Influence without Authority: Maximizing Information Coverage in Hypergraphs

Peiyan Li* Honglian Wang† Kai Li‡ Christian Böhm§

Abstract

In many social networks, besides peer-to-peer communication, people share information via groups. An interesting problem arises in this scenario: for such networks, which are the best groups to start information diffusion so that the number of eventually informed nodes can be maximized? In this study, we formulate a novel information coverage maximization problem in the context of hypergraphs, wherein nodes are connected by arbitrary-size hyperedges (i.e., groups). In contrast to the existing literature on influence maximization, which aims to find authority nodes with high influence, we are interested in identifying the key groups. To address this problem, we present a new information diffusion model for hypergraphs, namely Hypergraph-Independent-Cascade (HIC). HIC generalizes the popular independent cascade model to hypergraphs to allow capturing group-level information diffusion. We prove the NP-hardness of the proposed problem under HIC, and the submodular monotone property of the information coverage function. Further, inspired by the Degree Discount algorithm, we derive a new heuristic method named Influence Discount (InfDis). Extensive experiments provide empirical evidence for the effectiveness and efficiency of our approach.

1 Introduction

Information diffusion in online social networks has been a critical research area these years, which serves as the foundation for various applications such as influence maximization [5], opinion dynamics [19], and social recommendation [23]. While most previous studies are conducted on ordinary networks, they consider information diffusion as pairwise interactions of nodes and ignore the existing group structure in many social networks. For example, Facebook groups and WhatsApp groups; and some platforms are foundationally based on groups, like Discord and Clubhouse.

The group structure in social networks has taken on increasing importance in spreading information. A prominent characteristic of such networks is that, although group members may not be direct friends, they share the same information exposure, as all members can receive the same messages sent to the group. This characteristic promotes the efficiency of information diffusion. Figure 1 shows a concrete example. When a person belongs to multiple groups, she/he may forward messages from one group to another group or her/his friends. An interesting problem is: **in order to maximize the spread of information in a social network with groups, e.g., to promote a new product or transmit a message so that more people can get aware of it, which are the best groups to start information diffusion?**

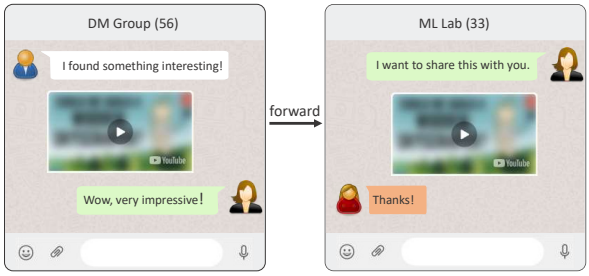

Figure 1: A motivating example of information diffusion between two groups.

In this study, we use hypergraphs to represent social networks with groups. We design a new information diffusion model for hypergraph networks. The objective is to select a fixed number of seed hyperedges to start information diffusion so that the number of eventually informed nodes is maximized. Our problem formulation differs from the existing literature on influence maximization (IM). The general IM studies aim to maximize the number of active nodes, while we aim to maximize the number of nodes covered by active hyperedges, which are the informed nodes. Next, we show that the computation of informed nodes can be simplified. We choose hyperedges as basic information diffusion units since all nodes in a hyperedge can get informed if a member passes information to that hyperedge, which gets activated simultaneously. Thus, the set of eventually informed nodes is a cover of all active hyperedges. Moreover, we select hyperedges (i.e., groups) for initial injection rather than a set of seed nodes to keep consistency.

Given the above points, we design HIC, a new infor-

*LMU Munich. lipeiyan@dbs.ifi.lmu.de
†KTH Royal Institute of Technology. honglian@kth.se
‡Huawei Noah's Ark Lab. likai210@huawei.com
§University of Vienna. christian.boehm@univie.ac.at

mation diffusion model for hypergraphs based on the Independent Cascade model (IC) [9]. Further, we formally define the problem of maximizing information coverage in hypergraphs under HIC. We prove the NP-hardness of this problem and the submodularity of the information coverage function, and we also show that computing the information coverage of a given seed set is #P-hard. To meet practical efficiency, we propose a new heuristic method named Influence-Discount (InfDis), which is inspired by the Degree Discount heuristics [5]. We assume that when propagation probability is small, the information coverage of a target hyperedge can be estimated within its immediate neighbors. Then we derive an iterative procedure of picking up a seed hyperedge by the estimated information coverage and re-estimating the information coverage of the seed's incident hyperedges. In summary, the contributions of this study are as follows:

- We formulate a new information coverage maximization problem in hypergraphs, which considers the group structure in social networks. In addition, we design a new information diffusion model for hypergraphs to model the spread of information among groups.
- We prove that the proposed problem is NP-hard, and computing the information coverage of a given seed set is #P-hard. We also show that the information coverage function is both monotone and submodular.
- We propose a new heuristic algorithm and test it on nine real-world hypergraph networks. The experimental results show that our method runs much faster than the naive greedy method (up to 10000 faster) and achieves comparable or even better results under most conditions.

2 Related Work

In this section, we first review influence maximization in graphs and hypergraphs. Then, we introduce the information coverage maximization problem, which is different from our setting. Further, we discuss potential solutions for information coverage maximization in hypergraphs.

2.1 Influence Maximization Influence maximization aims to find a seed set that will influence the maximum number of nodes. Most studies on influence maximization have been proposed based on two typical diffusion models, namely Independent Cascade (IC) [9] and Linear Threshold (LT) [10]. Under IC and LT, the influence maximization problem is NP-hard, and the property of monotone submodularity holds. Thus, some greedy algorithms [12, 15] are proposed to deal with this problem. The basic idea is to greedily select a seed that has the maximum increase of influence. A critical issue of these algorithms is the estimation of influence spread. The pioneer study [12] uses Markov Chain Monte Carlo (MCMC) simulations since the influence spread can be approximated arbitrarily close by increasing the number of MCMC simulations. In order to reduce the high computation costs of a large number of MCMC simulations, a few studies [13, 20] propose different strategies for MCMC simulations. For example, CELF [13] leverages the submodularity property of influence spread by skipping MCMC simulations of vertices that are known to be suboptimal; and CGA [20] tries to speedup single MCMC simulation by performing influence estimation in local communities.

It is worth noticing that MCMC-based algorithms are naturally limited in their speeds [2]. Trading space for speed is another choice. For example, the algorithms based on the reverse reachable sets (RR sets), including RIS [3], TIM [18] and IMM [17]. The construction of a RR set of a node v follows by (1) sampling the live edges; (2) preserving paths that start at v; and (3) the reverse reachable set is the set of nodes from which v is reachable. After sampling a sufficient number of RR sets, finding the seed set becomes a maximum cover problem.

