arXiv:2611.32738v1 [quant-ph] 24 Nov 2026

用于数据分析的量子社区检测

Vicente P. Soloviev[1] Brian Coyle[2,3] Stefano Scali[2,4] 和 Michal Krompiec[2]

[1] *Fujitsu Research of Europe Ltd., Pozuelo de Alarcón, 28224 Madrid, Spain* [*2] *Fujitsu Research of Europe Ltd., Slough SL1 2BE, UK* [3] *School of Informatics, University of Edinburgh, Edinburgh, UK* [4] *University of Exeter, Department of Physics and Astronomy, Stocker Road, Exeter EX4 4QL, UK*

社区检测为特征选择和聚类等数据分析任务提供了传统方法之外的另一种选择，而利用量子算法检测社区是实现量子加速的一个有前景的方向。我们研究了量子谱社区检测器（例如模块度矩阵算法 deteQt）在这两类任务处理流程中的应用。首先，我们通过递归二分将 deteQt 推广到多个社区，并证明它能够在合成数据集和小型数据集上恢复植入社区，从而识别冗余特征。随后，我们提出 projeQt：该方法通过量子奇异值变换对编码后的模块度矩阵应用谱滤波器，保留 deteQt 未使用的高阶特征向量。projeQt 作用于广义模块度矩阵，并通过对由量子态重叠构建的格拉姆矩阵进行 k-均值聚类，一次提取多个社区，无需递归。在 MNIST 上，我们展示了这两种处理流程：一方面选择信息性像素，另一方面直接通过聚类对图像进行分类；结果表明，它们的性能优于标准社区检测器。

一、引言

社区检测算法的目标是将图划分为节点组，使组内节点之间的连接比它们与网络其余部分的连接更加稠密 [1–3]。许多算法通过最大化模块度来实现这一目标 [4]，模块度将所提出组内的边密度与具有相同度的随机图中预期的边密度进行比较。精确地最大化模块度是 NP-困难的 [5]，因此实际方法都是近似方法。分裂法逐步移除介数最高的边，这些边位于社区之间 [6]，并在所生成的划分中保留模块度最高的划分 [4]。Louvain 等贪婪方法 [7]，及其改进方法 Leiden [8]，将每个节点移入使模块度增幅最大的相邻社区，然后将每个社区合并为单个节点，并在更小的图上重复这一过程，直到模块度不再增加。社区检测也可以表述为对随机块模型进行统计推断 [9–11]，而最大化模块度等价于对该模型的一个特殊情形进行最大似然推断 [12]。在本文中，我们将主要关注谱模块度方法，该方法利用模块度矩阵的主特征向量对图进行划分 [13, 14]。然而，上述所有经典社区检测方法都存在局限性。Louvain 作为一种贪婪方法，其得到的划分取决于节点的访问顺序，而不是取决于所达到的模块度值 [7]，尽管它在稀疏图上的运行时间接近线性 [7, 15]。基于模块度的方法还存在分辨率极限 [16]。在一个包含 m 条边的图中，设有两个社区，它们仅通过一条边相连，但每个社区的 $\sqrt{m}/2$ 内部边少于约

在这种情况下，将两个社区合并比分开具有更高的模块度，即使每个社区本身都是一个团。结果是，任何最大化模块度的方法都可能将两个彼此分离且不同的社区合并。另一方面，对于平均度较大的图，谱方法可以检测低至可检测性阈值的社区，该阈值是任何方法所能达到的极限 [17, 18]，但其代价高昂，因为通常需要在每次社区分裂时求出主特征向量。对于具有 N 个节点的图，使用幂方法时，在稠密图的第一次分裂中，求解特征值的主导阶复杂度为 $O(N^3)$；对于稀疏图上的整个递归过程，其复杂度为 $O(N^2 \log N)$，其中稀疏图即具有 $m = O(N)$ 条边的图，因此平均度 $2m/N$ 是有界的 [13, 14]。重复二分也可能无法找到划分为两个以上组时的最佳划分 [14, 19]。鉴于上述局限性，已有多项研究开发了用于处理社区检测的量子方法，我们将在第 IIB节中对其进行综述。

在这项工作中，我们聚焦于 deteQt 量子社区检测算法 [20]。与其他量子方法不同，deteQt 是一种谱模块度方法，需要量子奇异值变换（QSVT）等容错原语 [21]。该算法将模块度矩阵的主特征向量 B 制备为一个量子态，存储在 $\lceil \log_2 N \rceil$ 个量子比特中，而该特征向量振幅的符号用于标记各个社区。deteQt 通过振幅展平读取这些符号，并使用 QSVT 近似符号函数，随后比较所得结果

将量子态与一组探测态进行比较。原始方法采用了确定性消除，这是一种逐一精确移除候选社区划分的例程，直至只剩下一个划分。提出的应用场景是僵尸网络检测，其中一个小型社区的 k 节点隐藏在一个大得多的图中（$k \ll N/2$）。然而，按已发表的版本，它只返回一个二划分，因此最多只能将一个图划分为两个社区。我们在第IIB节介绍该算法。不过，本研究的目标并非直接改进量子社区检测

* vicente.perezsoloviev@富士通.com

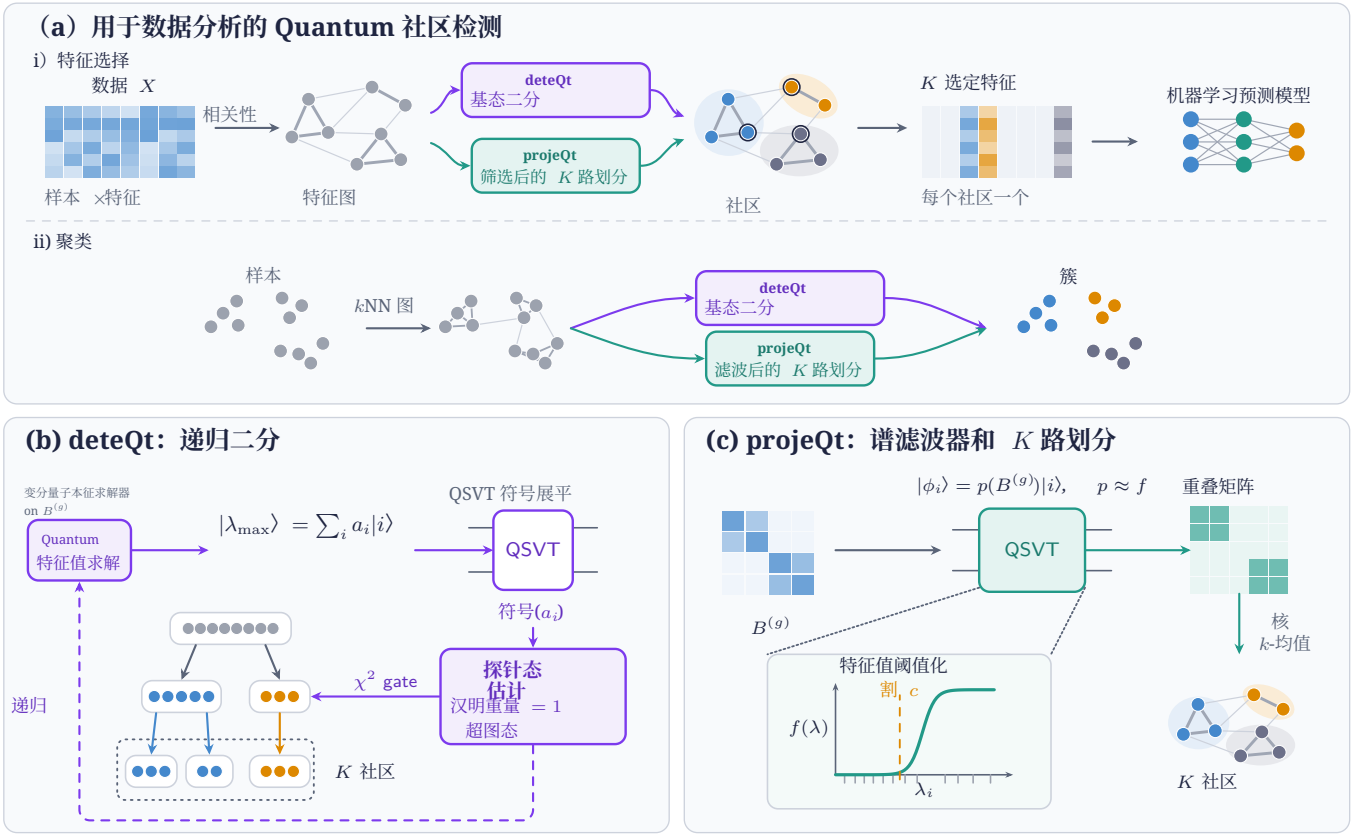

图 1：**面向数据分析的社区检测。(a)**在特征选择中，相关特征构成图，图社区检测会从每个社区中选出一个代表性特征，例如供机器学习预测器使用。在聚类中，样本构成一个 k近邻图，检测出的社区本身就是簇。社区检测通过 deteQt 和 projeQt 量子算法实现。(b)变分量子本征求解器制备当前组的模块度矩阵的主特征向量。振幅块编码和量子奇异值变换对其振幅应用符号函数，以提出一种二分方案。接受的组会递归拆分，直至达到 K 个社区。(c) projeQt 对该组的广义模块度矩阵应用 QSVT 谱滤波器，估计滤波后的节点态之间的重叠度，并用核 k-均值对其进行聚类，形成 K 个社区。

算法，而是研究它们在其他下游数据分析任务中的实用性，具体而言是特征选择和聚类。特征选择会在学习 [22, 23]之前移除冗余或无信息量的变量。为将社区检测应用于此，需将特征作为图的节点，并以特征两两之间的相关性为边赋权，使每个社区都由一组相关特征构成。从每个社区中保留一个特征，即可移除冗余特征，同时确保每组特征都有代表 [24, 25]。Louvain 等经典检测器已被用于此类任务 [26–28]。在聚类中，节点是数据点，它们通过近邻图中的边相连，而社区就是簇 [29]。如今，将 Louvain 算法应用于此类图，已成为根据测得的基因或蛋白质表达，将单个细胞归为不同细胞类型的标准技术 [30, 31]。谱聚类 [32] 则使用图拉普拉斯矩阵的特征向量来划分数据。谱模块度方法与之密切相关：当特征向量受到约束时

当权重取节点度时，用于最大化模块度的谱算法与归一化割划分算法相同，而后者可由谱聚类近似求解 [33]。已有研究提出了直接用于聚类的量子算法 [34, 35] ，以及用于以二次无约束二元优化形式表示的特征选择的量子算法 [36–38]。然而，据我们所知，迄今尚无量子社区检测算法将其与这些应用联系起来。

我们的贡献在于展示量子社区检测器在特征选择和聚类（第IV节）中的有效性。我们首先在具有预置社 [9] 区的随机块模型图上对其进行基准测试，信号强度涵盖可检测性阈值 [18], 的两侧；在社区间与社区内连接概率之比取中等值时，量子谱检测能够与Louvain区分开来。在特征选择方面，我们使用规模较小的糖尿病 [39]和乳腺癌 [40] 数据集，并展示量子社区检测能够恢复添加到糖尿病数据中的合成社区结构，以及选择出有

信息的特征，从而缩短糖尿病数据上的随机森林训练时间。随后，我们扩展到MNIST [41], ，分别在719节点的相关性图上对像素进行特征选择，并在最多包含2,000个节点的相似性图上直接对图像进行聚类。在整个过程中，我们将其与标准社区检测器进行比较：Louvain、 [7],Girvan–Newman算法、 [6], 贪婪模块度、 [42], Kernighan–Lin、 [43], Newman谱方法， [14] 以及用于特征选择的卡方、互信息和方差分析筛选器。

在此过程中，我们构建了一种处理流程，通过递归二分将 de-teQt 扩展到多社区划分（第 IIIB 节），并提出了 projeQt——一种新型量子社区检测器，它利用量子奇异值变换保留广义模块度矩阵的多个特征向量 [14], ，还可以通过一次划分将图分成多个社区（第 IIIC 节）。

II. 背景

A. 社区检测

本节重点介绍两种经典方法：模块度矩阵的谱二分和 Louvain算法，因为它们是与本研究最相关的主要技术。全文将采用一种特定的模型族来生成合成社区结构：常用的随机块模型（SBM） [9]。为方便后文引用，我们在此对其进行定义。对于含有 N 个节点的随机图，SBM 会生成 K个合成社区，每个社区包含 N/K 个节点。每个节点 i 都带有一个社区标签 $\sigma_i \in \{1,\dots,K\}$，而邻接矩阵 A 则通过概率方式构造，对每一对节点独立进行，具体如下：

$$P(A_{ij}=A_{ji}=1)=\begin{cases} p_{\text{in}}, & \sigma_i=\sigma_j, \\ p_{\text{out}}, & \sigma_i\neq\sigma_j, \end{cases} \quad \forall\, i\neq j, \qquad (1)$$
$$P(A_{ii}=1)=0 \quad \forall\, i.$$

比率 $r=p_{\text{in}}/p_{\text{out}}$ 设定了社区内连接与社区间连接的相对强度。对于 $r=1$，每对节点以相同的概率相连，因此该图没有社区结构。随着 r 增大（即社区内节点之间出现边的概率高于社区间节点之间出现边的概率），社区就越容易被识别。该图的平均度为

$$\bar{c}:=\mathbb{E}[k_i]=\left(\frac{N}{K}-1\right)p_{\text{in}}+\left(N-\frac{N}{K}\right)p_{\text{out}}, \qquad (2)$$

其中， $k_i=\sum_j A_{ij}$ 是节点 i的度。

1. *Modularity and spectral bipartitioning*

模块度由参考文献 [4], 提出，用于衡量某个划分中社区内部边所占比例相较于

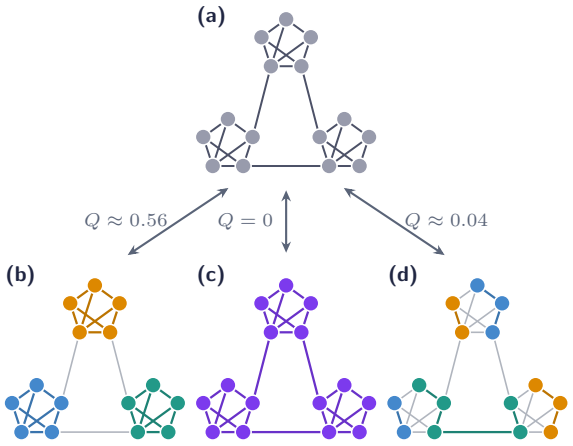

图2：模块度根据落在各社区内部的边数为图的划分评分，并将其与随机关联下的预期边数进行比较；也就是说，与节点度相同但连接经过随机置换的图进行比较。（**a**）该图有27条边。（**b**）–（**d**）该图的三种划分，每个社区各用一种颜色表示：同一社区中两个成员之间的边采用该社区的颜色，而社区之间的边为灰色。（**b**）模块度高：划分遵循三个连接紧密的组，24条边位于社区内部，而随机置换下的预期值为9条。（**c**）将全部27条边纳入同一个社区，恰好等于随机情况下的预期数量。（**d**）该划分横跨连接紧密的物理组：只有10条边位于社区内部，而偶然情况下的预期值为9条。

在保持度不变的零模型下预期比例的超出程度。在该零模型中，边被随机放置，同时保持每个节点的度不变。对于一个包含 N 个节点和 m 条边的图 $\mathcal{G}$，其邻接矩阵为 A_{ij}，节点度为 k_i，该划分的模块度为

