**DOI**：10.13718/j.cnki.xdzk.2023.08.006
佘美富，王逸伟，张建章，等. 超图环境下链路预测问题的探究 [J]. 西南大学学报(自然科学版)，2023，45(8)：61-75.

# 超图环境下链路预测问题的探究

佘美富[1]，　王逸伟[2]，　张建章[1]，　詹秀秀[1]，　刘闯[1]

1. 杭州师范大学 阿里巴巴复杂科学研究中心，杭州 311121；2. 合肥综合性国家科学中心数据空间研究院，合肥 230088

**摘要**：超图作为图的扩展，可以表示多种实体间的关系，使得其表达能力大大强于图，该优势吸引人们的关注并日益成为研究热点. 链路预测作为图数据挖掘中的常见任务，也在超图上扩展为超链路预测. 超链路预测通过已知超边或节点的属性来估计新超边出现的可能性，但是由于超边内节点数量的任意性，其可能的超边由 $O(n^2)$ 暴增至 $O(2^n)$，这大大增加了算法的复杂度. 本文使用下采样方法以减少候选超边集的大小，将图上的带重启的随机游走算法扩展到超图上. 还将图上的其他指标，如 CN、CE、Jaccard 等，扩展到超图进行比较. 结果表明，带重启的随机游走指标在精确率和召回率上要明显优于其他指标，并且观察到演化良好的超图其超边内部的联系强度随节点数的增加而增加，由此可知超链路预测的主要难点在于对小尺寸超边的预测.

**关　键　词**：超图；链路预测；超链路预测；带重启的随机游走；有限集合；算法

**中图分类号**：TP301.6；O158　　**文献标志码**：A　　开放科学(资源服务)标识码(OSID)：

**文 章 编 号**：1673－9868(2023)08－0061－15

# Exploration of Link Prediction Via Hypergraph

SHE Meifu[1]，　WANG Yiwei[2]，　ZHANG Jianzhang[1]，
ZHAN Xiuxiu[1]，　LIU Chuang[1]

1. *Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China;*
2. *Data Space Research Institute of Hefei Comprehensive National Science Center, Hefei 230088, China*

**Abstract**: As an extension of graphs, hypergraphs can represent relationships between multiple entities, making their expressive power much stronger than that of graphs, which has attracted attention and become a research hotspot. Link prediction, a common task in graph data mining, has also been extended as hyperlink prediction on hypergraphs. Hyperlink prediction estimates the likelihood of new hyperedges appearing based on known attributes of hyperedges or nodes, but due to the arbitrary number of nodes in hy-

收稿日期：2023－06－07
基金项目：浙江省自然科学基金项目(LQ22F030008).
作者简介：佘美富，硕士研究生，主要从事软件工程研究.
通信作者：刘闯，博士，教授，硕士研究生导师.

peredges, the possible hyperedges can exponentially increase from $O(n^2)$ to $O(2^n)$, greatly increasing the complexity of the algorithm. In this paper, we use a down-sampling method to reduce the size of the candidate hyperedge set and extend the graph's random walk with restart algorithm to hypergraphs. In addition, we extend other metrics on the graph, such as CN, CE, and Jaccard index, to hypergraphs. The results show that the metric is significantly better than other metrics in precision and recall, and we observe that the internal connectivity of hyperedges in well-evolved hypergraphs increases with the number of nodes, indicating that the main difficulty in hyperlinks prediction lies in predicting small hyperedges.

**Key words**: hypergraph; link prediction; hyperlink prediction; random walk with restart; finite set; algorithm

"关系"普遍存在于现实生活中的各实体之间，如人们之间的合作或社交、有机物之间的生化反应等. 数学上，人们习惯将上述各实体看作节点，实体间的关系看作节点对的连边，从而将其表达成统一而简洁的符号形式——图(亦称网络)，并辅之以严密的数学工具来分析和解决现有问题. 然而，相较于成对实体，现实中更多存在的是成群实体间的关系，如合作或社交小组、多个有机物共同参与的生化反应等. 边所连接节点的数量限制成为图表达能力的桎梏，使其只能展现成对实体之间的关系. 作为图的自然扩展，超图将边所连接的节点数量放松至任意，从而形成了不同于普通边的"超边". 这种结构上的突破使得超图能呈现任意数量的实体间关系，进而获得大大强于普通图的表达能力. 超图在表达能力上的优势吸引人们的关注并日益成为研究热点.

在构建图的过程中，由于观测手段、数据丢失等问题，不能保证实体间所有存在的关系都已被描述. 随着图的演化，节点之间会产生新的关系，预测缺失或未来即将出现的超边被称之为超图的链路预测，为了与普通图的链路预测问题进行区分，将其简称为"超链路预测". 由于关系的重要性，超链路预测日益成为人们的研究重点. 然而，超边大小的任意性导致人们在进行超链路预测时遇到了一些阻碍，首当其冲的便是候选超边数量的暴增. 具体来说，超边连接节点数量的放松导致候选超边数量的规模从 $O(n^2)$ 暴增至 $O(2^n)$，链路预测方法的输入节点数从固定值2变为任意值. 关于候选超边数量暴增的问题，对原始的候选超边集进行下采样是一个通用方法. Zhang 等[1]对原始的候选超边集按领域知识对各数据集进行采样，并定义了基于采样后的候选超边集的超链路预测问题. Patil 等[2]在完全随机采样和依超边大小分布采样的基础上，提出了基于 motif 和 clique 的抽样方法，并对这些方法所抽取的样本进行了可预测性研究. 另一种手段是根据启发式方法直接得到待预测超边的大小. Kumar 等[3]认为超边的产生应该遵循原始超图的结构规则，将原始超图的超边大小分布的采样值作为待预测超边的大小. Srinivasan 等[4]直接将大小的预测值嵌入生成对抗图中，利用对抗学习的机制来拟合最佳预测参数. 除候选超边数量暴增的问题外，超链路预测方法的输入节点数不固定也是一个棘手的问题，解决该问题的方式依赖于超链路预测方法本身的设计模式，偏重于机器学习、矩阵分解模型的方法通常会将待评分的候选超边编码成固定长度的特征向量，将其作为标准分类器的输入. 例如，Srinivasan 等[4]提出了一种新的特征聚合方法，该方法将节点与超边置于同等地位以进行对称式的卷积运算，证明了使用所提出方法进行特征学习，同一等价类的节点或超边之间有相同的特征表示，所学习到的特征在下游的链路预测任务上有着十分优异的表现. Yuan 等[5]提出一种基于张量的联合图嵌入方法，将成对链接和超链接同时编码到潜在空间上，该方法捕获并编码了高阶连接的依赖关系，并据此推断未观察到的超边. 而偏重于复杂图的方法则更加启发式地解决该问题，Benson 等[6]认为超边的相似性是所包含节点之间相似性的整体表现，提出了基于节点对计算候选超边相似性的一般方法. Kumar 等[3]则提出了一个迭代式算法，在每次迭代中进行寻找最优匹配节点并将所预测的超边大小作为迭代次数，最终输出符合预期的预测超边.

作为链路预测的延伸，人们会很自然地关注经典链路预测方法应用到超图上的可行性. 传统链路预测的理论已日趋成熟，诸如 Lorrain[7]、Ou[8]、Katz[9]、Liu[10]等的方法足以覆盖大部分场景. 在长时间的探

索后，人们普遍认为一个优秀的链路预测方法应具有这样的特征：在可解释性强、计算复杂度相对较低的情况下依然拥有较好的预测效果[11]. 依托成熟的链路预测理论，将链路预测方法应用到超图环境下是一个很自然的想法. 然而，图与超图之间巨大的拓扑结构差异使得将此方法移植至超图上时举步维艰. 近年来有极少数的研究基于经典链路预测方法的思想去开发对应的超链路预测版本. Pan 等[12]延伸了普通图的环思想，在超图上定义了节点环路与超边环路，对不同长度的环路进行加权求和作为逻辑回归的输入，从而对各候选超边进行概率预测. Srinivasan 等[4]受传统链路预测领域内 RA 算法的影响，提出了 HRA 算法以用于超链路预测.

基于上述研究，在探索的起点上，本文讨论了一个经典链路预测方法——带重启的随机游走(random walk with restart，简称 RWR)指标，如何扩展并应用于输入节点数可变的超链路预测场景中，以该方式为蓝本，扩展了其他传统链路预测指标，并以此作为基础方法. 引入了 9 个不同领域、不同规模的真实超图，并对这些超图的节点组进行抽样以生成部分基础方法的分数抽样分布，通过生成的抽样分布验证了所引入方法在解决超链路预测问题上的有效性，发现并解释了不同方法的分数分布在不同大小节点组上的差异. 然后，使用上述方法在所有真实超图上进行了标准的超链路预测实验，按照 Zhang 等[1]所提及的方式进行下采样以减少候选超边数量. 结果表明，带重启的随机游走指标在精确率和召回率上要明显优于其他指标，虽然没有一个方法在接收者操作特征曲线下面积(AUC)性能上对所有数据集表现一致优越，但各方法分组 AUC 的性能变化曲线却与对应的分数分布变化类似，且均呈现出随节点组大小递增的惊人一致性. 上述所有结果都暗示着大超边内部节点的连接强度要远远大于小超边.

