基于高阶反向影响采样的无向超图影响力最大化算法

芮晓彬[1,2,3]，吉嘉欣[1]，方强鹏[1]，时纪龙[1]，王志晓[1,2,3]

（1.中国矿业大学计算机科学与技术学院/人工智能学院，徐州 江苏 221116；2.矿山数字化教育部工程研究中心，徐州 江苏 221116；3.地下空间智能感知与应急物联江苏省产业技术工程化中心，徐州 江苏 221116）

摘要：现有无向超图影响力最大化研究中的传播模型普遍缺乏对超图高阶结构的深入分析与有效建模，难以刻画信息在群体层面的复杂传播机制；同时，现有算法多依赖节点的局部拓扑特征，难以反映信息的群体传播行为。为此，本文深入探究无向超图中的信息传播机制，对无向超图中两种典型信息群体传播过程进行了刻画，提出了无向超图独立级联传播模型；在此基础上，提出了高阶反向影响采样方法，并设计了同时适用于两种不同激活策略的无向超图影响力最大化算法，并通过理论分析证明该算法能够达到与最优解 $1-1/e-\varepsilon$ 的近似比。在 8 个真实超图数据集上的实验结果表明，本文算法的效果显著优于基线方法，影响力扩展度最高提升可达 42.85%，平均的运行时间仅为 CELF 算法的 0.7%，与基于启发式的算法基本持平。

关键词：影响力最大化；无向超图；独立级联模型；反向影响采样；社交网络

中图分类号：TP301.6

文献标志码：A

Influence maximization for undirected hypergraphs based on higher-order reverse influence sampling

Rui Xiaobin[1,2,3], Ji Jiaxin[1], Fang Qiangpeng[1], Shi Jilong[1], Wang Zhixiao[1,2,3]

1. School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Technology, Xuzhou 221116, China

2. Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China

3. Jiangsu Provincial Industrial Technology Engineering Center for Intelligent Sensing and Emergency IoT in Underground Space, Xuzhou 221116, China

Abstract: Existing studies on undirected hypergraph influence maximization generally lack in-depth analysis and effective modeling of hypergraph higher-order structures, making it difficult to characterize the complex group-level information diffusion mechanisms. Meanwhile, most existing algorithms rely heavily on local topological and fail to capture collective behaviors. To address these issues, this paper systematically investigates information diffusion in undirected hypergraphs, formalizes two typical group-based information diffusion processes, and proposes an Undirected Hypergraph Independent Cascade model. Based on this model and the classic Reverse Influence Sampling (RIS) framework, we further develop a hypergraph influence maximization algorithm based on higher-order Reverse Influence Sampling, and theoretically prove that it achieves a $(1-1/e-\varepsilon)$ approximation to the optimal solution. Experiments on eight real-world hypergraph datasets show that our method significantly outperforms all baseline approaches, improving the influence spread by up to 42.85%. Meanwhile, its average runtime is only 0.7% of the CELF algorithm and is comparable to heuristic-based methods.

Key words: Influence maximization, hypergraph, independent cascade model, reverse influence sampling, social network

---

收稿日期：XXXX–XX–XX；**修回日期：**XXXX–XX–XX

通信作者：王志晓，zhxwang@cumt.edu.cn

基金项目：国家自然科学基金资助项目（No.62402496）；江苏省基础研发基金资助项目（No.BK20242084）

Foundation Items: The National Natural Science Foundation of China (No. 62402496), The Basic Research Program of Jiangsu Province (No. BK20242084)

0 引言

在当今信息化社会中，社交网络平台的普及为广告营销、疾病防控、舆情监测等领域带来了前所未有的便利支持，但同时谣言、虚假广告、网络暴力等有害信息的扩散风险也显著提升。深入研究社交网络中的信息传播机制，不仅有助于促进有益信息的广泛传播，还能有效遏制有害信息的负面影响，并且在下游应用显示出广泛潜力[1-2]，吸引众多研究者开展影响力最大化问题的研究。

影响力最大化问题于21世纪初首次被Richards和Domingos[3]在病毒式营销的场景下提出，目标是在网络中寻找最优的种子节点集合以最大化信息传播范围。Kempe等人[4]提出两种基本的信息传播模型，即独立级联（Independent Cascade，IC）模型和线性阈值（Linear Threshold，LT）模型，并设计了经典的贪心算法。在后续研究中，如Leskovec等人[5]提出的CELF算法、Goyal等人[6]对CELF算法的优化以及Brogs等人[7]提出的RIS算法，虽提升了效率，但仍基于普通图模型。随着社交网络规模不断扩大和用户交互方式日益复杂，普通图已无法准确刻画现实世界中群体化、多层次的特征。由此，超图[①]（亦称作高阶网络）影响力最大化（Hypergraph Influence Maximization, HIM）应运而生。

信息传播模型是研究超图影响力最大化的基础，而普通图上的传统传播模型仅能支持用户间直接的传播行为模拟，无法反映信息通过群组讨论、社区互动等集体行为的传播过程。为解决这一问题，Suo等人[8]将传染病模型应用到超图中，并提出了Contact Process（CP）策略和Reactive Process（RP）策略，然而该模型中的节点可重新变为易感态，短时间内接受信息便被遗忘，与现实中信息的存留情况并不相符。后续Wang等人[9]剔除节点从感染态恢复为易感态的过程，提出了Susceptible Infected with Contact Process（SICP）模型和Susceptible Infected s with Reative Process（SIRP）模型，然而其沿用的CP策略仅允许节点选择一个所属超边并进行激活；RP策略忽略了超边特殊结构，默认激活节点的超边也均处于激活状态，不符合现实场景下用户选择性地将信息转发到一个或多个群组的行为特点。此外，Gangal等人[10]首次在超图中运用独立级联模型，规定超边仅允许被激活一次，无法模拟群体内信息的反复传播，缺乏对现实场景中群体激活情况的处理。综合来看，现有的超图信息传播模型存在超边激活机制不完善、状态转换不合理的问题，无法准确模拟现实场景中用户选择性地将信息转发到多个群组、群体内部反复讨论等复杂过程。

此外，影响力最大化问题被证明为NP难问题，大规模网络下的影响力估算更是属于#P难问题[4]，尚未形成可稳定求解的大模型通用框架。现有的超图影响力最大化算法多依托节点的拓扑属性设计启发式算法，难以反映信息的群体传播行为，无法准确评估节点的传播能力；同时，基于深度强化学习的方案也需大量计算成本并缺乏可靠的理论保障。尽管反向影响采样（Reverse Influence Sampling, RIS）技术在普通图中被证明能有效平衡影响力最大化问题的计算效率与理论保证，并在普通图场景下得到广泛应用与改进拓展[11-12]，但传统RIS依赖于确定的传播模型且仅需处理简单拓扑，难以直接应用于超图中的复杂高阶结构。

因此，本文对无向超图中的群体信息传播行为进行了明确的分析与刻画，提出了相应的传播模型，并结合RIS技术，提出了适用于无向超图的高阶反向影响采样方法，从而设计了高效可靠的超图影响力最大化算法，主要贡献有：

1）根据现实场景中用户与群体之间的交互特点，明确刻画信息的群体传播规则，提出了具有两种不同超边激活策略的无向超图独立级联模型，并剖析该模型与普通图独立级联模型之间的深层映射关系，为后续算法的理论分析打下基础；

2）基于上述传播模型，提出了高阶反向影响采样方法，构建以超边为媒介的超图反向可达集（Hypergraph-Reverse Reachable sets，H-RR sets），并设计了同时适配两种超边激活策略的超图影响力最大化算法—Hypergraph Influence Maximization Algorithm based on Higher-order Reverse Influence Sampling（HIM-HRIS）；

3）通过理论分析，本文提出的HIM-HRIS算

①不同于普通图中无向图可以看作有向图的子集，有向超图和无向超图具有本质区别，本文研究的超图特指无向超图。

法具有 1 - 1/e - ε的最优解近似比，通过计算所需超图反向可达集的数量，实现效率与准确性的平衡；实验验证表明，该算法选出的种子节点质量与算法效率均优于目前已有的超图影响力最大化算法。

1　相关工作

研究者们围绕超图的影响力最大化问题展开了算法层面的深入探索，现有方法主要可分为三类：基于节点拓扑信息的启发式方法、基于激活概率动态计算的方法以及基于深度强化学习的方法。

1.1　基于节点拓扑信息的方法

Xie 等人[13]在 SICP 模型下，提出一种超自适应度剪枝算法 HADP，使用节点度值衡量节点的影响力迭代选择当前节点度最高的节点，并采用自适应剪枝策略降低邻居节点的超度权重，减少影响重叠。相比于 HADP 算法仅考虑节点度，Gong 等人[14]将节点的所属超边数与邻居节点数之比定义为邻域系数，将 SICP 模型下超图上的影响力最大化问题转化为 SICP 变体模型下有向加权图上的影响力最大问题并提出 Adeff 算法解决变体问题。陈彬等人[15]在无向超图中运用线性阈值模型，结合节点环绕集，提出了潜力节点优先的预算贪婪算法，解决给定预算下无向超图的影响力最大化问题。EFS 算法[16]将超图类比为静电场，节点视为点电荷，从电荷初始化、库仑力计算、自优化更新三个步骤构建算法，实现多维度的节点影响力评估与种子选择。

