在图事务场景中基于无重复语义评估的基于图模式的关联规则

Basil Ell认知交互技术中心，比勒费尔德大学，德国；挪威知识驱动机器学习中心，奥斯陆大学，挪威；bell@techfak.uni-bielefeld.de

arXiv:2512.15308v1 [cs.DB] 17 Dec 2025

摘要

我们提出了适用于资源描述框架图等有向标记多重图的基于图模式的关联规则（GPAR）。GPAR既支持生成任务（扩展图），也支持评估任务（评估图的合理性）。该框架超越了相关形式体系，例如图函数依赖、图实体依赖、关系关联规则、图关联规则、多关系关联规则、路径关联规则和霍恩规则。给定一组图，我们在无重复语义下评估图模式，从而能够更有效地将图的拓扑结构纳入考量。我们定义了一个概率空间，并在概率框架下推导出置信度、提升度、杠杆率和确信度。我们还进一步分析这些度量指标与经典的基于项集的对应指标之间的关系，并确定了在何种条件下能够保留它们的特征性质。

关键词 关联规则 · 有向标记多重图 · RDF 图 · 推理 · 度量指标

1 引言

许多系统都可以自然地表示为图（具体而言，是有向标记多重图），包括社交网络（以个人、群体和组织为节点类型，以友谊、成员关系或影响关系为关系类型）、通信网络（个体、设备和服务器通过消息交换或物理连接关系相连）、引文网络（论文、作者和期刊通过引用、著作关系或出版关系相连）、交通网络（地点、路线和车辆通过服务、连接或换乘关系相连）、贸易网络（国家、企业和商品通过出口或进口关系相连）、生物相互作用网络（基因、蛋白质和复合物通过编码、相互作用或调控关系相连）、分子结构图（原子、分子和官能团通过成键、包含关系或空间邻近关系相连），以及化学反应网络（化合物、反应和催化剂通过输入、输出或催化关系相连）。

除了这些具体系统之外，形式化表示的知识同样可以建模为图。在形式化知识表示中，实体、概念和实例通常对应节点，而它们之间的语义关系则对应类型化边。这种结构是资源描述框架（RDF）等框架以及更广义的知识图谱的基础，其中信息以连接主语、谓词和宾语的三元组形式表达。

即使主要数据本身并非图结构数据——例如自然语言文本中的词元序列——只要将主要数据与注释（例如，文本中何处提及了人物类型的实体，或哪些实体之间表达了何种关系）、元数据（例如，文本提取自哪份文档、由哪位作者创建，或对注释创建方式的描述）以及背景知识（例如，被提及人物的职业，或所表达的关系具有对称性）相结合，就能得到图结构数据。

通过挖掘频繁模式和规则，研究这些图中的规律可能会很有意思。假设我们将两个频繁模式p_1和p_2组合成一条规则，可能会发现这样的情况经常发生：如果图的某个区域匹配模式p_1，那么同一区域也会匹配模式p_2。在评估任务中，如果我们发现p_1匹配而p_2不匹配，这可能表明该区域存在异常。在生成任务中，我们可能希望扩展图，使该区域匹配p_2。
有时我们处理的是单个图，这种情况称为单图设置；而在另一些情况下，我们处理的是一组同类图，这种情况称为图事务设置。例如，我们可能会研究多个不同的社交网络，或者同一社交网络在不同时期的多个快照。同样，我们也可以处理通信网络、引文网络、交通网络、贸易网络、生物网络、分子结构或化学反应网络的集合。此外，我们可以将一组自然语言句子转换为一组图，其中每个句子都附有注释和元数据，从而将语料库转换为一组图。

拥有一组图与拥有单个图有所不同。在某些情况下，我们可以将一组图合并为一个图，再对该图进行分析；而在另一些情况下，将这些图分开保留可能更为合适。例如，假设给定一组自然语言句子，我们将每个句子以及该句子所表达关系的信息表示为一个图，然后把这些图合并成一个单图。接着，我们从这个图中挖掘频繁模式，并计划用这些模式来分类一个表示句子的图所表达的关系，而该句子所表达的关系尚未知晓。我们不希望得到跨越多个句子的模式，因为在分类场景中，给定的只有一个句子。因此，我们要么需要限制模式挖掘，使其“局限于单个句子之内”，要么将这些图分开保留。因此，将图彼此分开可能很重要。

在传统数据挖掘中，诸如“如果一个集合包含集合 A 的所有成员，那么它也将包含集合 B 的所有成员”这样的预测关系，通常表示为关联规则，写作 $A \Rightarrow B$。关联规则在购物篮分析中广为人知，尤其具有吸引力，因为它们本质上易于解释。本文将关联规则的概念从基于项集的关联规则扩展到基于图模式的关联规则。
我们的主要贡献如下：

- 我们提出了有向标记多重图（其结构对应于资源描述框架图）的基于图模式的关联规则概念。

- 我们研究了在无重复语义下评估图模式的情形。与采用标准同态语义进行评估相比，这赋予了图模式更强的判别能力。

- 我们证明，在无重复语义下评估的基于图模式的关联规则可以用概率空间来表述。

- 我们定义了支持度、置信度、提升度、杠杆率和确信度等度量指标。这些指标在基于项集的关联规则中广为人知，既适用于单图设置，也适用于图事务环境；对于后者，我们还区分微平均得分和宏平均得分。

- 我们分析这些度量指标与基于项集的关联规则的经典度量指标之间的关系，指出它们共同的特征，以及它们在捕捉关联方面的差异。

本文结构如下。第2章介绍预备知识，包括基于项集的关联规则、其解释与质量指标、有向标记多重图、图模式以及图模式求值语义。第3章介绍基于图模式的关联规则的概念、其解释和潜在应用，包括生成任务和评估任务。第4章介绍基于图模式的关联规则的度量指标。首先，我们在第4.1节定义基础概率空间。第4.2节介绍单图场景下的度量指标，而第4.3节讨论图事务环境下的度量指标，包括微平均和宏平均两种变体。在第4.4节中，我们展示如何将单图场景和微平均图事务环境下的度量指标求值，重新表述为经典基于项集的关联规则的度量指标。第4.5节详细分析所有度量指标，重点说明其特征与经典度量指标的相似程度。第5章讨论相关工作，第6章总结全文并给出结论。附录包含关于所提出度量指标特征的证明。

2 预备知识

2.1 关联规则及其度量指标

关联规则 $A \Rightarrow B$ 由两个项集 A 和 B 构成，其中 A 和 B 均非空且不相交，并且 A 和 B 都是项集 I 的子集。 A 称为该规则的前件， B 称为后件。为了将这些关联规则与本文稍后介绍的基于图模式的关联规则（GPARs 或 GPA 规则）区分开来，我们在此将它们称为基于项集的关联规则（ISARs 或 ISA 规则）。基于项集的关联规则从一个非空的事务多重集 T（称为事务数据库）中挖掘得出，其中每个事务都是一个项集。基于项集的关联规则常见于购物篮分析 [2]。例如，$\{bread, cheese\} \Rightarrow \{butter\}$ 是一条关联规则。我们将规则 $A \Rightarrow B$ 在某个项集 S 中的含义形式化如下：

$$A \subseteq S \vdash_p B \subseteq S$$

该陈述表示，如果我们观察到 A 是 S 的子集，那么根据关联规则 $A \Rightarrow B$，我们可以概率蕴含地得出，我们能够观察到 B 是 S 的子集。在购物篮分析的语境中，我们可以设想一位正在购物的客户：在结账之前，商品会被添加到购物篮中（但不会从中移除）。由于我们不知道客户在购物过程结束时最终会选取的商品集合 S，因此我们只能根据截至目前已选取的商品进行观察，并预测客户在购物过程结束时会选取哪些商品。因此，如果我们观察到一位客户已将面包和奶酪添加到购物篮中，那么我们会预测该客户将把黄油添加到购物篮中。需要注意的是，该规则并不表示 $A \subseteq S$ 会导致 $B \subseteq S$ 成立。

我们将规则 $A \Rightarrow B$ with $B \subseteq A$ 视为平凡规则。该规则之所以平凡，是因为对于任何满足 $A \subseteq S$ 的项集 S，无需进一步观察 S，尤其无需观察 $S \setminus A$，我们就可以得出结论：$B \subseteq S$，也就是说，只要知道 A 是事务 S 的子集，就足以知道 B 是事务 S的子集；这对任何事务数据库都成立。因此，该规则无法让我们对 S作出非平凡预测。对于任何满足 $A \subseteq S$的项集S， $B \subseteq S$ 这一事实并不取决于 S，而是取决于 A 与 B之间的具体关系，即 $B \subseteq A$。请注意，在实践中，人们会舍弃满足 $A \cap B \neq \varnothing$ 和 $B \not\subseteq A$的规则，因为这些规则存在部分冗余。

在关联规则的研究中，人们提出了多种度量指标。我们介绍并讨论用于评估项集的知名度量指标（即绝对支持度和相对支持度），以及用于评估关联规则的知名度量指标（即置信度、提升度、杠杆率和确信度）。我们假设事务数据库 T非空。非空项集 S 的度量指标：

$$\begin{aligned}
\text{absolute-support}_T: \quad & \mathcal{P}(I) \longrightarrow \mathbb{N}, \\
& S \longmapsto |\{\, t \in T \mid S \subseteq t \,\}| \\
& = \textstyle\sum_{t \in T} \mathbb{1}_{\{S \subseteq t\}} \\
\text{relative-support}_T: \quad & \mathcal{P}(I) \longrightarrow [0,1] \cap \mathbb{Q}, \\
& S \longmapsto \frac{\text{absolute-support}_T(S)}{|T|} \\
& = \frac{\sum_{t \in T} \mathbb{1}_{\{S \subseteq t\}}}{|T|}
\end{aligned}$$

项集 S 的绝对支持度（absolute-support$_T$(S）是包含给定集合 S 中各成员的事务数量；而相对支持度[1]（ relative-support$_T$(S）是事务数据库中包含给定集合 S中各成员的事务所占的比例。通常，人们关注的是找出绝对支持度或相对支持度足够高的项集，然后据此构建规则。这两种度量指标都是反单调的，即对于任意事务数据库T以及满足 $S \subset S'$ 的任意两个集合 S 和 S'，都有 absolute-support$_T$(S) $\geq$ absolute-support$_T$(S')，且 relative-support$_T$(S) $\geq$ relative-support$_T$(S')。这一性质很重要，因为它可以在挖掘频繁项集时剪枝搜索空间：任何包含非频繁子集的项集（即该子集的支持度值低于用户定义的阈值）都不可能是频繁项集。无需评估幂集$\mathcal{P}$(I)中每个项集的支持度；一旦发现某个集合是非频繁的，就可以舍弃该集合幂集中所有作为其超集的成员。这就是 Apriori 原理 [2]。

[1]相对支持度由 Agrawal 和 Srika 提出，称为支持度 nt [2]。

基于项集的关联规则可以用概率空间来表述 [43]。尽管关联规则通常˜以项集和事务计数来定义，但也可以等价地在概率空间 $(\Omega\ \mathcal{F},\ P_T)$ 中表述如下：

- 样本空间 Ω 是事务的 T 多重集，因为我们从事务数据库 T中均匀随机地选取一个事务。
- σ-代数 $\mathcal{F}$ 是幂集 $\mathcal{P}(\Omega)$，因为 Ω 是有限的，因此 Ω 的每个子集都是可测事件。
- 经验概率测度 $\hat{}\,P_T\colon\ \mathcal{F} \to [0,1]$ 针对每个事件 $X \subseteq \Omega$ 定义为 $\hat{}\,P_T(X) = \frac{|X|}{|T|}$。给定一个项集 $A \subseteq \mathcal{I}$，我们将其关联到事件 $E_A\colon = \{\, t \in T \mid A \subseteq t \,\}$。因此，$P_T(E_A)$ 表示对于从 T中均匀随机选取的事务 t，$A \subseteq t$ 为真的概率。经验概率的测量如下：

$$\hat{P}_T(E_A) = \frac{|\{\, t \in T \mid A \subseteq t \,\}|}{|T|}$$

假设事务数据库 T是有限且非空的，则 $(\Omega\ \mathcal{F},\ \hat{}P_T)$ 是一个概率空间，因为它满足柯尔莫哥洛夫公理：

1. 非负性（公理1）。对于所有 $X \subseteq \Omega$、$|X| \geq 0$ 和 $|T| > 0$。因此，$\hat{}\,P_T(A) = \frac{|A|}{|T|}$是一个有理数，因而也是实数。此外，由于 $|A| \geq 0$，我们有 $\hat{}\,P_T(A) \geq 0$和 $\frac{|A|}{|T|} \geq 0$。2. 归一化（公理2）。由于 $\Omega = T$，可得 $\hat{}\,P_T(\Omega) = \frac{|\Omega|}{|T|} = \frac{|\Omega|}{|\Omega|} = 1$。3. 可数可加性（公理3）。

— —

设 $X = (X_1, \ldots, X_m)$ 是 Ω 的一列两两不相交的非空子集。那么

$$\hat{P}_T\left(\bigcup_{X_i \in X} X_i\right) = \frac{\left|\bigcup_{X_i \in X} (X_i)\right|}{|T|} = \frac{\sum_{X_i \in X} |X_i|}{|T|} = \sum_{X_i \in X} \frac{|X_i|}{|T|} = \sum_{X_i \in X} \hat{P}_T(X_i)$$

因此，相对支持度$_T(S)$ 等于 $\hat{}\,P_T(E_S)$。此外，我们可以如下描述经验联合概率和条件概率：

$$\hat{P}_T(E_A \cap E_B) = \quad \hat{P}_T(E_{A \cup B}) = \frac{|\{\, t \in T \mid A \subseteq t \wedge B \subseteq t \,\}|}{|T|}$$
$$\hat{P}_T(E_B \mid E_A) = \quad \frac{|\{\, t \in T \mid A \subseteq t \wedge B \subseteq t \,\}|}{|\{\, t \in T \mid A \subseteq t \,\}|}$$

$\hat{}P_T(E_A \cap E_B)$ 表示根据事务数据库 T 估算的经验概率：从 T 中均匀随机选取一个事务 $\hat{}t$，其中 $A \subseteq t$ 为真且 $B \subseteq t$ 为真。$P_T(E_B \mid E_A)$ 表示根据事务数据库 T 估算的经验条件概率：从 T 中均匀随机选取一个事务 t，在 $A \subseteq t$ 为真的条件下，$B \subseteq t$ 为真。