The above two categories of algorithms have approximation guarantees. However, they generally suffer from high time complexity or heavy memory consumption. To meet practical efficiency, researchers propose another line of influence maximization algorithms, which estimate the influence of nodes by some heuristics. For example, the ranking-based algorithm [6] and the diffusion model reduction algorithm [5]. The ranking of nodes can be easily derived from the graph, such as degree, betweenness centrality [8], and PageRank score [16]. The diffusion model reduction algorithms target to restrict the influence spread distance to speed up diffusion estimation. The underground assumption is that the influence spread can be estimated within a small scope. For example, the Degree Discount heuristics [5] restricts the influence spread of a node within its immediate neighbors. Although these heuristic influence maximization algorithms can not ensure theoretical guarantees, they obtain substantial performance improvements compared to the other two categories of methods.

In recent years, information diffusion in hypergraphs has aroused increasing attention. However, to the best of our knowledge, only a few studies [7, 25, 1] focus on influence maximization in networks with groups,

i.e., hypergraphs. Eftekhar et al. [7] propose an elaborate group diffusion model, which considers budget allocation and inter-group/intro-group diffusion. Following this line, Zhu et al. [25] further consider the crowd influence in social networks, i.e., when multiple group members are activated, there should be an additional influence towards an inactive node in the same group beyond direct influence from the active members. Antelmi et al. [1] consider information diffusion in hypergraphs as an iterative process between nodes and hyperedges, which may capture non-binary relations.

2.2 Information Coverage Maximization in Graphs The influence maximization problem only considers the maximization of active nodes. Wang et al. [21] point out the difference between active nodes and informed nodes and then proposes the information coverage maximization problem. In an information diffusion process, active nodes are message senders which try to activate their neighbors, and informed nodes are the neighbors who receive information from active nodes but remain inactive. The information coverage maximization problem aims to maximize the number of active nodes and informed nodes, which proved to be NP-hard. Along this line, Wang et al. [22] further consider the connection strength between active nodes and informed nodes, and propose the activity maximization problem; Ni et al. [14] generalize the information coverage problem with multiple sources of information. We omit the discussions for technical solutions of information coverage maximization since they are usually generalized from traditional influence maximization algorithms.

It should be noted that our information coverage maximization problem is defined in hypergraphs. Different from [21], we define "informed nodes" as nodes that have received the information. Specifically, it is a cover of all active hyperedges.

2.3 Discussion In this paper, we consider the problem of information coverage maximization in hypergraphs, wherein the sizes of hyperedges are arbitrary. We argue that the MCMC-based and RR-sets-based methods are not proper solutions for this problem. The reasons are as follows. The heavy computation of MCMC simulations can be even worse since the information coverage is calculated based on the number of nodes covered by the union of eventually influenced hyperedges. As for the RR-sets-based methods, they can not be directly generalized to hypergraphs since each hyperedge may contain multiple nodes. If ignoring the size of hyperedges and transferring the hypergraph to a simple graph where nodes correspond to hyperedges, there should be a conflict because an optimal seed set of hyperedges covering the maximal number of RR sets does not necessarily cover the maximal number of nodes. Moreover, the resulting simple graph would be relatively dense, and the number of RR sets needed to be sampled could be enormous. Taken together, we propose a new heuristic algorithm to solve this problem.

3 Preliminaries

3.1 Notations We define hypergraph by an ordered pair $H = (V, E)$, where $V = \{v_1, v_2, \cdots, v_n\}$ is the set of n nodes, and $E = \{e_1, e_2, \cdots, e_m\} \subseteq 2^V$ is the set of m hyperedges. Each hyperedge is a set of nodes. When $|e_i| = 2$ for all $i \in [1, m]$, the hypergraph H reduces to an ordinary graph with only pairwise interactions.

The degree of a hyperedge is defined by the number of nodes belonging to it, i.e., $d_e = |e|$. For two hyperedges e_i and e_j, we denote the intersection of them by $e_i \cap e_j$. If $e_i \cap e_j \neq \emptyset$, e_i is e_j's incident hyperedge and vice versa. Let $E^* \subseteq E$ be a subset of hyperedges in H, and E^* contains multiple hyperedges. We use $|E^*|$ to denote the number of hyperedges in E^*, and use $\bigcup_{e \in E^*} e$ to represent the set of nodes covered by E^*. A summary of symbols used in this paper is given in the supplement.

3.2 Hypergraph Representation We introduce three types of hypergraph representation in Figure 2. Let us begin with a bipartite representation of a hypergraph as illustrated in Figure 2 (b). Hyperedges can be seen as an independent set, and a node connects with other nodes through hyperedges. In the line graph representation, as Figure 2 (c) shows, hyperedges are transferred to nodes in a simple graph, and two hyperedges are connected if they share some common nodes.

3.3 Formal Problem Definition This study aims to solve the information coverage maximization problem in the context of hypergraph networks. We first review the classic IC model [12]. In the IC model, the diffusion process triggers a cascade of activation in discrete steps. Specifically, an active node can only activate its inactive neighbors once with an independent probability for each edge, and then it stays active and stops activation. This process ends when there is no activation at a certain step. We build our diffusion model named HIC based on the IC model. In the following, we first give the following definitions for ease of understanding and then introduce the diffusion process of HIC.

Definition 3.1. (Active Hyperedge) *Initially, all hyperedges are inactive. An inactive hyperedge turns active if (1) it is selected as a seed hyperedge or (2) it gets activated by an active incident hyperedge.*

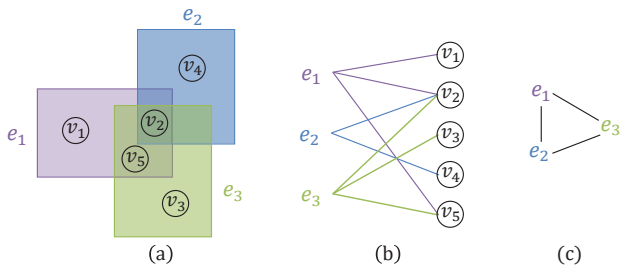

Figure 2: An illustration of hypergraphs. (a) A hypergraph with nodes $\{v_1, v_2, v_3, v_4, v_5\}$ and hyperedges $\{e_1, e_2, e_3\}$. (b) A bipartite graph representation with hyperedges on the left side and nodes on the right side. (c) A line graph expansion of a hypergraph is an ordinary graph where nodes represent hyperedges.

