	6 (6)	3	7 (6)	8	2	5 (6)	4
Rank difference	1	1	0	6	0	2	0	2
With ties	1	1	0	5	0	2	1	2

Table 4. Comparing the rankings of centralities induced by the Audience Selection (AS) and Influence Maximization (IM) problem on Pokec. Ranks in brackets show the tied ranks according to the cross-validation

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC
Avg. spread (%)	31.989	26.550	32.498	32.655	10.718	28.782	31.133	29.129
Rank by IM	3	7	2	1	8	6	4	5
nSRD	0.038	0.431	0.132	0.188	0.274	0.222	0.233	0.292
Rank by AS	1	8	2	3	6	4 (4.5)	5 (4.5)	7
Rank difference	2	1	0	2	2	2	1	2
With ties	2	1	0	2	2	1.5	0.5	2

one randomly until the size of the set became 500. We ran 5000 simulations to obtain the average spread of the measures. The same procedure was repeated on the Pokec data, but there the top 2000 nodes were considered. Tables 3 and 4 display the results.

In case of iWiW, there is a significant disparity in the performances of GDD under the Influence Maximization and Audience Selection problems. In Influence Maximization, GDD is the best-performing method, while in Audience Selection, it performs poorly compared to other methods. In case of Pokec, the rank difference between methods is not as massive as in iWiW. However, there are many smaller differences in the performance of methods.

Are these differences statistically significant? Luckily, the same CRRN test, that we used before can be applied, but now our input data is different. When we compared centralities, we looked at how they rank the sample sets. Now we want to compare Influence Maximization with Audience Selection, therefore we look at how they rank centralities. We can set either the ranking of IM or that of AS as the reference and compare their distance to the empirical SRD distribution of size $n = 8$ (since there are 8 centrality measures).

In case of iWiW, the distance between the IM and the AS ranking equals to a normalized SRD value of 0.375. In the SRD distribution, this corresponds to a p-value of 0.0685. If we consider ties the distance remains the same, however, the p-value becomes 0.07104 due to the change in the distribution. Assuming the usual 5% significance level, these two p-values mean that the AS ranking is indistinguishable from a random ranking compared to the IM ranking. In other words, the rankings significantly differ from each other and Hypothesis 1 holds.

In case of Pokec, the SRD values are 0.375 (in the strict ranking case) and 0.34375 (when ties are considered). The first value corresponds to a p-value of 0.0688, while the second to $p = 0.0380$. In other words, statistical significance depends on whether we allow ties or not. Nevertheless, an SRD score of 0.34375 is substantial even if it is not statistically significant. Considering that both

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Influence Maximization and Audience Selection are optimization problems, the difference is too big to treat them the same—especially, since they do not agree on the best method.

In conclusion, both our example and simulations on real data point toward that Hypothesis 1 holds: different methods perform well under the Influence Maximization model and the Audience Selection model.

8. Conclusion

We described a realistic decision situation that occurs in viral marketing campaigns. A company that organizes a series of public events wants to maximize social influence by selecting suitable locations/audiences. For this reason, the company must evaluate the collective impact of the selected audiences. As exact computation is infeasible, the potential is estimated by influence proxies (*e.g.,* network centralities). The question is which measure suits our needs the best? The Audience Selection problem resembles to the Influence Maximization at first glance. To prove that they are different, we proposed a testing framework that is capable to rank influence measures by their predictive power on sets. For this, we used a novel comparative statistical test, the Sum of Ranking Differences (SRD).

We demonstrated our results on two real-life social networks. We took samples from the node sets and created multiple rankings using various centrality measures. We compared these rankings to a reference ranking induced by the spreading potential of these sets under a Linear Threshold diffusion model. The algorithm whose ranking is the closest to the reference has the best predictive power. We compared this ranking to a ranking obtained by running a simulation with the top choices of these measures (in the manner of Influence Maximization).