接下来，我们将介绍基于事务计数以及等价地基于概率的度量指标。为了量化规则中两个项集之间关联的强度（其中第一个项集的支持度为 > 0），引入了置信度度量。[2] 其定义如下：

$$\text{置信度}_T\colon \{\, A \Rightarrow B \mid A, B \in \mathcal{P}(I) \setminus \{\varnothing\} \,\} \longrightarrow [0,1] \cap \mathbb{Q},$$
$$A \Rightarrow B \longmapsto \frac{\text{relative-support}_T(A \cup B)}{\text{relative-support}_T(A)}$$
$$= \frac{|\{\, t \in T \mid A \cup B \subseteq t \,\}|}{|\{\, t \in T \mid A \subseteq t \,\}|}$$
$$= \frac{\sum_{t \in T} \mathbb{1}_{\{A \subseteq t \wedge B \subseteq t\}}}{\sum_{t \in T} \mathbb{1}_{\{A \subseteq t\}}}$$
$$= \hat{P}_T(E_B \mid E_A)$$

[2]置信度度量由 Agrawal 和 Srikant [2] 提出。

置信度是规则作出的正确预测数量（即同时包含前件 A 和后件 B的事务数）占规则作出的所有预测数量（即包含前件 A的规则数）的比例。置信度高的规则更受青睐。需要注意的是，置信度既可以定义为 $A \cup B$ 的相对支持度与 A 的相对支持度之比，也可以定义为$A \cup B$ 的绝对支持度与 A的绝对支持度之比。规则 $A \Rightarrow B$ 的置信度可解释为经验条件概率 ^probability $P_T(E_B \mid E_A)$。

通常，人们关注支持度分数高于某个阈值且置信度分数高于某个阈值的规则，而这些阈值取决于具体应用。然而，高支持度和高置信度的关联规则未必能揭示真正有趣的事件关系 [7]。例如，假设发现规则 $\{butter\} \Rightarrow \{bread\}$ 同时具有足够高的支持度和置信度。但是，如果所分析的事务数据库来自一家几乎所有事务都包含面包的面包店，那么这条规则实际上可能并不有趣，因为无论是否购买黄油，人们都经常购买面包。因此，人们引入了提升度[3] 得分，即用规则的置信度除以后件的相对支持度。

$$
\begin{aligned}
\mathrm{lift}_T: \quad & \{\, A \Rightarrow B \mid A, B \in \mathcal{P}(I) \setminus \{\varnothing\} \,\} \longrightarrow [0, \infty) \cap \mathbb{Q}, \\
& A \Rightarrow B \longmapsto \frac{\mathrm{confidence}_T(A \Rightarrow B)}{\text{relative-support}_T(B)} \\
& = \frac{|\{\, t \in T \mid A \cup B \subseteq t \,\}| \cdot |T|}{|\{\, t \in T \mid A \subseteq t \,\}| \cdot |\{\, t \in T \mid B \subseteq t \,\}|} \\
& = \frac{\hat{P}_T(E_B \mid E_A)}{\hat{P}_T(E_B)} \\
& = \frac{\left[\sum_{t \in T} \mathbb{1}_{\{A \subseteq t \wedge B \subseteq t\}}\right] \cdot |T|}{\left[\sum_{t \in T} \mathbb{1}_{\{A \subseteq t\}}\right] \cdot \left[\sum_{t \in T} \mathbb{1}_{\{B \subseteq t\}}\right]}
\end{aligned}
$$

提升度分数的一个缺点是，它只使用相对数值。这意味着，对于事件 E_B 强烈依赖于事件 E_A 的规则，即使这些事件的绝对支持度可能非常小，它仍会产生较高的数值。因此，一些可能并不十分有趣的规则也可能具有较高的提升度分数。

例如，规则 $\{capers\} \Rightarrow \{anchovies\}$ 和规则 $\{apples\} \Rightarrow \{bread\}$ 可能具有相同的提升度分数，但项集 $\{capers\}$ 的绝对支持度可能远小于项集 $\{apples\}$ 的绝对支持度，而项集 $\{anchovies\}$ 的绝对支持度可能远小于项集 $\{bread\}$ 的绝对支持度。因此，人们可能会对后一个规则比前一个规则更感兴趣。

一种既考虑事件 E_A 和 E_B 之间的相关性，又考虑支持度分数的度量指标是杠杆率，由 Piatetsky-Shapiro [36], 提出，其定义如下:

$$
\begin{aligned}
\mathrm{leverage}_T: \quad & \{\, A \Rightarrow B \mid A, B \in \mathcal{P}(I) \setminus \{\varnothing\} \,\} \longrightarrow \left[-\tfrac{1}{4}, \tfrac{1}{4}\right] \cap \mathbb{Q}, \\
& A \Rightarrow B \longmapsto \text{rel.-support}_T(A \cup B) - \text{rel.-support}_T(A) \cdot \text{rel.-support}_T(B) \\
& = \frac{|\{\, t \in T \mid A \cup B \subseteq t \,\}|}{|T|} - \frac{|\{\, t \in T \mid A \subseteq t \,\}|}{|T|} \cdot \frac{|\{\, t \in T \mid B \subseteq t \,\}|}{|T|} \\
& = \frac{\sum_{t \in T} \mathbb{1}_{\{A \subseteq t \wedge B \subseteq t\}}}{|T|} - \frac{\sum_{t \in T} \mathbb{1}_{\{A \subseteq t\}}}{|T|} \cdot \frac{\sum_{t \in T} \mathbb{1}_{\{B \subseteq t\}}}{|T|} \\
& = \hat{P}_T(E_{A \cup B}) - \hat{P}_T(E_A) \cdot \hat{P}_T(E_B)
\end{aligned}
$$

杠杆率的一个缺点是该度量具有对称性。因此，杠杆率度量反映的是相关性，而非蕴含程度。确信度则是一种非对称度量，对蕴含的方向敏感，由 Brin 等人提出 [8]。确信度度量使用两种失效率:

1. 一方面，我们关注规则 $A \Rightarrow B$ 失效的概率，即经验^ ^失效率：$P_T(E_A \cap \neg E_B)$（其中 $\neg E_B = \Omega \setminus E_B$），它等于$P_T(A)$·置信度$_T(A \Rightarrow \neg B)$，而后者^等于 $P_T(A) \cdot (1 -$ 置信度$_T(A \Rightarrow B))$. 2. 另一方面，我们关注独立性假设下的预期失效率，即事件 E_A 和 E_B （或等价地，事件 E_A 和 $\neg E_B$）的预期失效率。在独立性假设下，^ ^ ^ ^ ^$P_T(E_A \cap \neg E_B) = P_T(E_A) \cdot P_T(\neg E_B) = P_T(E_A) \cdot (1 - P_T(E_B))$，它等于相对支持度$_T(A) \cdot (1 -$ 相对支持度$_T(B))$.

[3] T提升度量由 Brin 等人作为关联 提出。 [8].

规则的确信度定义为独立性假设下预期失效率与经验失效率之比：

$$\begin{aligned}\text{conviction}_T: \quad & \{\, A \Rightarrow B \mid A, B \in \mathcal{P}(I) \setminus \{\varnothing\} \,\} \longrightarrow \mathbb{Q}_{>0} \cup \{\infty\}, \\ & A \Rightarrow B \longmapsto \frac{1 - \text{relative-support}_T(B)}{1 - \text{confidence}_T(A \Rightarrow B)} \\ & = \frac{1 - \hat{P}_T(E_B)}{1 - \hat{P}_T(E_B \mid E_A)} \\ & = \frac{\left(\sum_{t \in T} \mathbb{1}_{\{A \subseteq t\}}\right) \cdot \left(|T| - \sum_{t \in T} \mathbb{1}_{\{B \subseteq t\}}\right)}{|T| \cdot \left(\left[\sum_{t \in T} \mathbb{1}_{\{A \subseteq t\}}\right] - \left[\sum_{t \in T} \mathbb{1}_{\{A \subseteq t \wedge B \subseteq t\}}\right]\right)}\end{aligned}$$

若 $\hat{P}_T(E_B) = 1$，则确信度未定义。当 $\hat{P}_T(E_B \mid E_A) = 1$ 时，或者等价地，当置信度$_T(A \Rightarrow B) = 1$ 时，确信度$_T(A \Rightarrow B) \to \infty$：由于分母从正侧趋近于零，该比值发散至无穷大。若置信度$_T(A \Rightarrow B) = 1$，我们将确信度$_T(A \Rightarrow B)$ 定义为 ∞。

请注意，对于规则 $A \Rightarrow B$ 和非空事务数据库，若 $\hat{P}_T(E_A) = 0$ $\hat{}$，则置信度分数未定义；若 $P_T(E_A) = 0$ 或 $P_T(E_B) = 0$，则提升度分数未定义。然而在实践中，人们会从非空事务数据库中导出项集，并根据支持度非零的项集构造规则。因此，不会出现前件的概率（即 $P_T(E_A)$）或后件的概率（即 $P_T(E_B$）为零的情况；若出现这种情况，置信度分数或提升度分数将无法定义。

2.2 图、图模式与图模式求值语义

令 $\mathcal{T} = \{t_1, t_2, \ldots\}$ 表示一个无限项集。有向标记多重图（DLM） g 是 $\mathcal{T} \times \mathcal{T} \times \mathcal{T}$ 的一个子集。我们将图中的一个元素称为三元组，并将三元组 $(t_i,\ t_j,\ t_k) \in g$ 解释为一条有向标记边，其中边的标签为项 t_j，边的起点是标签为项 t_i 的节点，边的终点是节点 t_k。给定一个图 g，$\mathcal{T}_g \subseteq \mathcal{T}$ 表示在 g 中出现的项集。[4]

我们要求，对于图 g，任意项 $t \in \mathcal{T}_g$ 在 g 中至多对应一个具有该标签的节点——因此，节点由其标签标识。根据我们对 DLM 的定义，对于图中由三元组 $(t_i,\ t_j,\ t_k)$ 创建的边，该边并非由边标签 t_j 标识，而是由三元组 $(t_i,\ t_j,\ t_k)$ 标识。这一要求有两个重要后果：i）图的拓扑结构由三元组集唯一确定；ii）要检验某个图 s 是否包含于某个图 g 中，我们无需执行复杂度更高的子图同构检验，而只需执行较简单的子集测试： $s \subseteq g$。

这一要求的另一个结果是，我们可以定义图关联规则（GAR）的概念，其中规则的形式为 $g_1 \Rightarrow g_2$ 且 g_1 且 g_2 均为 DLM。我们可以应用该规则来概率蕴含：如果图 s包含子图 g_1，即 $g_1 \subseteq s$，那么它也包含子图 g_2，即 $g_2 \subseteq s$：

$$g_1 \subseteq s \vdash_p g_2 \subseteq s$$

在图关联规则的语境中，三元组被视为一个项，因此项集 I 包含三元组，即 $I \subseteq \mathcal{T} \times \mathcal{T} \times \mathcal{T}$。对于每个事务都是一个图的事务数据库，其中的频繁项集挖掘等同于频繁子图挖掘；在 ISAR 环境中定义的度量指标也可应用于 GAR 环境。与 ISAR 类似，形式为 $g_1 \Rightarrow g_2$ 且满足 $g_2 \subseteq g_1$ 的 GAR 可视为平凡规则。此外，如果 $g_1 \cap g_2 \neq \varnothing$ 且 $g_2 \not\subseteq g_1$，则会舍弃该规则，因为它存在部分冗余。

图 1 展示了一个 DLM 示例，它可以出现在自然语言处理语境中，作为句子“ Alice Smith likes Bob and Charlie likes Bob.”的一种丰富表示。该图包含八个词元，每个词元对应一个节点。这些节点以项 t_1、.....、 t_8 作为标签/标识。每个词元节点通过一条后继边与下一个词元相连，例如 $(t_3$、 $next$、 $t_4)$，并通过一条标签边与句子中的实际词元字符串相连。更多信息表示命名实体识别任务的结果。例如，词元 t_1和 t_2 都是提及（pom）的一部分，该提及指向实体（moe）Alice。关系检测任务的结果表明，该句表达了实体 Alice 喜欢实体 Bob。因此，该图表示存在一个喜欢关系（ior）的实例，其中 Alice 是该关系实例（hori）的头实体，Bob 是尾实体（tori）。该关系实例是喜欢关系（ior）的一个实例。更多信息

[4] I.e., $\mathcal{T}_g = \bigcup_{(t_i,t_j,t_k) \in g} \{t_i, t_j, t_k\}$

$g = \{(t_1, label, "Alice"), (t_2, label, "Smith"), (t_3, label, "likes"), (t_4, label, "Bob"), (t_5, label, "and"), (t_6, label, "Charlie"), (t_7, label, "likes"), (t_8, label, "Bob")\ (t_1, next, t_2), (t_2, next, t_3), (t_3, next, t_4), (t_4, next, t_5), (t_5, next, t_6), (t_6, next, t_7), (t_7, next, t_8), (t_1, pom, em_1), (t_2, pom, em_1), (t_4, pom, em_2), (t_6, pom, em_3), (t_8, pom, em_4), (em_1, moe, Alice), (em_2, moe, Bob), (em_3, moe, Charlie), (em_4, moe, Bob), (Alice, hori, ri_1), (Bob, tori, ri_1), (Charlie, hori, ri_2), (Bob, tori, ri_2), (ri_1, ior, likes), (ri_2, ior, likes), (Alice, likes, Bob), (Alice, isA, Person), (Bob, isA, Person), (Charlie, isA, Person)\ \}$

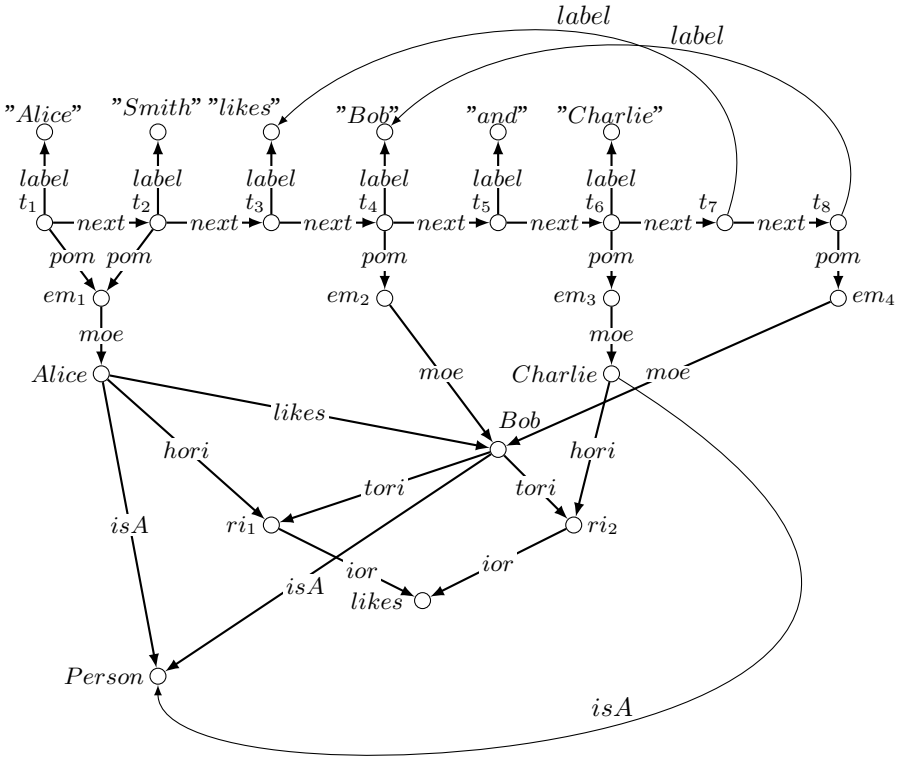


图 1：有向标记多重图（DLM）的示例，展示了其在自然语言处理场景中的一种可能形式。该图表示句子“Alice Smith 喜欢 Bob，Charlie 喜欢 Bob”，包含一些注释（例如提及了哪些实体以及表达了哪些关系）和一些背景知识（例如 Alice 是一个人物，且已知 Alice 喜欢 Bob）。 pom 表示提及部分， em 表示实体提及， moe表示实体提及， $hori$ 表示关系实例头部， $tori$ 表示关系实例尾部， ri 表示关系实例，而 ior 表示关系实例。

