在现实世界的超图中，超边是如何重叠的？模式，测量指标和生成算法

Geon Lee* KAIST AI Daejeon, South Korea geonlee0325@kaist.ac.kr
Minyoung Choe* KAIST AI Daejeon, South Korea minyoung.choe@kaist.ac.kr
Kijung Shin KAIST AI & EE Daejeon, South Korea kijungs@kaist.ac.kr

摘要

超图是图的一种推广，它自然地表示了多个体或对象之间的群体关系，这种关系在许多应用领域都很常见，如网络、生物信息学和社会网络。每条超边上节点数的灵活性是图和超图之间的结构差异的原因，也为超图提供了更强的表达能力。超边的重叠导致了复杂的高阶关系超越了成对关系。这也促使我们提出了图论中未曾考虑的新问题: 超边是怎样重叠的？它们的重叠有什么普遍的特征吗？什么潜在的过程会导致这样的重叠模式？

在这项工作中，我们仔细研究了来自不同领域的十三个真实世界的超图，并分享了关于超边重叠的有趣观察。为此，我们定义了原则性度量，并对实际超图和空模型中的超边重叠进行了统计比较。此外根据观察，我们提出了 HYPERLAP，一种真实的超图生成模型。HYPERLAP (a)真实: 它精确地再现了现实世界超图的重叠模式；(b)自动适配: 它的参数可以通过 HYPERLAP 自动调整，生成超图，特别是类似于给定的目标超图；(c)可伸缩: 它在几个小时内生成并适合一个超图，超图有 0.7 亿条超边。

Acm 参考格式:
李金，崔敏永，申基正。2021 年。在现实世界的超图中，超边是如何重叠的？- 模式、测量指标和生成算法。网络会议记录 2021(www’21) ，2021 年 4 月 19-23 日，卢布尔雅那，斯洛文尼亚 ACM，纽约，NY，美国，12 页。Https://doi.org/10.1145/3442381.3450010

1. 简介

在复杂的系统中，多个体或对象之间的群体交互是无处不在的: 合作者的协作、项目的共同参与、问答网站的群体交流等等。它们自然地被建模为一个超图，其中每个超边(即任意数量节点的子集)表示一个组交互。超图是普通图的推广，它自然地描述了成对的相互作用。

在现实世界的超图中，超边是相互重叠的，揭示了它们之间有趣的关系。由于每个超边大小的灵活性，即使一个固定数量的超边也可以以无限多种方式重叠。此外，这些关系是高阶的，将它们分解成成对关系会失去相当多的信息。超图的这一独特性质提出了一些在图中尚未被考虑到的重要问题: (1)在现实世界的超图中超边是如何重叠的？(2)是否存在区分现实世界超图和随机超图的非平凡模式？(3)我们如何通过简单的机制复制这些模式？

最近的实证研究部分地回答了这些问题，揭示了真实世界超图的结构和动力学模式。所发现的模式是关于连通分支[12] ，直径 [12,22] ，3 元组 [5] ，3-超边子超图[24] ，单纯闭包[5] ，时间相近超边之间的相似性[6]，相交超边[22]的数目等，这些模式直接或间接地受

到超边重叠的影响。此外，超边的重叠被用于超边预测[5,6,24]和实际超图的生成[12]。

在这项工作中，我们补充以前的研究，新的发现，措施，和现实的生成模型，关于超边的重叠。为此，我们仔细研究了来自 6 个不同领域的十三个真实世界的超图。具体来说，我们在三个不同的层次上分析它们中超边的重叠：节点子集，超边，和自我网络。然后我们用随机超图来验证我们的发现，我们随机重叠超边，同时保留节点的度数和超边的大小。我们的研究显示，真实世界超图中超边的重叠表现出以下特性：

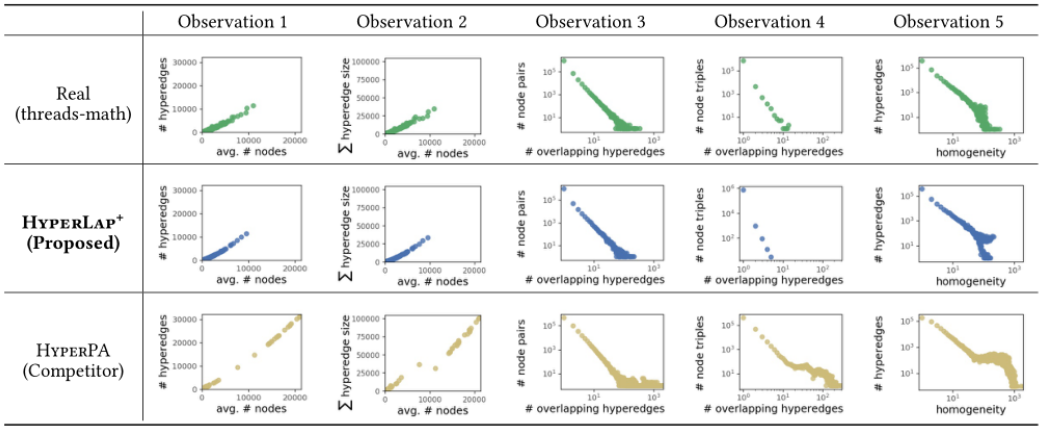
•实质性：在现实世界的超图中，每个自我网络中的超边往往比随机超图中的重叠更多。

•重尾性：在现实世界超图中，每一对或三元组节点上重叠的超边数目比随机超图更倾斜，尾巴更重。重叠超边的个数服从近幂律分布。

•同型性：在现实世界的超图中，包含在每个超边中的节点在结构上往往比随机图中的节点更相似(即，更多的超边在它们上重叠)。

为了研究真实世界的超图，我们设计了新颖的原则性测度。我们证明了我们的超边重叠度量满足三个直观清晰的公理，而广义密度度量不满足这三个公理。我们还引入了节点子集重叠度的度量，它揭示了有趣的近幂律分布行为，以及超边同质性的度量，它在实际超图生成中起着关键作用。

表 1: HYPERLAP+精确地再现了真实世界超图中超边的重叠。利用 HYPERLAP+创建的合成超图显示出。1)稠密网络，2)高度重叠的自我网络，3)重尾成对节点度分布，4)重尾三节点度分布，5)齐次超边。我们在[1]中提供了完整的结果。



什么潜在的过程可以导致超边系统地重叠显示上述模式？我们设计了 HYPERLAP 算法，一个随机超图生成模型，HYPERLAP 精确地再现了超边的真实重叠模式。此外，我们还提出了 HYPERLAP+，它可以自动调整 HYPERLAP 生成合成超图的参数，特别类似于给定的目标图(见表 1)。HYPERLAP 给出了在推理和预测超图演化过程中非常有用的直觉，当不可能收集或跟踪真实的超图时，它可以用来生成合成的超图，用于模拟和评估算法。HYPERLAP+可以用于匿名化不能公开的超图，以便共享它们。

我们的贡献归纳如下：

- 现实世界超图中的观察：我们发现了现实世界超图中超边重叠的三个独特特征，并用随机超图验证了它们
- 新的测度：我们定义了新的和原则性的测度关于超边的重叠在 3 个不同的水平。它们在调查和真实超图生成中起着关键作用。
- 现实生成模型：我们提出 HYPERLAP，一个随机的超图生成器，它可以重现现实的超边重叠。我们还提供了 HYPERLAP+ ，它能自动适应给定超图的 HYPERLAP 参数。根据经验，它们随着超边的数量线性增长。

可重复性：这项工作中使用的源代码和数据集等资源可以通过以下网站获得

https://github.com/young917/www21-hyperlap。

在第二部分，我们讨论相关的工作。在第三部分，我们描述了整个工作中使用的数据集和空模型。在第四部分，我们分享我们对实际超图中超边重叠的观察。在第 5 部分，我们提出了 HYPERLAP，一个真实的超图生成模型，并提供了实验结果。最后，我们在第六部分提出结论。

2. 相关工作

宏观结构模式[4,13,35,39] ，微观结构模式[32,33] ，动力模式[15,23,27]在现实世界成对图中已有广泛的研究，并提出了许多现实图生成器[8,14,25,27,44]来复制已发现的模式。在本节中，我们将重点讨论超图，并回顾以前对实际超图和现实超图生成器中的经验模式的研究。超图已被广泛应用于各个领域，包括计算机视觉[45]、生物信息学[17]、电路设计[20]、社会网络分析[41]和建议[30]。它们被用于各种分析和学习任务，包括分类[18,40]、聚类[3,28,29]和超边预测[5,43]。除了下面描述的实际的超图生成器之外，一些随机的超图模型[7,9,19,37]也被用于统计检验。

Benson 等[5]着重研究了单纯闭包事件(即包含一组节点的超边的第一次出现，每组节点的对在以前的超边中共同出现) ，并研究了它们的概率是如何受到来自不同域的现实世界超图的局部特征，如平均度的影响的。

