arXiv:2602.14708v1 [cs.DB] 16 Feb 2026

分布式数据织体的统一数学框架：范畴超图模型

T. SHASKA 和 I. KOTSIREAS

Abstract. 当前的分布式数据织体缺乏严谨的数学基础，往往依赖临时拼凑的架构，难以应对一致性、数据血缘和规模扩展方面的挑战。我们提出一种数据织体数学框架，通过基于超图的结构 $\mathcal{F} = (D,\ M,\ G,\ T,\ P,\ A)$，统一分布式系统中的异构数据管理。数据集、元数据、变换、策略和分析均建模于分布式系统之上 $\Sigma = (N,\ C)$，并通过超图编码多元关系 $G = (V,\ E)$。范畴方法将数据集视为对象、将变换视为态射，从而支持数据集成和联邦学习等操作。超图嵌入模张量范畴，通过编织幺半群结构捕捉关系对称性，并借助与 Hurwitz 空间的几何类比，丰富代数建模。我们证明了模式匹配和动态分区等关键任务具有 NP难性，并提出谱方法和基于对称性的对齐方案，以实现可扩展的解决方案。该框架利用稀疏关联矩阵和编织作用来执行容错操作，确保 CAP 和 CAL 定理下的一致性、完备性与因果性。该框架应用于集成数据库、实时分析和转换管道的多组件架构，支持高效的向量表示，并通过亚马逊卖家场景展示了其在现实环境中的可行性。这项工作在推进理论基础的同时，也为大规模数据生态系统提供了实用工具。

1. Introduction

来自云计算、物联网（IoT）、人工智能（AI）和分布式系统的数据迅速激增，已经使传统数据管理框架不堪重负。依赖刚性模式的集中式架构难以集成异构数据集、跨分布式节点扩展、实施治理或支持实时分析。例如，每天生成数 TB 传感器数据的物联网系统需要动态编排，以统一多样化的数据源、确保安全并及时提供洞察，而传统数据库无法应对这一挑战。数据编织通过提供统一的、元数据驱动的架构来解决这些问题，从而智能地管理复杂的数据生态系统，实现无缝集成、导航和分析。本文将数据织体形式化为一种数学结构，为现代数据驱动应用的自适应、安全系统设计提供严格且可扩展的框架。我们的模型借鉴 Fuad Aleskerov 在选择理论、网络中心性以及社会经济系统优化方面的贡献，将这些概念扩展到数据管理领域

2020 *Mathematics Subject Classification.* 68P15; 18M15; 05C65。

Key words and phrases. 数据织体、超图、范畴数据管理。

其中，超图关系映射网络依赖关系，范畴态射则促进分布式环境中的决策。我们的范畴视角还受到近期相关研究的进一步推动，这些研究将计算工具和变换视为结构化学习系统中的态射 [15]，并研究学习架构中的分级表示和层次化表示 [14, 17]。

我们将数据织体定义为一个元组 $\mathcal{F} = (D, M, G, T, P, A)$，在分布式系统 $\Sigma = (N, C)$ 上运行，其组件包括：

(1) D：时间索引数据集，涵盖物联网传感器读数或金融交易等来源。(2) M：元数据，为发现和溯源提供上下文及转换历史。(3) G：超图，用于编码数据集与元数据之间的多元关系。(4) T：变换，将数据映射到不同域，以实现集成和处理。(5) P：治理策略，通过访问控制确保安全与合规性。(6) A：分析函数，通过统计或机器学习模型生成洞见。(7) Σ：由节点 N 和通信链路 C 构成的分布式系统，用于托管和处理数据。

该框架利用超图 $G = (V, E)$ 对复杂依赖关系建模，利用范畴结构 DF 统一操作，并利用模张量范畴（MTC）通过编织幺半群结构捕获关系对称性。采用结构化向量表示和组合映射，与近期用于表示学习的分级方法和神经符号方法相契合；在这些方法中，层次结构和符号约束能够提升可解释性和鲁棒性[14, 16]。

我们的贡献涵盖数据织体设计的理论与实践进展。我们将数据织体形式化为一个数学元组 $\mathcal{F}$，在超图 G中集成数据集、元数据与分析。该超图嵌入模张量范畴（MTC）中，通过编织幺半群结构捕捉复杂的关系对称性，并借助与 Hurwitz 空间的新颖几何类比，丰富其代数建模，详见第5节。近期，关于结构化表示与嵌入的范畴和几何视角也已在分级学习系统及几何约化框架的背景下兴起 [5, 17], ，进一步为我们的代数视角提供了动机。

为统一数据织体操作，我们引入范畴结构 DF，将数据集建模为对象，将变换建模为态射。这种方法为数据集成和联邦学习等操作提供了严谨的框架，确保分布式系统中的操作相干性，相关讨论见第4节。我们的研究也补充了近期将学习流水线和计算工具视为分级及神经符号架构中组合态射的工作 [10, 15]。

我们还通过证明若干关键任务具有 NP难性来应对计算挑战，包括模式匹配（定理 7）和动态分区（定理 32）。为缓解这些瓶颈，我们提出谱方法和基于对称性的对齐，为大规模数据管理提供可扩展的解决方案，相关分析见

第 6 节。针对分布式系统鲁棒性，我们确保在 CAP 定理和 CAL 定理下的一致性、完备性与因果性。通过利用超图冗余和 MTC 编织，我们设计了容错操作，以在动态分区环境中保持相干性，如第 7 节所示。

实际上，我们将该框架应用于多组件架构，集成数据库、实时分析和转换管道。该系统支持通过 $\mathbb{R}^{11}$ 中的向量表示实现可扩展操作，详见第 8 节。对低维结构化嵌入的强调，与近期在学习和符号约化中采用几何表示和分次表示的方法相一致 [5, 14]。

本文结构如下。第2节定义数据织体元组 $\mathcal{F}$ 和分布式系统 Σ，并以数学方式精确地形式化时间索引数据集 D、元数据 M和超图 G 等组件。第 3 节详细介绍核心操作，包括数据集成、元数据驱动的导航和联邦学习，并确定其计算复杂度，证明模式匹配等问题具有 NP难性。第4节介绍范畴结构 DF，将数据集建模为对象、将变换建模为态射，从而为这些操作提供统一框架。第 5节描述超图 G、其在 $\mathbb{R}^n$中的向量表示，以及它向模张量范畴的嵌入，并由此提出与 Hurwitz 空间的新颖几何类比。第 6节分析计算瓶颈，例如 NP难分区，并提出谱聚类和基于对称性的对齐等缓解策略。第 7 节考察分布式环境中的一致性、完备性和因果性，并借助 CAP 和 CAL 定理确保操作稳健。第 8 节将该框架应用于多组件架构，集成数据库、实时分析和转换管道。最后，第 9 节总结本文的贡献并展望未来方向，探讨动态幺半范畴和拓扑不变量等扩展。

2. An Introduction to Data Fabrics

来自云计算、物联网 (IoT)、人工智能 (AI) 和分布式系统的数据迅速增长，已超出传统数据管理框架的应对能力。集中式或模式刚性系统难以集成异构数据集、扩展至分布式节点、实施治理或支持实时分析。数据织体提供了一种由元数据驱动的统一架构，能够智能地管理多样化的数据资产，实现无缝集成、导航和分析。本节将数据织体形式化为一种数学结构，详细阐述其组成部分，并为后续发展奠定基础。

我们的框架通过将数据编织视为嵌入超图的 d
带有张量对称性的范畴，为数学研究提供了精确基础，使其超越经验度量，迈向公理化结构。

我们将**数据织体**定义为一个元组 $\mathcal{F} = (D、 M、 G、 T、 P、 A)$，其运行于分布式系统 $\Sigma = (N、 C)$ 之上。元组 $\mathcal{F}$ 封装了数据资产、元数据、关系、变换、策略和分析，而 Σ则对分布式基础设施进行建模。下面，我们通过数学形式化表达式探讨各个组成部分，最后给出形式化定义和总结表1 。

组件描述		示例
D	时间索引数据集	销售记录（S_1）
M	带历史的元数据	医院，记录日期
G	关系超图	$e = (\{d_1, m_1\},\ \{d_3\})$
T	跨域变换	将温度转换为警报
P	治理策略	交易者的访问控制
A	分析函数	销售趋势预测
Σ	分布式节点和链路	云数据中心

TABLE 1. 数据织体组件总结

i)数据资产，记作

$$D = \{d_i(t)\}_{i,t}, \tag{1}$$

构成数据织体的核心，表示时间索引数据集。时间 t 处的每个数据集$d_i(t)$由模式 S_i 刻画，该模式是定义其结构的一组属性-类型对（例如价格、时间戳等属性），以及域Ω_i，该域为数值型（$\Omega_i \subseteq \mathbb{R}^k$）或分类型（例如 {电子产品、服装}）。形式上，

$$d_i(t) : T \to \Omega_i,$$

其中 $T \subseteq \mathbb{R}_{\geq 0}$ 为时间域，将时间戳映射到数据点。 D 的时间索引特性支持流数据。

模式 S_i 与域Ω_i 之间的异质性使集成变得复杂。为量化这一点，我们定义一种模式距离度量。

定义 1（模式距离度量）。 该模式距离度量由下式给出：

$$\mathrm{dist}(S_i, S_j) = \sum_{a \in S_i, b \in S_j} w(a,b) \cdot (1 - sim(a,b)), \tag{2}$$

其中 $sim(a,b) \in [0,1]$ 衡量属性相似度（例如通过本体对齐），而 $w(a,b)$ 对重要性进行加权。

命题 1. 模式距离$\mathrm{dist}(S_i, S_j)$ 在模式集合上定义了一个伪度量，且$\mathrm{dist}(S_i, S_j) = 0$ 当且仅当存在双射π： $S_i \to S_j$，使得对所有 $a \in S_i$，$sim(a, \pi(a)) = 1$ 。

证明。 非负性由 $1 - \mathrm{sim}(a,b) \geq 0$ 以及 $w(a,b) \geq 0$得出。由于求和遍历所有配对，因此满足对称性，且 $w(a,b) = w(b,a)$，$\mathrm{sim}(a,b) = \mathrm{sim}(b,a)$。零值条件等价于相似度度量下的完美对齐，这意味着本体下的模式同构。□

示例 1。 考虑模式 $S_1 = \{$价格、数量$\}$ (销售)和 $S_2 = \{$成本、存货$\}$(库存)。假设 sim(价格, 成本) $= 0.9$，sim(数量, 存货) $= 0.8$，sim(价格, 存货) $=0.1$，sim(数量, 成本) $= 0.2$，且权重为 $w = 1$。距离为：

$$\mathrm{dist}(S_1, S_2) = (1-0.9)+(1-0.1)+(1-0.2)+(1-0.8) = 0.1+0.9+0.8+0.2 = 2.0.$$

这一显式计算展示了度量对属性不匹配的敏感性。此外，对于高维模式，这一距离可以推广为矩阵形式：$\mathrm{dist}(S_i, S_j) = \mathrm{tr}(W(I - S))$，其中 S为相似度

矩阵，其元素为 $S_{a,b} = \mathsf{sim}(a, b)$，其中 $a \in S_i$、$b \in S_j$，而 W 为权重矩阵，其元素为 $W_{a,b} = w(a, b)$。

引理 1. 模式距离 $\mathrm{dist}(S_i, S_j)$ 的上界为 $0 \le \mathrm{dist}(S_i, S_j) \le |S_i| \cdot |S_j| \cdot \max w(a, b)$；当所有成对的 $sim(a, b) = 0$ 时，达到该上界。

证明。由于 $1 - \mathsf{sim}(a, b) \le 1$ 且 $w(a, b) \le \max w$，因此总和至多为 $|S_i| \cdot |S_j| \cdot \max w$。根据命题中的结论，下界由非负性得出。□

ii) **元数据**，$M = \{m_1, \ldots, m_k\}$，为数据资产提供上下文。每个 $m_j = (d_i, \alpha_j, \tau_j)$ 都将一个数据集 $d_i \in D$ 与属性 $\alpha_j \subseteq \mathcal{A}$（例如 {来源、格式}）及变换历史 τ_j 关联起来：$D \to \mathcal{H}$，其中 $\mathcal{H}$ 是由元组 $(t_k, t_{\mathrm{apply}})$ 构成的集合范畴，$t_k \in T$ 表示一个变换，t_{apply} 表示其时间戳。形式上，τ_j是一个将每个数据集映射到其历史记录集合的函数，满足 $\tau_j(d_i) \subseteq T \times \mathbb{R}_{\ge 0}$，从而确保只记录来自 T 的有效变换。

对 M 进行动态更新，尤其是流式处理 $d_i(t)$ 时，其复杂度与 $|M| \cdot |T|$ 成正比。为将其形式化，可将更新操作视为映射 u：$M \times D \to M$；对于 $d_i(t + \Delta t)$ 中的每个新数据点，历史记录 $\tau_j(d_i)$ 都会通过追加相关变换来扩展，这需要扫描 T 以识别适用的 t_k，由此得到上述复杂度。

引理 2。元数据更新函数 u 保持数据织体的一致性条件，确保每次更新后均满足 $\tau_j(d_i) \subseteq T \times \mathbb{R}_{\ge 0}$。

证明。每次更新都会追加元组 $(t_k, t_{\mathrm{apply}})$，其中 $t_k \in T$，从而通过构造保持子集关系。顺序更新下，该操作满足结合律，与变换的半群结构相一致。□

iii) **超图** $G = (V, E)$ 对数据与元数据之间的关系进行建模，其顶点为 $V = D \cup M$，超边为 $E \subseteq \mathcal{P}(V)$。与简单图不同，超边可以连接多个顶点，从而捕捉复杂依赖关系。

定义。2. 形式化地，G 是一个有向超图，其中每条超边 $e = (T_e, H_e)$ 都有一个尾部 $T_e \subseteq V$（输入）和头部 $H_e \subseteq V$（输出）。邻接集为：

$$\mathrm{Adj}(v) = \{e \in E \mid v \in T_e \cup H_e\}.$$

此外，该超图还可以通过其关联矩阵 $I \in \{0, 1\}^{|V| \times |E|}$来刻画，其中 $I_{v,e} = 1$ 当且仅当 $v \in e$，从而提供一种线性代数表示。

在分布式数据管理的超图计算上下文中，[19],这种结构与高阶网络模型相契合，其有向特性体现了因果流。

引理 3。若对于每一对顶点 $v, w \in V$，都存在一系列超边将它们连接起来，则超图 G 是弱连通的，从而满足数据织体定义中的连通性条件。

证明。弱连通性源于无向路径的存在，而无须考虑方向性。给定 E 上尾部与头部的并集，并假设底层图（将超边收缩为团）是连通的，则引理成立。在分配性条件下，这一性质仍然成立，因为 Σ 中的分区会映射为子超图。□

iv) 变换 $T = \{t_1, \ldots, t_m\}$ 是函数 $t_i : \Omega_i \to \Omega_j$，用于在各域之间映射数据，以实现集成、归一化或聚合。每个变换都满足：

$$t_i(d_i(t)) \in \Omega_j, \quad \text{loss}(t_i, d_i) = I(d_i; d_i) - I(t_i(d_i); d_i) \leq \epsilon, \tag{3}$$

其中，I 是互信息，$\epsilon > 0$ 界定损失。互信息$I(X;Y) = H(X) - H(X|Y)$，其熵为 H，满足链式法则和次可加性，从而确保复合变换 $t_2 \circ t_1$ 的累积损失有界：

$$\text{loss}(t_2 \circ t_1, d_i) \leq \text{loss}(t_1, d_i) + \text{loss}(t_2, t_1(d_i)).$$