这些内容来自背景知识，例如 Alice 属于 Person 类型（isA），以及已知 Alice 喜欢 Bob。该图展示了一种情形：主要数据本身并非图结构化数据，但将数据与注释和背景知识结合后，便可得到图结构表示。

RDF 图是由如下形式的 RDF 三元组组成的集合：$(s,\ p,\ o) \in (\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{I} \cup \mathcal{B} \cup \mathcal{L})$，其中 $\mathcal{I}$、 $\mathcal{B}$ 和 $\mathcal{L}$ 分别是由国际化资源标识符、空白节点和字面值构成的无限且两两不相交的集合。我们可以将 RDF 图的结构表示为 g

忽略 $\mathcal{I}$、$\mathcal{B}$ 和 $\mathcal{L}$ 之间的区别，并令项集合 $\mathcal{T}$ 为 $\mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$，即可将其表示为 DLM。RDF 数据模型[5] 可视为 DLM 的一种特化。在 RDF 图为基 RDF 图的条件下，图关联规则的概念可直接应用于 RDF 图；这意味着这些图不包含空白节点。若存在空白节点，则需要超越子图检查，执行子图同构[6] 和 简单蕴涵[7] 检查，因为空白节点被解释为经过存在量化的变量。

图关联规则存在重要局限。假设我们想从一组与图 1 所示图相似的图中推导出一条规则，该规则用于预测图中是否表达了一个人物喜欢另一个人物。此类规则（并不特别合适）的前件图可能包含以下三元组：(t_6, $next$, t_7)、(t_7, $label$, "$likes$") 和 (t_7, $next$, t_8)。项 t_6、t_7 和 t_8 是句子中词元的标识符。因此，只有在至少包含八个词元且第 7 个词元的标签为"likes"的句子中，才能检测到这一关系。因此，这条规则不具备一般性。此外，该规则没有考虑到：第 7 个词元的正前方和正后方都需要提及一个 Person 类型的实体；但要做到这一点，前件还需要包含 em_3 和 em_4 等项，而这些项是该句子特有的。通过将规则的前件和后件设为图模式，可以获得更好的泛化能力；我们将在下文中对其进行定义。

令 $\mathcal{V} = \{v_1, v_2, \ldots\}$ 表示一个无限变量集合。图模式是 $(\mathcal{T} \cup \mathcal{V}) \times (\mathcal{T} \cup \mathcal{V}) \times (\mathcal{T} \cup \mathcal{V})$ 的子集。我们将图模式中的元素称为三元组模式。给定一个图模式 p，$\mathcal{T}_p \subseteq \mathcal{V}$ 表示其中出现的项的集合，而 $\mathcal{V}_p \subseteq \mathcal{V}$ 表示其中出现的变量集合。[9]

将图模式 p 与图 g 匹配，所得结果是一个映射函数集合 $\Omega_{p,\ g}$，这些映射函数从 $\mathcal{V}_p \cup \mathcal{T}_p$ 映射到 $\mathcal{T}_g$，且每个映射函数都必须满足一组条件。需要满足哪些条件，取决于所选的求值语义。[10] 在最一般的情况下，采用称为同态语义 ***(hom)*** 的图模式求值语义时，求值结果集合 $\Omega^{hom}_{p\ g}$ 的特征如下：

$$\Omega^{hom}_{p,g} = \left\{ \mu \in \{ \mathcal{V}_p \cup \mathcal{T}_p \to \mathcal{T}_g \} \mid \forall t \in \mathcal{T}_p : \mu(t) = t \wedge \mu(p) \subseteq g \right\}$$

这里，$\{ \mathcal{V}_p \cup \mathcal{T}_p \to \mathcal{T}_g \}$ 表示定义域为 $\mathcal{V}_p \cup \mathcal{T}_p$、陪域为 $\mathcal{T}_g$ 的全函数集合。此外，$\mu(t)$ 表示 t 在 μ 下的像，而 $\mu(p)$ 表示将映射函数 μ 应用于图模式 p 中每个三元组模式的每个元素所得的结果，即每个三元组模式 (x_i, x_j, x_k) 都会被替换为 ($\mu(x_i)$ $\mu(x_j)$ $\mu(x_k)$)。我们将 $\mu(p)$ 称为与模式 p 和映射 μ 对应的嵌入。同态语义要求，对于图模式 p 和图 g 下的映射函数 μ，i) μ 将模式中的每个项映射到其自身；ii) 嵌入 $\mu(p)$ 是 g 的子图。

同态语义的一个缺点是，无法区分某些拓扑结构不同的图结构。例如，图模式 $\{(v_1, v_2, v_3)\}$ 可通过 $\mu_1 = \{(v_1, t_1), (v_2, t_2), (v_3, t_3)\}$ 匹配图 $g_1 = \{(t_1, t_2, t_3)\}$（该图不包含自环），也可通过 $\mu_2 = \{(v_1, t_1), (v_2, t_2), (v_3, t_1)\}$ 匹配图 $g_2 = \{(t_1, t_2, t_1)\}$（该图包含自环）。

此外，在同态语义下，嵌入的大小可能小于模式的大小。例如，设 g 为 $\{(t_1, t_2, t_3)\}$，p 为 $\{(v_1, t_2, v_2), (v_3, t_2, v_2)\}$。该模式将匹配包含 $\mu = \{(v_1, t_1), (v_2, t_3), (v_3, t_1)\}$ 的图。嵌入 $\mu(p)$ 为 $\{(t_1, t_2, t_3)\}$。尽管该模式由两个三元组模式组成，嵌入却只包含一个三元组。

为确保每个嵌入的结构与图模式的结构相对应，我们可以使用一种特定的基于同构的图模式求值语义，称为无重复语义（***nra***），其定义如下：

$$\Omega^{nra}_{p,g} = \left\{ \mu \in \Omega^{hom}_{p,g} \mid \forall (x_i, t_i), (x_j, t_j) \in \mu : t_i = t_j \Rightarrow x_i = x_j \right\}$$

在无重复语义下，映射函数必须是单射。因此，i）不能将两个变量绑定到同一项；ii）模式中的变量不能绑定到该模式中出现的项；iii）由 n 个三元组模式组成的模式，只能匹配由 n 个三元组组成的子图。

[5]参见 https://www.w3.org/TR/rdf11-concepts/[6]参见 https://www.w3.org/TR/rdf12-concepts/#graph-isomorphism[7]参见 https://www.w3.org/TR/rdf11-mt/#simpleentailment [8]即，$\mathcal{T}_g = \bigcup_{x_i\ x_j\ x_k \in g} [\{x_i, x_j, x_k\} \cap \mathcal{V}]_{(.\ .\ .)}$即，[9] $\mathcal{V}_p = \bigcup_{x_i\ x_j\ x_k \in g} [\{x_i, x_j, x_k\} \cap \mathcal{T}]_{(.\ .\ .)}$参见 Angles等人 [3]，了解图模式求值语义的概述。[10]

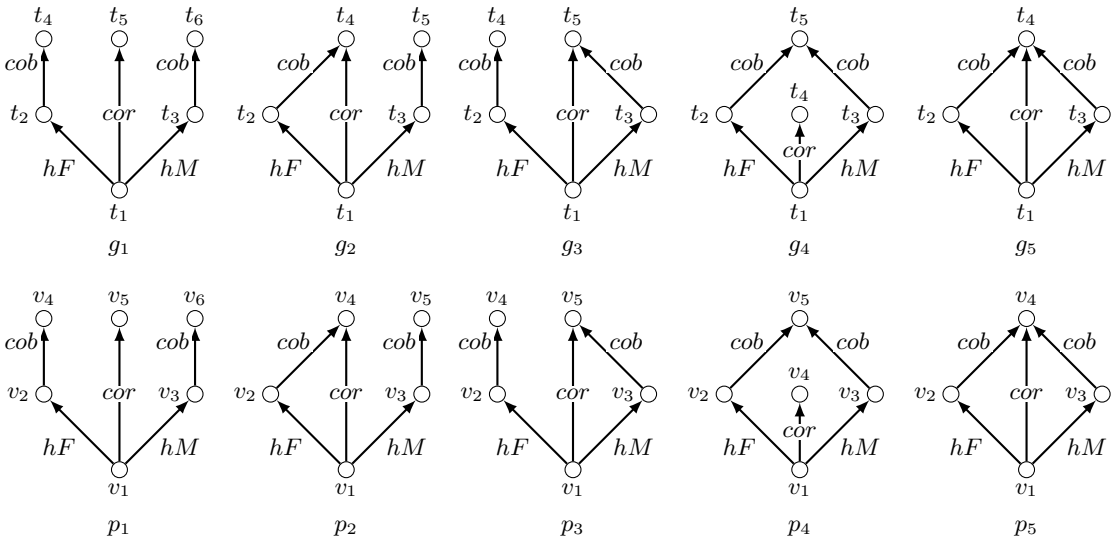


图 2：图 g_1、……、g_5 展示了人物（t_1）、其居住国以及其父亲和母亲的出生国所处的不同情况。边标签 *hF*、*hM*、*cob* 和 *cor* 分别表示 *has father*、*has mother*、*country ofbirth* 和 *country or residence*。通过将节点项替换为变量，可将这些图分别抽象为图模式，从而得到模式 p_1、……、p_5。

为了更好地理解何时使用*nra* 语义可能比使用*hom* 语义更有益，我们来看一个教育社会学领域的任务。我们可能想要了解，一个人是否居住在其父母的出生国，会对其受教育程度产生什么影响。图 2 展示了五种以图表示的情形。边标签*hF*、*hM*、*cob*和*cor* 分别表示父亲、母亲、出生国和居住国。例如，在图 g_1 中，人物（t_1）的父亲（t_2）出生于国家 t_4，母亲（t_3）出生于国家 t_6，而此人居住在国家 t_5。五个图 $g_1, \ldots, g_5$ 中的每一个，都通过将节点位置上的项替换为变量，抽象为一个图模式，由此得到模式 $p_1, \ldots, p_5$。

表1概述了在*hom* 语义（a）和*nra*语义（b）下，这些模式分别与哪些图匹配。例如，在*hom*语义下，p_2 与 g_2 和 g_5匹配；在*nra* 语义下，p_5 仅与 g_5匹配。在*hom* 语义下，大多数模式与多个图匹配，因此无法区分不同的情形。例如，在*hom* 语义下，模式 p_2 与图 g_2 匹配，该图表示人物居住在父亲的出生国而非母亲的出生国；同时，它也与图 g_5 匹配，该图表示人物的居住国与父母双方的出生国均相同。相比之下，在*nra* 语义下，每个模式恰好与其中一个图匹配。如果这些区别对于准确预测人物所获得的教育水平确实相关，那么*hom*语义将是不充分的，因为它无法作出相关区分——打个比方，如果在这种情形下这些区别确实重要，那么基于*hom* 语义下可作出的区分建立的模型，就没有在关键关节处刻画自然。相反，基于*nra* 语义下可作出的区分建立的模型，则在关键关节处刻画了自然。如果在*nra* 语义下可区分的模式比实际需要的更加具体，那么就不存在一个可以据此预测所关注情形的单一模式，而是存在一组模式。此时，模型可以对这些模式作析取，并在其中至少一个模式匹配时预测所关注的情形。

表1：概述在*hom* 语义（a）和*nra* 语义（b）下，模式 $p_1, \ldots, p_5$ 分别与图 $g_1, \ldots, g_5$ 中的哪些图匹配。

(a) *hom* 语义下的匹配

	g_1	g_2	g_3	g_4	g_5
p_1	✓	✓	✓	✓	✓
p_2		✓			✓
p_3			✓		✓
p_4				✓	✓
p_5					✓

(b) *nra*语义下的匹配

	g_1	g_2	g_3	g_4	g_5
p_1	✓				
p_2		✓			
p_3			✓		
p_4				✓	
p_5					✓

在无重复语义下，模式 p的形状与其每个嵌入 $\mu(p)$ 的形状相同。形式化地，给定一个模式 p及其一个嵌入 $\mu(p)$：

$$
\begin{aligned}
&\forall t \in \mathcal{T}_p : \\
&\qquad |\{\, (s,p,o) \in p \mid s = t \,\}| = |\{\, (s,p,o) \in \mu(p) \mid s = t \,\}| \\
&\qquad\quad \text{(i.e., same out-degree for term nodes in pattern and embedding)} \\
&\qquad \wedge |\{\, (s,p,o) \in p \mid o = t \,\}| = |\{\, (s,p,o) \in \mu(p) \mid o = t \,\}| \\
&\qquad\quad \text{(i.e., same in-degree for term nodes in pattern and embedding)} \\
\wedge\, &\forall v \in \mathcal{V}_p : \\
&\qquad |\{\, (s,p,o) \in p \mid s = v \,\}| = |\{\, (s,p,o) \in \mu(p) \mid s = \mu(v) \,\}| \\
&\qquad\quad \text{(i.e., same out-deg. for var. nodes in pattern and corresp. term node in emb.)} \\
&\qquad \wedge |\{\, (s,p,o) \in p \mid o = v \,\}| = |\{\, (s,p,o) \in \mu(p) \mid o = \mu(v) \,\}| \\
&\qquad\quad \text{(i.e., same in-deg. for var. nodes in pattern and corresp. term node in emb.)}
\end{aligned}
$$

因此，如果某个项在模式中的 n 个三元组模式里出现在主语位置（即该节点在模式中的出度为 n），那么该项在嵌入中的 n 个三元组里也出现在主语位置（即该节点在嵌入中的出度为 n）。入度的情况同理。此外，如果变量 v 在模式中的 n 个三元组模式里出现在主语位置（即该节点在模式中的出度为 n），那么相应的项 $\mu(v)$ 在嵌入中的 n 个三元组里也出现在主语位置（即该节点在嵌入中的出度为 n）。入度的情况同理。