$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(c_i,c_j), \qquad (3)$$

其中 $\delta(c_i,\ c_j)\ =1$ 如果节点 i 和 j 被划分到同一个社区，否则为 0。图2 展示了一个小型图上的模块度。数量 $k_ik_j/2m$ 是零模型中节点 i 和 j 之间的预期边数，因此模块度矩阵中的矩阵元素，

$$B_{ij}=A_{ij}-\frac{k_ik_j}{2m}, \qquad (4)$$

当一对节点的连接密度高于预期时为正，否则为负。对于编码为 $\boldsymbol{s}\in\{\pm1\}^N$ 的二社区划分，模块度采用二次型 [13]

$$Q=\frac{1}{4m}\boldsymbol{s}^{\top}B\boldsymbol{s}. \qquad (5)$$

对 $\{\pm1\}^N$ 进行最大化是 NP难 [5]。标准松弛将整数性约束替换为

$\|\boldsymbol{s}\| = 1$。在此约束下，$\boldsymbol{s}^\top B\boldsymbol{s}$ 的最大值由 B 的主特征向量取得，值为 $\lambda_{\max}(B)$。按符号对其各分量进行舍入，即可恢复一个近似二划分。

2. *Generalised modularity matrix*

当尝试使用模块度对先前发现的社区进行递归拆分时，g，我们会遇到一个问题。具体而言，如果将 g 视为一个新图并优化其模块度，就无法优化全局图的模块度，因为没有考虑 g 与 g [13, 14]外部节点之间的连接。Newman 的广义模块度矩阵 [14] 则根据将 g 划分后对整个图模块度产生的变化来为该划分评分：

$$B_{ij}^{(g)} = A_{ij} - \frac{k_i k_j}{2m} - \delta_{ij}\left(k_i^{(g)} - \frac{k_i\, d_g}{2m}\right), \tag{6}$$

其中，$k_i^{(g)} = \sum_{j\in g} A_{ij}$ 是从节点i 连接到 g 的边数，而 $d_g = \sum_{j\in g} k_j$ 是 g的总度。对角线项使得在 g上每一行的和都为零，

$$\sum_{j\in g} B_{ij}^{(g)} = 0. \tag{7}$$

因此，用符号$\boldsymbol{s} \in \{\pm 1\}^{N_g}$ 将 g 一分为二后，整个图的模块度变化为 $\boldsymbol{s}^\top B^{(g)}\boldsymbol{s}/4m$，而 $B^{(g)}$ 的主特征向量在与 B [14]相同的松弛条件下给出了最佳划分。当 g 是整个图时，$B^{(g)}$ 退化为 B。只有在划分能够提高模块度时才予以保留；为此，正的最大特征值是必要但不充分的条件 [13]。

3. *Limitations of spectral modularity methods*

如上所述，谱方法存在一些与本研究相关的局限性。首先是分辨率极限 [16]。根据式（3），合并两个社区A和 B 会使模块度变化为

$$\Delta Q = \frac{l_{AB}}{m} - \frac{d_A d_B}{2m^2}, \tag{8}$$

其中 l_{AB} 是它们之间的边数，而 d_A、d_B 是它们的总度。如果我们有两个社区，每个社区都有 l 条内部边，且两者之间只有一条边（因此 $d_A \approx d_B \approx 2l$），那么合并 A 和 B 将在$l < \sqrt{m/2}$时提高模块度。因此，无论这些小社区内部连接得多么紧密，它们最终都会发生合并。第二个问题是，将社区结构与模块度矩阵的主特征向量联系起来，只有在存在一定程度的结构时才有意义。要使谱方法发挥作用，最大特征值必须大于具有相同度结构但没有社区的随机图的最大特征值。为说明这一点，考虑一个邻接矩阵为 A 的图，该矩阵取自公式（1）的随机块模型（SBM），其中 $K = 2$、$p_{\text{in}} = c_{\text{in}}/N$和 $p_{\text{out}} = c_{\text{out}}/N$，因此 $\bar{c} \simeq (c_{\text{in}} + c_{\text{out}})/2$；令 B 为其公式（4）中的模块度矩阵。期望模块度矩阵 $\mathbb{E}[B]$只有一个非零特征值，即 $\lambda = (c_{\text{in}} - c_{\text{out}})/2$，其特征向量符号 $\text{sign}(\boldsymbol{v}_\lambda)$标记了各个社区。记 $B = \mathbb{E}[B] + X$为期望矩阵加上一个均值为零的随机矩阵$X := B - \mathbb{E}[B]$。根据公式（4），

$$X_{ij} = \underbrace{A_{ij} - \mathbb{E}[A_{ij}]}_{\text{independent entries}} - \underbrace{\left(\frac{k_i k_j}{2m} - \mathbb{E}\left[\frac{k_i k_j}{2m}\right]\right)}_{\text{low rank}}, \tag{9}$$

其中，第二项不会改变 X [17]的特征值范围。$A - \mathbb{E}[A]$ 对角线以上的矩阵元素相互独立，均值为零，且其每一行的总方差为 $\sum_j \text{Var}(A_{ij}) \simeq \bar{c}$。当$\bar{c} \gg \log N$时，这样的矩阵具有谱范数 $\|A - \mathbb{E}[A]\| \simeq 2\sqrt{\bar{c}}$ [44]，因此 X 的所有特征值都位于 $[-2\sqrt{\bar{c}}, 2\sqrt{\bar{c}}]$内。具有相同平均度的随机图具有 $\mathbb{E}[B] \approx 0$，因此 $\lambda_{\max}(B) \simeq 2\sqrt{\bar{c}}$。对于包含社区的情况，最大特征值为 [17]

$$\lambda_{\max}(B) \simeq \begin{cases} \bar{\lambda} + \bar{c}/\bar{\lambda}, & \bar{\lambda} > \sqrt{\bar{c}}, \\ 2\sqrt{\bar{c}}, & \bar{\lambda} \le \sqrt{\bar{c}}, \end{cases} \tag{10}$$

只有当$\bar{\lambda} > \sqrt{\bar{c}}$，也就是当 $c_{\text{in}} - c_{\text{out}} > 2\sqrt{\bar{c}}$时，它才会超过随机图。点$c_{\text{in}} - c_{\text{out}} = 2\sqrt{\bar{c}}$ 就是可检测性阈值 [18]。如果$c_{\text{in}} - c_{\text{out}} < 2\sqrt{\bar{c}}$，任何基于模块度矩阵的谱方法都无法找到这些社区。在本文后续部分，我们将使用多个主特征向量来提取社区结构。在这个例子中，特征值高于 $2\sqrt{\bar{c}}$ 的特征向量与这些社区相关，而低于该值的特征向量则不包含任何相关信息。

4. *The Louvain algorithm*

谱二分要求在递归的每个层级上，对每个被拆分的组 g，都求出 $B^{(g)}$ 的主特征向量。Louvain算法 [7] 速度更快。它通过两个交替进行的贪心阶段优化模块度。在第一阶段，依次考察每个节点 i，并尝试将其移至每个相邻社团 C。将节点 i移入社区 C所带来的模块度增益为

$$\Delta Q_i = \frac{1}{m}\left[k_{i,\text{in}}^{(C)} - \frac{k_i\, \Sigma_C}{2m}\right], \tag{11}$$

其中，$k_i^{(C)}$,在 $= \sum_{j\in C} A_{ij}$ 是从 i 到 C中节点的边权重之和，k_i 是节点 i的度，$\Sigma_C = \sum_{j\in C} k_j$ 是社区 C 的总度。从当前社区 C_0 中移除 i 会改变

模块度变化为 $-\Delta Q_i$ 对 $C_0 \setminus \{i\}$进行评估，因此移动的净增益为 $\Delta Q_i(C) - \Delta Q_i(C_0 \setminus \{i\})$。节点 i 会被重新分配到社区 $C^* = \mathrm{argmax}_C \Delta Q_i$，即能产生最大正净增益的社区；如果没有移动能够提高 Q，则保持不变。重复执行这一轮操作，直到移动任何单个节点都无法提高模块度。在第二阶段，将每个社区收缩为一个超级节点：自环表示内部边权重，社区间的边则进行合并。随后在该粗化图上再次执行第一阶段。两个阶段交替进行，直到划分稳定下来。在稀疏图上，其运行时间似乎与边数呈线性关系 [7, 15], ，并且该算法能够处理拥有数百万个节点的网络 [7]。Leiden算法 [8]通过保证最终划分中的每个社区都是连通的来改进 Louvain；而Louvain并不强制要求这一性质。

5. *Limitations of the Louvain algorithm*

由于 Louvain算法同样最大化模块度，因此它也继承了上文讨论的分辨率极限。此外，它还有另外两个局限。第一，由于其贪心特性，所选划分在很大程度上取决于访问节点的顺序 [7]。如上文所述，第一阶段中的节点移动只有在能够提高 Q时才会被接受。因此，不同的节点排序可能得到不同的局部模块度最大值，而模块度通常存在许多这样的局部最大值，且它们的取值几乎相同 [45]。因此，不同次运行可能返回具有相近模块度的不同划分。相比之下，谱二分仅取决于图本身。当 $\lambda_{\max}(B)$非简并时，划分是唯一的。第二，即使对于由式（1）生成的图，也没有证明存在任何条件能够保证标准 Louvain算法的划分反映这些社区；目前已知的唯一保证适用于一种经过修改的双组版本 [46]。即使是没有社区的随机图，即式（1）中 $r = 1$的情形，也存在模块度很高的划分 [47], ，而 Louvain仍会返回其中一个，尽管它毫无意义。谱二分在此处也会返回一个划分，但对于两个组，式（10）给出了该划分与社区相关的条件，即$c_{\mathrm{in}} - c_{\mathrm{out}} > 2\sqrt{\bar{c}}$。第IVA 节比较了二者在由式（1）生成的图上的表现。

B. 量子社区检测与deteQt

迄今为止，大多数用于社区检测的量子方法都采用以下两种方案之一：一是将模块度矩阵编码到对角哈密顿量中，并通过量子退火求出其基态 [48——53] 或使用变分电路 [54, 55], ；二是采用基于Louvain的算法，并结合基于Grover搜索的容错程序 [56]。此外，还有一些方法使用量子游走，根据游走的概率分布将节点分配到各个社区 [57, 58], ，而文献 [59]表明，通过量子相位估计一个由超图构建的矩阵（Kikuchi矩阵）的最大特征值，在判断超图是否具有社区结构时，可以实现四次加速。在这项工作中，我们关注deteQt量子算法。这是一种基于模块度矩阵谱特征的社区恢复（而不仅仅是存在性检测）算法。deteQt并未采用 N 量子比特上的类Ising编码，而是将主特征向量编码到 $\lceil$对数个$_2\, N\rceil$ 量子比特中，作为式（5）所述松弛问题的解。该算法包含以下组成部分和步骤。

a. 带符号的主特征向量 首先，需要制备 B、 $|\lambda_{\max}\rangle$ 的主特征向量。为了在量子计算机上访问该特征向量，问题被重新表述为基态制备问题，即寻找使 $\langle\psi|(-B)|\psi\rangle$ 最小的量子态；这一过程通过基态制备程序完成。参考文献 [20] 以 VQE [60] 作为示例程序，在 $n = \lceil \log_2 N\rceil$ 个量子比特上生成参数化酉算子 $U_{\max}(\boldsymbol{\theta})$，其中 $U_{\max}(\boldsymbol{\theta}^*)\ |0\rangle^{\otimes n} \approx |\lambda_{\max}\rangle$。下一步是将该量子态的实值振幅转换为 $\{\pm 1\}$ 符号，用以标记两个社区。在能够受控访问 $U_{\max}(\boldsymbol{\theta}^*)$ 的前提下，参考文献 [61] 中的方法构造出对角矩阵 $A = \mathrm{diag}(a_1, \ldots, a_{2^n})$ 的块编码，其中对角元素为各个振幅。该构造需要调用 $O(1)$ 次受控 $U_{\max}$。矩阵 M 的块编码是一种酉算子，其左上块为 M/α [21, 62]。应用 QSVT [21]˜，并使用次数为 d 的符号函数多项式，可得到 $A = \mathrm{diag}(\mathrm{sign}(a_1) \ldots, \mathrm{sign}(a_{2^n}))$。将其应用于 $H^{\otimes n}\,|0\rangle$ 个系统量子比特的均匀叠加态，QSVT 电路中辅助量子位处于 $|0\rangle$ 的分支是一个实等幅（REW）量子态，其中每个振幅都被压平为其符号

$$|\boldsymbol{s}\rangle \approx \frac{1}{\sqrt{2^n}} \sum_{i=1}^{2^n} s_i\,|i\rangle\,, \tag{12}$$

其中，当 N 不是二的幂时， $s_i = \mathrm{sign}(a_i) \in \{+1, -1, 0\}$ 给出 $i \le N$ 中节点 i 的社区标签；对于填充索引 $i > N$，则为 $s_i = 0$。多项式次数 d 控制符号近似在零附近的尖锐度。较大的 d 可提高保真度，但代价是需要更深的电路 [62]。

b. 探测态与线性组合酉算子准备好带符号特征向量后，第二阶段将确定它编码的是哪种二划分。候选集 $\mathcal{G}_{\mathrm{cand}}$ 由所有恰有 k 个矩阵元素等于 $+1$ 的 $\binom{N}{k}$ 符号向量组成，代表将 k 个节点分配到一个社区的所有可能方式。为了从中进行区分，我们根据汉明重量为 $k' = N/2 - k$ 的超图态 [63] 构造一组互补的探测态 $\mathcal{G}_{\mathrm{LCU}}$。其关键性质是，每个错误候选至少与一个探测态的重叠度为零。探测态的数量为 $|\mathcal{G}_{\mathrm{LCU}}| = \binom{N}{k'}$，并将它们嵌入酉算子线性组合（LCU）中：

$$|\psi_{\mathrm{hyp}}\rangle = \frac{1}{\sqrt{|\mathcal{G}_{\mathrm{LCU}}|}} \sum_{x \in \mathcal{G}_{\mathrm{LCU}}} |x\rangle\,|G(x)\rangle\,, \tag{13}$$

其中，$|G(x)\rangle$ 是探测态 x 对应的超图态。每次对辅助量子位进行测量都会返回一个探测态索引 x，其概率与 $|\langle G(x)|s\rangle|^2$ 成正比，也就是该探测态与带符号特征向量之间重叠度的平方。

c. 确定性消除 deteQt 利用这一零重叠特性，采用如下确定性消除策略。给定探针索引 x，每个满足 $G_c \in \mathcal{G}_{\text{cand}}$ $|\langle G_c|G(x)\rangle|^2 = 0$（或低于某一阈值）的候选项都会被消除，因为它不可能产生该结果。使用新的探针样本重复这一过程，候选集会逐渐缩小，直到只剩一个赋值或预算耗尽。该方法之所以有效，是因为正确划分与每个可采样探针都有非零重叠度，而每个错误候选项至少与一个探针的重叠度为零。因此，轮次数按 $\log_2|\mathcal{G}_{\text{cand}}| = \log_2\binom{N}{k}$的量级增长，前提是参考文献 [20] 假设每个样本能消除约一半剩余候选项。放宽零重叠要求后，就得到参考文献 [20], 考虑的第二种情况，即 小重叠情形。它需要$\binom{N}{k_{\text{LCU}}}$ 个汉明重量固定且较小的探针，k_{LCU}因此，当 $k_{\text{LCU}} = 1$ 时，我们有 $\binom{N}{1} = N$ 个探针。随后利用这些探针，消除与某个采样探针的重叠度低于特定阈值的候选项。