此类方法通过提取超图的静态拓扑特征如节点度、邻域系数设计启发式规则，以高效筛选种子节点，但未深入研究信息在超图网络中传播的动态过程，并且缺乏理论保障。

1.2　基于激活概率动态计算的方法

Gangal 等人[10]基于提出的超图独立级联模型，研究 Hyperedge Majority Influence Maximization（HEMI）问题，尝试选出使得最终影响的超边数量最多的种子集合。后续 M. A 等人[17]沿用该模型，将研究问题转变为选出种子集合使得最终影响的节点数量最多，并提出 Hyper IMRANK 算法解决该问题，通过迭代调整初始排序生成最终排名，但作为排序依据的激活概率值仅计算直接共享超边的节点对的原始概率。之后 Wang 等人[9]基于 SICP 模型和 SIRP 模型提出 MHPD 算法，通过构建感染概率矩阵，迭代更新节点在设定的跳数范围内的期望感染度，从而替代蒙特卡洛模拟计算出期望传播规模。HACE 算法[18]结合传播过程建模有效扩散能力与种子集局部感染风险，构建折扣接触能力指标，迭代选择该指标值最大的节点加入种子集。相比于 HACE 算法，HCLI 算法[18]更注重候选节点的整体感染风险，引入全局感染概率，精准建模候选节点被所有邻居感染的累积概率，并通过迭代持续更新，选择该指标最大的节点作为新种子。王志萍等人[19]提出基于遗传算法的低冗余超图影响力最大化算法 LR-HGA，对节点间的影响冗余和节点的实际传播值的充分考虑融入到遗传算法的选择操作和交叉操作中，以寻找全局最优解。

相比于基于节点拓扑信息的启发式方法，此类方法虽然利用节点的初始激活概率进行动态计算，更贴合信息扩散的真实过程，但仍然局限于一阶邻居或有限跳数内的节点信息，并且随着跳数增加，时间复杂度剧增。

1.3　基于深度强化学习的方法

Xu 等人[20]提出 HEDRL-IM 算法，将超图影响力最大化这一离散组合优化问题转化为 DQN 的连续权重优化问题，利用节点对超边的潜在激活能力解决稀疏奖励问题，结合进化算法和深度强化学习探索解空间与提升解质量。Wu 等人[21]提出的 IMH-DRL 算法需通过大量合成超图完成离线训练，让智能体学习通用的种子节点选择策略，训练完成后，将训练成果应用于真实超图。之后的 HIMH-DRL 算法[22]在异质超图上提出基于类型依赖的影响力增强机制，沿用离线训练到在线应用的模式，根据学习得出的异质超图的种子选择规律，选择种子节点。

以上方法虽尝试使用深度强化学习等手段解决超图影响力最大化问题，但其可靠性缺乏理论保障。并且深度强化学习类方法需预先生成大量合成超图进行离线训练，需要额外获得超图的嵌入信息，难以实现问题的高效求解。

综上可以看出，目前有关超图影响力最大化的研究缺乏对超图传播模型的深入剖析以及对实际情况中不同传播模式的考量；此外，现有研究设计的算法难以取得不同传播情况下对超图影响力最大化问题的高效精准求解和理论保障。

2　无向超图独立级联模型

2.1　无向超图

为了方便后续说明，本节首先给出无向超图的定义，即给定无向超图 $HG=(V,E,W)$，$V=\{v_1,v_2,v_3,\cdots,v_n\}$表示无向超图所包含$n$个节点的集合，$E=\{e_1,e_2,e_3,\cdots,e_m\}$表示无向超图中$m$条超边构成的集合。对于任意超边$e\in E$，满足$e\subseteq V$且$e\neq\varnothing$。$W$为权重函数，对于任意节点$v\in V$和超边$e\in E$，有$W(v,e)=\{p_{v\to e},p_{e\to v}\}$，其中$p_{v\to e}$表示节点$v$对所属超边$e$的成功激活概率，$p_{e\to v}$表示超边e对其所包含节点$v$的成功激活概率。

图1给出了一个无向超图的示例，其中$V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7\}$，$E=\{e_1,e_2,e_3,e_4\}$；$e_1=\{v_1,v_2,v_3,v_4,v_5\}$，$e_2=\{v_2,v_3,v_4,v_5\}$，$e_3=\{v_4,v_5,v_6\}$，$e_4=\{v_7\}$。在该图中，可以形象地理解为表示用户的节点$v_1,v_2,v_3,v_4,v_5$处在同一个社群$e_1$，信息在这个群体中可以进行广播；与此同时，节点$v_2$还存在于超边$e_2$中，因此，用户$v_2$可以将在社群$e_1$中获得的信息以某种方式传播至社群$e_2$。相比之下，节点$v_7$仅处于单个社群$e_4$中，且$e_4$也只包含这一个单独用户，那么假设用户$v_7$作为信息的初始拥有者，则信息在该无向超图中将无法继续传播。

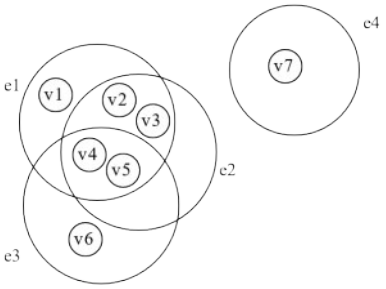

图1　无向超图

从无向超图的定义与上述实例可以得出，无向超图中的一条超边具备连接多个节点的能力，表示一个讨论群组或共享信息的用户集合，每个节点均可以是信息的发送者或接收者。因此，无向超图适用于每个社交用户处于多个群体的情况，能够更准确地刻画信息在社交网络间通过群体传播的形式。

2.2　面向无向超图的独立级联模型

普通图无法刻画信息在群体中的传播过程，而现有的超图信息传播模型适用的场景较为单一，无法准确模拟现实场景中用户选择性地将信息转发到多个群组、群体内部信息的反复讨论等复杂过程。本文深刻分析了信息在群组中的实际传播情况，总结信息在群体中传播的三个核心特点：

（1）信息由用户以一定概率传播到用户所在的社交群体中；

（2）社交群体中的信息以一定概率被该群体中的其他用户所接收。

（3）用户获知某条信息后，会长期维持信息的接收状态，不会因其他用户或群体的后续传播影响自身的已接受的信息的状态。

此外，在实际传播过程中，社交群体可能一次或者多次接受相同的信息。前者对应现实中转发通知信息的场景，例如一条通知信息由上级发送至群组中，再由群组中下级组织的各负责人转发到对应组织中，而群组中的其他人并不会重复发送该通知；后者刻画了热点信息或重大新闻在现实中传播的场景，例如针对某热点新闻，同一群组内的不同用户在不同时刻获知该信息后都可能向该群组转发。

基于以上信息传播特点，本文结合超图结构与普通图上的IC模型[4]，提出了面向无向超图的独立级联Undirected Hypergraph Independent Cascade（HIC）模型以及相应的两种传播策略One Touch（OT）和Multi Touch（MT），分别刻画上述两种不同的超图信息传播场景，具体传播规则如下：

给定一个无向超图$HG=(V,E,W)$，本文用$N(\mathrm{e})$表示超边$e\in E$所包含的节点集合，$N(v)$表示具体节点$v\in V$所在的超边集合。节点只存在未激活与激活两种状态，且一旦激活，状态就不再更改。每个节点在激活后有且仅有一次机会去分别激活其所属的处于非激活状态的超边，每条超边在激活后也仅有一次机会在下一时刻去分别激活其所包含的节点。OT策略下，每条超边被激活后会维持激活状态，即在传播过程中总计只能被激活一次；而MT策略下，超边向节点传播信息后会重置为未激活状态，即每条超边允许先后多次被激活。

以下举例说明HIC-OT与HIC-MT两种模型的具体传播过程，设定节点激活所属超边的概率为0.8，超边激活内部节点的概率为0.5。如图2所示，给定$HG=(V,E,W)$，$V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7\}$，$E=\{e_1,e_2,e_3\}$，$e_1=\{v_1,v_2,v_3\}$，$e_2=\{v_2,v_3,v_6,v_7\}$，$e_3=$

$\{v_3,v_4,v_5\}$，$S=\{v_1\}$，示例传播过程如下：

（1）$t=1$时刻，种子节点v_1以给定概率尝试激活e_1，假设成功激活，则在HIC-OT与HIC-MT模型中，e_1都将处于激活状态（图中用阴影表示）。

（2）$t=2$时刻，无论HIC-OT与HIC-MT模型，e_1都会以给定概率去尝试激活其包含的未激活节点v_2和v_3，假设在两种模型下都仅成功激活v_3。之后，HIC-OT模型中的e_1维持激活状态，而HIC-MT模型中的e_1则重置为未激活状态。

（3）$t=3$时刻，在HIC-OT模型中，虽然v_3同时属于e_1、e_2和e_3，但由于e_1是激活状态，所以v_3仅能尝试去激活e_2、e_3，假设尝试激活e_3成功；而在HIC-MT模型中，v_3则会分别以给定概率尝试激活e_1、e_2和e_3，假设尝试激活e_1和e_3成功。

（4）$t=4$时刻，在HIC-OT模型中，e_3尝试去激活其包含的未激活节点v_4和v_5，假设只有v_4被成功激活；而在HIC-MT模型中，假设e_3的激活情况与HIC-OT模型相同，e_1尝试去激活其包含的未激活节点v_2，假设该激活成功。之后，HIC-OT模型中的e_3维持激活状态，而HIC-MT模型中的e_1和e_3则均重置为未激活状态。

（5）$t=5$时刻，HIC-OT模型中的传播已结束；而在HIC-MT模型中，v_2未能激活e_1和e_2，但v_4成功激活e_3，而e_3又在$t=6$时刻进一步激活v_5，至此HIC-MT模型中的传播过程同样结束。

2.3　HIC模型与普通图IC模型的映射关系

本小节通过分析HIC-OT和HIC-MT模型与普通图IC模型的映射关系，深入剖析HIC模型的传播机制，为设计可靠的近似算法提供理论依据。

（1）在HIC-OT模型下，可将$HG(V,E,W)$映射为二部图$G(V',E',W)$，其中$V'=V\cup E$，表示V'包含n个原始节点和m个新增的“超边节点”（即将超图HG中的每一条超边都视作二部图G中的一个新节点），从而$|V'|=n+m$。此外，二部图G的边集E'遵循以下规则进行构建：对于每个节点$v\in V$及其所属超边$e\in N(v)$，对应一条包含两个权重$p_{v\to e}$和$p_{e\to v}$的无向边(v,e)，因此总边数为各条超边包含节点数的总和，即$|E'|=\sum_{e\in E}|N(e)|$。