In the iWiW network, the best-performing influence maximization algorithm fared relatively poorly on the Audience Selection test. On both networks, we found that the ranking of centralities in the Audience Selection problem substantially differed from a ranking obtained in the Influence Maximization problem. The discrepancy between the two rankings implies that we cannot blindly assume that the algorithm that finds the top spreaders in a network is the best in predicting the potential of an arbitrary set of agents.

The innovative use of SRD is one of the novelties of this paper. The test proved to be an excellent tool for comparing network centralities. Its use is not restricted to Influence Maximization, it can be applied in a wide variety of settings.

To examine the influence of ordinary users on the diffusion process, we utilized random sampling and compared Audience Selection and Influence Maximization on the randomly selected sample sets. All centrality measures pass the CRRN test, indicating that they order the sets significantly better than a random ranking. However, not every centrality is equally effective. PageRank and degree perform well on both networks. The effectiveness of degree may be attributed to the nature of the model. In a more realistic setting, sample sets should be based on locations, demographic groups, or other social traits. It remains to be seen how degree performs under a more realistic sampling technique.

Another simplifying assumption we had was to fix the activation function as the identity function, that is, every user in the seed set was activated. Although with random sampling this assumption imposes no bias, it becomes less realistic if the seed sets are composed in a more reasonable way (*e.g.,* based on location). A simplistic approach is to activate only a fix fraction of the targeted group. An even better way is to consider the adopter classification of users—innovators and early adopters in the seed set should become active with a higher probability.

Similarly to the activation function, different diffusion mechanisms need to be tested. Another popular diffusion model is independent cascade where each newly activated agent has one opportunity to infect its neighbors. We conjecture that Audience Selection and Influence Maximization rank centralities differently under other diffusion frameworks. Nevertheless, it would be interesting to examine the AS problem under the independent cascade model too.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Finally, we focused on comparing the Audience Selection to the Influence Maximization problem and also proposed a statistical framework to rank centralities in the former. A more comprehensive comparison study is needed to find the best method in the Audience Selection problem.

Competing interests. None.

Funding. The authors acknowledge financial help received from National Research, Development, and Innovation Office grant numbers KH 130502, K 138970, K 128573, K 138945. Balázs R. Sziklai is the grantee of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Data availability statement. The iWiW dataset is accessible upon request from the Databank of the Centre of Economic and Regional Studies,[1] while the Pokec dataset is downloadable from the Stanford Large Network Dataset Collection.[2]

Notes

1 https://adatbank.krtk.mta.hu/en/
2 https://snap.stanford.edu/data/#socnets

References

Andrić, F. L. (2018). Towards polypotent natural products: The Derringer desirability approach and nonparametric ranking for multicriteria evaluation of essential oils. *Journal of Chemometrics*, *32*, e3050.

Aral, S. (2011). Commentary–identifying social influence: A comment on opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*(2), 217–223.

Barbillon, P., Schwaller, L., Robin, S., Flachs, A., & Stone, G. D. (2020). Epidemiologic network inference. *Statistics and Computing*, *30*, 61–75.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, *2*(1), 113–120.

Bucur, D., & Holme, P. (2020, July). Beyond ranking nodes: Predicting epidemic outbreak sizes by network centralities. *PLOS Computational Biology*, *16*(7), 1–20.

Cai, J. L. Z., Yan, M., & Li, Y. (2016). Using crowdsourced data in location-based social networks to explore influence maximization. In *IEEE INFOCOM. 2016 - The 35th Annual IEEE International Conference on Computer Communications* (pp. 1–9).

Carnes, T., Nagarajan, C., Wild, S. M., & van Zuylen, A. (2007). Maximizing influence in a competitive social network: A follower's perspective. In *Proceedings of the Ninth International Conference on Electronic Commerce, ICEC '07* (pp. 351–360). New York: Association for Computing Machinery.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09* (pp. 199–208). New York: Association for Computing Machinery.

Eftekhar, M., Ganjali, Y., & Koudas, N. (2013). Information cascade at group scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13* (pp. 401–409). New York: Association for Computing Machinery,

Gensler, S., Völckner, F., Liu-Thompkins, Y., & Wiertz, C. (2013). Managing brands in the social media environment. *Journal of Interactive Marketing*, *27*(4), 242–256.