C. 特征选择

本工作应用社区检测的两个数据分析任务中，第一个是特征选择。特征选择从 N 可用的输入变量中选取一个子集，同时保留预测价值或分析价值。当 N 相对于样本数量较大时，对所有特征进行拟合会带来过拟合风险，并增加不必要的计算成本 [22, 23]。标准分类体系将方法分为：1）过滤方法，根据统计标准（互信息 [64, 65], 相关性、卡方检验）对特征进行评分，而不参考学习算法，但会遗漏多变量交互作用；2）封装方法，通过交叉验证的模型性能评估候选子集 [22, 66], ，但面临 2^N搜索空间；3）嵌入式方法，在训练过程中直接执行选择（ℓ_1正则化回归 [67, 68], 基于树的重要性 [69]），但会使选定特征依赖于单一模型类别的归纳偏置 [23]。除 mRMR等关注冗余的过滤方法 [65], ，以及条件判据 JMI和CMIM [70]、FCBF [71], （这些方法在评分函数中惩罚特征之间的依赖关系）外，这些方法族都不会显式表示特征之间的相关结构。接下来介绍的基于图的方法能够表示这种结构，也是本工作所采用方法的基础。

1. *Community detection for feature selection*

特征选择的另一种表述方式是构建一个图，其节点为特征，边则根据成对相似性（相关性、互信息、模糊粗糙依赖或类似指标）赋予权重 [24, 25]。相关的谱方法，如拉普拉斯评分 [72]和SPEC框架 [73] ，利用样本相似性图而非特征图对特征进行排序，但建立了图谱与特征相关性之间更广泛的联系。分组方法对特征图进行划分，并从每个组中保留一个或多个代表特征。这种基于图的特征选择理念可以直接融入社区检测。例如，文献 [26] 在特征上构建加权Spearman相关网络，其中仅保留与类别的互信息超过阈值的特征，随后不断添加边直至图连通，检测社区，并从每个社区中保留一个或多个特征。文献 [27] 以皮尔逊相关性为边赋予权重，使用ISCD算法寻找社区，并利用这些社区约束遗传算法，使其从每个社区中选取指定数量的特征。文献 [28] 基于鲁棒模糊粗糙模型构建特征交互图，在其上运行Louvain算法 [7] ，然后使用同一模型对每个社区内的特征进行排序。在生物信息学中，加权基因共表达网络分析（WGCNA） [74] 对基因相关网络采用类似的处理流程，并已成为高维组学数据的标准预处理步骤。该方法对相关性进行软阈值处理 [75], ，通过层次聚类寻找模块 [76], ，并用其特征基因或模块内枢纽基因概括每个模块 [77]。在给定邻接矩阵的情况下$|r_{ij}|$，枢纽基因是指与其模块中其他特征的平均绝对相关性最大的特征。等价地，对于距离 $1-|r_{ij}|$，枢纽基因与其模块中其他特征之间的平均距离（MAD）最小。我们在本工作中使用该特征作为代表。文献 [78] 表明，使用通常零模型对相关矩阵进行社区检测会产生偏差，因此改用随机矩阵理论中的零模型。这同样适用于我们的处理流程：该流程在经过阈值处理的相关图上使用通常的零模型（第IIIA节）。这些方法的成功取决于底层社区检测的质量，因此，对社区检测算法的任何改进也都能提升许多下游任务和应用的效果。

III. 数据分析处理流程

我们为多种可能的数据分析任务定义了一个共享处理流程。根据数据构建图（第 IIIA节）。deteQt（第 III B节）或 projeQt（第 IIIC节）的量子社区检测器将其划分为K 若干社区，并可选择进行经典优化。随后返回每个社区的一个代表性特征，或返回社区本身，分别用于特征选择或聚类

（第 IIID节）。两种方法都最大化式（3）的模块度，区别在于它们从 B的谱中读取划分的方式。deteQt 根据一个特征向量上的探针态读取二分结果，并递归执行。projeQt 保留高于某个割的谱部分，并通过对滤波后的状态进行核 k均值聚类来划分一个群组（图 1，附录 C）。

A. 图

a. 特征图 在进行特征选择时，节点代表特征；当两个特征的皮尔逊相关超过阈值时，就在它们之间连一条边。固定阈值存在两个问题。首先，它不会随样本量变化而调整：在包含 50 个样本的数据集中，$r = 0.1$的相关性只是噪声；但在包含 10，000 个样本的数据集中，这种相关性则具有高度显著性。其次，用户定义的阈值是一个超参数，如果选取不当，就会影响下游性能。为此，我们根据皮尔逊相关的双尾 t 检验推导阈值。给定样本量为 n 的样本，相关系数 r 的检验统计量为

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \tag{14}$$

其自由度为 $n-2$。在显著性水平 α 下对其求逆，可得在统计上可与零区分的最小相关性：

$$r_{\text{crit}} = \frac{t_{\text{crit}}(\alpha,\, n-2)}{\sqrt{t^2_{\text{crit}}(\alpha,\, n-2) + (n-2)}}, \tag{15}$$

其中，$t_{\text{crit}}(\alpha, n-2)$ 是t分布的双尾临界值。我们始终设定 $\alpha = 0.05$。当 $|r_{ij}| \geq r_{\text{crit}}$时，特征 i 与j 之间的边被纳入特征图 $\mathcal{G}$，其权重为 $|r_{ij}|$。注意，$|r_{ij}| = 1$ 涉及特征 i 与 j之间完全相关的成对比较。对于较大的数据集，这会生成更稠密的图，因为弱相关性承载着真实信号；而对于较小的数据集，则会生成更稀疏的图，因为弱相关性并不承载真实信号。相关矩阵使用所有可用样本计算，不使用标签信息。文献 [78] 表明，模块度的通常零模型在相关矩阵上存在偏差。我们仍采用通常的零模型。该阈值只会移除无法与噪声区分的边（第四节 B）。在弱相关性掩盖社区结构的数据集上，某次划分可能被接受但效果不佳，因为噪声边会扭曲特征向量。因此，处理流程从 τ_0开始，在移除低于权重阈值 τ 的边后于特征图上运行，并将剩余的边视为无权边。其接受检验（第三节 B）可以拒绝某次划分。当发生拒绝时，τ 增加 0.1，并在移除弱边后重新检验该子图。如果 τ 超过上限（$\tau > 1.0$），则保留该社区不再划分，并将 τ 返回到 τ_0。成功划分后不会重置该值，因此提高后的阈值会延续到下一组。该重试循环可以揭示在稠密图中不可见的结构。我们将 τ_0 视为一个超参数，其默认值为 r_{crit}（公式（15））。附录 A1 比较了对于小型数据集，τ_0、r_{crit} 和0。95这两个取值的影响。

b. 数据点图 对于聚类问题，节点代表数据点（可推广到任意类型的数据），每个节点都与其最近邻相连，且边不加权（权重为 1）（第 IVE 节）。正确定义每个节点的邻域会改变结果。由于每条边的权重均为 1，$\tau_0 = 0$ 因此阈值不会对此产生影响（第 IVA 节）。

B. deteQt用于 K 个社区

a. 一次二分 deteQt按照第 IIB 节所述协议将一个群体一分为二，使用 N 小重叠情形的单节点探针，$k_{\text{LCU}} = 1$。通过VQE电路 $U_{\max}$制备群体模块度矩阵的主特征向量 a_i 的振幅，并将其块编码到对角矩阵中，如 [20]所示。随后使用量子奇异值变换，对一个拟合于双曲正切函数($d = 5$)的10x次奇多项式进行处理，以近似符号函数 $\texttt{sign}(a_i) \in \{\pm 1\}$。本工作仅考虑 $k_{\text{LCU}} = 1$，因此每个探针对应一个特定汉明重量的 $= 1$量子态。因此，探针 j（仅翻转节点 j 的超图态）被抽取的概率为 $p_j \propto |\langle G_j|\psi\rangle|^2$。我们取 $S = 1000$ 个样本。对于每个节点 i，频率ν_i 定义为

$$\nu_i := \sum_{s=1}^{S} \mathbb{1}[j_s = i], \tag{16}$$

每次测量返回的是探针索引 j_s，而非社区标签。由于探针 i 对应于量子比特（节点）i被翻转的 HW$= 1$量子态，ν_i 统计 S 次抽取返回探针 i 的次数，以及 $\mathbb{E}[\nu_i] = Sp_i$。‘正确’的社区规模 k^*（不同于探针规模，k_{LCU}）可按参考文献 [20] 所述进行估计。deteQt 按 ν_i 对节点排序，将频率最高的 k^* 个节点分配到一个社区，并将其余节点分配到另一个社区。

b. 接受 由于测量过程并非确定性的，我们加入门控机制，以检测所提出社区的统计显著性。一旦提出社区划分，deteQt 就会检验其探针分布是否不同于均匀分布。对于探针 j，在 S 次抽取中的预期计数为 Sp_j，而均匀分布下的预期计数为 S/N。我们使用χ^2 拟合优度检验，将估计分布与均匀分布进行比较；当其 p-值不超过 0.05 时，接受该划分。与均匀分布足够接近的分布无法可靠地预测社区，因此拒绝此次拆分。在这种情况下，我们提高

阈值，τ 增加 $\Delta\tau$，以移除较弱的边，然后重复该过程。如果 τ 达到 1.0，则保持图不拆分。由所有图节点组成的社区（即 $k^* = 0$）也会被拒绝。

在本文的所有实验中，我们使用 $S = 1000$ 个样本。借助 $k_{\mathrm{LCU}} = 1$ 和精确符号，探针 j 在除 j 外的每个节点上的符号均为 $+1$，其重叠度为 $\langle \pi_j | \boldsymbol{s} \rangle = \sigma - 2s_j$，其中 $\sigma = \sum_i s_i$。对于平衡划分，$\sigma = 0$，因此所有探针的重叠度大小相同，分布也是均匀的。因此，精确符号舍入会导致测试拒绝平衡划分。五阶符号近似保留幅度大小，并打破这种退化。

c. 递归 为获得 K 个社区，deteQt 递归地应用被接受的二分。它将候选组保存在先进先出队列中，拆分下一个组，并将其两个部分添加到队尾。在阈值重试后仍无法拆分的组将被最终确定。当队列和已最终确定的组总共包含 K 个社区，或队列为空时，递归结束。

在组的每次划分中 g，deteQt 使用公式 $B^{(g)}$ (6) 中的广义模块度矩阵，并重新运行上述单次划分处理流程。请注意，在第一次社区划分后，我们不会修改图或阈值，τ。

如果递归在达到 K之前结束，deteQt 会强制进行进一步划分。在具有可用基于探针的二分方案的组中，它会选择 χ^2 p值最小的组，而不应用 0。05 的接受阈值。如果没有组具有基于探针的划分方案，则会进行随机划分。在后文的结果中，我们将展示，在达到某些图规模后，这种情况会频繁发生，而采用若干提议可以缓解这一问题。由于我们使用社区检测流程来执行进一步的数据分析任务，因此我们发现，至少对于此处选择的问题，没有必要使用参考文献 [20], 中的零重叠确定性消除流程。扩大探针量子态的维度，或者如我们在 projeQt 中所提议的那样，使用多个主特征向量而不仅仅是主特征向量，会更加有效和高效（第 IIIC节）。对小重叠度和低汉明重量的限制显著减少了所需的电路深度。

C. projeQt： k-路谱分裂

随着图规模增大，递归二分会在每一层累积误差，因为每次划分都在进行下一次划分之前被固定下来，而且可能无法找到将图划分为两个以上组的最佳方案 [14, 19]。此外，由于 deteQt 仅处理主特征向量，因此在不采用递归的情况下，无法明显地直接生成多个社区。我们提出的第二种方法是 projeQt，这是一种使用多个主特征向量、能够原生执行多社区检测的量子算法。projeQt 直接作用于正在划分的组的广义模块度矩阵 $B^{(g)}$ g，

个 N 节点（第 IIA2 节），通过补零填充至大小 2^n，$n = \lceil \log_2 N \rceil$，并按照文献 [21, 79, 80]所述方法生成如下量子态:

$$U \left|0\right\rangle_a \left|j\right\rangle_s = \left|0\right\rangle_a \frac{B^{(g)}}{\lambda} \left|j\right\rangle_s + \left|\perp\right\rangle, \tag{17}$$

其中，a 是一个辅助寄存器，s 是系统寄存器，二者各包含 n 个量子比特，而 $|\perp\rangle$ 在 $|0\rangle$之外有 a 。该量子态首先通过将标记为 j的列加载到以下量子态中来生成

$$U_R \left|0\right\rangle_a \left|j\right\rangle_s = \left|r\right\rangle_a \left|j\right\rangle_s, \qquad \left|r\right\rangle := \sum_k \frac{\|B^{(g)}_{k\cdot}\|}{\lambda} \left|k\right\rangle, \tag{18}$$

其中 $\lambda = \|B^{(g)}\|_F$，使得 $|r\rangle$ 包含$B^{(g)}$的行范数。随后加载行索引 i，以生成量子态

$$U_L \left|0\right\rangle_a \left|i\right\rangle_s = \left|i\right\rangle_a \left|b_i\right\rangle_s, \qquad \left|b_i\right\rangle = \sum_j \frac{B^{(g)}_{ij}}{\|B^{(g)}_{i\cdot}\|} \left|j\right\rangle, \tag{19}$$

其中 $|b_i\rangle$ 行 i 已归一化。 $U = U_L^\dagger U_R$ 充当一种块编码形式，因为在 $|0\rangle$上后选择辅助寄存器可得到

$$\begin{aligned} \left\langle 0\right|_a U \left|0\right\rangle_a &= \frac{B^{(g)}}{\lambda}, \\ \left\langle 0, i\right| U \left|0, j\right\rangle &= \left\langle i|r\right\rangle \left\langle b_i|j\right\rangle = \frac{B^{(g)}_{ij}}{\lambda}. \end{aligned} \tag{20}$$

执行这些操作的电路由R_y 和 CNOT 门组成。在制备好公式（17）的量子态后，我们使用量子奇异值变换来近似作用于 $B^{(g)}$ 的特征值的以下函数:

$$\begin{aligned} f(x) &= \frac{1}{2}\left[1 + \tanh\left(\kappa \frac{x - c}{\|B^{(g)}\|_F}\right)\right], \\ c &= \max\left(e, \frac{\lambda_{\max}}{2}\right), \qquad e = \frac{2\|B^{(g)}\|_F}{\sqrt{N}}. \end{aligned} \tag{21}$$

按如下方式投影出最高特征向量。记 $B^{(g)} = \sum_k \lambda_k \left|v_k\right\rangle \left\langle v_k\right|$，其中每个 $|\lambda_k| \leq \|B^{(g)}\|_F = \lambda$，因此每个 λ_k/λ 都属于 $[-1, 1]$。对式（17）所示编码应用 QSVT，会通过多项式变换每个特征值 λ_k/λ，而保持特征向量不变 [21]。该电路应用一个拟合于 $x \mapsto f(\lambda x)$ 的多项式 p，使得 $p(\lambda_k/\lambda) \approx f(\lambda_k)$。记 $p = p_{\mathrm{e}} + p_{\mathrm{o}}$ 为其偶部和奇部，并令 U_{e} 和 U_{o} 为相应的 QSVT 酉算子。LCU 在双量子位控制寄存器 ℓ 上制备关于 U_{e}、$U_{\mathrm{e}}^\dagger$、U_{o} 和 $U_{\mathrm{o}}^\dagger$ 的均匀叠加。令 a 为块编码辅助寄存器。伴随对会抵消不需要的虚部 QSVT 块；对联合全零寄存器进行后选择，

会在系统寄存器上留下一个算子：

$$\begin{aligned} |0_{\ell a}\rangle &\coloneqq |00\rangle_{\ell}\,|0\rangle_{a}\,, \\ \mathcal{S} &\coloneqq \frac{1}{4}\big(U_{\mathrm{e}}+U_{\mathrm{e}}^{\dagger}+U_{\mathrm{o}}+U_{\mathrm{o}}^{\dagger}\big), \\ \langle 0_{\ell a}|\,U_p\,|0_{\ell a}\rangle &= \langle 0|_{a}\,\mathcal{S}\,|0\rangle_{a} \\ &= \frac{1}{2}\left[p_{\mathrm{e}}\Big(\frac{B^{(g)}}{\lambda}\Big)+p_{\mathrm{o}}\Big(\frac{B^{(g)}}{\lambda}\Big)\right] \\ &= \frac{1}{2}p\Big(\frac{B^{(g)}}{\lambda}\Big)=\frac{1}{2}\sum_{k}p\Big(\frac{\lambda_k}{\lambda}\Big)\,|v_k\rangle\,\langle v_k|\,. \end{aligned} \tag{22}$$

