Influence without Authority: Maximizing Information Coverage in Hypergraphs

Peiyan Li[1], Honglian Wang[2], Kai Li[3], and Christian Böhm[4]

[1]Ludwig-Maximilians-Universität München,
[2]KTH Royal Institute of Technology,
[3]Huawei Noah's Ark Lab,
[4]University of Vienna.

lipeiyan@dbs.ifi.lmu.de

April 27, 2023

Backgroud

Social network with groups can be modeled as hypergraph.

Figure: Social networks with groups

1. https://web.facebook.com/groups;
2. https://www.whatsapp.com/;
3. https://discord.com/

Problem Formation

What are hypergraphs?

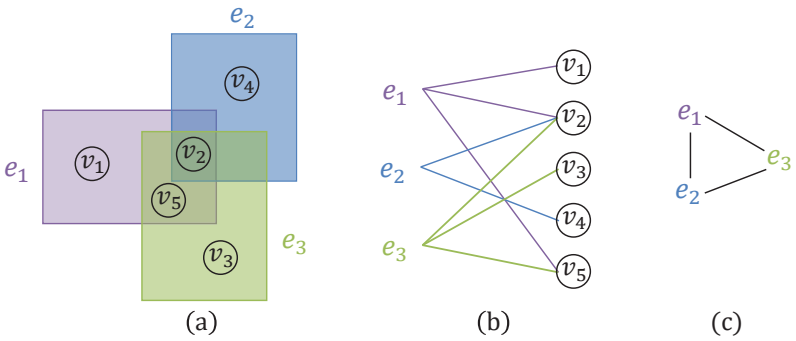


Figure: An illustration of hypergraphs. (a) A hypergraph with nodes $\{v_1, v_2, v_3, v_4, v_5\}$ and hyperedges $\{e_1, e_2, e_3\}$. (b) A bipartite graph representation with hyperedges on the left side and nodes on the right side. (c) A line graph expansion of a hypergraph is an ordinary graph where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

Problem Formation

1. How information diffuses in online social networks with groups (i.e., hypergraphs)?
2. In order to **maximize the spread of information** in a social network with groups, which are the best groups to start information diffusion?

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

Problem Formation

Modeling information diffusion in **hypergraph networks** based on a generalized independent cascade model.

Recall the classical Independent Cascade (IC) model:

1. Initialization with seed nodes.
2. When node u becomes active, it has one chance to activate each inactive neighbor v with probability p_{uv}.
3. Active nodes never deactivate.
4. The diffusion process ends when no nodes are activated at a certain step.

Problem Formation

The Hypergraph Independent Cascade model (HIC).
Assumption: all members in a group share the same information exposure.

Definition (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.

Definition (Activation)

An active hyperedge e_t has one change to 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.

Definition (Informed Node)

Initially, all nodes are uninformed. A node turns informed if it belongs to at least one activated hyperedge.

Problem Description

In order to **maximize the spread of information** in a social network with groups, which are the best groups to start information diffusion?

Definition (Information Coverage Maximization in Hypergraphs)

Find a small subset of groups (hyperedges) in a social network that could maximize the spread of information (i.e., the number of informed nodes).

Problem Analysis

Theorem

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

Theorem

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

Theorem

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

Detailed proofs please refer to our paper.

Recall the Simple Greedy Solution

Algorithm 1: Simple Greedy

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

Output: a hyperedge seed set S

1 Initialize $S = \emptyset$
2 **for** $i = 1, 2, \cdots, k$ **do**
3 Select $e \subseteq E \backslash S$ that maximizes $\sigma(S \cup \{e\}) - \sigma(S)$
4 Update $S = S \cup \{e\}$
5 **end**
6 **return** S, $\sigma(S)$

Compute the influence spread (information coverage) $\sigma(\cdot)$ is very time-consuming!

Our Solution

Algorithm 2 Our algorithm

Input: A hypergraph $H = (V, E)$, a budget k.
Output: A hyperedge seed set S.
$S \leftarrow \emptyset$.
Estimate influence coverage $h(e)$ for all $e \in E$.
$e = \arg\max_{e \in E} h(e)$.
$S \leftarrow \{e\}$, $E \leftarrow E \setminus \{e\}$.
while $|S| \leq k$ **do**
    for $u \in E$ **do**
        if $u \in \mathrm{Nei}(S)$ **then**
            Update $h(u)$
    $e = \arg\max_{e \in E} h(e)$.
    $S \leftarrow S \cup \{e\}$, $E \leftarrow E \setminus \{e\}$.

▷ Re-estimate $\sigma(e)$ for all $e \in E$.

We proposed a method to estimate the influence coverage $h(e)$ efficiently!