Gold, J., Pedrana, A. E., Sacks-Davis, R., Hellard, M. E., Chang, S., Howard, S., . . .Stoove, M. A. (2011). A systematic examination of the use of online social networking sites for sexual health promotion. *BMC Public Health*, *11*, 1–9.

Gong, K., & Kang, L. (2018). A new k-shell decomposition method for identifying influential spreaders of epidemics on community networks. *Journal of Systems Science and Information*, *6*(4), 366–375.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*, 489–515.

Hajdu, L., Bóta, A., Krész, M., Khani, A., & Gardner, L. M. (2020, March). Discovering the hidden community structure of public transportation networks. *Networks and Spatial Economics*, *20*(1), 209–231.

Héberger, K. (2010). Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry*, *29*(1), 101–109.

Héberger, K., & Kollár-Hunek, K. (2011). Sum of ranking differences for method discrimination and its validation: Comparison of ranks with random numbers. *Journal of Chemometrics*, *25*(4), 151–158.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Jung, K., Heo, W., & Chen, W. (2012). Irie: Scalable and robust influence maximization in social networks. In *2012 IEEE 12th International Conference on Data Mining* (pp. 918–923).

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03* (pp. 137–146). New York: Association for Computing Machinery.

Kim, C., Lee, S., Park, S., & Lee, S.-G. (2013). Influence maximization algorithm using markov clustering. In B. Hong, X. Meng, L. Chen, W. Winiwarter, & W. Song (Eds.), *Database Systems for Advanced Applications* (pp. 112–126). Berlin/Heidelberg: Springer.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, *6*, 888–893.

Kleinberg, J. M. (1999, September). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632.

Lengyel, B., Bokányi, E., Di Clemente, R., Kertész, J., & González, M. C. (2020). The role of geography in the complex diffusion of innovations. *Scientific Reports*, *10*, 15065.

Li, G., Chen, S., Feng, J., Tan, K.-L., & Li, W.-S. (2014). Efficient location-aware influence maximization. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14* (pp. 87–98). New York: Association for Computing Machinery.

Li, Y., Fan, J., Wang, Y., & Tan, K. (2018, October). Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *30*(10), 1852–1872.

Liu, Q., Xiang, B., Chen, E., Xiong, H., Tang, F., & Yu, J. X. (2014). Influence maximization over large-scale social networks: A bounded linear approach. In *Proc. of the 23rd ACM Int. Conf. on Information and Knowledge Management, CIKM '14* (pp. 171–180). New York: Association for Computing Machinery.

Lourenço, J. M., & Lebensztajn, L. (2018). Post-pareto optimality analysis with sum of ranking differences. *IEEE Transactions on Magnetics*, *54*(8), 1–10.

Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PLOS ONE*, *6*(6), 1–9.

Marchiori, M., & Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, *285*(3), 539–546.

Martin, C., & Niemeyer, P. (2019). Influence of measurement errors on networks: Estimating the robustness of centrality measures. *Network Science*, *7*(2), 180–195.

Michalak, T. P., Aadithya, K. V., Szczepański, P. L., Ravindran, B., & Jennings, N. R. (2013). Efficient computation of the shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research*, *46*, 607–650.

Moorthy, N. H. N., Kumar, S., & Poongavanam, V. (2017). Classification of carcinogenic and mutagenic properties using machine learning method. *Computational Toxicology*, *3*, 33–43.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Perra, N., & Fortunato, S. (2008). Spectral centrality measures in complex networks. *Physical Review E*, *78*, 036107.

Rahimkhani, K., Aleahmad, A., Rahgozar, M., & Moeini, A. (2015). A fast algorithm for finding most influential people based on the linear threshold model. *Expert Systems with Applications*, *42*(3), 1353–1361.

Riquelme, F., Gonzalez-Cantergiani, P., Molinero, X., & Serna, M. (2018). Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems*, *140*, 92–102.