由于 $\tfrac{1}{2}(1+\text{双曲正切 } z) = 1/(1+e^{-2z})$，每个特征向量的权重为

$$f(\lambda_k)=\frac{1}{1+e^{-2\kappa u_k}},\qquad u_k=\frac{\lambda_k-c}{\lambda},$$
$$\lim_{\kappa\to\infty} f(\lambda_k)=\begin{cases}1, & \lambda_k>c,\\ \tfrac{1}{2}, & \lambda_k=c,\\ 0, & \lambda_k<c.\end{cases} \tag{23}$$

对于远离割的固定特征值，向这些极限的收敛关于 κ 是指数级的：

$$\begin{aligned} 1-f(\lambda_k) &\le e^{-2\kappa u_k}, \quad u_k>0,\\ f(\lambda_k) &\le e^{-2\kappa|u_k|}, \quad u_k<0, \end{aligned} \tag{24}$$

公式（21）中的割取两个阈值中较大的一个。第一个阈值 e 设定了该群的谱尺度。由于$\|B^{(g)}\|_F^2=\sum_k\lambda_k^2$，$\|B^{(g)}\|_F/\sqrt{N}$ 是均方根特征值。我们设定$e = 2\|B^{(g)}\|_F/\sqrt{N}$，因此只有显著大于此值的特征值才能通过第一道截断。第二个阈值 $\lambda_{\max}/2$防止结构性很强的群保留高于该截断值的每个方向：它只保留特征值与最大特征值相当的方向。公式（21）中的所有项都可以通过以下方式获得：对于 e，廉价地计算 $B^{(g)}$ 的弗罗贝尼乌斯范数；或者对于 $\lambda_{\max}$，进行基态特征求解 $-B^{(g)}$，具体取决于 deteQt 的能力。换言之，我们不要求显式计算 $-B^{(g)}$ 的激发态。在下面的每个结果中，我们设定 $\kappa = 32$ 并使用 31 阶切比雪夫多项式

$$\begin{aligned} p_{31}(x) &= \sum_{q=0}^{31} a_q T_q(x) \approx f(\lambda x) \\ &= \frac{1}{2}\left[1+\tanh\left(32\left(x-\frac{c}{\lambda}\right)\right)\right], \qquad x\in[-1,1]. \end{aligned} \tag{25}$$

其中 T_q 是次数为 q 的切比雪夫多项式，系数 a_q 针对该目标变量进行拟合，并且绝对值不超过 1。未来的工作可以使用广义量子信号处理或量子奇异值变换的相位合成程序，直接实现这个非定宇称多项式，而无需进行奇偶拆分和四分支线性组合酉算子 [81, 82]。为了解理想谱滤波器本身的作用，写作

$$\omega_k \coloneqq \frac{1}{1+\exp\left[-2\kappa(\lambda_k-c)/\lambda\right]}.$$

相应的谱滤波器及其对节点i 的作用为

$$\begin{aligned} F_\kappa &= \sum_k \omega_k\,|v_k\rangle\,\langle v_k|\,, \\ F_\kappa\,|i\rangle &= \sum_k \omega_k\,\langle v_k|i\rangle\,|v_k\rangle\,, \\ |\phi_i\rangle &= \frac{F_\kappa\,|i\rangle}{\|F_\kappa\,|i\rangle\|}. \end{aligned} \tag{26}$$

有限多项式实现采用相同的构造方式，只是将滤波器值替换为相应的多项式值。每个节点的量子态都包含有关该节点如何参与社区结构的信息。因此，两个节点之间的相似度表明它们是否属于同一个社区。我们通过组装格拉姆矩阵 G来计算这种相似度，其分量为 G_{ij}：

$$\begin{aligned} G_{ij} &= \langle i|\,F_\kappa^2\,|j\rangle = \sum_k \frac{\langle i|v_k\rangle\,\langle v_k|j\rangle}{\left[1+\exp\left(-2\kappa(\lambda_k-c)/\lambda\right)\right]^2}, \\ \langle\phi_i|\phi_j\rangle &= \frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}} \end{aligned} \tag{27}$$

当 $\kappa\to\infty$趋于极限时，对于 $\lambda_k > c$，有 $\omega_k \to 1$，其他情况下为 0，因此

$$\begin{aligned} \lim_{\kappa\to\infty} G_{ij} &= \sum_{\lambda_k>c}\langle i|v_k\rangle\,\langle v_k|j\rangle\,, \\ \langle\phi_i|\phi_j\rangle &\longrightarrow \frac{\displaystyle\sum_{\lambda_k>c}\langle i|v_k\rangle\,\langle v_k|j\rangle}{\sqrt{\displaystyle\sum_{\lambda_k>c}|\langle v_k|i\rangle|^2\sum_{\lambda_k>c}|\langle v_k|j\rangle|^2}}. \end{aligned} \tag{28}$$

为估计格拉姆矩阵元素，我们对有序的非对角线元素对使用 N（$N-1$）符号 Hadamard 测试，对角线元素则由滤波量子态的范数给出（附录 1）。最后一步是使用所请求的社区数量进行核 k-均值聚类。仅当所得划分中每个社 $n_{\min}=$区至少包含 $\big(2,\ \lfloor N/(2K)\rfloor\big)$ 个节点时，我们才接受该划分；这可以防止划分产生单节点或非常小的社区，同时还要求整个图的模块度得到提升。若某个组的最大特征值 $\lambda_{\max}$ 不为正，则不会对其进行划分，因为它不存在能够带来正增益的谱方向。由于算法可能失败，我们在社区选择中加入了多种回退机制。首先，projeQt 可以采用递归方式，先选择 K 个社区，然后继续划分能够被划分的最大组，直到剩下所请求的 k^* 个社区。不同于 de- teQt，这并非严格要求，因为也可以直接将 $k^* = K$ 设置为初始值。其次，如果所请求的划分未被接受，projeQt 会在同一个格拉姆矩阵上重新运行核 k-均值聚类，并对得到的两个组应用 Kernighan–Lin [43],这一独立的基于图的细化方法。如果无法以这种方式划分任何组，则会针对模块度增益最大的组，采用核 k-均值聚类在滤波嵌入上返回的二分标签，

即使该增益为负。当没有任何组的滤波特征值高于其割值时，才会使用随机二分（附录 D）。

D. 从社区到输出

a. 特征选择 处理流程将特征图划分为 K 个社区后，会从每个社区中选取一个代表性特征。在社区 C中，我们计算每个特征与其他成员的平均距离，距离定义为 $1-|r_{ij}|$:

$$\mathrm{MAD}(i)=\frac{1}{|C|-1}\sum_{\substack{j\in C\\ j\neq i}}\bigl(1-|r_{ij}|\bigr). \tag{29}$$

代表性特征为 $\arg\min_{i\in C}\mathrm{MAD}(i)$，即平均距离最小的特征（MAD）。该特征与其余 C 成员的平均绝对相关性最高，也是社区中最居于中心位置的特征。这一准则与目标变量无关：它不会将选定特征与特定响应变量绑定，因此适用于监督式和无监督式场景。对于单成员社区（$|C|=1$），则直接选取唯一成员。使用选定特征在 80% 的样本上训练随机森林分类器，并报告其在其余 20% 样本上的准确率，该准确率按随机种子取平均（附录 D）。

b. 聚类 社区即为簇。我们根据簇与标签的最佳匹配计算准确率，并使用归一化互信息和调整兰德指数进行评分。

IV. 结果

我们在五种图规模和复杂度逐步增加的设置下评估该处理流程。首先，利用从公式（1）的随机块模型中生成的图，将两种方法与 Louvain 进行比较（第 IVA节）。随后，我们将该处理流程应用于三个真实数据集的特征选择：皮马印第安人糖尿病数据集 [39]（768 个样本、8 个特征，扩充至 17 个；第 IVB节）、威斯康星乳腺癌数据集 [40]（569 个样本、30 个特征；第 IVC节）以及 MNIST [41]（70,000 个样本、719 个像素特征；第 IVD节）。最后，我们通过构建数据点相似图而非特征相关图，将同一处理流程应用于 MNIST 的无监督聚类，并将规模扩展至 2,000 个节点（第 IVE节）。

A. 植入划分图上的谱方法与Louvain

我们比较了 deteQt、projeQt、Newman 的主特征向量方法 [13], 和 Louvain 在从式（1）的随机块模型中抽样得到的图上的表现。比率 $r=p_{\mathrm{in}}/p_{\mathrm{out}}$ 控制社区信号。我们使用具有 $N=200$、平均度为 20，且 $K=4$ 或 10 个植入社区的图

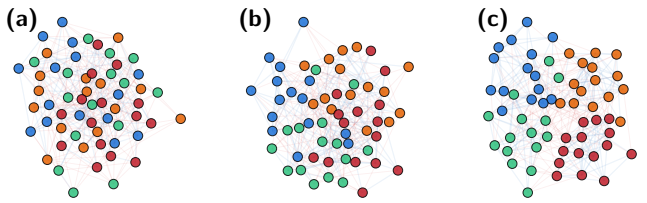

图 3：随机块模型图（式（1）， N =60， $K=4$，平均度为 10， $r=1.5$，3 和 5），其社区信号强度 $r=p_{\mathrm{in}}/p_{\mathrm{out}}$ 逐渐增加。节点按植入社区着色，蓝色边为社区内边-

社区，红色边表示社区间连接。**(a)** $r=1.5$，弱信号。**(b)** $r=3$，中等信号。**(c)** $r=5$，强信号。

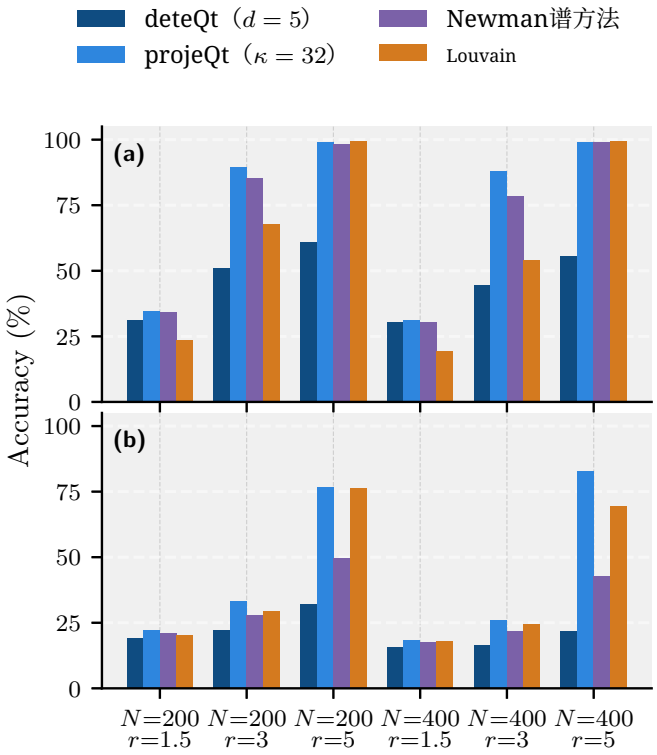

图4：根据公式（1）在随机块模型图上的社区恢复准确率，平均度为20，取五个随机种子的平均值。**(a)** $K=4$ 个合成社区。**(b)** $K=10$个合成社区。Louvain 自行选择社区数量，并根据该划分进行评分。

在 $r=1.5$、3 和 5 时。图 4 报告了经过最优一对一匹配后，被分配到其预设社区的节点比例。对于谱方法，要求其识别预设的 K，而 Louvain 会自动选择自己的社区数量。在 $r=1.5$时，所有方法的表现都接近随机水平。在 $r=3$时，对于 $K=4$ 在 $N=200$和 400 的情况，projeQt 分别比 Louvain 高出 22 和 34 个百分点；对于 $K=10$，则分别高出 4 和 2 个百分点。在 $r=5$时，projeQt、Newman 和 Louvain 几乎都找回了全部四个社区。对于 $K=10$，projeQt 和 Louvain

在 $N = 200$时表现相当，而在 $N = 400$时，projeQt 领先 13 个百分点。deteQt 在十二种设置中的八种低于 Louvain。我们还在一个具有 $N = 200$、$K = 5$和 $r = 3$的 SBM 图上，让每种方法分别运行 20 次。图 5 显示，projeQt 的平均准确率为 81。$4\% \pm 1.2\%$，Newman 为 57。$8\% \pm 0.8\%$，Louvain 为 52。$4\% \pm 9.0\%$，deteQt 为 37。$6\% \pm 2.2\%$。projeQt 在不同随机种子下的波动比 Louvain 小七倍。此外，Louvain 计算所得结果的方差更大。

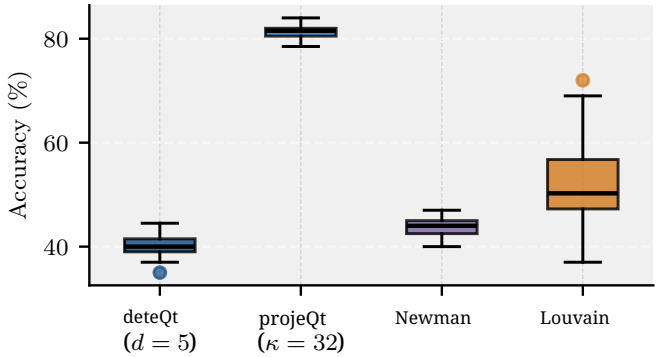

图 5：根据公式（1）生成的一个 SBM 图上的稳定性，其中 $N = 200$、$K = 5$和 $r = 3$，跨越 20 个随机种子。

B. 糖尿病特征图与选择

皮马印第安人糖尿病数据集 [39]包含 768 个样本、八个临床特征，以及一个表示是否患有糖尿病的二元标签。我们通过加入植入特征扩充该数据集：分别为妊娠次数、葡萄糖和血压添加三个带噪副本，使特征总数达到 17 个。这提供了一个可控的设定，其中预期的社区结构是预先已知的。图6展示了 projeQt 的一次电路运行，参数为$K = 3$。可以看到，每个植入的（带噪）特征都与其来源特征归为一组，符合预期。未重复的特征则被分配到其中一个$K = 3$ 选定社区中。对于特征选择，我们使用中位数绝对偏差（MAD），从每个社区中选取一个代表性特征（或质心），并使用选定特征训练随机森林分类器。我们将 deteQt 和 projeQt 与 Louvain、Girvan–Newman算法、 [6]，方差分析和卡方选择进行比较。为便于比较，我们要求每种方法返回 2 到 5 个特征。请注意，并非所有社区检测器都将选定的社区数量作为参数：当图7 中未生成指定数量的社区时，我们会进行随机划分或合并。图7比较了这些方法在 17 节点特征图上的表现。结果表明，对于这一问题，我们提出的方法可与最先进的社区检测器相媲美，有时甚至优于它们。最后，虽然选定的社区在绝对准确率方面表现良好，我们还研究了它们对模型复杂度本身的影响。