Benson 等人[6]考虑了现实超图中的序列(即相互关联的时序超边) ，并指出序列中的超图往往更类似于最近的超边而不是远处的超边。他们还发现，在每一对和三个节点上重叠的超边的数量在每个序列中往往比在零模型中更大。此外，作者建议在预测序列中的下一个超边时利用这两种模式。值得注意的是，在 4.2 节中，我们还研究了成对和三个节点上重叠的超边的数量。然而，我们(a)在超图层次上检验它们，(b)发现它们的近幂律分布，(3)将它们与保度随机超图进行比较。

Do 等人[12]考虑将一个真实世界的超图投影到多个成对图中，这样每个第 k 个图描述了节点的大小为 k 子集之间的相互作用。研究结果表明: 两两图具有(a)重尾度和奇异值分布，(b)巨连通分量，(c)小直径和(d)高聚类系数。受到这些观察的启发，作者提出了一个超图形生成器，叫做 HYPERPA [12]。在超边集中，节点的子集与新节点形成一个超边集，选择的概率与包含该子集的超边集的数量成正比。

Kook 等人[22]揭示了交叉超边的比例和真实世界超图的直径随时间逐渐减小，而超边的数量比节点的数量增加得更快。此外，他们还发现了四种结构模式，分别涉及(a)包含每个节点的超边的数目，(b)超边的大小，(c)两个超边之间的交点的大小，以及(d)入射矩阵的奇异值。为了重现这些模式，作者提出了一个超图形生成器，叫做 HYPERFF。对于每个新节点，超分类模拟森林火灾在超边上蔓延，新节点与每个烧毁的节点形成一个 大小为 2 的超边。然后再次模拟森林大火，扩展每个 2 号超边。

Lee 等人[24]提出了 26 个超图模体，它们是三条连通超边的连通模式，基于七个 Venn 图区域的空性。结果表明，在同一领域的现实世界超图中，h 基序的相对出现特别相似。

所有这些发现都直接或间接地与超边的重叠有关。在这项工作中，我们补充了以前的研究、新的发现、测量准则和更现实以及可扩展的生成器，所有这些都与超边的重叠有关。

3. 数据集和零模型

在这一部分，我们首先介绍一些注释和预备知识。然后，我们描述的数据集和零模型使

用本文。有关常用的记号，请参阅表 2。

表 2：常用符号。

Notation	Definition
$G = (V, E)$	hypergraph with nodes V and hyperedges E
$E = \{e_1, \ldots, e_{\|E\|}\}$	set of hyperedges
$E_{\{v\}}$	set of hyperedges that contain a node v
E_S	set of hyperedges that contain a subset S of nodes
L	number of levels in HyperLap
$w_1, \ldots, w_L$	weight of each level
$S_g^{(\ell)}$	set of nodes in a group g of level ℓ

3.1 预备知识和符号表示

我们回顾了超图的概念，然后是 Chung-Lu 模型，我们的零模型就是基于这个模型的。

超图: 一个超图$G = (V, E)$由一组节点V和一组超边$E \subseteq 2^V$组成。每个超边$e \subseteq V$都是全体节点集$|e|$个节点的非空子集。对于每个节点 v，我们表示包含 v 的超边集合为 $E_{\{v\}} \coloneqq \{e \in E : v \in e\}$，定义节点 v 的度为包含节点 v 的超边个数$\mathrm{d_v} = |E_{\{v\}}|$。如果两条超边$e_i$和$e_j$共享任何一个节点，我们就说它们是相交的，比如$e_i \cap ej \neq \emptyset$。

Chung-Lu 模型： Chung-Lu（CL）模型[10]是一个随机图模型，它生成的图中，节点的给定度序列预计会被保留。考虑一个图$\bar{G} = (\bar{V}, \bar{E})$，其中$\bar{E}$是成对边的集合。给定期望的度分布$\{d_1, d_2, \ldots, d_{|\bar{V}|}\}$，其中 d_i 是节点 i 的度，CL 模型通过在每对节点之间创建一条边来生成随机图，其概率与其度的乘积成正比。也就是说，对于每对(i, j)节点，使用概率$\frac{d_i d_j}{2M}$创建边 $\mathrm{e_{ij}}$，其中$M = \frac{1}{2}\sum_{k=1}^{|\bar{V}|} d_k$，假设$d_k < \sqrt{M}$对所有 k 都成立。如果我们将生成的图中每个节点 i 的度设为$\bar{d_i}$，则其期望值等于 d_i，即：

$$E[\tilde{d_i}] = \sum_{j=1}^{|\bar{V}|} \frac{d_i d_j}{2M} = d_i \sum_{j=1}^{|\bar{V}|} \frac{d_j}{2M} = d_i$$

当 CL 模型为所有可能的$O(|\bar{V}|^2)$节点对掷硬币时，快速 CL（FCL）模型[34]以与每个节点的程度成比例的概率对两个节点进行独立采样。然后，它在采样的一对节点之间创建一条边。此过程重复$|\bar{E}|$次，总时间复杂度为$O(|\bar{E}|)$。即使在 FCL 模型生成的图中，每个节点的期望度也等于d_i。

3.2 数据集

在删除重复或单一超边后，我们使用了六个不同域[5]中的十三个真实超图。有关超图的一些统计信息，请参阅表 3。

•电子邮件（email-Enron[21]和 email-Eu[26,42]）：每个节点都是一个电子邮件帐户，每个超边都是一组电子邮件的发送者和接收者。

•联系人（contact-primary [38]和 contact-high [31]）：每个节点都是一个人，每个超边缘都是个人之间的组交互。

•药物（NDC-classes 和 NDC-substances）：每个节点是一个类别标签（在 NDC-classes 中）或一种物质（在 NDC-substances 中），每个超边是药物的一组标签/物质。

•标签（tags-ubuntu 和 tags-math）：每个节点都是一个标签，每个超边都是一组附在问题上的标签。

•线程（threads-ubuntu 和 threads-math）：每个节点都是一个用户，每个超边都是一组参与线程的用户。

•合著（coauth-DBLP、coauth-geology [36]和 coauth-history [36]）：每个节点都是作者，每个超边都是出版物的一组作者。

表 3：来自 6 个域的 13 个真实超图的汇总统计：节点数| V |，超边数| E |，平均超边大小 avge∈E | E |，以及最大超边大小 maxe∈E | E |。

Dataset	|V|	|E|	$\text{avg}_{e\in E}\|e\|$	$\max_{e\in E}\|e\|$
email-Enron	143	1,459	3.13	37
email-Eu	986	24,520	3.62	40
contact-primary	242	12,704	2.41	5
contact-high	327	7,818	2.32	5
NDC-classes	1,149	1,049	6.16	39
NDC-substances	3,767	6,631	9.70	187
tags-ubuntu	3,021	145,053	3.42	5
tags-math	1,627	169,259	3.49	5
threads-ubuntu	90,054	115,987	2.30	14
threads-math	153,806	535,323	2.61	21
coauth-DBLP	1,836,596	2,170,260	3.43	280
coauth-geology	1,091,979	909,325	3.87	284
coauth-history	503,868	252,706	3.01	925

3.3 零模型：HyperCL（算法 1）

我们提出了 HyperCL，一种将 FCL 模型（见第 3.1 节）扩展到超图的随机超图生成器。在这项工作中，我们使用 HyperCL 生成的随机超图作为空模型。如算法 1 所述，在所考虑的现实世界超图中，节点的度分布和超边的大小分布作为输入。对于每个第 i 个超边$\tilde{e_i}$，将其节点独立采样，概率与每个节点的阶数成比例（即，每个节点 V 的概率为$d_v/\sum_{j=1}^{|V|} d_j$），直到超边的大小达到 s_i（第 4-6 行）。请注意，重复的节点将被忽略，以便每个第 i 个超边包含 s_i 个不同的节点。

在由 HyperCL 生成的超图中，超边的大小分布与输入大小分布完全相同，并且节点的度分布也期望与输入度分布相似。具体来说，如果我们假设$\sum_{j=1}^{|V|} d_j \gg \left(\max_{k\in\{1,\cdots,|E|\}} s_k\right)\cdot$

$\left(\max_{k\in\{1,\cdots,|V|\}} d_k\right)$，令$\widetilde{d_v}$等于生成的超图的度。

$$E[\widetilde{d_v}] = \sum_{\tilde{e}\in E} P[v \in \tilde{e}] \approx \sum_{\tilde{e}\in E}\left(|\tilde{e}| \cdot \frac{d_v}{\sum_{j=1}^{|V|} d_j}\right) = \frac{d_v}{\sum_{j=1}^{|V|} d_j}\sum_{\tilde{e}\in E}|\tilde{e}| = d_v$$

在[1]中，我们通过实验证明了由 HyperCL 生成的超图中度分布与输入度分布是接近的。

Algorithm 1: HyperCL: Random Hypergraph Generator

Input : (1) distribution of hyperedge sizes $\{s_1, ..., s_{|E|}\}$
(2) distribution of node degrees $\{d_1, ..., d_{|V|}\}$