上述映射保持了HIC-OT传播过程的等价性，即原超图HG中节点激活超边的过程对应为二部图

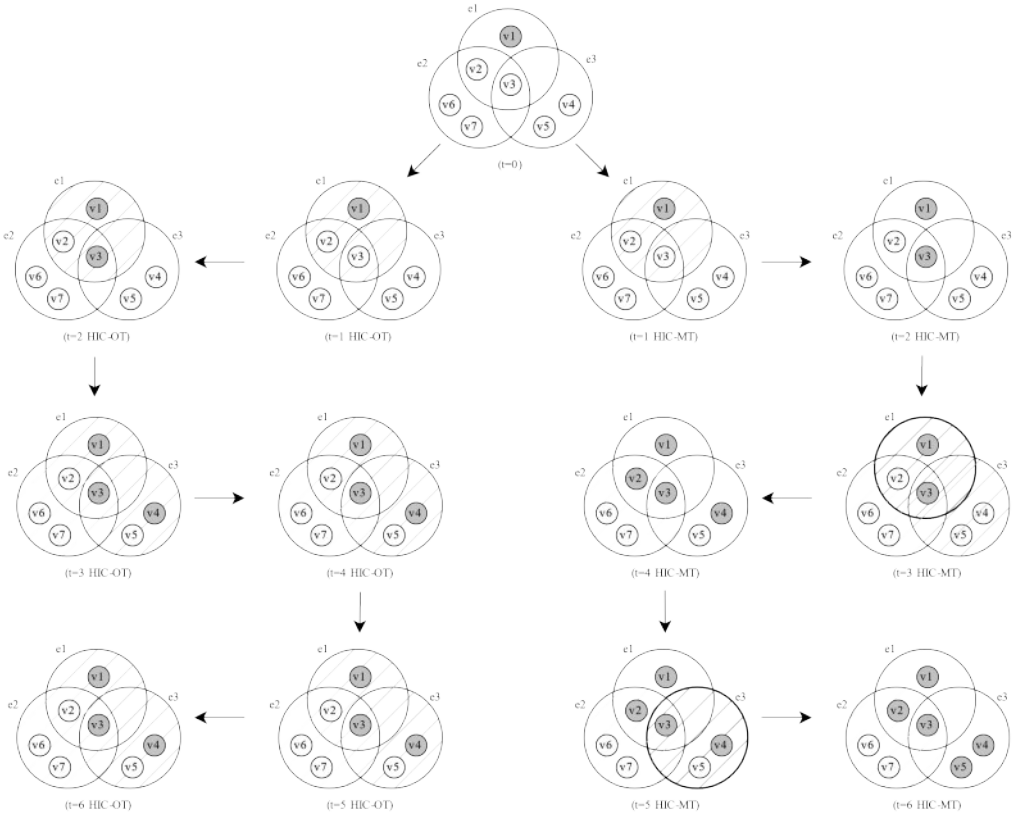

图2　无向超图独立级联模型的传播过程（加粗表示该超边被多次激活）

G中节点激活“超边节点”，原超图HG中超边激活节点的过程则对应为二部图G中“超边节点”激活普通节点。在该映射中，OT策略的特性（每条超边节点只能被激活一次）对应普通图IC模型中“每个节点仅能被激活一次”的规则。例如，图1所示例的超图在映射后的二部图如图3所示，该二部图包含节点集$V'=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7,e_1,e_2,e_3,e_4\}$以及对应的边集$E'=\{(v_1,e_1),(v_2,e_1),(v_3,e_1),(v_4,e_1),(v_5,e_1),(v_2,e_2),(v_3,e_2),(v_4,e_2),(v_5,e_2),(v_4,e_3),(v_5,e_3),(v_6,e_3),(v_7,e_4)\}$。

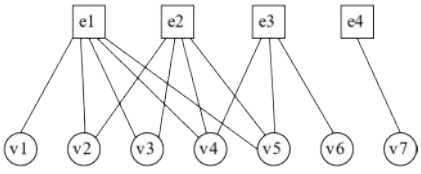

图3　图1示例的无向超图在HIC-OT模型下的二部图映射

（2）在HIC-MT模型下，可以将$HG(V,E,W)$映射为普通图$G(V,E',W')$，其中节点集V（$|V|=n$）保持不变，而边集E'则通过以下方式构建：对于任意两个节点u和v，若它们至少共享一条超边，即存在$e\in E$使得$u,v\in N(e)$，则在E'中建立无向边(u,v)，和HIC-OT模型的映射类似，每条边对应两个权重$W'(u,v)$和$W'(v,u)$。$W'(u,v)$表示u通过与v共享的所有超边去激活v的概率，即$W'(u,v)=1-\prod_{e\in(N(u)\cap N(v))}(1-p_{u\to e}*p_{e\to v})$，$|E'|$等于所有超边中节点组合数减去属于多个超边的节点组合数，即$|E'|=\sum_{e\in E}\binom{|N(e)|}{2}-\sum_{(u,v)\in V\times V}\max\{0,\varphi_{u,v}-1\}/2$，其中$\varphi_{u,v}=|\{e\in E:u,v\in N(e)\}|$。

上述映射将超图中的高阶传播关系转化为普通图中的二元关系，保留了MT策略允许节点通过不同超边多次尝试激活邻居的特性。例如，在如下图4的示例中，节点v_1和v_2仅共同属于一条超边e_1，则有$p_{v_1\to v_2}=p_{v_1\to e_1}*p_{e_1\to v_2}$，而节点$v_4$和$v_2$共同属于超边$e_1$和$e_2$，所以$W'(v_4,v_2)=p_{v_4\to v_2}=1-(1-p_{v_4\to e_1}*p_{e_1\to v_2})*(1-p_{v_4\to e_2}*p_{e_2\to v_2})$。

2.4　HIC模型下的无向超图影响力最大化

影响力最大化问题通常要求选择给定数量（通常用k表示）的节点构成种子节点集合S，通过将这些节点设置为初始激活节点，从而最大化影响力扩展度$\sigma(S)$。在普通图中，影响力扩展度$\sigma(S)$的含义为最终处于激活状态的节点数量期望。类似的，真实世界中的群组（超图中的超边）本质上是信息传播过程中作为信息扩散的媒介，因此不参与最终影响力扩展度的计算，即超图中的影响力扩展度$\sigma(S)$同样为最终处于激活状态的节点数量期望。

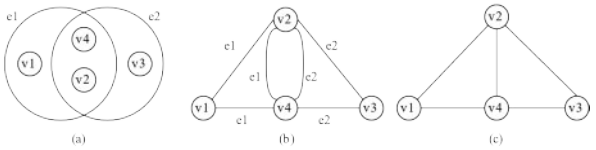

图4　在HIC-MT模型下将无向超图转换为普通图的过程。图(a)为无向超图，图(b)为保留连接节点间超边的示意图，图(c)为将图(b)中节点之间多条路径转换为一条路径的普通图。

基于影响力扩展度$\sigma(S)$和超图传播模型HIC，本节给出无向超图影响力最大化问题的定义，即：给定一个无向超图$HG=(V,E,W)$和种子节点数量k，无向超图影响力最大化问题要求找到包含k个种子节点的集合S，使得$\sigma(S)$最大，如下公式(1)所示：

$$S\in argmax_{S_0\in V,|S_0|\le k}\sigma(S),\#(1)$$

其中$\sigma(S)$具体表示种子节点集合S在无向超图HG中基于HIC-OT或HIC-MT模型传播信息后最终影响节点数量的期望。此外，根据公式(1)，该问题在选择种子集合时也仅允许选择节点，即由个体而非群组作为信息传播的发起者。

定理1　在HIC-OT和HIC-MT模型下，无向超图的影响力最大化问题是NP难的。

证明　首先，将最大集合覆盖（Max Set Cover，MSC）问题规约成HIC-OT模型下的超图影响力最大化问题。已知MSC问题是经典NP问题，给定基本集$U=\{u_1,u_2,\cdots,u_m\}$，子集族$F=\{F_1,F_2,\cdots,F_n\}$和整数k，MSC问题要求找到F中的k个子集，使其并集覆盖U中的元素数量最大化。上述问题可映射为超图$HG_{SC}=(V_{SC},E_{SC},W_{SC})$，具体步骤如下：构造节点集$V_{SC}$大小为$n+m$，包含两类节点，将基本集$U$的每一个元素$u_j$都转化为超图中的一个节点$v_{u_j}$，将子集族$F$的每一个子集$F_i$都转化为超图中的一个节点$v_{F_i}$。对于任意$F_i\in F$能够覆盖元素$u_i$，构造一条超边$e_i$包含$F_i$和$u_i$所对应的节

点，且设置超边影响节点的概率和节点影响超边的概率均为1。

在HIC-OT模型下，超边仅能被激活成功一次，种子集合S激活超边e_{ij}成功当且仅当$v_{Fi}\in S$且$u_j\in F_i$，覆盖等价性成立，即$\sigma(S)=|\bigcup_{v_{Fi}\in S}F_i|$。因此判定该SMC问题的解等价于判定影响力最大化问题的解。因为SMC问题是一个NP难的问题，所以HIC-OT模型下的超图影响力最大化问题是一个NP难的问题。在HIC-MT模型下，超边可被多次激活，但由于激活概率为1，且节点v_{u_j}一旦被激活后状态不可逆，多次激活不会改变覆盖结果，与OT策略下的覆盖等价性完全一致，同理得证。■