成本 $\text{cost}(t_i)$各不相同（例如，线性映射为 $O(n)$，复杂映射为 $O(n^2)$）。优化需要兼顾效率和保真度：

$$\min_{t_i \in T} \text{cost}(t_i) + \lambda \text{loss}(t_i, d_i),$$

其中 $\lambda > 0$。如果损失是凸的，这就是一个凸优化问题，可通过梯度下降求解，需进行 $O(|T|)$次迭代，每次迭代的成本为 $O(n)$。

引理4。损失函数 $\text{loss}(t_i, d_i)$ 非负，且满足数据处理不等式，从而确保任何变换都不会增加互信息。

证明。根据互信息的定义，$I(t_i(d_i); d_i) \leq I(d_i; d_i)$，因为处理过程不会增加信息。非负性显然成立，而该不等式则由马尔可夫链性质得出。□

v) 治理策略 $P = \{p_1, \ldots, p_l\}$ 强制实施安全与合规性，其中每个 $p_i = (c_i, a_i)$ 都包含一个谓词 $c_i : D \times \mathcal{U} \to \{0, 1\}$（例如，用户许可级别）和一个操作 a_i（例如，授予访问权限）。用户上下文空间 $\mathcal{U}$ 包含角色或时间戳。

请求 $r(d_i, u)$ 在满足以下条件时会获准：

$$r(d_i, u) = 1 \iff \bigwedge_{p_j = (c_j, a_j) \in P} c_j(d_i, u).$$

这一合取在谓词上构成一个布尔格，确保访问决策具有单调性。评估复杂度为 $O(|P| \cdot |N|)$。

定理3。请求 $r(d_i, u)$ 的策略评估复杂度为 $O(|P| \cdot |N|)$。

证明。要评估 $r(d_i, u)$，需在节点 N上计算合取 $\bigwedge_{p_j = (c_j, a_j) \in P} c_j(d_i, u)$。每个谓词 c_j 需要 O次操作，假设上下文查找（例如，检查用户角色或数据属性）耗时为常数。对于 $|P|$ 条策略，在单个节点上评估所有谓词需要 $O(|P|)$。在分布式系统中，由于数据或策略可能分布在各处，验证可能需要检查每个 $|N|$ 节点上的 d_i 或策略条件，因此最坏情况下的复杂度为 $O(|P| \cdot |N|)$。□

vi) 分析函数 $A = \{a_1, \ldots, a_p\}$ 用于生成洞见，其中 $a_i : D \to \mathbb{R}^k$（例如回归）或 $\mathcal{C}$（例如分类）。

每个 a_i 由 θ_i参数化，并通过以下方式进行优化：

$$\theta_i = \arg\min_{\theta} \mathcal{L}(a_i(D, \theta), y),$$

其中，$\mathcal{L}$是损失函数（例如均方误差）。假设 $\mathcal{L}$是凸且利普希茨连续的，则通过梯度下降，优化过程在 $O(1/\epsilon)$次迭代内收敛，每步为 $O(|\theta_i| \cdot |D|)$。复杂度通常为 $O(|\theta_i| \cdot |D|)$。

引理 5. 在 $\mathcal{L}$的凸性条件下，针对 θ_i 的优化在$O(1/\epsilon)$ 步内达到 ϵ-最优。

证明。 根据标准凸优化理论，步长为 $1/L$（利普希茨常数）的梯度下降会收敛为 $\mathcal{L}($ $\theta^{(k)}) - \mathcal{L}^* \leq O(1/k)$，从而得到该界。□

yii) **分布式系统** $\Sigma = (N, C)$承载数据织网，其中节点 N 存储$D_n \subseteq D$，链路为 $C \subseteq N \times N$。邻接矩阵为：

$$A_\Sigma(n_i, n_j) = \begin{cases} w(n_i, n_j) & \text{if } (n_i, n_j) \in C, \\ \infty & \text{otherwise,} \end{cases}$$

其中， $w(n_i, n_j)$表示延迟或带宽。该矩阵定义了一个加权图，其最短路径可通过弗洛伊德-沃肖尔算法在 $O(|N|^3)$内计算，或者针对稀疏 C，使用 Dijkstra 算法在$O(|N|^2 \log |N|)$内计算。

负载均衡：

$$\text{load}(n) = |D_n| + \sum_{a_i \in A_n} \text{cost}(a_i),$$

负载函数构成一个向量 $\mathbf{l} \in \mathbb{R}^{|N|}$，重新分配后达到均衡需要满足 $\|\mathbf{l}\|_\infty \leq \lambda$，并通过线性规划进行优化。

命题 2。 该分布式系统 Σ 通过由 A_Σ导出的最短路径距离诱导出一个度量空间，并满足延迟上限的三角不等式。

证明。 最短路径距离 $d(n_i, n_j)$ =取满足条件的路径上的最小值$\sum w$，满足 $d(n_i, n_k) \leq d(n_i, n_j) + d(n_j, n_k)$，因为路径的连接会产生一条有效路径。□

定义。4（数据织体）。 数据织体是一个 $\mathcal{F} = (D, M, G, T, P, A)$ 定义在$\Sigma = (N, C)$之上，其中：

(1) $D = \{d_i(t)\}_{i,t}$：带有模式 S_i 和域 Ω_i的时间索引数据集（公式（*1*））。 (*2*) $M = \{m_1, \ldots, m_k\}$：用于上下文和历史记录的元数据 $m_j = (d_i, \alpha_j, \tau_j)$（引理 *2*）。 (*3*) $G = (V, E)$：具有 $V = D \cup M$、 $E \subseteq \mathcal{P}(V)$的有向超图（定理 *2*）。 (*4*) $T = \{t_1, \ldots, t_m\}$：变换 $t_i : \Omega_i \to \Omega_j$，带有损失约束（公式（*3*））。 (*5*) $P = \{p_1, \ldots, p_l\}$：用于合规性的策略 $p_i = (c_i, a_i)$（第 *2* 节）。 (*6*) $A = \{a_1, \ldots, a_p\}$：由 θ_i 参数化的分析函数（第 *2* 节）。 (*7*) $\Sigma = (N, C)$：承载 $D_n \subseteq D$ 的分布式系统（第 *2* 节）。

条件：

i) 一致性：$\tau_j(d_i) \subseteq \{(t_k, t_{apply}) \mid t_k \in T\}$。*ii)* 连通性： G 确保从 $d_i \in D$ 到 $m_j \in M$存在路径。*iii)* 合规性： $r(d_i, u) = 1 \iff \bigwedge_{p_j \in P} c_j(d_i, u)$。*iv)* 分布性： $D = \bigcup_{n \in N} D_n$，并进行本地处理和聚合。

定理 5. 元组 $\mathcal{F}$ 在变换 T的复合运算下构成一个半群，其单位元为平凡变换 $id : \Omega_i \to \Omega_i$。

证明。 集合范畴 T在复合运算下封闭，因为 $t_2 \circ t_1 : \Omega_i \to \Omega_k$ 对 $t_1 : \Omega_i \to \Omega_j$,$t_2 : \Omega_j \to \Omega_k$成立，且函数复合满足结合律。恒等映射 id 满足单位公理。其余分量 D, M, G, P, A 在该运算下保持不变，从而确保 $\mathcal{F}$上的半群结构。□

3. Operations of Data Fabrics

基于第 $\mathcal{F} = (D, M, G, T, P, A)$节中给出的分布式系统 $\Sigma = (N, C)$上的数据编织元组的形式化定义，其组成部分构成了基础结构。本节将详述实现数据织网功能的各项操作。这些操作——数据集成、元数据驱动的导航、可扩展性与分布、治理与安全、溯源跟踪以及联邦学习——利用超图 G、变换 T和分析函数 A 来管理和分析分布式数据。这些操作的复杂性凸显了 NP难优化和延迟约束等挑战，相关内容将在第6节进一步讨论。下文将通过示意图展示操作流程，见图1。

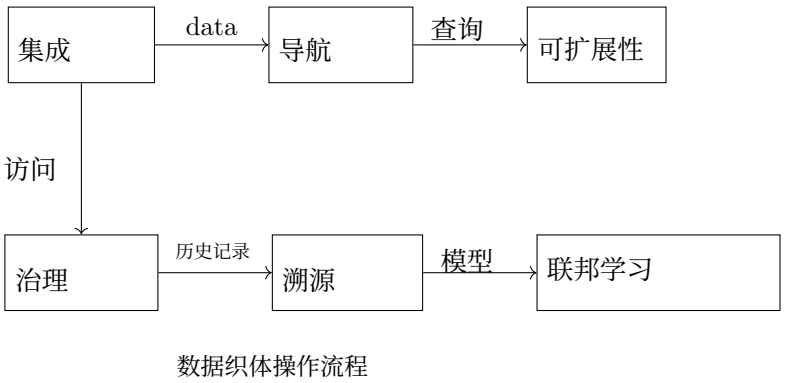


Figure 1. 各项操作之间的交互，展示了从数据集成到联邦学习的数据流。

i) **数据集成** 通过对齐模式与域，统一异构数据集，并直接基于公式（S_i, S_j）中定义的模式距离度量 dist(S_i, S_j)，该度量通过加权相似度和量化不匹配程度。

定义。6 (数据集成**)。** 数据集成 是一个映射 $\phi : D \to D' \subseteq D$，其中 D' 是一个统一数据集，通过选择变换构建 $t \in T$，使得对于每个 d_i，$d_j \in D$，$t(d_i) \in \Omega_j$，满足模式兼容性并使集成成本最小化。形式上，ϕ 被实现为第 4 节将要介绍的范畴结构中的态射复合。

形式上，对于数据集 $d_i, d_j \in D$，我们寻求一个变换 $t \in T$，使其最小化：

$$\min_{t \in T} \left(\sum_{d_i, d_j \in D} \text{dist}(S_i, t(S_j)) + \lambda \text{cost}(t) \right),$$

约束条件：

$$t(d_i) \in \Omega_j, \quad \text{compat}(S_i, t(S_j)) \geq \theta,$$

其中：

(1) $\text{dist}(S_i, S_j) = \sum_{a \in S_i, b \in S_j} w(a, b) \cdot (1 - \text{sim}(a, b))$ is the schema distance from Equation (1), satisfying the pseudo-metric properties as established there,

(2) cost(t) 是应用 t 的计算复杂度（例如，对于线性映射为 $O(n)$，对于复杂连接为 $O(n^2)$，其中 $n = |d_i|$）；(3) $\lambda > 0$ 用于平衡模式对齐与计算效率，通常通过交叉验证进行调优，以最小化总体损失；(4) compat(S_i, S_j) $\in [0, 1]$ 通过本体映射衡量语义兼容性，其定义为二分匹配下对齐属性所占的比例；(5) $\theta \in (0, 1]$ 是确保对齐具有意义的兼容性阈值，其中 $\theta = 0.5$ 通常用于部分重叠。

如果 dist 和 cost 是凸函数，则该优化问题是凸的，可使用内点法等求解器高效求解，复杂度为 $O(|T|^{3/2})$。

术语模式匹配 指寻找一个映射 $\pi : S_i \to S_j$，使属性相似度最大：

$$\max_{\pi: S_i \to S_j} \sum_{a \in S_i} \mathrm{sim}(a, \pi(a)),$$

其中 sim(a, b) $\in [0, 1]$ 表示相似度分数，通常通过字符串度量（例如莱文斯坦距离）或语义嵌入计算。该过程是集成的核心环节，但由于其 NP难性，计算上颇具挑战性，近期关于基于提示的匹配的研究也证实了这一点 [6]。

定理7。 模式匹配问题，即最大化 $\sum_{a \in S_i} sim(a, \pi(a))$，是 *NP*难的。

证明。我们通过将子图同构问题归约到模式匹配问题，证明其 NP难性。设 $G_1 = (V_1, E_1)$ 和 $G_2 = (V_2, E_2)$ 是具有 $|V_1| \le |V_2|$ 的图范畴，其中我们要寻找一个与 G_1 同构的 G_2 子图。按如下方式构造模式 S_i 和 S_j：

(1) 对每个顶点 $v \in V_1$，创建一个属性 $a \in S_i$；对每个顶点 $v' \in V_2$，创建一个属性 $b \in S_j$。(2) 定义相似度：sim(a, b) $= 1$ 当且仅当存在一个双射 $\pi : S_i \to S'_j \subseteq S_j$，使得对于所有 ($v$, w) $\in E_1$，都存在一条对应的边 ($\pi(v)$ $\pi(w)$) $\in E_2$，且 $\pi(a) = b$；否则，sim(a, b) $= 0$。

最大化 $\sum_{a \in S_i}$ sim(a, $\pi(a)$) 需要找到一个映射 π，使 S_i 与 S_j 的一个子集对齐，这等价于在 G_2 中找到一个与 G_1 同构的子图。该归约是多项式的，因为构造 S_i、S_j，且相似度函数的计算需要 $O(|V_1| + |V_2| + |E_1|)$。

由于子图同构是 NP难的，模式匹配也是 NP难的。判定版本（是否存在相似度为 $\ge k$ 的映射？）属于 NP，因为验证映射需要多项式时间。□

例2。 为明确这一约简，考虑一个子图同构问题，其中 G_1 是一个三角形，其边为 $V_1 = \{v_1, v_2, v_3\}$, $E_1 = \{(v_1, v_2)、(v_2, v_3)$ 和 $(v_3, v_1)\}$，而 G_2 的边为 $V_2 = \{u_1, u_2, u_3, u_4\}$ 和 $E_2 = \{(u_1, u_2)、(u_2, u_3)、(u_3, u_1)、(u_3, u_4)\}$。

构造模式 $S_i = \{a_1, a_2, a_3\}$，分别对应于 v_1、v_2、v_3 和 $S_j = \{b_1, b_2, b_3, b_4\}$，以及 u_1、u_2、u_3、u_4。若映射 $a_i \to b_j$ 保持 G_1 在 G_2 中的边结构，则相似度 $sim(a_i, b_j) = 1$，否则为 *0*。最大化 $\sum_{a \in S_i} sim(a, \pi(a))$ 即可找到映射π： $\{a_1, a_2, a_3\} \to \{b_1, b_2, b_3\}$ 如果 u_1、u_2、u_3 在 G_2 中构成一个三角形，这就对应于子图同构任务。

引理6。 对模式匹配进行常数因子近似是 *NP*难的。

证明。 除非 P $=$ NP，否则子图同构问题不存在常数因子近似，因为边映射的微小变化会大幅降低同构性。由于该归约在二元相似性下保留了这一性质（$\mathsf{sim}(a,b) \in\{0,1\}$），在常数因子范围内近似 $\sum_{a\in S_i}\mathsf{sim}(a,\pi(a))$ 仍然是 NP 难的。□

命题 3。 集成代价函数 $\sum\mathrm{dist}(S_i,t(S_j))+\lambda$成本$(t)$ 在 t 中是凸的，前提是 dist 和成本均为凸函数，从而可通过凸优化技术确保获得全局最优解。

证明。 和的凸性源于其关于 λ 的线性以及各项的凸性。对于离散的 T，可通过枚举获得最小值；但对于 t的连续参数化，基于梯度的方法会收敛到全局最小值。□

示例 *3*。对于亚马逊卖家网络中的实际集成，考虑将销售数据集 d_1*(t)* 与模式为 $S_1=\{$产品 *ID*、价格、数量$\}$、域 $\Omega_1=\mathbb{R}\times\mathbb{N}$ 的库存数据集 d_2*(t)* 进行集成，其中 $S_2=\{$产品 *ID*、成本、存货$\}$，$\Omega_2=\mathbb{R}\times\mathbb{N}$。变换 $t_1\in T$ 对齐属性（例如，t_1：通过缩放因子 α 将价格 $\to$ 变换为成本，其中 t_1*(*价格*)* $=\alpha\cdot$价格），生成模式为 $S_3=\{$产品 *ID*、价格、数量、存货$\}$的统一数据集 d_3*(t)*。模式距离 *dist*(S_1,S_2)（如公式（*1*）所定义）指导 t_1 的选择，但模式匹配的*NP*难性使得在大型数据集上必须采用启发式算法（第*6.1*节）。