Ristovski, J. T., Janković, N., Borčić, V., Jain, S., Bugarčić, Z., & Mikov, M. (2018). Evaluation of antimicrobial activity and retention behavior of newly synthesized vanilidene derivatives of Meldrum's acids using QSRR approach. *Journal of Pharmaceutical and Biomedical Analysis*, *155*, 42–49.

Saleem, M. A., Kumar, R., Calders, T., Xie, X., & Pedersen, T. B. (2017). Location influence in location-based social networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17* (pp. 621–630). New York: Association for Computing Machinery.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*(3), 269–287.

Sheikhahmadi, A., Nematbakhsh, M. A., & Shokrollahi, A. (2015). Improving detection of influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, *436*, 833–845.

Shi, Q., Wang, C., Chen, J., Feng, Y., & Chen, C. (2019). Location driven influence maximization: Online spread via offline deployment. *Knowledge-Based Systems*, *166*, 30–41.

Skibski, O., Rahwan, T., Michalak, T. P., & Yokoo, M. (2019). Attachment centrality: Measure for connectivity in networks. *Artificial Intelligence*, *274*, 151–179.

Song, C., Hsu, W., & Lee, M. L. (2016). Targeted influence maximization in social networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16* (pp. 1683–1692). New York: Association for Computing Machinery.

Staudacher, J., Sziklai, B. R., Olsson, L., Horn, D., Pothmann, A., Ali Tugay Sen, A. G., & Héberger, K. (2023). *rSRD: Sum of Ranking Differences Statistical Test*. CRAN. R Package Manual.

Sun, L., Chen, A., Yu, P. S., & Chen, W. (2020). Influence maximization with spontaneous user adoption. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20* (pp. 573–581). New York: Association for Computing Machinery.

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Suri, N. R., & Narahari, Y. (2008). Determining the top-k nodes in social networks using the shapley value. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3, AAMAS '08* (pp. 1509–1512). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
Sziklai, B. R. (2018). How to identify experts in a community? *International Journal of Game Theory*, *47*, 155–173.
Sziklai, B. R. (2021). Ranking institutions within a discipline: The steep mountain of academic excellence. *Journal of Informetrics*, *15*(2), 101133.
Sziklai, B. R., Baranyi, M., & Héberger, K. (2022). Testing rankings with cross-validation arXiv:2105.11939.
Sziklai, B. R., & Héberger, K. (2020). Apportionment and districting by sum of ranking differences. *PLOS ONE*, *15*(3), 1–20.
Sziklai, B. R., & Lengyel, B. (2022). Finding early adopters of innovation in social networks. *Social Network Analysis and Mining*, *13*(1), 4.
Wang, X., Lan, Y., & Xiao, J. (2019, July). Anomalous structure and dynamics in news diffusion among heterogeneous individuals. *Nature Human Behaviour*, *3*(7), 709–718.
Wang, X., Zhang, X., Zhao, C., & Yi, D. (2016, December). Maximizing the spread of influence via generalized degree discount. *PLOS ONE*, *11*(10), 1–16.
Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.
Weng, L., Menczer, F., & Ahn, Y.-Y. (2014). Predicting successful memes using network and community structure. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.
West, C. (2018). Statistics for analysts who hate statistics, Part VII: Sum of Ranking Differences (SRD)s. *LCGC North America*, *36*, 2–6.
Willett, P. (2013). Combination of similarity rankings using data fusion. *Journal of Chemical Information and Modeling*, *53*(1), 1–10.
Was, T., & Skibski, O. (2023). Axiomatic characterization of pagerank. *Artificial Intelligence*, *318*, 103900.
Yan, Q., Huang, H., Gao, Y., Lu, W., & He, Q. (2017). Group-level influence maximization with budget constraint. In *International Conference on Database Systems for Advanced Applications* (pp. 625–641). Cham: Springer.
Yang, R., Wang, B.-H., Ren, J., Bai, W.-J., Shi, Z.-W., Wang, W.-X., . . .Zhou, T. (2007). Epidemic spreading on heterogeneous networks with identical infectivity. *Physics Letters A*, *364*(3), 189–193.