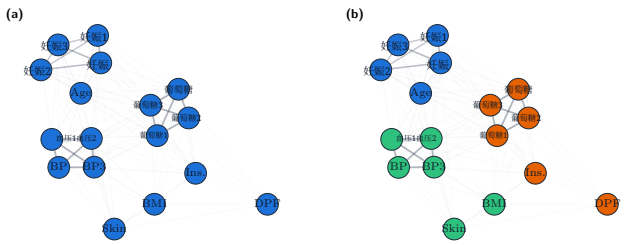

图 6：使用 projeQt 社区检测得到的增强型糖尿病数据集（17 个特征）的特征相关图。节点表示特征，边连接满足 $|r_{ij}| \geq r_{\text{crit}}$ 的特征对。文本标签标识每个源特征及其三个合成副本（如有）。(a) 未划分的输入图。(b)projeQt 电路返回的三个社区，其中 $K = 3$。projeQt 正确地将带噪特征与其来源分组；由于我们只指定了 $K = 3$，未重复的特征也被纳入了其中一个社区。

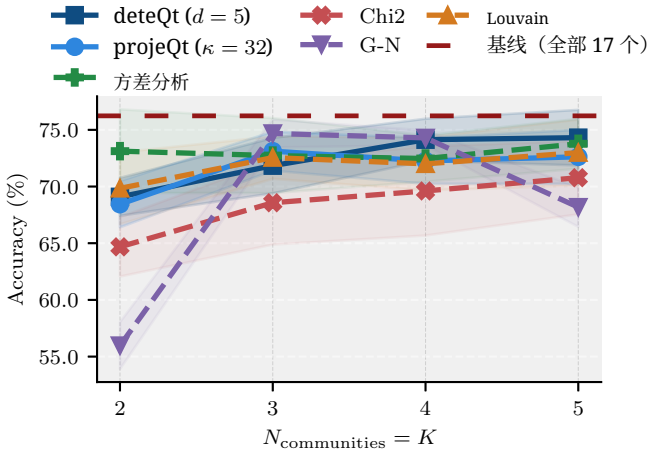

图 7：增强版 Diabetes 数据集上的分类准确率随社区数量变化的情况。

(表I）。需要注意的是，尽管特征选择使模型拟合速度更快，但其预测速度却比使用全部 17 个特征时更慢，因为生成的决策树从 42，500 个树节点增长到 51，$200 - 92$，800 个树节点。拟合这些数据所需的更大模型可能源于这样一个事实：我们还将独立特征与重复特征分组，并且每个子集只选择一个特征。因此，我们也丢弃了有用的特征信息。尽管如此，在社区检测方法中，projeQt 在预测和拟合方面都不逊于更快的方法，而 deteQt 仍然具有竞争力。

C. 乳腺癌数据集上的特征选择

其次，我们在没有合成特征的威斯康星乳腺癌数据集上进行测试 [40]。该数据集包-

表 I：增强型糖尿病数据集上的随机森林成本（250 棵决策树，154 个样本的测试集划分），各随机种子上的均值 ± 标准差，单位为 CPU 时间的毫秒数。预测时间为每个拟合后的随机森林 20 次调用的中位数，拟合时间仅测量一次。基线使用全部 17 个特征。平均行报告 $\bar{t} = \frac{1}{4}\sum_{K=2}^{5} t_K$，并假设 K 间独立运行，将标准差传播得到：$\sigma_{\bar{t}} = \frac{1}{4}\sqrt{\sum_{K=2}^{5} \sigma_K^2}$，其中 σ_K 是每个 K 上的随机种子标准差。

K	基线	方差分析	χ^2	互信息	Girvan–Newman算法	Louvain	projeQt	deteQt
				Prediction				
2	11.6	12.4 ± 0.5	12.4 ± 0.4	12.6 ± 0.2	13.2 ± 0.3	12.3 ± 0.5	**12.2** ± 0.3	12.4 ± 0.3
3	11.6	**11.9** ± 0.2	12.2 ± 0.3	12.1 ± 0.3	12.0 ± 0.2	12.3 ± 0.3	12.1 ± 0.3	12.4 ± 0.4
4	11.6	12.0 ± 0.2	11.9 ± 0.1	12.4 ± 0.2	12.1 ± 0.2	12.1 ± 0.4	11.9 ± 0.4	**11.9** ± 0.3
5	11.6	**11.7** ± 0.3	12.0 ± 0.2	11.9 ± 0.2	11.8 ± 0.1	11.9 ± 0.2	11.8 ± 0.4	12.0 ± 0.3
平均值	**11.600**	**12.000** ± 0.162	12.125 ± 0.137	12.250 ± 0.115	12.275 ± 0.106	12.150 ± 0.184	**12.000** ± 0.177	12.175 ± 0.164
				Fitting				
2	279	212 ± 6	199 ± 8	214 ± 4	217 ± 5	200 ± 9	**198** ± 5	204 ± 5
3	279	204 ± 3	201 ± 1	207 ± 4	202 ± 3	**199** ± 3	201 ± 4	204 ± 6
4	279	237 ± 5	234 ± 3	245 ± 4	238 ± 3	227 ± 5	**220** ± 6	223 ± 5
5	279	227 ± 4	235 ± 4	233 ± 4	223 ± 3	225 ± 4	**217** ± 6	248 ± 41
平均值	279.00	220.00 ± 2.32	217.25 ± 2.37	224.75 ± 2.00	220.00 ± 1.80	212.75 ± 2.86	**209.00** ± 2.66	219.75 ± 10.51

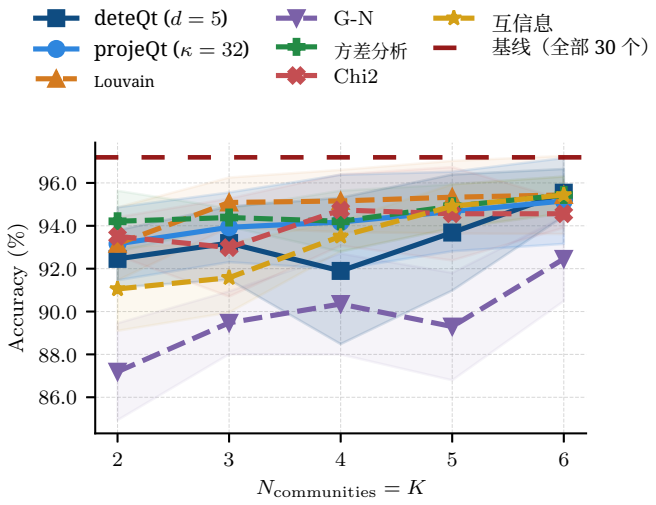

图 8：乳腺癌数据集上的分类准确率随社区数量的变化。对于无法先验指定社区数量的情况，我们仍通过随机划分来达到所需数量。

含 569 个样本和 30 个连续特征，这些特征根据细针穿刺物的数字化图像计算得出，并带有二元诊断标签。这 30 个特征由 10 个基础测量值（半径、纹理等）组成，每个测量值分别以均值、标准误和最差值的形式报告，从而在相同测量类型的特征之间形成自然的相关结构。特征图使用全部 30 个特征构建，并采用加权边和统计推导阈值（第 IIIA节）。与之前一样，我们将 deteQt、projeQt、Louvain 和 Girvan–Newman算法用于社区检测，并与方差分析、卡方检验和互信息过滤基线方法进行比较，同时在选定特征上训练随机森林分类器（500 个估计器）。

同样，我们发现 deteQt 和 projeQt 在识别有用特征方面表现不俗。与糖尿病数据集一样，我们将这些方法与基线模型进行比较；基线模型将全部 30 个可用特征输入分类器，而社区检测方法的表现与该基线相差不到 3%。

D. MNIST 上的特征选择

为了在更大的特征图上测试处理流程，我们将其应用于 MNIST [41] 图像数据集。我们将 784 个像素视为特征，并构建相关图。移除零方差像素后，剩下 719 个特征，这在实际意义上相当于移除了所有图像中常见的透明背景。如此规模的全连接图包含 258,121 条边；使用皮尔逊相关矩阵（在所有 $70k$ 数据样本上计算）进行剪枝后，得到包含 162,474 条边的稠密图，占完整图的 63%。我们使用统计推导阈值$r_{\text{crit}} = 0.0074$ 来进行此操作（第 IIIA节）。

该图中的社区对应于空间相关像素的组。与上文一样，我们为每个社区选取一个代表性像素。随后，针对每种社区检测方法，我们使用选定的像素训练一个随机森林分类器（100 个估计器）。图 9 展示了 projeQt、deteQt（探测态二分、QSVT 多项式 $d = 5$）、Louvain 和随机选择特征的准确率。结果表明，在特征数量较少时，我们的方法与 Louvain 和随机选择的表现相当，有时略逊一筹。这是因为结果依赖于随机种子，从误差条（表示一个标准差）中可以看出这一点。然而，当节点数超过 ~ 50 时，我们的两种方法都会趋于稳定，并以几个百分点的优势胜过 Louvain。

然而，分析 MNIST 上的方法并与前两个数据集进行比较后，可以发现一个有趣的特点。当被移除的

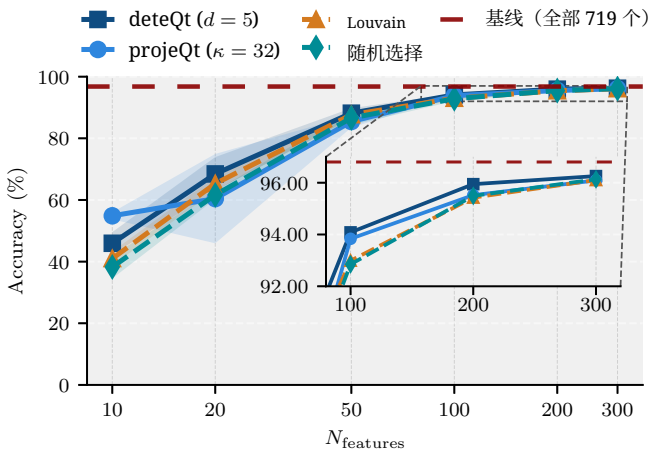

图9：MNIST 上的分类准确率与选定特征数量的关系（70,000 个样本的相关图，719 个节点）。显示五个随机种子的均值和一个标准差。我们再次将其与包含全部 719 个像素的完整特征数据集进行比较。通过 $N_{\text{features}} \approx 200$，我们恢复到了距这一最高性能 2% 以内的水平。

冗余是集中而非分散的（第 IV B 节和 IVC节中 17 个和 30 个节点的图似乎就是这种情况），社区结构可能会更有帮助，但这里并非如此。因此，MNIST 实验凸显了处理流程在特征预算较大时的局限性。从 $K = 50$开始，projeQt 的谱判据便不再支持进一步划分，尚未达到 K 时，若某个组中没有高于 projeQt 割值的特征值，就会随机将其一分为二：在 $K = 50$ 时，projeQt 约三分之一的社区来自此类二分；在 $K = 300$ 时，这一比例达到三分之二。Louvain 找到约十个社区，并通过随机二分达到从 20 开始的所有测试预算。deteQt 则通过对探测态进行强制划分来继续细分；在 $K = 200$时，这类划分占其社区的三分之一到大多数。特征预算较大时，projeQt 和 Louvain 的许多社区都是从大型组中随机选出的子集，因此，将它们与随机选择进行比较，无法单独衡量社区结构的贡献。未来的工作可以将所选特征数量与社区数量解耦，并在每个社区中选择多个代表，而不是强行继续划分。也可以用基于组内多样性或预测相关性的划分取代随机二分。其他图构造方式和自适应特征预算或有助于处理分散的冗余；而受控实验则可确定社区结构、代表选择和强制划分在何种情况下有助于下游性能。请注意，这些主要是对数据分析处理流程本身的扩充，而非对底层量子社区检测算法的改动。图 10 展示了完整的 70，000 个样本 MNIST 数据集上，分类准确率与随机森林模型规模（全部 100 个估计器的树节点总数）之间的关系。在 $K = 10$ projeQt 其准确率最高，模型规模也最大。在$K = 20$ 中，deteQt 的平均准确率最高，平均模型规模最小，最多比 Louvain 小 5%；不过，模型规模差异，以及使用 20 和 50 个特征时的准确率差异，都处于不同随机种子结果的波动范围内，而

在某些预算下，projeQt 的模型比 Louvain 的更大，而在另一些预算下则更小。

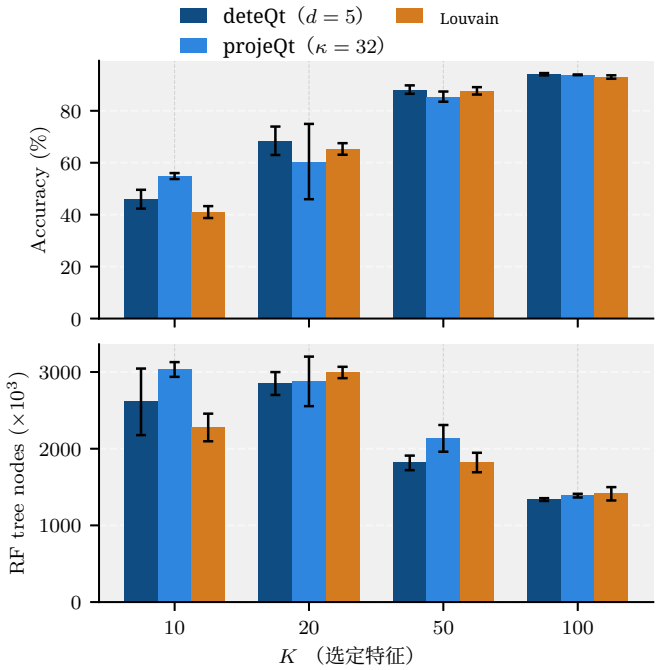

图10：完整 MNIST（70，000 个样本）上的分类准确率（左）和随机选择森林模型大小（右）随选定特征数量 K 的变化。结果展示了处理流程经过 5 次独立运行后的均值和标准差。

E. 扩展至无监督式聚类

最后，我们展示量子社区检测流程在无监督聚类问题中的应用。如上所述，这里我们将特征相关图替换为数据点相似度图。为展示这一点，我们将该流程应用于MNIST数据集的直接聚类：每个 N 图像都是一个节点，并在PCA降维空间中构建 k近邻图（$n_c = 40$ 个主成分、$k = 15$个邻居、无权边）。我们在这一问题上对deteQt、projeQt以及几种经典社区检测方法进行基准测试，方法同上。这里，社区数量是不同图像类别的数量，即$K = 10$。目标是识别 $K = 10$ 个簇，使每个簇 i 包含代表同一数字的图像。在本例中，如上所述，贪婪模块度和Louvain会自行选择社区数量，但这里我们不会随机划分，直到

选定指定数量的社区。谱Newman算法可以接受指定的社区数量作为输入，但如果无法进一步获得模块度增益，则可能无法生成所需数量的社区。对于这些方法，我们直接按照其标准实现，根据输出划分对其进行评分。对于Kernighan–Lin算法，通过依次二分其最大的群组，生成10个类别。

这里的图构建方式不同于特征选择情形。生成的 k 近邻图是稀疏且无权的（N 个节点，生成了 $\sim 7.5N - 15N$条边），而特征相关图则是稠密且带权的。在稀疏图中，谱结构较弱，因此与第 IV B 节至 IVD节中稠密相关图上的情况相比，谱方法与贪婪方法之间的区别不那么明显。