Our Solution: Influence Discount Heuristic

How to estimate the information coverage of a target hyperedge e?
Assumption: when propagation probability is small, the information coverage of a target hyperedge can be estimated within **its immediate neighbors**.

t_e – number of nodes in e that have appear in S (t_e can be 0)
p – the independent propagation probability

We discuss the following two cases:

- 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. $P[\sigma(S) = \sigma(S \cup e)] = 1-(1-p)^{t_e}$
- 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. $P[\sigma(S) \neq \sigma(S \cup e)] = (1-p)^{t_e}$

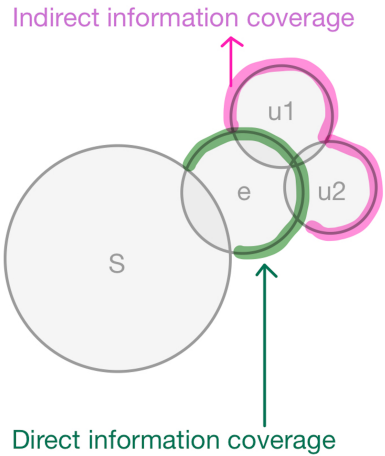
Our Solution: Influence Discount Heuristic

Compute the information coverage gain of selecting e into S:

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

$f(e)$: the direct information coverage
$g(e)$: the indirect information coverage

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



Our Solution: Influence Discount Heuristic

How to calculate the information coverage of a target hyperedge e?

t_e – number of nodes in e that have appear in S (t_e can be 0)
d_e – hyperedge degree, number of nodes in e

When the hyperedge e is not activated by S, there are two kinds of additional information coverage.

Definition (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:

$$f(e) = d_e - t_e \tag{2}$$

Our Solution: Influence Discount Heuristic

Approximate the Indirect Information Coverage:

Definition (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.

Theorem

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

$$g^*_{(e)} \leq \sum_{u \in Nei(e) \setminus S} [1 - (1-p)^{t_u}] \cdot (d_u - t_u) \tag{3}$$

Experiments

All code, real-world datasets and the supplement are available at this URL*.

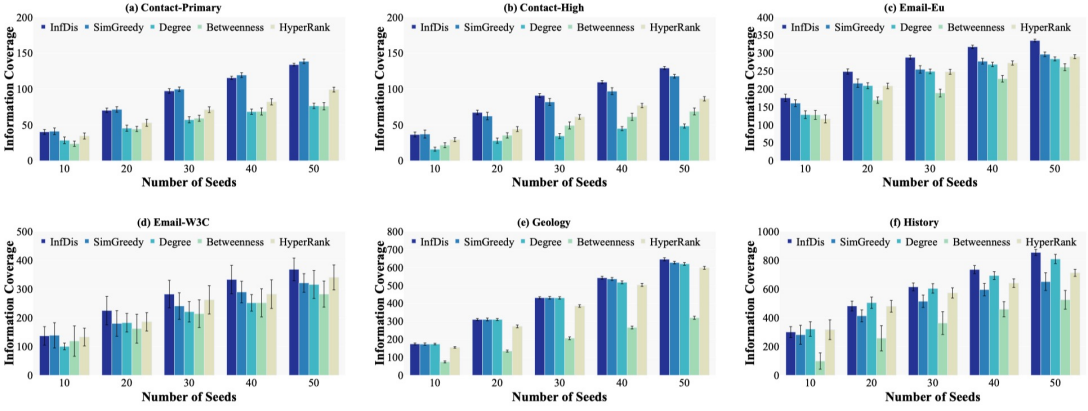
Table: Statistics of datasets. n and m are the number of nodes and hyperedges, respectively; d_{max} is the max hyperedge degree while d_{avg} is the average; q_{max} is the max number of neighbors (i.e., incident hyperedges) of a hyperedge, while q_{avg} is the average.

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

*https://github.com/KXDY233/InfDis

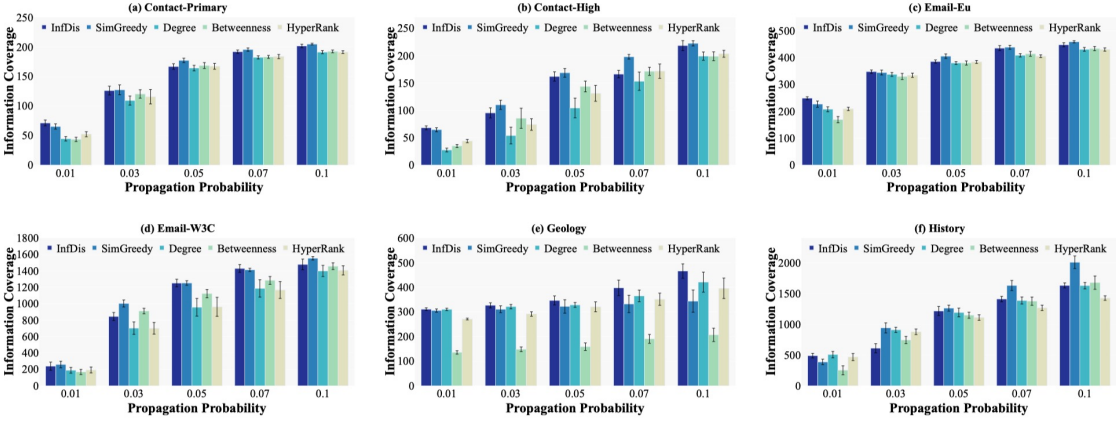
Experiments

Setup 1: fix the individual propagation probability p, vary the number of selected seeds.



Experiments

Setup 2: fix the number of selected seeds, vary the individual propagation probability p.



Runtime

Table: 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 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

Conclusion

- We design a new information diffusion model for hypergraphs to model the spread of information among groups.
- We formulate a new information coverage maximization problem in hypergraphs based on the designed diffusion model. Further, we give detailed analysis of the defined problem.
- We propose a new heuristic algorithm based on the Degree Discount solution, which has high efficiency and achieves competitive results.

Potential Directions

Consider complex group effects, e.g.,

- Groups of different sizes may incur different costs
- Users in different groups may have different probabilities of being informed.

Group-based rumor control / source detection.

The End