Appendix A: Tables

Table A1. Average centrality values of the 24 randomly selected seed sets of iWiW. Avg. spread denotes the percentage of nodes the sample sets managed to infect on average over 5000 runs

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (%)
S1	19.686	0.2197	3.6226	32.924	10.152	22.522	0.9846	0.2853	3.270
S2	20.132	0.2210	3.6639	33.660	10.532	22.964	0.9784	0.2971	3.350
S3	19.500	0.2205	3.6198	32.976	10.248	22.362	0.9833	0.2887	3.270
S4	20.058	0.2197	3.6984	32.947	10.078	23.014	1.0170	0.2875	3.330
S5	19.664	0.2199	3.6625	33.037	10.226	22.570	1.0114	0.2879	3.290
S6	18.300	0.2193	3.4616	32.580	10.100	20.970	0.9470	0.2742	3.075
S7	20.972	0.2212	3.7904	34.108	10.606	23.932	1.0245	0.3011	3.462
S8	20.848	0.2215	3.8019	35.055	10.838	23.942	1.0265	0.3010	3.443
S9	18.948	0.2195	3.5252	32.467	10.036	21.768	0.9607	0.2821	3.162
S10	19.538	0.2210	3.6277	33.875	10.560	22.334	0.9836	0.2938	3.265
S11	19.486	0.2196	3.5936	33.431	10.334	22.254	0.9698	0.2908	3.239
S12	20.284	0.2213	3.7126	33.744	10.498	23.136	0.9973	0.2985	3.380
S13	20.166	0.2195	3.7371	33.069	10.204	23.148	1.0296	0.2962	3.350
S14	20.076	0.2193	3.7118	33.288	10.058	23.000	1.0267	0.2831	3.330
S15	20.268	0.2215	3.7231	33.826	10.500	23.184	1.0070	0.2990	3.367

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table A1. continued.

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (%)
S16	20.658	0.2217	3.8264	34.869	10.786	23.680	1.0436	0.3040	3.425
S17	20.764	0.2207	3.7817	33.879	10.446	23.678	1.0211	0.3045	3.455
S18	20.544	0.2215	3.7619	34.239	10.602	23.438	1.0093	0.2983	3.413
S19	19.662	0.2201	3.6355	33.450	10.382	22.472	0.9882	0.2971	3.281
S20	19.072	0.2189	3.5767	32.822	9.964	21.864	0.9807	0.2759	3.199
S21	19.874	0.2212	3.6521	34.361	10.672	22.836	0.9790	0.2936	3.290
S22	19.534	0.2204	3.6190	33.562	10.380	22.308	0.9852	0.2918	3.275
S23	20.382	0.2212	3.7190	34.284	10.534	23.276	0.9981	0.2951	3.380
S24	19.646	0.2207	3.6318	33.962	10.560	22.532	0.9814	0.2904	3.292

Table A2. iWiW ranking matrix. Data compiled from Table A1. Spreads closer than 0.005% to each other were considered as a tie. nSRD stands for normalized SRD

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (rn.)
S1	11	7	7	4	6	9	10	5	7
S2	15	15	13	13	16	13	4	17	15.5
S3	5	12	6	6	9	7	8	8	7
S4	13	8	14	5	4	15	18	6	13.5
S5	10	9	12	7	8	11	17	7	11
S6	1	2	1	2	5	1	1	1	1
S7	24	19	22	19	21	23	20	22	24
S8	23	21.5	23	24	24	24	21	21	22
S9	2	4	2	1	2	2	2	3	2
S10	7	16	8	16	18.5	6	9	13	5
S11	4	6	4	10	10	4	3	10	4
S12	18	20	16	14	14	16	13	19	18.5
S13	16	5	19	8	7	17	23	15	15.5
S14	14	3	15	9	3	14	22	4	13.5
S15	17	23	18	15	15	18	15	20	17
S16	21	24	24	23	23	22	24	23	21
S17	22	13	21	17	13	21	19	24	23
S18	20	21.5	20	20	20	20	16	18	20
S19	9	10	10	11	12	8	12	16	9
S20	3	1	3	3	1	3	6	2	3

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table A2. continued.