变换复合的非交换性质，其中 $t_1\circ t_2\neq t_2\circ t_1$，映照出量子群（例如 $U_q(\mathfrak{sl}_2)$ 的代数结构，其余积定义了非交换张量积。这一类比表明，量子群作用可以对集成中的顺序敏感依赖进行建模，并有望为选择 $t_i\in T$的优化策略提供参考，详见第 6.1节。

ii) **元数据驱动导航** 通过遍历超图来解析查询G，利用定理 2 中定义的邻接结构以及引理 2中的元数据关联。

定义。**8** (元数据驱动导航)。 元数据驱动导航 通过遍历 G，重新解析查询 $q:D\to\{0,1\}$，找到最短路径 $p\in$*Paths*(v_s,v_t)，从源顶点 $v_s\in V$（例如元数据）到目标顶点 $v_t\in V$（例如数据集），并使路径代价最小：

$$\mathrm{argmin}_{p\in Paths(v_s,v_t)}\sum_{(u,v)\in p}w(u,v),$$

其中 $w(u,v)\geq 0$ 表示延迟或语义距离，通常根据公式 (1)中的模式相似度 *sim* 推导得出。

G 中的路径是超边序列 e_1，e_2，.....，其中 $H_{e_i}\cap T_{e_{i+1}}\neq\emptyset$。导航采用经过修改的迪杰斯特拉算法，并针对超图作如下调整：

(1) 用（v_s,0）初始化优先队列，并设置距离 $\mathsf{dist}[v]=\infty$，但不设置 $\mathsf{dist}[v_s]=0$。（2）对于每个顶点 u，遍历超边 $e=(T_e,H_e)\in\mathsf{Out}(u)$，若 $\mathsf{dist}[u]+w(u,v)<\mathsf{dist}[v]$，则更新 $v\in H_e$ 的 $\mathsf{dist}[v]$，其中 $w(u,v)=1-\mathsf{sim}(\alpha_u,\alpha_v)$，用于元数据属性。

(3) 对 G 使用稀疏表示，其复杂度为 $|E| = O(|V|\log |V|)$，最终得到复杂度 $O(|E| + |V|\log |V|)$；这是因为边遍历占主导，而优先队列的更新复杂度为对数级。

复杂度源于处理每条超边（$O(|E|)$）并通过优先队列（$O(|V|$对数 $|V|)$）更新距离，假设通过压缩稀疏行格式实现高效访问。这与近期的超图导航算法相一致；这些算法强调分布式数据的高阶交互 [19]。对 G进行动态更新（例如添加新数据集或元数据）会增加计算开销，从而影响可扩展性（第6.2节）。

命题 4。元数据驱动导航中的路径代价受 G 的底层图范畴的直径限制，该直径定义为最长最短路径长度；在稀疏超图中，其至多为 O（对数$|V|$）。

证明。在度受 O(对数$|V|$) 限制的稀疏超图中，根据扩张性质，直径为O(对数 $|V|$)。代价总和 $\sum w(u,\ v)$ 随之成立，因为每条边的权重均不超过 1（由 $sim \in [0,1]$ 得出）。□

例如，在亚马逊卖家网络中，查询

$$q(d_i) = (\text{category} = \text{electronics} \wedge \text{quantity} > 100)$$

从一个具有属性 $\alpha_1 =$ {类别：电子产品} 的元数据顶点 $m_1 \in M$ 开始。导航沿着超边 $e_1 = (\{m_1\}, \{d_1\})$ 到达销售数据集 $d_1(t)$，随后沿着 $e_2 = (\{d_1, m_2\}, \{d_3\})$ 到达聚合数据集 $d_3(t)$，其中 m_2 包含时间约束（例如，$\alpha_2 =$ {日期：$2025-04-01$}）。为简化起见，赋予权重 $w(u,v) = 1$，路径代价为 2，表示遍历的超边数量。在更精细的模型中，权重可以纳入来自 Σ 的邻接矩阵（第 2 节）的延迟。

iii)**可扩展性与分布**将数据资产划分到分布式系统中 Σ，确保实现高效的负载均衡，定义见第2节。

定义。**9**（可扩展性与分布）。可扩展性与分布将数据资产 $D = \bigcup_{n\in N} D_n$ 分区，其中 $D_n \subseteq D$ 位于节点 $n \in N$ 上，并将分析 $a \in A$ 计算为聚合：

$$a(D) = \bigoplus_{n\in N} a(D_n),$$

同时最小化计算成本和通信成本，其中负载函数 $load(n) = |D_n| + \sum_{a_i\in A_n}$ 成本(a_i) 来自第 2节。

形式化地说，分区问题旨在最小化：

$$\min_{\{D_n\}} \left(\sum_{n\in N} \text{cost}(a(D_n)) + \sum_{(n_i,n_j)\in C} \text{comm}(D_{n_i}, D_{n_j}) \right),$$

其中：

(1) cost($a(D_n)$) $= O(|D_n|)$ 是在 D_n 上进行分析的计算成本；(2) comm(D_{n_i}, D_{n_j}) 是通信成本，与链路 (n_i, n_j) $\in C$ 上的数据传输量成正比，通常根据 A_Σ 建模为 $|D_{n_i} \cap D_{n_j}| \cdot w(n_i,\ n_j)$；(3) $\bigoplus$是聚合算子（例如，对回归系数求和，对分类标签取并集）。

例如，在亚马逊卖家网络中，销售数据集 $d_1(t)$ 被分区到各区域服务器 $\{D_{n_1}, D_{n_2}, \dots\}$ 上，其中每个节点 $n_i \in N$ 都会计算本地销售总额 $a(D_{n_i}) = \sum_{x \in D_{n_i}} x.$ 数量。随后通过求和进行全局聚合，得出总销售额 $a(D) = \sum_{n \in N} a(D_{n_i})$。这种优化会在本地计算成本（与 $|D_{n_i}|$ 呈线性关系）和通信成本之间进行权衡，而通信成本取决于节点间的数据依赖关系。

如以下定理所示，分区问题是 NP难，这与近期关于分布式数据分配的证明一致 [1]。

定理10。 分区问题的最小化目标为

$$\sum_{n \in N} cost(a(D_n)) + \sum_{(n_i, n_j) \in C} comm(D_{n_i}, D_{n_j}),$$

是 *NP*难。

证明。 我们将已知为 NP难的图划分问题归约为数据分区问题。给定一个边权重为 $w(e)$ 的图 $G = (V, E)$，将顶点 V 映射为数据集 D，将边 E 映射为数据依赖关系，并将边权重映射为通信成本 comm。将数据集 D 分配给节点 N 以最小化节点间通信，相当于对 G 进行分区，以最小化跨分区边的权重总和，这是一个 NP难问题。该归约是多项式的，因为将顶点映射为数据集、将边映射为依赖关系所需时间为 $O(|V| + |E|)$，由此证明数据分区问题是 NP难的。□

引理7。 分区问题不允许常数因子近似，除非 $P = NP$，因为底层图划分在近似意义下是困难的。

证明。 除非 P = NP，否则一般图的图划分问题不存在常数因子近似，而该归约保留了这一难度，因为成本函数与边割相对应，且没有失真。□

这一计算复杂度将在第6.2节进一步分析，它限制了动态数据集的处理效率 $d_i(t)$，因此需要采用自适应分区策略来平衡负载并最大限度地减少通信开销，例如使用超图拉普拉斯算子进行谱聚类，正如第5节预览的那样。

iv) **治理与安全** 对数据资产和操作实施策略，将第2节中的基于谓词的定义加以扩展。

定义。11 (治理与安全)。 治理与安全 实施一组策略 $P = \{p_1, \dots, p_l\}$，其中每项策略 $p_i = (c_i, a_i)$ 由一个谓词 c_i：$D \times \mathcal{U} \to \{0, 1\}$ 和一个操作 a_i 组成，用于授予对访问请求 $r(d_i, u)$ 的访问权限，并通过差分隐私等机制确保分析过程中的隐私 $a \in A$。</*output*>

如果满足以下条件，则授予用户上下文 $u \in \mathcal{U}$（例如，角色、凭证、时间戳）对数据集 $d_i \in D$ 的请求 $r(d_i, u)$：

$$r(d_i, u) = 1 \iff \bigwedge_{p_j = (c_j, a_j) \in P} c_j(d_i, u).$$

谓词 $c_j(d_i, u)$ 用于评估用户授权或数据敏感性等条件。例如，在亚马逊卖家网络中，一名角色为“管理员”的经理 u 请求访问销售数据集 $d_1(t)$。策略集合 P 包含诸如 $c_1(d_1, u) = 1$ 若 $u.$ 角色为 $=$ 管理员之类的谓词，确保只有获得授权的用户才能访问敏感数据。

差分隐私保护分析中的各个数据点，其形式化定义如下：

$$P(a(D) \mid D) \leq e^{\epsilon} P(a(D') \mid D') + \delta,$$

对于相邻数据集 D，D'（相差一条记录），隐私参数为ϵ，$\delta > 0$。例如，在亚马逊卖家数据网络中计算 d_1（t）的平均销售额时，会使用差分隐私来确保不会泄露单笔交易详情，从而在提供汇总洞察的同时维护客户隐私。正如近期分布式治理研究所分析的那样，差分隐私下的效用损失上限为 O（$1/\epsilon$） [21]。

策略评估复杂度为：

$$O(|P| \cdot |N|),$$

因为每项策略 $p_j \in P$ 都要在所有节点 N上进行检查，假设谓词评估（例如角色或凭证查询）为常数时间。这一可扩展性挑战将在第6.3节进一步探讨。

定理 12. 对请求 $r(d_i, u)$ 进行策略评估的复杂度为 $O(|P| \cdot |N|)$。

证明。 要评估 $r(d_i,\ u)$，需在节点 N 上计算合取 $\bigwedge_{p_j=(c_j,\ a_j)\in P} c_j(d_i,\ u)$。每个谓词 c_j 需要 $O(1)$ 次操作，假设上下文查询（例如检查用户角色或数据属性）为常数时间。当有 $|P|$ 项策略时，在单个节点上评估所有谓词需要 $O(|P|)$。在分布式系统中，由于数据或策略可能分布在各处，验证时可能需要检查 $|N|$ 个节点中的每个节点上的 d_i 或策略条件，最坏情况下复杂度为 $O(|P| \cdot |N|)$。□

命题 5. 策略合取 $\bigwedge c_j$ 具有单调性：添加策略只会限制访问，从而在扩展时维护安全。

证明. 布尔函数的合取会随着项数增加而单调不增，因为每增加一个 c_j ，就会增加一个约束，使结果可能从 1 变为 0，但不会反向变化。□

这一性质确保随着 P 的增长，治理依然稳健，并与 2025年治理框架中的分布式趋势相呼应 [20]。

v)**溯源跟踪** 利用元数据 τ_j 重构变换历史，源自引理 2 以及定理 2中的超图路径。

定义. 13 (溯源跟踪)。 溯源跟踪 构建数据集 $d_i \in D$ 的轨迹，定义如下：

$$\mathrm{tr}(d_i) = \{t_j \in T \mid t_j \ \textit{was applied to produce or modify}\ d_i\},$$

使用变换历史τ_j ： $D \to \mathcal{H}$ ，该历史来自元数据 $m_j \in M$ 和超图 G。

形式上，对于元数据描述符 $m_j = (d_i, \alpha_j, \tau_j) \in M$，其溯源为：

$$\mathrm{tr}(d_i) = \{t_k \in T \mid (t_k, t_{\text{apply}}) \in \tau_j(d_i)\}.$$

超图 G 通过提供连接 d_i、其元数据 m_j及源数据集的路径来支持这一点，这些路径由超边构成。例如，在亚马逊卖家网络中，追溯销售预测数据集 $d_6(t)$ 的溯源，需要识别应用于原始销售数据集$d_1(t)$ 的转换操作，包括 $t_1 \in T$ (数据清洗)和 $t_2 \in T$ (聚合)。元数据 $m_3 \in M$ 包含 $\tau_3(d_6) = \{(t_1,\ 2025-04-01),\ (t_2,\ 2025-$

$04-02)\}$，而超图包含一条超边 $e_1 = (\{d_1, m_2\}, \{d_6\})$，其中$m_2$ 记录了额外的上下文（例如，时间范围）。

溯源跟踪的复杂度为： $O(|T| \cdot |E|)$，因为重建 $\tau_j(d_i)$ 需要遍历每个变换 $e \in E$ 的所有超边 $t_k \in T$。对 T 或 G进行动态更新，例如在流式场景中新增变换或数据集，也会进一步影响实时性能（第6.4节）。

定理 14。 数据集 d_i 的溯源跟踪复杂度为 $O(|T| \cdot |E|)$。

证明。 要计算 $\mathbf{tr}(d_i)$，请从元数据 $m_j \in M$中获取变换历史记录 $\tau_j(d_i)$，其中最多列出了 $|T|$ 个变换 $t_k \in T$。验证每个变换 t_k 需要检查所有包含 d_i （位于尾部 T_e或头部 H_e）的超边 $e \in E$，每个变换需要进行 $O(|E|)$ 次操作。对于 $|T|$ 个变换，总复杂度为 $O(|T| \cdot |E|)$。□

引理 8. 追踪函数 $\mathrm{tr}(d_i)$ 在复合下具有单调性：若 $t_3 = t_2 \circ t_1$，则 $\mathrm{tr}(t_3(d_i)) \supseteq \mathrm{tr}(t_1(d_i)) \cup \mathrm{tr}(t_2(t_1(d_i)))$。

证明。 复合会在τ_j中追加历史记录，同时保留所有先前的变换，因为链中的每个 t_k 都会按顺序记录。□

vi)**联邦学习** 将分析分布到 Σ，并以第 2 节中的解析函数为基础 A from Section 2。

定义。15 (联邦学习)。 联邦学习 在分布式数据$D = \bigcup_{n \in N} D_n$上计算解析函数 $a \in A$，其中每个节点 $n \in N$ 使用本地数据 D_n训练参数为 θ_n 的本地模型，并聚合结果：

$$a(D) = \bigoplus_{n \in N} a(D_n, \theta_n).$$

本地模型参数 θ_n 通过梯度下降进行更新：

$$\theta_n \leftarrow \theta_n - \eta \nabla \mathcal{L}(a(D_n, \theta_n)),$$

其中， $\eta > 0$ 是学习率， $\mathcal{L}$ 是损失函数（例如，分类采用交叉熵，回归采用均方误差）。全局模型聚合本地更新，通常采用平均值：

$$\theta = \frac{1}{|N|} \sum_{n \in N} \theta_n.$$

例如，在亚马逊卖家网络中，区域服务器 $n_i \in N$ 训练本地模型$a(D_{n_i}, \theta_{n_i})$，根据本地销售数据 $D_{n_i} \subset d_1(t)$预测销售趋势。每个 D_{n_i}都包含带有 {价格，数量}属性的销售记录，模型则输出预测的销售量。本地参数 θ_{n_i} （例如神经网络权重）通过取平均值进行聚合，以形成全局预测器 θ，从而确保隐私，因为原始数据始终保留在本地。

本地模型训练每次迭代的计算复杂度为： $O(|\theta_n| \cdot |D_n|)$，这反映了对 $|D_n|$ 个数据点进行梯度计算的成本；对于具有 $|\theta_n|$ 个参数的模型，每个数据点需要执行 $O(|\theta_n|)$次操作。各节点聚合 θ_n的通信成本与 $|N| \cdot |\theta_n|$成正比，因为每个节点都会发送其参数向量。近期分析指出，异构

模型 [23]。概念漂移是指数据分布 $P(D_n)$随时间变化，可通过统计检验进行检测：

$$D = \sup_x |F_t(x) - F_{t'}(x)|,$$

其中 $F_t(x)$ 和 $F_{t'}(x)$ 分别是在时刻 t 和 t' 的累积分布函数。检测复杂度为 $O(|D_n| \log |D_n|)$，通常通过 Kolmogorov-Smirnov 检验实现。这些挑战，包括通信开销和漂移适应，将在第 6.4 节中进一步分析。

定理 16. 联邦学习中的本地模型训练每次迭代的复杂度为 $O(|\theta_n| \cdot |D_n|)$ 。

证明。 对于每个节点 $n \in N$，计算梯度 $\nabla \mathcal{L}(a(D_n, \theta_n))$ 需要在 $|D_n|$ 个数据点上计算解析函数 $a(D_n, \theta_n)$。对于包含 $|\theta_n|$ 个参数的模型，每次计算需要 $O(|\theta_n|)$ 次操作（例如，在神经网络中计算前向传播和反向传播）。通过梯度下降更新 θ_n 的复杂度为 $O(|\theta_n|)$，因为这涉及对参数进行向量运算。因此，每次迭代的总复杂度为 $O(|\theta_n| \cdot |D_n|)$。□