需要注意的是，出于实际考虑，映射函数只需将变量映射到项即可。因此，只需满足 $\mu \in \{\mathcal{V}_p \mapsto \mathcal{T}_g\}$，而不必满足 $\mu \in \{\mathcal{V}_p \cup \mathcal{T}_p \to \mathcal{T}_g\}$。由于每个映射函数 $\mu \in \Omega^{nra}_{p\,g}$. 都会将每个项 $t \in \mathcal{T}_p$ 映射到其自身，因此，如果不作此处理，这些项到项的映射就会针对每个映射函数重复存储，造成内存空间浪费。此外，这些项到项的映射可以从 $\mathcal{T}_p$推导出来。因此，实际上，映射函数只存储变量到项的映射。

3 基于GraphPattern的关联规则

3.1 预备考虑

在 ISAR 环境中，一个项集要么是另一个项集的子集，要么不是。当我们转而讨论基于图模式的规则时，会遇到这样一种情况：一个图模式可能与某个图匹配多次。我们可以忽略匹配的实际次数，只关心是否存在匹配。因此，我们要处理的条件是 $\Omega^{nra}_{p,\ g} \neq \varnothing$。这样，对于给定的图 g，我们可以将形式为 $p_1 \Rightarrow p_2$ 的规则解释如下：

$$\Omega^{nra}_{p_1,g} \neq \varnothing \vdash_p \Omega^{nra}_{p_2,g} \neq \varnothing$$

因此，在图 g 与模式 p_1匹配的条件下，我们可以概率蕴含图 g 也与模式 p_2匹配。然而，这一概念存在局限性。我们会错失指定约束的机会，无法规定 p_2 的一次匹配需要如何对应 p_1的一次匹配。例如，我们希望表达：对于 μ_1 的每一次匹配，

p_1 在 g 上（即 $\mu_1 \in \Omega_{p_1,g}^{nra}$），存在 p_2 在 g 上的一次匹配 μ_2（即 $\mu_2 \in \Omega_{p_2,g}^{nra}$），且位于同一位置（这意味着嵌入 $\mu_1(p_1)$ 与 $\mu_2(p_2)$ 相互重叠）。
为了突破这一限制，我们可以定义约束，例如：要使 p_1 和 p_2 的两个匹配相对应，必须满足以下条件：p_1 的匹配 μ_1 中，绑定到某个变量的项，必须等于 p_2的匹配 μ_2 中绑定到某个变量的项。这样，我们便超越了仅仅表达模式之间相关性的范畴，转而描述逻辑蕴含。

3.2 形式化与解释

设 p_1 和 p_2 为图模式，分别具有 $\mathcal{V}_{p_1} \neq \varnothing$ 和 $\mathcal{V}_{p_2} \neq \varnothing$；再设 $V_1 = (v_{1,1}, \ldots, v_{1,n})$ 和 $V_2 = (v_{2,1}, \ldots, v_{2,n})$ 分别为来自 $\mathcal{V}_{p_1}$ 和 $\mathcal{V}_{p_2}$ 的等长、非空且不重复的变量序列。我们将这两个列表中的元素称为连接变量，并将 (p_1, p_2, V_1, V_2) 称为基于图模式的关联规则。我们在某个图 g 的上下文中，将基于图模式的关联规则 (p_1, p_2, V_1, V_2) 的含义形式化如下：

$$\Omega_{p_1,g}^{nra} \neq \varnothing \vdash_p \forall \mu_1 \in \Omega_{p_1,g}^{nra} : \exists \mu_2 \in \Omega_{p_2,g}^{nra} : \mu_1(V_1) = \mu_2(V_2)$$

稍微滥用一下记号，$\mu(V)$ 表示将序列 V 中的每个变量 v 替换为 $\mu(v)$ 后得到的结果，从而形成一个不重复的项序列。规则的含义是：如果前件模式 p_1 匹配图 g（即 $\Omega_{p_1,g}^{nra} \neq \varnothing$），那么我们就概率蕴含，对于前件图模式 p_1 在 g（即 $\forall \mu_1 \in \Omega_{p_1,g}^{nra}$）上的每个匹配 μ_1，都存在后件图模式 p_2 在 g 上的一个匹配 μ_2，使得项序列 $\mu_1(V_1)$ 等于项序列 $\mu_2(V_2)$；这意味着，在两个映射中，对应的变量都绑定到同一个项。因此，借助 $\mu_1(V_1) = \mu_2(V_2)$，我们表达了 p_1 和 p_2 的两个匹配 μ_1 和 μ_2 之间的对应关系。

需要注意的是，该规则并未表达这样的含义：某个 $\mu_1 \in \Omega_{p_1,g}^{nra}$ 会导致 某个 $\mu_2 \in \Omega_{p_2,g}^{nra}$ 满足 $\mu_1(V_1) = \mu_2(V_2)$ 这一条件。

3.3 示例

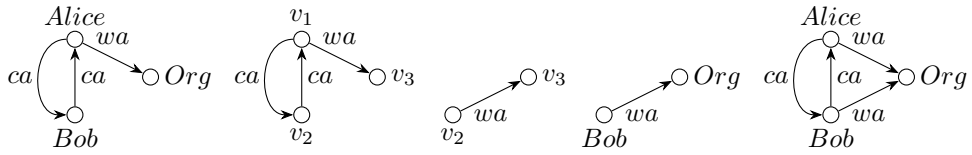
图 3–5 展示了基于图模式的关联规则示例。图 3 所示的规则中，变量 v_2 和 v_3 由模式 p_1 和 p_2 共享。图 g 描述了 Alice 和 Bob 是合著者，且 Alice 在 Org 任职。该规则表示：如果两个实体互为合著者，且其中一个实体与某个对象存在 worksAt 关系，那么另一个实体也与该对象存在 worksAt 关系。模式 p_1 与图 g 恰好匹配一次，得到映射 μ。应用该规则可以为图添加一条边，表示 Bob 在 Org 任职。该边是将映射 μ 应用于模式 p_1（即 $\mu(p_2)$）得到的。其工作原理将在第 3.7 节中解释。

图 4 所示的规则中，变量 v_2 由模式 p_1 和 p_2 共享。图 g 描述了 Alice 认识 Bob，且 Bob 认识 Alice。该规则表示：如果两个实体通过某种关系相互连通，那么这种关系很可能是对称的。模式 p_1 与图 g 恰好匹配一次，得到映射 μ。应用该规则可以为图添加一条边，表示 knows 关系很可能是对称的。请注意，与前一个示例不同，这里使用了一个位于谓词位置的变量。此外，还要注意，项 knows 既作为节点标签出现，也作为边标签出现。

此外，值得注意的是，应根据规则正确性的现有证据为规则评分，然后针对每个用例决定在应用或舍弃规则时所需的最低证据水平。因此，我们在第4节中为基于图模式的关联规则引入度量指标。最后，应用从数据中学习到的规则体现了归纳推理，但从数据中学习到或手动创建的规则也可能实际上是演绎规则。因此，基于图模式的关联规则能够表达演绎规则，并可用于逻辑演绎任务。

在图5所示的规则中，变量 v_1 在模式 p_1 和 p_2之间共享。图 g 表示 m_1 既是分子，也是羧酸。模式 p_1 表示某物的类型为羧酸。模式p_2 描述了羧酸的定义：一个表示分子的节点（v_1）通过含有原子关系连接到四个节点（v_2–v_5），其中 v_2 和 v_4 是氧原子，v_3 是碳原子，v_5 是氢原子。v_2 和 v_3 通过双键相连，v_3 和 v_4 通过单键相连，v_4 和 v_5 通过单键相连。请注意，与前面的示例不同，该规则的后件更为复杂，由多个三元组模式组成；而前两个示例的后件仅由一个三元组模式组成。相比之下，基于项集的关联规则的后件通常恰好包含一项。此外，在那些示例中，情况是$\mathcal{V}_{p_2} \subseteq \mathcal{V}_{p_1}$。虽然图3和图4所示规则的显而易见的应用是链接预测，但图5所示规则的应用可以是合理性分析：任何类型为羧酸的分子都应与后件模式匹配，

$$
\begin{aligned}
p_1 &= \{(v_1, coauthorOf, v_2)、(v_2, coauthorOf, v_1)、(v1, worksAt, v_3)\} \\
p_2 &= \{(v_2, worksAt, v_3)\} \\
V_1 &= (v_2,\ v_3) \\
V_2 &= (v_2, v_3) \\
r &= (p_1, p_2, V_1, V_2) \\
g &= \{(Alice, coauthorOf, Bob)、(Bob, coauthorOf, Alice)、(Alice, worksAt, Org)\} \\
\mu &= \{(v_1,\ Alice),\ (v_2,\ Bob),\ (v_3,\ Org)\} \\
\mu(p_2) &= \{(Bob, worksAt, Org)\} \\
g \cup \mu(p_2) &= \{(Alice, coauthorOf, Bob), (Bob, coauthorOf, Alice), (Alice, worksAt, Org), \\
&\quad (Bob,\ worksAt,\ Org)\}
\end{aligned}
$$



g p_1 p_2 $\mu(p_2)$ $g \cup \mu(p_2)$

图 3：基于图模式的关联规则示例。在可视化图中，项 coauthorOf 简写为 ca，项 worksAt 简写为 wa。

否则，如果该模式不匹配，就会被视为异常。此处，复杂后件的实用性变得更加明显。

3.4 对应关系的类型

本节将介绍图 g中模式 p_1 和 p_2 的两个匹配之间不同类型的对应关系。

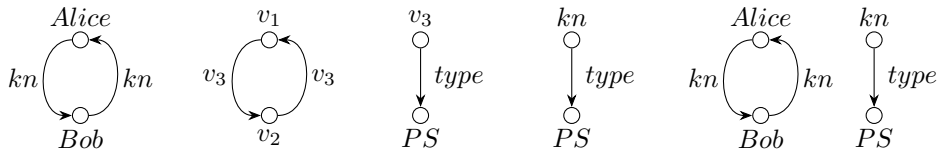
如果存在一个项 $t \in \mathcal{T}_{p_1} \cap \mathcal{T}_{p_2}$，它在两个模式中都出现在节点位置（即三元组模式的主语位置或宾语位置），那么对于每个 $\mu_1 \in \Omega_{p_1,g}$ 和 $\mu_2 \in \Omega_{p_2,g}$，都有 $t \in \mathcal{T}_{\mu_1(p_1)} \cap \mathcal{T}_{\mu_2(p_2)}$。因此，嵌入 $\mu_1(p_1)$与嵌入 $\mu_2(p_2)$重叠，因为它们共享一个节点项。这可以视为两个模式的匹配的嵌入之间的一种先验对应关系，因为该对应关系独立于 μ_1 和 μ_2而存在。我们将探讨两种有趣的非先验对应关系。

1. 给定 g 上 p_1 的一个匹配 μ_1（即 $\mu_1 \in \Omega_{p_1,\ g}^{nra}$）、$g$ 上 p_2 的一个匹配 μ_2（即 $\mu_2 \in \Omega_{p_2,\ g}^{nra}$），以及一对连接变量（$v_1$，$v_2$），其中 v_1 在 p_1 中出现在节点位置，且 v_2 在 p_2 中出现在节点位置，那么，如果两个变量绑定到同一个项，即 $\mu_1(v_1) = \mu_2(v_2)$，嵌入 $\mu_1(p_1)$ 和 $\mu_2(p_2)$ 便具有一个共同节点。对应关系体现在共享一个节点项上。从拓扑上看，这两个嵌入在图中发生了重叠。然而，p_1 和 p_2 的所有嵌入对可能并不都以相同方式对应。

2. 给定 g 上 p_1 的一个匹配 μ_1（即 $\mu_1 \in \Omega_{p_1,\ g}^{nra}$）、$g$ 上 p_2 的一个匹配 μ_2（即 $\mu_2 \in \Omega_{p_2,\ g}^{nra}$），以及一对连接变量（$v_1$，$v_2$），其中 v_1 在 p_1 中仅出现在谓词位置，且 v_2 在p_2中仅出现在谓词位置，那么嵌入 $\mu_1(p_1)$ 和 $\mu_2(p_2)$ 之所以对应，是因为它们都包含一条具有相同标签的边。仅凭这一点不足以构成拓扑重叠。

3. 给定 g 上 p_1 的匹配 μ_1（即 $\mu_1 \in \Omega_{p_1,\ g}^{nra}$），以及 g 上 p_2 的匹配 μ_2（即 $\mu_2 \in \Omega_{p_2,\ g}^{nra}$），并给定一对连接变量（$v_1$，$v_2$），其中 v_1 仅出现在 p_1 的谓词位置，而 v_2 仅出现在 p_2的节点位置（如图4所示规则的情况），那么嵌入 $\mu_1(p_1)$ 和 $\mu_2(p_2)$ 在以下意义上对应：$\mu_1(v_1) = \mu_2(v_2)$，其中 $\mu_1(v_1)$ 是一条边，$\mu_2(v_2)$ 是一个节点。仅凭这一点还不足以构成拓扑重叠。

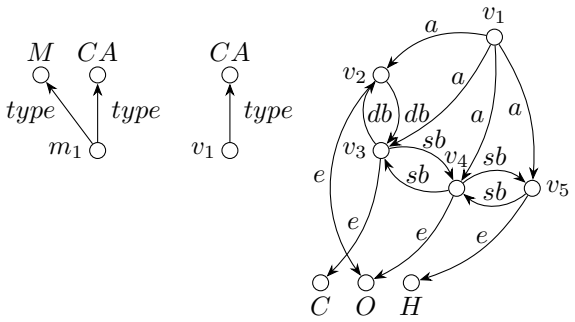
$$
\begin{aligned}
p_1 &= \{\ (v_1, v_2, v_3)\ ,\ \ (v_3, v_2, v_1)\ \} \\
p_2 &= \{(v_3,\ \ type,\ \ ProbablySymmetric)\} \\
V_1 &= (v_3) \\
V_2 &= (v_3) \\
r &= (p_1, p_2, V_1, V_2) \\
g &= \{(Alice, knows, Bob), (Bob, knows, Alice)\} \\
\mu &= \{(v_1, Alice), (v_2, knows), (v_3, Bob)\} \\
\mu(p_2) &= \{(knows,\ \ type,\ \ ProbablySymmetric)\} \\
g \cup \mu(p_2) &= \{(Alice, knows, Bob), (Bob, knows, Alice), (knows, type, ProbablySymmetric)\}
\end{aligned}
$$



g p_1 p_2 $\mu(p_2)$ $g \cup \mu(p_2)$

图4：基于图模式的关联规则示例。在可视化图中，项“认识”缩写为 kn，项“可能对称”缩写为 PS。

$$
\begin{aligned}
p_1 &= \{(v_1,\ \ type,\ \ CarboxylicAcid)\} \\
p_2 &= \{(v_1, hasAtom, v_2), (v_1, hasAtom, v_3), (v_1, hasAtom, v_4), (v_1, hasAtom, v_5), \\
&\quad (v_2, doubleBond, v_3), (v_3, doubleBond, v_2), (v_3, singleBond, v_4), (v_4, singleBond, v_3), \\
&\quad (v_4, singleBond, v_5), (v_5, singleBond, v_4), (v_2, element, O), (v_3, element, Carbon), \\
&\quad (v_4, element, Oxygen), (v_5, element, Hydrogen)\} \\
V_1 &= (v_1) \\
V_2 &= (v_1) \\
r &= (p_1, p_2, V_1, V_2) \\
g &= \{(m_1, type, Molecule), (m_1, type, CarboxylicAcid)\}
\end{aligned}
$$



g p_1 p_2

图5：基于图模式的关联规则示例。在可视化中，项缩写如下：Molecule 缩写为 M，羧酸缩写为 CA，含有原子缩写为 a，双键缩写为 db，单键缩写为 sb，元素缩写为 e，碳缩写为 C，氧缩写为 O，Hygrogen 缩写为 H。