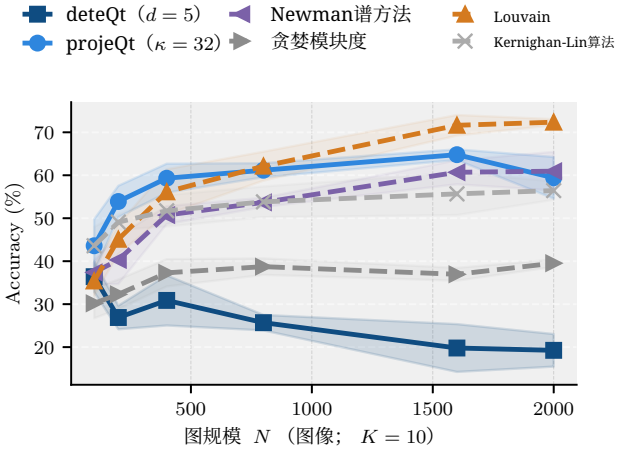

图 11：MNIST 上的聚类准确率随图规模的变化 N（k近邻相似度图）。结果展示了每种受测方法经过 5 次独立运行后的平均表现。

图 11显示了聚类准确率随图规模 N变化的情况，最多可处理 2000 个图节点（可通过 $\log_2(N) \approx 10$ 量子比特运行），结果取 5 个随机种子的平均值。随着 N规模增大，deteQt 的表现变差，这可能是因为它被限制在最简单的情形，即 $k_{\mathrm{LCU}} = 1$。deteQt 根据逐渐变得更加稀疏的子图中的一个特征向量决定每次划分，并且随着 N 规模增大，需要进行更多次划分。在这些问题上，Louvain 的表现优于我们的方法，除非我们在 projeQt 和 deteQt 输出的社区选择结果上增加类似 Louvain 的迭代过程。这可能表明，特征选择是更有前景的问题类别，或者聚类问题需要更加密集的数据点。

V. 结论

在本研究中，我们探讨了量子谱社区检测作为数据分析工具的应用，将其用于特征选择和聚类。我们通过递归二分，将 deteQt 从单个二划分扩展到任意数量的社区，并纳入了广义模块度矩阵。我们还提出了 projeQt，它将量子谱滤波器应用于模块度矩阵。该方法能够同时保留多个主特征向量，并在单个步骤中将一个图划分为多个社区，而无需制备激发态。

在合成数据集和真实数据集上，这两种方法均可与成熟的经典社区检测方法相竞争。projeQt 在信号强度适中时，比 Louvain 更准确地恢复植入社区。它还能将冗余特征正确归入其来源特征所在的组，并且在特征预算较小的 MNIST 数据集上以及中等规模的聚类图上表现良好。deteQt 在特征选择方面具有竞争力，但在合成图和大型聚类问题上的表现较弱。两种方法都未能始终优于冗余感知特征筛选器。在最大的图上，Louvain 的贪婪优化目前仍比谱准则提取出更多结构。所有结果均通过经典模拟获得，因此我们不声称实现了量子加速。相反，这些结果表明，这些量子子程序的输出可作为数据分析流程的有效输入。

其余一些限制存在于处理流程中，而非 Quantum 算法本身。未来的工作可以将选定特征的数量与社区数量解耦，并以更具信息性的标准取代强制随机切分。进一步的方向包括为 deteQt 制定统计上稳健的接受检验、以有原则的方式选择 projeQt 的谱切分，以及构造更高效的电路和读出。这些改进将同样惠及经典和量子谱检测器。

声明

a. 资金支持。 作者感谢 Fujitsu Research of Europe 和 Fujitsu Limited 提供的支持。

b. 竞争利益。 作者声明不存在竞争利益。

c. 数据可用性。 所有数据集均为公开基准数据集：皮马印第安人糖尿病数据集和来自 UCI机器学习资料库的威斯康星乳腺癌数据集，以及 MNIST。

d. 代码可用性。本文结果所用的代码由 OpenQARP 编写 [83]。OpenQARP 代码库位于 https://github.com/OpenQARP/openqarp。我们还使用了 Net- workX [84]进行图生成和经典社区检测。

e. 大语言模型的使用。— 作者使用大语言模型工具（Claude，Anthropic；Codex，OpenAI）协助软件开发和代码审查，并编辑稿件文本及其LaTeX 结构。作者审阅并核实了所有输出，并对本作品的内容承担全部责任。

[1] Santo Fortunato。图中的社区检测。*Physics Reports*，486(3–5):75–174，2010。doi:10.1016/j.physrep.2009.11.002。URL https://doi.org/10.1016/j.physrep.2009.11.002。[2] Santo Fortunato 和 Darko Hric。网络中的社区检测：用户指南。*Physics Reports*，659: 1–44，2016。doi:10.1016/j.physrep.2016.09.002。URL https://doi.org/10.1016/j.physrep.2016.09.002。[3] Santo Fortunato 和 M. E. J. Newman。网络社区检测的 20 年。*Nature Physics*，18(8): 848–850，2022。doi:10.1038/s41567-022-01716-7。URL https://doi.org/10.1038/s41567-022-01716-7。[4] M. E. J. Newman 和 M. Girvan。发现并评估网络中的社区结构。 *Physical Review E*，69 2 10 1103 69 026113 ( ):026113，2004。doi: . /PhysRevE. . . URL https://doi.org/10.1103/PhysRevE.69.026113。[5] Ulrik Brandes、Daniel Delling、Marco Gaertler、Robert G¨orke、Martin Hoefer、Zoran Nikoloski 和 Dorothea Wagner。关于模块度聚类。 *IEEE Transactions on Knowledge and Data Engineering* 20 2 172 188, , ( ): – 。2008 10 1109 2007 190689 doi: . /TKDE. . . URL https: //doi.org/10.1109/TKDE.2007.190689。[6] Michelle Girvan 和 M. E. J. Newman。社会与生物网络中的社区结构。*Proceed-ings of the National Academy of Sciences*，99(12):7821–7826，2002。doi:10.1073/pnas.122653799。URL https: //doi.org/10.1073/pnas.122653799。[7] Vincent D. Blondel、Jean-Loup Guillaume、Renaud Lam-biotte 和 Etienne Lefebvre。在大型网络中快速展开社区。 *Journal of Statistical Mechanics: Theory and Experiment* 2008 10 10008, 2008 , ( ):P 。doi:10.1088/1742-5468/2008/10/P10008。URL https: //doi.org/10.1088/1742-5468/2008/10/P10008。[8] Vincent A. Traag、Ludo Waltman 和 Nees Jan van Eck。从 Louvain 到 Leiden：确保社区具有良好的连通性。 *Scientific Reports*，9:5233，2019。doi:10.1038/s41598-019-41695-z。URL https://doi.org/10.1038/s41598-019-41695-z。[9] Paul W. Holland、Kathryn Blackmond Laskey 和 Samuel Leinhardt。随机块模型：初探。*Social Networks* 5 2 109 137, , ( ): – 。1983 10 1016 0378 8733 83 90021 7 doi: . / - ( ) - 。URL https://doi.org/10.1016/0378-8733(83)90021-7。[10]Brian Karrer 和 M. E. J. Newman。网络中的随机块模型与社区结构。*Physical Review E*，83(1):016107，2011。doi:10.1103/PhysRevE.83.016107。URL https://doi.org/10.1103/PhysRevE.83.016107。[11] Tiago P. Peixoto。大型网络中的分层块结构与高分辨率模型选择。*Physical Review X*，4(1):011047，2014。doi:10.1103/PhysRevX.4.011047。URL https://doi. org/10.1103/PhysRevX.4.011047。[12] M. E. J. Newman。模块度优化与最大似然方法在社区社区检测。 *Physical Review E*，94(5):052315，2016。doi:10.1103/PhysRevE.94.052315。URL https://doi.org/10.1103/PhysRevE.94.052315。[13] M. E. J. Newman。网络中的模块度与社区结构。 *Proceedings of the National Academy of Sciences* 103 23 8577 8582, 2006 , ( ): – 。doi:10.1073/pnas.0601602103。URL https://doi.org/10.1073/pnas.0601602103。[14] M. E. J. Newman。利用矩阵的特征向量发现网络中的社区结构。*Physical Review E*，74(3):036104，2006。doi:10.1103/PhysRevE.74.036104。URL https://doi. org/10.1103/PhysRevE.74.036104。[15] V. A. Traag。更快地展开社区：加速Louvain算法。 *Physical Review E*，92(3): 032801，2015。doi:10.1103/PhysRevE.92.032801。URL https://doi.org/10.1103/PhysRevE.92.032801。[16] Santo Fortunato和Marc Barth´elemy。社区检测中的分辨率极限。*Proceedings of the National Academy of Sciences* 104 1 36 41, 2007 , ( ): – 。doi:10.1073/pnas.0605965104。URL https://doi.org/10.1073/pnas.0605965104。[17] Raj Rao Nadakuditi和M. E. J. Newman。图谱与网络中社区结构的可检测性。*Physical Review Letters*，108(18):188701，2012。doi:10.1103/PhysRevLett.108.188701。URL https: //doi.org/10.1103/PhysRevLett.108.188701。[18]Aurelien Decelle、Florent Krzakala、Cristopher Moore和Lenka Zdeborov´a。模块化网络随机块模型的渐近分析及其算法应用。*Physical Review E* 84 6 066106, , ( ): 。2011 10 1103 84 066106 doi: . /PhysRevE. . 。URL https: //doi.org/10.1103/PhysRevE.84.066106。[19] Xiao Zhang和M. E. J. Newman。网络中的多路谱社区检测。 *Physical Review E*，92 5 052808，2015 10 1103 92 052808 ( ): 。doi: . /PhysRevE. . 。URL https://doi.org/10.1103/PhysRevE.92.052808。[20] Chukwudubem Umeano、Stefano Scali和Oleksandr Kyriienko。通过确定性消除进行量子社区检测。*Physical Review A* 112 5 052422, , ( ): 。2025 10 1103 7 9 doi: . /s cm-c qy。URL https://doi.org/ 10.1103/s7cm-c9qy。[21] Andr´as Gily´en、Yuan Su、Guang Hao Low和Nathan Wiebe。量子奇异值变换及其后续：量子矩阵运算的指数级改进。在 *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*，第193–204页，2019年。doi:10.1145/3313276.3316366。URL https://doi.org/10.1145/3313276.3316366。[22] Isabelle Guyon和Andr´e Elisseeff。变量与特征选择简介。*Journal of Machine Learning Research* 3 1157 1182, 2003 , : – 。URL https: //jmlr.org/papers/v3/guyon03a.html。[23] Jundong Li、Kewei Cheng、Suhang Wang、Fred Morstat- ter、Robert P. Trevino、Jiliang Tang和Huan Liu。特

征选择：数据视角。 *ACM Computing Surveys* 50 6 94 1 94 45, 2017 10 1145 3136625 ，( ): : – : 。doi: . / 。URL https://doi.org/10.1145/3136625。[24] Cihan Kuzudisli、Burcu Bakir-Gungor、Nurten Bulut、Bahjat Qaqish和Malik Yousef。基于特征分组的特征选择方法综述。 *PeerJ*，11 10 7717 15666 :e15666，2023。doi: . /peerj. . URL https://doi.org/10.7717/peerj.15666。[25] Wael Etaiwi和Lubna Albahar。基于图的特征选择：方法与应用的综合综述。在 *Intelligent Systems and Ap-plications*，《网络与系统讲义》，第404–419页。施普林格·自然瑞士，2025年。doi:10.1007/978-3-032-07109-5 27。URL https://doi.org/10.1007/978-3-032-07109-5_27。[26]Miloˇs Savi´c、Vladimir Kurbalija、Zoran Bosni´c和Mirjana Ivanovi´c。基于特征相关网络中社区检测的特征选择。 *Computing* 101 10 1513 1538, ，( ): – 。2019 10 1007 00607 019 00705 8 doi: . /s - - - 。URL https://doi.org/10.1007/s00607-019-00705-8。[27] Mehrdad Rostami、Kamal Berahmand和Saman Forouzandeh。一种用于特征选择的新型基于社区检测的遗传算法。 *Journal of Big Data*，8 2，2021 10 1186 40537 020 00398 3 : 。doi: . /s - - - 。URL https://doi.org/10.1186/s40537-020-00398-3。[28] Duanyang Feng、Wenjian Liu、Sheng Peng、Yuankun Liu和Wanlei Zhou。CoRoFR：社区检测特征图通过鲁棒模糊粗糙集改进特征选择。 *Advanced Intelligent Systems*，7(12): e202500044，2025。doi:10.1002/aisy.202500044。URL https://doi.org/10.1002/aisy.202500044。[29] Satu Elisa Schaeffer。图聚类。*Computer Science Review*，1(1):27–64，2007。doi:10.1016/j.cosrev.2007.05.001。URL https://doi.org/10.1016/j.cosrev.2007.05.001。[30] Jacob H. Levine、Erin F. Simonds、Sean C. Bendall、Kara L. Davis、El-ad D. Amir、Michelle D. Tadmor、Oren Litvin、Harris G. Fienberg、Astraea Jager、Eli R. Zunder、Rachel Finck、Amanda L. Gedman、Ina Radtke、James R. Downing、Dana Pe'er和Garry P. Nolan。对AML进行数据驱动的表型解剖揭示了与预后相关的祖细胞样细胞。*Cell*，162(1): 184–197，2015。doi:10.1016/j.cell.2015.05.047。URL https://doi.org/10.1016/j.cell.2015.05.047。[31] Malte D. Luecken和Fabian J. Theis。单细胞RNA测序分析的当前最佳实践：教程。*Molecular Systems Biology*，15(6):e8746，2019。doi:10.15252/msb.20188746。URL https://doi.org/10.15252/msb.20188746。[32] Ulrike von Luxburg。谱聚类教程。*Statistics and Computing*，17(4):395–416，2007。doi:10.1007/s11222-007-9033-z。URL https://doi.org/10.1007/s11222-007-9033-z。[33] M. E. J. Newman。社区检测与图划分的谱方法。 *Physical Review E*，88(4): 042822，2013。doi:10.1103/PhysRevE.88.042822。URL https://doi.org/10.1103/PhysRevE.88.042822。[34] Iordanis Kerenidis和Jonas Landman。量子谱聚类。 *Physical Review A* 103 4 042415, ，( ): 。doi: . /2021 10 1103 103 042415 PhysRevA. . 。URL https: //doi.org/10.1103/PhysRevA.103.042415。[35]Iordanis Kerenidis、Jonas Landman、Alessandro Luongo和Anupam Prakash。q-means：一种量子算法，在用于无监督式机器学习的量子算法，在 *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* 2019 ，。URL https://proceedings.neurips.cc/paper/2019/hash/16026d60ff9b54410b3435b403afd226-Abstract.html。[36] Sascha M¨ucke、Raoul Heese、Sabine M¨uller、Moritz Wolter 和 Nico Piatkowski。量子计算机上的特征选择。 *Quantum Machine Intelligence* 5 1 11, ，( ): 。2023 10 1007 42484 023 00099 doi: 。/s - - -z。URL https: //doi.org/10.1007/s42484-023-00099-z。[37] Maurizio Ferrari Dacrema、Fabio Moroni、Riccardo Nembrini、Nicola Ferro、Guglielmo Faggioli 和 Paolo Cre- monesi。利用量子退火机进行排序和分类的特征选择。发表于 *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*，第 2814–2824 页，2022 年。doi:10.1145/3477495.3531755。URL https://doi.org/10.1145/3477495.3531755。[38] Christa Zoufal、Ryan V. Mishmash、Nitin Sharma、Ni- raj Kumar、Aashish Sheshadri、Amol Deshmukh、Noelle Ibrahim、Julien Gacon 和 Stefan Woerner。用于无约束黑盒二元优化的变分量子算法：特征选择应用。 *Quantum*，7 909，2023 10 22331 2023 01 26 909：。doi: 。/q- - - -。URL https://doi.org/10.22331/q-2023-01-26-909。[39] Jack W. Smith、James E. Everhart、W. C. Dickson、William C. Knowler 和 Robert S. Johannes。使用 ADAP 学习算法预测糖尿病的发病。 *Proceedings of the Annual Symposium on Computer Application in Medical Care*，第 261–265 页，1988 年。URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/。[40] W. Nick Street、William H. Wolberg 和 Olvi L. Man- gasarian。用于乳腺肿瘤诊断的核特征提取。 *Proceedings of SPIE – Biomedical Image Processing and Biomedical Visualization* 1905 861 870, ，：–。1993 10 1117 12 148698 doi: 。/。。URL https://doi.org/ 10.1117/12.148698。[41] Yann LeCun、L´eon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。 *Proceedings of the IEEE* 86 11 2278 2324, ，( ): –。1998 10 1109 5 726791 doi: 。/。。URL https://doi.org/ 10.1109/5.726791。[42] Aaron Clauset、M. E. J. Newman 和 Cristopher Moore。在超大规模网络中发现社区结构。*Physical Review E*，70(6):066111，2004 年。doi:10.1103/PhysRevE.70.066111。URL https://doi.org/10.1103/PhysRevE.70.066111。[43] B. W. Kernighan 和 S. Lin。用于图划分的高效启发式过程。 *The Bell System Technical Journal* 49 2 291 307, ，( ): –。1970 doi:10.1002/j.1538-7305.1970.tb01770.x。URL https://doi.org/10.1002/j.1538-7305.1970. tb01770.x。[44]Afonso S. Bandeira 和 Ramon van Handel。具有独立元素的随机矩阵范数的精确非渐近界。 *The Annals of Probability*，44 4 2479 2506，2016 10 1214 15 1025（）：–。doi: 。/ -AOP 。URL https://doi.org/10.1214/15-AOP1025。[45] Benjamin H. Good、Yves-Alexandre de Montjoye 和 Aaron Clauset。实际情境中的模块度最大化性能。 *Physical Review E* 81 4 046106, ，( ): 。2010 10 1103 81 046106 doi: 。/PhysRevE.。URL https: //doi.org/10.1103/PhysRevE.81.046106。