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (rn.)
S21	12	18	11	22	22	12	5	12	11
S22	6	11	5	12	11	5	11	11	7
S23	19	17	17	21	17	19	14	14	18.5
S24	8	14	9	18	18.5	10	7	9	11
SRD	19	96	31	99	111	24	95	72	0
nSRD	0.066	0.333	0.108	0.344	0.385	0.083	0.330	0.250	0.000

Table A3. Average centrality values of the 24 randomly selected seed sets of Pokec. Avg. spread denotes the percentage of nodes the sample sets managed to infect on average over 5000 runs

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (%)
S1	15.049	0.1920	3.5720	26.153	7.9635	14.855	0.9884	0.2649	7.170
S2	15.639	0.1919	3.6628	26.777	8.0790	15.260	1.0147	0.2674	7.398
S3	14.982	0.1921	3.5307	26.478	8.0980	14.759	0.9814	0.2622	7.133
S4	15.357	0.1922	3.6022	26.602	8.1625	14.837	1.0020	0.2666	7.288
S5	15.366	0.1924	3.6297	26.759	8.1615	15.364	1.0114	0.2683	7.342
S6	15.458	0.1922	3.6073	26.446	8.1155	15.063	0.9963	0.2664	7.345
S7	15.725	0.1921	3.6189	26.788	8.2080	15.163	0.9955	0.2693	7.421
S8	15.447	0.1922	3.6306	26.529	8.1140	15.402	1.0000	0.2732	7.334
S9	15.319	0.1917	3.6074	26.218	7.9845	14.930	1.0067	0.2682	7.265
S10	15.638	0.1926	3.6573	26.905	8.1930	15.424	1.0092	0.2709	7.409
S11	15.237	0.1917	3.5871	26.491	8.0540	14.723	0.9925	0.2632	7.269
S12	14.983	0.1917	3.5674	26.305	8.0355	14.775	0.9932	0.2655	7.157
S13	14.904	0.1919	3.5509	26.006	7.9805	14.634	0.9843	0.2653	7.106
S14	15.353	0.1923	3.6075	26.657	8.0835	15.132	1.0016	0.2622	7.296
S15	15.364	0.1923	3.5989	26.965	8.2150	15.051	0.9867	0.2679	7.309
S16	14.719	0.1918	3.4987	25.978	7.9945	14.254	0.9681	0.2636	7.033
S17	15.156	0.1923	3.5715	26.372	8.0805	14.960	0.9809	0.2680	7.206
S18	15.213	0.1923	3.6048	26.174	7.9650	15.039	0.9967	0.2642	7.241
S19	15.962	0.1919	3.7730	27.138	8.1415	15.360	1.0759	0.2695	7.544
S20	15.960	0.1926	3.7120	27.083	8.2355	15.491	1.0262	0.2787	7.527
S21	15.934	0.1926	3.7204	27.348	8.2590	15.279	1.0392	0.2726	7.523
S22	14.764	0.1916	3.5011	25.968	7.9370	14.535	0.9652	0.2590	7.071
S23	15.060	0.1916	3.5549	26.246	8.0035	14.565	0.9883	0.2658	7.170
S24	15.615	0.1925	3.6550	26.825	8.1960	15.352	1.0145	0.2701	7.398

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

Table A4. Pokec ranking matrix. Data compiled from Table A3. Spreads closer than 0.005% to each other were considered as a tie. rSRD stands for normalized SRD