命题 6. 在 $\nabla \mathcal{L}$满足利普希茨连续性的条件下，联邦梯度下降以 $O(1/k)$ 的速率收敛至一个驻点，其中假设本地梯度的方差有界，k 为迭代次数。

证明。根据联邦平均分析，全局更新满足下降引理；结合光滑性假设和聚合步骤 [12]，可得其收敛速率。□

表 2 概述了亚马逊卖家网络中的各项操作、其计算复杂度及示例，为其数学性质和实际应用提供了简明参考。

操作	复杂度	亚马逊卖家示例
数据集成	NP难	统一销售和库存模式
元数据驱动导航	$O(\|E\| + \|V\| \log \|V\|)$	查询热销电子产品
可扩展性与分布	NP难	将销售分区到各服务器
治理与安全	$O(\|P\| \cdot \|N\|)$	授权经理访问销售
溯源跟踪	$O(\|T\| \cdot \|E\|)$	将销售预测追溯至原始数据
联邦学习	$O(\|\theta_n\| \cdot \|D_n\|)$	按地区预测销售趋势

TABLE 2. 数据织体操作、复杂性与示例概要

4. A Categorical Perspective

范畴论是数学的一个分支，它将结构及其关系抽象为对象和态射，为统一数据织体的操作提供了强大的框架。通过将数据集建模为对象、将变换建模为态射，范畴论提供了分析数据织体组件（$\mathcal{F} = (D, M, G, T, P, A)$）及其操作（第3节）的视角。本节介绍范畴论的基础知识，将数据织体形式化为一个范畴 DF，并展示函子映射和自然变换如何整合操作，以及如何应对一致性等挑战

和动态模式更新（第6节）。我们给出严谨的定义、定理和基础性著作参考，最终实现数据编织框架的范畴统一。

定义。17. 一个范畴 $\mathcal{C}$ 由以下部分组成：

(1) 对象的集合 $Ob(\mathcal{C})$。*(2)* 对于 $A, B \in Ob(\mathcal{C})$ 中的每一对对象，都有一个态射集合 $Hom_{\mathcal{C}}(A, B)$。*(3)* 一个复合运算 $\circ : Hom_{\mathcal{C}}(B, C) \times Hom_{\mathcal{C}}(A, B) \to Hom_{\mathcal{C}}(A, C)$，其中对于 $f : A \to B$, $g : B \to C$，有 $g \circ f : A \to C$。*(4)* 对于每个对象 A，都有一个恒等态射 $id_A : A \to A$。

这些满足以下公理：

(1) Associativity: *For* $f : A \to B$, $g : B \to C$, $h : C \to D$,

$$(h \circ g) \circ f = h \circ (g \circ f).$$

(2) 恒等：对于 $f : A \to B$,

$$f \circ id_A = f, \quad id_B \circ f = f.$$

定理 18。范畴 $\mathcal{C}$ 中的复合运算满足结合律。

证明。根据定义，对于范畴 $\mathcal{C}$ 中的态射 $f : A \to B$、$g : B \to C$、$h : C \to D$，结合性公理表明 $(h \circ g) \circ f = h \circ (g \circ f)$。这直接源于范畴的结构，确保复合的顺序不会影响结果，因为复合的定义满足 $\mathrm{Hom}_{\mathcal{C}}$ 中所有态射的这一性质。□

范畴的示例包括：

(1) **集合范畴**：对象是集合，态射是函数，复合是函数复合，恒等态射是恒等函数。(2) **图范畴**：对象是图，态射是图同态，复合是同态复合，恒等态射保持图结构。(3) **拓扑范畴**：对象是拓扑空间，态射是连续函数，复合和恒等态射采用标准定义。

函子在范畴之间映射并保持其结构，对于将数据织体关联到其超图表示至关重要。

定义。19 (函子)。函子 $F : \mathcal{C} \to \mathcal{D}$ 在范畴 $\mathcal{C}$ 和 $\mathcal{D}$之间指定：

(1) 将每个对象 $A \in Ob(\mathcal{C})$ 映射到对象 $F(A) \in Ob(\mathcal{D})$。*(2)* 将每个态射 $f : A \to B$（属于 $\mathcal{C}$）映射到态射 $F(f) : F(A) \to F(B)$（属于 $\mathcal{D}$），s使得：

(1) 对于 $f : A \to B$, $g : B \to C$，有 $F(g \circ f) = F(g) \circ F(f)$；*(2)* 对每个 $A \in Ob(\mathcal{C})$，有 $F(id_A) = id_{F(A)}$。

定理 20。函子 $F : \mathcal{C} \to \mathcal{D}$ 保持复合和恒等态射。

证明。根据定义，F对于态射 $f : A \to B$,$g : B \to C$，满足 $F(g \circ f) = F(g) \circ F(f)$，从而确保复合得以保持。类似地，$F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$ 保持恒等态射。这些性质是函子的公理，由其定义直接保证。□

自然变换提供了一种比较函子的方法，用于描述数据织体不同表示形式之间的关系。

定义。21 (自然变换**)。** 一个自然变换 $\eta : F \to G$定义在函子 $F, G : \mathcal{C} \to \mathcal{D}$ 之间，它为每个对象 $A \in Ob(\mathcal{C})$ 指定一个态射$\eta_A : F(A) \to G(A)$，且该态射属于 $\mathcal{D}$；对于 $\mathcal{C}$ 中的每个态射 $f : A \to B$，下图均交换：

$$\begin{array}{ccc} F(A) & \xrightarrow{\eta_A} & G(A) \\ \downarrow{\scriptstyle F(f)} & & \downarrow{\scriptstyle G(f)} \\ F(B) & \xrightarrow{\eta_B} & G(B) \end{array}$$

即，$\eta_B \circ F(f) = G(f) \circ \eta_A$。

范畴论的抽象性尤其适用于数据织体，其中数据集（D）可以是对象，变换（T）可以是态射，而集成和导航等操作可以建模为复合或函子映射。基础参考文献包括 [13]（全面论述），[3]（通俗易懂的入门介绍），以及 [18]（数据库和数据管理方面的应用）。

定义。22 (数据织体范畴**)。** 数据织体范畴 DF定义如下：

（*1*）对象：$D = \{d_i(t)\}_{i,\ t}$，即按时间索引的数据资产（公式（*1*））。（*2*）态射：$T = \{t_1, \ldots, t_m\}$，其中 $t_i : d_i \to d_j$是一个变换$t_i : \Omega_i \to \Omega_j$（公式（*3*））。（*3*）复合：对于 $t_1 : d_i \to d_j$，$t_2 : d_j \to d_k$，复合 $t_2 \circ t_1 : d_i \to d_k$由 $(t_2 \circ t_1)(x) = t_2(t_1(x))$ 定义。（*4*）恒等：对于每个 $d_i \in D$，恒等态射$id_{d_i} : d_i \to d_i$，其中$id_{d_i}(x) = x$。

定理 23. DF是一个范畴，具有满足结合性的复合与恒等态射。

证明。 为验证DF是一个范畴：

（1）结合性：对于态射 $t_1 : d_i \to d_j$，$t_2 : d_j \to d_k$，$t_3 : d_k \to d_l$，复合即函数复合：$(t_3 \circ t_2) \circ t_1(x) = t_3(t_2(t_1(x))) = t_3 \circ (t_2 \circ t_1)(x)$，由函数复合的结合性可知其满足结合性。（2）恒等：对于 $t_i : d_i \to d_j$，恒等i$\mathrm{d}_{d_i}(x) = x$ 满足 $t_i \circ \mathrm{id}_{d_i} = t_i$，因为 $t_i(\mathrm{id}_{d_i}(x)) = t_i(x)$，类似地id$_{d_j} \circ t_i = t_i$。对于$\mathrm{id}_{d_j} : d_j \to d_j$，$\mathrm{id}_{d_j}(t_i(x)) = t_i(x)$。

因此，DF 满足范畴公理。 □

例如，在供应链织体中，$d_i \in D$ 是原始货运数据，d_j 是归一化数据，d_k 是汇总库存摘要。变换 $t_1 : d_i \to d_j$（归一化）和 $t_2 : d_j \to d_k$（聚合）复合为 $t_2 \circ t_1 : d_i \to d_k$，表示完整的数据处理流水线。

超图 G（定理 2）通过映射到超图范畴的函子进行建模。

定义 24（超图范畴）。范畴 $\mathcal{HG}$ 具有：*(1)* 对象：超图 $G = (V, E)$，其中 V 是顶点集，$E \subseteq \mathcal{P}(V)$ 是超边集。

(2) 态射：超图同态 ϕ：$G_1 \to G_2$，其中 ϕ：$V_1 \to V_2$映射顶点，使得对每个 $e \in E_1$，$\phi(e) = \{\phi(v) \mid v \in e\} \in E_2$。*(3)* 复合：同态的标准函数复合。*(4)* 恒等：恒等态射 id_G：$G \to G$，其中 $id_G(v) = v$。

函子 F：DF $\to \mathcal{HG}$ 将数据织体映射为其超图表示：

(1) 对象：$d_i \in D \mapsto v_i \in V$，其中 v_i 是 G 中的一个顶点。(2) 态射：t_i：$d_i \to d_j \mapsto e \in E$，其中 $e = (T_e, H_e)$，且 $T_e = \{v_i, m_k\}$、$H_e = \{v_j\}$，而 $m_k \in M$ 是与 t_i 相关联的元数据。

该函子保持复合：

$$F(t_2 \circ t_1) = F(t_2) \circ F(t_1),$$

作为一个序列，$t_2 \circ t_1$：$d_i \to d_k$ 映射为一条从 v_i 到 v_k、经由 v_j 的超边路径。例如，供应链织体中的 $t_1 \circ t_2$ 对应于 G 中的一条路径，该路径连接运输数据、库存和预测，从而实现血缘追踪。

定理 25. 函子 F：DF $\to \mathcal{HG}$ 保持复合与恒等态射。

证明。对于态射 t_1：$d_i \to d_j$, t_2：$d_j \to d_k$, $F(t_1) = e_1 = (\{v_i, m_{k1}\}, \{v_j\})$,$F(t_2) = e_2 = (\{v_j, m_{k2}\}, \{v_k\})$。复合态射 $t_2 \circ t_1$ 映射为超边路径e_1，e_2，其中 $H_{e_1} \cap T_{e_2} = \{v_j\} \neq \emptyset$，在 $\mathcal{HG}$ 中表示为复合态射。因此，$F(t_2 \circ t_1) = F(t_2) \circ F(t_1)$。对于恒等态射 id_{d_i}，$F(\mathrm{id}_{d_i}) = \mathrm{id}_{v_i}$，这是 $\mathcal{HG}$ 中的恒等态射，从而保持恒等态射。□

DF 的范畴结构与函子 F 统一了数据织网的操作（第3节）：

（1）数据集成：在 DF 中复合态射。对于 t_1：$d_i \to d_j$，集成通过串联变换来对齐模式，同时最小化模式距离（定理 6）。（2）元数据驱动导航：通过 $\mathcal{HG}$ 中的 F遍历超边路径，寻找数据集与元数据之间的最短路径（定理 8）。（3）溯源跟踪：重构 DF 中的态射序列，得到 $\mathrm{tr}(d_i)$（定理 13）。（4）联邦学习：将本地分析建模为 DF 中的态射，并将聚合视为跨节点的函子操作（定理 15）。

以医疗健康数据织物为例：患者记录 d_i 通过 t_1（清理）转换为规范化数据 d_j，再通过 t_2（模型推理）转换为预测结果 d_k。DF 中的序列 $t_2 \circ t_1$ 映射为 G 中的超边路径，支持集成（对齐记录）、导航（定位相关数据）、溯源（追踪变换）和联邦学习（分布式模型训练）。

自然变换 用于刻画数据织体不同函子表示之间的模型关系，以解决分布式系统中的一致性约束（第6节）。

定义。26 (数据织物自然变换**)。** 函子之间的自然变换η：$F \to G$ 会为每个数据集 $d_i \in D$ 指定一个态射 η_{d_i}：$F(d_i) \to G(d_i)$，该态射属于 $\mathcal{HG}$，且对于每个变换 t_i：$d_i \to d_j$，满足：

$$\eta_{d_j} \circ F(t_i) = G(t_i) \circ \eta_{d_i}.$$

例如， F 和 G 可能会将 DF 映射为 G的不同超图表示（例如，元数据粒度不同）。自然变换 η 确保导航路径（通过 F）与变换序列（通过 G）保持一致，从而在分布式节点 $N \in \Sigma$之间维护一致性。这对于溯源跟踪等操作至关重要，因为转换历史必须在不同表示之间保持一致。

定理 27. 自然变换 η： $F \to G$ 确保数据织体操作的交换图。

证明。 根据定义，对于 t_i： $d_i \to d_j$, $\eta_{d_j} \circ F(t_i) = G(t_i) \circ \eta_{d_i}$。这确保了 $\mathcal{HG}$ 中的图可交换，也就是说，DF 中的变换（由 F和 G映射）保留了 G的关系结构。对于导航等操作， $F(t_i)$ 和 $G(t_i)$表示超边路径，而 η 确保路径等价，从而维持操作一致性。□

范畴视角为数据织体带来了多项益处：

(1) 统一性：操作被抽象为复合或函子映射，从而简化其分析与实现。(2) 一致性：自然变换用于建模一致性约束，确保分布式节点之间保持一致（第 6节）。(3) 可扩展性：到 $\mathcal{HG}$ 的函子映射支持高效导航和溯源跟踪，不过，对 DF 的动态更新（例如模式演化$S_i(t)$）会带来挑战（第 6.1节）。

例如，在物联网编织中，传感器数据变换是 DF 中的态射，并映射为 G中的超边，从而支持实时分析（第 6.4 节）。然而，NP难的模式匹配（第6.1 节）和延迟约束（第 6.4 节）等挑战，需要采用动态范畴或伴随函子等范畴扩展，相关探讨见 [18]。

范畴方法还通过抽象 Σ 中的节点交互，为分布式系统设计提供指导。未来的工作可以利用幺半范畴或拓扑斯理论来建模治理策略 P 或分析函数 A，并在 [4] 提出的框架基础上继续推进。

5. Hypergraph Embeddings and Representations

超图 $G = (V, E)$ 是数据织体的核心组成部分

$$\mathcal{F} = (D, M, G, T, P, A),$$

对数据资产 D 和元数据 M之间的多元关系进行编码。本节对三个关键特征进行了严谨的数学处理：数据资产的向量嵌入、邻接矩阵的构造，以及将G 嵌入模张量范畴（MTC）。这些内容建立在第 3 节的操作（如集成和导航）以及第 4 节的范畴框架（将数据集视为对象，将变换视为态射）之上。

5.1. 数据资产的向量表示。 为了将数据资产 $D = \{d_i(t)\}_{i,\ t}$ 集成到超图 G 及其范畴嵌入中，我们将定义在域 Ω_i 上、具有模式 S_i的每个数据集 $d_i(t)$表示为有限维空间中的向量。这种表示能够容纳数值型、类别型和混合类型

数据，同时捕捉时间动态，并支持集成和导航等操作。

对于数值数据，其中 $\Omega_i \subseteq \mathbb{R}^k$，数据集 $d_i(t)$ 由点 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 构成，其中每个 $\mathbf{x}_j \in \mathbb{R}^k$。代表向量是质心：

$$\mathbf{v}_i(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j \in \mathbb{R}^k,$$

假设 Ω_i 为凸域。对于非凸域，则使用中心点：

$$\mathbf{v}_i(t) = \arg \min_{\mathbf{x}_j \in d_i(t)} \sum_{l=1}^{n} \|\mathbf{x}_j - \mathbf{x}_l\|_2,$$