## 1　问题定义

### 1.1　超图

给定超图 $H=\langle V, E\rangle$，其中 $V=\{v_1, \cdots, v_n\}$，是 $n$ 个节点组成的节点集，$E=\{e_1, \cdots, e_m\}$，是由 $m$ 个超边组成的超边集. 每一条超边 $e_j$ 由若干个节点组成，表示节点之间的相互作用，用 $|e_j|$ 表示超边 $e_j$ 的大小，即其所包含的节点数. $V$ 的所有节点组合构成了该节点集的幂集 $2^V$，取节点组 $e\in 2^V$，则 $e$ 的大小可取值 1，2，3，…，$n$. 因此有超图的超边集 $E\subset 2^V$. 数学上，使用关联矩阵 $\boldsymbol{B}=(b_{ij})_{n\times m}$ 表示超图，其行表示节点，列表示超边，当且仅当节点 $v_i$ 属于(亦称关联) 超边 $e_j$ 时，对应元素 $b_{ij}=1$，否则 $b_{ij}=0$.

### 1.2　超链路预测

对于上述超图，想象因“某种原因”使得部分超边丢失，导致其仅有部分超边能够被观察到. 则 $E$ 被分为已观察超边集 $E^o$(亦称正边集) 和遗失超边集 $E^{uo}$，且有 $E^{uo}=E\backslash E^o$，这些遗失超边与不能够形成超边的节点组(亦称负边集) 混杂在一起组成了候选集 $D$，记负边集合为 $E^n$，且有 $D=E^{uo}\cup E^n$，$\Phi=E^{uo}\cap E^n$. 基于此定义超图上的链路预测任务：利用正边集 $E^o$，从候选集 $D$ 中尽可能多地寻找到属于 $E^{uo}$ 的边. 该任务等价于一个二分类问题，在实施过程中，首先利用链路预测方法对 $D$ 中所有样本进行评分，根据评分结果设立一个阈值，将评分高于阈值的样本判定为正样本，否则为负样本.

### 1.3　下采样

这里，候选样本集 $E^s$ 的选取也值得讨论. 在超图链路预测的环境下，超边大小的任意性导致了候选集(所有未形成边的节点组)$2^V\backslash E^o$ 的大小呈 $O(2^{|V|})$ 上升，即所谓的“极端类不平稳问题”(extreme class imbalance，简称 ECI)[2]. 从大量的候选边中寻找各个遗失超边无异于大海捞针. 为此，需要对候选集进行下采样以减小推断空间的大小. 而下采样所采样出的候选样本需要符合“实际情况”，这里将“实际情况”理解为候选样本的大小分布应该与已观察超边集合的大小分布一致[2]. 为此，根据已观察超边集 $E^o$ 的超边大小分布，对候选样本空间 $2^V\backslash E^o$ 进行下采样以生成候选样本集 $E^s$，从而保证候选样本集 $E^s$ 的大小分布与 $E^o$ 一致.

## 2 方法

超链路预测问题考虑的是超图中一组节点是否能形成超边，因此首先研究节点对之间的相似性. 本文提出用超图上带重启的随机游走来定义节点对之间的相似性，还拓展了基于普通图的一些节点对相似性方法来定义超图中节点对的相似性. 基于节点对的相似性再刻画节点组的相似性，以此作为超链路的预测方法.

### 2.1 节点对的相似性

#### 2.1.1 带重启的随机游走指标

带重启的随机游走是链路预测里的一个经典方法，该方法基于图上的带重启随机游走，并将节点对之间的平稳互达概率作为链路预测分数[11]. 为将上述指标扩展到超图，需要先定义超图上的随机游走. 假设游走者在时刻 $t(t=0,1,\cdots)$ 位于节点 $v_i$，考虑一种无偏的情况，游走者首先随机游走到包含该节点的任意一条超边上，并在 $t+1$ 时刻随机游走到关联该超边的节点 $v_j$. 这里定义的超图随机游走可简单表述为上述过程的迭代. 该游走者在 $t$ 到 $t+1$ 期间，从节点 $v_i$ 转移到节点 $v_j$ 的概率为：

$$p_{ij}=\sum_{k=1}^{m}\frac{b_{ik}}{\sum\limits_{x=1}^{m}b_{ix}}\frac{b_{jk}}{\sum\limits_{y=1}^{n}b_{yk}} \tag{1}$$

式(1)计算了从 $v_i$ 到 $v_j$ 的单步转移概率，其前一项和后一项分别描述了“游走到超边”和“游走到节点”的过程. 由于超图中有 $n$ 个节点，相应地有 $n^2$ 个单步转移概率，将这些概率值组织成概率转移矩阵 $\boldsymbol{P}=(p_{ij})_{n\times n}$. 依式(1)可将该矩阵进行分解：

$$\boldsymbol{P}=\boldsymbol{D}_{vE}^{-1}\boldsymbol{B}\boldsymbol{D}_{eV}^{-1}\boldsymbol{B}^{\mathrm{T}} \tag{2}$$

$\boldsymbol{D}_{vE}\in\mathbb{R}^{n\times n}$，$\boldsymbol{D}_{eV}\in\mathbb{R}^{m\times m}$，均为对角阵，对角线元素分别为各节点所关联的超边数和各超边所包含的节点数.

定义状态向量 $\boldsymbol{\pi}_t^i\in\mathbb{R}^{n\times 1}$，其第 $j$ 个元素 $\boldsymbol{\pi}_t^i[j]$ 表示初始时游走者在节点 $v_i$，经过 $t$ 步随机游走后位于节点 $v_j$ 的概率. 显然，该向量表示了一个随机分布，满足 $\sum\limits_{k=1}^{n}\boldsymbol{\pi}_t^i[k]=1$. 由于初始时游走者具有确定的位置状态 $v_i$，初始状态向量 $\boldsymbol{\pi}_0^i$ 的第 $i$ 个元素为1，其余元素均为0. 由式(3)所述的状态转移方程描述了相邻时刻状态向量之间的关系：

$$\boldsymbol{\pi}_{t+1}^i=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\pi}_t^i \qquad t=0,1,\cdots \tag{3}$$

需要注意的是，并未限制在每一步随机游走的过程中游走者不能跳回上一步节点，这是为了使状态转移矩阵 $\boldsymbol{P}$ 能分解为式(2)所示的矩阵运算. 在此基础上添加重启机制，即在每步游走时，有一定可能跳回到初始节点 $v_i$，跳回概率为 $c(0<c<1)$. 则状态转移方程可调整为：

$$\boldsymbol{\pi}_{t+1}^i=(1-c)\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\pi}_t^i+c\boldsymbol{\pi}_0^i \qquad t=0,1,\cdots \tag{4}$$

当随机游走过程进入稳态时，其状态向量保持不变，记 $\boldsymbol{\pi}_\infty^i$ 为平稳状态向量，此时 $\boldsymbol{\pi}_{t+1}^i=\boldsymbol{\pi}_t^i=\boldsymbol{\pi}_\infty^i$，代入式(4)可解得：

$$\boldsymbol{\pi}_\infty^i=c[\boldsymbol{I}-(1-c)\boldsymbol{P}^{\mathrm{T}}]^{-1}\boldsymbol{\pi}_0^i \tag{5}$$

因此，可以定义节点对 $v_i$，$v_j$ 的平稳解 $RWR$ 分数：

$$RWR(v_i,v_j)=\boldsymbol{\pi}_\infty^i[j] \tag{6}$$

式(6)将随机游走者从初始节点 $v_i$ 出发并到达 $v_j$ 的平稳概率作为节点对的相似性分数，其大小表征了从 $v_i$ 到 $v_j$ 的可达性，内在编码了超图的拓扑结构. 很明显，该分数并不具有对称性，即 $RWR(v_i,v_j)\neq RWR(v_j,v_i)$.

2.1.2　基础方法扩展

受普通图链路预测方法和已有超图研究的启发，本文设计了一些超图上刻画节点对相似性的方法，以下列举这些基础方法基于节点对输入的定义.

1) 共同邻居指标(Common Neighbors，简称 CN). 共同邻居指标是传统链路预测中最简单也是最经典的指标. 该指标基于“共同邻居越多的节点对越相似”这一思想，计算公式为[13]：

$$CN(v_i, v_j) = |N_{v_i} \cap N_{v_j}| \tag{7}$$

式中 $N_{v_i}$ 是节点 $v_i$ 的邻居集合，在超图中，它是指与节点 $v_i$ 共存于同一超边内的其他所有节点.