定理2　在HIC-OT和HIC-MT模型下，无向超图影响力最大化问题中的影响力扩展度$\sigma(S)$均为非负的单调次模函数。

证明　给定集合$S\subseteq T\subseteq V$，记$\Delta(v|S)=\sigma(S\cup\{v\})-\sigma(S)$；同理，$\Delta(v|T)=\sigma(T\cup\{v\})-\sigma(T)$。记$P_v$为所有从$v$出发的路径集合，$\Delta(v|S)=\sum_{P\in P_v}\Pr(P\text{ is active})*1_{\{P\cap S=\varnothing\}}$，其中$1_{\{P\cap S=\varnothing\}}$为指示函数。对于一条路径$P\in P_v$，将路径$P$分解为$P=(v_0,e_0,v_1,e_1,\cdots,v_t,e_t,u)$，OT策略下，路径中的超边$e_t$只能被其前驱节点$v_t$成功激活一次，MT策略下，$e_t$可以被多次激活，因此可能重复出现，两种策略仅会引起路径所包含的元素的不同，路径P是活跃的概率只取决于超图的结构和概率设置，与种子集S或T无关。由于$S\subseteq T$，所以有$\{P|P\cap S=\varnothing\}\supseteq\{P|P\cap T=\varnothing\}$，所以$1_{\{P\cap S=\varnothing\}}\geqslant 1_{\{P\cap T=\varnothing\}}$，并且$\Pr(P\text{是活跃的})$独立于$S$和$T$，所以$\Delta(v|S)\geqslant\Delta(v|T)$，

定理得证。■

3　算法设计

本节介绍本文提出的基于高阶反向影响采样的无向超图影响力最大化算法（Hypergraph Influence Maximization Algorithm based on Hyper Reverse Influence Sampling，HIM-HRIS），详细说明算法思想和算法过程，并给出关于HIM-HRIS算法近似比保障的理论分析。

3.1　HIM-HRIS算法概述

整体而言，HIM-HRIS算法分为高阶反向影响采样部分和种子节点选择部分，初始化空种子集合后，将超图与传播规则传入高阶反向影响采样模块，输出超图反向可达集集合；该结果传入种子节点选择模块，结合预算约束完成节点择优筛选，最终输出最优种子节点集合S，具体如算法1所示。

算法1 HIM-HRIS算法

输入：无向超图$HG=(V,E,W)$，策略OT/MT，预算k

输出：种子节点集合S

1. 初始化$S=\varnothing$
2. $HR\leftarrow$高阶反向影响采样
3. $S\leftarrow$种子节点选择
4. *RETURN S*

3.2　基于HIC的高阶反向影响采样

传统的反向影响采样方法常用于普通图中的IC模型和LT模型，通过提前生成大量的反向可达集（Reverse Reachable sets, RR sets）对节点的影响力进行准确评估，进而对算法的效果提供保障。然而，由于超图与普通图的结构差异以及超图中信息传播规则的变化，普通图中的RIS方法无法直接应用于无向超图中。不过，本文在2.3节详细论述了本文所提出的HIC模型与普通图IC模型之间的映射关系，因此可以将传统的普通图上的反向影响采样方法通过映射关系应用在无向超图中。

简单来说，给定一个无向超图$HG=(V,E,W)$，超图反向可达集的含义如下：节点v的一个随机超图反向可达集，记作$R(v)$，是在HG上的一次随机传播实例中可以激活v的所有元素的集合。

具体来说，基于HIC传播模型，给定一个无向超图$HG=(V,E,W)$，从任意一个节点$v\in V$出发的一次高阶反向影响采样过程本质上为从v出发，模拟HIC的反向传播过程，即首先依据给定传播概率模拟本次能够成功激活节点v的超边，随后再模拟本次能够激活该超边的节点，持续迭代，直至不再产生新的激活节点或超边为止；最终，本文将本次反向影响采样中的这些节点和超边构成的集合记为$R(v)$，其中节点构成的集合记为$R_V(v)$，超边构成的集合记为$R_E(v)$，即$R(v)=R_V(v)\cup R_E(v)$，$R_V(v)\cap R_E(v)=\varnothing$。综上，基于HIC模型的一次详细反向影响采样过程可描述为：

（1）在V中随机选择一个节点v（通常被称为

根节点）进行一次反向影响采样，此时$R(v)=\{v\}$，$R_V(v)=\{v\}$，$R_E(v)=\varnothing$。另外，初始化最新激活节点集合$V_{new}=\{v\}$，最新激活超边集合$E_{new}=\varnothing$；

（2）OT策略下，$\forall u\in V_{new}$，以$p_{e\to v}$的概率对所有$e\in N(u)\backslash R_E(v)$进行采样，若成功，则将超边$e$加入$R_E(v)$和$E_{new}$；MT策略下，$\forall u\in V_{new}$，以$p_{e\to u}$的概率对所有$e\in N(u)$进行采样，若成功，则将$e$加入$R(v)$，$R_E(v)$和$E_{new}$；此后，将$V_{new}$置为空集；

（3）$\forall e\in E_{new}$，以$p_{u\to e}$的概率对所有$u\in N(e)\backslash R_V(v)$进行采样，若采样成功，则将节点$u$加入$R(v)$，$R_V(v)$和$V_{new}$；此后，将$E_{new}$置为空集；

（4）重复上述过程，直至第2步或第3步不再有任何元素加入$R(v)$。

上述高阶反向影响采样过程最终返回一个以v为根节点的随机超图反向可达集H-RR set，其物理含义为：在此次随机传播实例中，该H-RR set中的所有节点都能够激活根节点v。不难看出，由于普通图IC模型与超图HIC模型之间存在映射关系，传统的反向影响采样技术通过简单调整便可应用于超图中，形成了高阶反向影响采样方法。

图5给出了H-RR sets的示例。无向超图$HG=(V,E,W)$的结构如图5（a）所示，其中$V=\{v_1,v_2,v_3,v_4,v_5\}$，$E=\{e_1,e_2,e_3\}$，$e_1=\{v_1,v_2,v_3\}$，$e_2=\{v_3,v_4,v_5\}$，$e_3=\{v_2,v_3\}$，图5（b）、（c）和（d）以子图形式给出了以$v_1$为根节点的3个H-RR set的例子，其中图5（b）和（c）在HIC-OT模型下进行反向采样，图5（c）在HIC-MT模型下进行反向采样，加粗表示该超边被多次激活。图5（b）中v_1反向传播到了e_1，但没有进一步反向传播到e_1包含的节点，这个H-RR set为$\{v_1\}$；图5（c）v_1通过e_1反向传播到v_3，在HIC-OT模型下e_1已加入$R_E(v_1)$，所以v_3仅能尝试反向激活e_2和e_3，成功反向激活e_3后没能再反向激活新的节点，这个H-RR set为$\{v_1,v_3\}$；图5（d）中v_1通过e_1反向传播到v_3，在HIC-MT模型下可以尝试反向激活e_1,e_2和e_3，假设成功反向激活e_1和e_3，那么e_1可尝试反向激活v_2，激活成功后，v_2无法通过超边再反向传播到其他节点，最终这个H-RR set为$\{v_1,v_3,v_2\}$。

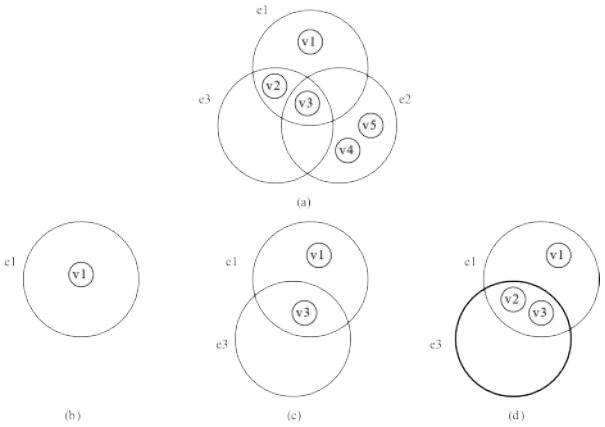

图5 超图反向可达集的示例

3.3 种子节点选择

基于以上对无向超图反向可达集的分析，可将无向超图影响力最大化问题转化为种子节点对于最终生成的反向可达集集合的最大覆盖问题。

给定$HG=(V,E,W)$，预算k和高阶反向采样生成的θ个随机H-RR sets构成的集合HR，种子节点选择过程包括以下几个步骤：

（1）遍历所有未被选中的节点$v\in V$，计算节点v覆盖的有效的H-RR set的数量$count(v)$；

（2）从H-RR sets中选择覆盖集合数量最多即$count(v)$最大的H-RR sets的节点v_r，并将其加入种子集合S；

（3）从当前H-RR sets中移除所有包含已选择节点v_r的超图反向可达集，消除影响力重叠；

（4）重复上述过程，直至选择k个种子节点。

3.4 HIM-HRIS算法近似比理论分析

定理3 给定$HG=\{V,E,W\}$，在HIC-OT模型和HIC-MT模型下，对于任意种子集合$S\in V$，任意节点$v\in V$，种子集合S激活节点v的概率$ap(S,v)$等于S与以v为根的一个随机超图反向可达集$R(v)$相交的概率，即$ap(S,v)=\Pr\{S\cap R(v)\neq\varnothing\}$。从而种子集合$S$的影响力扩展度满足

$$\sigma(S)=n^*\Pr\{S\cap R\neq\varnothing\}=n^*\Pr\{S\cap R_V\neq\varnothing\},\#(2)$$

其中R是一个根从V中随机选取的超图反向可达集，R_V是仅包含R中节点元素的集合。

证明　在HIC-OT模型下，结合3.3节将超图HG转换为二部图$G(V',E',W)$，给出基于活跃边图[4]的HIC-OT模型的等价定义，给定二部图每条边上的两个权重（即$p_{v\to e}$和$p_{e\to v}$），HIC-OT模型是由下列元素构成的一个随机传播模型的特例：

1）二部图$G=(V,E,W)$，节点集$V=V^V\cup V^E$，V^V表示原超图中的节点集，其中的节点记为v^V，V^E表示原超图中超边转节点集，其中的节点记为v^E；