计算时间为 $O(n^2 k)$。这能为相似性度量保留集中趋势。

对于分类数据，其中 $\Omega_i = \{c_1, \dots, c_m\}$，我们通过 ϕ 嵌入：$\Omega_i \to \mathbb{R}^d$（例如，独热编码 $\phi(c_j) = \mathbf{e}_j \in \mathbb{R}^m$，或采用 $d \ll m$ 的学习嵌入，并归一化至 $\|\phi(c_j)\|_2 = 1$）。代表向量是频率加权平均：

$$\mathbf{v}_i(t) = \sum_{j=1}^{m} f_j \phi(c_j) \in \mathbb{R}^d,$$

其中 $f_j = |\{\mathbf{x} \in d_i(t) \mid \mathbf{x} = c_j\}|/n$。

对于混合类型数据，结合数值型$\mathbf{x}_{\text{num}} \in \mathbb{R}^k$ 和分类属性，向量为：

$$\mathbf{v}_i(t) = (\mathbf{x}_{\text{num}}, \phi(c_{j_1}), \dots, \phi(c_{j_l})) \in \mathbb{R}^{k+ld},$$

其中总维度标准化为 ≤ 15，以提高运算效率。

时间动态通过$\mathbf{v}_i$ 建模：$T \to \mathbb{R}^p$（其中 $p = k$、d或 $k + ld$）。对于流式更新：

$$\mathbf{v}_i(t + \Delta t) = (1 - \alpha)\mathbf{v}_i(t) + \alpha \mathbf{x}_{\text{new}},$$

其中 $\alpha \in$ 取值范围为(0,1)，并将$\mathbf{x}_{\text{new}}$ 进行适当嵌入，在 $O(p)$时间内完成。

元数据 $m_j = (d_i, \alpha_j, \tau_j) \in M$ 向量化如下：

$$\mathbf{v}_{m_j} = (\phi(\alpha_j), \mathbf{v}_{\tau_j}) \in \mathbb{R}^{q+s},$$

其中 $\phi(\alpha_j) \in \mathbb{R}^q$ 嵌入属性，$\mathbf{v}_{\tau_j} \in \mathbb{R}^s$ 编码历史记录（例如，平均变换参数），从而确保稀疏性和有界维度。

5.2. **邻接结构的构造**。超图的邻接结构 $G = (V, E)$通过有向超边定义连通性，从而实现对于数据集成、导航和溯源跟踪等操作至关重要的多元关系。我们形式化顶点 $V = D \cup M$ 如何经由超边 E相连，并将该结构奠基于范畴框架的第4节。

超边 $e \in E$ 是一个有向对 $e = (T_e, H_e)$，其中 $T_e \subseteq V$ 为尾部（输入顶点），$H_e \subseteq V$ 为头部（输出顶点）。顶点 $v \in V$ 的邻接集为：

$$\text{Adj}(v) = \{e \in E \mid v \in T_e \cup H_e\},$$

含入向分量与出向分量：

$$\text{In}(v) = \{e \in E \mid v \in H_e\}, \quad \text{Out}(v) = \{e \in E \mid v \in T_e\}.$$

关联矩阵定义为：

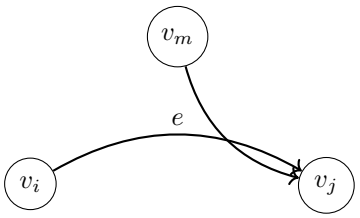
$$I_T \in \{0, 1\}^{|V| \times |E|}, \quad (I_T)_{v,e} = 1 \text{ if } v \in T_e, \quad 0 \text{ otherwise},$$

$$I_H \in \{0,1\}^{|V|\times|E|}, \quad (I_H)_{v,e} = 1 \text{ if } v \in H_e, \quad 0 \text{ otherwise.}$$

对于较大的 G，强制实施稀疏性， I_T 和 I_H 中的非零项以$O(|V|\log|V|)$为界，从而能够以压缩稀疏行（CSR）格式高效存储，每项的访问复杂度为 $O(1)$。

超边根据操作依赖关系生成，并与 DF 中的态射保持一致：

(1) 集成：变换 $t_i \in T$映射 $\{d_{i_1},\ldots,d_{i_n}\} \to d_j$诱导$e = (\{v_{i_1},\ldots,v_{i_n},v_{m_k}\},\ \{v_j\})$，其中 $m_k \in M$记录了 t_i在 $\tau_k(d_j)$ 中的情况。(2) 导航：共享 α_k 中属性的数据集 d_i， d_j构成 $e = (\{v_i,v_j\},\ \{v_{m_k}\})$.
(3) 溯源：与来源 $\{d_{i_1},\ldots,d_{i_n}\}$关联的派生数据集 d_j产生$e = (\{v_{i_1},\ldots,v_{i_n},v_{m_j}\},\ \{v_j\})$。(4) 联邦学习：对 $D_n = \{d_{i_1},\ldots,d_{i_k}\}$进行本地分析并生成 θ_n，从而创建 $e = (\{v_{i_1},\ldots,v_{i_k},v_{m_n}\},\ \{v_{\theta_n}\})$。



有向超边 $e = (\{v_i,v_m\},\ \{v_j\}) \in E$

FIGURE 2. 超图 G的结构，其中顶点 v_i、 $v_m \in V$（数据集或元数据）以及编码多路关系的超边 $e \in E$ 。

通过将每个顶点限制为 $O(\log|V|)$ 条超边来确保稀疏性，从而得到$|E| = O(|V|\log|V|)$ 。具体做法是优先保留具有较高互信息的关系：

$$I(t_i(d_i); d_j) \geq \epsilon,$$

其中 $\epsilon > 0$通过余弦相似度 $\cos(\mathbf{v}_i(t), \mathbf{v}_j(t)) \in [-1,1]$进行近似，并在 $O(p)$中计算。通过确保每个数据集顶点 v_i 都至少可从一个元数据顶点v_{m_j}到达来维持连通性，从而支持导航查询。 G中的路径是序列（e_1、 e_2、.....、 e_m），其中 $H_{e_i} \cap T_{e_{i+1}} \neq \emptyset$，通过适用于超图的迪杰斯特拉算法进行枚举，其复杂度为 $O(|E| + |V|\log|V|)$，与导航效率相符。

5.3. **映射到模张量范畴**。 为以代数方式捕捉超图的多元依赖关系，我们将 $G = (V,\ E)$ 嵌入模张量范畴（MTC）中。MTC 是一种半单、球面带状融合范畴，具有编织幺半结构和非退化模性。此嵌入扩展了第 4 节的范畴框架，其中数据集和变换是 DF 中的对象和态射，从而为操作提供精确的代数模型。

定义. 28. 一个模张量范畴 $\mathcal{C}$ 是半单的球面带状融合范畴，配备有：

(1) 简单对象 的有限集 $\{X_i\}$，在张量积 $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$下封闭，单位对象为 **1**。*(2)* 态射 $Hom_{\mathcal{C}}(X, Y)$，其复合与恒等满足范畴公理。*(3)* 编织，即自然同构$c_{X,\,Y}$： $X \otimes Y \to Y \otimes X$，满足六边形公理：

$$c_{X,Y\otimes Z} = (1_Y \otimes c_{X,Z}) \circ (c_{X,Y} \otimes 1_Z),$$
$$c_{X\otimes Y,Z} = (c_{X,Z} \otimes 1_Y) \circ (1_X \otimes c_{Y,Z}).$$

(4) 一个带扭结，即自然同构 θ_X ： $X \to X$，与编织相容：

$$\theta_{X\otimes Y} = c_{Y,X} \circ c_{X,Y} \circ (\theta_X \otimes \theta_Y).$$

(5) 一个非退化的 S-矩阵，定义如下：

$$S_{X,Y} = tr(c_{Y,X} \circ c_{X,Y}),$$

其中 *tr* 表示量子迹，且 S 在 $\mathbb{C}$上可逆。

嵌入映射将表示数据集 $d_i(t) \in D$ 或元数据 $m_j \in M$的顶点 $v_i \in V$映射为简单对象 $X_i \in \mathcal{C}$。向量表示 $\mathbf{v}_i(t) \in \mathbb{R}^p$ 或$\mathbf{v}_{m_j} \in \mathbb{R}^q$ 用于表示 X_i，例如，作为配备张量积结构的向量空间 $\mathbb{C}^p$ 。张量积对联合依赖关系建模：

$$X_i \otimes X_j \cong \bigoplus_k N_{ij}^k X_k,$$

其中，若存在某个 $e \in E$满足 $v_i,\ v_j \in T_e$，则融合系数 $N_{ij}^k = 1$取值为 $v_k \in H_e$，否则为 0，以反映超边连通性。

每条超边 $e = (T_e,\ H_e)$，其中 $T_e = \{v_{i_1}, \ldots, v_{i_n}\}$, $H_e = \{v_{j_1}, \ldots, v_{j_m}\}$，对应一个态射：

$$f_e : X_{i_1} \otimes \cdots \otimes X_{i_n} \to X_{j_1} \otimes \cdots \otimes X_{j_m}.$$

对于数据集成（定理 6），其中 $H_e = \{v_j\}$， f_e： $\otimes_{k=1}^n X_{i_k} \to X_j$ 编码变换矩阵；对于导航（定理 8）， f_e 投影到元数据对象上。

带扭结刻画了循环依赖：

$$\theta_{X_i} = \sum_{e \in \mathrm{Loop}(v_i)} w_e f_e,$$

其中 Loop(v_i) $\subseteq E$ 是包含 v_i 的超边环集合，而 w_e =$I(t_e(d_i); d_j)$ 用于变换 t_e，否则表示频率，满足：

$$\theta_{X_i \otimes X_j} = c_{Y,X} \circ c_{X,Y} \circ (\theta_X \otimes \theta_Y).$$

The S-矩阵量化连通性：

$$S_{X_i,X_j} = \sum_{e \in E_{i,j}} \mathrm{tr}(c_{X_j,X_i} \circ c_{X_i,X_j}),$$

其中 $E_{i,\,j} = \{e \in E \mid v_i, v_j \in T_e \cup H_e\}$。对于大型超图， S 通过拉普拉斯算子近似：

$$L = I_T I_T^T - I_H I_H^T,$$

其特征值反映连通性，并在 $O(|V| \log |V|)$ 中通过随机化 SVD 计算得出。

定理 29。 超图 G 通过忠实函子Φ: $\mathcal{HG} \to \mathcal{C}$嵌入 *MTC* $\mathcal{C}$ ，其中顶点对应简单对象，超边对应态射，并通过编织和 S-矩阵保留邻接关系。

证明。定义 $\Phi : \mathcal{HG} \to \mathcal{C}$:

(1) 对象：将 $G = (V, E)$ 映射到 $\mathcal{C}_G \subseteq \mathcal{C}$，其中包含简单对象 $\{X_i \mid v_i \in V\}$。(2) 态射：将超图同态 $h : G_1 \to G_2$ 映射到 $\Phi(h) : \mathcal{C}_{G_1} \to \mathcal{C}_{G_2}$，并发送 $X_i \to X_{h(i)}$, $f_e \to f_{h(e)}$。

对于超边 $e = (T_e, H_e)$, $\Phi(e) = f_e : \otimes_{v_k \in T_e} X_k \to \otimes_{v_l \in H_e} X_l$。带有 $H_{e_1} \cap T_{e_2} \neq \emptyset$的路径 (e_1, e_2) 映射到 $f_{e_2} \circ f_{e_1}$，并保持复合。编织c_{X_i, X_j} 满足六边形公理，用于刻画对导航至关重要的超边对称性（定理 8）。扭结 θ_{X_i} 编码环，并支持溯源（定理 13）。 S-矩阵通过 G的连通性确保非退化性，从而量化顶点间的相互作用。由于 Φ保留对象、态射及其关系结构，因此它是忠实的，并将 G嵌入 $\mathcal{C}$。□

MTC 的编织$c_{X_i,\ X_j}$ 源自量子群的 R 矩阵，例如 $U_q(\mathfrak{sl}_2)$ 在单位根处的 R 矩阵，它编码了非交换对称性 [22]。这种量子群结构构成了 S矩阵和融合规则的基础，这表明量子启发式算法可以利用这些代数对称性，优化导航和溯源跟踪等操作，相关内容将在第6.2 节和第6.4节进一步探讨。

5.4. 编织作用与几何类比。MTC 中的编织$\mathcal{C}$是一种自然同构$c_{X,\ Y}$: $X \otimes Y \to Y \otimes X$ 满足六边形公理，用于描述 G中超边的对称性。对于超边 $e = (T_e,\ H_e)$，当 $T_e = \{v_i, v_j, \dots\}$时，编织可确保输入顺序不变性，从而保持态射 f_e。这对于元数据驱动导航至关重要，在这种导航中，由元数据关联的数据集构成$e = (\{v_i, v_j\},\ \{v_{m_k}\})$，交换输入仍能保持该关系：

$$f_e \circ c_{X_i, X_j} = f_e.$$

这种不变性支持置换不变查询以及溯源跟踪中的变换序列。

为使几何类比更加严谨，我们回顾一下：模张量范畴会在拓扑量子场论（TQFT）中导出模函子，为带标记曲面赋予有限维向量空间，为配边赋予线性映射，并满足粘合公理，从而在构形空间的同调群上产生映射类群的射影表示。在 Reshetikhin-Turaev 等三维拓扑量子场论中，模函子编码了亏格为 g、带有 n 个穿孔的曲面上的共形块，其中的编织来自穿孔周围的半单值。在 Hur$_{d\ 空间,\ n}$中，以 $\mathbb{P}^1$ 的 d 次覆盖（具有 n 个分歧点及固定分歧型）为参数时，对于简单分歧，辫群 $B_n = \pi_1(\text{Conf}_n(\mathbb{C})/S_n)$ 对连通分量具有传递作用（Hurwitz 定理）。对于具有分歧点 $p = (p_1、\ \dots\dots、\ p_n) \in \text{Conf}_n(\mathbb{C})$ 的覆盖 $\phi: C \to \mathbb{P}^1$ ，单值 ρ_p：$\pi_1(\mathbb{P}^1 \setminus p,\ b) \to S_d$ 在 B_n 作用下共轭：

$$\rho'(\gamma) = \beta\rho(\gamma)\beta^{-1}, \quad \beta \in S_d,$$

其中，共轭源自 $\text{Conf}_n(\mathbb{C})$ 中的路径。这与 MTC 编织类似，其中$c_{X_i,\ X_j}$ 诱导保持复合的同构，类似于映射类群作用下的模函数映射。

形式上，编织满足杨-巴克斯特方程：

$$(c_{X,Y} \otimes \mathrm{id}_Z)(\mathrm{id}_X \otimes c_{Y,Z})(c_{X,Z} \otimes \mathrm{id}_Y) = (\mathrm{id}_Y \otimes c_{X,Z})(c_{X,Y} \otimes \mathrm{id}_Z)(\mathrm{id}_X \otimes c_{Y,Z}),$$

对应于辫群关系。对于多顶点边 $e = (\{v_{i_1}, v_{i_2}, v_{i_3}\}, \{v_j\})$：

$$f_e \circ (c_{X_{i_1},X_{i_2}} \otimes \mathrm{id}_{X_{i_3}}) = f_e,$$

根据自然性和相干性。

在模空间 $\mathcal{M}_{g,\,n}$ 中，映射类群 $\mathrm{Mod}_{g,\,n}$ 通过德恩扭转作用于 Teichm¨uller 空间；对于 $g = 0$，这可简化为 B_{n+1}。来自模张量范畴的模函子表示 $\mathrm{Mod}_{g,\,n}$，其中德恩扭转是量子群表示中卡西米尔算子的指数。在数据织网中，编织模拟依赖关系中的“扭结”，并通过相干性保持输出。

在代数上，模张量范畴是半单带状范畴，其 S-矩阵非退化；其模函子通过拓扑量子场论中的柯比着色，将 B_n 表示扩展到 $SL(2,\ Z)$。由量子群在 level k 处构造的酉模张量范畴产生有限维表示，并为联邦学习保持范数。