2) 共存超边指标(Coexist Edge，简称 CE). 不同于普通图，在超图环境中，由于边大小的任意性，两个节点往往会被多个超边同时包含. 统计包含节点对的超边个数便得到共存超边指标，其计算公式为：

$$CE(v_i, v_j) = |E_{v_i} \cap E_{v_j}| \tag{8}$$

式中 $E_{v_i}$ 是包含节点 $v_i$ 的超边集合.

3) Jaccard 指标(简称 JC). Jaccard 指标[14] 最初用来计算两集合之间的相似性. 在超图中，用该指标计算节点对之间的相似性定义如下：

$$JC(v_i, v_j) = \frac{|E_{v_i} \cap E_{v_j}|}{|E_{v_i} \cup E_{v_j}|} \tag{9}$$

4) Adamic-Adar 指标(简称 AA). Adamic-Adar 指标[15] 在 CN 的基础上考虑了共同邻居的权重，权重是共同邻居的邻居数对数的倒数. 在超图中，对于节点，可以将共存于同一超边的其他节点看作“点邻居”，将节点所关联的超边看作“边邻居”. 对于超边，可以将所包含的节点看作点邻居，将具有共同关联节点(即相交) 的超边看作边邻居. 由于邻居可分为点邻居和边邻居，点邻居、边邻居的不同组合对应 4 种不同的 AA 指标，分别为 $VV_{\mathrm{AA}}$、$VE_{\mathrm{AA}}$、$EV_{\mathrm{AA}}$、$EE_{\mathrm{AA}}$. 计算公式如下：

$$\begin{aligned}
VV_{\mathrm{AA}}(v_i, v_j) &= \sum_{z \in N_{v_i} \cap N_{v_j}} \frac{1}{\ln k_{vV}(z)} \\
VE_{\mathrm{AA}}(v_i, v_j) &= \sum_{z \in N_{v_i} \cap N_{v_j}} \frac{1}{\ln k_{vE}(z)} \\
EV_{\mathrm{AA}}(v_i, v_j) &= \sum_{z \in E_{v_i} \cap E_{v_j}} \frac{1}{\ln k_{eV}(z)} \\
EE_{\mathrm{AA}}(v_i, v_j) &= \sum_{z \in E_{v_i} \cap E_{v_j}} \frac{1}{\ln k_{eE}(z)}
\end{aligned} \tag{10}$$

指标名称的第一个和第二个字母分别代表共同邻居和共同邻居的邻居的性质. 例如，$VE_{\mathrm{AA}}$ 表示节点对的共同邻居为“点邻居”，而共同邻居的邻居为“边邻居”，也即计算与节点对均为邻居的节点权重，权重由该邻居所关联的超边数量计算所得. $k_{vV}(z)$、$k_{vE}(z)$ 统计了节点 $z$ 的邻居节点个数和关联超边个数，而 $k_{eV}(z)$、$k_{eE}(z)$ 则统计了超边 $z$ 所包含的节点个数及超边 $z$ 所相交的超边个数，一般将 $k_{vE}$ 和 $k_{eV}$ 看作节点和超边的度[16].

5) $\alpha$ 阶局部游走指标. $\alpha$ 阶局部游走指标定义在游走的基础之上，将节点对之间 $\alpha$ 阶游走的数量作为分数. 在普通图中，形如 $v_{i_0}, v_{i_1}, \cdots, v_{i_{\alpha-1}}$ 这样，节点个数为 $\alpha$ 且相邻节点在图中互为邻居的节点序列被称之为一次 $\alpha$ 阶游走. 扩展到超图上[17]，超图上的一次 $\alpha$ 阶游走可定义为一个节点、超边交替出现的序列：

$$v_{i_0}, e_{i_0}, v_{i_1}, e_{i_1}, \cdots, e_{i_{\alpha-1}}, v_{i_\alpha} \tag{11}$$

其长度等于 $\alpha$，即超边出现的次数. 同样，序列中的相邻元素在对应超图中相互关联. 节点对之间的 $\alpha$ 阶游走数量，即 $\alpha$ 阶局部游走指标可表示为：

$$WK_{\alpha}(v_i, v_j) = (\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})^{\alpha}_{ij} \tag{12}$$

### 2.2 节点组的相似性

上述所提及方法均只计算了超图内节点对的相似性，现在将其适用范围扩展至节点组. 给定超链路预测方法 $f$ 及超图 $H=\langle V, E\rangle$，对于节点对 $v_i, v_j \in V$，其输出分数为 $f_H(v_i, v_j)$，记节点组 $\tilde{e} \in 2^V \backslash V$，则对应的超链路预测分数由式(13)给出：

$$f_H(\tilde{e})=\frac{1}{\binom{|\tilde{e}|}{2}} \sum_{\substack{v_i, v_j \in \tilde{e} \\ i \neq j}} f_H(v_i, v_j) \tag{13}$$

该式首先计算了每个节点对的连接强度，并将所有节点对的连接强度的算术平均值作为边缘密度. 例如，$\tilde{e}$ 的 $RWR$ 分数可计算为：

$$RWR_H(\tilde{e})=\frac{1}{\binom{|\tilde{e}|}{2}} \sum_{\substack{v_i, v_j \in \tilde{e} \\ i \neq j}} RWR_H(v_i, v_j) \tag{14}$$

在社交领域，式(14)所蕴含的思想直观且有效：若一组成员之间两两关系亲密，那么他们更有可能组成一个团体.

## 3 实验

### 3.1 数据集

本文引入了9个真实超图，其领域覆盖了食谱、生物、合作、社交等方面. 各数据集的详细统计信息见表1. 表1由真实数据构建的超图拓扑结构性质，其中 $n$，$m$，$\langle k_{vE}\rangle$，$\langle k_{vV}\rangle$，$\langle|e|\rangle$，$d(H)$，$|E_2|$，$|E_3|$，$|E_4|$，$|E_5|$，分别表示超图的节点数，超边数，节点超度的平均值，节点一阶邻居数的平均值，超边大小的平均值，密度[18]，超边大小分别为2,3,4,5的超边数量.

**表1 各数据集的详细统计信息和超图拓扑结构性质**

| 数据集 | chuancai | iAB_RBC_283 | iJO1366 | CoreComplex | Cora-Co-citation | Cora-Co-reference | DBLP-Co-authorship | cat-edge-music-blues-reviews | email-Eu |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | 438 | 342 | 1 805 | 2 314 | 1 330 | 1 961 | 4 695 | 1 112 | 1 005 |
| $m$ | 835 | 465 | 2 546 | 1 338 | 1 503 | 875 | 2 561 | 694 | 25 791 |
| $\langle k_{vE}\rangle$ | 3.40 | 4.83 | 5.55 | 2.85 | 3.46 | 2.35 | 3.06 | 9.44 | 87.83 |
| $\langle k_{vV}\rangle$ | 5.42 | 10.09 | 16.92 | 29.51 | 6.23 | 37.4 | 11.72 | 166.97 | 58.31 |
| $\langle\|e\|\rangle$ | 1.79 | 3.55 | 3.94 | 4.93 | 3.06 | 5.26 | 5.62 | 15.13 | 3.42 |
| $d(H)$ | 0.012 | 0.030 | 0.009 | 0.013 | 0.005 | 0.019 | 0.002 | 0.15 | 0.058 |
| $\|E_2\|$ | 248 | 63 | 430 | 361 | 572 | 309 | 179 | 86 | 13 188 |
| $\|E_3\|$ | 71 | 16 | 129 | 411 | 447 | 169 | 331 | 52 | 5 080 |
| $\|E_4\|$ | 28 | 109 | 674 | 158 | 306 | 98 | 447 | 39 | 2 357 |
| $\|E_5\|$ | 9 | 149 | 581 | 95 | 178 | 90 | 411 | 33 | 1 401 |
| 领域 | 食谱 | 生物 | 生物 | 生物 | 引文 | 引文 | 合作 | 兴趣 | 社交 |

1) chuancai[1]：菜谱数据集，节点由食材构成，一道川菜所需的一组食材构成一条超边.

2) iAB_RBC_283，iJO1366[1]：人类、大肠杆菌代谢反应数据集. 超边表示生物体内的代谢反应，边内的节点表示参与该反应的反应物.

3) CoreComplex：本文构建了人类蛋白质相互作用数据集. 超边代表人体内的蛋白质相互作用，其包含的节点表示参与相互作用的蛋白质.

4) Cora-Co-citation，Cora-Co-reference[19]：这两个是引文数据集，两者的节点均表示机器学习论文. 在Cora-Co-citation超图中，超边连接了被同一篇论文所引用的所有论文. 而在Cora-Co-reference中，引用

了同一篇论文的论文组成一条超边.