3.5 平凡规则

根据对基于图模式的关联规则（p_1, p_2, V_1, V_2）的解释，对于每个图 $g \in G$，对于每个映射函数 $\mu_1 \in \Omega^{nra}_{p_1,g}$，都存在一个对应的映射函数 $\mu_1 \in \Omega^{nra}_{p_1,g}$，其中“对应”意味着 $\mu_1(V_1) = \mu_2(V_2)$。与平凡的基于项集的关联规则类似，一旦我们知道 $\mu_1 \in \Omega^{nra}_{p_1,g}$，就可以得出结论：无需进一步观察 g，尤其无需观察 $g \setminus \mu_1(p_1)$，便存在一个对应的映射 $\mu_2 \in \Omega^{nra}_{p_2,g}$。因此，该规则无法让我们对 g 作出非平凡预测。对于任何存在 μ_1 的图，都存在一个对应的 μ_2；这一事实并不取决于 g，而是取决于 p_1 与 p_2 之间的特定关系。在 GPAR 环境中，决定规则是否平凡的 p_1 与 p_2 之间的关系，比在 ISAR 环境中稍微复杂一些。我们将先介绍几种特定类型的关系，然后再给出关于使规则变得平凡的 p_1 与 p_2 之间关系的一般论述。所有类型的关系都有一个共同点，即以下情况成立：

$$\forall g \in \mathcal{G} : \forall \mu_1 \in \Omega^{nra}_{p_1,g} : \exists \mu_2 \in \Omega^{nra}_{p_2,\mu_1(p_1)} : \mu_1(p_1) \subseteq \mu_2(p_2)$$

其中 $\mathcal{G}$ 是图的无限集合。注意，p_2 并非针对 g 进行评估，而是针对 g 的一个子图进行评估，即与匹配 μ_1 相对应的 p_1 的嵌入 $\mu_1(p_1)$。因此，$\mu_2 \in \Omega^{nra}_{p_2,\ \mu_1(p_1)}$。$p_2$ 在 g 上的匹配 μ_2 所对应的嵌入（即 $\mu_2(p_2)$），是 p_1 在 g 上的匹配 μ_1 所对应嵌入（即 $\mu_2(p_2)$）的子图。因此，g 中未包含在 p_1 的任何嵌入中的部分都无关紧要。我们先验地知道，如果 g 的一个子图是与 p_1 的某个匹配相对应的嵌入，那么该子图就与 p_2 匹配。只要知道 μ_1 是一个匹配，就足以确定存在相应的匹配 μ_2。

平凡规则最简单的一种情况是满足 $p_1 = p_2$ 的规则。鉴于 $p_1 = p_2$，我们可以将规则为平凡规则的条件改写为：

$$\forall g \in \mathcal{G} : \forall \mu_1 \in \Omega^{nra}_{p_1,g} : \exists \mu_2 \in \Omega^{nra}_{p_2,g} : \mu_1(p_1) \subseteq \mu_2(p_2)$$

to

$$\forall g \in \mathcal{G} : \forall \mu_1 \in \Omega^{nra}_{p_1,g} : \exists \mu_2 \in \Omega^{nra}_{p_1,g} : \mu_2(p_1) \subseteq \mu_1(p_1)$$

这一表达　式在 $\mu_1 = \mu_2$ 时成立；这之所以可能，是因为 $\Omega^{nra}_{p_1,\ g} = \Omega^{nra}_{p_2,\ g}$。鉴于 $\mu_1 = \mu_2$，我们得到

$$\forall g \in \mathcal{G} : \forall \mu_1 \in \Omega^{nra}_{p_1,g} : \mu_1(p_1) \subseteq \mu_1(p_1)$$

这显然成立。

请注意，在目前关于平凡规则的讨论中，我们忽略了变量序列 V_1 和 V_2。这些序列能够强制两个映射之间建立特定的对应关系。然而，它们并不影响这样一个事实：通过针对与 p_1 匹配相对应的嵌入来评估 p_2，可以获得 p_2 的匹配。因此，变量序列对于区分平凡规则与非平凡规则并无影响。

除了 $p_1 = p_2$ 的情况（这会导致这样一种情形：只要知道 μ_1 是一个匹配，就足以知道存在相应的匹配 μ_2）之外，如果 p_2 与 p_1 子图同构（其中 $p_2 \subseteq p_1$ 是一种特殊情况），且 p_2 是 p_1 的（部分）实例化，或上述情况的组合，我们也处于同样的情形。

通俗地说，如果我们能通过单射一致地重命名 p_2 中的变量，使结果与 p_1的某个子图完全相同，那么 p_2 就与 p_1 子图同构。形式上，如果存在一个全单射函数 $m_1 \in \{\mathcal{V}_{p_2} \mapsto \mathcal{V}_{p_1}\}$，使得 $m_1(p_2) \subseteq p_1$，那么 p_2 就与 p_1 子图同构；其中，$m_1(p_2)$表示按照 m_1替换 p_2中的每个变量后得到的结果。

以下规则（此处省略变量序列，因为它们与此处讨论无关）是一个平凡规则的示例，其中 p_2 与 p_1子图同构，因为存在一个具备必要性质的变量到变量的重命名函数 m_1：

$$\begin{aligned} p_1 &= \{(v_1, t_1, v_2), (v_2, t_2, v_3)\} \\ p_2 &= \{(v_1, t_2, v_2)\} \\ m_1 &= \{(v_1, v_2), (v_2, v_3)\} \\ m_1(p_2) &= \{(v_2, t_2, v_3)\} \end{aligned}$$

非正式地说，如果能够一致地使用单射，将模式 p_2 中的项替换为变量（这些变量尚未出现在 p_2 中），使两个模式变得完全相同，那么模式 p_2 就是模式 p_1 的实例化。形式化地说，如果存在一个映射函数 $m_2 \in \{\mathcal{T}_{p_2} \mapsto \mathcal{V}_{p_1} \setminus \mathcal{V}_{p_2}\}$，使得 $m_2(p_2) = p_1$，那么 p_2 就是 p_1 的实例化；其中，$m_2(p_2)$ 表示根据 m_2 替换 p_2 中的项后得到的结果。

以下规则是一个平凡规则的示例，其中 p_2 是 p_1 的实例化，因为存在一个具有所需性质的项到变量重命名函数 m_2：

$$\begin{aligned} p_1 &= \{(v_1, t_1, v_2)\} \\ p_2 &= \{(v_1, t_1, t_2)\} \\ m_2 &= \{(t_2, v_2)\} \\ m_2(p_2) &= \{(v_1, t_1, v_2)\} \end{aligned}$$

最后，将两类关系结合起来，非正式地说，如果后件是某个模式的实例化，且该模式与前件子图同构，则该规则是平凡的。形式上，规则 (p_1, p_2, V_1, V_2) 是平凡的，当且仅当存在一个全单射函数 $m \in \{(\mathcal{V}_{p_2} \cup \mathcal{T}_{p_2}) \mapsto (\mathcal{V}_{p_1} \setminus \mathcal{T}_{p_1})\}$，其中项可以映射为变量，也可以映射为其自身，使得 $m(p_2) \subseteq p_1$；其中 $m(p_2)$ 表示根据 m 替换 p_2 中每个变量和每个项后的结果。

以下规则是一个平凡规则的示例，其中 p_2 是某个模式的实例化，且该模式与 p_1 子图同构，因为存在一个具有所需性质的项与变量重命名函数 m：

$$\begin{aligned} p_1 &= \{(v_1, t_1, v_2), (v_2, t_2, v_3)\} \\ p_2 &= \{(v_1, t_2, t_3)\} \\ m &= \{(v_1, v_2), (t_2, t_2), (t_3, v_3)\} \\ m(p_2) &= \{(v_2, t_2, v_3)\} \end{aligned}$$

3.6 简化记法

到目前为止，我们已经描述了一个基于图模式的关联规则，其中包含两个图模式 p_1 和 p_2，以及两个非重复变量序列 V_1 和 V_2。这两个图模式可能是通过两个独立的挖掘过程推导出来的。因此，两个变量 $v_i \in \mathcal{V}_{p_1}$ 和 $v_j \in \mathcal{V}_{p_2}$ 相同或不同，可能并不携带任何信息。于是，通过序列 V_1 和 V_2，我们可以建立变量之间的对应关系，并遍历 V_1 和 V_2 的所有可能序列，从而遍历我们可以建立的所有可能对应关系。不过，我们也可以重命名模式中的变量，然后假设两个变量 $v_i \in \mathcal{V}_{p_1}$ 和 $v_j \in \mathcal{V}_{p_2}$ 相同或不同是有意义的。

给定形式为 (p_1, p_2, V_1, V_2) 的基于图模式的关联规则 r，我们可以通过以下两个步骤将其改写为形式为 (p_1, p_2'') 的基于图模式的关联规则 r'。

1. 步骤1。我们需要注意，在重写后的规则 r 中，不应出现某个变量同时出现在两个模式中，而原始规则 r 未表达这种对应关系。因此，我们将 p_2 转换为 p_2'。令 $V_o = \mathcal{V}_{p_1} \cap (\mathcal{V}_{p_2} \setminus V_2)$[11] 表示这样一个变量集合：这些变量同时出现在 p_1 和 p_2 中，但未通过 V_2 表达对应关系。对于每个变量 $v \in V_o$，使用一个既不出现在 p_1 中、也不出现在 p_2' 中的新变量姓名，并将 v 在 p_2 中的每次出现替换为该新姓名。2. 步骤2。我们需要根据通过 V_1 和 V_2 表达的对应关系来替换 p_2' 中的变量，使两个模式中名称相同的变量表示对应关系。因此，我们将 p_2' 转换为 p_2''。给定 $V_1 = (v_{1,1}, \ldots, v_{1,n})$ 和 $V_2 = (v_{2,1}, \ldots, v_{2,n})$，令 $V_1 \triangledown V_2$ 表示集合 $\{(v_{1,i}, v_{2,i}) \mid v_{1,i} \in V_1, v_{2,i} \in V_2, 1 \leq i \leq n\}$。对于每一对 $(v_i, v_j) \in V_1 \triangledown V_2$，我们将 v_j 在 p_2'' 中的每次出现替换为 v_i。对 p_2' 的转换结果为 p_2''。

执行上述两个步骤后，我们将规则 (p_1, p_2, V_1, V_2) 转换为 (p_1, p_2'')。图 6 展示了规则 r 以及将其化简为 r'的一种可能结果。请注意，此处定义的两个步骤并不能保证结果唯一，因为新变量的命名方式并未指定。图的左侧给出了初始规则 r，并展示了模式 p_1 和 p_2。根据连接变量对所确定的模式变量对应关系，以虚线表示，并标注建立对应关系的连接变量对。图的右侧给出了所得规则 r'，并展示了模式 p_1 和 p_2''。虚线表示两个模式之间通过同名变量建立的对应关系。在此例中，$V_o = \mathcal{V}_{p_1} \cap (\mathcal{V}_{p_2} \setminus V_2) = \{v_1, v_2, v_3\} \cap (\{v_1, v_2, v_3, v_4\} \setminus \{v_2, v_3\}) = \{v_1\}$。因此，在步骤1中，我们需要将 v_1 替换为 p_2 中的一个新变量。我们将其替换为 v_5。由此得到图模式 $p_2' = \{(v_5, t_3, v_2), (v_2, t_4, v_3), (v_3, t_5, v_4)\}$。接着，在步骤2中，按照 $V1 \triangledown V2$，我们将 v_3 替换为 v_2，并将 v_2 替换为 v_3，从而得到 $p_2'' = \{(v_5, t_3, v_3), (v_3, t_4, v_2), (v_2, t_5, v_4)\}$。因此，我们得到了规则 $r' = (p_1, p_2'')$。

[11] 我们稍微宽泛地使用记号，偶尔将变量序列视为变量集合。

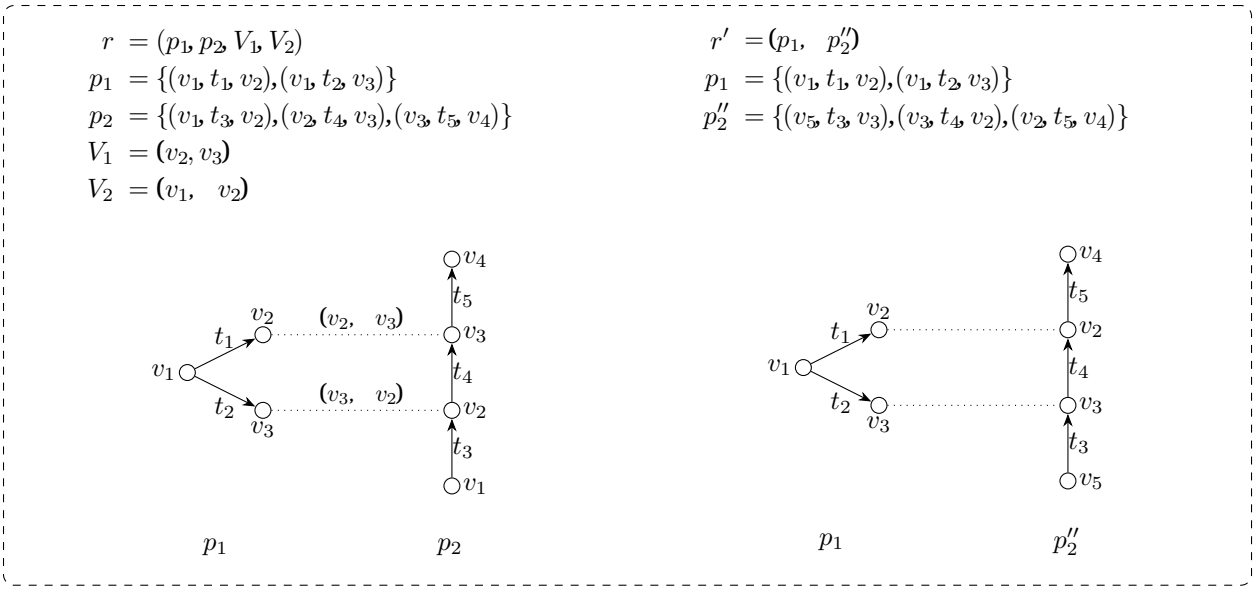


图 6: 形式为 $r = (p_1, p_2, V_1, V_2)$ 的规则示例，以及将其重写为 $r' = (p_1, p_2''$ 的结果).