Output: random hypergraph $\tilde{G} = (\tilde{V}, \tilde{E})$

$\tilde{V} \leftarrow V$ and $\tilde{E} \leftarrow \emptyset$
for each $i = 1, ..., |E|$ **do**
  $\tilde{e_i} \leftarrow \emptyset$
  while $|\tilde{e_i}| < s_i$ **do**
    $v \leftarrow$ select a node with prob. proportional to the degree
    $\tilde{e_i} \leftarrow \tilde{e_i} \cup \{v\}$
  $\tilde{E} \leftarrow \tilde{E} \cup \{\tilde{e_i}\}$
return $\tilde{G} = (\tilde{V}, \tilde{E})$

4. 观察结果

在本节中，我们将研究真实超图中超边的重叠模式，并通过与由 HyperCL 获得的随机超图中的重叠模式进行比较来验证它们。我们研究了三个不同层次上的超边重叠，我们的观察结果总结如下。

•（L1）自我网络级别：在真实超图中，每个节点的自我网络中的超边重叠往往比在随机超图中更为明显。

•（L2）节点对/三重级别：每对或三重节点上重叠的超边数遵循近似（截断）幂律分布。此外，与随机超图相比，真实世界超图中重叠超边的数量更偏斜，尾部更重。

•（L3）超边级别：与随机超图相比，超边在真实超图中往往包含结构更相似的节点（即更多超边重叠的节点）。

4.1 L1.自我网络级别

自我网络的密度：我们首先研究现实世界超图中的自我网络。我们将节点 v 的自我网络定义为包含 v（即$E_{\{v\}} \coloneqq \{e \in E: v \in e\}$）的超边集。为了定量地测量自我网络中的超边基本上是如何重叠的，我们首先考虑真实世界和随机超图中的自我网络的密度（参见定义 1），这导致观察 1。而一组超边$\mathcal{E}$的密度可以定义为超边的数量除以诱导节点 V 的功率集的大小

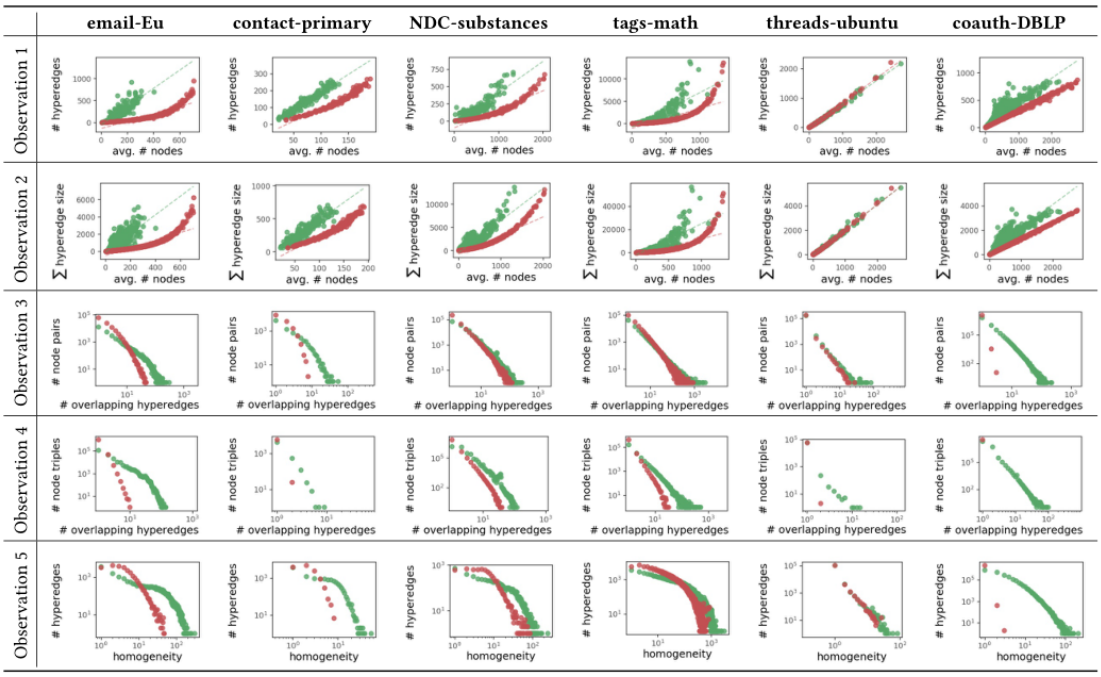
（即$\frac{|\mathcal{E}|}{2^{|V|}-1}$），我们在这项工作中遵循[16]中的定义。

定义 1（密度[16]）。给定一组超边$\mathcal{E}$，该集的密度 ρ ($\mathcal{E}$)定义为：$\rho(\mathcal{E}) \coloneqq \frac{|\mathcal{E}|}{|\cup_{e\in\mathcal{E}} e|}$

观察 1。*现实世界超图中的自我网络往往比随机超图中的自我网络更密集。*

具体而言，如表 4 第一行中的图所示，当考虑具有相同数量超边的自我网络时，它们在真实超图中的节点往往少于随机超图中的节点。因此，在真实超图中，密度（定义为超边数与节点数之比）往往高于随机超图中的密度。在这些图中，回归线的斜率接近于平均自我网络密度，在现实世界的超图中比在随机超图中更陡。

表 4：真实超图中的超边与随机超图中的超边明显重叠。我们检查（Obs.1）每个自我网络的密度，（Obs.2）每个自我网络的重叠，（Obs.3）每对节点上重叠的超边的数量，（Obs.4）每三个节点上重叠的超边的数量，以及（Obs.5）每个超边的同质性。关于观察结果 5，我们通过将超边同质性的连续值组合成最近的整数，对其进行预处理。我们在[1]中提供了完整的结果。



原则性度量：重叠度：但是，“密度”并没有充分考虑超边的重叠。考虑两组超边：$\mathcal{E}1 = \{\{a, b, c\}, \{a, b, c, d\}, \{a, b, c, d, e\}\}$和$\mathcal{E}2 = \{\{v, w, x\}, \{x, y\}, \{y, z\}\}$。直观地说，$\mathcal{E}1$的重叠程度比$\mathcal{E}2$更高，但由相同数量的节点和超边组成的两个集合的密度是相同的。

为了解决这个问题，我们首先提出了三个公理，这三个公理是超边重叠的任何合理度量都应该满足的。然后，我们提出了满足所有公理的新度量——重叠度。这三条公理在公理 1、2 和 3 中被形式化。

公理 1（超边数）。考虑两组包含相同大小的超边的超边集$\mathcal{E}$和$\mathcal{E}'$，以及相同数量的不同节点。然后，具有更多超边的集合比其他集合重叠更多。正式地：

$$\left(\left(|\mathcal{E}| < \left|\mathcal{E}'\right|\right) \wedge \left(|e| = \left|e'\right| = n, \forall e \in \mathcal{E}, \forall e' \in \mathcal{E}'\right)\right.$$

$$\wedge\left(\left|\bigcup_{e\in\mathcal{E}} e\right| = \left|\bigcup_{e'\in\mathcal{E}'} e'\right|\right)\Rightarrow f(\mathcal{E}) < f\left(\mathcal{E}'\right).$$

公理 2（不同节点的数量）。考虑两个超边$\mathcal{E}=\{e_1,\cdots,e_n\}$和$\mathcal{E}'=\{e_1',\cdots,e_n'\}$，具有相同数量的超边和相同的超边的大小分布。然后，包含较少不同节点的集合比其他集合重叠更多。正式地：

$$\left(\left(|\mathcal{E}| < \left|\mathcal{E}'\right|\right)\wedge\left(|e_i| = \left|e_i'\right| = n, \forall i\in\{1,\ldots,\mathrm{n}\}\right)\right.$$

$$\wedge\left(\left|\bigcup_{e\in\mathcal{E}} e\right| = \left|\bigcup_{e'\in\mathcal{E}'} e'\right|\right)\Rightarrow f(\mathcal{E}) < f\left(\mathcal{E}'\right).$$

公理 3（超边的大小）。考虑两组超边$\mathcal{E}=\{e_1,\cdots,e_n\}$和$\mathcal{E}'=\{e_1',\cdots,e_n'\}$，它们具有相同数目的不同的节点和相同数量的超边。然后，具有较大超边的集合比其他集合重叠更多。正式地：

$$\left(\left(|\mathcal{E}| = \left|\mathcal{E}'\right| = n\right)\wedge\left(|e_i| < \left|e_i'\right|\right)\wedge\left(\left|e_j\right| \le \left|e_j'\right|, \forall j\in\{1,\cdots,n\}\backslash\{i\}\right)\right.$$

$$\wedge\left(\left|\bigcup_{e\in\mathcal{E}} e\right| = \left|\bigcup_{e'\in\mathcal{E}'} e'\right|\right)\Rightarrow f(\mathcal{E}) < f\left(\mathcal{E}'\right).$$