2）每个节点的状态空间$\sum=\{0,1\}$；

3）由随机活跃边图L组成的有限传播概率空间Ω，其中随机生成一个特定L的概率围为$\Pr(L)=\prod_{(u,v)\in E(L)}p(u,v)*\prod_{(u,v)\in E(G)\setminus E(L)}(1-p(u,v))$，其中$E(G)$和$E(L)$分别表示图$G$和图$L$中的边；

4）在图G中传播事件的离散时间序列$\{(t,v)|t=1,2,3\cdots,v\in V_V\}$；

5）传播函数$F_v\left(X_{v,t},\left\{X_{u,t}|u\in N^-(v)\right\},L\right)=\|\left\{\vee_{u\in N_L^-(v)}X_{u,t}=1\vee X_{v,t}=1\right\}$，其中$N^-(v)$表示$v$的入邻居，$X_{v,t}$表示$v$在$t$时刻的激活状态，$\|\{\varepsilon\}$当事件$\varepsilon$为真时值为1，反之为0，即当节点$v$自己或有一个在$L$中的入邻居在前一时刻是激活的，那么$v$在当前时刻也是激活的。种子集合$S_0=\left\{v\in V_V|X_{v,0}=1\right\}$。用$S_t$表示从一次传播中截止$t$时刻的活跃节点集合，用$\Gamma_t(L,S)$表示从节点集合$S$沿图$L$中有向边$t$步内能到达的节点集合，$0\leqslant t\leqslant m+n-1$，首先证明$S_t=\Gamma_t(L,S_0)$。

在由S_0和活跃边图L决定的一次传播实例中，设$v^V\in\Gamma_t\left(L,S_0\right)$，则存在$v_0^V\in S_0$和至少一条活$L$中的路径$\left(v_0^V,v_1^E,v_2^V,\cdots,v_\tau^V\right)$，$0\leqslant\tau\leqslant t$，使$v_\tau^V=v^V$，则有$X_{v_1^E,1}=F_{v_1^E}(X_{v_1^E,1},\{X_{u,1}|u\in N^-(v_1^E)\},L)=\|\{V_{u\in N_L^-(v_1^E)}(X_{u,0}=1\wedge(u,v_1^E)\in E(L))\vee X_{v_1^E,0}=1\}$，因为$v_0^V\in N^-(v_1^E)$，且$X_{v_0^V,0}=1$，$(v_0^V,v_1^E)\in E(L)$，所以$X_{v_1^E}=1$。以此类推，$v=v_\tau^V$在时刻$\tau$是活跃的，因此$v^V\in S_t$。对于任意$t$时刻最终活跃节点$v\in S_t$，一定能找到长度小于等于$t$的激活序列$(v_0^V,v_1^E,v_2^V,v_3^E\cdots,v_{2\tau}^V)$，$0\leqslant\tau\leqslant t/2$，使得$v_0^V\in S_0$，$v_\tau^V=v$，$v_{2i-2}^V$激活$v_{2i-1}^E$，$v_{2i-2}^E$激活$v_{2i-1}^V$，对所有的$i=1,2,\cdots,\tau$成立，因此$(v_0^V,v_1^E,v_2^V,v_3^E\cdots,v_{2\tau}^V)$是$L$中的一个路径，从而$v=v_\tau^V$可以从一个种子节点$v_0\in S_0$在$L$中到达，即$v\in\Gamma_t(L,S_0)$。因此$S_t=\Gamma_t(L,S_0)$。

对于一个随机活跃边图L，种子集合S激活v^V等价于S在L中可以到达v^V，等价于存在$u\in S$，使得u在L中可达v，即$v\in\Gamma(L,\{u\})$。根据H-RR set的定义，$v\in\Gamma(L,\{u\})$等价于$u\in HR(v)$。因此S激活v等价于存在节点$u\in S$，$u\in HR(v)$，即$S\cap HR(v)\neq\varnothing$。因此，当$R$的根$v$均匀选取时，$\Pr\{S\cap R\neq\varnothing\}=\sum_{v\in V}\frac{1}{n}\Pr\{S\cap R(v)\neq\varnothing\}=\frac{1}{n}\sum_{v\in V}ap(S,v)=\frac{1}{n}\sigma(S)$。

相较于HIC-OT模型，在HIC-MT模型下，除了随机活跃边图L的概率分布公式有所不同，其余证明过程类似。综上所述，超边在HIC-OT模型和HIC-MT模型下仅作为转播媒介，不影响对于种子节点影响力的计算，定理得证。■

此外，可以从反向影响采样集合覆盖的角度理解超边在HIC模型下作为传播媒介的作用，对于$\forall e\in R_E(v)$，假设e在t时刻加入$R_E(v)$，那么在$t-1$时刻，一定存在$v\in R_V(v)$且$v\in e$，在最终的超图反向可达集$R(v)$中仅保留$R_V(v)$即可刻画反向可达集对根节点v的影响。因此得出下述公式(3)：

$$\begin{aligned}\Pr\{S\cap R\neq\varnothing\}&=\Pr\left\{S\cap\left(R_E\cup R_V\right)\neq\varnothing\right\}\\&=\Pr\{S\cap R_V\neq\varnothing\}\end{aligned}\quad\#(3)$$

基于上述理解，如果一个包含k个节点的种子集合S能够覆盖最多的H-RR sets，那么这k个节点构成的集合S具备将信息传播至最大范围的能力。并且H-RR sets数量越庞大，最终选出种子节点集合的准确性越理想；但随着H-RR sets数量的增加，生成H-RR sets的时间成本也随之增加。因此，需要确定合适的H-RR sets数量，从而平衡时间消耗与算法准确性。

定理4　定义S^*是$\sigma(S)$下的最优解，即满足$S^*\in argmax_{S\in V,|S|\leqslant k}\sigma(S)$。令$OPT=\sigma\left(S^*\right)$，对于任意$\delta_1,\delta_2>0$，$\varepsilon>0$，和任意$\varepsilon_1\epsilon\left(0,\varepsilon/(1-1/e)\right)$，给定$HG=(V,E,W)$，节点数量为$n$，种子节点集合大小为$k$，$\Omega$表示所有随机元素$R_0$构成的概率空间，令

$$\theta_1=\frac{2n\cdot\ln\left(\frac{1}{\delta_1}\right)}{OPT\cdot\varepsilon_1^2},\theta_2=\frac{\left(2-\frac{2}{e}\right)\cdot n\cdot\ln\left(\binom{n}{k}/\delta_2\right)}{OPT\cdot\left(\varepsilon-\left(1-\frac{1}{e}\right)\cdot\varepsilon_1\right)^2}\#(4)$$

则 $\forall\theta>\theta_1$，$Pr_{R_0\sim\Omega}\{\widehat{\sigma_\theta}(S^*,R_0)\geqslant(1-\varepsilon_1)\cdot OPT\}\geqslant 1-\delta_1$；$\forall\theta>\theta_2$，对于每个相对于$\varepsilon$来说坏的$S$，即$\sigma(S)<(1-1/e-\varepsilon)*OPT$，有$Pr_{R_0\sim\Omega}\{\widehat{\sigma_\theta}(S,R_0)\geqslant(1-1/e)(1-\varepsilon_1)\cdot OPT\}\leqslant\delta_2/\binom{n}{k}$。

证明 给出Chernoff界定义如下：假设$X_1X_2,\cdots,X_t$是t个$[0,1]$上的相互独立的随机变量，且$\exists\mu\epsilon[0,1]$使得$E[X_i]=\mu$对任意$i\in[t]$均成立。令$Y=\sum_{i=1}^{t}X_i$，则对于$\forall\gamma>0$，有

$$\Pr\{y-t\mu\geqslant\gamma\cdot t\mu\}\leqslant\exp\left(-\frac{\gamma^2}{2+\frac{2}{3}\gamma}t\mu\right)\#(5)$$

对于$\forall 0<\gamma<1$，有

$$\Pr\{y-t\mu\leqslant-\gamma\cdot t\mu\}\leqslant\exp\left(-\frac{\gamma^2}{2}t\mu\right)\#(6)$$

用$R_0[\theta]$表示生成的θ个随机独立的超图反向可达集，对于所有节点子集S，$X_i^{R_0}(S)$表示$R_0[\theta]$的第i个集合是否与S相交，由公式(2)，$X_i^{R_0}(S)=\sigma(S)/n\in[0,1]$，应用Chernoff界，有

$$Pr_{R_0\sim\Omega}\{\widehat{\sigma_\theta}(S^*,R_0)<(1-\varepsilon_1)*OPT\}$$

$$=Pr_{R_0\sim\tau}\{\mathrm{n}*\frac{\sum_{i=1}^{\theta}X_i^{R_0}(s^*)}{\theta}<(1-\varepsilon_1)*\sigma(S^*)\}$$

$$=Pr_{R_0\sim\Omega}\left\{\sum_{i=1}^{\theta}X_i^{R_0}(s^*)-\theta*\frac{\sigma(S^*)}{m+n}<-\varepsilon_1\left(\theta*\frac{\sigma(s^*)}{n}\right)\right\}$$

$$\leqslant\exp\left(-\frac{\varepsilon_1^2}{2}\theta*\frac{\sigma(s^*)}{n}\right)$$

$$\leqslant\left(-\frac{\varepsilon_1^2}{2}*\frac{2n*\ln\left(\frac{1}{\delta_1}\right)}{OPT*\varepsilon_1^2}*\frac{\sigma(s^*)}{n}\right)=\delta_1$$