5) DBLP-Co-authorship[20]：这是一个计算机领域的论文合作数据集. 节点表示论文作者，超边连接了同一篇文章的所有作者.

6) cat-edge-music-blues-reviews[21]：这是一个共评论数据集. 节点是亚马逊购物网上的用户，在一个月内评论过相同布鲁斯音乐产品的用户组成一条超边.

7) email-Eu[6, 22-23]：线上社交数据集. 节点表示欧洲科研机构的电子邮箱，超边是一次邮件发送，由邮件发送者和所有接收者所组成. 原始数据集中，email-Eu 图的超边带有时间戳，这里将多个带时间戳的超边合并成一个超边以静态化原始数据集.

### 3.2　方法有效性验证与超图结构探索

算法 1：指标分数抽样算法

输入：超图：$H=\langle V, E\rangle$，指标：$f$，节点组大小：$b$

输出：任意节点组分数样本集：$S_{arb}$，超边节点组分数样本集：$S_{edge}$

```
S_arb, S_edge ← [], [];
//获得指定大小的超边子集
E_b ← E.get_subset(b);
for e in E_b do
    //从节点集中不放回抽取 |e| 个节点
    ẽ = V.random_choice(|e|);
    S_arb.append(f_H(ẽ));
    H' ← ⟨V, E\e⟩;
    S_edge.append(f_H'(e));
end do
```

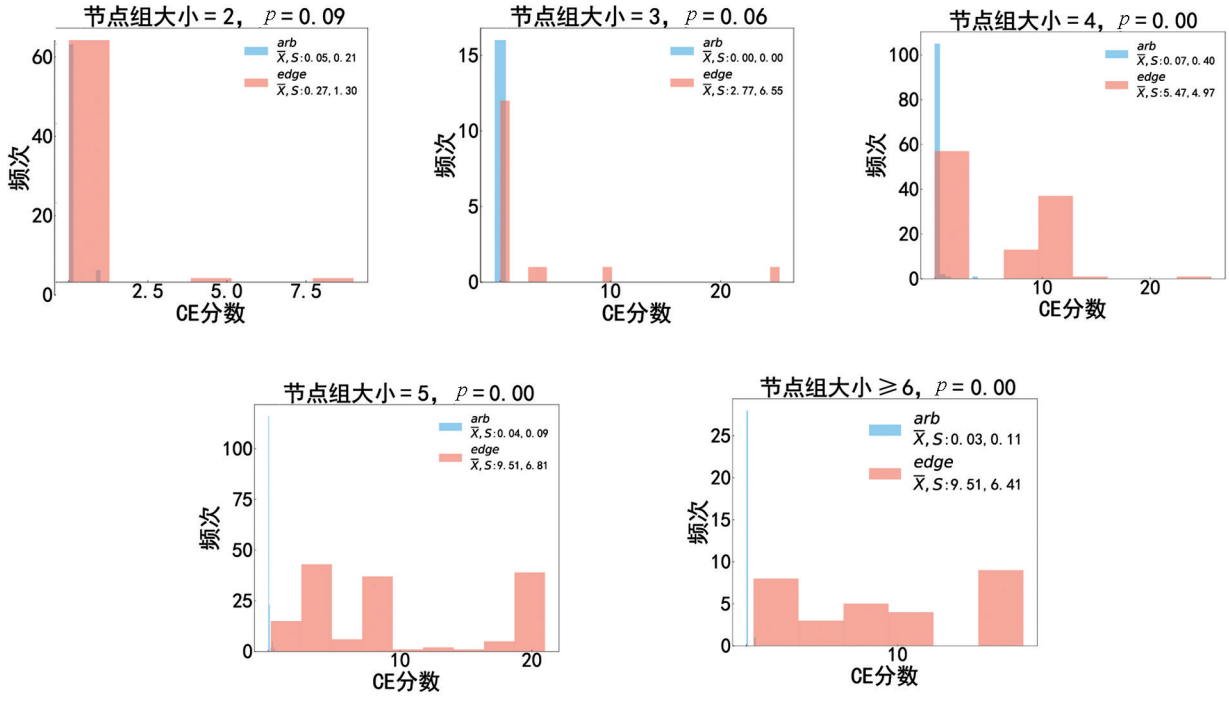
Result：$S_{arb}$，$S_{edge}$

如何验证第 2 章所列举的链路预测方法在超图环境下的有效性，即指标 $f$ 是否有能力将有潜力形成超边的节点组和普通节点组区分开来. 一个自然的想法是：随机选取一条超边，同时从超图节点集中随机选取一个相同大小且无重复元素的节点组，前者的 $f$ 指标分数大于后者的概率越大，指标 $f$ 就有效.

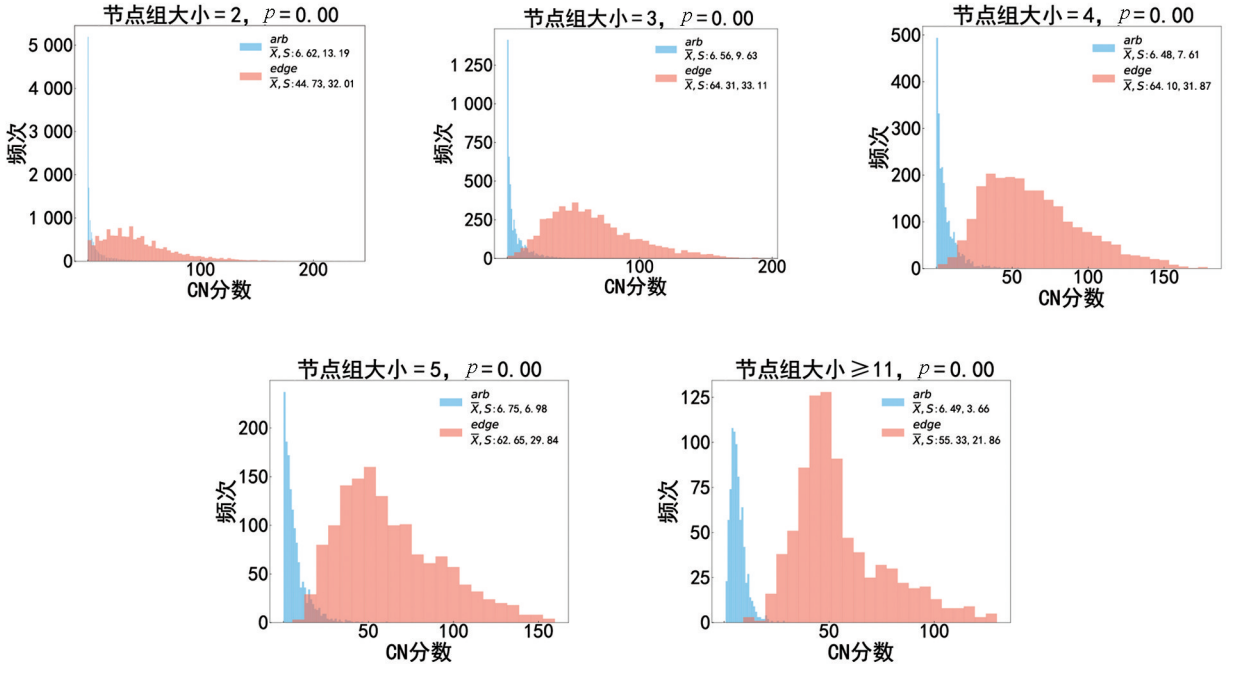
基于上述想法，分别对所扩展的 CN、CE 和 $WK_2$ 方法进行抽样分析. 具体来说，对于任意一个真实图 $H=\langle V, E\rangle$，首先将 $2^V\backslash V$ 内(不考虑大小为 1 的超边)所有节点组分为两大类：超边节点组和任意节点组. 顾名思义，一个超边节点组 $e\in E$，其本身构成了一条超边. 而任意节点组 $\tilde{e}\in 2^V\backslash V$ 取自超图所有可能的超边集合. 利用抽样算法分别得到超图的超边节点组和任意节点组的分数样本集合，并利用这些样本生成超边节点组和任意节点组在超图上的指标分数分布，最后比较两组分布和统计特征的差异性. 考虑到超图上的高阶特性，进一步细化抽样工作：抽样算法 1 以节点组大小为区分分开进行抽样，从而可生成不同大小节点组的分数分布. 注意到在对超边节点组进行抽样时，算法 1 首先将该节点组形成的超边所移除以模拟该超边形成前的环境.