纽结理论：闭辫产生链路；诸如 Jones $V_L(q) = \mathrm{tr}(\rho(\beta))$（其中 β 辫）的不变量可以检测异常。

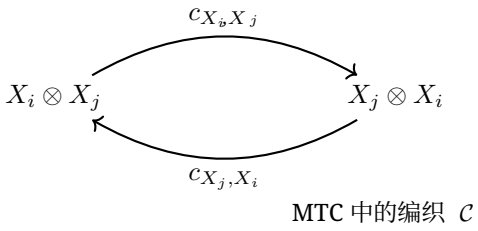


FIGURE 3. 编织对称性建模。

引理9。 编织 $c_{X_i,\ X_j}$ 保留超边语义。

证明。 设 f_e : $\bigotimes_{k\in T_e} X_k \to \bigotimes_{l\in H_e} X_l$。根据 c 的自然性，该图交换：

$$\begin{array}{ccc} \bigotimes T_e & \xrightarrow{f_e} & \bigotimes H_e \\ {\scriptstyle c_{X_i,X_j}\otimes \mathrm{id}}\downarrow & & \downarrow{\scriptstyle \mathrm{id}} \\ \bigotimes T_e & \xrightarrow{f_e} & \bigotimes H_e. \end{array}$$

因此 $f_e \circ (c_{X_i,\ X_j} \otimes \mathrm{id}) = f_e$。对于任意置换，相干定理确保编织生成一致的同构（幺半范畴的 Mac Lane 相干定理）。在导航中，辫群作用下的等价路径；在溯源中，由于幺半结构通过函子性保留复合，序列保持不变。□

5.5. **嵌入与应用**。 为了在统一的代数结构中捕捉关系对称性，尤其是面向未来的量子数据应用（其中编织幺半对称性可用于建模纠缠和拓扑不变量），我们将超图 G嵌入模张量范畴（MTC） $\mathcal{C}$。该嵌入扩展了第 4 节中的范畴框架，为 NP 困难任务（第 6 节）中的基于对称性的优化以及

CAP/CAL 下的容错操作（第 7 节）。借助模张量范畴，该框架得以利用编织结构进行置换不变处理、利用纽结不变量进行异常检测，并利用模函子支持潜在的量子规模计算；这与量子数据中心和后量子安全的发展趋势相契合。

该嵌入通过忠实函子形式化表示 $\Phi:\ \mathcal{HG}\to\mathcal{C}$，其中 $\mathcal{HG}$ 是以超图为对象的有向超图范畴 $G=(V,E)$，以保持尾部和头部的超图同态为态射（定理 22）：

(1) 对象： $\Phi(G)=\bigoplus_{v_i\in V}X_i$，其中每个 X_i 都是对应于顶点 v_i 的 $\mathcal{C}$ 中的简单对象，构成一个子范畴 $\mathcal{C}_G\subseteq\mathcal{C}$。(2) 态射：超图同态 h： $G_1\to G_2$ 映射到 $\Phi(h):\mathcal{C}_{G_1}\to\mathcal{C}_{G_2}$，将 $X_i\to X_{h(i)}$ 和 $f_e\to f_{h(e)}$发送到相应位置，其中 f_e 是与超边 e 关联的态射。

命题 7。 函子 Φ 是忠实的，并保持态射的单射。

证明。 由于 $\mathcal{C}$ 是半单且带状的，简单对象 X_i 具有不同的量子维数，而态射 f_e 由简单路径基底中的矩阵元确定。超图同态 h 会在对象和超边上诱导唯一的标记，确保 Φ 将 $\mathrm{Hom}_{\mathcal{HG}}(G_1,\ G_2)$ 单射到 $\mathrm{Hom}_{\mathcal{C}}(\mathcal{C}_{G_1},\ \mathcal{C}_{G_2})$ 中，因为非同构超图会产生不同的融合规则。□

这种嵌入的应用包括：

（1）导航：态射复合 $f_{e_2}\circ f_{e_1}$ 表示超边路径，与最短路径查询相一致。在模张量范畴中，编织确保路径在同构意义下等价，从而可通过模不变量实现高效遍历，并利用 S 矩阵对角化将复杂度降至 $O(|V|\log|V|)$。（2）溯源：带扭结 $\theta_{X_i}\in\mathrm{Aut}(X_i)$ 满足 $\theta_{X\otimes Y}=c_{Y,\ X}\circ c_{X,\ Y}\circ(\theta_X\otimes\theta_Y)$，用于建模循环依赖。对于环 $d_i\to d_j\to d_i$，迹 $\mathrm{tr}(\theta_{X_i\otimes X_j})$ 产生能够检测非平凡环的不变量，从而支持变换序列。（3）联邦学习：态射 f_e 聚合本地分析，同时通过模张量范畴中的枢轴结构保护隐私，其中对偶 X^* 支持伴随操作，以实现梯度共享。这确保了分布式更新中的相干性，而量子扩展则有望利用任意子融合实现安全多方计算。

S矩阵 $S_{ij}=\frac{1}{\mathcal{D}}\,\mathrm{tr}(c_{X_j,\ X_i^*}\circ c_{X_i,\ X_j})$，其中 $\mathcal{D}$ 表示总量子维数，可通过对超图拉普拉斯算子进行随机化奇异值分解来近似 $L=I-D^{-1/2}AD^{-1/2}$(由关联关系构造的 A)，其计算时间为 $O(|V|\log|V|)$，误差为 ϵ，采用秩-k 近似（对于稀疏 G，k ¡¡ —V—）。这提升了大型数据织网的可扩展性（第6.2节）。

6. Computational Challenges

数据编织框架 $\mathcal{F}=(D,\ M,\ G,\ T,\ P,\ A)$，在分布式系统 $\Sigma=(N,\ C)$（第 2 节）上运行，可支持数据集成、元数据驱动导航和联邦学习等复杂操作（第 3 节）。然而，其数学复杂性根植于超图 G （第 5.2 节）及其嵌入模张量范畴（MTC）之中（第5.3 节），由此带来了重大的计算挑战。数据资产的异质性以及跨越

分布式节点、通过策略执行实现的治理，以及实时处理需求，都会阻碍互操作性、效率、合规性和响应能力。本节在范畴结构 DF（定理 22）和 MTC 的辫状幺半框架（第 5.4 节）中分析这些挑战，并给出严格的复杂度界和 NP 困难性证明。我们提出利用超图邻接性质和 MTC 对称性的缓解策略，并探讨其与拓扑方法和谱方法的联系，以解决这些瓶颈，确保与操作层面和范畴层面的基础保持一致。

6.1. **异质性**。 数据资产中的异质性 $D = \{d_i\ (t)\ \}_{i,\ t}$，其特征在于多样的模式 S_i 与域 Ω_i，使数据集成和元数据驱动的导航变得复杂（定理 6 与定理 8）。对齐不同的表示需要变换 $t_i \in T$，这些变换被建模为数据织体范畴 DF 中的态射（定理 22）。超图 G，其顶点为 $V = D \cup M$ 与超边 E，并使用向量表示对这些关系进行编码（第5.1节），但随机分布和动态模式演化带来了计算上的障碍。

定义 30（异构数据集成）。 异构数据集成统一数据集 d_i、$d_j \in D$，这些数据集具有不同的模式 $S_i \neq S_j$ 与域Ω_i、Ωj，经由变换 $t_i \in T$，在随机条件下最小化期望损失，同时确保语义兼容性。

该优化问题寻求一个变换 $t_i \in T$，以最小化：

$$\min_{t_i \in T} \mathbb{E}[\mathcal{L}(t_i(d_i), d_j) \mid P(d_i, d_j)],$$

其中，$\mathcal{L} = W_2(P_{t_i(d_i)}\ P_{d_j})$ 是分布之间的 2-瓦瑟斯坦距离。对于 $\mathbb{R}^k$ 中的 n 个点分布，计算 W_2 需要求解一个最优传输问题，其复杂度为： $O(n^3 \log n)$，该复杂度源于对 $n \times n$ 传输矩阵进行线性规划；对于高维 Ω_i，这会给可扩展性带来压力。在 MTC 嵌入（第 5.3 节）中，t_i 对应一个态射 $f_e : X_i \to X_j$，而损失与融合系数 N_{ij}^k 相关；对于动态模式，这些系数通过迭代计算，复杂度为 $O(|E|)$。

动态模式匹配对于不断演变的 $S_i(t)$ 至关重要，其目标是最大化属性相似度：

$$\max_{\pi : S_i \to S_j} \mathbb{E}[\mathrm{sim}(a, \pi(a)) \mid \mathrm{ont}(S_i)],$$

其中 $\mathrm{ont}(S_i)$ 是一个带噪声的本体，而 $\mathrm{sim}(a, b) \in [0, 1]$。如定理 6所示，这是 NP难问题，因为它将子图同构归约为模式匹配。范畴论视角将 $S_i(t)$ 建模为 DF 中的动态对象，但调整态射需要在每次更新时执行 $O(|S_i|)$，这对大型模式构成瓶颈。对于非结构化数据，通过最优传输获得的嵌入（第 5.1 节）具有 $O(n\log\ n)$ 的复杂度，影响联邦学习（定理 15）。Sinkhorn 算法将其降至 $O(n^2)$，其准确度权衡受以下界限约束：

$$I(t_i(d_i); d_j) \leq I(d_i; d_j) - \mathcal{L}(t_i),$$

其中 I 表示互信息。MTC 的编织作用（第 5.4节）表明，可以采用置换不变态射来简化模式对齐，并有望利用拓扑不变量降低复杂度，相关探讨见第 9节。

MTC 的融合系数源自量子群表示，为模式匹配提供了一种受量子启发的方法。通过将数据集建模为

类似量子群的表示 $U_q(\mathfrak{sl}_2)$，变换 t_i 充当交织算子，有望利用非交换对称性降低对齐含噪本体的复杂度，详见第9节。

6.2. **可扩展性**。 可扩展性对于处理大型数据集至关重要 D ， Σ这也是云端分析平台所要求的。具有稀疏关联矩阵的超图 G， I_T， I_H （第5.2节）支持高效导航，但动态工作负载和节点可变性会对分布效率构成挑战（定理9）。

定义。31 （可扩展数据分布）。可扩展数据分布对$D = \bigcup_{n\in N} D_n$进行分区，其中 $D_n \subseteq D$ 驻留在节点 $n \in N$上，以最小化计算成本 $\sum_{n\in N}$成本$(a(D_n))$ 以及通信成本 $\sum_{(n_i,\ n_j)\in C}$ $\mathrm{comm}(D_{n_i}, D_{n_j})$。

分区问题的优化目标是：

$$\min_{\{D_n\}} \left(\sum_{n\in N} \mathrm{cost}(a(D_n)) + \sum_{(n_i,n_j)\in C} \mathrm{comm}(D_{n_i}, D_{n_j}) \right),$$

其中，对于分析 $a \in A$， $\mathrm{cost}(a(D_n)) = O(|D_n|)$，而$\mathrm{comm}(D_{n_i},\ D_{n_j})$与数据传输量成正比。这是NP难问题，因为图划分可以归约为数据划分。

定理3。 分区问题是*NP*难。

证明。 将图划分问题归约为数据划分。给定一个带有边权重 $w(e')$ 的图范畴$G' = (V',\ E')$，将顶点 V' 映射为数据集 D，将超边 E' 映射为依赖关系，并将权重映射为通信开销。将 D 划分为 $|N|$ 个子集以最小化节点间通信，对应于将 G' 划分以最小化边割权重：

$$\sum_{(v_i,v_j)\in E', v_i\in D_{n_k}, v_j\in D_{n_l}, k\neq l} w(v_i, v_j).$$

图划分是NP难的，且该归约为多项式时间，可在 $O(|V'| + |E'|)$ 中构造 D及依赖关系。因此，数据划分是NP难的。 □

对于 $|N|$个节点，穷举分区的复杂度为 $O(|N|^D$，启发式方法的近似复杂度为$O(|N|^2)$。在MTC框架（第 5.3 节）中，分区将张量积 $X_i \otimes X_j$分解，但对于大型 $|V|$ ，计算融合系数 $O(|E|)$ 需要 $d_i(t)$。针对流式处理

$$\mathrm{time}(\mathrm{repartition}) = O(|D| \log |N|),$$

借助并行执行，可降至 $O(|D|\log\ |N|/|N|)$，但同步成本会按 $O(|N|)$ 的规模增长。负载不均衡，其中：

$$\max_{n\in N} \mathrm{load}(n) \gg \lambda, \quad \mathrm{load}(n) = |D_n| + \sum_{a_i\in A_n} \mathrm{cost}(a_i),$$

需要预测模型，其复杂度为 $O(|N| \cdot t)$，时间范围为 t。基于超图拉普拉斯算子的谱聚类：

$$L = I_T I_T^T - I_H I_H^T,$$

将划分复杂度降低至：

$$O(|V| \log |V| + |N| \log |N|),$$

通过计算 L的顶部特征向量，但近似比通常超过 1.5，影响导航效率（定理 8）。

MTC 的 S矩阵源自量子群表示，为拉普拉斯算子提供了谱类比 L，这表明可以采用量子启发式划分算法。利用源自量子群（例如 $U_q(\mathfrak{sl}_2)$ 的非退化S矩阵，我们可以提高聚类效率，并有望将近似比降至 1.5 以下，详见第9节。

6.3. **治理**。 治理通过策略确保合规性与安全P，这对访问控制（定理 11）等操作至关重要。超图 G 将策略与数据集关联起来，但在 Σ 范围内扩展策略执行的计算成本很高。

定义。33 (治理策略执行)。 治理策略执行会评估策略 $P = \{p_1, \ldots, p_l\}$，其中 $p_i = (c_i, a_i)$ 的谓词为 c_i： $D \times \mathcal{U} \to \{0, 1\}$，以授予对访问请求 $r(d_i, u)$ 的访问权限，并确保分析$a \in A$中的差分隐私。

正在评估请求 $r(d_i, u)$：

$$\bigwedge_{p_j \in P} c_j(d_i, u),$$

需要 $O(|P| \cdot |N|)$ 步，因为每个谓词 c_j 都要在 $O(1)$ 中的 $|N|$ 个节点上进行检查，假设角色查找为常数时间。在 MTC 中，策略是约束交互的态射，但评估规模随 $|N|$线性增长。差分隐私：

$$P(a(D) \mid D) \leq e^{\epsilon} P(a(D') \mid D') + \delta,$$

会造成效用损失：

$$\mathcal{L}_{\text{utility}} \propto \frac{1}{\epsilon},$$

对 ϵ、δ 进行优化需要 $O(|D|)$，从而影响联邦学习（定理 15）。分布式哈希表可将验证降至 $O(|P| \cdot \log |N|)$，但会与大规模 $|T|$的溯源跟踪（定理 13）发生冲突。并行评估可达到 $O(|P|/|N|)$，通信开销为 $O(|N| \cdot \log |P|)$。编织作用（第 5.4 节）表明，可以通过利用 G中的关系对称性来对策略进行对称化。

引理 10. 并行策略评估将复杂度降至 $O(|P|/|N|)$，前提是负载均匀分布。

证明。 将 $|P|$策略分配到 $|N|$节点上，并为每个节点分配 $|P|/|N|$谓词。每次谓词评估的开销为 $O(1)$，因此每个节点的开销为 $O(|P|/|N|)$。假设同步开销可以忽略不计，则均匀负载可确保没有节点超过此上限；分布式系统模型（第 2 节）验证了这一假设。 □

6.4. **实时处理**。 对流式处理 $d_i(t)$进行实时处理对动态应用至关重要，其要求如下：

$$\text{time}(t(d_i(t))) + \text{time}(a(t(d_i(t)))) \leq \delta,$$

其中， δ 是延迟上限，在 DF 中建模为态射（定理 22）。

定义。34 (实时处理)。 实时处理会在延迟上限δ内对流数据 $d_i(t) \in D$ 应用转换 $t \in T$ 和分析 $a \in A$，以适应动态分布和概念漂移。

延迟复杂度如下：

$$\text{time}(t) = O(|d_i(t)| \cdot k), \quad \text{time}(a) = O(|\theta| \cdot |d_i(t)|),$$

