子图匹配中的结构等价

Dominic Yang、Yurun Ge、Thien Nguyen、DenaliMolitor、JacobMoorman、AndreaL.Bertozzi，电气电子工程师学会会员。

arXiv:2301.03161v1 [cs.DS] 9 Jan 2023

摘要—对称性在子图匹配中发挥着重要作用，既体现在对相关图的描述上，也体现在它对搜索过程的干扰上。本研究探讨如何量化这些影响，以及如何利用对称性提高子图同构算法的效率。我们提出结构等价的严格定义，并确立可安全地利用结构等价生成更多解的条件。我们说明如何调整标准搜索例程，以利用这些对称性加速搜索并紧凑地描述解空间。随后，我们调整一种最先进的求解器，并进行一系列全面测试，以展示这些方法在标准基准测试集上的效用。我们将这些方法扩展到多层图，并展示了从交通系统、社交媒体、对抗性攻击和知识图谱中提取的大型多重网络上的结果。

索引词— 子图同构、子图匹配、多重网络、结构等价、图结构

I. 引言

子图同构问题（也称子图匹配问题）给定一个小图（**模板**），并要求在一个更大的（**世界图**）图中找到它作为子图。这个问题已得到广泛研究，尤其是在模式识别领域。综述 [20], [9],和 [28] 介绍了各种各样的技术及其应用，包括手写识别 [2], 、人脸识别 [4],、生物医学应用 [3], 、数独谜题和对抗性活动[51]。近年来，子图匹配也成为模体发现 [23], [37], 中的一个环节，通过它可以发现频繁子图，用于社交网络和生化数据等领域的图分析。此外，子图匹配也与知识图谱搜索相关，可通过查询知识数据库来补全不完整事实陈述 [8], [47]。

网络广泛存在于各种应用中；因此，检测网络中显现的有趣结构，即子图，具有重要意义。我们研究了多种网络中的子图匹配，包括模拟网络和真实网络、单通道网络和多通道网络，其规模从数百个节点到数百万个节点不等。这些数据集包括生化反应 [21], 、模式识别 [12], 、交通网络[13],、社交网络 [16],和知识图谱 [52]。

本文研究精确子图同构：给定一个模板 $G_T = (V_T, E_T)$ 和一个世界图 $G_W = (V_W, E_W)$，求

DominicYang、Yurun Ge、Denali Molitor、JacobMoorman 和 Andrea L. Bertozzi 任职于加利福尼亚大学洛杉矶分校数学系，地址为美国加利福尼亚州洛杉矶市，邮编 90095。电子邮箱：domyang@math.ucla.edu、yurun@math.ucla.edu、dmolitor@math.ucla.edu、jdmoorman@math.ucla.edu、bertozzi@math.ucla.edu。

Thien Nguyen 任职于波士顿马萨诸塞州 02115 东北大学计算机科学系。电子邮箱：nguyen.thien@northeastern.edu

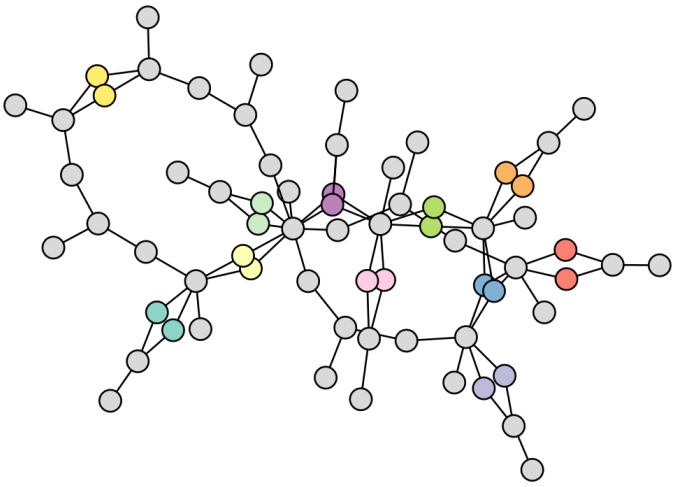


图1. 表示一个生化反应系统的图，取自 [21]。非灰色节点中，颜色相同的节点在结构上等价。

一个映射 $f : V_T \to V_W$，它既是单射，又尊重 G_T 的结构。要满足后一性质，我们要求：如果 $(t_1, t_2) \in E_T$，那么必须有 $(f(t_1)\ f(t_2)) \in E_W$。如果满足这一条件，我们就称 f 是**保边**的。我们将子图同构定义如下：

定义 1。给定一个模板 $G_T = (V_T, E_T)$ 和一个世界图 $G_W = (V_W, E_W)$，若映射 $f : V_T \to V_W$ 是子图同构，当且仅当 f 是单射且保边。

本文中，我们将“子图同构”和“子图匹配”这两个术语交替使用。相关术语还包括**子图同态**，它放宽了单射要求；以及**诱导子图同构**，它还要求映射保非边（如果 $(u, v) \notin E_T, (f(u)\ f(v)) \notin E_W$）。

我们关注子图匹配问题（SMP）[51]:

定义 2（**子图匹配问题**）。给定模板图 G_T 和世界图 G_W，求出从 G_T 到 G_W 的所有子图同构。

如果至少存在一个子图同构，我们就称该问题**可满足**。

仅仅是寻找一个子图同构就是 NP 完全问题 [6],，这表明不存在能够在所有图上高效找出全部子图同构的算法。尽管如此，在开发用于检测子图同构的算法方面，已经取得了显著进展 [18], [22], [27], [45]。与其他 NP 完全问题一样，相关文献主要关注

精确子图同构的枚举，重点是对解空间进行完整的树搜索。最先进的算法通常侧重于迭代构建部分匹配，并采用启发式方法优化变量排序和剪枝搜索树中的分支。

我们关注 SMP 解空间的完整枚举与刻画。这对实际应用至关重要。设想一个代表犯罪团伙交易的模板，而世界图则是更广泛的交易网络。如果存在数千个匹配，那么任意一个匹配都很可能是假阳性，这说明有必要通过向模板添加更多细节来进一步缩小搜索范围。我们的工作识别并刻画了匹配问题中由对称性导致的冗余结构。我们利用这些对称性生成解空间的压缩版本，既节省空间，也有助于理解问题的解。

以对称性为例，观察图1中的模板图，它来自一个生化反应系统 [21]；注意，在任何解中，每对颜色相同的节点都可以互换，因为它们的邻居完全相同。由于图中有 11 对这样的节点，对于找到的任意一个同构，我们只需互换节点，就能生成 $2^{11} = 2048$ 个更多解。通过在子图搜索中避免冗余解，我们可以显著缩短搜索时间（潜在地缩短 2048 倍甚至更多）。这种简单形式的对称性称为**结构等价**。

更广义的等价性概念可以进一步加速搜索。在图 2 中，黄色节点和蓝色节点各自都是结构等价的。然而，如果我们先将 A 与 1 匹配，那么只需将 B 和 C 分别与 2、3、4 或 5 中的任意节点匹配，就能完成一个同构。4 和 5 上额外相连的边掩盖了这样一种可能：可以用 2 或 3 替换 4 或 5。通过识别何时可以忽略这些额外边，我们同样能够大幅减少工作量。我们将第二种等价性称为**候选等价**。在本文正文中，我们将对这些术语作出正式定义，并说明如何将它们应用于树搜索算法。这两种等价性概念大致可分为两类：**静态等价性**，描述从问题描述中即可看出的等价性；以及**动态等价性**，描述在搜索过程中发现的等价性。结构等价属于前一类，而候选等价属于后一类。需要指出的是，这些形式的等价性取决于图的结构，因此，如果将一个模板图换成另一个模板图，就不能再使用它们。

A. 相关工作

Ullmann 提出的第一个重要子图同构算法 [1] 会根据已匹配世界图顶点的邻居生成候选顶点。广泛使用的 VF2算法 [7] 在此基础上进行了改进：它选择一种匹配顺序，优先匹配与已匹配模板顶点相邻的模板顶点，并根据顶点度数添加剪枝规则。作者最近又发表了 VF3算法 [31]，进一步扩展了这一方法。

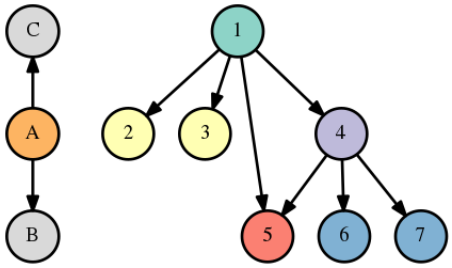


图 2. 子图同构问题示例：左侧为模板图，右侧为世界图。同色节点在结构上等价。

在另一种方法中，Solnon [10] 强调了人工智能中的约束传播范式，她的 LAD 算法会在内部为每个模板顶点存储候选列表。通过反复应用基于顶点邻域结构的过滤器，她的算法能够有效剪除树搜索中的分支。她在 [30] 中进一步拓展了这项工作，纳入了更强大的过滤器。其他求解器，包括 Glasgow [24], [45], 、SND [19], 和 ILF [11], 分别探讨了如何基于其他图属性增强过滤器，这些属性包括路径数量、团或更复杂的邻域结构，以加强过滤效果。

已有大量研究致力于利用对称性压缩图并计数同构。TurboIso [18] 利用模板图中的基本对称性，并根据候选区域的选择及对这些区域的探索来优化匹配顺序。CFL-Match [27] 提出了一种匹配顺序，该顺序基于将模板图分解为核心（即高度连通子图）和森林，再将森林进一步分解为森林和叶节点。BoostIso[25] 利用世界图中的对称性，并提出了一种方法，通过采用其方法来加速其他基于树搜索的方法。ISMA算法[17] 利用模板的基本双侧对称性和旋转对称性来加速子图搜索；此后，这项工作扩展为ISMAGS算法 [22], 以纳入模板图的一般自同构对称性。

一个密切相关的问题是近似子图匹配问题。该问题将精确同构中的严格保边约束替换为对边和标签不匹配的惩罚，并以最小化该惩罚为目标。惩罚的具体形式因应用而异，相关方法也多种多样，其中许多借鉴了精确匹配问题。这些方法涉及各种技术，包括过滤、 [40], [47], A*搜索 [39], 索引 [32],以及连续技术 [44]。本文不考虑这一问题。

B. 论文结构

本文展示了如何通过利用搜索开始时即可观察到的静态等价形式，以及随着搜索进行而揭示的动态等价形式，加速面向树搜索的方法。第二节中，我们正式定义结构等价，并说明如何将其融入子图搜索。

第三节中，我们介绍候选等价，以说明如何在子图搜索过程中发现此前未曾显现的等价关系。第四节中，我们介绍节点覆盖等价，这是一种易于计算的替代等价形式，并将所有等价概念统一到一个层次结构中。第五节中，我们改造格拉斯哥求解器 [45]，使其能够利用等价关系，并将其应用于一组基准测试，以评估各个等价级别[1]的性能。第六节中，我们展示如何利用等价关系简洁地表示和可视化大规模解集。第七节中，我们扩展算法，使其能够处理多层多重图，并展示该算法在多种结构更复杂的网络上全面描绘解空间的能力。

本文从 [51]中汲取了通用子图树搜索结构的灵感，并与其共享了许多针对多通道网络的测试用例。在我们此前的工作[42],中，我们提出了更简单的候选等价和候选结构概念，并在少量多通道网络上进行了测试。通过这些先前的研究，我们观察到解空间具有很高的组合复杂度，因此需要一种能够利用对称性压缩解空间并加速搜索的方法。在本研究中，我们通过提出几种新的等价性概念，为其在子图搜索中的有效性奠定严谨的理论基础，建立解空间的紧凑表示，并在大量真实和合成数据集上对这些方法进行实证评估，从而拓展了上述两篇论文的工作。

II. 结构等价

结构等价是网络中一种易于理解的性质；如果网络具有这种性质，便可利用它大幅加快子图搜索。直观而言，若两个顶点可以在不改变图结构的情况下彼此“交换”，则它们在结构上等价。这种等价性常见于同时与同一顶点相邻的叶节点。

定义 3. 在图 $G = (V, E)$中，如果两个顶点v, w 满足以下条件，则称它们结构等价 （记作 $v \sim_s w$）：

1) *For* $u \in V, u \neq v, w,$
   a) $(u, v) \in E \Leftrightarrow (u, w) \in E$
   b) $(v, u) \in E \Leftrightarrow (w, u) \in E$
2) $(v, w) \in E \Leftrightarrow (w, v) \in E$