#### 3.2.1　分数分布分析

由于空间有限，图 1 仅展示部分数据集的部分方法分数分布，还展示了分布的样本均值 $\overline{X}$、样本标准差 $S$ 以及两节点组分布 $t$ 检验的 $p$ 值. 由于各数据集中超边数量总体呈“小超边多、大超边少”的幂律分布趋势，在生成分数分布时，对小尺寸的节点组单独进行统计，而对大尺寸节点组进行联合统计. 同时绘制了各分数分布的误差棒图，图 2 进行了部分展示. 实验结果表明 CN 及其衍生方法(VVAA、VEAA)、CE 及其衍生方法(EVAA，EEAA，JC)、$WK_2$ 和 $WK_3$ 的分布形状较为类似，其差别在于取值范围及精度，以下着重分析可解释性较强的 CN、CE 和 WK 方法的分数分布.




a. iAB_RBC_283的CE分数分布




b. email-Eu的CN分数分布

红色为超边节点组，蓝色为任意节点组.

**图1 部分抽样分布**

各数据集任意节点组的分数分布都极其类似，其分数严重集中在低分值，且均值方差几乎没有变化，本文将其作为对照组来观察同尺寸的超边节点组分数分布. 在生物代谢图上，超边节点组所有分数抽样分布的变化都十分统一：开始时节点组的分数基本集中于低分值上，随着节点数的增加，其样本分数的取值范围逐渐增大且更加均匀地分布于各个区间. 观察iAB_RBC_283的CE分数分布(图1a)，该数据集的节点组CE分数表示节点组内各反应物平均共同参与的反应数. 可以看到，2,3-超边节点组的分数并未显著

大于同尺寸任意节点组($p>0.05$)，而 4-超边节点组相当一部分超边节点组的分数落在了 10 左右，当节点组大小为 5 时，超边节点组的分数更均匀地落在了大分数上，这表明多反应物的代谢反应显著依赖于反应物两两之间的可反应程度，且依赖程度随代谢反应规模的增加而增加. 蛋白质代谢图 CoreComplex 的各分数分布也有类似变化，但并未表现得像前两者显著.

cat-edge-music-blues-reviews 和 DBLP-Co-authorship 图的超边分别以“兴趣”和“合作”关系将代表各用户的相关节点相连. 这两个图的构成具有一定的相似性，即由代表人的节点因为共同事件所联系到一起，CE、CN 分别表征了人们之间共同参与事件及共同参与者的数量. 而两者超边节点组分数分布也比较类似. 以 cat-edge-music-blues-reviews 图为例，超边节点组在 CN 与 CE 上的分数分布表明：节点组增大时，虽然人们之间共同购买的产品并未显著变化，但用 CN 所统计的用户的共同兴趣者却大大增加. 对于邮件收发图 email-Eu，除 2-超边节点组外，email-Eu 图各方法分数分布在其余大小的节点组上几乎没有差别，即均表现为均值类似的正态分布(图 1b). 且各方法分数分布的均值总体先上升后保持平稳(略有上升或下降)，造成这种现象的一个解释是“朋友在工作中产生”，工作环境下的多次群发促成了日后代表亲密关系的“一对一沟通”.

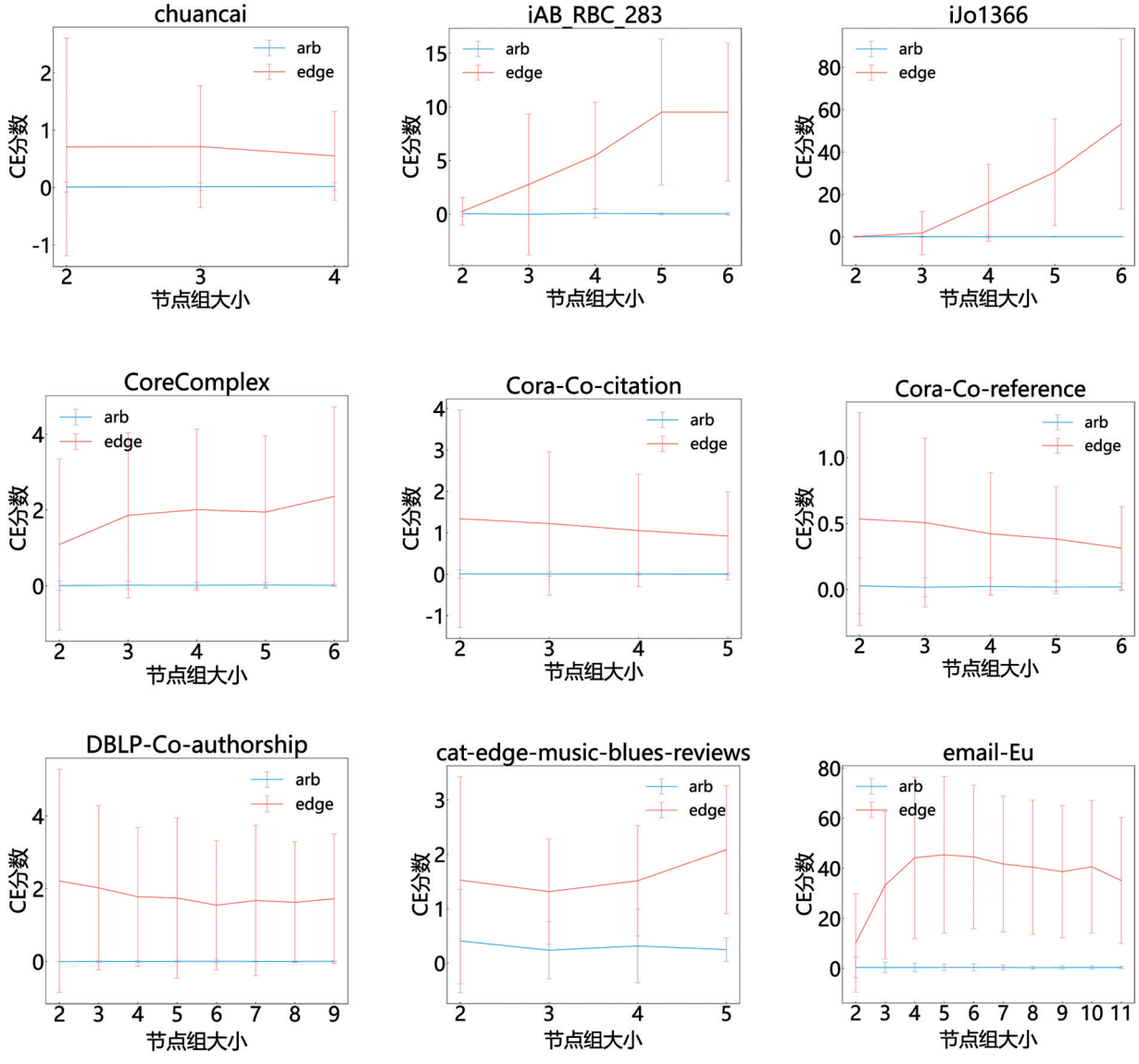



图 2　各数据集 CE 分数分布的误差棒图

3.2.2 结构分析

总体上看，由于超边形成机理相同，同一领域数据集的超边节点组分数分布相似. 而不同数据集超边节点组的各方法分数抽样分布与均值方差变化呈现两种趋势，要么没有明显变化，要么初始时显著上升. 这说明两点：第一，各方法以多视角衡量了节点组之间节点对的亲密程度. 第二，超图的“派系闭合假说”认为，大关系在形成前，其内部节点之间的联系已经非常紧密. 基于此，演化良好的超图，大尺度超边节点组内部节点对的亲密程度必然大于2-超边节点组. 如若不然，超边节点组内的各个亲密节点对更倾向于生成各种细密的2-超边而没有继续扩张的趋势. 甚至于在大部分数据集中，2-超边节点组和2-任意节点组之间分数分布几乎没有差异. 一般认为，方法对超边节点组和任意节点组的评分差异越大，该方法预测能力越强. 上述情况会使得各方法很难将大小为2的超边节点组和其他节点组区分开来从而导致误判. 例如，两个引文图Cora-Co-citation和Cora-Co-reference的所有方法的超边节点组分数抽样分布几乎不受节点组大小的影响，由于没有明显的演化规律和超边节点组与任意节点组之间的分数差异不明显，所提出的方法对这两个图的链路预测效果将不会太理想. 总的来说，通过抽样及之后的分析，在验证了已抽样方法的有效性——即它们的确能区分正负样本的同时，还发现在大多数情况下，这种区分能力和节点组的大小有关，即对大尺寸节点组的区分能力要远远强于小尺寸节点组. 由于真实超图的超边大小分布往往展现为幂律，即小尺寸超边占绝大多数，因此，对小尺寸、甚至于2-节点组的无力是制约超链路预测方法性能的主要桎梏.

### 3.3 超链路预测

本文使用经典的二分类全局指标——接收者操作特征曲线下面积(AUC)，与局部指标——精确率($precision@k$)和召回率($recall@k$)，来评估第2节所列举方法在各真实图上的表现. 对单个数据集及特定指标的每次实验，采用五折交叉验证的方式进行并重复若干次，记录每次实验的交叉验证平均值. 参数设置上，RWR的游走重启概率$c$固定为0.15.