	Degree	Harmonic	PageRank	GDD (0.05)	*k*-core	LTC (0.7)	Shapley (G1)	TC	Avg. spread (rn.)
S1	6	10	4	8	2	9	8	7	6.5
S2	20	7	17	21	10	17	21	13	18.5
S3	4	12	11	3	13	6	4	2	4
S4	13	15	14	11	18	8	16	12	12
S5	15	20	16	17	17	21	19	17	16.5
S6	17	14	10	13	15	14	12	11	16.5
S7	21	11	18	16	21	16	11	18	21
S8	16	13	13	18	14	22	14	23	15
S9	11	5	6	14	5	10	17	16	10.5
S10	19	24	20	20	19	23	18	21	20
S11	10	3	12	9	9	5	9	4	10.5
S12	5	4	8	6	8	7	10	9	5
S13	3	9	3	4	4	4	5	8	3
S14	12	17	15	15	12	15	15	3	13
S15	14	19	21	10	22	13	6	14	14
S16	1	6	2	1	6	1	2	5	1
S17	8	16	9	7	11	11	3	15	8
S18	9	18	5	12	3	12	13	6	9
S19	24	8	23	24	16	20	24	19	24
S20	23	22	22	22	23	24	22	24	22.5
S21	22	23	24	23	24	18	23	22	22.5
S22	2	2	1	2	1	2	1	1	2
S23	7	1	7	5	7	3	7	10	6.5
S24	18	21	19	19	20	19	20	20	18.5
SRD	11	124	38	54	79	64	67	84	
nSRD	0.038	0.431	0.132	0.188	0.274	0.222	0.233	0.292	

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23

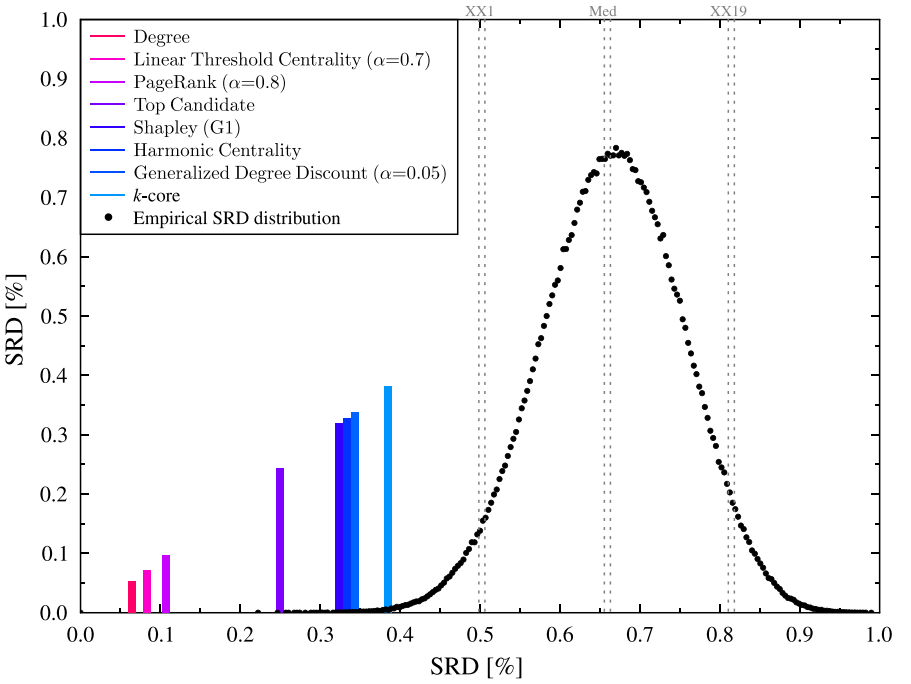

Figure A1. Parameter selection for Linear Threshold Centrality and Generalized Degree Discount. Smaller SRD values indicate better alignment with the reference, which was the spreading potential of the sample sets on iWiW (XX1: 5% threshold, Med: Median, XX19: 95% threshold).

Cite this article: Sziklai B. R. and Lengyel B. (2024). Audience selection for maximizing social influence. *Network Science* **12**, 65–87. https://doi.org/10.1017/nws.2023.23

Downloaded from https://www.cambridge.org/core. IP address: 18.218.245.163, on 03 May 2025 at 02:44:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/nws.2023.23