根据此定义，结构等价顶点的邻居（不包括顶点本身）必须相同。以下命题验证了这是一种等价关系。

命题 4. $\sim_s$ 是一种等价关系。

证明。本文中所述命题的所有证明均见附录。

□

利用这一关系，我们可以将任意图的顶点划分为结构等价类，并互换每个类中的成员，而不改变图的基本结构。在最坏情况下，判断两个顶点是否等价，只需在一次 $O(|V|)$ 操作中比较它们的邻居；但对于稀疏图，通常会更快。通过两两比较顶点来计算这些等价类，最坏情况下需要 $O(|V|^3)$ 次操作。算法 1 展示了如何实现广度优先搜索算法，利用图的稀疏性加快计算。由于对于每个已访问的顶点 v，将其邻居划分为等价类需要 $O(deg(v)^2)$，因此最坏情况下，该算法需要 $O(\sum_v deg(v)^2) \approx O(|V|deg(v)^2)$，其中 $deg(v)^2$ 表示所有 v 的 $deg(v)^2$ 的平均值。对于稀疏图，$deg(v) \ll |V|$，因此这会比朴素的两两比较快得多。

算法 1计算等价类的例程

```
1: 函数 FINDEQCLASSES(G = (V, E))
2:     令 Q为一个队列
3:     选择第一个顶点 v 放入 Q
4:     令 EQ = {}
5:     令已访问集合 = {}
6:     当 Q 不为空时 执行
7:         从 Q中移出 v
8:         将 v 添加到已访问集合
9:         将 N(v)划分为等价类，划分到 EQ
10:        将邻居类中的代表元添加到 Q
11:    检查第一个顶点 v 是否属于某个类，若是则将其添加进去
12:    否则，将其添加到自己的类中
13:    返回 EQ
```

A. 可互换性与同构计数

现在我们证明，给定任意一个子图同构，我们可以互换模板图中的任意两个顶点，同时仍保持子图同构。在此之前，我们先正式定义可互换性的含义。

定义 5. 如果对于所有子图同构 $f : V_T \to V_W$，通过互换 v 和 w 得到的映射 g：中任意两个模板图顶点 v、$w \in V_T$ 都满足可互换，那么：

$$g(u) = \begin{cases} f(w) & u = v \\ f(v) & u = w \\ f(u) & otherwise \end{cases}$$

也是一个子图同构。

如果对于所有子图同构 f，当 v'、w' 都属于 f 的像，且其原像分别为 v、w 时，映射 g：中的任意两个世界图顶点 v'、$w' \in V_W$ 都满足可互换：

$$g(u) = \begin{cases} w' & u = v \\ v' & u = w \\ f(u) & otherwise \end{cases}$$

是一个同构。如果像中只有一个 v'，那么由下式给出的h

$$h(u) = \begin{cases} w' & u = v \\ f(u) & otherwise \end{cases}$$

[1]我们算法的实现可在以下代码仓库中找到：
https://github.com/domyang/glasgow-subgraph-solver。

也是一个同构。

在本文后文中，我们将通过将可互换性限制为仅适用于同构的某些子集来进一步限定这一定义。确认模板结构等价下模板顶点可互换性的命题如下：

命题 6. 给定图 $G_T = (V_T, E_T)$、$G_W = (V_W, E_W)$，如果 $v, w \in V_T$ 结构等价，那么它们在任何子图同构中都可以互换。

因此，互换两个模板顶点的像不会改变子图同构。由于换位可以生成所有排列，我们得到以下结果：

命题 7. 如果我们可以将 $V_T = C_1, \ldots, C_n$ 划分为结构等价类，并且至少存在一个子图同构，那么至少存在

$$\prod_{i=1}^{n} |C_i|!$$

个子图同构。

我们还可以将这种结构等价性应用于世界图，以展示类似的可互换性。

命题 8. 如果 $v', w' \in V_W$ 结构等价，那么在任何子图同构 f 中，它们都可以互换。

如果我们在问题中同时应用模板和世界结构等价，自然会想知道现在能从单个解生成多少个解。以下命题给出了答案：

命题 9. 设 $f : V_T \to V_W$ 是一个子图同构。假设我们将模板图 $V_T = \bigcup_{i=1}^{n} C_i$ 和世界图 $V_W = \bigcup_{j=1}^{m} D_j$ 划分为结构等价类。令 $C_{i,j} = C_i \cap f^{-1}(D_j)$ 表示 C_i 中映射到等价类 D_j 中世界顶点的模板顶点集合。那么共有

$$\prod_{i=1}^{n} |C_i|! \prod_{j=1}^{m} \prod_{i=1}^{n} \binom{|D_j| - \sum_{k=1}^{i-1} |C_{k,j}|}{|C_{i,j}|}$$

通过根据命题 *6* 和命题 *8* 互换等价模板顶点或世界顶点而生成的同构。

B. 树搜索的应用

现在我们将展示如何调整任意树搜索算法以纳入等价性。树搜索算法通过构造模板顶点到世界顶点的部分匹配来推进，并在每一步将下一个模板顶点赋值给其候选世界顶点之一，从而扩展匹配。如果在任意时刻匹配无法扩展（由于出现矛盾或找到完整匹配），则将最后赋值的模板顶点重新赋值给下一个候选顶点。模板顶点到世界顶点的每一种可能赋值都对应树中的一个节点，而从树根到叶节点的一条路径则对应顶点的一个完整映射。

树搜索是一种递归例程，如算法 2 所述。在此过程中，我们维护一个二元 $|V(T)| \times |V(W)|$ 矩阵 $cands$，其中 $cands[i, j]$ 若世界顶点 j 是模板顶点 i 的候选对象，则值为 1，否则为 0；此外，我们还维护一个从模板顶点到世界顶点的映射，用于描述哪些顶点已经匹配。第 2 至 4 行中，在所有顶点都完成匹配后，我们报告一次匹配。第 5 行调用应用筛选器，根据部分匹配中目前所做的假设排除候选项。第 7 行中，我们保存当前状态，以便回溯后返回；第 11 至 14 行中，我们遍历当前顶点的所有候选项，尝试进行匹配，直至全部检查完毕。随后在第 18 行中，我们恢复先前的状态并进行回溯。

算法 2树搜索的通用例程

```
1: 函数 SOLVE(部分匹配, 候选项)
2:     如果MatchComplete(部分匹配) 则
3:         ReportMatch(部分匹配)
4:         返回
5:     应用筛选器(部分匹配, 候选项)
6:     令 u = 获取下一个模板顶点()
7:     令候选项复制 =候选项.copy()
8:     如果使用世界等价 那么
9:         重新计算等价关系(部分匹配, 候选项)
10:    令世界顶点集合 =生成世界顶点(候选项，等价关系)
11:    对于 v 在世界顶点集合中执行
12:        部分匹配.匹配(u, v)
13:        求解(部分匹配，候选项副本)
14:        部分匹配.取消匹配(u, v)
15:        如果使用模板等价性则
16:            对于未匹配的 u' ~ u 执行
17:                设置候选项[u', v] = 0
18:    令 cands =cands 副本
19:    若使用世界等价 则
20:        恢复等价()
21:    返回
```

模板等价可以显著加速树搜索；根据命题 6，我们可以交换等价模板顶点的赋值，从而找到另一个同构。如果我们有一个部分匹配，其中包含模板顶点 $u_1 \sim u_2$，并且刚刚考虑了将候选对象 w 映射到 u_1，那么我们可以忽略将 u_2 映射到 w 的分支，因为我们可以通过交换一个将 u_1 映射到w 的同构来生成这些同构。第 16 行和第 17 行展示了如何将这一思路融入树搜索（如果不采用这些做法，我们就会使用标准的树搜索）。

为了将世界等价融入搜索，我们修改搜索过程，使得搜索中每个模板顶点只会被赋值给等价类中的一个代表元。我们可以通过修改生成世界顶点，使其从当前模板顶点的候选项中，为每个等价类仅选取一个代表顶点。

请注意，执行树搜索后，找到的解将代表可通过互换生成的解类。我们需要进行一些记账，以确定哪些赋值可以互换，从而计算不同解的数量。我们将实际找到的解（而非通过互换顶点生成的解）称为**代表元**

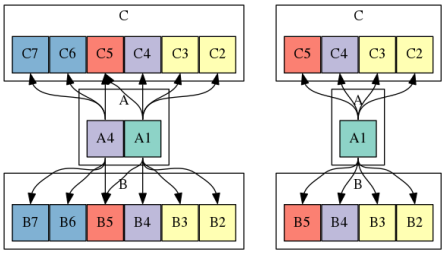


图 3. 将模板节点 A 赋值给世界节点 1 之前和之后，图 2 中各图的候选结构。

解。对于给定的代表解，通过互换等价顶点可以生成的解集合称为一个**解类**。用稀疏的解集表示庞大的解空间的能力，使我们能够紧凑地描述大规模解空间。

III. 候选等价性

上一节讨论的等价性是一种静态等价形式，仅考虑子图搜索开始时提供的信息。然而，随着子图搜索的进行，我们可能能够根据已有赋值推导出的信息，进一步舍弃节点和边。例如，在图 2 中，将 A 赋值给 1 后，我们可以舍弃节点 6 和 7 以及边 (4, 5)，因为如果 A 和 1 已经匹配，它们就不可能包含在匹配中。舍弃这些节点后，我们发现，就已有匹配而言，2、3、4 和 5 实际上可以互换。为了利用这种动态等价形式，我们需要引入一种辅助结构，以考虑我们对每个顶点候选项的了解 u（记作 $C[u]$）。

定义*10.* 给定模板图 $G_T = (V_T, E_T)$、世界图 $G_W = (V_W, E_W)$，以及每个 $u \in V_T$ 的候选集 $C[u] \subset V_W$，候选结构是有向图 $G_C = (V_C, E_C)$，其中顶点 $V_C = \{(u, c) : u \in V_T, c \in V_W\}$ 是模板顶点 - 候选项对，并且 $((u_1, c_1), (u_2, c_2)) \in E_C$ 当且仅当 $(u_1, u_2) \in E_T$ 且 $(c_1, c_2) \in E_W$。

候选结构既表示每个模板节点的候选项信息，也表示模板邻接关系与世界邻接关系如何相互作用。它去除了多余信息，从而揭示出仅观察原始图时不易发现的等价性。值得注意的是，这种数据结构类似于 [27] 中提出的紧凑路径索引（CPI）。不过，CPI 仅针对模板图的一棵给定有根生成树定义，而我们的候选结构则考虑了模板图的所有边。对于图 2 中的示例，假设候选集已缩减至最小候选集，使得 $C[A] = \{1, 4\}$ 和 $C[B] = \{2, 3, 4, 5, 6, 7\}$。那么，这些图的候选结构如图 3 左侧所示。此时，候选结构中尚无明显可利用的等价性。然而，一旦我们决定将节点 A 映射到 1，候选结构就会缩减为图 3 右侧所示的图。

从图中可以清楚看出，作为 B 和 C 的候选项，节点 2、3、4 和 5 在结构上等价。同样，如果将 A 指定为 4，那么在候选结构中，节点 5、6 和 7 将在结构上等价。我们希望确定在什么情况下这能确保它们可互换。我们引入以下定义：

定义 11. 给定候选结构 $G_C = (V_C, E_C)$、c_1、$c_2 \in V_W$，我们称 c_1 相对于 $u \in V_T$ 与 c_2 候选等价，记作 $c_1 \sim_{c,u} c_2$，当且仅当 c_1、$c_2 \notin C[u]$，或者 c_1、$c_2 \in C[u]$ 且 $(c_1, u) \sim_s (c_2, u)$。

很容易证明，如果候选集是完备的（对于每个模板节点 u，如果存在一个匹配将 u 映射到世界节点 v，那么 $v \in C[u]$），那么如果 $c_1 \sim_s c_2$，则 $c_1 \sim_{c,\, u} c_2$ 对所有模板顶点 u 都成立。

可互换性的确切判据要复杂一些。例如，在图 2 中，4 既是 A 的候选对象，也是 B 和 C 的候选对象，因此我们不能简单地将 4 与相对于 B 而与 4 候选等价的节点互换。为了处理更复杂的可互换性概念，我们引入一些术语。我们称子图同构 f 是从候选结构 G_C 导出的，当且仅当对于任意 $v \in V_T$，$(v,\ f(v)) \in V_C$（即，$f(v)$ 是 v 的候选对象）。如果对于从候选结构 G_C 导出的所有同构，世界顶点 w_1，w_2 都可以互换且保持同构，则称它们是 G_C**-可互换的**。