令$\varepsilon_2=\varepsilon-(1-1/\mathrm{e})*\varepsilon_1$，令$S$是相对于$\varepsilon$来说坏的解，则有

$$Pr_{R_0\sim\Omega}\{\widehat{\sigma_\theta}(S,R_0)\geqslant(1-1/e)(1-\varepsilon_1)*OPT\}$$

$$=Pr_{R_0\sim\Omega}\left\{\sum_{i=1}^{\theta}X_i^{R_0}(S)-\theta*\frac{\sigma(S)}{m+n}\geqslant\frac{\theta}{n}*\left(\left(1-\frac{1}{e}\right)*(1-\varepsilon_1)*OPT\right)-\sigma(S)\right\}$$

$$\leqslant Pr_{R_0\sim\Omega}\left\{\sum_{i=1}^{\theta}X_i^{R_0}(S)-\theta*\frac{\sigma(S)}{n}\geqslant\frac{\theta}{n}*\varepsilon_2*OPT\right\}$$

$$=Pr_{R_0\sim\Omega}\left\{\sum_{i=1}^{\theta}X_i^{R_0}(S)-\theta*\frac{\sigma(S)}{n}\geqslant\left(\varepsilon_2*\frac{OPT}{\sigma(S)}\right)*\theta*\frac{\sigma(S)}{n}\right\}$$

$$\leqslant\exp\left(-\frac{\left(\varepsilon_2*\frac{OPT}{\sigma(S)}\right)^2}{2+\frac{2}{3}\left(\frac{OPT}{\sigma(S)}\right)}\right)*\theta*\frac{\sigma(S)}{n}$$

$$\leqslant\exp\left(-\frac{\varepsilon_2^2*OPT^2}{2\sigma(S)+\frac{2}{3}(\varepsilon_2*OPT)}\right)*\theta*\frac{1}{n}$$

$$\leqslant\exp\left(-\frac{\varepsilon_2^2*OPT^2}{2\left(1-\frac{1}{e}-\varepsilon\right)*OPT+\frac{2}{3}(\varepsilon_2*OPT)}\right)*\theta*\frac{1}{n}$$

$$\leqslant\exp\left(-\frac{\left(\varepsilon-\left(1-\frac{1}{e}\right)*\varepsilon_1\right)^2*OPT}{2-\frac{2}{e}}\right)*\frac{\left(2-\frac{2}{e}\right)\cdot n\cdot\ln\left(\binom{n}{k}\delta_2\right)}{OPT\cdot\left(\varepsilon-\left(1-\frac{1}{e}\right)\cdot\varepsilon_1\right)^2}*\frac{1}{n}$$

$$=\delta_2/\left(\frac{n}{k}\right)\blacksquare$$

根据联合界可得，对于任意满足$\theta\geqslant\theta_1$和$\theta\geqslant\theta_2$的θ，算法1选出的种子节点集合S以$1-\delta_1-\delta_2$的概率，满足$\sigma(S)\geqslant(1-1/e-\varepsilon)OPT$。在此基础上，本文的概率的参数设置采用与IMM算法[23]的相同的方式。对OPT折半猜测，并设置$\delta_1=\delta_2=\frac{1}{4n^l}$，$\varepsilon_1=\varepsilon\cdot\frac{\alpha}{\left(1-\frac{1}{e}\right)\cdot\alpha+\beta}$，其中，$\alpha$与$\beta$满足如下公式(7)：

$$\alpha = \sqrt{l\ln n + \ln 4},$$
$$\beta = \sqrt{\left(1 - \frac{1}{\mathrm{e}}\right)\cdot\left(\ln\binom{n}{k} + l\ln n + \ln 4\right)} \#(7)$$

在该参数设置下，可保证算法1选出的种子节点集合S以$1 - 1/2n^l$的概率满足与最优解至少$1 - 1/e - \varepsilon$的近似程度，此时θ_1与θ_2相等，为

$$\theta_1 = \theta_2 = \frac{\left(2n\cdot\left(\left(1 - \frac{1}{e}\right)\cdot\alpha + \beta\right)\right)^2}{OPT\cdot\varepsilon^2} \#(8)$$

4　实验

4.1　数据集

本文使用8个来自不同领域的真实超图数据集，这些数据集规模不等，且节点度、超边度等拓扑属性均有所不同，具体数据集信息如表1所示。

4.2　对比算法

MHPD-Heuristic[9]：一种基于启发式节点评估的影响力最大化算法，其核心思想是通过多跳传播深度（MHPD）方法计算每个节点在局部邻域内的预期传播范围。

HACE[18]：建模候选节点的有效扩散能力和被现有种子集接触的概率，通过折扣接触能力这一指标迭代选择k个种子节点。

HCLI[18]：建模候选节点的有效扩散能力和被所有邻居接触的概率，进行联合计算作为候选节点的边际收益指标。

表**1** 数据集的拓扑属性。n表示节点数，m表示超边数，deg表示节点平均一阶度数，d^H表示节点平均超边度，d^E表示超边包含平均节点数，c表示超图聚集系数，d表示平均最短路径，ε表示超图直径，ρ表示从超图所派生普通网络的边密度。

数据集	n	m	deg	d^H	d^E	c	d	ε	ρ
Algebra	423	1268	78.9	19.53	6.52	0.8	1.95	5	0.19
Email	143	1542	25.17	32.5	3.01	0.59	2.07	4	0.18
Geometry	580	1193	164.79	21.53	10.47	0.82	1.75	4	0.28
H-Com	1290	341	195.56	9.2	34.79	0.53	1.9	4	0.15
iAF1260b	1668	2351	13.26	5.46	3.87	0.56	2.67	7	0.01
NDC	1161	1088	10.72	5.55	5.92	0.61	3.5	9	0.01
Restaurant	565	601	79.75	8.14	7.66	0.54	1.98	5	0.14
S-Bills	294	29157	155.65	789.62	7.96	0.84	1.46	2	0.53

HEDRL-IM[20]：一种进化深度强化学习算法，通过DQN将超图IM问题转化为网络权重优化问题，结合EA的全局探索与DRL的局部利用优势，进而选择种子节点。关键参数均采取原论文设置，学习率为0.001，折扣因子为0.8，批次大小为64，种群大小为100，交叉概率和变异概率为0.8和0.2。

Hyper IMRANK （H-IMRANK）[17]：将IMRANK应用到超图中，仍然旨在通过迭代估算节点的边际影响，迭代地对节点排序直到节点排名收敛为止，获取最终稳定的排名序列中分值最高的k个节点。

Hyper Degree Heuristic Method （H-Degree）：基于超边度的启发式算法。计算各个节点的超边度并选取前k条超边度最大的节点。

Degree Heuristic Method（Degree）：基于度的启发式算法，首先计算所有节点的度值，然后选取前k个度最大的节点。

Hyper CELF（H-CELF）：将CELF算法[5]应用到超图中，通过动态迭代计算所有节点的边际影响力，直至选出影响力最大的k个种子节点。

为验证HIM-HRIS算法的有效性，本文分别在HIC-OT和HIC-MT模型下，计算模拟8个数据集下的各算法选取的k个种子节点的影响力范围。设置k从1到30，蒙特卡洛模拟次数为5000，采取两种概率设置方式，分别是**权重设置**：设置节点影响所属超边概率$p_{v\to e}$为$1/N(v)$，超边影响所包含的节点概率$p_{e\to v}$为$1/N(e)$；**常数设置**：设置$p_{v\to e}$和$p_{e\to v}$为某固定的常数。在S-Bills数据集上，受限于HEDRL-IM的高空间复杂度，在算法运行过程中发生内存溢出（实验环境为96GB内存），因此无对应实验结果。

实验表明，在iAF1260b、NDC和Restaurant数据集上，由于网络结构相对稀疏且传播范围有限，HIM-HRIS与其他对比算法之间的差距并不显著HEDRL-IM算法在S-Bills数据集上内存溢出，而HIM-HRIS仍能够稳定运行；MHPD-Heuristic算法由于仅考虑限定跳数内的节点的预期传播范围，而HIC-MT模型允许超边被多次激活，在种子数量较少时，对于节点的影响力评估误差被进一步拉大，种子质量较差；基于简单拓扑结构的算法反而由于超边可被重复激活提升了算法性能，但整体上HIM-HRIS通过精准的集合覆盖计算，在k较小时

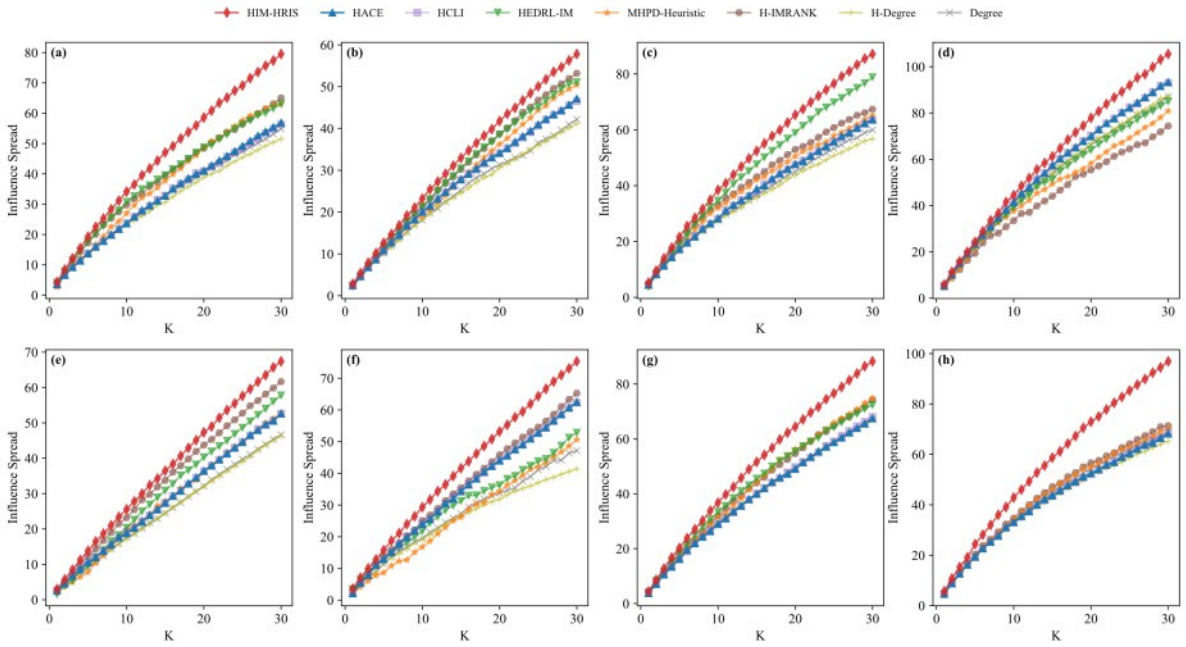

图6　HIC-OT模型下各算法的影响力范围（权重设置）。(a)Algebra; (b)Email; (c)Geometry; (d)H-Com; (e)iAF1260b; (f)NDC; (g)Restaurant; (h)S-Bills