3.3.1 预测结果分析

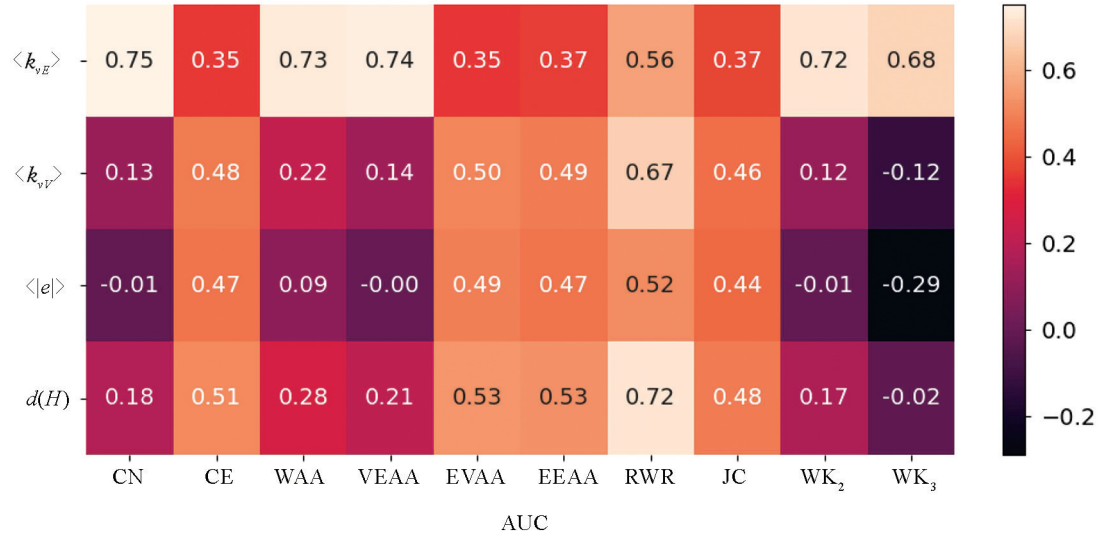
预测结果的AUC评价展示在表2，可以看到，并未有任何一种方法在所有数据集上表现得最好，这便是超链路预测环境下的“天下没有午餐定律”[24]. 令人惊喜的莫过于CE，该方法定义十分简单，却在iAB_RBC_283、iJO1366、CoreComplex、Cora-Co-reference、DBLP-Co-authorship上表现优异. 将CE应用于这些数据集还能得到更加直观的物理解释，例如，对于合作图DBLP-Co-authorship，以CE方法的视角来看，当一群作者相互之间合作过很多次时，他们更有可能组团进行合作. 在数据集chuancai和email-Eu上，CN方法有着不错的性能表现. 以email-Eu为例，CE方法计算了节点组内节点对之间的平均邮件收发事件数，而CN方法计算了各自参与的邮件收发事件的其他共同参与者数量，这表明，有着更多工作伙伴关系的成员组之间更有可能产生邮件收发事件. 在普通图的链路预测中，奇数步局部游走指标在代谢图上效果良好[25]，如果将CE看作“$WK_1$”，其效果在生物数据集上表现最优也就不足为奇了. JC作为CE的加权，整体上前者的表现不如后者，这也说明了链路预测方法的准确度并非与其计算复杂度呈正比. 对于在超图环境中所扩展的4类AA方法，若分别将VVAA、VEAA看作CN的改进，EVAA、EEAA看作CE的改进，则这两组方法内部之间的性能表现都十分接近. 且将点邻居个数看作共同邻居

或共存超边的权重时($VV_{AA}$ 之于 CN、$EV_{AA}$ 之于 CE)，整体有着更为明显的性能提升，这可能是由于点邻居的数量相对于边邻居更多，致使其权重的粒度更加细致从而有着很好的区分度. RWR 在各数据集上的 AUC 表现似乎与邻域方法、特别是 CE 的表现正相反，如前者在 chuancai、cat-edge-music-blues-reviews、email-Eu 上具有较优秀的预测效果，而后者在除这些以外的其他数据集上有着较好的表现，这表明在超链路预测任务中，节点间的全局特征与局部特征都能提供有效信息. 同时本文对某一方法作用于所有数据集上的链路预测性能(AUC 分数)与所有数据集的某一特定统计特征进行了相关性计算，热力图见图 3. 总体上看，超图的“规模”越大，其可预测性越好，注意到$\langle k_{vE}\rangle$，即所谓超图的平均超度[26]与方法的预测性能之间呈显著正相关，这意味着该指标的大小直接展示了数据集可利用信息的多少.

**表 2　各方法在不同数据集上的 AUC 对比**

| Datasets | chuancai | iAB_RBC_283 | iJO1366 | CoreComplex | Cora-Co-citation | Cora-Co-reference | DBLP-Co-authorship | cat-edge-music-blues-reviews | email-Eu |
|---|---|---|---|---|---|---|---|---|---|
| CE | 0.650 | 0.887 | **0.870** | 0.828 | 0.645 | 0.704 | 0.832 | 0.911 | 0.890 |
| CN | 0.684 | 0.802 | 0.796 | 0.789 | 0.678 | 0.650 | 0.826 | 0.757 | 0.956 |
| JC | 0.648 | 0.874 | 0.867 | 0.825 | 0.645 | 0.706 | 0.831 | 0.893 | 0.886 |
| VVAA | 0.685 | 0.821 | 0.805 | 0.803 | **0.680** | 0.663 | 0.829 | 0.794 | 0.959 |
| VEAA | 0.679 | 0.823 | 0.806 | 0.772 | 0.677 | 0.642 | 0.820 | 0.764 | 0.959 |
| EVAA | 0.650 | **0.888** | **0.870** | **0.829** | 0.645 | **0.707** | **0.833** | 0.920 | 0.891 |
| EEAA | 0.647 | **0.888** | **0.870** | 0.811 | 0.639 | 0.686 | 0.826 | 0.914 | 0.890 |
| RWR | 0.730 | 0.826 | 0.806 | 0.790 | 0.569 | 0.674 | 0.771 | **0.972** | **0.965** |
| $WK_2$ | 0.686 | 0.800 | 0.801 | 0.790 | 0.677 | 0.651 | 0.822 | 0.756 | 0.936 |
| $WK_3$ | 0.741 | 0.778 | 0.790 | 0.735 | 0.630 | 0.577 | 0.772 | 0.670 | 0.909 |

注：表 2 中加粗数据表示对各数据集最佳方法的突出显示.




横坐标为方法，纵坐标为统计特征，该图表征了真实数据集的统计特征与方法预测性能的相关性.

**图 3　相关性热力图**

预测结果的精确率及召回率曲线分别展示在图4、图5，两者的阈值上限取各数据集测试集样本数量的一半. 不同于在AUC表现上的“百家争鸣”，精确率及召回率曲线上扩展的RWR方法在所有数据集上均取得了不错的效果，说明所引入超图随机游走方式的有效性，即该过程确实能够捕捉到超图的拓扑结构. 局部游走指标($WK_2$、$WK_3$)效果普遍较差，两者的效果随着阈值$k$的增加而快速下降，这可能是因为在扩展局部游走时并未考虑“游走宽度”[17]，从而丢失了超边的“高阶特性”. 总的来说，各方法虽然在性能上有所差异，但正如本章第2节所言，所扩展的方法在应用于超链路环境中时，都能够对正负样本有效地进行预测.