Definition 3.2. (Informed Node) *Initially, all nodes are uninformed. A node turns informed if it belongs to at least one activated hyperedge.*

Definition 3.3. (Activation) *An active hyperedge e_t can activate an inactive hyperedge e_s if (1) e_s is a incident hyperedge of e_t, and (2) at least one of their common nodes $u \in e_t \cap e_s$ "forwards information" from e_t to e_s.*

Let $E^0 = S$ be the seed hyperedges (i.e., E^0 is initially activated), and $E^\tau \subseteq E$ be the set of hyperedges that are activated in the τ-th step. If there is a node $v \in e_t \cap e_s$, where $e_t \in E^\tau$ and e_s is not activated, v can "forward information" from e_t to e_s with an independent probability p such that e_s is activated. If multiple active hyperedges intersect with an inactive hyperedge e_s, e.g., $\sum_{e_{t_j} \in E^\tau} |e_{t_j} \cap e_s| = l$ and $l < |e_s|$, then in the $(\tau+1)$-th step, e_s can be activated (i.e., $e_s \in E^{\tau+1}$) with probability $1-(1-p)^l$.

Since we consider hyperedges as the basic information diffusion units, the information cascades of HIC have the following three properties. (1) Each active hyperedge has only one chance to activate its incident hyperedges. (2) Each hyperedge can only be activated once and can not be deactivated. (3) The diffusion process ends when no hyperedges are activated at a certain step. Moreover, the information coverage is quantified by the number of informed nodes, which equals the number of unique nodes in all active hyperedges. Altogether, we formally define the information coverage maximization problem as follows: given a hypergraph H, a number k, and the information diffusion model HIC, how to select an initial hyperedge set S of cardinality k, to maximize the number of nodes $\sigma(S)$ eventually informed by the initial set S? Specifically,

$$\max_{S \subseteq E} \quad \sigma(S) \qquad s.t. \quad |S| = k. \tag{3.1}$$

3.4 Problem Analysis In this part, we analyze the properties of the proposed problem by giving the following theorems.

Theorem 3.1. *The information coverage maximization problem in hypergraphs under HIC is NP-hard.*

Proof. We use a reduction from the maximum coverage problem [11]. Consider the decision version of it: given a collection of sets $P = \{P_1, P_2, \cdots, P_m\}$ of a ground set $U = |\bigcup_{P_i \in P} P_i| = \{u_1, u_2, \cdots, u_n\}$, decide whether there exists a subset $P' \subseteq P$ of size k such that the number of covered elements $|\bigcup_{P_i \in P'} P_i|$ is at least K.

We now create a mapping from an instance of maximum coverage to an instance of our problem: build a hypergraph with m hyperedges $\{e_1, \cdots, e_m\}$ and n nodes $\{v_1, \cdots v_n\}$; hyperedge e_i corresponds to set P_i, and node v_j is on hyperedge e_i if $u_j \in P_i$; the activation probability p between any two hyperedges is 0. The mapping can be done in polynomial time with respect to m and n.

Observe that there exists a size k subset of P that covers at least K elements of U if and only if there exists a size k hyperedge seed set that eventually inform at least K nodes. Therefore, by solving the information maximization problem with parameter $p = 0$ and K, we can solve the maximum coverage problem. Thus we proved the information coverage maximization problem under HIC is NP-hard. ∎

Theorem 3.2. *Given a hypergraph H and an initial hyperedge seed set S, computing the information coverage $\sigma(S)$ under HIC is #P-hard.*

Proof. As proven in reference [4], it is #P-hard to compute the exact coverage given a seed set in directed graphs under the IC model. This can be easily extended to the #P-hardness in the case of undirected graphs by considering undirected edges as anti-parallel bi-directional edges. It is the same with hypergraphs under HIC since the group-wise information diffusion is a special case of pairwise interactions if we see hyperedges as nodes in an ordinary graph. ∎

Theorem 3.3. *For the information coverage maximization problem in hypergraphs under HIC, the information coverage function $\sigma(S)$ is both monotone and submodular.*

Proof. The monotonicity of $\sigma(S)$ is straightforward. To prove the submodularity, we borrow the concept of "live-arc graph" introduced in [12]. Although not being explicitly stated in [12], this concept also applies to undirected graphs. For an undirected hypergraph $H = (V, E)$, we construct a random live-arc graph H_L with nodes in H_L representing hyperedges. Let $Prob(H_L)$ be the probability that H_L is selected from all possible live-arc graphs, and $R_{H_L}(S)$ be the set of hyperedges that can be reached from the seed hyperedge set S in H_L, then the information coverage is the expected number of nodes covered by $R_{H_L}(S)$. Let $Q_{H_L}(S) = |\bigcup_{e \in R_{H_L}(S)} e|$, i.e., the number of nodes covered by $R_{H_L}(S)$, we have:

$$\sigma(S) = \sum_{all\ possible\ H_L} Prob(H_L) \cdot Q_{H_L}(S) \tag{3.2}$$

Note that the submodularity of an non-negative linear combination of submodular functions also holds, so we only need to prove the submodularity of $Q_{H_L}(S)$. We consider two sets of hyperedges M and N such that $M \subseteq N$. For an arbitrary hyperedge e, $Q_{H_L}(M \cup \{e\}) - Q_{H_L}(M)$ is the number of nodes covered by $R_{H_L}(e)$ that are not in $R_{H_L}(M)$, and $Q_{H_L}(N \cup \{e\}) - Q_{H_L}(N)$ denotes the number of nodes covered by $R_{H_L}(e)$ that are not in $R_{H_L}(N)$. Since $R_{H_L}(M) \subseteq R_{H_L}(N)$, we have $Q_{H_L}(M \cup \{e\}) - Q_{H_L}(M) \geq Q_{H_L}(N \cup \{e\}) - Q_{H_L}(N)$. Thus, the submodularity of $Q_{H_L}(S)$ holds and $\sigma(S)$ is both monotone and submodular. □

4 Proposed Approach

As discussed in the related work, while MCMC-based and RR-set-based influence maximization algorithms work well, they are not proper solutions to the problem of information coverage maximization in hypergraphs. In the following, we present InfDis, a new heuristic method inspired by the Degree Discount algorithm [5].

4.1 Influence Discount Heuristic In this part, we introduce the Influence Discount (InfDis) heuristic. The notion of "influence" represents the coverage of information in this study. Let u be an incident hyperedge of e, and e should be selected to the seed set in the current step. Then in the next step, the information coverage of u should be discounted due to the intersections between u and the new seed set. Following this idea, we derive an iterative procedure of picking up a seed hyperedge and re-estimating the information coverage of the seed's incident hyperedges. Next, we show how to calculate the information coverage of a target hyperedge e, i.e., the expected number of additional nodes informed by selecting e as seed.