请注意，表 5 中列出的密度和四个其他广泛使用的测量值并不满足所有公理。因此，我们提出重叠度（见定义 2）作为超边重叠度的度量，它满足定理 1 中形式化的所有公理。

表 5：重叠度合理地度量了超图重叠的程度，满足所有公理，而其他公理不满足。详情见附录 A。

Metric	Axiom 1	Axiom 2	Axiom 3
Intersection	✗	✗	✗
Union Inverse	✗	✓	✗
Jaccard Index	✗	✗	✗
Overlap Coefficient	✗	✗	✗
Density	✓	✓	✗
Overlapness (Proposed)	✓	✓	✓

定义 2（重叠）。给定一组超边 E，该集的重叠度$\mathrm{o}(\mathcal{E})$定义如下：

$$o(\mathcal{E}) \coloneqq \frac{\sum_{e\in\mathcal{E}}|e|}{|\cup_{e\in\mathcal{E}} e|}$$

定理 1（重叠的可靠性）。重叠度$o(\cdot)$满足公理 1、2 和 3。

证明。见附录 A。

在重叠中，考虑的是超边的大小之和，而不是超边的数量。值得注意的是，超边集的重叠度相当于集合中不同节点的平均度。此外，如果我们指定每个超边的大小作为其权重，则重叠度相当于加权密度。在前面的例子中，重叠与我们的直觉是一致的。也就是说，对于$\mathcal{E}1=$

$\{\{a, b, c\}, \{a, b, c, d\}, \{a, b, c, d, e\}$
和 $\mathcal{E}2 = \{v, w, x\}, \{x, y\}, \{y, z\}, o(E1) = 12/5 > o(E2) = 7/5$ 是成立的。

自我网络的重叠度：我们测量现实世界和随机超图中自我网络的重叠度，这导致观察 2。如表 4 第二行中的图所示，现实世界超图中的自我网络往往比随机超图中的自我网络具有更高的重叠度。回归线的斜率接近平均 egonet 重叠度，在真实超图中比在随机超图中更陡。

观察 2。*现实超图中的自我网络比随机超图中的自我网络具有更高的重叠度。*

跨域比较：此外，我们计算了超图 G 中自我网络的密度和重叠度的显著性，定义为：

$$sig_{\rho}(G) \coloneqq \frac{\bar{\rho}(G) - \bar{\rho}(G')}{\max\limits_{g \in \omega(G), g' \in \omega(G')} |\rho(g) - \rho(g')|}$$

$$sig_{o}(G) \coloneqq \frac{\bar{o}(G) - \bar{o}(G')}{\max\limits_{g \in \omega(G), g' \in \omega(G')} |o(g) - o(g')|}$$

其中 G′ 是 G 的随机超图；$\bar{\rho}(\cdot)$和$\bar{o}(\cdot)$分别是平均自我网络密度和重叠度；$\omega(\cdot)$是自我网络的集合。如图 1 所示，来自同一领域的真实超图在自我网络的密度和重叠度方面具有相似的意义，这表明它们的超边在自我网络级别上具有相似的重叠模式。

4.2 L2.节点对/三节点组级

给定一对或三对节点，有多少超边在它们处重叠？换句话说，有多少条超边包含这对或三重边？虽然度通常定义为包含每个单独节点的超边的数量，但这里我们将此概念扩展到节点对和节点三元组。具体来说，如果我们让$E_S := \{e \in E : S \subseteq e\}$是在子集$S \subseteq V$上重叠的超边集然后，将每个节点对$\{i, j\}$的度定义为$d^{(2)}(\{i, j\}) := |E\{i, j\}|$，将每个节点三元组$\{i, j, k\}$的度定义为$d^{(3)}(\{i, j, k\}) := |E\{i, j, k\}|$。对或三元组的程度也可以解释为对或三元组中节点之间的结构相似性。直观地说，节点在结构上更相似，因为它们一起包含在更多的超边中。

表 6: 在每一对或三个节点上重叠的超边数的分布是重尾的，接近于截断的幂律分布。这一主张得到了报告的对数似然比的支持，当对三种重尾分布(幂律、截断幂律和对数正态)分别与指数分布进行拟合时。

Dataset	Pair of Nodes (Obs. 3)			Triple of Nodes (Obs. 4)		
	pw	tpw	logn	pw	tpw	logn
email-Enron	-0.36	**4.22**	**3.50**	**1.91**	**3.88**	**3.47**
email-Eu	**0.66**	**1.48**	1.29	**0.21**	**0.77**	**0.63**
contact-primary	**0.64**	**1.40**	1.35	**0.01**	**0.48**	**0.48**
contact-high	**0.75**	**0.81**	**0.79**	-1.04	-	**0.80**
NDC-classes	**13.49**	**15.74**	**14.78**	**24.37**	**31.53**	**29.19**
NDC-substances	**38.68**	**43.87**	**42.55**	**102.90**	**116.45**	**109.77**
tags-ubuntu	**39.66**	**41.55**	**41.25**	**17.03**	**17.84**	**17.79**
tags-math	**3.82**	**4.49**	**4.47**	**26.97**	**29.26**	**29.07**
threads-ubuntu	**3.79**	**3.97**	**3.97**	**0.34**	**0.80**	**0.73**
threads-math	**14.25**	**14.78**	**14.68**	-1.04	-0.09	-1.12
coauth-DBLP	**19.23**	**22.47**	**22.31**	**5.75**	**5.84**	**5.83**
coauth-geology	**45.20**	**53.39**	**52.92**	**9.69**	**13.73**	**13.01**
coauth-history	**3.74**	**3.81**	**3.91**	-0.36	**1.42**	**1.27**

通过检查节点对和三元组的度分布，而不是单个节点的度分布，可以更深入地了解节点作为一个集合如何形成超边。在表 4 的第三列和第四列中，我们提供了 $d^{(2)}$和 $d^{(3)}$在真实超图和相应随机超图中的分布。观察结果 3 和 4 总结了我们的研究结果。

观察 3。*与随机超图相比，真实超图中**节点对**上重叠的超边数量（即每对节点的度）更偏斜，尾部更重。该分布类似于截断的幂律分布。*

观察 4。*与随机超图相比，真实超图中**三节点组**上重叠的超边数（即每三个节点的阶数）更偏斜，尾部更重。该分布类似于截断的幂律分布。*

除了目视检查之外，我们还计算了三种典型的重尾分布(幂律、截断幂律和对数正态)与指数分布的对数似然比，如文献[2,11]所建议的。如果比率大于 0，则给定的分布与相应的重尾分布比指数分布更相似。如表 6 所示，除了一种情况外，至少有一种重尾分布的比率为正，而且在大多数情况下，截断的幂律分布的比率最高。这些结果支持了节点对和节点三元组的度分布是重尾分布，类似于截断的幂律分布的结论。

事实上，这些结果是直观的。一对或三个节点相互作用的频率越高，它们再次相互作用的可能性就越大。例如，共同撰写多篇论文的研究人员可能有共同的兴趣，这可能导致未来更多的合作。

4.3 L3.超边水平

形成超边的节点是如何相互关联的？在现实世界的超图中，每个超边都是由随机选择的节点组成的，这是不可能的。它们之间存在强相关性，形成一个超边。为了研究这种依赖性，我们使用定义 3 中定义的超边的同质性，来衡量这些节点的结构相似程度。

定义 3(超边的同质性)。超边 $e \in E$ 的同质性定义如下:

$$Homogeneity(e) := \begin{cases} \frac{\sum_{\{u,v\}\in\binom{e}{2}} |E_{\{u,v\}}|}{\binom{|e|}{2}}, & \text{if } |e| > 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

此处$\binom{e}{2}$是 e 和 $|E_{\{u,v\}}|$中节点对的集合，是在 u 和 v 对上重叠的超边的个数(即，这对 $\{u,v\}$的度数)。注意，在等式(1)中，两个节点之间的结构相似性是根据在它们上重叠的超边的数量来测量的，我们在第 4.2 节中讨论了这一点。式(1)可以很容易地扩展到三个或更多的节点。

表 4 最后一行的图显示了实际超图和相应的随机超图中超边的同质性。如观察 5 所述，现实世界超图中每个超边的同质性大于随机超图中的同质性。此外，我们还验证了同质性的分布是重尾分布(见表 7) ，如前面的小节。

观察 5。*与随机超图相比，现实超图中的超边在结构上往往包含更多的相似节点(即有许多超边重叠的节点)。*

超边的同质性在生成真实超图中起着关键作用，如下一节所述。

5. 超图生成

我们已经证明了真实超图中超边的重叠模式与随机超图中的重叠模式有着明显的区别。在这一部分，我们提出了 HYPERLAP，这是一个可扩展的、真实的超图生成模型，可以复制真实的超边重叠模式。在提出 HYPERLAP 之后，我们提出了 HYPERLAP+ ，它自动调整超图的参数，从而生成类似于给定目标超图的超图。然后对 HYPERLAP 和 HYPERLAP+进行实验评价。