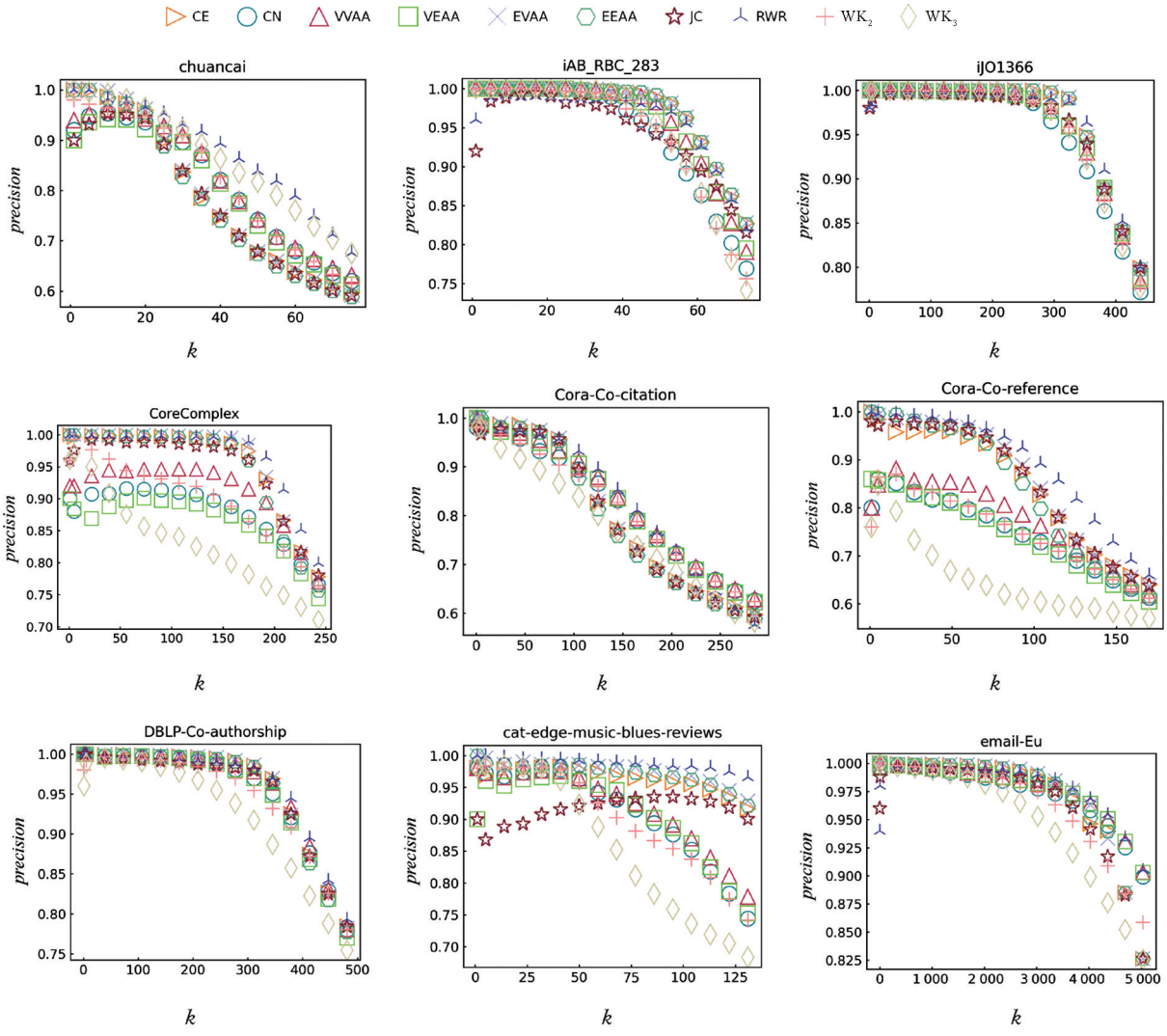



**图4 精确率曲线**

### 3.3.2 分组评估

延续抽样时的方式，在链路预测完成后，对各方法应用于各数据集的表现分节点组大小进行评估，评估结果见图6. 与抽样类似的是，各方法2-节点组上的链路预测效果明显低于其他节点组. 例如，对于iAB_RBC_283，2-超边节点组的CE、CN分数分布并未显著大于任意节点组，相对应，很多方法对该数据集2-节点组进行链路预测的AUC效果甚至低于0.5. 对于大尺寸节点组，各方法在整体上表现一致. 这一现象不仅与抽样结果相吻合，还从链路预测角度体现了大超边内部节点的强链接程度. 与抽样不同的是，AUC的效果评估曲线在绝大部分数据集上呈现绝对递增的趋势，即随着节点组节点数的增加，各方法在数据集上的链路预测效果越来越好.

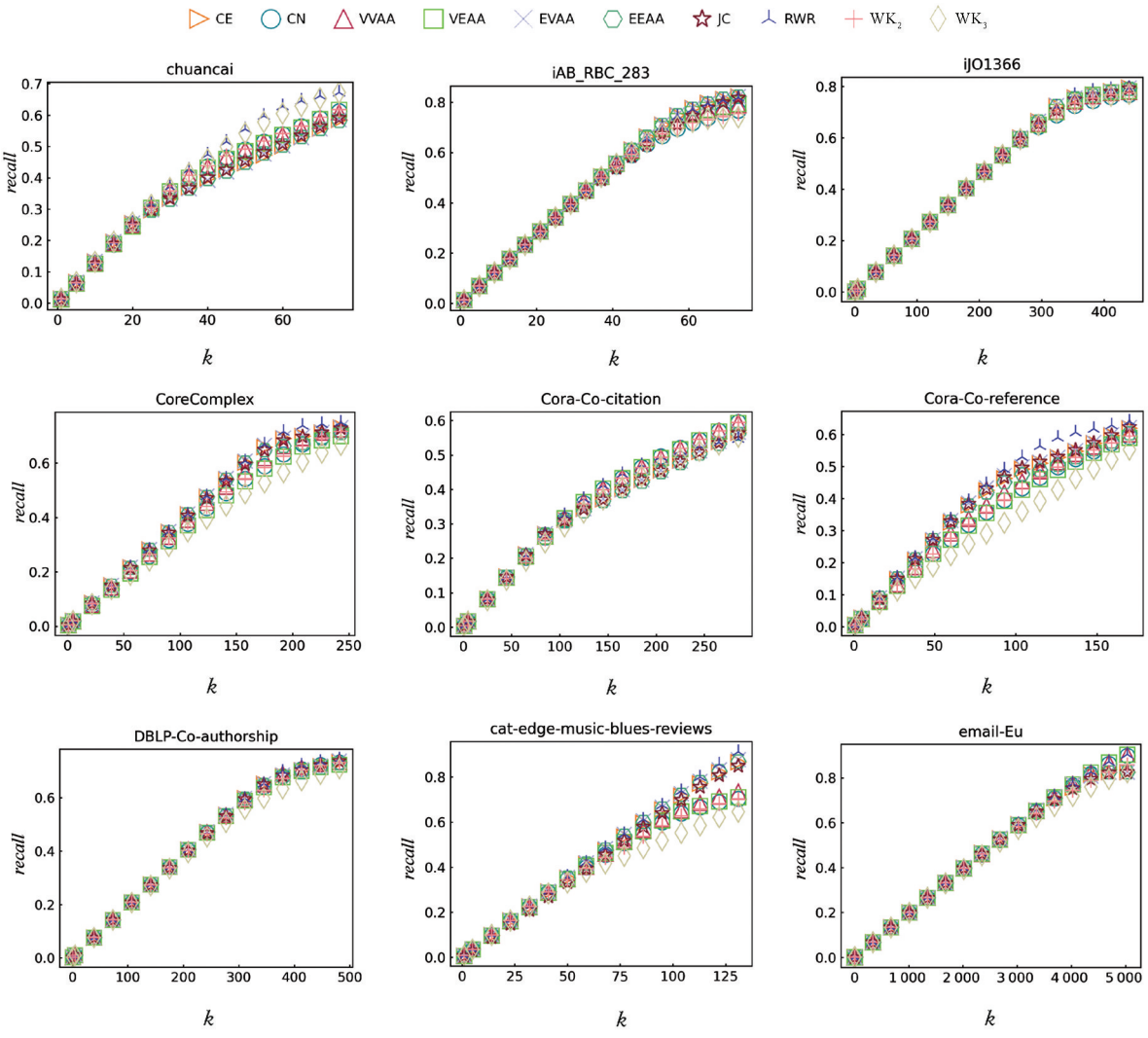



**图 5　召回率曲线**

## 4　总结与展望

本文给出了将链路预测方法扩展到超图环境的一般方式，并基于此方式扩展了 11 种方法. 对所扩展的方法在真实超图上进行分数抽样并生成了对应的分数抽样分布，验证了扩展方法的有效性，发现了不同数据集的超边演化规律，认为演化良好的超图其超边内部的联系强度随节点数增加而增加，由此推论超链接预测的主要难点在于对小尺寸超边的预测. 然后，将所扩展方法直接应用于真实超图的超链路预测，链路预测结果表明，不同方法适用于不同的超图环境，但在局部精确率和召回率上，RWR 有着更好的性能. 还根据节点组大小对预测效果分别进行评估，结果表明在同一数据集内，扩展方法的性能随着节点组节点数的增加而增加.

在未来的研究中，希望引入除式(13)外更多的扩展方式以达到性能的进一步提升. 对于具体的链路预测方法，在普通链路预测指标的基础上，希望加入更多超图独有的高阶性质，如“宽度”[17]、“高阶随机游走”[27]等，另外，更希望所设计的方法能有效地应对当前方法在预测小尺寸超边上的无力. 最后，关于超边下采样算法的讨论正如火如荼[2, 28-29]，期待在未来，使用最新的下采样算法所得到更加“仿真”的负样本能更好地评估方法的性能.