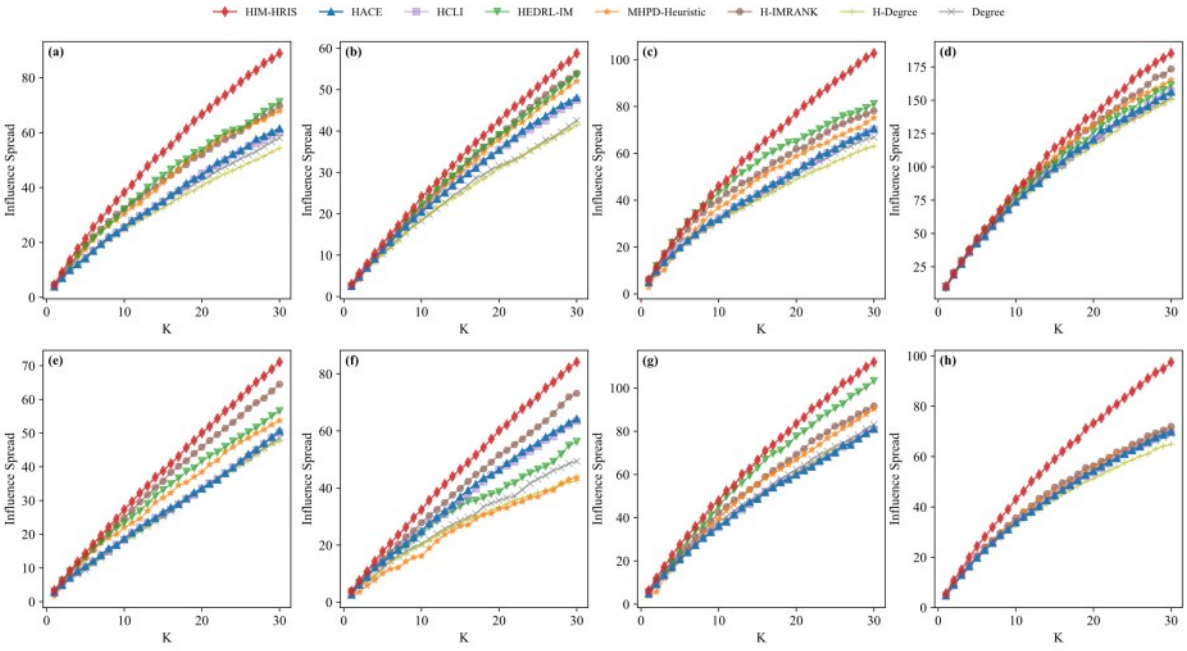

图7　HIC-MT模型下各算法影响力范围（权重设置）。(a)Algebra; (b)Email; (c)Geometry; (d)H-Com; (e)iAF1260b; (f)NDC; (g)Restaurant; (h)S-Bills

仍能够选取高质量的种子节点集合。尽管其在HIC-OT模型下性能略逊于HCLI，但在HIC-MT模型下能够获得影响力效果最优的种子集合。在其余数据集中，HIM-HRIS也均表现出明显优势。

总体而言，在不同数据集、不同传播策略以及不同的传播概率设置下，HIM-HRIS都能选择高质量种子节点集合，且在不同的数据集中表现稳定。

在运行效率方面，比较了在8个数据集上k设置为5，$p_{v\to e}$和$p_{e\to v}$均设置为0.01的情况下，HIM-HRIS算法与对比算法分别在HIC-OT和HIC-MT模型下的运行时间（单位：秒），实验结果（表2）表明HIM-HRIS的与运行效率明显优于H-CELF算法，在HIC-OT模型下8个数据集的平均加速比达210.03倍，在HIC-MT模型下8个数据集的

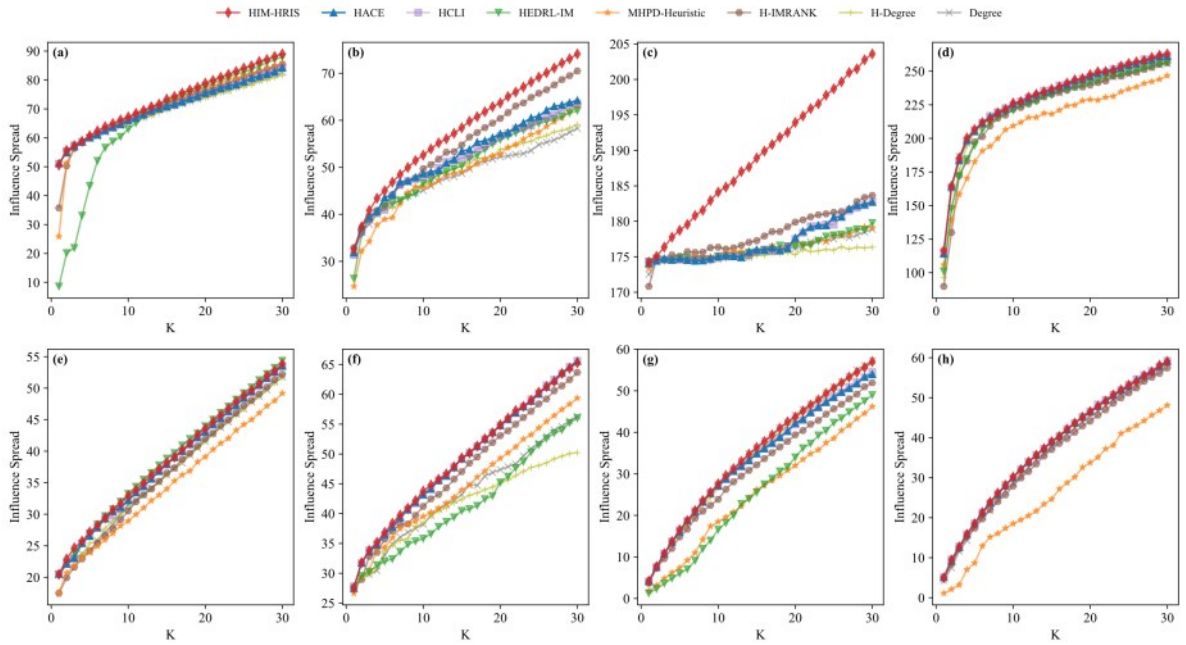

图 8　HIC-OT 模型下各算法影响力范围（常数设置）。(a)Algebra; (b)Email; (c)Geometry; (d)H-Com; (e)iAF1260b; (f)NDC; (g)Restaurant; (h)S-Bills

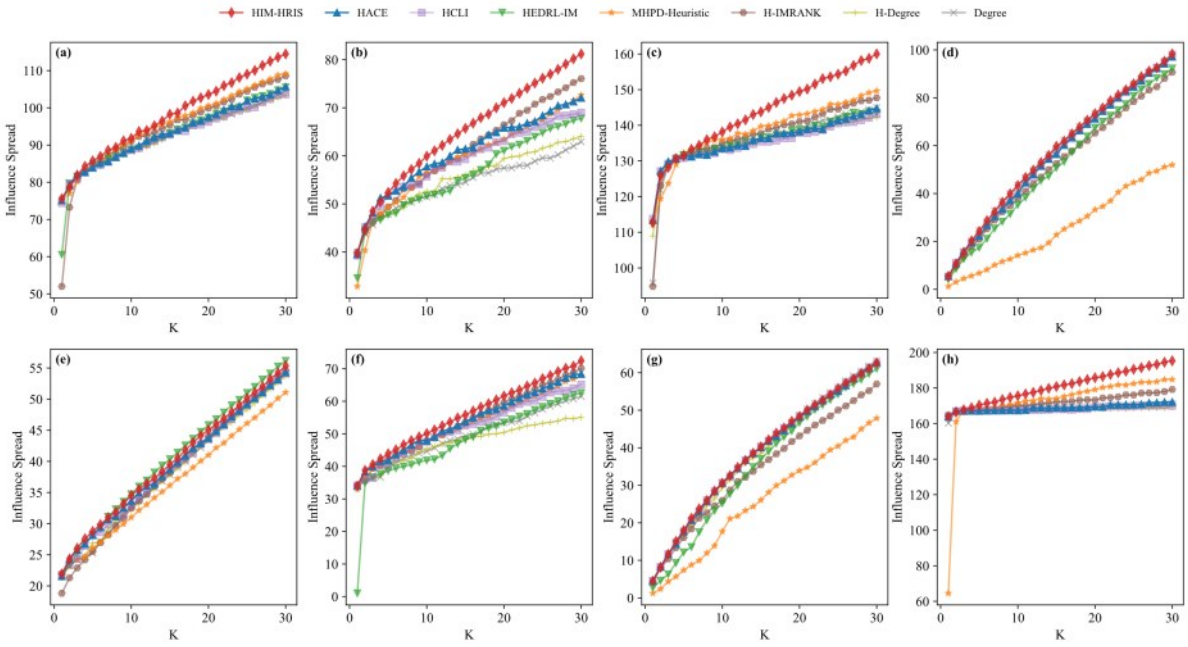

图 9　HIC-OT 模型下各算法影响力范围（常数设置）。(a)Algebra; (b)Email; (c)Geometry; (d)H-Com; (e)iAF1260b; (f)NDC; (g)Restaurant; (h)S-Bills