5.1 HyperLap: 多级 HyperCL

Algorithm 2: HyperLap: Realistic Hypergraph Generator

Input : (1) distribution of hyperedge sizes $\{s_1, \cdots, s_{|E|}\}$
(2) distribution of node degrees $\{d_1, \cdots, d_{|V|}\}$
(3) number of levels L ($\leq \log_2 |V|$)
(4) weights of each level $\{w_1, \cdots, w_L\}$

Output: synthetic hypergraph $\hat{G} = (\hat{V}, \hat{E})$

```
/* Initialization */
V̂ ← {1, ··· , |V|} and Ê ← ∅
/* Hierarchical Node Partitioning */
S_1^(L), ··· , S_{2^{L-1}}^(L) ← uniformly partition V̂ into 2^{L-1} groups
for each level ℓ = L − 1, ··· , 1 do
    for each group g = 1, ··· , 2^{ℓ-1} do
        S_g^(ℓ) = S_{2g-1}^(ℓ+1) ∪ S_{2g}^(ℓ+1)
/* Hyperedge Generation */
for each i = 1, ··· , |E| do
    ℓ ← select a level with prob. proportional to the weight
    S_g^(ℓ) ← select a group at level ℓ uniformly at random
    ê_i ← ∅
    while |ê_i| < s_i do
        v ← select a node from S_g^(ℓ) with prob. proportional
          to the degree
        ê_i = ê_i ∪ {v}
    Ê = Ê ∪ {ê_i}
return Ĝ = (V̂, Ê)
```

我们提出了 hyperlap，一个真实的超图生成模型，算法 2 描述了它的伪代码。Hyperlap 的核心思想是将 hypercl 扩展到多个层次。回想一下 hypercl 本身并不能准确地重现真实的重叠图案，如第 4 节所示。

Hyperlap 的描述: hyperlap 是 hypercl 的一个多级扩展，需要两个额外的输入: (1)每个级别的级数 l 和(2)权重$\{w1, \ldots, wl\}$.[1] 目前，我们假设参数已经给出，如何设置参数将在下一小节讨论。超重叠由层次节点划分步骤和超边生成步骤组成。

第一步。分层节点划分(第 3-7 行)。超重叠第一分割节点成组在每个级别。具体地说，在每一层$l \in \{1, \ldots, L\}$，它随机分割节点2^{l-1}群，用$S_1^{(l)}, \ldots, S_{2^{l-1}}^{(l)}$表示，同时满足以下条件:

(1) $S_i^{(\mathrm{l})} \cap S_j^{(\mathrm{l})} = \emptyset$ for all $i \neq j \in \{1, \cdots, 2^{\mathrm{l}-1}\}$,

(2) $\left|\cup_{i=1}^{2^{\mathrm{l}-1}} S_i^{(\mathrm{l})}\right| = |V|$,

(3) $\left|S_i^{(\mathrm{l})}\right| = \left\lfloor \frac{|V| \cdot i}{2^{\mathrm{l}-1}} \right\rfloor - \left\lfloor \frac{|V| \cdot (i-1)}{2^{\mathrm{l}-1}} \right\rfloor$ for all $i \in \{1, \cdots, 2^{\mathrm{l}-1}\}$,

(4) $S_i^{(\mathrm{l})} = S_{2i-1}^{(\mathrm{l}+1)} \cup S_{2i}^{(\mathrm{l}+1)}$ for all $\mathrm{l} < L$, and $i \in \{1, \cdots, 2^{\mathrm{l}-1}\}$。

第一和第二条件确保在每个级别上，每个节点恰好属于一个组。第三个条件是，每个层次的群体大小几乎是一致的。最后一个条件是，群体是分等级的。也就是说，如果节点在一个层次上在同一组中，那么它们在所有较低层次上在同一组中。注意，节点在更高的层次上被更精细地划分为更小的子集。在最低层 1，存在一个单一的组，这是相同的节点 V 的整个集合，而在最高层 L，存在大多数组的数目是 $2^{\mathrm{L}\text{-}1}$。

步骤 2。超边生成(第 8 - 16 行)。一旦我们在上一步分层划分节点，对于每个第 i 个超边 $\hat{e}i$, hyperlap 首先选择一个级别，其概率与每个级别的权重成正比。也就是说，每个级别ℓ的选择概率与$w\ell$成正比。在选定的级别ℓ上，hyperlap 均匀随机地选择一组$S_{\mathrm{g}}^{(\mathrm{l})}$。然后，对构成$\hat{e}i$的节点进行独立抽样，概率与每个节点的度成正比，直到超边的大小达到si。也就是说，我们不是考虑所有的节点，而是将节点划分为多个组，并将超边可以包含的节点限制为一个组中的节点。请注意，在更高级别的同一组中生成的超边更有可能相互重叠，因为在更高级别的每个组中有更少的节点。实际上，由于$\hat{e}i$不能从一个大小小于si的组中生成，所以我们选择层次ℓ使$\ell \leq log_2 \frac{|v|}{si} + 1$。

Hyperlap 的度保持: 在 hypelap 生成的超图中，超边的尺寸分布与输入尺寸分布完全相同。具体地说，$|\hat{e}i| = si$ 保持所有 $i \in \{1, \ldots, |e|\}$。节点的度分布也应该与输入度分布相似。为了证明这一点，我们首先提供引理 1，我们的分析是基于它的。

引理 1。对于层次l上的每个群 $S_g^{(l)}$ ，从 $S_g^{(l)}$生成超边 e 的概率是

$$P\left[e \subseteq S_g^{(\mathrm{l})}\right] = \frac{w_{\mathrm{l}}}{W_e} \cdot \frac{1}{2^{\mathrm{l}-1}} \quad (2)$$

其中，We是适当级别的权重之和。也就是说，$We = \sum_{\mathrm{k}=1}^{Le} \mathrm{w_k}$，其中$L_e = \left\lfloor \log_2 \frac{|V|}{s_i} + 1 \right.$

证明。给定任何超边 e，HyperLap 首先随机选择与给定权重成比例概率的合适级别。因此，该水平的概率ℓ 要选择的是$w\ell$ /We一旦确定了等级，2 个$\ell - 1$个级别的组ℓ 随机均匀选择，即概率为$1/2^{\ell-1}$.由 S 生成 e 的概率$S_{\mathrm{g}}^{(\mathrm{l})}$是两个概率的乘积，因此等式（2）成立。

对于每个节点v，设$\hat{d}_v^{(\ell)}$是在级别生成的超边中包含节点v的超边数ℓ. 然后，输出超图中v的度$\hat{d}v$是的和所有ℓ级别超图中$\hat{d}_v^{(\mathrm{l})}$的和，即$\widehat{d_v} = \sum_{\mathrm{l}=1}^{L} \widehat{d_v^{(\mathrm{l})}}$设$d_{\max} \coloneqq \max_{k \in \{1, \cdots, |V|\}} d_k$和$s_{\max} = \max_{k \in \{1, \cdots, |E|\}} s_k$。假设$|V| \gg 2^{L-1} \cdot d_{max}$和$\sum_{j \in S_g^{(\mathrm{l})}} d_j \gg d_{\max} \cdot s_{\max}$ 对所有$S_g^{(l)}$都成立。那么，

$$\mathbb{E}[\hat{d}_v] = \sum_{\ell=1}^{L} \mathbb{E}[\hat{d}_v^{(\ell)}] = \sum_{\ell=1}^{L} \sum_{e \in E} P[e \subseteq S_g^{(\ell)}(v)] \cdot P[v \in e | e \subseteq S_g^{(\ell)}(v)]$$
$$\approx \sum_{e \in E} \sum_{\ell=1}^{L_e} \left(\frac{w_\ell}{W_e} \cdot \frac{1}{2^{\ell-1}} \right) \left[|e| \cdot \left(\frac{d_v \cdot 2^{\ell-1}}{\sum_{j=1}^{|V|} d_j} \right) \right]$$
$$= \frac{d_v}{\sum_{j=1}^{|V|} d_j} \cdot \sum_{e \in E} \left(|e| \cdot \sum_{\ell=1}^{L_e} \frac{w_\ell}{W_e} \right) = d_v \cdot \frac{\sum_{e \in E} |e|}{\sum_{j=1}^{|V|} d_j} = d_v,$$

此处$S_g^{(l)}$(v)是一个包含v水平的群体ℓ。也就是说，正如我们在[1]中经验证实的那样，预计$\hat{d}v$将接近于dv。

HyperLap 背后的直觉：在本节中，我们提供了一些理由，说明为什么我们希望 HyperLap 能够准确地再现第 4 节中发现的超边的真实重叠模式。

•对于属于同一小组的一对或三对节点，在它们处重叠的超边数量预计会很高。因此，在每对或每三重上重叠超边的数量的分布预计是倾斜的。

•由于超边可以在一个包含结构相似节点的小组内形成，因此每个超边的同质性预计很高。此外，由于组的大小不同，超边的同质性预计会因生成它们的组的大小而不同.