Let S be the set of seed hyperedges selected; for an unselected hyperedge e, assuming there are t_e nodes in e that appear in S. We discuss the following two cases:

(a) The hyperedge e is activated by S, which has a probability of $1 - (1-p)^{t_e}$. Under such a case, selecting e to the seed set does not contribute additional information coverage.

(b) The hyperedge e is not activated by S, which has a probability of $(1-p)^{t_e}$. Under such a case, selecting e to the seed set S contributes additional information coverage.

To quantify the additional information coverage in case (b), we assume that the information coverage of selecting a hyperedge into the seed set can be estimated within the hyperedge's immediate neighbors when the propagation probabilities are small. Under this assumption, each hyperedge has a scope of information coverage, including itself and its immediate neighbors. Therefore, we can further divide the additional information coverage into two categories, i.e., direct/indirect information coverage. Here are the two definitions.

Definition 4.1. (Direct Information Coverage) *The direct information coverage $f(e)$ of selecting e into the seed set S is the number of new nodes covered by the new seed set:*

$$f(e) = d_e - t_e \tag{4.3}$$

Definition 4.2. (Indirect Information Coverage) *The indirect information coverage $g^*_{(e)}$ of selecting e into the seed set S is the expected number of new nodes covered by the hyperedges activated by e.*

To approximate $g^*_{(e)}$, we use the following theorem.

Theorem 4.1. *The indirect information coverage $g^*_{(e)}$ follows:*

$$g^*_{(e)} \leq \sum_{u \in Nei(e) \setminus S} [1 - (1-p)^{t_u}] \cdot (d_u - t_u) \tag{4.4}$$

Proof. Let $S^e = S \cup \{e\}$, and let the set of immediate neighbors of e be $Nei(e)$. Then for each hyperedge $u \in Nei(e) \setminus S$, the probability that u is activated by the seed set S^e is $1 - (1-p)^{t_u}$, where $t_u = |u \cap \bigcup_{e^* \in S^e} e^*|$ is the number of nodes of u that are already in S^e. Similarly, the activation of u contributes a number of $d_u - t_u$ newly informed nodes. Thus, we derive the inequality in Eq. 4.4, and the equality holds when there are no intersections between the immediate neighbors of e. □

For convenience, we denote the upper bound of $g^*_{(e)}$ by $g(e)$, which is the right part of inequality 4.4, and we

use $h(e)$ to denote the expected information coverage by selecting a hyperedge e as seed. Then $h(e)$ can be approximated by:

$$h(e)^* = (1-p)^{t_e} \cdot [f(e) + g(e)] \tag{4.5}$$

In the special case when $t_e = 0$, i.e., no neighbors of e is selected as seed, the expect information coverage $h(e) = f(e) + g(e)$.

The procedure of InfDis is as follows. Initially, when the seed set S is empty, we can get an approximated information coverage for each hyperedge. After selecting a new hyperedge e to S based on the ranking of information coverage, we should re-estimate the information coverage of incident hyperedges of e due to the intersections between e and its immediate neighbors. Then InfDis derives an iterative procedure by performing ranking-based seed selection and information coverage re-estimation in turn.

4.2 Time Complexity In the phase of finding seeds, the major computation load is the estimation of indirect information coverage. Assuming the max number of neighbors of a hyperedge is q_{max}, then there are at most $(m + kq_{max})$ estimations of indirect information coverage. There are two major parts in each estimation, including the union of at most k subsets and at most q_{max}-times intersection operations. Each intersection operation costs $O(kz)$, where z is the max hyperedge degree. Thus the time complexity of a single estimation is $O(kn + q_{max}kz)$. Moreover, the time complexity of InfDis is $O(kmn + nk^2q_{max} + mzkq_{max} + zk^2q_{max}^2)$.

5 Experiments

In this section, we present a series of experiments designed to evaluate the performance of InfDis through the following contexts: effectiveness under various numbers of seeds and different propagation probabilities, and time-efficiency of the proposed algorithm. Details regarding the datasets and comparison algorithms can be found in the supplement. All code, real-world datasets and the supplement are available at this URL[1].

5.1 Experimental Protocol There are five algorithms evaluated in our experiments. We need to set the independent propagation probability p, the number of seed hyperedges k, and the number of MCMC simulations R. Among the five algorithms, SimGreedy is the only one that outputs the seeds and their corresponding information coverage, which is estimated by a number of MCMC simulations. The other four algorithms only find seeds. For a fair comparison, we design the same evaluation phase for the four algorithms by using the same number of MCMC simulations. Due to the heavy computation load of MCMC simulations, we set R to 100 and limit the step of information propagation to 10 for the experiments' feasibility. The experimental results of information coverage are reported by mean and standard deviation.

We design two sets of experiments to evaluate the effectiveness of the proposed algorithm. (1) The first set of experiments adopts different numbers of seed hyperedges and a fixed independent propagation probability for every single node. For instance, we set the individual propagation probability p to 0.001 for Flickr, and set p to 0.01 for the other hypergraphs. The value of $p = 0.01$ is also used in the experiments of [12, 5, 24]. Although $p = 0.01$ and $p = 0.001$ may seem small, we observe that even with such small probabilities for information propagation, the number of eventually covered nodes could be large. Moreover, if increasing p to a larger value, the experimental results could be less discriminative because nearly all nodes would be covered. As for the number of seed hyperedges, we vary it between $\{10, 20, 30, 40, 50\}$. (2) To further analyze the sensitivity regarding p, we fix the number of seeds and modify the individual propagation probability. To be specific, the number of seeds k is set to 20; for Flickr, we vary p in $\{0.001, 0.003, 0.005, 0.007, 0.01\}$; and we vary p in $\{0.01, 0.03, 0.05, 0.07, 0.1\}$ for the other datasets.

5.2 Experimental Results In Figure 3, we summarize the estimation results of the information coverage with various numbers of seed hyperedges. InfDis achieves the best results in most cases, better than the simple greedy algorithm and the other three heuristics. Among the four comparative methods, SimGreedy generally ranks second and shows stable results. The other three heuristics' performances are unstable, and Betweenness shows the worst results in most cases since the hyperedge degree is completely ignored. The degrees of hyperedges have a substantial impact on the experimental performance. Selecting high-degree hyperedges as seeds achieves relatively better results when the max degree of hyperedges is large (e.g., Email-Eu, Email-W3C, and Stackoverflow). Specifically, Degree and HyperRank tend to assign higher scores for hyperedges with large degrees. When the degrees of hyperedges are generally small, Degree and HyperRank cannot select hyperedge seeds with a large scope of information coverage.

To further analyze the sensitivity, we conduct the second set of experiments by fixing the number of seeds and varying the propagation probabilities. The results are reported in Figure 4.

The first four hypergraphs are relatively small.