以下命题给出了一个判断可互换性的简单准则：

命题 12. 假设给定一个特定的候选结构 $G_C = (V_C, E_C)$，对于某个模板顶点 u，有 $c_1, c_2 \in C[u]$ 和 $c_1 \sim_{c,u} c_2$。假设 c_1 和 c_2 不是任何其他顶点的候选项。那么，c_1 和 c_2 是 G_C-可互换的。

该命题为利用候选等价提供了一种简单的方法。在我们的树搜索中，为给定顶点 u 生成候选顶点时，我们会为每个候选等价类找到代表元，且这些代表元不会作为其他顶点的候选项。如果一个等价类中有顶点出现在其他候选集中，那么我们就无法利用等价性，必须检查该类中的每个成员。此外，随着我们不断进行匹配并排除候选项，更多世界顶点将变得等价，因此，像算法 2 第 9 行所做的那样，在每次匹配前重新计算等价关系是有利的。取消匹配时，我们需要在第 20 行恢复之前的等价关系。

如果我们有 $f(v) = c_1$ 和 $f(w) = c_2$，并且想用 c_2 替换 c_1，就需要更强的条件；也就是说，它们相对于 v 和 w 都等价。在树搜索过程中，我们无法确切知道每个顶点会映射到什么，因此我们考虑一个更强的条件：

定义 13. 给定候选结构 $G_C = (V_C, E_C)$，如果对于所有 $u \in V_T$，$c_1 \sim_{c,u} c_2$，$c_1 \in V_W$ 都与 $c_2 \in V_W$ 完全候选等价，则称 $c_1 \in V_W$ 与 $c_2 \in V_W$ 是完全候选等价，记为 $c_1 \sim_c c_2$。

注意，如果对于某个 u，$c_1 \sim_{c,\, u} c_2$，并且 c_1，c_2 不是其他任何顶点的候选对象，那么 $c_1 \sim_c c_2$。这一条件

使我们能够互换世界顶点，同时仍满足子图同构条件。以下命题证明了这一点：

命题14。 假设给定一个特定的候选结构 $G_C = (V_C, E_C)$，且有 $c_1, c_2 \in V_W$ 和$c_1 \sim_c c_2$。那么， c_1 和 c_2 是 G_C-可互换的。

IV. 节点覆盖等价性

文献 [51],提出了另一种等价性概念，该概念涉及节点覆盖的使用。**节点覆盖** 是一个节点子集，移除这些节点及其关联边后，图将完全断开。文献 [51] 中的方法是，先为节点覆盖中的所有顶点构建部分匹配，然后为节点覆盖以外的所有节点分配候选对象。将覆盖以外所有节点的候选集缩减为与覆盖中节点有足够连接的候选对象后，剩下要确保它们彼此不同。

我们通过一些定义将其形式化。部分匹配是从模板图的一个子图到世界图的子图同构。我们将映射列为有序对的列表 $M = \{(v_1, w_1) \ldots, (v_n, w_n)\}$。如果对于每个 $(v_i, w_i) \in M$，若 $(v_i, v) \in V_T$，则 $(w_i, c) \in V_W$；并且如果 $(v, v_i) \in V_T$，则 $(c, w_i) \in E_T$，那么模板顶点 — 候选对象对 (v, c) 与部分匹配 M 可连接。如果在任何扩展部分匹配 M 的子图同构中，两个世界顶点 w_1, w_2 可以互换，我们就称 w_1和 w_2 是 M-可互换的。

由于该问题要简单得多，因此更容易在顶点上获得某种等价关系。

定义15. 设 M 是节点覆盖N上的部分匹配 M，并假设对于所有 $u \in V_T \setminus N$，候选集 $C[u]$ 完全由所有可与 M 相连的世界顶点组成。若对于所有 $u \in V_T \setminus N$，$w_1 \in C[u]$ 当且仅当 $w_2 \in C[u]$，则两个世界顶点 w_1、w_2 关于 M 是节点覆盖等价的，记作 $w_1 \sim_{N,M} w_2$。

例如，考虑图4中的模板和世界图。一旦节点覆盖中的节点 B 和 D 分别映射到 2 和 5，其余节点的候选项便具有世界图中对应的颜色。接着，我们只需将这些候选项分别归入等价类。请注意，图中以红色标出的边正是阻碍结构等价的因素，而节点覆盖方法实际上忽略了这些边，从而揭示了这些顶点之间的等价性。

命题16。 假设我们有模板图 N 的一个节点覆盖，并且部分匹配 M 在 N 上，而两个尚未匹配的世界顶点 w_1、w_2 满足 $w_1 \sim_{N,\ M} w_2$。那么 w_1 和 w_2 是 M-可互换的。

节点覆盖等价很容易检验，并且能够捕获其他方法所提出的等价性中相当大的一部分。这通常是因为可互换节点由兄弟叶节点组成，而它们通常不在节点覆盖中。

我们注意到，本文讨论的方法（命题 9 中针对模板和世界图的基本结构等价除外）无法同时纳入模板等价和世界图等价。允许模板

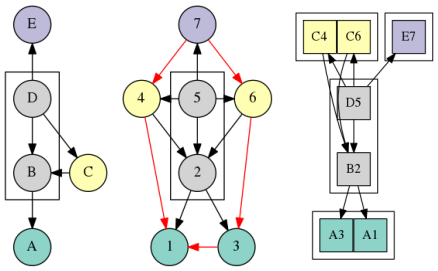


图 4. 从左到右依次为：模板、世界图和可能的候选结构。方框标出的顶点构成模板的一个节点覆盖，以及该节点覆盖在世界图中的像。世界图中颜色相同的节点是节点覆盖等价的。红色边是多余边；移除这些边后，等价性便显现出来。

和世界节点互换，并采用动态的世界图等价类，会大大增加计数过程的复杂性。一种有助于同时使用这两种等价形式的方法是进行树搜索，一次分配整个模板等价类，而不是单个模板节点。附录D节讨论了这种用于分配节点覆盖之外剩余节点的方法。

A. 等价层级

节点覆盖等价与完全候选等价之间存在如下命题所述的关系：

命题 17。 假设 N 是 V_T 的一个节点覆盖， M 是 N 上的部分匹配，候选集已缩减为可连接顶点，并且 w_1、$w_2 \in V_T \setminus N$。那么 $w_1 \sim_{N,\ M} w_2 \Leftrightarrow w_1 \sim_c w_2$。

因此，在指定节点覆盖之前，我们可以利用候选等价来避免冗余分支；一旦匹配了节点覆盖中的所有节点，就可以检查节点覆盖等价：这是一个更简单的条件。

节点覆盖等价与完全候选等价相一致，这一点在图4右侧所示的候选结构中显而易见。从该候选结构可以看出，黄色节点与绿色节点完全候选等价，因为它们拥有相同的邻居；它们也满足节点覆盖等价，因为它们分别只作为模板中对应黄色节点和绿色节点的候选项出现。

各种等价概念构成一个层次结构。世界节点的结构等价要求最严格，并蕴含其他所有形式的等价。命题 17 指出，在温和条件下，完全候选等价与节点覆盖等价是一回事。我们还可以将相对于模板顶点的候选等价纳入其中：它是完全候选等价所蕴含的一种较弱条件，但并不能保证可互换性。以下命题总结了这些发现：

命题 18。 假设命题 *17* 的条件成立。给定模板顶点 t 和世界顶点 w_1、 w_2，我们有$w_1 \sim_s w_2 \Rightarrow w_1 \sim_{N,\ M} w_2 \Leftrightarrow w_1 \sim_c w_2 \Rightarrow w_1 \sim_{c,\ t} w_2$。在前三种等价性下， w_1 和 w_2 具有可互换性。

根据该命题，我们观察到，完全候选等价和节点覆盖等价提供了最紧凑的解空间，因为它们所需的条件最弱，同时仍能保证可互换性。然而，与更简单的等价类型相比，确定完全候选等价的成本可能过高。我们将在第五节中利用真实数据和模拟数据讨论这些权衡。

V. 实验

为展示等价性对 SMP 的实用价值，我们对先进的树搜索子图同构求解器 Glasgow 进行了改造， [45]，采用算法 2 中所述的修改。我们考察七种等价模式：无等价（NE，默认）、模板结构等价（TE）、世界结构等价（WE）、模板与世界结构等价（TEWE）、命题 12 中的候选等价（CE）、命题 14 中的完全候选等价（FE）以及节点覆盖等价（NC）。每种等价模式都单独集成到格拉斯哥求解器中。

对于涉及模板等价（TE、TEWE）的每个测试类别，我们计算模板结构等价类；对于涉及世界等价（WE、TEWE、CE、FE、NC）的每个测试类别，我们在开始时使用算法 1 计算世界结构等价类。随后，我们按照算法 2 对模板等价和世界等价进行相应修改。对于需要在树搜索的每个节点重新计算等价类的算法，为提高速度，我们仅重新计算当前模板节点的候选项中的节点的等价性。我们直接根据定义检查每对候选对象之间的等价性。

我们考虑 [43] 中基准测试套件的单通道网络。数据集的基本参数列于表I中。*SF* 由100个实例组成，这些实例使用幂律随机生成，旨在构成无标度网络。*LV* 是随机生成图的多样化集合，满足各种性质（连通、双连通、三连通、二部、平面等）。*SI* 是随机生成实例的集合，分为四类：有界价、修正有界价、四维网格和 Erd˝os–R´enyi 图。*images-cv*、*meshes-cv*和*images-pr*数据 [12], [26] 集是真实实例，表示取自模式识别文献的分割图像和三维物体网格。生化 数据集 [21]包含取自真实生化反应系统的匹配问题。*phase* 数据集 [36] 由随机生成的 Erd˝os–R´enyi 图组成，其参数已知会使最先进的求解器难以处理。

我们还纳入了一组问题，其中模板图是一个较小的 Erd˝os–R´enyi 图，世界图由圣母大学网站上的网页构成，有向边表示页面之间的链接 [5]。在这些实例中，模板是随机生成的，包含 n_t 个节点和 e_t 条边，其中 $5 \le n_t \le 15$ 和 $n_t \le e_t \le 3n_t$。世界图相当稀疏，包含 325,729 个顶点和 1,497,135 条边。我们将这组问题称为 *www* 数据集。对于 n_t 的每个取值，我们收集了 50 个模板图，总计 550 个模板。

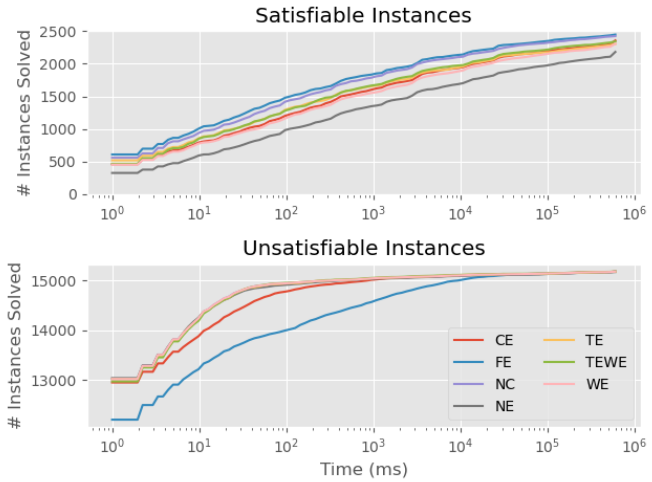


图 5. 在给定时间内解决的可满足（上）和不可满足（下）实例数量。统计结果汇总了所有单通道基准数据集。对于可满足实例，“解决”意味着已完整枚举解空间。

对于每个实例，我们针对每个等价级别运行算法，并将求解器配置为统计所有解。我们记录找到的代表解数量、通过互换可生成的解总数，以及该实例的总运行时间。如果搜索在 600 秒后仍未完成，我们便终止本次运行，并记录未完成运行的统计数据。对于每次运行，我们测量压缩率，即代表解数量除以找到的解总数。该数值表示所选等价形式将解空间规模缩小的倍数。所有实验均在配备 3 GHz 处理器、25 MB 缓存和 125 GB 内存的 Intel Xeon Gold 6136 上进行。

在表 II 中，我们记录了在 600 秒内由各算法完整枚举解空间的可满足问题所占比例。对于生化、*LV*和*si* 数据集，使用任何形式的等价关系，都能使完全解决的问题占比提高 5%–10%。我们还注意到，完全等价的表现始终最佳，其次是节点覆盖等价。这并不令人意外，因为命题 17 指出，完全等价是范围最广的等价形式。