我们在某个图 g的语境下，将形式为 (p_1, p_2) 的基于图模式的关联规则形式化定义如下:

$$\Omega^{nra}_{p_1,g} \neq \varnothing \vdash_p \forall \mu_1 \in \Omega^{nra}_{p_1,g} : \exists \mu_2 \in \Omega^{nra}_{p_2,g} : \forall v \in \mathcal{V}_{p_1} \cap \mathcal{V}_{p_2} : \mu_1(v) = \mu_2(v)$$

请注意，对于图 3–5 中展示的示例规则，后件模式中的变量采用了便于理解规则的命名方式，因此对应的变量名称相同。在这些示例中，给定规则 (p_1, p_2, V_1, V_2)，只需省略两组变量序列，即可得到其简化形式 (p_1, p_2)。

3.7 应用

我们简要介绍基于图模式的关联规则如何用于两类任务：生成任务和评估任务。在生成任务中，需要生成图可能的扩展；而在评估任务中，则根据图的概率为其打分。

请注意，这与语言模型 [28] 类似：语言模型既可以通过预测最可能的后续内容来扩展词元序列，也可以预测词元序列的概率。

3.7.1 生成任务

给定一个图 g 和形式为 (p_1, p_2) 的规则，其中 p_1 与该图匹配（即 $\Omega^{nra}_{p_1,g} \neq \varnothing$）。考虑 p_2 中的所有变量都是连接变量的情况（即 $\mathcal{V}_{p_2} \subseteq \mathcal{V}_{p_1}$）。那么，对于每个 $\mu \in \Omega_{p_1,g}$，都有 $\mu(p_2)$ 不包含变量，因为所有变量都已通过 μ 替换。因此，我们通过以下操作 e_r得到一个（可能的）扩展图 $g' = e_r(g)$（包含 $|g'| \geq |g|$）：

$$e_r(g) = g \cup \bigcup_{\mu \in \Omega^{nra}_{p_1,g}} \mu(p_2)$$

图经由该操作扩展后，还可能使用相同的操作进一步扩展，如图 7 所示。给定规则 r，令 g^0_r 为 g，并令 g^n_r（包含 $n > 1$）为扩展图 g^{n-1}_r 后的结果，即:

$$g^n_r = e_r(g^{n-1}_r)$$

于是，我们可以将图 g 在规则 r 下的闭包 $Cl_r(g)$ 定义如下:

$$Cl_r(g) = \bigcup_{n \in \mathbb{N}} g^n_r$$

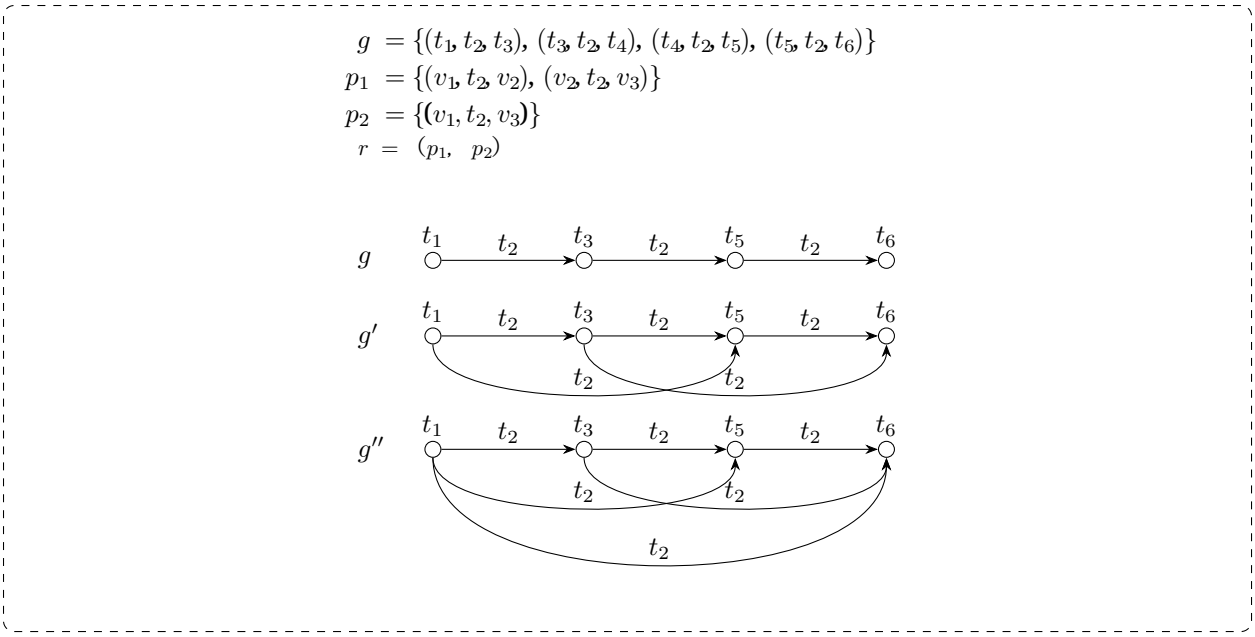


图 7：一个示例，展示了如何使用一条规则经过两步迭代扩展一个图。该规则以简化形式给出，即 (p_1, p_2)。

由于我们只考虑有限图和基于有限图模式的规则，因此经过有限步后即可达到不动点。不过请注意，在实践中，达到不动点未必是目标。相反，我们可能只关注特定区域内的特定扩展，而非所有区域中的所有可能扩展。对于图 3 和图4 所示的规则，$\mathcal{V}_{p_2} \subseteq \mathcal{V}_{p_1}$。因此，对于每个 $\mu \in \Omega^{nra}_{p_1,\, g}$，$\mu(p_2)$ 都是一个图。因此，我们可以使用这些规则将 g 扩展为 $g \cup \mu(p_2)$。

如果 p_2 包含更多变量，那么对于每个 $\mu \in \Omega^{nra}_{p_1,\, g}$ 都有 $\mu(p_2)$ 是一个包含变量的图模式，而扩展 g' 不会是图，而是图模式。不过，这并不意味着这些规则对生成任务没有意义。只是需要引入另一个步骤，将剩余变量替换为项，从而得到一个图。此外，如果要将图迭代扩展为图模式，就需要扩展图模式求值的定义，使图模式能够针对另一个图模式进行求值。图5所示的规则就是这种情况。

我们在生成任务概述中未涉及的是如何选择规则。我们不仅需要选择前件模式与图匹配的规则，还可能希望选择能够扩展图中特定区域的规则（因此需要对前件模式设置约束）、能够以特定方式扩展图的规则（因此需要对后件模式设置约束），或者根据质量指标选择规则。第4章将介绍基于图模式的关联规则的质量指标。对于具体用例，需要研究如何根据这些度量指标得出的得分来选择规则。

知识图谱中的链接预测是一项广为人知的任务，旨在推断知识图谱中实体之间缺失或可能存在的关系。这里区分三种链接预测任务：头项预测（HTP）、关系项预测（RTP）和尾项预测（TTP），具体如下：

- (HTP) 给定知识图谱 g 和不完整三元组 $(?, r, t)$，其中 r 是关系，t 是尾实体（目标）项，任务是识别最可能的头实体（源）项 $s \in V$，其中 V 是项的词汇表。
- (RTP) 给定知识图谱 g 和不完整三元组 $(s, ?, t)$，其中 s 是源术语，t 是尾实体（目标）项，任务是识别最可能的关系项 $r \in V$，其中 V 是项的词汇表。
- (TTP) 给定知识图谱 g 和不完整三元组 $(s, r, ?)$，其中 s 是源术语，r 是关系项，任务是识别最可能的尾实体（目标）项 $t \in V$，其中 V 是项的词汇表。

对于这些任务中的每一项，将预测三元组（s，r，t）添加到 g 中，都能补充图中缺失的、合理的新信息。

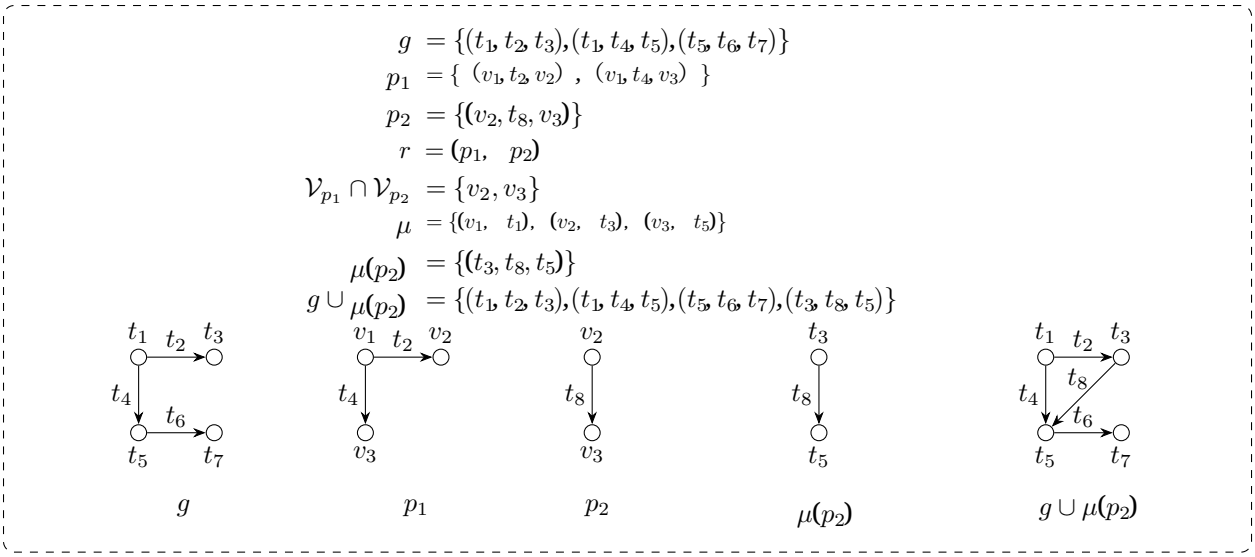


图 8：尾项预测（TTP）示例，其中预测的尾节点项（即 t_5）不包含在 $\mathcal{T}_{p_2}$ 中。规则以简化形式给出，即（p_1, p_2）。

请注意，近期基于规则的方法（如SAFRAN [35] 和AnyBURL [33] ）在链接预测任务上取得的结果，已可与基于神经嵌入的机器学习模型的方法相媲美。

在基于图模式的关联规则进行链接预测的情境中，我们可以区分两种情况：规则中包含要插入不完整三元组的项（例如，规则中包含项 t，这样我们就可以将 $(s, r, ?)$ 补全为 (s, r, t)）；或者规则中不包含该项。因此，如果给定特定词汇表，并希望针对其中的每个项从大量规则中选出能够添加该项的规则，那么仅使用后件模式中包含该项的规则可能并不足够。

图 8 展示了一个使用图 g 和规则进行尾项预测的示例。给定 $\mu = \{(v_1, t_1), (v_2, t_3), (v_3, t_5)\} \in \Omega^{nra}_{p_1, g}$，通过将 $\mu(p_2) = \{(t_3, t_8, t_5)\}$ 添加到 g，我们用三元组 (t_3, t_8, t_5) 扩展 g；因此，预测的尾项（目标项）是 t_5。在此示例中，t_5 不包含在 p_2 中（即 $t_5 \notin \mathcal{T}_{p_2}$）。这表明，即使目标项不包含在后件模式中，相应规则仍可用于将该项添加到图中以扩展图。图 9 展示了一个使用图 g 和规则进行关系项预测的示例。给定$\mu = \{(v1, t2), (v2, t3), (v3, t5), (v4, t6), (v5, t7)\} \in \Omega^{nra}_{p_1, g}$，通过将 $\mu(p_2) = \{(t_3, t_8, t_7)\}$ 添加到 g，我们用三元组 (t_3, t_8, t_7) 扩展 g；因此，预测的关系项是 t_8。在此示例中，t_8 包含在 p_2 中（即 $t_8 \in \mathcal{T}_{p_2}$）。对于 HTP、RTP 和 TTP 链接预测任务，后件模式仅由一个三元组模式构成即可满足要求。然而，即使在生成任务的情境中，允许使用更复杂的后件也可能很有用，这样就可以通过较大的步骤扩展图。例如，对于图 1 所示的这类图，可以设想一个前件，用于识别某个关系是在已识别实体中的哪些实体之间表达的，然后一步将一个新的关系实例添加到图中，并描述该关系实例的头实体和尾实体。

3.7.2 评估任务

我们提出了一些关于如何根据一组 G 由一组图推导出的、形式为（p_1，p_2）的规则来预测给定图 g 的概率的思路。

我们利用以下性质：

- p_1 与 G 中的许多/少数图匹配（即考虑到 $|\{g' \in G \mid \Omega^{nra}_{p_1, g'} \neq \varnothing\}|$）。
- 如果 p_1 与图 $g' \in G$匹配，那么它经常/很少匹配（即考虑到 $|\Omega^{nra}_{p_1, g'}|$）。
- 规则所做的预测通常/很少已经包含在图 $g' \in G$中。如果预测 $\mu(p_2)$不包含变量，那么当 $\mu(p_2) \subseteq g'$时，该预测包含在图 $g' \in G$ 中；如果预测 $\mu(p_2)$包含变量，那么当 $\Omega^{nra}_{\mu(p_2), g'} \neq \varnothing$时，该预测“包含”在图 $g' \in G$中。