[1] https://github.com/KXDY233/InfDis

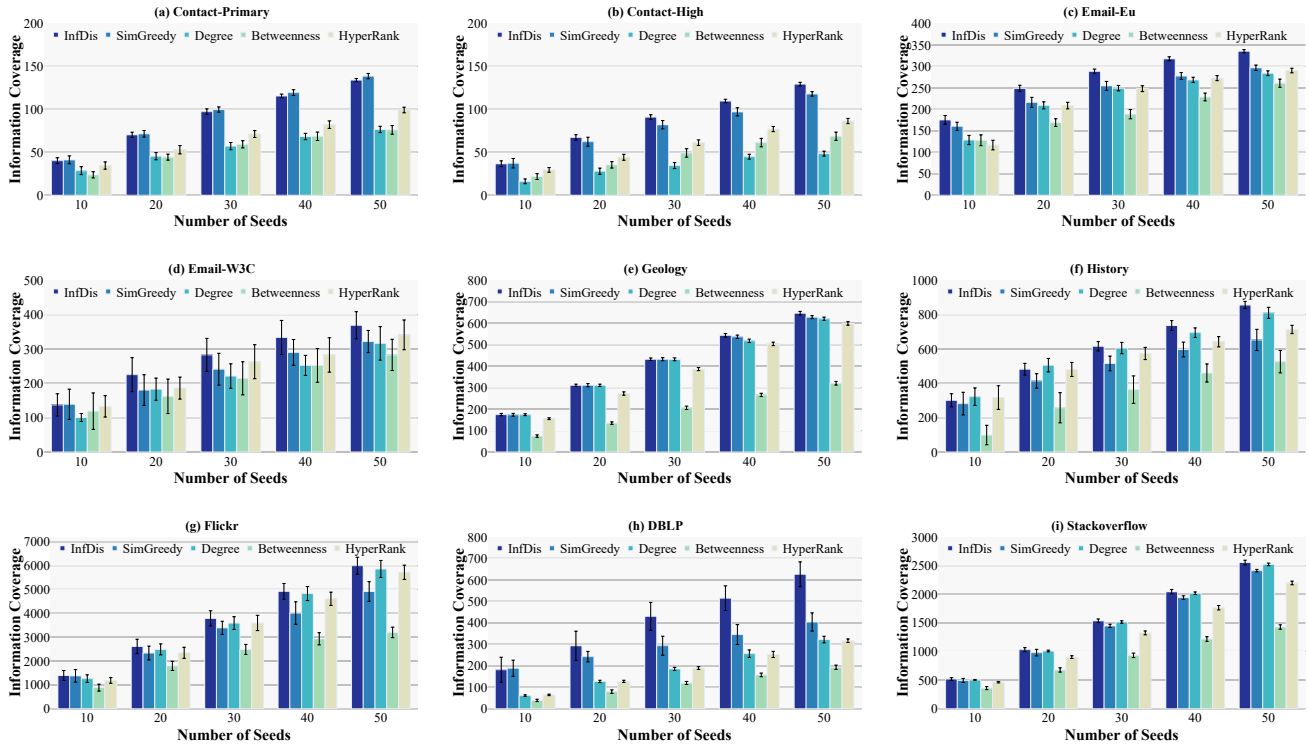

Figure 3: Information coverage of different algorithms on the nine hypergraph networks. The individual propagation probability p is set to 0.001 for Flickr, and 0.01 for all other datasets. The x-axis is the number of seed hyperedges, and the y-axis is the number of covered nodes of the eventually influenced hyperedges.

When p increases to $\{0.05, 0.07.0.1\}$, with only 20 seeds, information diffusion in 10 steps is able to cover most nodes in these hypergraphs. InfDis and SimGreedy can find key hyperedges that result in extensive information coverage under different propagation probabilities. The two algorithms show the best results in general. Moreover, in Geology and Stackoverflow, InfDis performs much better than SimGreedy. That is a supervising case since when increasing the propagation probability, the underlying assumption of InfDis tends to be violated. A potential reason is that, when increasing p, the information coverage of the incident hyperedges of the selected seed tends to degenerate faster, then InfDis is more likely to select "remote" hyperedges as seeds. This property may help InfDis find seeds from a large search space.

To sum up, InfDis shows its effectiveness and robustness through a wide range of experiments.

5.3 Runtime Comparison To evaluate the time efficiency, we compare the running time of different algorithms and show the results in Table 1. SimGreedy is the only method that outputs seeds and estimated information coverage simultaneously among the five algorithms. Thus, we also include the average evaluation time of the other four methods for a fair comparison. InfDis achieves the second-best running time on the nine datasets, which is only inferior to Degree, and it is almost 10000 times faster than SimGreedy. Interestingly, the runtime of InfDis for History and Stackoverflow are nearly the same, while the scale of the two hypergraphs varies widely. Recall that the main computation load of InfDis is the estimation of indirect influence, which ma-

Table 1: Running time (sec) of different algorithms for finding seeds, and average running time of MCMC evaluations. ($k = 20$; $p = 0.001$ for Flickr and $p = 0.01$ for all other datasets; $R = 100$; 10 steps of information cascades; and 16 threads for SimGreedy and Evaluation)

	Algorithms					
Dataset	InfDis	SimGreedy	Degree	Betweenness	HyperRank	Evaluation
Contact-Primary	0.043	254.224	0.007	0.344	0.062	4.043
Contact-High	0.029	261.760	0.007	0.328	0.047	4.543
Email-Eu	0.078	859.751	0.004	0.610	0.094	4.703
Email-W3C	0.068	2704.953	0.007	1.252	0.131	4.041
Geology	0.008	162.703	0.007	2.715	0.078	4.035
History	0.109	869.664	0.007	2.734	0.124	3.811
Flickr	6.052	21264.742	0.015	172.207	23.735	6.288
DBLP	0.336	64823.312	0.010	338.939	889.903	10.885
Stackoverflow	0.139	1322.316	0.015	90.129	935.826	7.463