图 5 展示了在给定时间内解决的实例数量，并表明在可满足问题的解空间枚举中，完全等价表现最佳，其次是节点覆盖等价和其他形式的等价关系。不可满足问题的图表则显示了使用完全候选等价的缺点：如果不存在解，就不需要花费额外计算时间检查等价性。在 TE、WE 和 NC 例程中，检查等价性的操作开销很低，因此与 NE 例程相比，解决的不可满足问题数量差异很小。

图 6 展示了通过比较 NE 和 FE 例程的运行时间，数据集之间及各数据集内部的差异。我们观察到，这种差异主要体现在生化、*SI*和

表I 基准数据集统计

数据集	# 实例	模板 # 节点 Min	模板 # 节点 Max	模板 # 边 Min	模板 # 边 Max	模板 密度 Min	模板 密度 Max	世界图 # 节点 Min	世界图 # 节点 Max	世界图 # 边 Min	世界图 # 边 Max	世界图 密度 Min	世界图 密度 Max
SF	100	180	900	478	5978	0.006	0.165	200	1000	592	7148	0.006	0.159
LV	6105	10	6671	10	209000	0.001	1.000	10	6671	10	209000	0.001	1.000
SI	1170	40	777	41	12410	0.005	0.209	200	1296	299	34210	0.004	0.191
images-cv	6278	15	151	20	215	0.019	0.190	1072	5972	1540	8891	4.89e-4	0.003
meshes-cv	3018	40	199	114	539	0.022	0.146	201	5873	252	15292	4.40e-4	0.022
images-pr	24	4	170	4	241	0.017	0.667	4838	4838	7067	7067	0.001	0.001
生化	9180	9	386	8	886	0.012	0.423	9	386	8	886	0.012	0.423
phase	200	30	30	128	387	0.294	0.890	150	150	4132	8740	0.370	0.782
www	3850	5	15	5	45	0.071	0.750	325729	325729	1497135	1497135	1.41e-5	1.41e-5

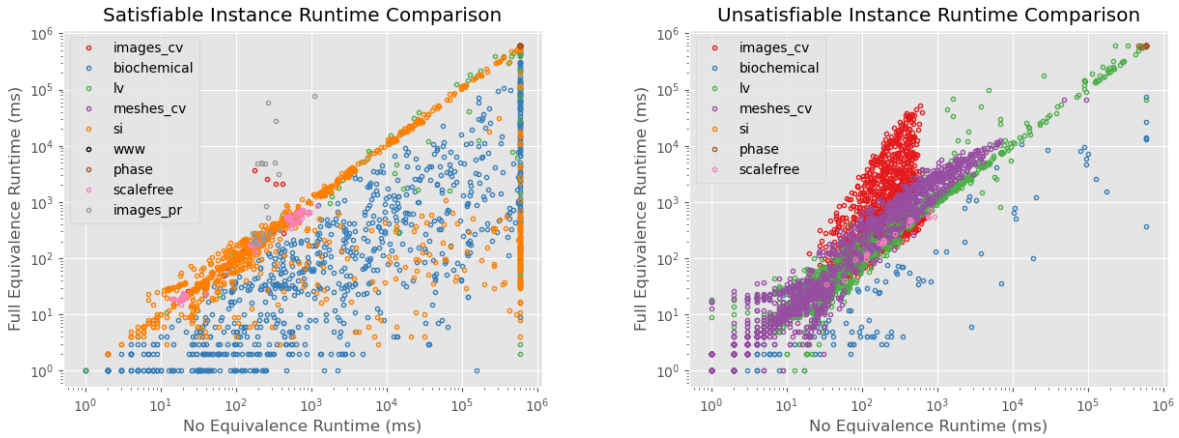


图 6. 对于可满足问题（左）和不可满足问题（右），比较无等价性与完全等价性运行中完全枚举的各项运行时间。请注意，*phase*、*www* 和 *meshes cv* 问题的任何实例都无法终止，且都会耗尽 600 秒的运行时间，因此在图中较难辨认，因为它们都占据右上角的同一位置。

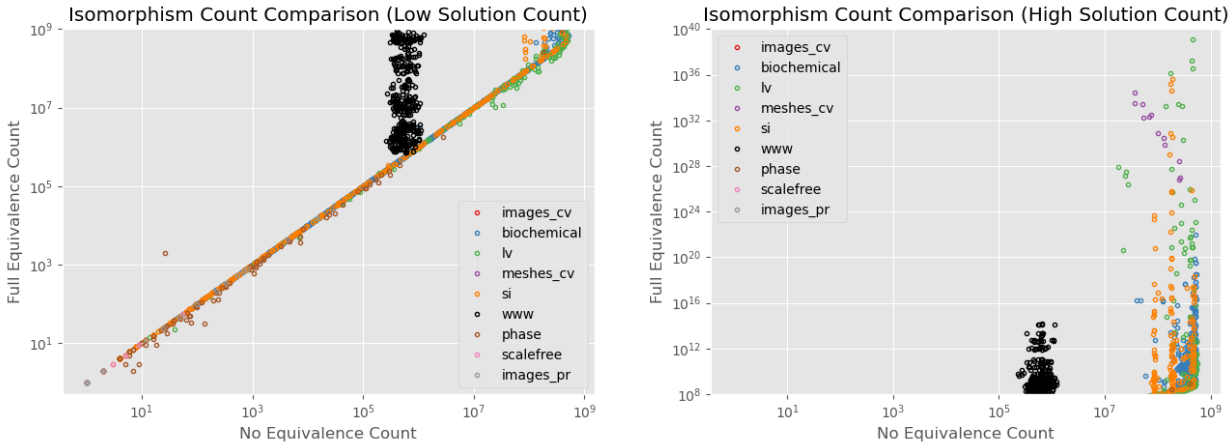


图 7. 对于同构数量较少（$< 10^9$）的问题（左）和同构数量较多（$\geq 10^9$）的问题（右），比较无等价性与完全等价性运行中的同构计数。请注意各图所选用的刻度。对于 110 个实例，采用完全等价性的求解器找到了超过 10^{40} 个同构（最多的实例有 $\approx 10^{384}$ 个同构），这些实例未显示在图中。

表 II 完全枚举的可满足实例比例

数据集	NE	TE	WE	FE	TEWE	CE	NC
生化	0.73	0.78	0.83	0.86	0.81	0.84	0.85
LV	0.10	0.10	0.14	0.15	0.13	0.14	0.15
无标度	1.00	1.00	1.00	1.00	1.00	1.00	1.00
images-cv	1.00	1.00	1.00	1.00	1.00	1.00	1.00
meshes-cv	0.00	0.00	0.00	0.00	0.00	0.00	0.00
si	0.83	0.92	0.83	0.94	0.90	0.85	0.93
images-pr	1.00	1.00	1.00	1.00	1.00	1.00	1.00
phase	0.00	0.00	0.00	0.00	0.00	0.00	0.00
www	0.00	0.00	0.00	0.00	0.00	0.00	0.00

LV 对于这些问题，完全等价例程的表现远胜于非等价例程，通常相差好几个数量级。另一方面，*images-cv*、*meshes-cv*和*images-pr* 数据集则更具挑战性，尤其是对于不可满足问题。这可能是因为没有可压缩的解空间，导致等价性检查被浪费。

图 7 包含两张图表，用于展示使用 NE 和 FE 例程时解的数量变化。两张图的坐标轴范围各不相同，以突出不同方面。左图表明，对于同构数量少于10^9 的问题，通常只需 10 分钟，就足以在不使用等价性的情况下枚举完整个解空间。对于使用原始格拉斯哥求解器的任何问题，数字 10^9 可作为在 10 分钟内找到的解数量的近似上限。我们注意到，对于*www*数据集，基础格拉斯哥求解器最多只能找到约10^6 个解，数量小得多。这是因为大量时间都花在加载庞大的世界图上。右图展示了 FE 例程找到的解比 NE 例程多出好几个数量级的问题。在这些高度对称的问题中，基于等价性的方法对于全面理解解空间至关重要。我们观察到，对于生化、*si*、*lv*和*www*数据集，使用完全等价时找到的解多出好几个数量级。FE 和 NE 找到的解数量之间最大差距未在图中显示——FE 找到了约 10^{384}个解，而 NE 找到了约 10^9 个解：相差 375 个数量级。

图 8 展示了各数据集在不同等价级别下的平均压缩率。正如预期，FE、NC 和 CE 的压缩效果最佳，其次是 TEWE；至于 TE 还是 WE 效果更好，则取决于数据集。对于生化、*lv*、*meshes cv*和*www* 数据集，使用 CE、FE 或 NC 方法时，我们观察到解空间的规模平均缩小了一个数量级或更多。对于这些数据集中某些特别对称的问题，使用 FE 或 NC 时，解空间规模甚至可以缩小数十个乃至数百个数量级。对于这些情况，必须纳入等价性，才能较为充分地理解子图同构问题的解集。

VI. 紧凑解表示

如前文所述，子图匹配的解空间通常具有组合复杂性。然而，

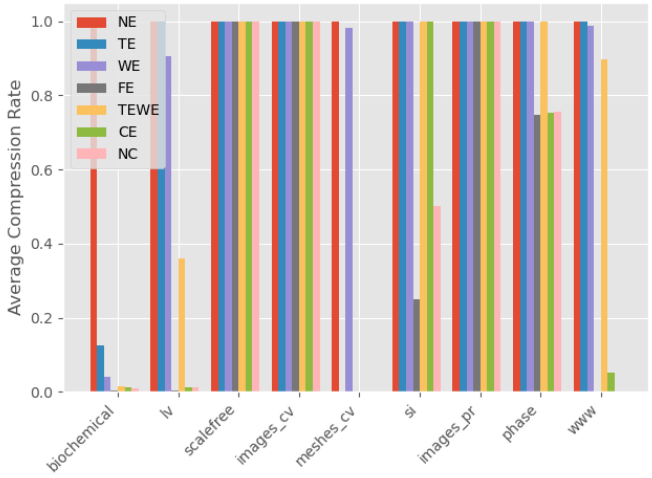


图 8. 各数据集和等价类型的平均压缩率。

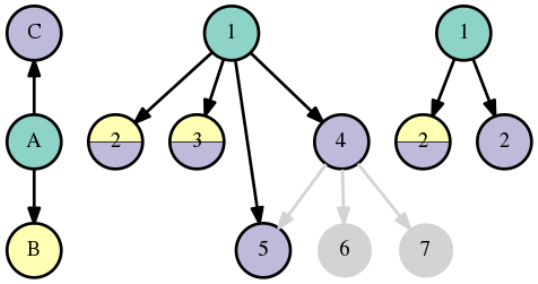


图 9. 将图 2 中的模板（左）和世界图（中）重新着色，以表示解（$B \to \{2、3\}$ $A \to 1$、$C \to \{2、3、4、5\}$）。每个世界节点的颜色与可映射到该节点的模板节点颜色相同；若没有节点可映射到该节点，则将其标为灰色。右侧的图通过移除非参与节点，并将同色节点合并为一个带有数量标注的节点，压缩了世界图。

结构等价的概念提供了以紧凑、直观的方式绘制解空间的方法。我们先以图 2 中的简单示例进行说明。我们从一个解的基本表示开始：$(A \to 1,\ B \to 2,\ C \to 3)$，它将模板顶点与世界顶点配对，并在此基础上纳入等价性。

如果采用模板结构等价，则等价的模板节点可以互换。我们通过将模板等价类与世界顶点配对来表示这一点；生成一个解相当于从每个等价类中选取一个唯一的代表元。在我们的示例中，由于 B 和 C 等价，我们写作 $(A \to 1,\ \{B,\ C\} \to 2,\ \{B,\ C\} \to 3)$。世界等价也以类似方式表示：表示 $(A \to 1,\ B \to \{2,\ 3\}\ C \to 4)$ 表示 B 可以与 2 或 3 匹配。

表 III 展示了图 2 中这个简单问题在不同等价级别下代表解数量的差异。候选对象等价和节点覆盖等价的数量是在假设首先分配 A 的情况下计算的。由于完整的候选对象等价和节点覆盖等价级别采用了最宽泛的等价概念，因此压缩效果最佳。

如果我们使用候选对象等价或节点覆盖等价，那么每次赋值之前都会重新计算等价类。因此，模板节点可能会与一个先前已被赋值、但因重新计算等价关系而规模增大的世界等价类配对。