随后可以将这些属性用于预测图的概率：

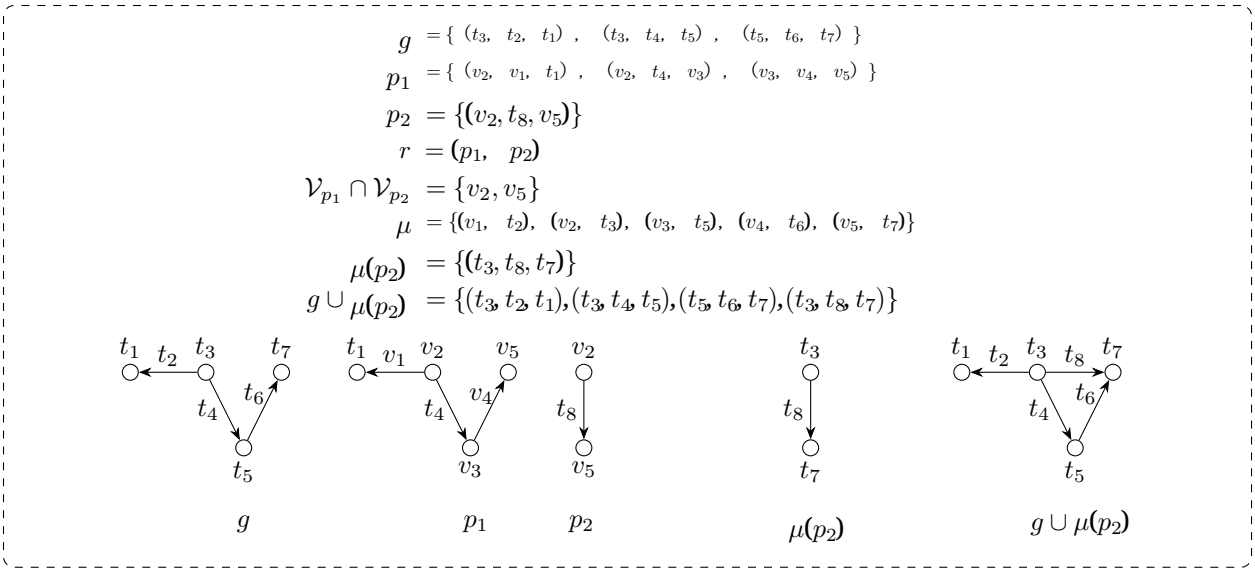


图9：关系项预测（RTP）示例，其中预测的关系项（t_8）包含在 $\mathcal{T}_{p_2}$中。规则以简化形式给出，即（p_1， p_2）。

1. 如果某条规则中， p_1 与 G 中的许多图匹配，但与 g 不匹配，则可能表明 g 的概率较低。2. 如果某条规则中， p_1 与 G中的许多图匹配，并且与 g 匹配，则可能表明 g 的概率较高。3. 如果某条规则中， p_1 经常与 G 中的许多图匹配，但不与 g 匹配或很少匹配，则可能表明 g 的概率较低。4. 如果某条规则中， p_1 经常与 G 中的许多图匹配，并且也经常与 g 匹配，则可能表明g 的概率较高。5. 如果某条规则经常与 G 中的图匹配，并且该规则所作的预测经常出现在这些图中，那么，如果 p_1 与该图匹配，但预测内容未包含在图中，这可能表明 g 的概率较低。6. 如果某条规则经常与 G 中的图匹配，并且该规则所作的预测经常出现在这些图中，那么，如果p_1 与该图匹配，且预测内容包含在图中，则这可能表明 g 的概率较高。

4 基于图模式的关联规则的度量指标

本节介绍基于图模式的关联规则的置信度、提升度、杠杆率和确信度，并说明这些度量指标在多大程度上与基于项集的关联规则中的对应指标具有相同特征。

4.1 集合与概率空间

在定义概率空间并介绍度量指标之前，需要讨论一些核心概念并定义集合。

给定一个图 g 和一条基于图模式的关联规则（p_1， p_2， V_1， V_2）（其中 $|V_1| = |V_2| = n$），如果存在一个长度为 n 的非重复项序列 T，使得 μ_1（V_1）$= \mu_2$（V_2）$= T$，则映射函数$\mu_1 \in \Omega_{p_1, g}^{nra}$ 与映射函数 $\mu_2 \in \Omega_{p_2, g}^{nra}$ 相对应。μ_1与 μ_2之间的具体对应关系由 T刻画。为了量化两个模式之间对应关系的强度，我们需要定义：i）能够刻画两个模式的两个映射之间对应关系的项序列集合；以及 ii）实际刻画两个模式的两个映射之间对应关系的项序列集合。

考虑由集合 $\mathcal{T}_g$ 创建的长度为 n 的非重复项序列集合，即图 g 中出现的项的集合。如果模式 p_1 包含项（即 $\mathcal{T}_{p_1} \neq \varnothing$），那么对于包含属于 $\mathcal{T}_{p_1}$ 的项的项序列，我们可以先验地知道，不存在映射函数 $\mu \in \Omega_{p_1, g}^{nra}$ 使得 $\mu(V_1) = T$，因为

这违反了无重复语义。同样，对于原则上存在映射函数 $\mu_1 \in \Omega^{nra}_{p_1,\,g}$ 的项序列，我们可以先验地知道，不存在映射函数 $\mu_2 \in \Omega^{nra}_{p_2}$ 使得 T 表征 μ_1 与 μ_2 之间的对应关系，前提是 T 包含属于 $\mathcal{T}_{p_2}$ 的项。

因此，我们引入集合 $\mathcal{T}^n_{g,\,p_1,\,p_2}$，它是由 $\mathcal{T}_g \setminus (\mathcal{T}_{p_1} \cup \mathcal{T}_{p_2})$ 中的项构成的长度为 n 的非重复项序列集合。原则上，这些项序列可以刻画两个映射 $\mu_1 \in \Omega^{nra}_{p_1,\,g}$ 与 $\mu_2 \in \Omega^{nra}_{p_2,\,g}$ 之间的对应关系。我们将该集合定义如下：

$$\mathcal{T}^n_{g,p_1,p_2} := \{\,(t_1,\ldots,t_n) \in \bigtimes_{i=1}^{n} (\mathcal{T}_g \setminus (\mathcal{T}_{p_1} \cup \mathcal{T}_{p_2})) \mid \forall i,j \in [1,n] : i \neq j \Rightarrow t_i \neq t_j \,\}$$

请注意 $\mathcal{T}^n_{g,\,p_1,\,p_2} = \mathcal{T}^n_{g,\,p_2,\,p_1}$。我们按如下方式计算 $\mathcal{T}^n_{g,\,p_1,\,p_2}$ 的基数：

$$|\mathcal{T}^n_{g,p_1,p_2}| = \prod_{i=0}^{n-1} \Big(|\mathcal{T}_g \setminus (\mathcal{T}_{p_1} \cup \mathcal{T}_{p_2})| - i \Big) \tag{1}$$

我们引入一个谓词：给定一个图 g、项序列 T、模式 p 和变量序列 V，其真值按如下方式求得：

$$m_g(T,p,V) \equiv \exists \mu \in \Omega^{nra}_{p,g} : \mu(V) = T$$

我们称，在给定图 g、模式 p 和变量序列 V 的情况下，映射函数 $\mu \in \Omega^{nra}_{p\,g,}$ 与项序列 T 对应（或者，同样地，项序列 T 与映射函数 $\mu \in \Omega^{nra}_{p\,g,}$ 对应），当且仅当 $m_g(T,\ p,\ V)$ 为真。因此，对于给定的图 g、模式 p、变量序列 V 和项序列 T，如果存在一个与 T 对应的映射函数，则该谓词的值为真。

给定一个图 g、两个图模式 p_1 和 p_2，使得 $\Omega^{nra}_{p_1,g} \neq \varnothing$、$\Omega^{nra}_{p_2,g} \neq \varnothing$，且 $\mathcal{T}^n_{g,p_1,p_2} \neq \varnothing$，以及一个基于图模式的关联规则（$p_1, p_2, V_1, V_2$），我们可以如下构造概率空间 $(\Omega\ \mathcal{F}, P_g)$：

- 样本空间 Ω 是集合 $\mathcal{T}^n_{g,p_1,p_2}$ ——我们从 $\mathcal{T}^n_{g,p_1,p_2}$ 中均匀随机选取一个项序列 T，即原则上可以用于刻画两个匹配 $\mu_1 \in \Omega^{nra}_{p_1,g}$ 和 $\mu_2 \in \Omega^{nra}_{p_2,g}$ 之间对应关系的项序列集合。

- σ-代数 $\mathcal{F}$ 是幂集 $\mathcal{P}(\Omega)$，因为 Ω 是有限的，因此 Ω 的每个子集都是可测事件。

- 经验概率测度 $\hat{}\,P_g$：$\mathcal{F} \to [0,1]$ 对每个事件 $X \subseteq \Omega$ 的定义为 $\hat{}\,P_g(X) = \frac{|X|}{|\mathcal{T}^n_{g,p_1,p_2}|}$。给定项序列集合 $E \subseteq \mathcal{T}^n_{g,p_1,p_2}$，我们将其与事件 $E_{g,p_1,V_1;p_2} := \{\,T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1)\,\}$（以及事件 $E_{g,p_2,V_2;p_1} := \{\,T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_2,V_2)\,\}$）关联起来。因此，项序列 T 从 $\mathcal{T}^n_{g,p_1,p_2}$ 中均匀随机选出后，对应于 p_1 的匹配 μ（即 $\mu \in \Omega^{nra}_{p_1,g}$）且满足 $\mu(V_1) = T$ 的概率 $\hat{}\,P(E_{g,p_1,V_1;p_2})$，按如下方式测量：

$$\hat{P}_g(E_{g,p_1,V_1;p_2}) = \frac{|\{\,T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1)\,\}|}{|\mathcal{T}^n_{g,p_1,p_2}|}$$

事件 $E_{g,p_p,V_2;p_1}$ 的定义方式等价，经验概率 $\hat{}\,P_g(E_{g,p_2,V_2;p_1})$ 的测量方式也等价。

要求 $\mathcal{T}^n_{g,p_1,p_2} \neq \varnothing$ 意味着原则上这两个模式都必须能够匹配该图。一个显而易见的例子是，任何非空模式都无法匹配空图。但我们只考虑所有图均非空的情况。然而，在 $n = |V_1| = |V_2|$ 下，如果 $n > |\mathcal{T}_g \setminus (\mathcal{T}_{p_2} \cup \mathcal{T}_{p_2})|$，那么 $\mathcal{T}^n_{g,p_1,p_2} = \varnothing$，因为不存在足够多的项供变量绑定，从而无法构造长度为 n 的非重复项序列。

假设 $\mathcal{T}^n_{g,p_1,p_2}$ 非空，则 $(\Omega\ \mathcal{F}, \hat{}\,P_g)$ 是一个概率空间，因为它满足柯尔莫哥洛夫公理：

1. 非负性（公理 1）。
   对于所有的 $X \subseteq \Omega$、$|X| \geq 0$ 和 $|T| > 0$。因此，$\hat{}\,P_g(X) = \frac{|X|}{|\mathcal{T}^n_{g,p_1,p_2}|}$ 是一个有理数，因而也是实数。此外，由于 $|X| \geq 0$，我们有 $\hat{}\,P_g(X) \geq 0$ 和 $\frac{|X|}{|\mathcal{T}^n_{g,p_1,p_2}|} \geq 0$。

2. 归一化（公理2）。

Because $\Omega = \mathcal{T}^n_{g,p_1,p_2}$, it follows that $\hat{P}_g(\Omega) = \frac{|\Omega|}{|\mathcal{T}^n_{g,p_1,p_2}|} = \frac{|\Omega|}{|\Omega|} = 1$.

3. 可数可加性（公理3）。

令 $X = (X_1、\ldots\ldots、X_m)$ 为 Ω 的一列两两不相交的非空子集。则

$$\hat{P}_g\left(\bigcup_{X_i \in X} X_i\right) = \frac{\left|\bigcup\limits_{X_i \in X}(X_i)\right|}{|\mathcal{T}^n_{g,p_1,p_2}|} = \frac{\sum\limits_{X_i \in X}|X_i|}{|\mathcal{T}^n_{g,p_1,p_2}|} = \sum_{X_i \in X}\frac{|X_i|}{|\mathcal{T}^n_{g,p_1,p_2}|} = \sum_{X_i \in X}\hat{P}_g(X_i)$$

我们可以如下描述经验联合概率和条件概率:

$$\hat{P}_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) = \frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\,\}|}{|\mathcal{T}^n_{g,p_1,p_2}|}$$

$$\hat{P}_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2}) = \frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\,\}|}{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1)\,\}|}$$

$\hat{P}_g(E_{g,p_1,V_1;p_2})$ 表示从图 g 中估计得到的经验概率，即随机选取的项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$ 满足 $m_g(T,p_1,V_1)$ 的概率；这意味着存在一个在 g（即 $\mu \in \Omega^{nra}_{p_1,g}$）上的 p_1 的匹配，该匹配由 $\hat{}$ T（即 $\mu(V_1) = T$）所刻画。 $P_g(E_{g,p_2,V_2;p_1})$ 可以作等价解释。 $P_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1})$ 表示从图 g 中估计得到的经验概率，即随机选取的项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$ 同时满足 $\hat{}m_g(T,p_1,V_1)$ 和 $m_g(T,p_2,V_2)$ 的概率。最后， $P_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2})$ 表示从图 g 中估计得到的经验条件概率，即在 T 满足 $m_g(T,p_1,V_1)$ 的条件下，随机选取的项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$ 满足 $m_g(T,p_2,V_2)$ 的概率。

请注意，此处事件的定义与基于项集的关联规则语境下的定义有一个必要区别。在 ISAR 环境中，对于规则 $A \Rightarrow B$，我们能够在不了解后件的情况下定义前件的相对支持度（反之亦然）。然而，此处由于我们使用集合 $\mathcal{T}^n_{g,\,p_1,\,p_2}$，因此需要同时知道前件模式和后件模式，以便舍弃那些我们可以预先确定无法表征对应关系的项序列。因此，我们称之为事件 $E_{g,\,p_1,\,V_1;\,p_2}$，而不是 $E_{g,\,p_1,\,V_1}$。为表明模式 p_2 稍次要（因为我们主要关心 p_1 的匹配），但又不能忽略（因为没有 p_2，我们就无法指定 $\mathcal{T}^n_{g,\,p_1,\,p_2}$），我们将次要的 p_2 放在 $E_{g,\,p_1,\,V_1;\,p_2}$ 的分号之后。

4.2 单图设置下的度量指标

给定一个图 g 和一条规则 $(p_1,\ p_2,\ V_1,\ V_2)$，其中 $|V_1| = |V_2| = n$。令 $\mathcal{P}^n$ 表示图模式集合，其中每个模式都至少包含 n 个变量，并且对每个 $p \in \mathcal{P}^n$ 都有 $|\mathcal{V}_p| \geq n$；令 $\mathcal{V}^n$ 表示长度为 n 的非重复变量序列集合。

对于规则（$p_1,\ p_2,\ V_1,\ V_2$）和具有 $\mathcal{T}^n_{g,\,p_1,\,p_2} \neq \varnothing$ 的图 g，当我们定义两个图模式之一（例如 p_1）的绝对支持度和相对支持度时，也需要考虑另一个图模式（例如 p_2）。此外，我们还需要考虑连接变量的顺序，例如 V_1。因此，我们评估形如（$p_1,\ V_1; p_2$）的元组的支持度。这里， p_1 是我们主要关注的模式，但我们也需要考虑第二个图模式，即 p_2。