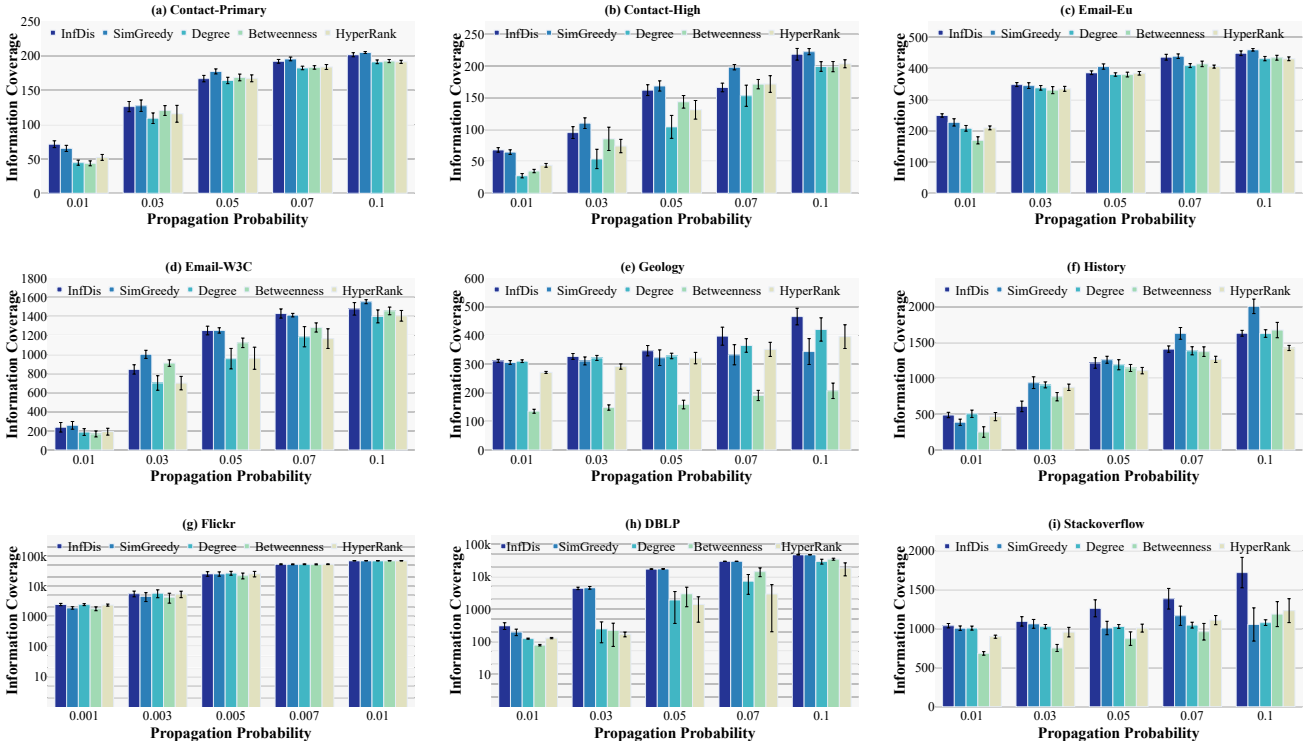

Figure 4: Information coverage of different algorithms on the nine hypergraph networks. The number of seed hyperedges is set to 20 for all datasets. The x-axis is the propagation probabilities, and the y-axis is the number of covered nodes of the eventually influenced hyperedges. Especially, results of DBLP and Flickr are reported in log-scale due to the large range of values.

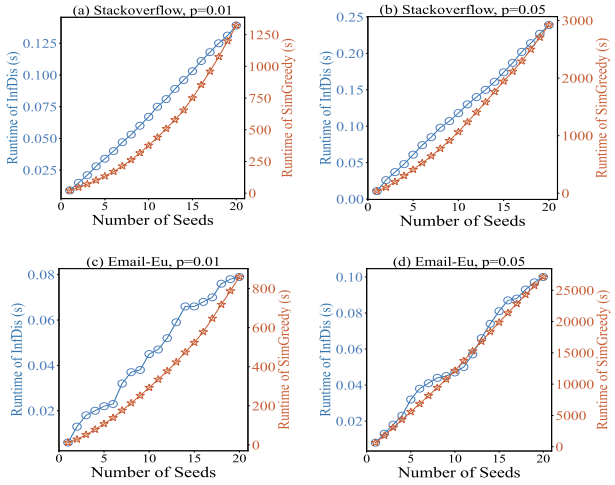

Figure 5: Runtime of InfDis and Greedy on Stackoverflow and Email-Eu.

jorly depends on the number of incident hyperedges associated with the selected seed. Thus, the running time of InfDis does not necessarily increase as the scale of the hypergraph increases. Furthermore, as shown in the time complexity analysis, besides m and n, the number of incident hyperedges also impacts InfDis's runtime. Especially, it is in quadratic form. To some extent, it may have a larger impact on the runtime. As a concrete example, Stackoverflow's q_{max} and q_{avg} are much smaller than Flickr's, and InfDis runs 43 times faster on Stackoverflow than on Flickr.

We further report the runtime of InfDis and SimGreedy on Stackoverflow and Email-Eu in Figure 5. It shows that InfDis's runtime nearly increases linearly with the number of seeds on Stackoverflow, while there is no apparent trend on Email-Eu. Recall that for InfDis, the runtime for finding the next seed depends on the number of re-estimations. This number is closely related to the local density of the currently selected seed, and it may vary for different seeds. In general, SimGreedy takes increasing time to find the next seed because it needs to simulate the information coverage of an increasing number of seeds.

6 Conclusion

In this work, we formulated a novel information coverage maximization problem in hypergraphs, which targets maximizing the number of eventually informed nodes by selecting a set of seed hyperedges for initializing information propagation. To address it, we proposed HIC, a new information diffusion model for hypergraphs. We formally defined our problem under HIC and analyzed its properties. Then we derived a new heuristic method called Influence Discount (InfDis). The experiments on the nine real-world hypergraph networks verified the effectiveness and efficiency of InfDis.

Currently, the diversity of individuals is not considered. This concept can be implemented by setting the propagation probability of individuals to be different. InfDis also applies to this setting, by slightly changing the calculation of expected information coverage. Our primary target in this work was to incorporate the group structure in studying information diffusion, thus we only gave the simple case with equal individual propagation probability. A potential future direction is to consider complex group effects, e.g., groups of different sizes may incur different costs, and users in these groups may have different probabilities of being informed.

Acknowledgments

This research was supported by CSC 201906070155 and EC H2020 RIA project SoBigData++ (871042). The authors would thank Sijing Tu for valuable discussions.

References

[1] A. Antelmi, G. Cordasco, C. Spagnuolo, and P. Szufel, *Social influence maximization in hypergraphs*, Entropy, 23 (2021), p. 796.

[2] A. Arora, S. Galhotra, and S. Ranu, *Debunking the myths of influence maximization: An in-depth benchmarking study*, in SIGMOD, 2017, pp. 651–666.

[3] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, *Maximizing social influence in nearly optimal time*, in SODA, 2014, pp. 946–957.

[4] W. Chen, C. Wang, and Y. Wang, *Scalable influence maximization for prevalent viral marketing in large-scale social networks*, in KDD, 2010, pp. 1029–1038.