[46]Vincent Cohen-Addad、Adrian Kosowski、Fred- erik Mallmann-Trenn 和 David Saulpic。Louvain 在随机块模型中的能力。发表于 *Advances in Neural Information Processing Systems*，卷 33，页 4055–4066。Curran Associates, Inc.，2020 年。URL https://proceedings.neurips.cc/paper_files/paper/2020/file/29a6aa8af3c942a277478a90aa4cae21-Paper.pdf。[47] Roger Guimer`a、Marta Sales-Pardo 和 Lu´ıs A. Nunes Amaral。随机图和复杂网络中的波动所产生的模块度。*Physical Review E*，70(2)：025101，2004。doi:10.1103/PhysRevE.70.025101。URL https://doi.org/10.1103/PhysRevE.70.025101。[48] Hayato Ushijima-Mwesigwa、Christian F. A. Negre 和 Susan M. Mniszewski。使用量子退火在 D-Wave系统上进行图划分。发表于 *Proceedings of the Second International Workshop on Post Moores Era Supercomputing (PMES'17)*，第22–29页。ACM，2017 年。doi:10.1145/3149526.3149531。URL https://doi.org/10.1145/3149526.3149531。[49]Christian F. A. Negre、Hayato Ushijima-Mwesigwa 和 Susan M. Mniszewski。使用量子退火在 D-Wave系统上检测多个社区。*PLOS ONE*，15(2):e0227538，2020 年。doi:10.1371/journal.pone.0227538。URL https:// doi.org/10.1371/journal.pone.0227538。[50] Ehsan Zahedinejad、Daniel Crawford、Clemens Adolphs 和 Jaspreet S. Oberoi。使用量子硬件进行符号图中的多社区检测，2019 年。URL https://arxiv.org/pdf/1901.04873。[51] Marina Fern´andez-Campoamor、Corey O'Meara、Giorgio Cortiana、Vedran Peric 和 Juan Bernab´e-Moreno。使用量子退火进行电网中的社区检测，2021 年。URL https://arxiv.org/pdf/2112. 08300。[52] Joan Falc´o-Roget、Kacper Jurek、Barbara Wojtarowicz、Karol Capala 和 Katarzyna Rycerz。通过 D-Wave Advantage 量子处理单元中的递归式和层次退火，实现复杂网络中的模块度最大化与社区检测。*IEEE Trans-actions on Network Science and Engineering*，13:7394–7410，2026 年。doi:10.1109/TNSE.2026.3668505。[53] David Bucher、Daniel Porawski、Benedikt Wimmer、Jonas N¨ußlein、Corey O'Meara、Naeimeh Mohseni、Giorgio Cortiana 和 Claudia Linnhoff-Popien。评估量子优化在动态自力更生社区检测中的应用。*IEEE Transactions on Smart Grid*，16 (2):1339–1350，2025 年。doi:10.1109/TSG.2024.3483657。[54] Ruslan Shaydulin、Hayato Ushijima-Mwesigwa、Ilya Safro、Susan Mniszewski 和 Yuri Alexeev。小型量子计算机上的网络社区检测。*Advanced Quantum Technologies*，2(9):1900029，2019 年。doi:10.1002/qute.201900029。[55] Felix G. Gemeinhardt、Robert Wille 和 Manuel Wimmer。量子k社区检测：算法提议与跨架构评估。*Quantum Information Processing*，20(9):302，2021 年。doi:10.1007/s11128-021-03239-1。[56] Chris Cade、Marten Folkertsma、Ido Niesen 和 Jordi Weggemans。用于社区检测的量子算法及其实测运行时间。*Quantum Information and Computation*，24(5&6)：361–410，2024。doi:10.26421/QIC24.5-6-1。

[57]Kanae Mukai 和 Naomichi Hatano。用于社区检测的复杂网络离散时间量子行走。*Physical Review Research*，2:023378，2020。doi:10.1103/PhysRevResearch.2.023378。[58] Md Samsur Rahaman 和 Supriyo Dutta。使用Szegedy量子行走进行网络社区检测，2026。URL https://arxiv.org/pdf/2601.21152。[59] Alexander Schmidhuber 和 Alexander Zlokapa。用于社区检测的四次量子加速，2025。URL https://arxiv.org/abs/2510.08494。[60] Alberto Peruzzo、Jarrod McClean、Peter Shadbolt、Man- Hong Yung、Xiao-Qi Zhou、Peter J. Love、Al´an Aspuru- Guzik 和 Jeremy L. O'Brien。光子量子处理器上的变分特征值求解器。*Nature Communications* 5 4213, 2014 10 1038 5213，：。doi:。/ncomms。[61] Arthur G. Rattew 和 Patrick Rebentrost。量子振幅的非线性变换：指数级改进、泛化与应用，2023。URL https://arxiv.org/pdf/2309.09839。[62] Guang Hao Low 和 Isaac L. Chuang。通过量子比特化进行哈密顿量模拟。*Quantum*，3:163，2019。doi:10.22331/q-2019-07-12-163。[63]Matteo Rossi、Marcus Huber、Dagmar Bruß 和 Chiara Macchiavello。量子超图态。*New Journal of Physics*，15:113022，2013。doi:10.1088/1367-2630/15/11/113022。[64]R. Battiti。使用互信息选择监督神经网络学习中的特征。*IEEE Transactions on Neural Networks* 5 4 537 550, 1994，()：–。doi:10.1109/72.298224。[65]Hanchuan Peng、Fuhui Long 和 Chris Ding。基于最大依赖、最大相关性和最小冗余互信息准则的特征选择。*IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 8 1226 1238, 2005，()：–。doi:10.1109/TPAMI.2005.159。[66] Ron Kohavi 和 George H. John。用于特征子集选择的封装器。*Artificial Intelligence* 97 1 2 273 324, 1997，(–)：–。doi:。/S-( )-X。[67] Robert Tibshirani。通过套索进行回归收缩与选择。*Journal of the Royal Statistical Society Series B* 58 1 267 288, 1996，()：–。doi:10.1111/j.2517-6161.1996.tb02080.x。[68] Hui Zou 和 Trevor Hastie。通过弹性网络进行正则化与变量选择。*Journal of the Royal Statistical Society Series B* 67 2 301 320, 2005，()：–。doi:10.1111/j.1467-9868.2005.00503.x。[69] Leo Breiman。随机森林。*Machine Learning*，45(1): 5–32，2001。doi:10.1023/A:1010933404324。[70] Gavin Brown、Adam Pocock、Ming-Jie Zhao 和 Mikel Luj´an。条件似然最大化：信息论特征选择的统一框架。*Journal of Machine Learning Research*，13:27–66，2012。URL https://jmlr.org/papers/v13/brown12a.html。[71] Lei Yu 和 Huan Liu。通过分析相关性与冗余实现高效特征选择。*Journal of Machine Learning Research* 5 1205 1224, 2004：–。URL https://jmlr.org/papers/v5/yu04a.html。[72] Xiaofei He、Deng Cai 和 Partha Niyogi。用于特征选择的拉普拉斯得分。载于 *Advances in Neu-ral Information Processing Systems 18 (NeurIPS)*，第507–514页，2005。URL https://papers.nips.cc/paper/2909-laplacian-score-for-feature-selection。

[73]Zheng Zhao 和 Huan Liu。用于监督式与无监督式学习的谱特征选择。载于*Proceedings of the 24th International Conference on Machine Learning (ICML)*，1151–1157，2007，第–页。doi:10.1145/1273496.1273641。[74]Peter Langfelder 和 Steve Horvath。WGCNA：用于加权相关性网络分析的 R 软件包。*BMC Bioinformatics*，9:559，2008。doi:10.1186/1471-2105-9-559。[75] Bin Zhang 和 Steve Horvath。加权基因共表达网络分析的通用框架。*Statistical Applications in Genetics and Molecular Biology*，4(1)：文章17，2005。doi:10.2202/1544-6115.1128。[76] Peter Langfelder、Bin Zhang 和 Steve Horvath。从层次聚类树中定义簇：用于 R 的 Dynamic Tree Cut 软件包。*Bioinformatics* 24 5 719 720, , （）：–。2008 10 1093 563 doi:。/bioinformatics/btm。[77] Peter Langfelder、Paul S. Mischel 和 Steve Horvath。何时枢纽基因选择优于标准荟萃分析？*PLoS ONE*，8(4):e61505，2013。doi:10.1371/journal.pone.0061505。[78] Mel MacMahon 和 Diego Garlaschelli。相关性矩阵的社区检测。*Physical Review X*，5(2):021006，2015。doi:10.1103/PhysRevX.5.021006。[79] Iordanis Kerenidis 和 Anupam Prakash。量子推荐系统。载于*8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* 67 *Leibniz International Proceedings in Informatics (LIPIcs)* 49 1 49 21，2017，卷号，页码：–：。doi:10.4230/LIPIcs.ITCS.2017.49。[80] B. David Clader、Alexander M. Dalzell、Nikitas Stamatopoulos、Grant Salton、Mario Berta 和 William J. Zeng。对经典数据矩阵进行块编码所需的量子资源。*Engineering* 3 2022 10 1109 2022 3231194 *IEEE Transactions on Quantum*，，doi:。/TQE。。[81] Danial Motlagh 和 Nathan Wiebe。广义量子信号处理。*PRX Quantum*，5:020368，2024。doi:10.1103/PRXQuantum.5.020368。[82] Christoph S¨underhauf。广义量子奇异值变换，2023。URL https://arxiv.org/pdf/2312.00723。[83] Stefano Scali、Vicente P. Soloviev、Antonio M´arquez Romero、Brian Coyle、Giuseppe Buonaiuto、Annie Paine、Jonathan H. Fetherolf、Marcos Diez Garc´ía、Michal Krompiec 和 Josh Kirsopp。Openqarp：用于量子应用研究的模块化框架，2026。URL https://arxiv.org/abs/2609.15697。[84] Aric A. Hagberg、Daniel A. Schult 和 Pieter J. Swart。使用 NetworkX 探索网络结构、动态特性和功能。见*Proceedings of the 7th Python in Science Conference* 11 15，2008，，页码 – 。网址 https: //www.osti.gov/servlets/purl/960616。

附录 A：消融研究

1. 阈值对图构建的影响

首先，图12 比较了边权重阈值机制（第 IIIA节）在 Diabetes 数据集上的两种设置。两次运行使用相同的近似符号函数（$d = 5$）和探测态二分。我们将固定值 0。95 与统计推导出的阈值进行比较，r_{crit}，在此实例中为 $r_{\text{crit}} = 0.071$，保留了 136 条边中的 114 条。相比之下，固定阈值仅保留 15 条边。由于阈值递增规则，较高的固定值 0。95 可确保失败的划分不会在更高阈值下重试，而使用 r_{crit} 时则允许重试。当 $K = 2$ 时，两者在此情况下表现相当。然而，当使用 $K \geq 3$时，$\tau_0 = r_{\text{crit}}$ 表现更佳。除非另有说明，本文所有实验均使用临界阈值。

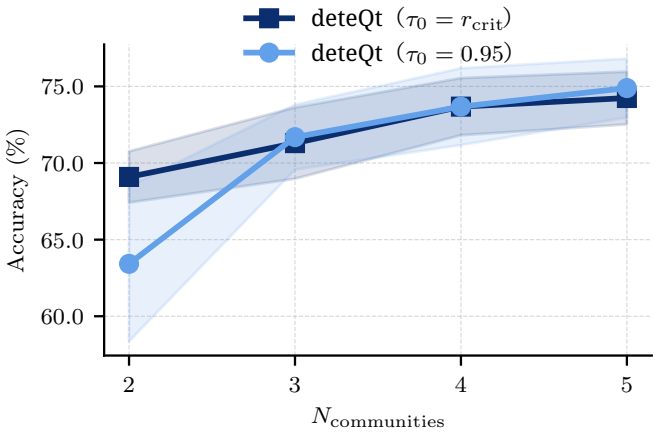

图 12：阈值处理对糖尿病数据集的影响。使用 $\tau_0 = r_{\text{crit}}$ 时，被拒绝的划分会在更高阈值下重试。使用 $\tau_0 = 0.95$ 时，被拒绝的划分不会在更高阈值下重试；强制完成仍可能根据其探针证据对组进行划分。

2. deteQt 中的低度量子奇异值变换

接下来，图 13 比较了 deteQt 在乳腺癌数据集上的分类准确率，考察了两种情形。第一种情形是，使用度为 d 的 QSVT 多项式对其振幅进行舍入；第二种情形是，将其舍入为精确符号。‘原始’特征向量等价于经过重新缩放的度为 1 的奇多项式（或者，直接使用基态制备电路的输出并绕过符号细化）。我们观察到，随着 d 增加，通过 QSVT 得到的近似符号逐渐接近精确符号舍入的准确率，而后者又优于直接使用原始符号。到 $d = 15$ 时，近似结果趋于饱和，因此该问题只需使用中等深度的电路。

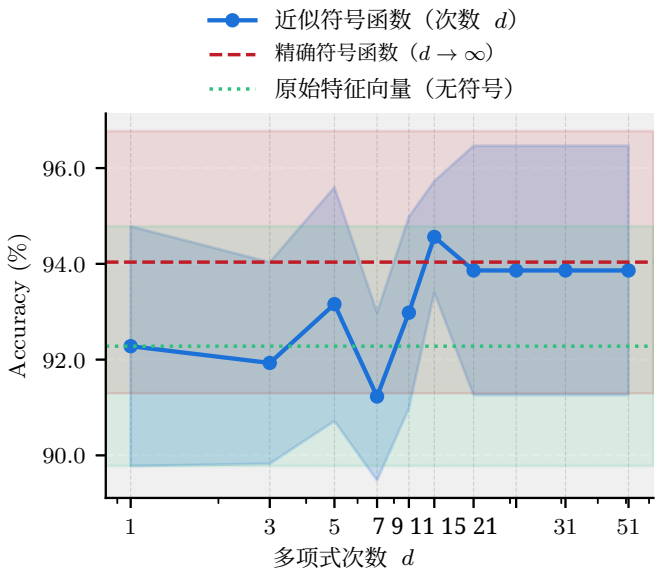

图 13：deteQt 中原始特征向量、次数为 d 的量子奇异值变换符号多项式和精确符号的乳腺癌分类准确率，在 K =4 时的结果。对于不超过 $d = 21$ 的度，我们通过直接仿真电路进行评估；对于更高的度，则改用对有效多项式作用的经典仿真。阴影表示五个随机种子所得结果的一个标准差。