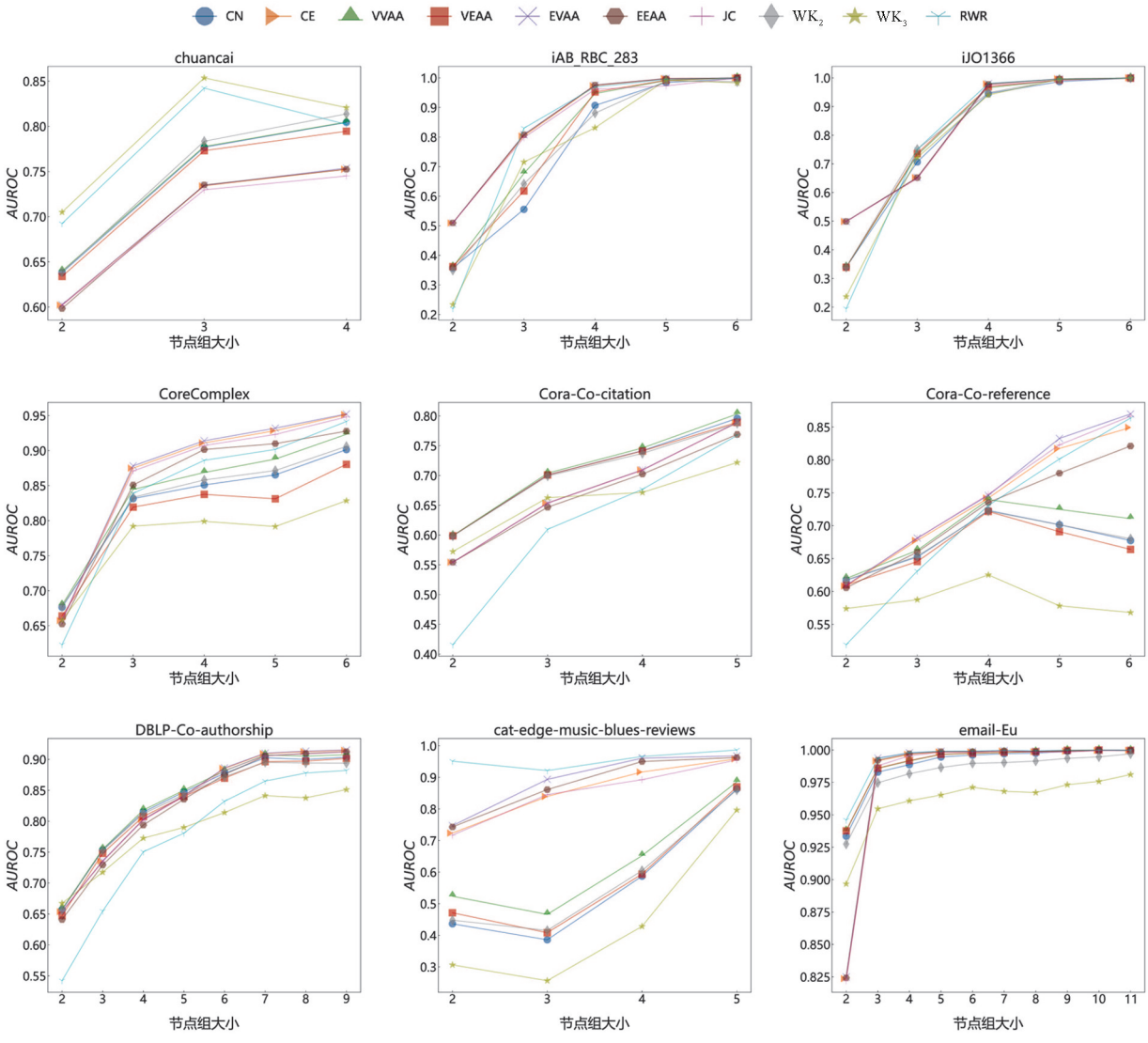



**图6 AUROC随节点组大小变化图**

**参考文献:**

[1] ZHANG M H, CUI Z C, JIANG S L, et al. Beyond Link Prediction: Predicting Hyperlinks in Adjacency Space [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[2] PATIL P, SHARMA G, MURTY M N. Negative Sampling for Hyperlink Prediction in Networks [M] //Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2020: 607-619.

[3] KUMAR T, DARWIN K, PARTHASARATHY S, et al. HPRA: Hyperedge Prediction Using Resource Allocation [C] // Proceedings of the 12th ACM Conference on Web Science. July 6-10, 2020, Southampton, United Kingdom. New York: ACM, 2020: 135-143.

[4] SRINIVASAN B, ZHENG D, KARYPIS G. Learning over Families of Sets-Hypergraph Representation Learning for Higher Order Tasks [M] //Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Philadelphia, PA: Society for Industrial and Applied Mathematics, 2021: 756-764.

[5] YUAN Y B, QU A N. High-Order Joint Embedding for Multi-Level Link Prediction [J]. Journal of the American Statistical Association, 2022: 1-15.

[6] BENSON A R, ABEBE R, SCHAUB M T, et al. Simplicial Closure and Higher-Order Link Prediction [J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(48): E11221-E11230.

[7] LORRAIN F, WHITE H C. Structural Equivalence of Individuals in Social Networks [J]. The Journal of Mathematical Sociology, 1971, 1(1): 49-80.

[8] OU Q, JIN Y D, ZHOU T, et al. Power-Law Strength-Degree Correlation from Resource-Allocation Dynamics on Weighted Networks [J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007, 75(2 Pt 1): 021102.

[9] KATZ L. A New Status Index Derived from Sociometric Analysis [J]. Psychometrika, 1953, 18(1): 39-43.

[10] LIU Z, ZHANG Q M, LÜ L Y, et al. Link Prediction in Complex Networks: A Local Naïve Bayes Model [J]. EPL (Europhysics Letters), 2011, 96(4): 48007.

[11] 吕琳媛，周涛. 链路预测 [M]. 北京：高等教育出版社，2013.

[12] PAN L M, SHANG H J, LI P Y, et al. Predicting Hyperlinks via Hypernetwork Loop Structure [J]. EPL (Europhysics Letters), 2021, 135(4): 48005.

[13] 吕琳媛. 复杂网络链路预测 [J]. 电子科技大学学报，2010，39(5)：651-661.

[14] JACCARD P. Etude Comparative de La Distribution Florale Dans Une Portion des Alpes et des Jura [J]. Bull Soc Vaudoise Sci Nat, 1901, 37: 547-579.

[15] ADAMIC L A, ADAR E. Friends and Neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230.

[16] WANG Q, YAN G Y. IHRW: An Improved Hypergraph Random Walk Model for Predicting Three-Drug Therapy [J]. bioRxiv, 2021: 2021. 02. 25. 432979.

[17] AKSOY S G, JOSLYN C, ORTIZ MARRERO C, et al. Hypernetwork Science via High-Order Hypergraph Walks [J]. EPJ Data Science, 2020, 9(1): 16.

[18] SHARMA G, PATIL P, MURTY M N. C3MM: Clique-Closure Based Hyperlink Prediction [C] //Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. July 11-17, 2020. Yokohama, Japan. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 3364-3370.

[19] SEN P, NAMATA G, BILGIC M, et al. Collective Classification in Network Data [J]. AI Magazine, 2008, 29(3): 93.

[20] Ley M. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives [C] //String Processing and Information Retrieval: 9th International Symposium, SPIRE 2002 Lisbon, Portugal, September 11 - 13, 2002 Proceedings 9. Springer Berlin Heidelberg, 2002: 1-10.

[21] NI J M, LI J C, MCAULEY J. Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects [C] //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 188-197.

[22] YIN H, BENSON A R, LESKOVEC J, et al. Local Higher-Order Graph Clustering [C] //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13-17, 2017, Halifax, NS, Canada. New York: ACM, 2017: 555-564.

[23] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph Evolution: Densification and Shrinking Diameters [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2-es.

[24] GHASEMIAN A, HOSSEINMARDI H, GALSTYAN A, et al. Stacking Models for nearly Optimal Link Prediction in Complex Networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(38): 23393-23400.

[25] KOVÁCS I A, LUCK K, SPIROHN K, et al. Network-Based Prediction of Protein Interactions [J]. Nature Communications, 2019, 10(1): 1240.

[26] ZHANG Z K, LIU C A. A Hypergraph Model of Social Tagging Networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2010, 2010(10): P10005.

[27] CARLETTI T, BATTISTON F, CENCETTI G, et al. Random Walks on Hypergraphs [J]. Physical Review E, 2020, 101(2): 022308.

[28] HWANG H, LEE S, PARK C, et al. AHP: Learning to Negative Sample for Hyperedge Prediction [C] //Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. July 11-15, 2022, Madrid, Spain. New York: ACM, 2022: 2237-2242.

[29] YANG D Q, QU B Q, YANG J, et al. Revisiting User Mobility and Social Relationships in LBSNS: A Hypergraph Embedding Approach [C] //WWW '19: The World Wide Web Conference. May 13-17, 2019, San Francisco, CA, USA. New York: ACM, 2019: 2147-2157.

**责任编辑**　任剑乔