[5] W. Chen, Y. Wang, and S. Yang, *Efficient influence maximization in social networks*, in KDD, 2009, pp. 199–208.

[6] S. Cheng, H. Shen, J. Huang, W. Chen, and X. Cheng, *Imrank: influence maximization via finding self-consistent ranking*, in SIGIR, 2014, pp. 475–484.

[7] M. Eftekhar, Y. Ganjali, and N. Koudas, *Information cascade at group scale*, in KDD, 2013, pp. 401–409.

[8] L. C. Freeman, *A set of measures of centrality based on betweenness*, Sociometry, (1977), pp. 35–41.

[9] J. Goldenberg, B. Libai, and E. Muller, *Talk of the network: A complex systems look at the underlying process of word-of-mouth*, Marketing letters, 12 (2001), pp. 211–223.

[10] M. Granovetter, *Threshold models of collective behavior*, American journal of sociology, 83 (1978), pp. 1420–1443.

[11] D. S. Hochbaum and A. Pathria, *Analysis of the greedy approach in problems of maximum k-coverage*, Naval Research Logistics, 45 (1998), pp. 615–627.

[12] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, in KDD, 2003, pp. 137–146.

[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, *Cost-effective outbreak detection in networks*, in KDD, 2007, pp. 420–429.

[14] Q. Ni, J. Guo, C. Huang, and W. Wu, *Information coverage maximization for multiple products in social networks*, Theoretical Computer Science, 828 (2020), pp. 32–41.

[15] N. Ohsaka, T. Akiba, Y. Yoshida, and K.-i. Kawarabayashi, *Fast and accurate influence maximization on large networks with pruned monte-carlo simulations*, in AAAI, 2014.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web.*, tech. rep., Stanford InfoLab, 1999.

[17] Y. Tang, Y. Shi, and X. Xiao, *Influence maximization in near-linear time: A martingale approach*, in SIGMOD, 2015, pp. 1539–1554.

[18] Y. Tang, X. Xiao, and Y. Shi, *Influence maximization: Near-optimal time complexity meets practical efficiency*, in SIGMOD, 2014, pp. 75–86.

[19] S. Tu and S. Neumann, *A viral marketing-based model for opinion dynamics in online social networks*, in WWW, 2022, pp. 1570–1578.

[20] Y. Wang, G. Cong, G. Song, and K. Xie, *Community-based greedy algorithm for mining top-k influential nodes in mobile social networks*, in KDD, 2010, pp. 1039–1048.

[21] Z. Wang, E. Chen, Q. Liu, Y. Yang, Y. Ge, and B. Chang, *Maximizing the coverage of information propagation in social networks*, in IJCAI, 2015.

[22] Z. Wang, Y. Yang, J. Pei, L. Chu, and E. Chen, *Activity maximization by effective information diffusion in social networks*, TKDE, 29 (2017), pp. 2374–2387.

[23] M. Ye, X. Liu, and W.-C. Lee, *Exploring social influence for recommendation: a generative model approach*, in SIGIR, 2012, pp. 671–680.

[24] C. Zhou, P. Zhang, W. Zang, and L. Guo, *On the upper bounds of spread for greedy algorithms in social network influence maximization*, TKDE, 27 (2015), pp. 2770–2783.

[25] J. Zhu, J. Zhu, S. Ghosh, W. Wu, and J. Yuan, *Social influence maximization in hypergraph in social networks*, TNSE, 6 (2018), pp. 801–811.

Supplement - Influence without Authority: Maximizing Information Coverage in Hypergraphs

Peiyan Li* Honglian Wang† Kai Li‡ Christian Böhm§

1 Symbols

The main symbols used in this paper and the supplement:

Symbol	Interpretation
$H = (V, E)$	Hypergraph
$V = \{v_1, \cdots, v_n\}$	Set of nodes
$E = \{e_1, \cdots, e_m\}$	Set of hyperedges
$d_e = \|e\|$	Number of nodes in e
$e_i \cap e_j$	Intersection of two hyperedges
$E^* \subseteq E$	Subset of hyperedges
$\|E\|$	Number of hyperedges
$\bigcup_{e \subseteq E} e$	The set of nodes covered by E
p	Individual propagation probability
S	Seed hyperedges
k	Number of hyperedges in S
E^τ	Hyperedges activated in the $\tau-$step
$\sigma(S)$	The final information coverage

2 The General Greedy Framework

We present a general greedy framework for our problem, which is also adopted by many influence maximization algorithms. As illustrated in Algorithm 1, it starts with an empty set S, and alternatively adds a hyperedge e into S such that the marginal gain of the information coverage function $\sigma(\cdot)$ is maximized. This process terminates when there are k hyperedges in S. This general greedy framework has its theoretical guarantee when the $\sigma(\cdot)$ function is non-negative monotone submodular, which holds under HIC as shown in Theorem 3.3. Therefore, Algorithm 1 solves the information coverage maximization problem with an approximation ratio of $1 - 1/e$.

However, this framework requires evaluating the information coverage of $O(km)$ hyperedges. Besides, for each hyperedge, SimGreedy should run a large number of MCMC simulations to approximate the accurate information coverage, which is infeasible for large hypergraphs. In this study, we introduce a new heuristic method to reduce the computational cost.

Algorithm 1: Simple Greedy

Input: a hypergraph $H = (V, E)$, the influence function $\sigma(\cdot)$, a number k
Output: a hyperedge seed set S

```
Initialize S = ∅
for i = 1, 2, ··· , k do
    Select e ⊆ E\S that maximizes σ(S ∪ {e}) − σ(S)
    Update S = S ∪ {e}
end
return S, σ(S)
```

3 Pseudocode and Acceleration Strategy of InfDis

Algorithm 2 is the pseudocode of InfDis. Similar to the idea of CELF [6] and CELF++ [5], we can accelerate InfDis by skipping estimations of a part of hyperedges that are known to have suboptimal information coverage. For instance, the first round of estimation (line 4-8 in Algorithm 2) is calculated when the seed set is empty, and the approximated information coverage of each hyperedge achieves the best value of itself. When the seed set is nonempty, the information coverage of an unselected hyperedge will decrease due to the potential intersections between this hyperedge and the seed set. Thus, in the step of reestimation (line 13-18 in Algorithm 2), we can skip a part of neighbors with small information coverage.