表 III 图 2 中玩具问题各等价级别的代表解数量

等价级别	# 代表解	示例解
NE	18	$A \to 1, B \to 2, C \to 3$
TE	9	$A \to 1, \{B, C\} \to 2, \{B, C\} \to 3$
WE	10	$A \to 1, B \to \{2, 3\}\ C \to 4$
TEWE	6	$A \to 1,\ \{B,\ C\} \to 5,\ \{B,\ C\} \to \{6,\ 7\}$
CE	5	$A \to 1, B \to 2, C \to \{3, 4, 5\}$
FE	2	$A \to 1, B \to \{2, 3, 4, 5\}\ C \to \{2, 3, 4, 5\}$
NC	2	$A \to 1, B \to \{2, 3, 4, 5\}\ C \to \{2, 3, 4, 5\}$

例如，如果我们首先将模板节点 B 赋值给等价类 {2，3}，就必须将 A 赋值给 1。最后，我们重新计算等价关系，发现 {2，3，4，5} 构成一个等价类，并将最后一个模板节点 C 赋值给它。因此，我们得到解类 ($B \to$ {2， 3} $A \to 1$， $C \to$ {2， 3， 4， 5})。我们在图9中绘出了这个解类，其中每个模板节点及其对应的候选项都用相同颜色表示。所有参与代表性解的节点和边构成的子图就是**解诱导的世界子图**。最后一张图展示了**压缩的解诱导世界子图**，其中删除了所有不参与的节点和边，并将颜色相同的节点合并为“超节点”，其标签标明了合并的节点数。从这张图中，我们可以观察到参与其中的世界节点之间保留着原始模板图的图结构。

我们使用这些图形表示来描绘数据集中各种对称特征。例如，我们在图10中绘制了生化数据集中一个示例的模板子图和世界子图。从图中可以看出，等价性可能源自多个不同方面。其中一个方面是模板图中大量结构等价节点对被标为相同颜色。另一个来源是模板图中的叶节点，它们可以映射到世界图中的大型等价类。通过使用等价关系引导的子图搜索，我们可以准确揭示这些复杂性出现的位置。压缩的解诱导世界图如图11所示，清晰展示了解中每个世界节点相对于模板图所发挥的作用。

VII. 多重网络中的应用

A. 多层多重图匹配

分析人员通常希望将带属性的信息编码到图的节点和边中，并允许节点之间发生不止一种交互。例如，交通网络中的枢纽之间可能存在多种出行方式（例如火车和地铁）。[51]。形式上，如果我们有K 个不同的边标签，那么一个**多层多重图**是一个 K+1元组（V， E^1， E^2，... E^K），其中 V是顶点集合，而$E^i : V \times V \to \mathbb{Z}_{\geq 0}$ 是一个函数，用于规定两个顶点之间标签为 i 的边的数量。直观地说，多层多重图是 K 个共享同一节点集的多重图的集合。边函数的索引 i 是“通道” i，我们将由边函数E^i 给出的边称为通道 i中的边，并将图（V， E^i）称为通道 i中的图。

多路子图同构 $f : V_T \to V_W$会保留每个通道中的边数。给定模板 $(V_T, E_T^1, \ldots, E_T^K)$ 和世界图 $(V_W, E_W^1, \ldots, E_W^K)$，对于任意$u, v \in V_T$，我们要求 $E_W^i(f(u)\ f(v)) \geq E_T^i(u, v)$，也就是说， $f(u)$ 与 $f(v)$ 之间必须有足够的边，以支持 u 与 v 之间的边。等价关系的定义也可自然延伸：如果对于每个通道，针对每个 $u \neq v, w$, $E^i(v, u) = E^i(w, u)$、 $E^i(u, v) = E^i(u, w)$，且 $E^i(v, w) = E^i(w, v)$，均有 i ，我们就称 $v \sim_s w$ 。其他形式的等价关系及相关定理也都可以类似地推广。

近期，针对多路子图同构算法的开发已取得显著进展。[29]提出了一种基于多通道图中邻域结构的索引方法。[46] 扩展了单通道软件包 [14] ，使其能够处理多通道情形，并着重利用智能顶点排序来寻找同构。[38]采用约束规划方法过滤候选项，并在 [51] 中对其进行了扩展。[41]也采用了类似的筛选方法。 [44]将该问题松弛为连续优化问题，求解后再投影回原始空间。

B. 多路实验

我们评估了针对格拉斯哥求解器所作等价性增强的性能，该求解器经过调整，可处理多路子图同构问题。这些调整仅对基础算法作了最小改动，以确保只有在匹配保留每个通道中的边时才会产生匹配。为进一步减少候选项，我们还使用 [38] 中的统计过滤器和拓扑筛选器进行预过滤，并将每个通道中的子图作为格拉斯哥算法所用的补充图保留下来。

我们考虑的数据集包括来自 [51] 的那些数据集，它们既涵盖真实案例，也涵盖合成数据。真实案例包括大不列颠的交通网络 [13],、航空网络 [15],、基于 Twitter 上与希格斯玻色子相关的交互构建的社交网络[16], ，以及 COVID 数据 [52]。对于交通网络和 Twitter 网络，模板是从世界图中提取的。合成数据集包括用于表示个人之间电子邮件、电话、金融交易以及其他交互的示例，这些数据均作为 DARPA-MAA 计划的一部分生成 [33]–[35]。需要检测的子图同构可能涉及一组参与对抗活动的参与者，包括人口贩运和洗钱。表 IV 描述了这些不同子图同构问题的统计数据。有关这些特定数据集的更多详情，请参见

多层数据集比上一节中的单通道图大得多，其中最大的世界图拥有数十万个节点和数亿条边。合成数据集根据生成数据集的组织机构分为三组：PNNL [34], GORDIAN [35],和 IvySys Technologies [33]。

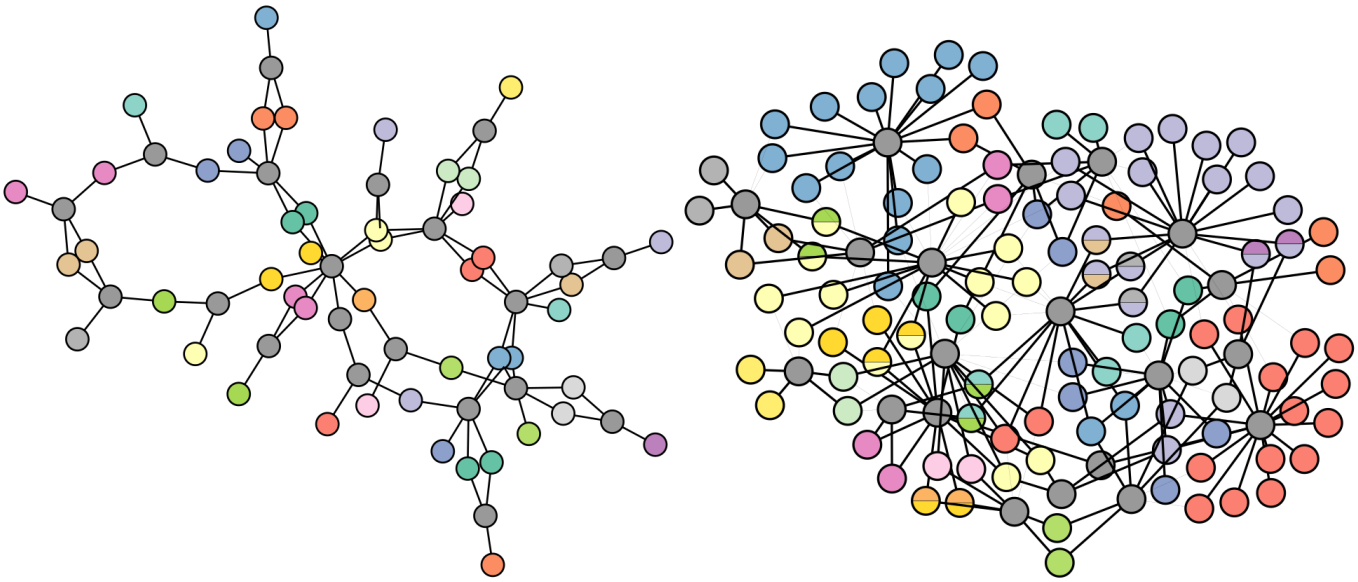


图10. 生化反应 [21] 模板图（左）以及某个由 9.18×10^{13} 个解组成的解类所诱导的世界子图（右）。深灰色节点是只有一个候选对象的节点。世界子图中颜色相同且非灰色的节点完全候选等价。具有两种或更多颜色的节点在子图搜索早期阶段属于一个类，之后该类被合并到另一个类中。压缩解表示的所有解，都可以通过将一种颜色的模板节点映射到具有相同颜色的世界节点来生成。

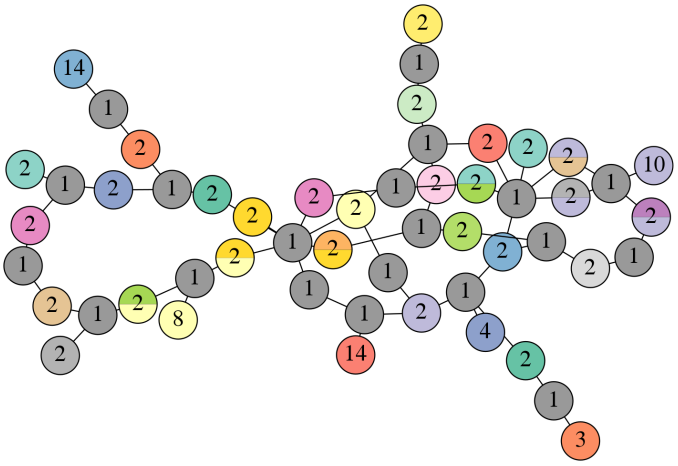


图11. 图10中的世界图，其中等价节点合并为超节点，数字表示类的大小。

表 IV 多层图数据

数据集	模板 节点	模板 边	世界图 节点	世界图 边	通道
英国交通	53	56	262377	475502	5
希格斯推特	115	2668	456626	5367315	4
航空公司	37	210	450	7177	37
PNNL RW	74	35	158	6407	3
PNNL v6-b0-s0	74	1620	22996	12318861	7
PNNL v6-b5-s0	64	1201	22994	12324975	7
PNNL v6-b1-s1	75	1335	22982	12324340	7
PNNL v6-b7-s1	81	1373	23011	12327168	7
GORDIAN v7-1	156	3045	190869	123267100	10
GORDIAN v7-2	92	715	190869	123264754	10
IvySys v7	92	195	2488	5470970	3
IvySys v11	103	387	1404	5719030	5
新冠肺炎	28	38	87580	1736985	9
Twitter - ER	5-15	4-31	456626	5367315	4

在我们的实验中，我们考察了单通道情况下使用的同样七种等价模式，但将时间限制设为一小时，以尽可能统计更多解。这些实验在同一台计算机上运行，使用的是我们改编的格拉斯哥求解器。枚举所有解所需的时间列于表 V 中，而使用给定方法找到的解数量列于表 VI 中。快速查看这些时间可以发现，在少数情况下（航空公司、GORDIAN 和希格斯推特），使用完全等价进行枚举的速度比其他任何方法都快一个数量级。表中的解数量也体现了这一加速效果：FE 找到的解明显多得多。其他方法只能找到全部解中的很小一部分。NC 方法通常在解的数量方面以及

找到了所有结果，并计算了枚举它们所需的时间。鉴于命题 18，这一结果是合理的。TE 似乎是第三好的方法，这可以归因于其实现简单，且无需重新计算等价关系。WE 和 CE 与其他方法相比缺乏竞争力。这些数据集各具特点，说明了为什么某些等价性层级比其他层级效果更好。我们将详细讨论其中几个数据集。

1) PNNL： PNNL 模板和世界图 [34]是根据真实数据生成的，用于建模特定的通信、旅行和交易模式，随后将模板嵌入世界图中。对于这些实例，应用初始筛选后，计数问题几乎就完全解决了，而解空间则由模板中的等价关系所确定。例如，观察图 12 所示的模板。