•由于每个节点 v 的自我网络中的超边可能包含与 v 属于同一小组的节点，因此其密度和重叠度预计较高。

5.2 HyperLap+：参数选择

Algorithm 3: HyperLap$^+$: Automatic Parameter Selection

Input : (1) input hypergraph $G = (V, E)$
(2) update resolution p
Output: synthetic hypergraph $\hat{G} = (\hat{V}, \hat{E})$

$\hat{G} = (\hat{V}, \hat{E}) \leftarrow$ run HyperCL using the distributions in G
for each level $\ell = 2, \cdots, L$ **do**
  $i^* \leftarrow \arg\min_{i \in \{1, \cdots, 1/p\}} HHD\left(G, \text{update}(\hat{G}, p \cdot i, \ell)\right)$
  $\bar{G} \leftarrow \text{update}(\hat{G}, p \cdot i^*, \ell)$
  if $HHD(G, \bar{G}) < HHD(G, \hat{G})$ **then** $\hat{G} \leftarrow \bar{G}$
  else break
return $\hat{G} = (\hat{V}, \hat{E})$

update$(\hat{G} = (\hat{V}, \hat{E}), q, \ell)$
  $\bar{G}(\bar{V}, \bar{E}) \leftarrow \hat{G}(\hat{V}, \hat{E})$
  remove $(q \cdot 100)\%$ of the hyperedges created at level $\ell - 1$
  create the same number of hyperedges at level ℓ
  return $\bar{G} = (\bar{V}, \bar{E})$

给定一个输入超图G，我们如何设置 HyperLap 的参数（即级别 L 的数量和每个级别的权重$\{w1,\cdots,wL\}$），以便生成一个合成超图$\hat{G}$，特别是与目标真实世界超图相似的超图？应仔细调整参数，因为生成的超图的结构属性因其设置而异。为此，我们提出了 HyperLap+，它可以自动调整参数。

超边齐性目标：HyperLap+使用输入超图G和生成超图$\hat{G}$之间的超边齐性距离HHD（G，$\hat{G}$）作为其目标函数。它被定义为G和$\hat{G}$的超边齐性分布之间的 Kolmogorov-Smirnov D 统计量。即：

$$HHD\left(G,\hat{G}\right)=\max_{x}\left\{\left|F(x)-F^{'}(x)\right|\right\} \qquad (3)$$

其中F和F'分别是超图G和$\hat{G}$的累积超边齐性分布。然后，假设给定了级别L的数量，HyperLap+的目标是找到使超边性距离最小化的级别权重。也就是说，HyperLap+旨在解决以下优化问题：

$$\max_{w1,\ldots,wL} HHD\left(G,\hat{G}\right)$$

其中我们假设$w1+\cdots+wL=1$，因为只有权重之间的比率才重要。

优化方案：定义了目标后，我们描述了 HyperLap+如何将其最小化。为避免空组，请使用$L\leq log2\,|V|$应保持不变，并且L的级别数初始化为$\lfloor log2\,|V|\rfloor$。

由于有无限多个级别权重$w1,\cdots,wL$的组合，我们提出了一种算法 3 中描述的高效贪婪优化方案，其中在较低级别创建的超边的一部分被替换为在较高级别新创建的超边的一部分，重复，直到等式（3）收敛。

具体来说，hyperlap 先通过 hypercl 生成一个超图，等价于 $L=1$的 hyperlap (第 1 行)。这相当于把 $w1$设置为 1，把 wl 设置为 0。然后在每个 层次l 从 2 到 L，我们寻找在层次$l-1$创建的超边的一个最优分数，用那些在 层次l 新创建的超边替换(第 3 行)。请注意，只有大小 $\frac{|v|}{2^{l-1}}$或更小的超边才能被替换。如果替换严格减小超边同质距离，则超重叠 更新当前的合成超图(第 5 行)。这相当于减少了 w_{l-1}和增加了同样数量的 w_l。否则，我们返回当前的合成超图(第 6 行)。我们将更新分辨率 p 调整为 0.05。我们注意到生成的超图的质量对 p 的选择不敏感。

5.3 生成超图质量的经验评价

超重叠生成的超图能多好地再现输入超图的结构性质？我们通过比较它们与四个强基线: hypercl，hyperpa [12] ，hyperff [22]和天真地调整 hyperlap 来评估它的有效性。我们在附录 b 中描述了详细的实验设置。

为了度量由实际超图和生成超图导出的分布之间的相似性，我们使用 kolmogorov-smirnov D-统计量，定义为 $D=max_x\left|f'(x)-f(x)\right|$，其中$x$是所考虑的随机变量的值，$f'$和 f 是实际和相应生成的分布的累积分布函数。

观察 1 和 2: 在表 8 中，我们报告了真实世界超图和相应的合成超图中自我网络密度和自我网络重叠度分布之间的 D-统计量。

Hyperlap+生成由自我网络组成的超图，这些自我网络在结构上与真实世界的超图最相似。具体来说，hyperlap 给出了比最近提出的 hyperpa 更相似的自我网络密度分布和 2.35 倍的自我网络重叠分布。

观察 3 和 4: 我们用可视化和统计学的方法来检验超重叠 生成的超图是否遵循观察值 3 和 4。在表 9 中，我们说明了在每一对和每三个节点上重叠的超边的数量分布。与 hypercl

相比，hyperlap 能更好地再现节点对和节点三元组的度数。这在表 10 中得到了统计上的证实，超圈 给出了最小的 D-统计量。此外，这些分布在大多数数据集中是重尾分布，从至少有一个似然比是正值这一事实中可以看出(见第 4.2 节有关统计检验的详细信息)。

表 7: 实际超图中超边同质性的分布与 HyperLap+生成的超边同质性的分布是重尾的。对数似然比按表 6 计算。

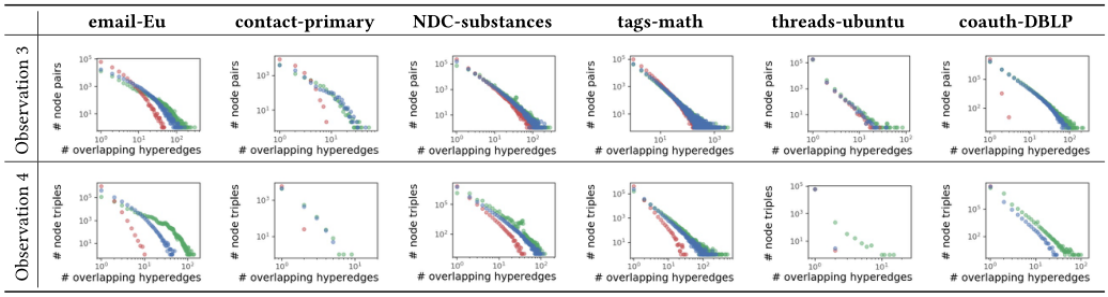
Dataset	Real-World Data			Generated		
	pw	tpw	logn	pw	tpw	logn
email-Enron	-1.09	-0.26	-0.38	-2.71	-0.43	-4.76
email-Eu	**0.90**	**0.90**	**0.91**	-3.00	**3.13**	**2.08**
contact-primary	**2.19**	**2.30**	**2.22**	**0.67**	**2.26**	**1.90**
contact-high	**1.55**	**1.55**	**1.95**	**2.50**	**4.72**	**3.65**
NDC-classes	0.00	**0.39**	**0.18**	-0.47	**0.87**	**0.52**
NDC-substances	**0.64**	**1.22**	**1.13**	**1.87**	**2.90**	**2.58**
tags-ubuntu	**2.25**	**2.25**	**2.26**	-2.01	**7.00**	**6.19**
tags-math	-17.66	-7.93	**2.62**	**3.53**	**6.56**	**6.07**
threads-ubuntu	**4.58**	**7.70**	**6.55**	**3.92**	**4.25**	**3.94**
threads-math	-0.72	**9.00**	**6.69**	**4.30**	**12.10**	**10.53**
coauth-DBLP	**4.01**	**4.31**	**4.20**	**10.65**	**25.23**	**22.82**
coauth-geology	**4.29**	**5.52**	**5.37**	**1.75**	**8.06**	**7.00**
coauth-history	-	-	**1.73**	**3.98**	**4.31**	**4.02**

表 8:HyperCL（H-CL）、HyperPA（H-PA）、HyperFF（H-FF）、HyperLap（H-LAP）和 HyperLap+（H-LAP+）五种模型生成的现实世界超图和相应超图中（1）egonet 密度、（2）egonet 重叠和（3）超边同质性分布之间的 D 统计。HyperLap+最准确地再现分布。