附录B：基线方法

本文比较了几种经典的和标准的社区检测与特征选择方法，下面将对此作简要介绍。

1. 经典社区检测基线

*a. Girvan–Newman*算法。 Girvan–Newman算法 [6] 反复移除具有最高边 介数的边，将图划分为社区，并选择介数最大的边 $b(e) = \sum_{s<t} \sigma_{st}(e)/\sigma_{st}$，其中 σ_{st} 统计从 s 到 t的最短路径 $\sigma_{st}(e)$ 统计经过 e的路径。对于特征选择比较，我们取在 $K-1$ 次划分后的划分。

b. 贪婪模块度。Clauset–Newman– Moore方法 [42] 从每个节点对应一个社区开始，并反复合并能够带来最大模块度增量的社区对 $\Delta Q_{ab} = Q(\mathcal{P}_{a\cup b}) - Q(\mathcal{P})$，其中$Q(\mathcal{P}) = \sum_{C\in\mathcal{P}}[l_C/m - (d_C/2m)^2]$， l_C 是内部边权重， d_C 是 C中的度之和。该方法自行选择社区数量，并根据该划分在聚类结果中进行评分。

*c. Kernighan–Lin*算法。Kernighan–Lin算法 [43]通过在两个部分之间移动节点来细化二分划分，从而减少割 $C(A, B) = \sum_{i\in A, j\in B} w_{ij}$。作为独立的基线方法，它会反复对当前最大的组进行二分，直到得到 K 个组。我们将 Kernighan–Lin算法用作 projeQt 中对所提出的二分方案进行细化的一个步骤。

*d. Newman*谱方法。Newman方法 [14] 递归地使用 $s_i = \text{sign}(v_{\max,i})$对组进行二分，其中 $B^{(g)}v_{\max} = \lambda_{\max}v_{\max}$，其中 $v_{\max}$ 是主特征向量，然后应用顶点移动细化。所提议划分的模块度增益为 $\Delta Q = s^{\mathsf{T}}B^{(g)}s/(4m)$。在固定-$K$ 比较中，只要存在可接受的谱划分，就会继续划分。如果无法达到 K，则使用带种子的随机二分。如果产生的组数超过K ，则合并最小的组。在聚类比较中，如果不再接受进一步划分，也可以在原生划分处停止。

2. 特征选择基线

a. 方差分析 F-检验。单因素方差分析 F 统计量根据组间变异与组内变异之比对特征进行排序。它会独立评估每个特征，其中 $F_j = [SS_{B,j}/(C-1)]/[SS_{W,j}/(n-C)]$, 其中$SS_{B,\ j}$和$SS_{W,\ j}$ 分别是特征 j的组间平方和与组内平方和。

b. 卡方检验。卡方得分使用 $\chi^2_j = \sum_{b,c}(O_{bc}-E_{bc})^2/E_{bc}$ 衡量某个特征的分箱值与类别标签之间的依赖关系，其中包括观测与期望分箱标签计数。根据该得分对特征进行排序。按照检验的要求，输入值会被平移为非负数。

c. 互信息。互信息根据每个特征与目标变量之间的估计依赖关系对特征进行排序，$I(X_j;\ Y) = \sum_{x,y} p(x,y)\log[p(x,y)/(p(x)p(y))]$, 使用 scikit-learn 中的最近邻估计器。

d. mRMR。最小冗余最大相关性方法首先选择 $I(X_j;Y)$ 最大的特征，然后添加满足 $j^* = \arg\max_{j\notin S}[I(X_j;\ Y) - |S|^{-1}\sum_{s\in S} I\ X_j\ X_s\ ]$, S [65] (;) 的特征，其中为已选特征集合。

附录 C：统一流程

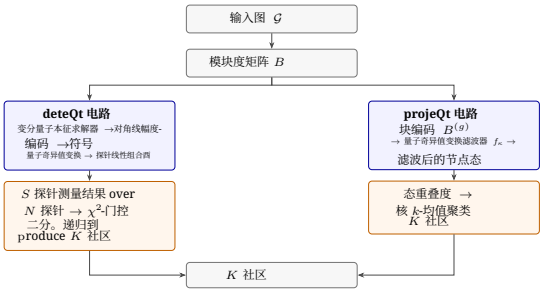

图 14：用于数据分析的社区检测统一流程。蓝色方框表示电路阶段，橙色方框表示经典后处理。deteQt 测量探针测量结果并选择一次二分，然后递归执行，直到找到指定数量的社区。projeQt 测量态重叠，并应用核 k-均值聚类，直接输出指定数量的社区；也可以选择递归执行。

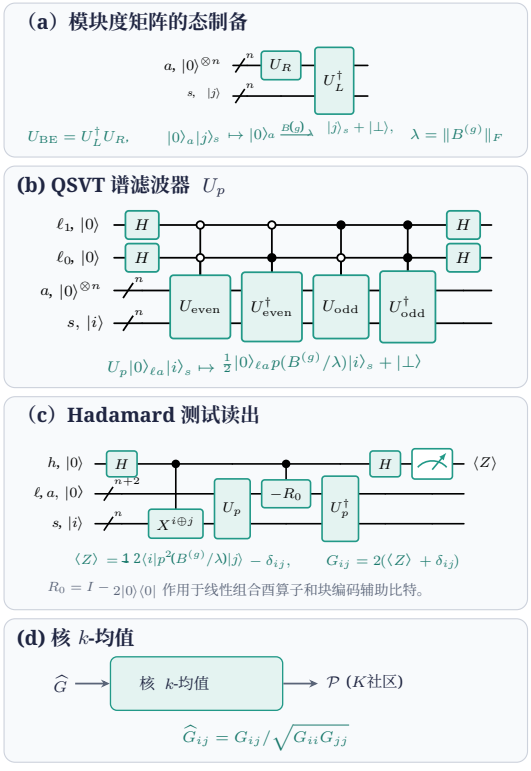

图 15：**projeQt 算法中的电路组件**。**a**）用于制备模块度矩阵的行和列的电路。**b**）通过线性组合酉算子实现奇、偶宇称的 QSVT 谱阈值滤波器。**c**）格拉姆矩阵元素 G_{ij} 的 Hadamard 测试读出。**d**）对格拉姆矩阵 G 应用核 k-均值聚类，以识别 K 社区。

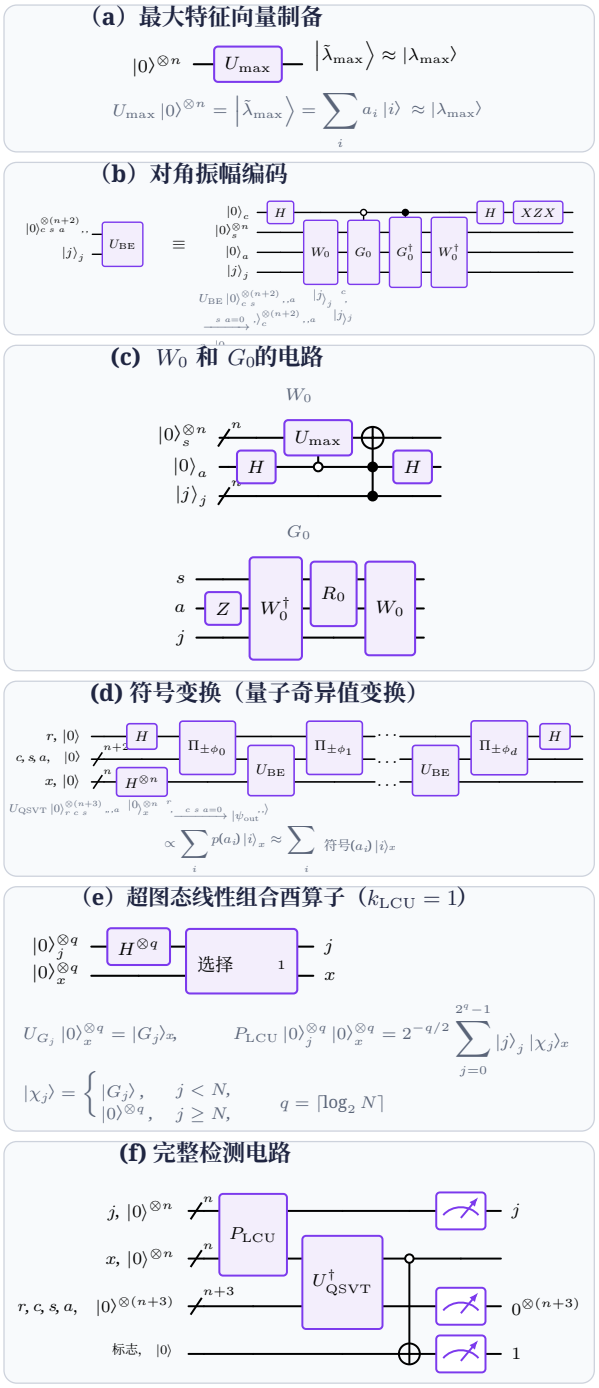

图 16：**deteQt 算法的电路组件** [20]，**限定于** $k_{\mathrm{LCU}}=1$。**a**）主特征向量的制备 $|\lambda_{\max}\rangle$，例如通过 VQE。**b**）块编码，将 $|\lambda_{\max}\rangle$ 的振幅置于对角线上。**c**）子电路 W_0 和 G_0 用于该块编码（详见参考文献 [20, 61] ）。**d**）对符号函数进行 QSVT 近似，并将其应用于 $|\lambda_{\max}\rangle$的振幅。**e**）超图 $k_{\mathrm{LCU}}=1$ 的 LCU 制备（汉明重量为 1 ）。**f**）计算带符号特征向量与候选探针态之间重叠度的完整检测电路。

算法 1：从输入图中进行projeQt社区检索

要求： 图 $\mathcal{G}$ 和计数 $2 \le K \le k^* \le |V(\mathcal{G})|$
确保： 将 $V(\mathcal{G})$ k^* 划分为 $\mathcal{P}$ 个社区

$N_0 \leftarrow |V(\mathcal{G})|$, $n_{\min} \leftarrow \max(2, \lfloor N_0/(2k^*) \rfloor)$
$\mathcal{P} \leftarrow \{V(\mathcal{G})\}$
$\mathcal{P}' \leftarrow \textsc{projeQt}(\mathcal{G}, V(\mathcal{G})\ K, n_{\min})$ ▷请求 K 社区
如果 $\mathcal{P}'$ 被接受 **那么**
  $\mathcal{P} \leftarrow \mathcal{P}'$
当 $|\mathcal{P}| < k^*$ **时执行**
  已接受 ←假
  对于 按大小递减的社区，其中 $g \in \mathcal{P}$，且 $|g| > \max(3, n_{\min})$ **执行**
    $R \leftarrow k^* - |\mathcal{P}| + 1$
    如果 $R \ge 3$ **那么**
      $K_g \leftarrow$ 最小值$(R,\ 5)$
    else
      $K_g \leftarrow 2$
    $\mathcal{P}' \leftarrow \textsc{projeQt}$（$\mathcal{G}$、 g、 K_g、 $n_{\min}$）
    如果 $\mathcal{P}'$ 被接受**则**
      $\mathcal{P} \leftarrow$ （$\mathcal{P} \setminus \{g\}$） $\cup \mathcal{P}'$
      已接受 ← 为真
      中断
  如果未接受**则**
    对每个 g，使用 $\widehat{G}$ 上的双向核 k-均值和 $\lambda_{\max}(B^{(g)}) > c$构造一个二分
    如果至少存在一个提议**则**
      选择最高的ΔQ 提议，即使其为负数
    else
      随机选择对最大的社区进行二分
    用此二分替换选定的社区
返回 $\mathcal{P}$

算法 2：针对一个社区的 projeQt 拆分提议

过程 $\textsc{projeQt}(\mathcal{G}, g, K_g, n_{\min})$
需要： 图 $\mathcal{G}$、社区 g、请求数量 K_g、最小规模 $n_{\min}$
确保： g的接受的划分，或失败
  构造 $B^{(g)}$，如公式（6）所示
  使用 VQE（变分量子本征求解器）估计 $\lambda_{\max}(B^{(g)})$，即在 $-B^{(g)}$ 的基态下，如 deteQt 中所示
  $\lambda \leftarrow \|B^{(g)}\|_F$， $e \leftarrow 2\lambda/\sqrt{|g|}$
  $c \leftarrow$ 最大值（e， $\lambda_{\max}$（$B^{(g)}$）/2），如公式（21）
  选择度为d 的多项式 p_d 近似 f，如公式（21）
  如果 $\lambda_{\max}$（$B^{(g)}$） $> 10^{-12}$ 和 $|g| > 2$，**则**
    根据公式（ $U_{\mathrm{BE}} = U_L^{\dagger} U_R$ ）从公式（ $B^{(g)}/\lambda$ 中的行与列加载器构造 [21, 79, 80]
    对于每个节点 i在 g **执行**
      准备并归一化 $p_d(B^{(g)}/\lambda)\,|i\rangle$ 作为 $|\phi_i\rangle$
    form $\widehat{G}_{ij} = \langle \phi_i | \phi_j \rangle$
    如果 $|g| \ge 2K_g$ **则** ▷ 每个社区需要两个节点； $n_{\min}$ 将在下文强制执行
      $\mathcal{P}' \leftarrow$ 对.....进行核 k 均值聚类 $\widehat{G}$ 与 K g 社区
      如果每个社区 $C \in \mathcal{P}'$ 都有 $|C| \ge n_{\min}$ 和 ΔQ（$\mathcal{P}'$） > 0 **则**
        返回 $\mathcal{P}'$
    如果 $|g| > 2$ **则**
      $\mathcal{P}_2 \leftarrow$ 双向核 k-均值算法应用于 $\widehat{G}$
      通过 Kernighan–Lin算法进行细化
      如果两个社区满足 $n_{\min}$ 且 $\Delta Q(\mathcal{P}_2) > 0$ **则**
        返回 $\mathcal{P}_2$
  返回 失败

为完整起见，算法 1 给出了从输入图中进行社区检索的完整 projeQt 算法。它展示了一种递归实现：projeQt 首先请求 K 个社区，然后不断拆分当前社区，直到达到目标 k^*。递归是可选的；正如正文所述，初始拆分可以直接请求 k^* 个社区。电路组件如图 15所示。算法 2展示了 projeQt 量子社区检测器的一个实例，该实例通过谱阈值处理、重叠度估计和经典优化来提出社区。

附录 D：实验超参数

表 II 给出了第 IV节中每项结果背后的运行设置。

表 II：运行设置。 N_0 是整个图的大小， G 是当前组的数量。

设置	值
deteQt	
探针量子态测量-结果 S	1000
k_{LCU}	1
符号多项式	奇数，度为 5，拟合于双曲正切$(10x)$
分割门	χ^2 在 Sp_j 上对比均匀分布， $p \leq 0.05$
τ_0，步长，上限	r_{crit} （单位权重图 0），0.1，1.0
拆分顺序	先进先出
VQE 深度	$L = 8 - 32$ 层 HEA 试探态
projeQt	
尖锐度 κ	32,
多项式次数	31
割 c	$\max(2\|B^{(g)}\|_F/\sqrt{N},\ \lambda_{\max}/2)$
核 k-均值	k-均值++开始，20 次重新启动
每次拆分的部分数	K 先执行，然后如果满足条件则取 $\min(K-G+1,\ 5)$ ≥ 3，否则取 2
最小分区大小	$\max(2,\ \lfloor N_0/2K \rfloor)$ 个节点
Kernighan–Lin算法	最多 20 轮次
evaluation	
随机森林	250 棵决策树（糖尿病）,500（乳腺癌），100（MNIST）
训练集/测试集划分	80/20，未分层