表 V 多通道问题解空间枚举时间（秒）。实验最长运行一小时后超时。Twitter-ER 数据集的结果是对一组问题取平均值，运行十分钟后超时。

算法 数据集	CE	FE	NC	NE	TE	TEWE	WE
英国交通	3600	3600	3600	3600	3600	3600	3600
希格斯推特	3600	**369**	456	3600	3600	3600	3600
航空公司	0.34	**0.24**	1985	3600	1329	3600	3600
PNNL RW	3600	3600	3600	3600	3600	3600	3600
PNNL v6-b0-s0	41.6	42.1	41.8	**41.3**	41.7	41.4	41.6
PNNL v6-b1-s1	241	**240**	241	**240**	242	**240**	**240**
PNNL v6-b5-s0	62.6	58.1	58.9	**58.0**	62.3	62.3	63.0
PNNL v6-b7-s1	1133	200	201	3600	**188**	211	1138
GORDIAN v7-1	3600	**327**	3600	3600	3600	3600	3600
GORDIAN v7-2	3600	**316**	3600	3600	3600	3600	3600
IvySys v7	3600	3600	3600	3600	3600	3600	3600
IvySys v11	3600	3600	3600	3600	3600	3600	3600
新冠肺炎	3600	3600	3600	3600	3600	3600	3600
Twitter - ER	514.7	505.4	**494.8**	536.0	536.7	544.2	541.3

表VI 一小时内找到的多通道问题的解的数量。Twitter-ER 数据集的结果是对一组问题取平均值，并在十分钟后超时。

算法 数据集	CE	FE	NC	NE	TE	TEWE	WE
英国交通	1.48e+11	**2.34e+15**	4.97e+08	1.27e+07	2.48e+12	2.02e+12	1.17e+07
希格斯推特	1.38e+14	**3.23e+14**	**3.23e+14**	5.65e+06	6.44e+06	5.72e+06	6.95e+06
航空公司	**3.67e+09**	**3.67e+09**	**3.67e+09**	3.55e+09	3.65e+09	8.11e+08	2.35e+09
PNNL RW	3.50e+09	2.78e+11	**4.72e+11**	8.59e+08	2.01e+10	5.00e+09	8.70e+08
PNNL v6-b0-s0	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**
PNNL v6-b1-s1	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**
PNNL v6-b5-s0	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**	**1.15e+03**
PNNL v6-b7-s1	**3.14e+08**	**3.14e+08**	**3.14e+08**	8.57e+07	**3.14e+08**	**3.14e+08**	**3.14e+08**
GORDIAN v7-1	1.11e+12	**9.13e+12**	1.35e+10	1.61e+07	7.84e+09	5.34e+09	1.58e+07
GORDIAN v7-2	2.14e+11	**1.35e+16**	1.15e+15	1.72e+07	3.88e+08	3.19e+08	1.65e+07
IvySys v7	2.04e+14	**8.04e+96**	2.09e+90	1.75e+09	2.67e+47	5.84e+45	5.39e+07
IvySys v11	7.41e+10	**3.64e+89**	5.09e+66	1.77e+09	4.43e+72	6.64e+71	4.88e+07
新冠肺炎	5.27e+14	**7.45e+21**	3.11e+20	9.63e+06	9.53e+06	4.51e+07	1.31e+07
Twitter - ER	3.52e+08	8.84e+09	**1.40e+11**	7.17e+05	7.41e+05	6.14e+05	6.78 e+05

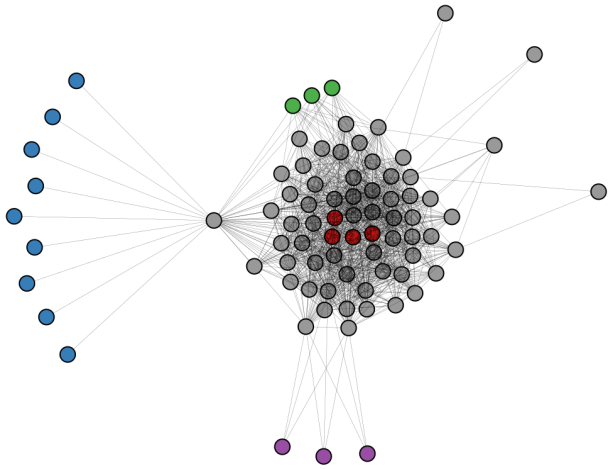


图 12. PNNL v6-b7-s1 的模板图。相同颜色的非灰色节点彼此结构等价。

生成的解的数量等于单个代表元解通过排列模板节点所生成的解的数量。我们有一组 9 个节点、一组 4 个节点，以及两组各有 3 个可互换节点，这意味着任意一个解都能生成 9!4!3!3! 个更多解。PNNL 问题的所有变体都体现了这种情况。

2) GORDIAN：GORDIAN 数据集 [35] 的模板和世界图规模远大于 PNNL 数据集；这些图以基于智能体的方式分别生成，以匹配特定人群的日常活动和出行模式。只有 FE 方法能够完整枚举解空间，但 NC 和 CE 方法也接近完整枚举。图 13 展示了一个解类的对称性；模板图中有一大组叶节点。将中心节点映射到一个候选对象后，只需将叶节点映射到该候选对象的邻居。这些图展示了节点特异性与对称性之间的权衡：边较少的模板节点具有很高的对称性，

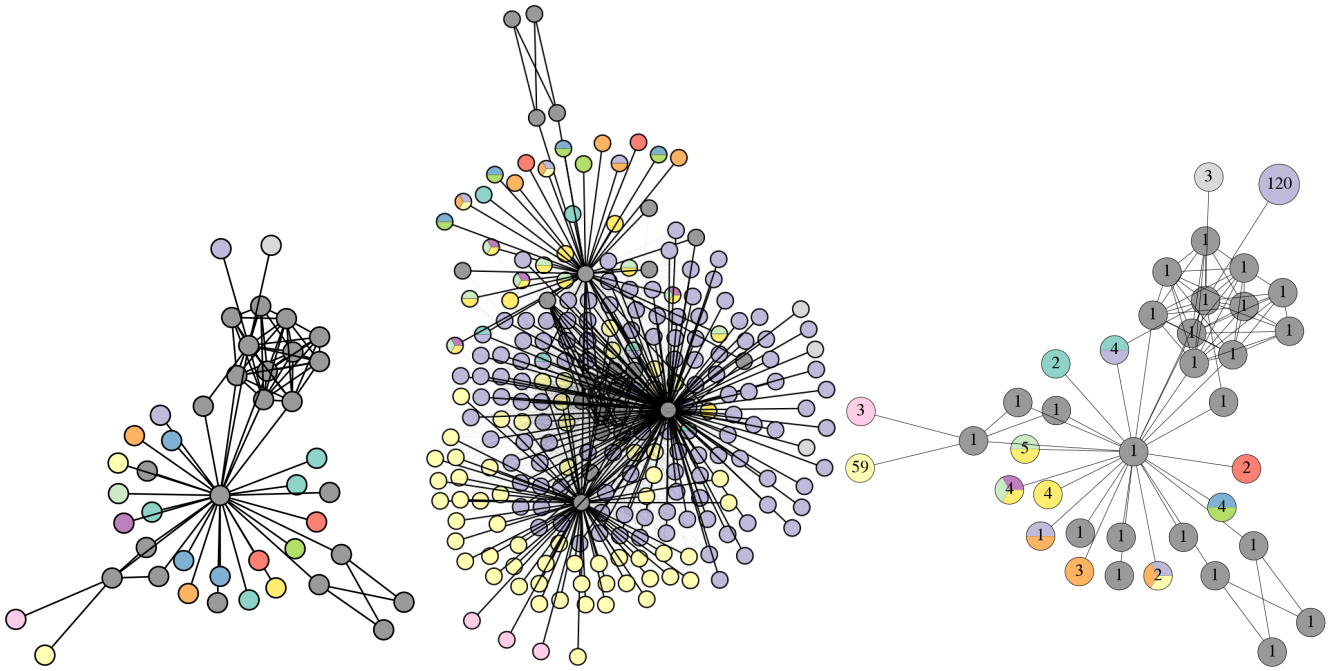


图13. 对于一个解类，其可为 GORDIAN v7-2 [35] 生成约 3×10^{12} 个解：模板（左）、解诱导的世界子图（中）以及压缩的解诱导世界子图（右）。颜色相同的世界节点完全候选等价，且是同色模板节点的候选项。该压缩解所表示的所有解，都可以通过将每个有颜色的节点映射到同色世界节点组中的某一组来生成。

而稠密子图的候选项受限，对称性也极低。图 13 右侧的图展示了由该解类导出的世界图压缩版本，可据此生成 3×10^{12} 解。

3) IvySys：Ivysys 模板和世界图 [33]分别生成，以匹配安然电子邮件数据集的度分布和电子邮件行为，并且具有最复杂的解空间。所有方法都未能成功枚举全部解。解空间的庞大与图的规模形成鲜明对比，后者仅有数千个节点。复杂性源于模板叶节点数量众多，如图14所示，其中描绘了一个可生成 7.82×10^{103} 个解的解类。图15展示了该解对应的世界子图的压缩表示，以及一个显示某些模板节点候选项的维恩图。

TE 求解器为 IvySys v7 找到了惊人的 10^{47} 个解。然而，使用 FE 方法仍会大幅增加解的数量，因为它会将这些较大的模板等价类映射到更大的世界等价类。基于等价性的子图搜索至关重要，因为 NE 方法仅找到 1.75×10^{9} 个解，比 FE 搜索少 90 个数量级。此外，典型的子图搜索会依次为每组叶节点赋值，这意味着只会探索最后一组的候选项。纳入对称性后，便能更全面地了解解空间。

4）新冠肺炎：最后，我们将算法应用于一个查询问题：查询表示多种生化实体之间已知因果关系的知识图谱。这个问题源于人们希望从研究文献中自动提取因果知识。在 [52], a 知识图谱由多个来源汇编而成，包括 COVID-19 开放研究数据集 [48]，、Blender 知识图谱 [49]，和比较毒理基因组学数据库 [50]。 [52] 的作者随后创建了一个查询，用以表示 SARS-CoV-2 可能如何引发一条导致 COVID-19 患者出现细胞因子风暴的路径，同时将该查询泛化，以检测路径中其他可能的混杂因素。

将问题重新表述为多通道子图同构问题后，模板节点和世界节点代表生化实体。其中一些模板节点已明确指定，另一些则标记为化学品、基因或蛋白质。该问题中的 9 个通道代表实体之间各种已知类型的交互，例如激活。解是为每个节点赋值，使其满足所需的化学交互关系。

如表 V 和表 VI 所示，这个问题有大量解，而纳入等价性极大地增强了我们理解解空间的能力。图 17 展示了一个解类的模板及候选项集合的维恩图，并揭示了候选项数量庞大的未指定模板节点。这些信息有助于分析人员确定路径中的混杂因素，并建议添加标签信息或交互，以便更好地限定整个解空间。

5）希格斯推特埃尔德˝–雷´尼实验：最后，我们使用小型埃尔德˝–雷´尼图作为模板，并以最大的图——希格斯推特数据集——作为世界图，进行与单通道图实验类似的实验。我们通过叠加对应于各个通道的 4 个不同图来生成一个多通道模板图，每个图均按埃尔德˝–雷´尼图生成，其中 $p = \log_{8n_t}^{n_t}$，n_t 表示模板节点的数量。选择该值 p，使得