Dataset	Density of Egonets (Obs. 1)					Overlapness of Egonets (Obs. 2)					Homogeneity of Hyperedges (Obs. 5)				
	H-CL	H-PA	H-FF	H-LAP	**H-LAP⁺**	H-CL	H-PA	H-FF	H-LAP	**H-LAP⁺**	H-CL	H-PA	H-FF	H-LAP	**H-LAP⁺**
email-Enron	0.545	0.202	0.391	0.405	**0.125**	0.517	0.398	0.398	0.391	**0.111**	0.498	0.241	0.656	0.191	**0.136**
email-Eu	0.724	-	0.402	0.577	**0.310**	0.534	-	0.639	0.432	**0.197**	0.505	-	0.688	0.247	**0.168**
contact-primary	0.896	0.537	0.975	0.334	**0.128**	0.867	0.471	0.942	0.285	**0.095**	0.430	0.236	0.484	**0.142**	0.188
contact-high	0.948	0.529	0.880	0.522	**0.345**	0.874	0.431	0.703	0.486	**0.296**	0.423	0.196	0.336	**0.120**	0.178
NDC-classes	0.694	0.785	0.731	0.696	**0.635**	0.302	0.715	0.406	**0.231**	0.248	0.274	0.410	0.484	0.272	**0.225**
NDC-substances	0.451	-	0.801	0.426	**0.366**	0.321	-	0.338	0.243	**0.157**	0.377	-	0.740	0.262	**0.108**
tags-ubuntu	0.522	**0.162**	0.216	0.410	0.300	0.432	**0.117**	0.398	0.487	0.210	0.245	0.136	0.844	0.105	**0.011**
tags-math	0.496	0.350	0.561	**0.195**	0.227	0.460	0.325	0.709	**0.151**	0.186	0.337	0.217	0.921	0.086	**0.015**
threads-ubuntu	0.159	0.856	-	0.163	**0.159**	0.299	0.953	-	0.300	**0.297**	0.020	0.291	-	0.016	**0.011**
threads-math	0.137	0.492	-	**0.120**	0.135	0.232	0.714	-	0.235	**0.229**	0.060	0.368	-	0.102	**0.019**
coauth-DBLP	0.228	-	-	0.227	**0.132**	0.302	-	-	0.267	**0.244**	0.715	-	-	0.540	**0.026**
coauth-geology	0.200	-	-	0.202	**0.138**	**0.248**	-	-	0.252	0.266	0.624	-	-	0.481	**0.044**
coauth-history	**0.087**	-	-	0.090	0.089	**0.316**	-	-	0.321	0.324	0.154	-	-	0.125	**0.020**
Average	0.468	0.489	0.619	0.335	**0.237**	0.439	0.515	0.566	0.313	**0.219**	0.358	0.261	0.644	0.206	**0.088**

-: out of time (taking more than 10 hours) or out of memory

表 9:HyperLap+精确再现了每对和每三个节点上重叠超边的数量分布，而 HyperCL 在许多情况下失败。它们服从重尾分布，就像真实的分布一样。



观察 5: 从表 8 的结果可以看出，**实际中超边均匀性的分布与超重叠生成的对应超图之间的数据统计量极小。**由于 hyperlap+的目标是减少 HHD，因此它自然比 hypercl 更好地再

现超边均匀性，而 hypercl 在参数天真地设置的情况下，其性能优于 hyperlap。这一结果表明了所提出的优化方案的有效性。如表 7 所示，超重叠生成的超图中超边均匀性的分布是重尾的(统计检验的细节见第 4.2 节)。

表 10: 实际超图中每对和每三个节点的重叠超边数目分布与由五种模型生成的对应超图之间的 d 统计量: hypercl (h-cl)、 hyperpa (h-pa)、 hyperff (h-ff)、 hyperlap (h-lap)和 hyperlap (h-lap)。超重叠 最精确地再现了分布，这些分布遵循重尾分布。

Dataset	Pair of Nodes (Obs. 3)								Triple of Nodes (Obs. 4)							
	Distance from Real (D-statistics)					Heavy-tail Test			Distance from Real (D-statistics)					Heavy-tail Test		
	H-CL	H-PA	H-FF	H-LAP	**H-LAP⁺**	pw	tpw	logn	H-CL	H-PA	H-FF	H-LAP	**H-LAP⁺**	pw	twp	logn
email-Enron	0.143	**0.056**	0.217	0.075	0.139	-2.37	-0.29	-1.53	0.089	0.295	0.136	**0.061**	0.072	-0.22	**0.38**	**0.24**
email-Eu	0.225	-	0.352	0.162	**0.066**	**0.24**	**2.75**	**2.53**	0.480	-	0.516	0.337	**0.206**	**0.41**	**2.11**	**1.96**
contact-primary	0.196	0.062	0.223	0.070	**0.051**	**9.53**	**15.74**	**13.92**	0.137	0.061	0.110	0.053	**0.031**	-1.86	-1.27	**1.23**
contact-high	0.277	**0.062**	0.141	0.127	0.067	-3.09	-0.95	-0.06	0.210	**0.131**	0.182	0.182	0.193	-3.95	-	**0.50**
NDC-classes	0.273	0.197	0.196	0.246	**0.172**	**12.15**	**14.42**	**14.04**	0.376	**0.167**	0.405	0.349	0.286	**3.22**	**7.92**	**7.34**
NDC-substances	0.272	-	0.244	0.251	**0.202**	**33.69**	**40.13**	**39.66**	0.521	-	0.591	0.492	**0.453**	**45.30**	**55.38**	**54.99**
tags-ubuntu	0.091	**0.019**	0.182	0.034	0.033	**42.33**	**43.70**	**43.55**	0.148	0.067	0.191	**0.020**	0.074	**14.25**	**15.57**	**15.43**
tags-math	0.095	0.066	0.278	0.073	**0.011**	**42.75**	**45.60**	**45.41**	0.209	**0.053**	0.286	0.113	0.079	**21.38**	**23.12**	**22.99**
threads-ubuntu	0.011	0.137	-	**0.008**	0.009	**1.28**	**1.75**	**1.75**	**0.004**	0.130	-	**0.004**	**0.004**	-1,346	-1.72	-1.72
threads-math	0.041	0.163	-	**0.014**	0.033	**15.79**	**16.66**	**16.52**	0.006	0.138	-	**0.001**	0.005	-1.49	-0.98	**0.96**
coauth-DBLP	0.224	-	-	0.191	**0.032**	**55.86**	**74.95**	**73.45**	0.215	-	-	0.214	**0.192**	**2.87**	**6.73**	**6.46**
coauth-geology	0.178	-	-	0.157	**0.040**	**31.13**	**45.08**	**44.06**	0.086	-	-	0.085	**0.069**	-0.10	**1.10**	**0.84**
coauth-history	0.033	-	-	0.030	**0.009**	**1.74**	**1.77**	**1.63**	**0.001**	-	-	**0.001**	**0.001**	-0.86	-	**0.57**
Average	0.158	0.095	0.229	0.110	**0.066**				0.193	0.130	0.302	0.147	**0.128**			

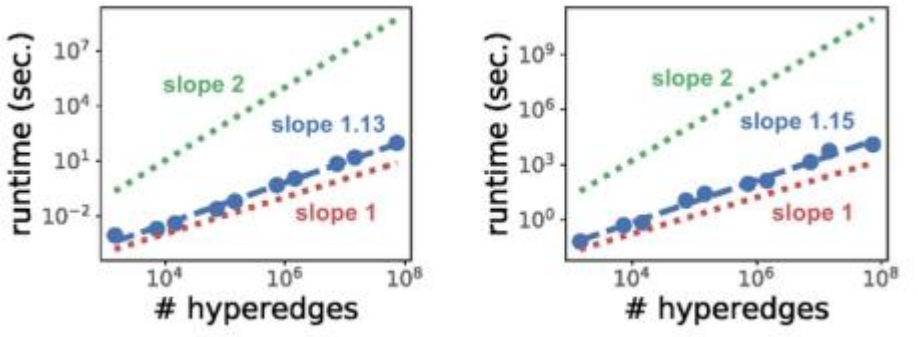
-: out of time (taking more than 10 hours) or out of memory

5.4 Hyperlap 和 Hyperlap+的可扩展性

本部分从理论和实验两个方面分析了 Hyperlap 和 Hyperlap+的可扩展性。值得注意的是，我们根据经验证明 Hyperlap 和 Hyperlap+ 的标度几乎与所考虑的超图的大小成线性关系。

事实上，虽然一些基线在特定的数据集中是难以处理的，但是 Hyperlap 和 Hyperlap+可以在所有考虑过的数据集中执行。Hyperpa 的可伸缩性在很大程度上取决于超边的大小，因此不适用于包含大型超边的超图(即 email-eu、ndc-substances、coauth-dblp、coauth-geology 和 coauth-history)。Hyperff 依赖于节点的数量，不能在具有多个节点的大型数据集中工作(即，threads-ubuntu、 threads-math、 coauth-dblp、 coauth-geology 和 coauth-history)。