4 Evaluation Setup

4.1 Datasets We collected and processed nine datasets from two online data collections, SNAP[1], and ARB[2]. Contact-Primary and Contact-High are interactions between students at a primary school and a high school, respectively. Email-Eu and Email-W3C are email records of two different organizations, where each node represents an email address, and each hyperedge is a tuple of sender and receivers (the directed links between sender

*LMU Munich. lipeiyan@dbs.ifi.lmu.de
†KTH Royal Institute of Technology. honglian@kth.se
‡Huawei Noah's Ark Lab. likai210@huawei.com
§University of Vienna. christian.boehm@univie.ac.at

[1] http://snap.stanford.edu/data/
[2] https://www.cs.cornell.edu/~arb/data/

Algorithm 2: Influence Discount

Input: A hypergraph $H = (V, E)$, the number of seed hyperedges k, and the propagation probability of each node p
Output: the seed set of hyperedges S

function MAIN(H,p,k)
  Initialize $S = \emptyset$
  Compute the hyperedge degree $d_e = |e|$ for each $e \subseteq E$
  for *each hyperedge* $e \subseteq E \backslash S$ **do**
    Initialize $t_e = 0$ and $f(e) = d_e$
    $g(e) =$ INDIRECTINFLUENCE(S, e)
    $h(e) = f(e) + g(e)$
  end
  while $|S| < k$ **do**
    Select $e \subseteq E \backslash S$ that maximizes $h(e)$
    $S = S \cup \{e\}$
    $Q = \bigcup_{e^* \in S} e^*$
    for *each hyperedge* $w \subseteq Nei(e) \backslash S$ **do**
      $t_w = |w \cap Q|$
      $f(w) = d_w - t_w$
      Get $g(w) =$ INDIRECTINFLUENCE(S, w)
      $h(w) = (1-p)^{t_w} \cdot [f(e) + g(e)]$
    end
  end
  return S
end function

function INDIRECTINFLUENCE(S, e)
  Initialize $g(e) = 0$ and $Q = \bigcup_{e^* \in S} e^*$
  for *each hyperedge* $u \subseteq Nei(e) \backslash S$ **do**
    $t_u = |u \cap Q|$
    $g(e) = g(e) + [1 - (1-p)^{t_u}] \cdot (d_u - t_u)$
  end
  return $g(e)$
end function

and receivers are omitted in the two datasets). Geology, History, and DBLP are co-author hypergraph networks, where each hyperedge denotes co-authors of a paper and nodes represent authors. Flickr is collected from an image-sharing website, where each node represents a user, and each hyperedge corresponds to a user-defined group. In the Stackoverflow dataset, each node denotes a user, and each hyperedge contains users participating in the same thread. We removed isolated nodes and only kept the largest connected component for all the nine hypergraph networks. Further statistics of these datasets are listed in Table 1.

4.2 Baselines Since the proposed problem is new, there are no existing algorithms targeting this problem. To some extent, selecting the best seed set is to find the most influential hyperedges. Thus, we incorporate ranking-based heuristics as comparative algorithms, namely Degree, Betweenness, and HyperRank [2]. These three ranking methods quantify the importance of hyperedges from different perspectives. Moreover, we use the simple greedy algorithm given in Algorithm 1 as the baseline approach. We do not run other MCMC-based or RR-sets-based influence maximization algorithms on the bipartite graph expansion and the line graph expansion of hypergraphs. The reasons are as follows. (1) Although some methods [1, 7] for hypergraph influence maximization have been proposed, they use the bipartite graph expansion, and they target to find influential nodes, not key hyperedges. Moreover, they can not fit in HIC since their diffusion models are specially designed for the bipartite graph. (2) SimGreedy is an representative MCMC-based algorithm for our problem. (3) In the line graph expansion, the hyperedge degree is completely ignored, and the resulting ordinary graphs are too dense to proceed for representative RR-sets-based algorithms [3, 9, 10].

- **Simple Greedy (SimGreedy).** The simple greedy method is generally adopted for solving the monotone submodular maximization problem.
- **Degree Heuristic (Degree).** Ranking hyperedges based on the degrees of hyperedges.
- **Betweenness Heuristic (Betweenness) [4].** Betweenness is a measure of centrality based on shortest paths. We calculate the betweenness of hyperedges in the bipartite graph representation of hypergraphs (as illustrated in Figure 2 (b)).
- **Hyperedge-based Ranking (HyperRank) [2].** This

Table 1: Statistics of datasets. n and m are the number of nodes and hyperedges, respectively; d_{max} is the max hyperedge degree while d_{avg} is the average; q_{max} is the max number of neighbors (i.e., incident hyperedges) of a hyperedge, while q_{avg} is the average.

Data set	n	m	d_{max}	d_{avg}	q_{max}	q_{avg}
Contact-Primary	209	976	4	2.194	84	31.311
Contact-High	258	867	4	2.261	83	25.821
Email-Eu	490	1565	25	2.445	177	42.118
Email-W3C	1840	1841	23	2.19	236	58.905
Geology	5125	1604	21	4.647	23	4.772
History	5531	2195	24	4.300	151	12.523
Flickr	96497	9171	100	22.747	652	77.258
DBLP	141886	102502	6	3.227	202	13.441
Stackoverflow	702589	69922	50	10.336	24	1.626

ranking method is a variation of PageRank [8], where hyperedges with large degrees tend to be ranked higher.

We implemented the proposed algorithm and the four comparative algorithms in Python and ran the experiments on a Windows Laptop with AMD Ryzen 7 5800H 16 Core 3.2GHZ CPU and 32GB memory.

References

[1] A. Antelmi, G. Cordasco, C. Spagnuolo, and P. Szufel, *Social influence maximization in hypergraphs*, Entropy, 23 (2021), p. 796.

[2] A. Bellaachia and M. Al-Dhelaan, *Multi-document hyperedge-based ranking for text summarization*, in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 1919–1922.

[3] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, *Maximizing social influence in nearly optimal time*, in SODA, 2014, pp. 946–957.

[4] U. Brandes, *A faster algorithm for betweenness centrality*, Journal of mathematical sociology, 25 (2001), pp. 163–177.

[5] A. Goyal, W. Lu, and L. V. Lakshmanan, *Celf++ optimizing the greedy algorithm for influence maximization in social networks*, in Proceedings of the 20th international conference companion on World wide web, 2011, pp. 47–48.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, *Cost-effective outbreak detection in networks*, in KDD, 2007, pp. 420–429.

[7] A. MA and A. Rajkumar, *Hyper-imrank: Ranking-based influence maximization for hypergraphs*, in CODS-COMAD, 2022, pp. 100–104.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web.*, tech. rep., Stanford InfoLab, 1999.

[9] Y. Tang, Y. Shi, and X. Xiao, *Influence maximization in near-linear time: A martingale approach*, in SIGMOD, 2015, pp. 1539–1554.

[10] Y. Tang, X. Xiao, and Y. Shi, *Influence maximization: Near-optimal time complexity meets practical efficiency*, in SIGMOD, 2014, pp. 75–86.