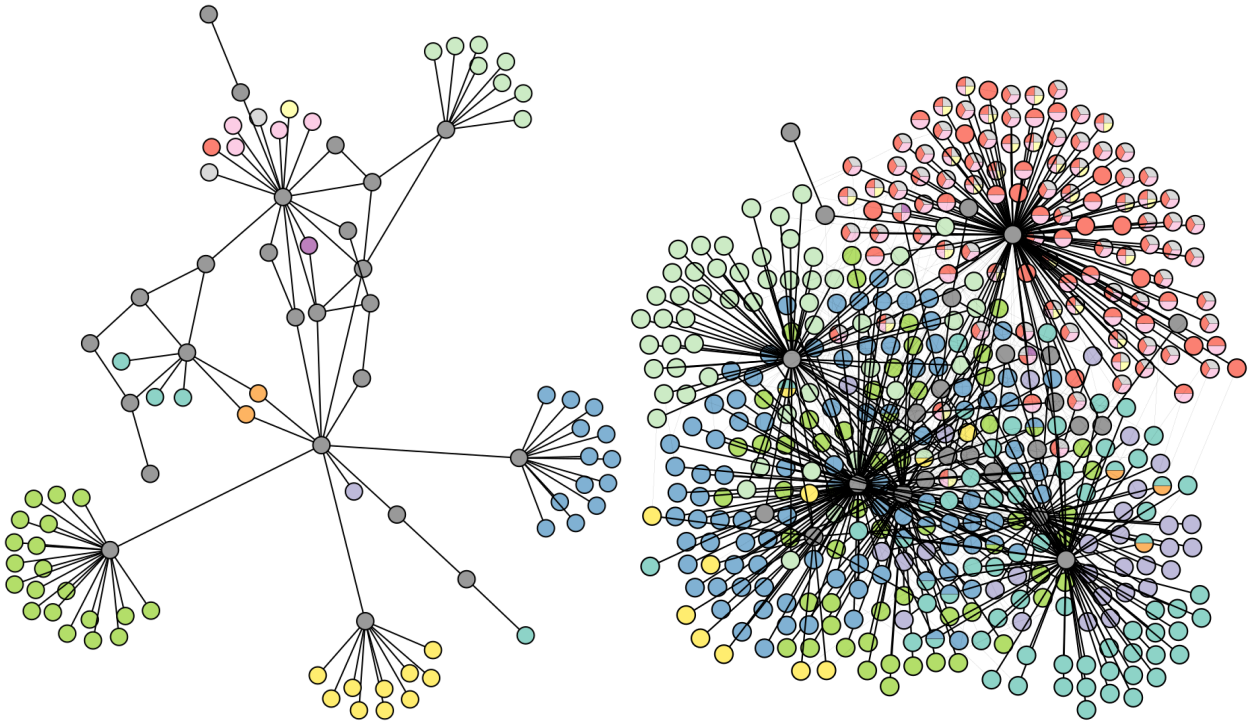


图14. 模板（左）和某个解类对应的解诱导世界子图（右），由此可生成 IvySys v7 [33] 的 7.82×10^{103} 个解。相同颜色的世界节点完全候选等价，且是相同颜色模板节点的候选项。通过将每个彩色节点映射到一组相同颜色的世界节点，即可生成此压缩解所表示的所有解。

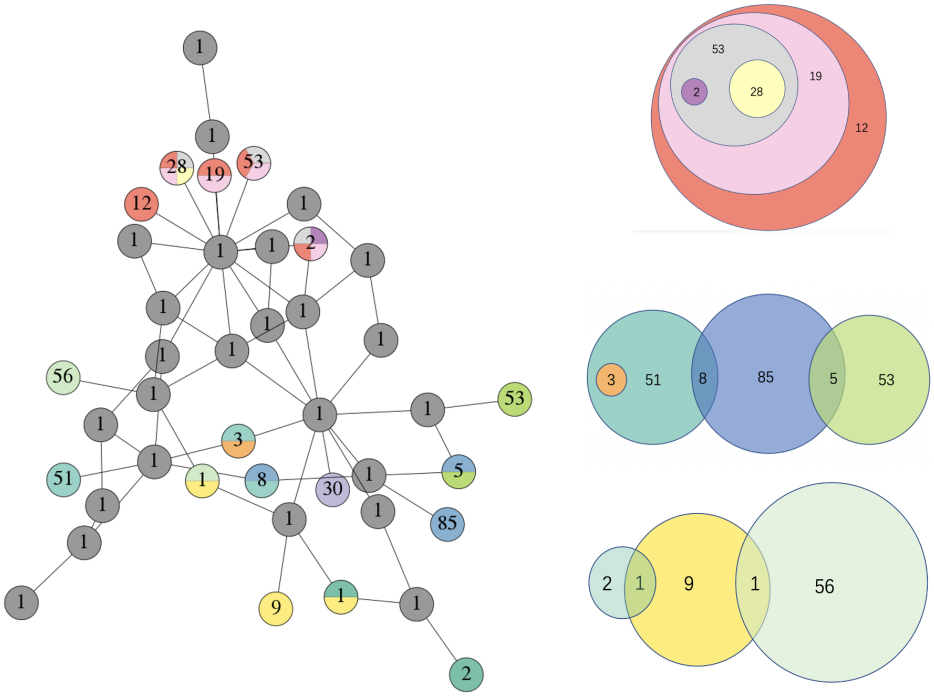


图15. IvySys v7 [33] 的压缩解诱导世界图（左）以及世界图中相交候选集的维恩图表示（右），对应某个可生成 7.82×10^{103} 个解的解类。维恩图各区域中的数字表示世界图中节点覆盖等价类的大小。通过将模板中的每个彩色节点映射到维恩图中颜色相同的集合，即可生成此压缩解所表示的所有解。

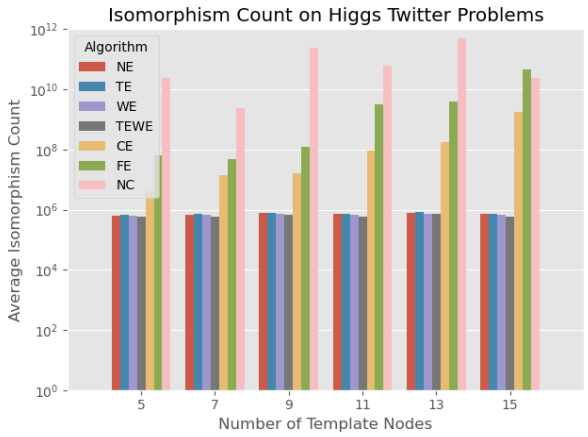


图16. 对于每个等价层级找到的子图同构的平均数量，其中模板为小型 Erd˝os–R´enyi图，世界图为*Higgs Twitter* 图。

该图以较高概率为连通图。我们以这种方式为每个 $n_t = 5$、7、9、11、13、15生成45个连通图。随后，我们计算每种方法在10分钟内统计到的同构数量。对于这些问题，我们在运行算法之前，预先计算 Higgs Twitter 图的世界结构等价类。图16展示了各等价方法在不同模板尺寸下的平均同构数量。总运行时间和同构数量的总体平均值列于表 V 和表 VI 的 *Twitter-ER* 数据集中。

从这些结果中可以看出，NC、FE 和 CE 方法找到的解显著多于基础例程，而其他等价方法并未优于 NE 方法。就找到的同构数量和总耗时而言，NC 的表现都是最佳的；我们推测这是因为它计算开销较低，并且能够捕获大部分等价性。FE 和 CE 的表现差了几个数量级，而其余方法 TE、WE 和 TEWE 相比基础方法没有显著优势；事实上，涉及世界结构等价时，表现反而更差。我们认为，这些方法提升不大的原因在于，多通道 Erd˝os–R´enyi 图中的节点不太可能结构等价。总的来说，这些实验表明，即使使用随机生成的模板图，将等价性纳入算法也能带来显著改进。不过，某些等价模式可能更适用于特定类别的图，因此必须谨慎选择，确保所选的等价层级确实有助于解决问题。

VIII. 结论

在本研究中，我们建立了静态和动态等价性概念的理论，并提出了在保持同构的情况下可以互换节点赋值的条件。只需对纳入等价性进行树搜索的子图同构程序进行最小修改，我们就能大幅减少求解问题所需的时间，并获得解空间的紧凑刻画。对于对称性最小的实例，能够获得的收益很少；但对于具有大型对称结构的问题，利用等价性对于理解庞大的解空间至关重要。具体而言，我们证明了 FE 方法和 NC 方法在捕捉问题中存在的等价性方面均表现良好，从而实现了对解空间的最大压缩。我们将这些方法集成到当前最先进的求解器 Glasgow 中，证明了我们的结果适用于标准子图求解器；此外，我们还将这些方法扩展到了更复杂的多层多重图问题空间。

未来的研究方向包括将这些等价性概念应用于非精确搜索，并提出非精确形式的等价性。我们还希望进一步了解如何将自同构等价与本文讨论的不同等价性概念结合起来。

致谢

本材料基于美国空军研究实验室和 DARPA 根据协议编号 FA8750-18-2-0066 资助的研究。尽管重印件上标有版权声明，美国政府仍获准为政府用途复制和分发重印件。本文所述观点和结论仅代表作者意见，不应被解释为必然代表美国空军研究实验室、DARPA 或美国政府明示或暗示的官方政策或认可。

这项工 作也得到了 NSF DMS-202 号资助的支持 7277.

参考文献

[1] J. Ullmann，“子图同构算法”，美国计算机协会会刊（*JACM*），第23卷，第1期，第31–42页，1976年。[2] A. Sanfeliu和K.-S. Fu，“用于模式识别的属性关系图之间的距离度量”，*IEEE*系统、人类与控制论汇刊，第3期，第353–362页，1983年。[3] A. C. Dumay、R. J. van der Geest、J. J. Gerbrands、E. Jansen和J. H. Reiber，“应用于动脉造影图冠状动脉节段标注的一致性不精确图匹配”，载于第*11*届*IAPR*国际模式识别会议。第三卷。会议*C*：像、语音与信号分析，IEEE 计算机学会，第1卷，1992年，第439–442页。[4] L. Wiskott、N. Kr¨uger、N. Kuiger和C. von der Malsburg，“通过弹性束图匹配进行人脸识别”，*IEEE*模式分析与机器智能汇刊，第19卷，第7期，第775–779页，1997年。[5] R. Albert、H. Jeong和A.-L. Barab´asi，“万维网的直径”，*Nature*，第401卷，第6749期，第130–131页，1999年。[6] M. R. Garey和D. S. Johnson，计算机与难解性。纽约：W.H. Freeman，2002年，第29卷。[7] L.P. Cordella、P. Foggia、C. Sansone和M. Vento，“用于匹配大型图的一个（子）图同构算法”，*IEEE*模式分析与机器智能汇刊，第26卷，第10期，第1367–1372页，2004年。

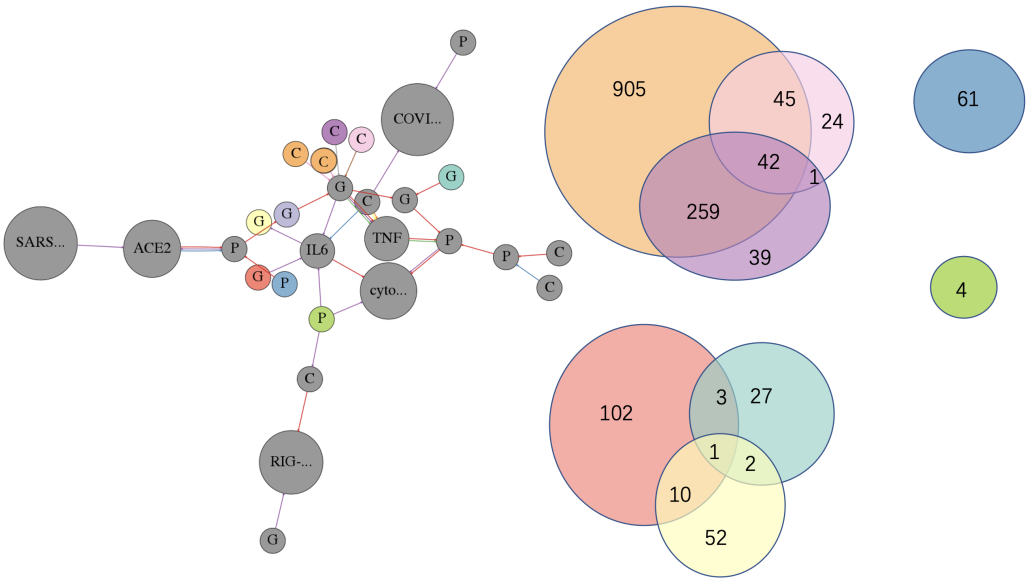


图17. 新冠肺炎-19 [52] 模板（左）以及世界图中候选集的维恩图（右），由此可生成属于同一解类的 2.6×10^{18} 解。维恩图中的每个区域代表一个节点覆盖等价类，区域中的数字表示该类的大小。部分模板节点在开始时已指定，其他节点则仅被标记为 C、P 或 G，分别表示化学品、蛋白质和基因。可通过将同一颜色的非灰色模板节点映射到维恩图中相同颜色区域内的世界节点来生成这些解。