其中， k 表示变换维度， $|\theta|$ 表示模型大小，对于大型 $|d_i(t)|$，通常超过δ 。概念漂移检测：

$$D = \sup_x |F_t(x) - F_{t'}(x)|,$$

通过柯尔莫哥洛夫-斯米尔诺夫检验，其规模为：

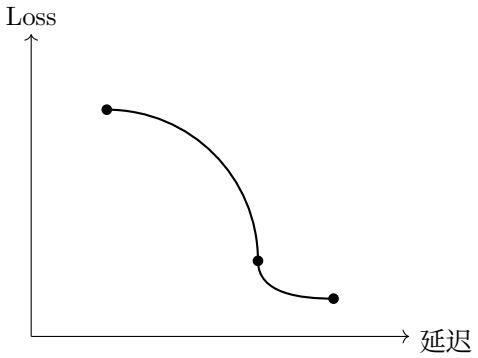
$$O(n \log n),$$

模型适配需要 $O(|\theta|^2)$。在 $|N|$ 个节点上进行并行检测，复杂度可降至 $O(n \log n/|N|)$，同步成本为 $O(|N|)$。针对延迟、损失和资源进行优化：

$$\min_{\theta,t,a} \left(\text{latency}(t,a), \mathcal{L}(a), \text{resource}(t,a)\right),$$

得到一个复杂度为 $O(|T| \cdot |A|)$ 的帕累托前沿，可通过贪心近似降至 $O(|T| + |A|)$，但延迟最多会增加 20%。MTC 的 S-矩阵近似（第 5.3 节）将导航复杂度降至 $O(|V| \log |V|)$，从而提升实时分析能力。

源自量子群表示的 S矩阵，或可启发用于实时处理的量子算法；借助 MTC 的代数结构优化变换与分析流水线，在量子计算环境中有望实现亚线性复杂度 [22], ，如第 9 节所述。



延迟-损失优化的帕累托前沿。

FIGURE 4. 实时处理的帕累托前沿，平衡延迟与损失之间的权衡。

数据织体的计算挑战——异质性、可扩展性、治理和实时处理——源于其数学结构的复杂性，尤其是超图 G 及其 MTC 嵌入。异质性带来的 $O(n^3 \log n)$ 瓦瑟斯坦计算和 NP难的模式匹配，阻碍了集成。受 $O(|N|^2)$ 分区限制，可扩展性面临负载不均衡和重新分区成本。治理在 $O(|P| \cdot |N|)$ 下难以扩展，同时面临隐私与效用权衡；实时处理则难以满足 $O(|d_i(t)| \cdot k)$ 延迟界限。范畴框架 DF 和 MTC 嵌入提供了缓解策略：编织

对称性可简化模式匹配，谱方法可降低分区复杂度，随机化奇异值分解可增强实时分析。未来工作应探索用于策略执行的幺半范畴、用于漂移检测的拓扑不变量，以及用于并行框架，以确保数据织体在大规模应用中的可扩展性和响应能力。

此外，支撑 MTC 结构的量子群，有望为这些挑战带来新的思路。它们的非交换作用可以建模动态模式对齐，其谱性质则可以增强漂移检测算法，详见第9。

7. Consistency, Completeness, Causality

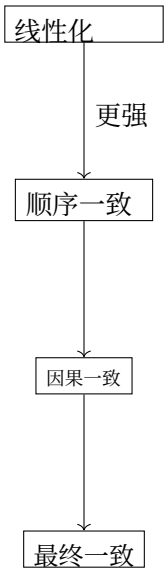
数据织体 $\mathcal{F} = (D, M, G, T, P, A)$ 运行于分布式系统$\Sigma = (N, C)$（第 2 节）之上，依赖于一致性、完备性与因果性，以保证数据集成、元数据驱动导航和联邦学习等可靠操作（第 3 节）。这些性质以超图 $G = (V, E)$（第 5.2 节）和范畴结构 DF（定理 22）为基础，用以缓解分布式管理方面的挑战（第6节）。模张量范畴（MTC）嵌入（第 5.3 节）对关系动力学进行建模，从而在网络故障与数据流中确保相干性。本节利用稀疏关联矩阵（第 5.2 节）和模张量范畴编织（第 5.4 节）将这些性质加以形式化，并在图5和图7中给出可视化。

7.1. 分布式系统与数据织体。 分布式系统 $\Sigma = (N, C)$ 构成了数据织体的骨干，其中节点 N 托管分区数据资产 $D_n \subseteq D$，链路 $C \subseteq N \times N$ 则实现通信。这种架构与超图 G相吻合：顶点 $V = D \cup M$ 表示数据集和元数据，超边 E 则建模集成与导航等操作所涉及的多元依赖关系。正常情况下，Σ是连通的，可由邻接矩阵 $A_\Sigma \in \{0,1\}^{|N|\times|N|}$表示，从而确保高效的查询路由和转换执行。然而，网络分区可能导致 N的子集断连，因此需要通过关联矩阵 I_T，I_H中的多条路径，在 G 中实现冗余。MTC 嵌入将节点映射为简单对象，将链路映射为态射，并通过编织对称性，在分区情况下保留关系结构。

7.2. 一致性。 一致性确保数据资产 D 上的操作在各节点 N之间表现一致，这对于准确集成和联邦学习至关重要。形式上，一致性要求：

(1) 原子性：对 $d_i \in D$ 的操作（例如通过 $t \in T$ 进行的更新）在所有 Σ 上看起来都是瞬时完成的。(2) 顺序性：操作具有全序，就像在单个节点上执行一样，并满足：$\forall d_i \in D$，$\forall n$，$m \in N$，$\text{state}_n(d_i) = \text{state}_m(d_i)$，其中 $\text{state}_n(d_i)$ 表示 d_i 在节点 n 上的状态。

如图5所示，按照 CAP 定理的约束，线性一致性这一最强模型要求序列立即达成一致。对于集成ϕ：$D \to D'$，线性一致性可确保查询 $q(d_i)$ 即时反映 $\phi(d_i)$，从而保持模式对齐。然而，同步延迟会按 $O(|N|)$ 的速度增长，影响实时处理。



一致性模型层次结构 [8]。

FIGURE 5.一致性模型之间的关系，从最强（线性化）到最弱（最终一致）。

为量化各种权衡，我们应用 CAL 定理（见 [11]），该定理将一致性C、可用性 A 和延迟 L 与极大加代数（$a\oplus b=\max(a,b)$、 $a\otimes b=a+b$）联系起来：

$$C \oplus A \leq L.$$

此处， $C=O(|N|)$ 用于线性一致性， $A\approx 0$ 用于强可用性，且 L 包含超图遍历 $O\left(|E|+|V|\log |V|\right)$ 。稀疏矩阵 I_T， I_H 优化 L，平衡 C 与 A。在 MTC $\mathcal{C}$ 中，一致性即编织$c_{X_i,\ X_j}$下的态射不变性，从而确保置换不变更新。

最终一致性允许暂时的不一致，并使副本趋于收敛：

$$\lim_{t\to\infty} \mathrm{state}_n(d_i,t)=\mathrm{state}_m(d_i,t).$$

超图 G 借助依赖边提供辅助，并由元数据 m_j 跟踪历史记录 τ_j以实现语义收敛。

引理 11. 在 *CAL* 下，一致性延迟 C 以图直径 $\mathrm{diam}(\Sigma)$ 为上界。

证明。 在极大加法代数中， $C=\max_{n,m} L_{n,m}$，其中 $L_{n,m}$ 为路径延迟；根据最短路径性质， $C\leq \mathrm{diam}(\Sigma)\cdot \max L_e$，其中边 e 属于 C。□

7.3. **可用性**。 可用性确保数据织体能够响应请求，这对于导航和实时分析至关重要。形式化地：

(1) 弱可用性：在正常条件下，系统最终会作出响应：

$$\forall q: D\to\{0,1\}, \exists t, q(d_i,t) \text{ returns}.$$

(2) 强可用性：在故障或分区期间仍能响应：

$$\forall q, \exists n\in N\setminus N_{\text{failed}}, q(d_i) \text{ succeeds}.$$

根据 CAL 定理，可用性 A 与一致性 C 和延迟 L之间存在权衡$C + A \leq L$。强可用性 $A \approx 0$ 通过 G中的冗余路径实现：

$$|\operatorname{Paths}(v_i, v_j)| > 1,$$

由 I_T, I_H 在 $O(|E|)$ 中计算得到。导航查询可在备选方案上成功，复杂度为$O(|E| + |V| \log |V|)$。策略 P在不降低 A 的前提下强制执行访问，并使用 $O(|P| \log |N|)$ 中的分布式协议。在 DF 中，可用性经由用于路由的自然变换，确保查询的态射存在。

引理 12. 冗余 $\operatorname{rank}(I_T + I_H) \geq |V| - k$ 可容忍 k 个故障。

证明. 秩条件确保在移除 k 列（失效超边）后仍保持连通性；依据矩阵扰动，供查询使用的路径仍然存在。 □

7.4. 分区容错. 分区容错使数据织体能够在 Σ 被划分为互不连通的分量并隔离节点子集 N 时继续运行。形式上， Σ 具有分区容错，如果：

$$\forall P_1, P_2 \subseteq N, P_1 \cup P_2 = N, P_1 \cap P_2 = \emptyset, \exists D_{P_1}, D_{P_2} \text{ s.t. } \operatorname{ops}(D_{P_1}), \operatorname{ops}(D_{P_2}) \text{ succeed},$$

其中 ops 为查询或更新等操作。超图 G 通过识别备选路径支持分区容错，并带有冗余：

$$\operatorname{rank}(I_T + I_H) \geq |V|,$$

确保连通性（第 5.2节）。CAL 定理通过量化分区期间的延迟，为分区容错提供依据 L ；分区期间，对 D_{P_1} 和 D_{P_2}执行的操作若满足以下条件，则会成功：

$$L \geq \max(C_{P_1}, A_{P_1}) + \max(C_{P_2}, A_{P_2}),$$

反映分区内的独立处理。策略 P 确保分区期间的合规性，限制对敏感数据的访问（第 6.3节），其复杂度为$O(|P| \cdot \log |N|)$ 。分区 $D = \bigcup_{n \in N} D_n$ 可最大限度地减少依赖关系，与 MTC 张量积分解相符（第5.3节），其中张量积$X_i \otimes X_j$ 用于对独立的数据资产建模。MTC 中由量子群融合规则驱动的张量积分解，可以描述非局部的分区依赖关系，从而实现受量子启发的路由策略，增强断连组件间的容错能力；这一设想将在第9节作为未来研究方向提出。

定理 35。 *CAP* 定理指出，分布式系统无法同时保证线性一致性、强可用性和分区容错；最多只能满足其中两项。

证明。 假设某系统能够保证全部三项属性。考虑将 N 拆分为 N_1， N_2，其中 $d_i \in D$ 复制在节点 $n_1 \in N_1$、 $n_2 \in N_2$上。对 n_1 执行的更新 $u(d_i)$必须立即反映出来（一致性，即线性一致性， $C = 0$），并且可供访问（强可用性， $A = 0$）。分区容错要求 n_2 作出响应，但在无法通信的情况下， n_2 无法反映 u（d_i），这会违反一致性；或者必须拒绝请求，从而违反可用性。CAL 定理对此进行了量化：$C + A \leq L$；而在分区期间满足 $L = \infty$ ，若 $C = A = 0$，则会导致矛盾。因此，至多只能满足两项属性，正如 [7]中形式化证明的那样。□

CAP 定理经 CAL 定理补充后，对操作产生影响：

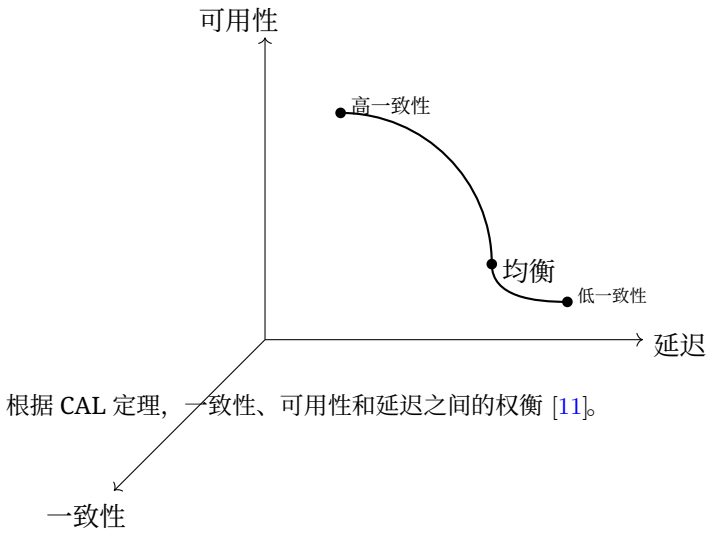


FIGURE 6. 数据织体的 CAL 定理权衡。

数据集成： 线性一致性（$C = O(|N|)$）可确保 $\phi : D \to D'$的模式精确对齐，但会增加延迟 L，从而影响实时处理（第 6.4 节）。

可扩展性： 分区容错支持分布式分析 $a(D) = \bigoplus_{n \in N} a(D_n)$，最终一致性可降低 C，最大限度减少 L (第6.2节)。

实时处理： 强可用性（$A \approx 0$）优先保障快速查询响应，这对于满足延迟上限至关重要 δ，同时通过优化 L(第6.4节)。

在 [2, 11]，中分析的延迟与一致性权衡，通过 G 影响导航效率。

7.5. **完备性**。 完备性确保 D 涵盖满足查询 q 所需的所有数据集： $D \to \{0, 1\}$:

$$\exists d_i \in D, q(d_i) = 1.$$

元数据 $m_j = (d_i, \alpha_j, \tau_j) \in M$ 编目属性与变换历史，从而支持导航（定理 8）以验证覆盖范围。超图 G 通过超边将数据集连接到元数据，从而确保:

$$\forall q, \exists v_i \in V, v_{m_j} \in V, e = (\{v_i\}, \{v_{m_j}\}) \in E, q(v_i) = 1.$$

溯源跟踪（定理 13）验证变换 $t \in T$ 保持完备性:

$$\mathrm{tr}(d_j) = \{t_k \in T \mid t_k \text{ applied to } d_j\},$$

防止数据丢失。在动态结构中，通过更新来维持完备性M，复杂度为 $O(|M| \cdot |T|)$（引理2）。在 DF 中，完备性对应于对象覆盖，并通过自然变换确保查询相关态射存在（定理26）。

7.6. **因果性**。因果性强制形成偏序 $\prec$ 于 G之上，确保操作遵循依赖关系，这对集成和联邦学习至关重要。形式上， $d_i \prec d_j$若 d_i 通过变换 $t \in T$ 或超边影响 d_j：

$$d_i \prec d_j \iff \exists t \in T, t(d_i) = d_j \text{ or } \exists e = (\{v_i, \dots\}, \{v_j\}) \in E.$$

查询遵循以下顺序：

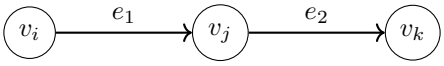
$$q(d_j) \implies \text{check } \{d_i \mid d_i \prec d_j\}.$$

溯源可重建 τ_j，超边编码 $\prec$，从而确保因果一致：副本按顺序观察具有因果关系的操作，而并发操作的顺序则可能不同。

这支持联邦学习，其中本地模型 θ_n 依赖于在因果上先行的 D_n。MTC 模张量范畴编织确保组合具有置换不变性，并通过以下方式保留依赖关系：

$$f_e \circ c_{X_i, X_j} = f_e,$$

复杂度为 $O(|E|)$。



因果路径 $v_i \prec v_j \prec v_k$。

FIGURE 7. G中的因果依赖关系。

引理 13。 因果一致性确保查询结果遵循 $\prec$，验证复杂度为 $O(|E|)$。

证明。 对于 $q(d_j)$，遍历入超边 $e \in \text{In}(v_j)$以查找前驱，每个顶点需要 $O(|\text{In}(v_j)|)$。在稀疏度为 $|\text{In}(v)| \leq O(\log |V|)$的情况下，对 E 进行完整验证的复杂度为 $O(|E|)$。MTC 编织通过相干性保留顺序语义。□