根据每一关的级别和重量，运行超重圈需要多少时间？假设所有集合和映射都使用哈希表实现。对于每个超边 e，在 $o(1)$时间内选择层 l 和组。此外，由于每个节点都是独立采样的，$|e|$ 节点是在 $o(|e| \cdot (1+\varepsilon))$时间内选择的，因此 ε 是由于碰撞的可能性(即对于一个超边选择多个节点)。ε 取决于节点的度数和超边的大小。我们注意到经验上 ε 在考虑的数据集中非常小。因此，生成 $|e|$ 超边需要$o(\sum e \in e(|e| \cdot (1+\varepsilon)))$时间。在 hyperlap 中，我们考虑替换步骤。在每个层次上，最多$\frac{1}{p}|E| = o(|e|)$超边被(临时)替换，取 $o(\sum e \in e(|e| \cdot (1+\varepsilon)))$时间。由于最大能级数为 $log_2|v|$，因此 hyperlap 总共需要$o(log_2|v|\sum e \in e(|e| \cdot (1+\varepsilon)))$时间。



(a) HyperLap (generation) (b) HyperLap⁺ (generation & fitting)

图 2：超重叠和超重叠 与所考虑的超图的大小成线性关系。

在图 2 中，我们使用不同大小的合成超图来测量 hyperlap 和 hyperlap+的运行时间。它们是通过将最小的超图-email-Enron 上升 5 到 5 万倍，使用 hyperlap。hyperlap 和 hyperlap+几乎与所考虑的超图的大小成线性关系。特别是 hyperlap+在几个小时内，生成并符合一个合成的超图，它有 7 亿条超边。我们在附录 b 中描述了详细的实验设置。

6. 结论

在本文中，我们研究了来自 6 个领域的十三个实际超图的超边重叠的结构性质。为此，我们定义了一些原则性的测度，并基于观察，我们开发了一个现实的超图生成模型。我们将我们的贡献总结如下。

•现实世界超图中的观测: 我们发现了现实世界超图中超边重叠的三个独特性质。我们使用随机超图来验证这些性质，其中节点的度数和超边的大小都得到了很好的保持。

•新的度量: 我们提出超边的重叠性和同质性。我们通过一个公理化的方法来证明重叠性是一个原则性的度量。同质性揭示了一个有趣的重叠模式，在此基础上我们发展了一个现实的生成模型。

•现实的生成模型: 我们提出 hyperlap，一种超图生成模型，它能精确地重现现实世界超图中超边的重叠模式。我们还提供了 hyperlap+ ，它可以自动拟合超重叠的参数到给定的图中。他们在几个小时内生成并适应了一个有 7 亿条超边的超图。

重现性：这项工作中使用的源代码和数据集可在 https://github.com/young917/www21-hyperlap 获得。

致谢 金志秀博士富有成果的讨论。这项研究由韩国国家研究基金会(nrf)资助，韩国政府(msit)(第号)。Nrf-2020r1c1c1008296)和韩国政府资助的信息与通信技术规划与评估研究所(iip)补助金(编号: 20190-00075，人工智能研究生院计划(kaist))。

参考文献

[1] 2021. Online Appendix. https://github.com/young917/www21- hyperlap.

[2] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 2014. powerlaw: a Python package for analysis of heavy-tailed distributions. PloS one 9, 1 (2014), e85777.

[3] Ilya Amburg, Nate Veldt, and Austin Benson. 2020. Clustering in graphs and hypergraphs with categorical edge labels. In WWW.

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. Science 286, 5439 (1999), 509–512.

[5] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. Simplicial closure and higher-order link prediction. PNAS 115, 48 (2018), E11221–E11230.

[6] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2018. Sequences of sets. In KDD.

[7] Berge C. 2013. Hypergraphs. Vol. 45. North Holland, Amsterdam.

[8] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. 2004. R-MAT: A recursive model for graph mining. In SDM.

[9] Philip S Chodrow. 2020. Configuration models of random hypergraphs. Journal of Complex Networks 8, 3 (2020), cnaa018.

[10] Fan Chung and Linyuan Lu. 2002. The average distances in random graphs with given

expected degrees. PNAS 99, 25 (2002), 15879–15882.

[11] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. SIAM review 51, 4 (2009), 661–703.

[12] Manh Tuan Do, Se-eun Yoon, Bryan Hooi, and Kijung Shin. 2020. Structural patterns and generative models of real-world hypergraphs. In KDD.

[13] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the internet topology. ACM SIGCOMM computer communication review 29, 4 (1999), 251–262.

[14] Nikhil Goyal, Harsh Vardhan Jain, and Sayan Ranu. 2020. GraphGen: A Scalable Approach to Domain-agnostic Labeled Graph Generation. In WWW.

[15] Cesar A Hidalgo and Carlos Rodríguez-Sickert. 2008. The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387, 12 (2008), 3017–3024.

[16] Shuguang Hu, Xiaowei Wu, and TH Hubert Chan. 2017. Maintaining densest subsets efficiently in evolving hypergraphs. In CIKM.

[17] TaeHyun Hwang, Ze Tian, Rui Kuangy, and Jean-Pierre Kocher. 2008. Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction. In ICDM.

[18] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. 2019. Dynamic Hypergraph Neural Networks.. In IJCAI.

[19] Michał Karoński and TomaszŁuczak. 2002. The phase transition in a random hypergraph. J. Comput. Appl. Math. 142, 1 (2002), 125–135.

[20] George Karypis, Rajat Aggarwal, Vipin Kumar, and Shashi Shekhar. 1999. Multilevel hypergraph partitioning: applications in VLSI domain. TVLSI 7, 1 (1999), 69–79.

[21] Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In European Conference on Machine Learning. Springer.

[22] Yunbum Kook, Jihoon Ko, and Kijung Shin. 2020. Evolution of Real-world Hypergraphs: Patterns and Models without Oracles. ICDM (2020).

[23] Dongjin Lee, Kijung Shin, and Christos Faloutsos. 2020. Temporal locality-aware sampling for accurate triangle counting in real graph streams. The VLDB Journal 29, 6 (2020), 1501–1525.

[24] Geon Lee, Jihoon Ko, and Kijung Shin. 2020. Hypergraph Motifs: Concepts, Algorithms, and Discoveries. PVLDB 13 (2020), 2256–2269. Issue 11.

[25] Jure Leskovec and Christos Faloutsos. 2007. Scalable modeling of real graphs using kronecker multiplication. In ICML.

[26] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In KDD.

[27] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. TKDD 1, 1 (2007), 2–es.

[28] Pan Li and Olgica Milenkovic. 2017. Inhomogeneous hypergraph clustering with applications. In NeurIPS.

[29] Pan Li and Olgica Milenkovic. 2018. Submodular hypergraphs: P-Laplacians, cheeger inequalities and spectral clustering. In ICML.

[30] Mingsong Mao, Jie Lu, Jialin Han, and Guangquan Zhang. 2019. Multiobjective ecommerce recommendations based on hypergraph ranking. Information Sciences 471 (2019), 269–

287.

[31] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. 2015. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. PloS one 10, 9 (2015), e0136497.

[32] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. Science 303, 5663 (2004), 1538–1542.

[33] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. Science 298, 5594 (2002), 824–827.

[34] Ali Pinar, Comandur Seshadhri, and Tamara G Kolda. 2012. The similarity between stochastic kronecker and chung-lu graph models. In SDM.

[35] Kijung Shin, Tina Eliassi-Rad, and Christos Faloutsos. 2018. Patterns and anomalies in k-cores of real-world graphs with applications. Knowledge and Information Systems 54, 3 (2018), 677–710.

[36] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In WWW.

[37] Despina Stasi, Kayvan Sadeghi, Alessandro Rinaldo, Sonja Petrovic, and Stephen Fienberg. 2014.β models for random hypergraphs with a given degree sequence. In COMPSTAT.

[38] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, JeanFrançois Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. 2011. High-resolution measurements of face-to-face contact patterns in a primary school. PloS one 6, 8 (2011), e23176.

[39] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of‘smallworld’networks. Nature 393, 6684 (1998), 440–442.

[40] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. 2019. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In NeurIPS.

[41] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudre-Mauroux. 2019. Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. In WWW.

[42] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In KDD.

[43] Se-eun Yoon, Hyungseok Song, Kijung Shin, and Yung Yi. 2020. How Much and When Do We Need Higher-order Informationin Hypergraphs? A Case Study on Hyperedge Prediction. In WWW.

[44] Jiaxuan You, Rex Ying, Xiang Ren, William L Hamilton, and Jure Leskovec. 2018. GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models. In ICML.

[45] Jun Yu, Dacheng Tao, and Meng Wang. 2012. Adaptive hypergraph learning and its application in image classification. TIP 21, 7 (2012), 3262–3272.