[8] S. Auer、C. Bizer、G. Kobilarov、J. Lehmann、R. Cy- ganiak和Z. Ives，“Dbpedia：开放数据网络的核心”，载于语义网络，施普林格，2007年，第722–735页。[9] D. Conte、P. Foggia、C. Sansone和M. Vento，“模式识别和计算机视觉应用为何以及如何使用图”，载于计算机视觉与模式识别中的应用图论，A. Kandel、H. Bunke和M. Last编，施普林格柏林海德堡，2007年，第85–135页。[10] C. Solnon，“基于AllDifferent的过滤用于子图同构”，人工智能，第174卷，第12–13期，第850–864页，2010年。[11] S. Zampelli、Y. Deville和C. Solnon，“使用约束编程求解子图同构问题”，约束，第15卷，第3期，第327–353页，2010年。[12] G. Damiand、C. Solnon、C. De la Higuera、J.-C. Janodet和 ´E. Samuel，“nd开放组合图的子同构多项式算法”，计算机视觉与图像理解，第115卷，第7期，第996–1010页，2011年。[13] W. Fan，“为社交网络分析重新审视图模式匹配”，载于第*15*届国际数据库理论会议论文集，美国计算机协会，2012年，第8–21页。[14] V. Bonnici、R. Giugno、A. Pulvirenti、D. Shasha和A. Ferro，“一种子图同构算法及其在生化数据中的应用”，*BMC*生物信息学，第14卷，第7期，第1–13页，2013年。[15] A. Cardillo、J. G´omez-Gardenes、M. Zanin、M. Ro- mance、D. Papo、F. Del Pozo和S. Boccaletti，“网络特征从多重性中的涌现”，《科学报告》，第3卷，第1期，第1–6页，2013年。

[16] M. De Domenico、A. Lima、P. Mougel 和 M. Musolesi，《科学谣言的剖析》，《科学报告》，第 3 卷，第 1 期，第 1–9 页，2013年。[17] S. Demeyer、T. Michoel、J. Fostier、P. Audenaert、M. Pickavet 和 P. Demeester，《基于索引的子图匹配算法（isma）：利用优化搜索树在大型网络中快速枚举子图》，《公共科学图书馆·综合》，第 8 卷，第 4 期，e61183，2013年。[18] W.-S. Han、J. Lee 和 J.-h. Lee，《Turboiso：实现大型图数据库中超快速、稳健的子图同构搜索》，载于*SIGMOD*会议，2013年。[19] G. Audemard、C. Lecoutre、M. Samy-Modeliar、G. Goncalves 和 D. Porumbel，《针对子图同构问题的基于评分的邻居支配》，载于约束编程原理与实践国际会议，施普林格，2014年，第 125–141 页。[20] P. Foggia、G. Percannella 和 M. Vento，《过去十年模式识别中的图匹配与学习》，《模式识别与人工智能国际期刊》，第 28 卷，第 01 期，第 1 450001 页，2014年。[21] S. Gay、F. Fages、T. Martinez、S. Soliman 和 C. Sol- non，《关于子图满同态问题》，《离散应用数学》，第 162 卷，第 214–228 页，2014年。[22] M. Houbraken、S. Demeyer、T. Michoel、P. Audenaert、D. Colle 和 M. Pickavet，《具有一般对称性的基于索引的子图匹配算法（ismags）：利用对称性加速子图枚举》，《公共科学图书馆·综合》，第 9 卷，第 5 期，e97896，2014年。[23] P. Ribeiro 和 F. Silva，《发现彩色网络模体》，载于复杂网络*V*，施普林格，2014年，第 107–118 页。

[24] C. McCreesh 和 P. Prosser，《使用补充图的并行回溯跳转子图同构算法》，载于约束编程原理与实践国际会议，施普林格，2015年，第 295–312 页。[25] X. Ren 和 J. Wang，《利用顶点关系加速大型图中的子图同构》，*VLDB*基金会会刊，第 8 卷，第 5 期，第 617–628 页，2015年。[26] C. Solnon、G. Damiand、C. De La Higuera 和 J.-C. Janodet，《子地图同构与最大公共子地图问题的复杂性》，模式识别，第 48 卷，第 2 期，第 302–316 页，2015年。[27] F. Bi、L. Chang、X. Lin、L. Qin 和 W. Zhang，《通过延后笛卡尔积实现高效子图匹配》，载于*2016*年数据管理国际会议论文集，美国计算机协会，2016年，第 1199–1214 页。[28] F. Emmert-Streib、M. Dehmer 和 Y. Shi，《图匹配、网络对齐与网络比较的五十年》，《信息科学》，第 346 卷，第 180–197 页，2016年。[29] V. Ingalalli、D. Ienco 和 P. Poncelet，《Sumgra：通过高效索引查询多重图》，载于数据库与专家系统应用国际会议，施普林格，2016年，第 387–401 页。[30] L. Kotthoff、C. McCreesh 和 C. Solnon，《子图同构算法组合》，载于学习与智能优化，P. Festa、M. Sellmann 和 J. Vanschoren 编，Cham：施普林格国际出版，2016年，第 107–122 页。[31] V. Carletti、P. Foggia、A. Saggese 和 M. Vento，《推出 VF3：一种新的子图同构算法》，载于模式识别中的基于图表示国际研讨会，施普林格，2017年，第 128–139 页。[32] Y. Liang 和 P. Zhao，《图数据库中的相似性搜索：一种多层索引方法》，载于*2017*年*IEEE*第*33*届数据工程国际会议（*ICDE*），电气电子工程师学会，2017年，第 783–794 页。[33] K. O. Babalola、O. B. Jennings、E. Urdiales 和 J. A. DeBardelaben，《生成合成电子邮件数据集的统计方法》，载于*2018*年*IEEE*国际大数据会议（*Big Data*），2018年10月，第 3986–3990 页。[34] J. A. Cottam、S. Purohit、P. Mackey 和 G. Chin，《包含对抗活动的多通道大型网络仿真》，载于*2018*年*IEEE*国际大数据会议（*Big Data*），2018年10月，第 3947–3950 页。[35] K. Karra、S. Swarup 和 J. Graham，《对合成社会接触网络复杂性与真实性的实证评估》，载于*2018*年*IEEE*国际大数据会议（*Big Data*），2018年10月，第 3959–3967 页。[36] C. McCreesh、P. Prosser、C. Solnon 和 J. Trimble，《子图同构何时会变得极其困难，以及原因》

这对于图数据库而言十分重要，”《人工智能研究杂志》，第61卷，第723–759页，2018年。[37] G. Micale、R. Giugno、A. Ferro、M. Mongiovi、D. Shasha和A. Pulvirenti，“寻找显著标记图模式的快速分析方法”，《数据挖掘与知识发现》，第32卷，第2期，第504–531页，2018年。[38] J. D. Moorman、Q. Chen、T. K. Tu、Z. M. Boyd和A. L. Bertozzi，“多重网络上子图匹配的过滤方法”，载于*2018*年*IEEE*大数据国际会议（*Big Data*），IEEE，2018年，第3980–3985页。[39] H. Jin、X. He、Y. Wang、H. Li和A. L. Bertozzi，“多重网络上的噪声子图同构”，载于*2019*年*IEEE*大数据国际会议（*Big Data*），IEEE，2019年，第4899–4905页。[40] A. Kopylov和J. Xu，“噪声多重网络上非精确子图匹配的过滤策略”，载于*2019*年*IEEE*大数据国际会议（*Big Data*），2019年，第4906–4912页。[41] L. Liu、B. Du、H. Tong、等，“G-finder：近似属性子图匹配”，载于*2019*年*IEEE*大数据国际会议（*Big Data*），IEEE，2019年，第513–522页。[42] T. Nguyen、D. Yang、Y. Ge、H. Li和A. L. Bertozzi，“结构等价性在多通道多重图子图同构中的应用”，载于*2019*年*IEEE*大数据国际会议（*Big Data*），IEEE，2019年，第4913–4920页。[43] C. Solnon，“子图同构求解器的实验评估”，载于模式识别中的基于图表示国际研讨会，施普林格，2019年，第1–13页。[44] D. Sussman、Y. Park、C. E. Priebe和V. Lyzinski，“噪声诱导子图检测的匹配滤波器”，《*IEEE*模式分析与机器智能汇刊》，第1–1页，2019年，ISSN：0162-8828。DOI：10.1109/TPAMI.2019.2914651。[45] C. McCreesh、P. Prosser和J. Trimble，“格拉斯哥子图求解器：使用约束规划处理困难子图同构问题变体”，载于图变换国际会议，施普林格，2020年，第316–324页。[46] G. Micale、V. Bonnici、A. Ferro、D. Shasha、R. Giugno和A. Pulvirenti，“Multiri：标记多重图中的快速子图匹配”，*arXiv*预印本 *arXiv:2003.11546*，2020年。[47] T. K. Tu、J. D. Moorman、D. Yang、Q. Chen和A. L. Bertozzi，“非精确属性子图匹配”，*IEEE*大数据大会、对抗性活动分析的图技术（*GTA3 4.0*）研讨会论文集，第2575–2582页，2020年。[48] L. L. Wang、K. Lo、Y. Chandrasekhar、R. Reas、J. Yang、D. Eide、K. Funk、R. Kinney、Z. Liu、W. Merrill、等，“CORD-19：COVID-19 开放研究数据集”，*ArXiv*，2020年。[49] Q. Wang、M. Li、X. Wang、N. Parulian、G. Han、J. Ma、J. Tu、Y. Lin、H. Zhang、W. Liu、等，“COVID-19文献

知识图谱构建和药物再利用报告生成”，*arXiv*预印本*arXiv:2007.00576*，2020年。

[50] A. P. Davis、T. C. Wiegers、J. Wiegers、C. J. Grondin、R. J. Johnson、D. Sciaky 和 C. J. Mattingly，《CTD 解剖学：从解剖学视角分析化学物质诱导的表型与暴露及其对环境健康研究的启示》，《当代毒理学研究》，第 2 卷，第 128–139 页，2021 年。

[51] J. D. Moorman、T. Tu、Q. Chen、X. He 和 A. Bertozzi，《多重网络上的子图匹配》，《电气电子工程师学会网络科学与工程汇刊》，第 8 卷，第 2 期，第 1367–1384 页，2021 年。

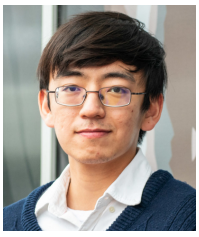
[52] J. Zucker、K. Paneri、S. Mohammad-Taheri、S. Bhar-gava、P. Kolambkar、C. Bakker、J. Teuton、C. Hoyt、K. Oxford、R. Ness 和 O. Vitek，《利用结构化生物学知识进行反事实推断：病毒发病机制案例研究》，《电气电子工程师学会大数据汇刊》，第 7 卷，第 01 期，第 25–37 页，2021 年 1 月。



Dominic Yang 是加利福尼亚大学洛杉矶分校数学系的博士候选人。他于2018年获得加利福尼亚大学戴维斯分校应用数学和计算机科学学士学位。他的研究兴趣包括图算法、组合优化、混合整数规划和深度学习。



Yurun Ge 是加利福尼亚大学洛杉矶分校数学系的博士候选人。他于2018年获得上海交通大学数学学士学位。他的研究兴趣包括图算法和网络分析。



Thien Nguyen 是东北大学计算机科学系的博士候选人。他于2020年获得加利福尼亚大学欧文分校计算机科学学士学位。他的研究兴趣包括鲁棒学习的优化与算法。



Denali Molitor 于2021年获得加利福尼亚大学洛杉矶分校数学博士学位。她于2014年获得科罗拉多学院数学学士学位。她的研究兴趣包括数值线性代数和随机优化。



Jacob D. Moorman 于2021年在加利福尼亚大学洛杉矶分校完成数学博士学位。他于2016年获得新泽西理工学院数学和计算机科学学士学位。他的研究兴趣包括网络分析、线性代数、随机优化算法、模式识别和高维数据分析。



Andrea L. Bertozzi， 电气电子工程师学会会士 在普林斯顿大学完成了所有数学学位。她是一位应用数学家，专长于非线性偏微分方程以及用于机器学习的图模型。1991年至1995年，她在芝加哥大学担任L. E. Dickson讲师和美国国家科学基金会博士后研究员。1995年至1996年，她是阿贡国家实验室的Maria Geoppert-Mayer杰出学者。1995年至2004年，她在杜克大学任教，最初担任数学副教授随后担任数学与物理学教授。自2003年以来，她一直在加利福尼亚大学洛杉矶分校任教，现任数学与机械及航空航天工程杰出教授，并担任Betsy Wood Knapp创新与创造力讲席。她于2010年当选美国艺术与科学院院士，并于2010年成为工业与应用数学学会（SIAM）会士。她于2013年成为美国数学学会会士，并于2016年成为美国物理学会会士。2018年，她当选美国国家科学院院士。她获得的荣誉包括1995年的斯隆研究奖、1996年的总统科学家与工程师早期职业奖、2009年的工业与应用数学学会科瓦列夫斯卡娅奖，以及2019年的工业与应用数学学会克莱因曼奖。2014年，她与A. Flenner共同获得工业与应用数学学会杰出论文奖，以表彰她们在机器学习基于几何图的算法方面的工作。她于2017年获得西蒙斯数学+ X研究者奖。