平均加速比达 427.15 倍。HIM-HRIS 算法在所有数据集上的运行速度均远快于 HEDRL-IM；与其他启发式算法相比，HIM-HRIS 的运行时间虽略高于 H-Degree、Degree 等基于简单拓扑的启发式方法，但其在传播影响范围上的性能优势显著，实现了效果与效率的良好权衡；同时相较于 HACE、HCLI、MHPD-Heuristic 等复杂启发式算法，HIM-HRIS 在多数数据集上仍保持相当或更优的运行效率，综合表现更为均衡稳定。

5　结束语

本文针对社交网络中信息的群体传播行为，系统研究了无向超图中的影响力最大化问题。通过深刻分析无向超图中的信息传播过程，本文提出了具有两种不同超边激活策略的无向超图独立级联模

表2　HIC-OT/HIC-MT模型下8个数据集上各算法（MHPD-Heuristic简写为MHPD）的运行时间，单位：秒。

数据集	HIM-HRIS	H-CELF	HEDRL-IM	HACE	HCLI	MHPD	H-IMRANK	H-Degree	Degree
Algebra	5.2/4.2	1675.9/1591.1	1247.3/1107.8	2.2/0.7	2.5/0.7	2.1/2.4	1.7/2.0	0.02/0.03	0.6/0.7
Email	1.3/1.1	161.9/276.2	773.1/519.2	0.3/0.1	0.4/0.1	0.2/0.2	0.3/0.3	0.01/0.01	0.3/0.3
Geometry	6.6/5.5	3251.8/3908.7	3528.3/1898.4	4.2/1.6	6.5/4.4	3.9/3.6	5.5/5.2	0.03/0.04	0.9/0.8
H-Com	18.3/10.2	4919.3/18430.5	2695.8/2402.2	7.5/4.3	18.8/3.9	20.3/23.4	5.1/4.4	0.06/0.07	1.0/0.9
iAF1260b	20.1/16.6	2871.3/3885.6	1288.3/1233.6	9.4/12.2	9.7/10.1	39.1/28.4	5.6/4.7	0.10/0.07	0.9/0.7
NDC	11.4/9.5	583.0/605.3	1712.6/1668.7	3.0/2.2	3.23/2.1	18.2/13.6	3.1/2.6	0.05/0.04	0.6/0.5
Restaurant	5.1/4.2	1074.1/1437.7	1476.9/1027.5	4.2/0.8	3.11/1.0	4.2/3.0	1.1/0.9	0.03/0.02	0.4/0.3
S-Bills	179.5/220.5	15114.6/10375.6	内存溢出	39.2/10.4	64.36/9.9	1.2/0.9	1620.6/1147.6	0.04/0.02	22.4/12.2

型，并揭示了其与普通独立级联模型的本质映射关系。突破了普通图仅能模拟节点间两两交互的局限，精准适配现实社交网络中“用户选择性转发至多群组”、“热点信息在群组中反复传播”、“通知类信息在群组中单次传播”等多样化场景。在此基础上，本文将反向影响采样技术拓展至超图，形成高阶反向影响采样方法；通过生成给定数量的超图反向可达集，将无向超图影响力最大化这一NP难问题转化为可高效求解的最大覆盖问题，从而设计了具有理论性能保证的HIM-HRIS算法，本文通过理论分析证明该算法能够达到与最优解$1 - 1/e - \varepsilon$的近似比。在8个真实网络数据集上的实验验证表明HIM-HRIS在传播效果方面优于现有方法，并在计算效率上明显优于同样具有理论保障的传统贪心算法，平均运行时间仅为其约0.7%，实现了求解效率与传播效果的双重优化，验证了本文所提模型与算法的有效性和实用性。

基于本文的研究成果，未来可从以下几个方向进一步深入探索：第一，拓展传播模型的适用场景，考虑节点异质性、信息时效性等现实因素，构建更具泛化性的超图传播模型，适配更复杂的社交网络环境；第二，结合新兴技术，探索HIC模型与大模型、强化学习等方法的融合路径，进一步提升算法的实用性与扩展性。

参考文献：

[1] 曾志林,张超群,吴国富,等.一种预测未知节点的融合影响力最大化的知识可迁移GNN模型[J]. 中文信息学报, 2025, 39(02): 89-99+110.
ZENG Z L, ZHANG C Q, WU G F, et al. A knowledge transferable GNN model integrating influence maximization for predicting unknown nodes[J]. Journal of Chinese Information Processing, 2025, 39(02): 89-99+110.

[2] 陈晋音,张敦杰,林翔,等.基于影响力最大化策略的抑制虚假消息传播的方法[J].计算机科学, 2020, 47(S1): 17-23+33.
CHEN J Y, ZHANG D J, LIN X, et al. False message propagation suppression based on influence maximization[J]. Computer Science, 2020, 47(S1): 17-23+33.

[3] RICHARDSON M, DOMINGOS P M. Mining knowledge-sharing sites for viral marketing[C]// Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002: 61-70.

[4] KEMPE D, KLEINBERG J. Maximizing the spread of influence through a social network[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM,2003:137-146.

[5] LESKOVEC J, KRAUSE A, GUESTRIN C, et al. Cost-effective outbreak detection in networks[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2007:420-429.

[6] GOYAL A, LU W, LAKSHMANAN L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]//International Conference Companion on World Wide Web, 2011: 47-48.

[7] BORGS C, BRAUTBAR M, CHAYES J, et al. Maximizing social influence in nearly optimal time[C]//In Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, 2014: 946-957.

[8] SUO Q, GUO J L, SHEN A Z. Information spreading dynamics in hypernetworks[J]. Physica A: Statistical Mechanics and its Applications, 2018, 495: 475–487.

[9] WANG H S, PAN Q T, TANG J. MHPD: An efficient evaluation method for influence maximization on hypergraphs[J]. Communications in Nonlinear Science and Numerical Simulation, 2024, 139, 102868.

[10] GANGAL V, RAVINDRAN B, NARAYANAM R. HEMI: Hyperedge majority influence maximization[C]//Proceedings of the 2nd International Workshop on Social Influence Analysis. 2016, 1622: 38-47.

[11] 王璿, 张瑜, 周军锋, 等. 基于社交网络的影响力最大化算法[J]. 通信学报, 2022,43(8):151-163.
WANG X, ZHANG Y, ZHOU J F, et al. Influence maximization algorithm based on social network[J]. Journal on Communications, 2022, 43(8): 151-163.

[12] 顾秋阳, 吴宝, 孙兆洋, 等. 基于改进灰狼优化的复杂网络重要节点识别算法[J]. 通信学报, 2021,42(6):72-83.
GU Q Y, WU B, SUN Z Y, et al. Key node identification algorithm for complex network based on improved grey wolf optimization[J]. Journal on Communications, 2021, 42(6): 72-83.

[13] XIE M, ZHAN X X, LIU C, et al. An efficient adaptive degree-based heuristic algorithm for influence maximization in hypergraphs[J]. Information Processing & Management, 2023, 60(2): 103161.

[14] GONG X L, WANG H C, WANG X Y, et al. Influence maximization on hypergraphs via multi-hop influence estimation[J]. Information Processing & Management, 2024, 61(2): 103683.

[15] 陈彬. 基于超图的社交网络中的影响力传播问题的算法研究[D]. 北京邮电大学, 2022.
CHEN B. Research on influence propagation algorithm in social networks based on hypergraph[D]. Beijing University of Posts and Telecommunications, 2022.

[16] LI S Y, LI X. Influence maximization in hypergraphs: A self-optimizing algorithm based on electrostatic field[J]. Chaos, Solitons and Fractals, 2023, 174: 113888.

[17] ATHUL M A, ARUN R. Hyper-IMRANK: Ranking-based influence maximization for hypergraphs[C]//CODS-COMAD, 2022, 100-104.

[18] WU L Y, LI C, QU B, et al. Adaptive overlap penalization and probabilistic modeling in hypergraph influence maximization[J]. Information Processing & Management, 2026, 63(4): 104594.

[19] 王志萍,赵嘉乐,刘凯,等.基于遗传算法的低冗余超图影响力最大化[J].复杂系统与复杂性科学,2025,22(02):97-104.
WANG Z P, ZHAO J L, LIU K, et al. Genetic algorithm-based low redundant hypergraph influence maximization[J]. Complex Systems and Complexity Science ,2025,22(02):97-104.

[20] XU L, MA L J, LIN Q Z, et al. Influence maximization in hypergraphs based on evolutionary deep reinforcement learning[J]. Information Sciences, 2025, 698: 121764.

[21] WU J, LI D. Modeling and maximizing information diffusion over hypergraphs based on deep reinforcement learning[J]. Physica A, 2023, 629: 129193.

[22] SUN Y H, WU J, SONG N A, et al. Deep reinforcement learning-based influence maximization for heterogeneous hypergraphs[J]. Physica A, 2025, 660: 130361.

[23] TANG Y, SHI Y, XIAO X. Influence maximization in near-linear time: A martingale approach[C]//In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), 2015: 1539 – 1554.

芮晓彬（1992– ），男，江苏徐州人，博士，中国矿业大学副教授，主要研究方向为社交网络分析、图数据挖掘和影响力最大化。

吉嘉欣（2003– ），女，江苏淮安人，中国矿业大学在读硕士研究生，主要研究方向为社交网络影响力最大化。

方强鹏（2000– ），男，江苏扬州人，中国矿业大学在读博士研究生，主要研究方向为社交网络影响力阻断最大化和公平性算法。

时纪龙（2000– ），男，山东济宁人，中国矿业大学在读硕士研究生，主要研究方向为社交网络影响力阻断最大化。

王志晓（1979– ），男，河南平顶山人，博士，中国矿业大学教授，主要研究方向为社交网络分析、图数据挖掘和图表示学习。