7.7. **容错**。 容错将分区容错扩展至处理 N中的节点故障，确保即使发生硬件或软件故障，运行仍能连续进行。Paxos 或 Raft 等共识协议 [9] 将更新复制到 D，通过法定人数实现一致性，在异步网络中的消息复杂度为 $O(|N| \log |N|)$。这使得以下功能成为可能：

(1) 治理：策略 P将请求重新路由 $r(d_i$至在线节点，借助分布式目录在 $O(|P| \log |N|)$ 时间内完成（第 6.3 节）。(2) 导航： G 通过冗余路径重定向查询，这些路径的计算时间为 $O(|E| + |V| \log |V|)$（定理 8）。(3) 分析：联邦学习从幸存的 θ_n中聚合数据，每轮耗时 $O(|\theta_n||D_n|)$（定理 15）。

超图的关联矩阵优化冗余：

$$\text{rank}(I_T I_H^T) \geq |V| - |N_{\text{failed}}|,$$

确保 $|V| - |N_{\text{failed}}$连通分量，尽管共识延迟 $O(\log |N|)$ 轮次可能与实时界限（第 6.4节）冲突。元数据 M 通过心跳跟踪节点状态，从而能够在 $O(|M|)$ 中进行重路由。MTC 的 S矩阵量化了谱连通性，提示可采用基于特征值的路由进行故障检测（第 5.3节）。

引理14。 当秩为 $\text{rank}(I_T + I_H) \geq |V| - k$ 时，系统可容忍 k 个节点故障，同时保持查询可解析性。

证明。 矩阵 $I_T + I_H$ 表示无向关联图；秩 $\geq |V| - k$ 意味着故障后至多有 k 个不连通分量（根据秩-零化度定理）。查询通过剩余路径得到解析，因为冗余确保存活子图中存在替代路径。□

8. Framework Implementation in Distributed Architectures

本节将数据编织框架 $\mathcal{F} = (D, M, G, T, P, A)$ 应用于$\Sigma = (N, C)$所构成的多组件架构，集成面向行和面向列的数据库、实时分析、搜索索引、变更日志、数据仓库、视图/控制器及转换管道。在第5 节的理论基础（例如向量表示和 MTC 嵌入）之上，我们展示这些组件如何映射到 $\mathcal{F}$，推导优化向量，构建用于操作的超图，并解决可扩展性问题；同时借鉴第 6 节的策略，并确保符合 CAP/CAL 的一致性（第 7节）。图8 和 9展示了这种集成。

i)架构组件及其与数据织体的映射 该架构的组件映射到 $\mathcal{F}$ 元素的方式如下，依据超图G和范畴数据编织：

(1) 行式数据库：存储元组，并映射到具有模式的 D $S_i \subseteq \mathcal{A}$。(2) 列式数据库：管理时间序列，也映射到具有时间模式的 D 。(3) 实时分析数据库：计算聚合，映射到 A和M。(4) 搜索索引：正排/倒排索引映射到 M ，以便快速检索。(5) 变更日志：跟踪变更，映射到 M中的转换历史。(6) 数据仓库：存储历史数据，映射到 D 和 M。(7) 视图和控制器：路由查询，映射到 P。(8) 转换管道：应用 $t_i \in T$，并记录在 M中。

部署在 N 上，并通过链路 C连接，可确保分布式一致性。
该示意图（图 8）展示了映射到 $\mathcal{F}$ 的组件：

(1) 行式/列式数据库、数据仓库： D。

(2) 搜索、变更日志、分析： M, A。

(3) 视图/控制器： P。(4) 管道： T。

ii)**向量表示与维度** 如第 5.1 节所述，数据集和元数据被向量化为 $\mathbb{R}^{11}$，该向量表示由 2 个数值特征、4 个分类特征（独热编码）、3 个时间特征（傅里叶）和 2 个元数据特征（主成分分析）构成。对于销售数据集 $d_i(t)$，$\mathbf{v}_i(t) \in \mathbb{R}^{11}$ 可在 $O(11)$ 中实现相似性搜索。

示例4。 销售与库存的整合使用向量上的余弦相似度，而模张量范畴编织确保顺序不变性。

iii)超图构建及其在操作中的作用 超图 G通过依赖关系（集成、导航，

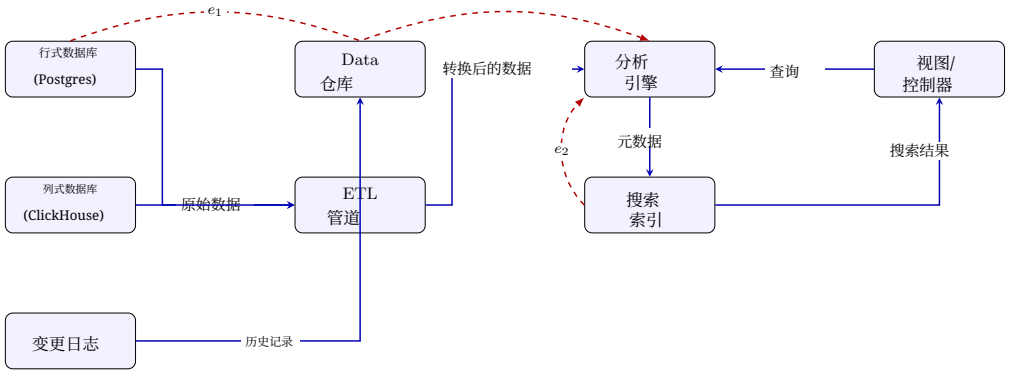
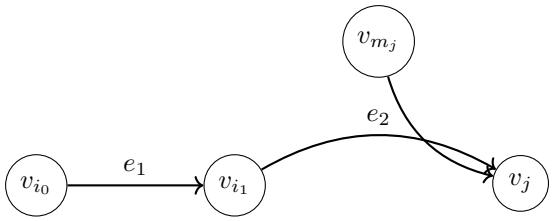


FIGURE 8. 架构组件映射至 $\mathcal{F}$，其中数据流以实线蓝色表示，超边以红色虚线表示。



用于溯源追踪的超边路径。

FIGURE 9. 用于操作的超图连通性。

溯源），确保稀疏性 $|E| = O(|V|$对数 $|V|)$。路径的遍历方式与第 5.2 节相同，复杂度为 $O(|E|+|V|$对数 $|V|)$。

操作的应用方式如下：

(1) 集成：通过态射变换模式。(2) 导航：查询遍历编织路径。

(3) 溯源：利用不变量重建历史记录。

iv)可扩展性与性能考量 分区采用拉普拉斯算子 L 上的谱聚类，复杂度为 $O(|V|\log |V|)$。在 N 上进行并行处理，可将局部计算量降至 $O(|D_n|)$。

引理 15。 遍历复杂度为 $O(|E|+|V|\log |V|)$，且在分区后保持不变。

证明。 采用经调整的 Dijkstra 算法和稀疏边可得此界；分区可保持局部稀疏性。 □

9. CONCLUDING REMARKS

数据编织框架形式化为分布式系统 $\Sigma = (N, C)$ 上的元组 $\mathcal{F} = (D,\ M,\ G,\ T,\ P,\ A)$，为管理异构分布式数据生态系统建立了一种变革性的数学范式。本文已

将丰富的理论进展与实际应用编织在一起，统一了基于超图的连通性、范畴结构、模张量范畴（MTC）、量子群表示，以及受弦理论启发的拓扑学见解。通过综合这些视角，该框架不仅应对现代数据管理的复杂性，还为数据密集型领域的未来创新开辟了一条大胆路径，尤其是在决策、优化和数据分析方面；这些领域与 Fuad Aleskerov 在社会、经济和政治系统数学建模方面的多元贡献相呼应。

该框架的理论贡献核心在于严谨的代数与拓扑基础。超图 $G = (V, E)$ 的顶点 $V = D \cup M$ 代表数据资产和元数据，并通过稀疏关联矩阵编码多元关系，从而支持高效导航与溯源追踪。范畴结构 DF 将数据集建模为对象、将变换建模为态射，为数据集成、元数据驱动导航和联邦学习等操作提供统一语言。MTC 嵌入凭借其编织幺半结构，通过编织作用 c_{X_i,X_j} 捕捉关系对称性，确保运算具有置换不变性。本文各处引入并阐述的这些基础要素，利用谱方法和基于对称性的对齐，解决模式匹配和分区等 NP难问题。

量子群表示的引入，尤其是通过 $U_q(\mathfrak{sl}_2)$ 实现的引入，标志着一项重大进展。通过将数据集建模为有限维模、将变换建模为交织算子，该框架捕捉了分布式系统中固有的非交换依赖。由量子群表示推导出的 MTC 融合规则和 S 矩阵可实现谱优化，将分区的复杂度降至 $O(|V|\log |V|)$，并通过代数对称性增强模式对齐。表示论一节详述的这些见解，为建模动态数据关系和优化计算任务提供了有力工具。此外，这种受量子启发的方法也为在新兴量子数据库中使用模张量范畴提供了理论依据；在这些数据库中，非交换代数结构用于建模数据存储与查询中的量子叠加和纠缠。在量子数据库中，数据集可表示为希尔伯特空间中的量子比特或量子多能级比特，变换则作为保持相干性的酉算子发挥作用。编织幺半群范畴有助于处理量子并行性，从而能够在指数级庞大的状态空间上高效进行查询优化。这对于将数据编织扩展至量子计算环境，以及解决高维数据分析中的优化问题，至关重要。

作为对这一代数方法的补充，弦理论提供了拓扑与几何视角，以全新方式诠释数据织体操作。数据集被建模为 D 膜，模式则作为边界条件，变换则作为连接它们的开弦。超图路径构成编织世界面，辫群B_n 表示 $\pi(\sigma_i) = c_{X_i,\ X_{i+1}}$ 则编码数据流对称性。结不变量（例如琼斯多项式）可检测循环依赖等异常，为流水线优化提供了一种新机制。这些在弦理论章节中探讨的拓扑类比，将异质性转化为几何对齐问题，将可扩展性转化为流形优化任务，从而丰富了

框架的表达能力，并使其与决策模型中的先进优化技术相契合。

在实际应用中，该框架通过多组件架构实现，集成行式和列式数据库、实时分析、搜索索引、变更日志、数据仓库、视图/控制器以及转换管道。该架构利用 $\mathbb{R}^n$ 中作为 $U_q(\mathfrak{sl}_2)$-模和D膜的向量表示，并将超图遍历建模为编织世界面。数据集成和导航等操作受益于交织算子和辫群对称性，其复杂度达到 $O(|E|+|V|\log\ |V|)$。通过量子启发式谱聚类和几何D膜对齐提升可扩展性，同时利用纽结不变量确保稳健的异常检测。这一实际实现展示了该框架弥合理论严谨性与现实适用性的能力，解决计算挑战，并确保分布式环境中的一致性、完备性和因果性，这与Aleskerov开创的集体选择和网络优化数学模型颇为相似。

展望未来，数据编织框架为研究和部署开辟了令人振奋的新方向。动态幺半范畴的集成可用于建模时间治理约束，从而增强实时系统中的策略执行。受弦理论与拓扑量子场论联系启发的拓扑数据分析（TDA），有望通过使用持久同调追踪演化分布，检测流数据中的概念漂移。量子启发式算法可利用 MTC 的 S矩阵和融合规则，在量子计算环境中实现分区和分析的次线性复杂度；还可扩展至量子数据库，其中 MTC 编织将量子电路分解建模，用于容错数据查询。辫群在数据流建模中的作用表明，可以构建可扩展的异常检测框架，适用于物联网、人工智能和云端分析平台中的复杂管道，也可用于优化全球贸易网络或类似 Aleskerov 应用的灾害响应协议。这些方向以论文中的代数和拓扑创新为基础，使数据织体成为下一代数据管理系统的基石。

总之，这项工作代表了分布式数据管理范式的转变，将超图连通性、范畴统一、量子群表示和弦论拓扑融为一体。通过以统一的数学视角处理异质性、可扩展性和实时处理问题，该框架不仅解决了关键的计算挑战，也为决策和优化领域的变革性应用奠定了基础，体现了 Aleskerov 多方面贡献的精神。随着数据生态系统日益复杂，数据织体以其深厚的理论内涵和实用价值脱颖而出，蓄势塑造大规模、数据密集型领域的未来。

References

[1] *Leveraging quantum computing for optimal data allocation in distributed systems*, 《量子计算与量子启发技术在数据密集型系统中的应用》第二届研讨会论文集（ACM）（2025年6月）。[2] Daniel Abadi, *Problems with CAP, and Yahoo's little known NoSQL system*, 2010年。[在线；访问日期：4月22日。 2025]。[3] Steve Awodey, *Category theory*, 第2版, 《牛津逻辑指南丛书》，第52卷，牛津大学出版社，英国牛津，2010年。

[4] Peter Buneman、David A. Holland 和 Ryan Schultz, *Categorical data integration for computational science* **318** 2016 49 65 , 《理论计算机科学电子札记》（ ），–。第七届生命科学数据集成国际研讨会（DILS 2016）论文集。[5] Elira Curri 和 Tony Shaska, *Polynomials Galois groups and database-driven arithmetic*, 《人工智能与数学研究》，2026年，第231–294页。MR5023832[6] Longyu Feng、Huahang Li 和 Chen Jason Zhang, *Prompt-matcher: Leveraging large models to reduce uncertainty in schema matching results* 2408 14507 2024 , arXiv 预印本 arXiv:。（ ）。[7] Seth Gilbert 和 Nancy A. Lynch, *Brewer's conjecture and the feasibility of consistent available partition-tolerant web services* **33** 2002 2, 51 59 , 《SIGACT News》（ ），第 – 期。[8] Jepsen LLC, *Consistency Models*, 2025年。[在线；访问日期：4月23日。 2025]。[9] Martin Kleppmann, *A critique of the cap theorem*, 2015年。[10] Ilias Kotsireas 和 Tony Shaska, *A metrosymbolic framework for geometric reduction of binary forms* 2026, 173 193 , 《人工智能与数学研究》，第 – 页。MR 5023829
[11] Edward A. Lee、Ravi Akella、Soroush Bateni、Shaokai Lin、Marten Lohstroh 和 Christian Menard, *Consistency vs. availability in distributed real-time systems*, arXiv 2301.08906（2023），可在 2301.08906 查阅。[12] Boyuan Li、Shaohui Zhang 和 Qiuying Han, *Federated learning with joint server-client momentum* **15** 2025 99385 , 《科学报告》（《自然》）（5月），。[13] Saunders Mac Lane, *Categories for the working mathematician*, 《数学研究生教材》，第5卷，施普林格出版社，纽约州纽约市，1971年。[14] Tony Shaska, *Graded neural networks*, 《数学科学数据科学国际期刊》（2025年），可在 2502.17751 查阅。[15] *Internalizing tools as morphisms in graded transformers* 2025 2511. （ ），可在 查阅。[16] *Artificial neural networks on graded vector spaces*, 《人工智能与数学研究》，2026年，第1–71页。MR5023824[17] *Graded transformers: A symbolic-geometric approach to structured learning*, 《数学科学数据科学国际期刊》，即将出版（2026年），可在 2507.20108 查阅。[18] David I. Spivak, *Category theory for the sciences*, 麻省理工学院出版社，马萨诸塞州剑桥市，2014年。[19] Jian Tang 和 Kai Ma, *Hypergraph kolmogorov–arnold networks for station level meteorological forecasting* 2025 , 《物理学 A：统计力学及其应用》（ ）。[20] DATAVERSITY 编辑团队, *Data governance trends in 2025*, DATAVERSITY（2024年12月）。[21] GeeksforGeeks 编辑团队, *Data governance in distributed systems*, GeeksforGeeks（2025年7月）。[22] Vladimir G. Turaev, *Quantum invariants of knots and 3-manifolds*, 德格鲁伊特出版社，1994年。[23] Dezhong Yao、Yutong Dai、Zonghao Guo、Xiaokang Zhou、Yudong Ruan、Changsheng Li、Yanzhao Wu 和 Li Jin, *Fedhm: Efficient federated learning for heterogeneous models vialow-rank factorization*, 《人工智能》344（2025年），104333。