我们将如下定义一对项（例如 $(p_i, V_i; p_j)$，其中 $p_i = p_1$ 和 $V_i = V_1$，或 $p_i = p_2$ 和 $V_i = V_2$）的绝对支持度和相对支持度（针对单个图 g 或图的多重集 G）：

$$\begin{aligned}
\text{absolute-support}_g: \quad & (\mathcal{P}^n \times \mathcal{V}^n \times \mathcal{P}^n) \longrightarrow \mathbb{N}, \\
& (p_i, V_i; p_j) \longmapsto |\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_i, V_i)\,\}| \\
& = \textstyle\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_i,V_i)\}} \\
\text{relative-support}_g: \quad & (\mathcal{P}^n \times \mathcal{V}^n \times \mathcal{P}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\
& (p_i, V_i; p_j) \longmapsto \frac{|\{\, T \in \mathcal{T}^n_{g,p_i,p_j} \mid m_g(T, p_i, V_i)\,\}|}{|\mathcal{T}^n_{g,p_i,p_j}|} \\
& = \frac{\sum_{T \in \mathcal{T}^n_{g,p_i,p_j}} \mathbb{1}_{\{m_g(T,p_i,V_i)\}}}{|\mathcal{T}^n_{g,p_i,p_j}|} \\
& = \hat{P}_g(E_{g,p_i,V_i;p_j})
\end{aligned}$$

给定项序列集合$\mathcal{T}^n_{g,p_1,p_2}$，这些项序列原则上可以刻画两个映射$\mu_1 \in \Omega^{nra}_{p_1,g}$ 和 $\mu_2 \in \Omega^{nra}_{p_2,g}$之间的对应关系。对于一个（模式，变量序列）元组（$p_i, V_i$），单个图 g 的绝对支持度定义为其中满足 $m_g(T, p_i, V_i)$为真的项序列数量。

给定项序列集合$\mathcal{T}^n_{g,p_1,p_2}$，这些项序列原则上可以刻画两个映射$\mu_1 \in \Omega^{nra}_{p_1,g}$ 和 $\mu_2 \in \Omega^{nra}_{p_2,g}$之间的对应关系。对于一个（模式，变量序列）元组（$p_i, V_i$），单个图 g 的相对支持度定义为满足 $m_g(T, p_i, V_i)$为真的项序列数量，与原则上可以刻画两个映射 μ_1 和 μ_2之间对应关系的项序列数量之比。

支持度指标具有反单调性。给定一个（模式，变量序列）元组（p_i，V_i），当我们通过添加一个或多个三元组模式，将 p_i 扩展为 p_i' 时，对于每个变量 $v \in V_i$，额外的三元组可能会增加（但绝不会移除）关于 v 的约束——因此，不可能存在这样的项序列 T，使得存在映射 $\mu' \in \Omega^{nra}_{p_i',g}$ 满足 $\mu'(V_i) = T$，但不存在映射 $\mu \in \Omega^{nra}_{p_i,g}$ 满足 $\mu(V_i) = T$。不过，可能出现$|\Omega^{nra}_{p_i}| \le |\Omega^{nra}_{p_i',g}|$的情况，但这并不意味着支持度指标不具有反单调性。

单图场景下的置信度、提升度、杠杆率和确信度度量指标，定义方式与其 ISAR 对应指标类似。置信度定义如下：

$$\begin{aligned}
\text{confidence}_g: & (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\
& (p_1, p_2, V_1, V_2) \longmapsto \frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_1, V_1) \wedge m_g(T, p_2, V_2)\,\}|}{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_1, V_1)\,\}|} \\
& = \frac{\displaystyle\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\}}}{\displaystyle\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1)\}}} \\
& = \hat{P}_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2})
\end{aligned}$$

如果 p_1 与图 g 不匹配，则置信度得分$_g$ 未定义（即 $\Omega^{nra}_{p_1,g} = \varnothing$，或者等价地，$\hat{P}_g(E_{g,p_1,V_1;p_2}) = 0$）。我们定义提升度如下：

$$\begin{aligned}
\text{提升度}_g: & (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,\infty) \cap \mathbb{Q}, \\
& (p_1, p_2, V_1, V_2) \longmapsto \frac{\text{confidence}_g(p_1, p_2, V_1, V_2)}{\text{relative-support}_g(p_2, V_2; p_1)} = \frac{\hat{P}_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2})}{\hat{P}_g(E_{g,p_2,V_2;p_1})} \\
& = \frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_1, V_1) \wedge m_g(T, p_2, V_2)\,\}| \cdot |\mathcal{T}^n_{g,p_1,p_2}|}{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_1, V_1)\,\}| \cdot |\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_2, V_2)\,\}|} \\
& = \frac{\left[\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\}}\right] \cdot |\mathcal{T}^n_{g,p_1,p_2}|}{\left[\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1)\}}\right] \cdot \left[\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_2,V_2)\}}\right]}
\end{aligned}$$

如果 p_1 或 p_2 与图 g 不匹配（即 $\Omega^{nra}_{p_1,\ g} = \varnothing$ 或 $\Omega^{nra}_{p_2,\ g} = \varnothing$；等价地，$\hat{}\hat{}P_g(E_{g,\ p_1,\ V_1;\ p_2}) = 0$ 或 $P_g(E_{g,\ p_2,\ V_2;\ p_1}) = 0$），则提升度$_g$ 的得分未定义。杠杆率定义如下：

$$\begin{aligned}
&\text{leverage}_g : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [-\tfrac{1}{4}, \tfrac{1}{4}] \cap \mathbb{Q}, \\
&(p_1, p_2, V_1, V_2) \longmapsto \frac{|\{T \in \mathcal{T}^n_{g,p_1,p_2} \mid M_g(T, p_1, V_1 \wedge m_g(T, p_2, V_2)\}|}{|\mathcal{T}^n_{g,p_1,p_2}|} - \\
&\frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_1, V_1)\,\}|}{|\mathcal{T}^n_{g,p_1,p_2}|} \cdot \frac{|\{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T, p_2, V_2)\,\}|}{|\mathcal{T}^n_{g,p_1,p_2}|} \\
&= \frac{\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\}}}{|\mathcal{T}^n_{g,p_1,p_2}|} - \\
&\frac{\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1)\}}}{|\mathcal{T}^n_{g,p_1,p_2}|} \cdot \frac{\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_2,V_2)\}}}{|\mathcal{T}^n_{g,p_1,p_2}|} \\
&= \hat{P}_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) - \hat{P}_g(E_{g,p_2,V_2;p_1}) \cdot \hat{P}_g(E_{g,p_2,V_2;p_1})
\end{aligned}$$

如果 $\mathcal{T}^n_{g,\ p_1,\ p_2} = \varnothing$，杠杆率的得分$_g$ 未定义，这意味着在 g 上无法匹配两个模式。我们将确信度定义如下：

$$\begin{aligned}
&\text{conviction}_g : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow \mathbb{Q}_{>0} \cup \{\infty\}, \\
&(p_1, p_2, V_1, V_2) \longmapsto \frac{\hat{P}_g(E_{g,p_1,V_1;p_2}) \cdot (1 - \hat{P}_g(E_{g,p_2,V_2;p_1}))}{1 - \hat{P}_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2})} \\
&= \frac{\left(\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1)\}}\right)^2 \cdot \left(|\mathcal{T}^n_{g,p_1,p_2}| - \sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_2,V_2)\}}\right)}{|\mathcal{T}^n_{g,p_1,p_2}|^2 \cdot \left(\left[\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1)\}}\right] - \left[\sum_{T \in \mathcal{T}^n_{g,p_1,p_2}} \mathbb{1}_{\{m_g(T,p_1,V_1) \wedge m_g(T,p_2,V_2)\}}\right]\right)}
\end{aligned}$$

在 $P_g(E_{g,p_2,V_2;p_1} \upharpoonright E_{g,p_2,V_2;p_1}) = 1$（即，置信度$_g(p_1, p_2, V_1, V_2) = 1$）的情况下，确信度$_g(p_1, p_2, V_1, V_2) \to \infty$：由于分母从正值一侧趋近于零，该分数趋于无穷大。当置信度 $_g(p_1, p_2, V_1, V_2) = 1$ 时，我们将确信度$_g(p_1, p_2, V_1, V_2)$定义为 ∞。如果 $P_g(E_{g,p_2,V_2;p_1}) = 1$，则确信度未定义。

实际上，人们从非空图中推导出图模式，并从具有非零ˆ支持度的图模式构造规则。因此，不会出现这样的情况：前件的概率（即 $P_g(E_{g,\ p_1,\ V_1;\ p_2})$）或后件的概率（即 $\hat{}\ P_g(E_{g,\ p_2,\ V_2;\ p_1})$）为零，从而导致置信度分数、提升度分数或杠杆率分数未定义。

4.3 图事务环境的度量指标

我们有两种等价的方式来定义图事务环境的度量指标。以置信度度量为例。第一种方法是对每个图对应的单独概率求和，例如：

$$\text{confidence}_G(p_1, p_2, V_1, V_2) = \frac{\sum_{g \in G} \hat{P}_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1})}{\sum_{g \in G} \hat{P}_g(E_{g,p_1,V_1;p_2})}$$

第二种方法包括调整事件的定义，并定义概率 $\hat{}\ P_G(E_{G,\ p_2,\ V_2;\ p_1})$ 和 $\hat{}P_G(E_{G,\ p_1,\ V_1;\ p_2} \cap E_{G,\ p_2,\ V_2;\ p_1})$，并如下定义置信度：

$$\text{confidence}_G(p_1, p_2, V_1, V_2) = \frac{\hat{P}_G(E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1})}{\hat{P}_G(E_{G,p_1,V_1;p_2})}$$

为了定义事件 $E_{G,\ p_1,\ V_1;\ p_2}$，我们需要将样本空间从 $\mathcal{T}^n_{g\ ,\ p_1,\ p_2}$ 调整为 $\mathcal{T}^n_{G,\ p_1,\ p_2}$，定义如下：

$$\mathcal{T}^n_{G,p_1,p_2} := \{\, (i, T) \mid g_i \in G \wedge T \in \mathcal{T}^n_{g_i,p_1,p_2} \,\}$$

这里， i 是图 $g_i \in G$的标识符。然后，我们可以定义事件 $E_{G,\ p_1,\ V_1;\ p_2}$:

$$E_{G,p_1,V_1;p_2} = \{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\}$$

事件 $E_{G,\ p_2,\ V_2;\ p_1}$ 的定义等价。我们将事件的概率定义如下:

$$\hat{P}(E_{G,p_1,V_1;p_2}) = \frac{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\}|}{|\mathcal{T}^n_{G,p_1,p_2}|}$$

事件 $E_{G,\ p_2,\ V_2;\ p_1}$ 的概率定义等价。然后，我们可以如下定义相对支持度:

$$\begin{aligned} \text{relative-support}_G : \quad & (\mathcal{P}^n \times \mathcal{V}^n \times \mathcal{P}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\ & (p_i, V_i; p_j) \longmapsto \hat{P}(E_{G,p_i,V_i;p_j}) \end{aligned}$$

我们定义经验联合概率和条件概率如下:

$$\begin{aligned} \hat{P}_G(E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1}) = & \quad \frac{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \wedge m_{g_i}(T,p_2,V_2) \,\}|}{|\mathcal{T}^n_{G,p_1,p_2}|} \\ \hat{P}_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_1,V_1;p_1}) = & \quad \frac{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \wedge m_{g_i}(T,p_2,V_2) \,\}|}{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\}|} \end{aligned}$$

$\hat{}P_G(E_{G,p_1,V_1;p_2})$ 表示从图的多重集 G 中估计得到的经验概率：对于随机选取的一个二元组，该二元组由图的标识符 i 和项序列 T 组成（原则上，该项序列可能对应于图中标识符为 i 的图上 p_1 的匹配和 p_2 的匹配），项序列满足 $m_{g_i}(T,p_1,V_1)$；这意味着存在一个 g_i 上 p_1 的匹配（即 $\mu \in \Omega^{nra}_{p_1,g_i}$），且该匹配由 T（即 $\mu(V_1) = T$）表征。 $P_G(E_{G,p_2,V_2})$ 也可以作等价解释。 $P_G(E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1})$ 表示从图的多重集 G 中估计得到的经验概率：对于随机选取的一个二元组，该二元组由图的标识符 i 和项序列 T 组成（原则上，该项序列可能对应于图中标识符为 i 的图上 p_1 的匹配和 p_2 的匹配），项序列同时满足 $m_{g_i}(T,p_1,V_1)$ 和 $m_{g_i}(T,p_2,V_2)$。最后，$\hat{}P_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_1,V_1;p_2})$ 表示从图的多重集 G 中估计得到的经验条件概率：对于随机选取的一个二元组，该二元组由图的标识符 i 和项序列 T 组成（原则上，该项序列可能对应于图中标识符为 i 的图上 p_1 的匹配和 p_2 的匹配），在项序列满足 $m_{g_i}(T,p_1,V_1)$ 的条件下，该项序列满足 $m_{g_i}(T,p_2,V_2)$。

我们构造一个概率空间 $(\Omega\ \mathcal{F},\ \hat{}P_G)$，其中 $\Omega = \mathcal{T}^n_{G,\ p_1,\ p_2}, \mathcal{F} = \mathcal{P}(\Omega)$，以及上文定义的经验概率测度 $\hat{}P_G$。显然，在这种情况下柯尔莫哥洛夫公理同样成立。基于事件$E_{G,\ p_1,\ V_1;\ p_2}$和 $E_{G,\ p_2,\ V_2;\ p_1}$，我们可以为图事务设置定义以下度量指标:

$$\begin{aligned} \text{confidence}_G \ : & (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\ & (p_1,p_2,V_1,V_2) \longmapsto \hat{P}_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_1,V_1;p_2}) \\ & = \frac{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \wedge m_{g_i}(T,p_2,V_2) \,\}|}{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\}|} \end{aligned}$$

如果 p_1 与 G 中的任何图都不匹配（即，$\forall g \in G : \Omega^{nra}_{p_2,\ g} = \varnothing$，或者等价地，$\hat{}\forall g \in G :\ P_g(E_{g,\ p_1,\ V_1;\ p_2}) = 0$），则置信度$_G$ 的得分未定义。

$$\begin{aligned} \text{提升度}_G : & (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,\infty) \cap \mathbb{Q}, \\ & (p_1,p_2,V_1,V_2) \longmapsto \frac{\hat{P}_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_1,V_1;p_2})}{\hat{P}_G(E_{G,p_2,V_2;p_1})} \\ & = \frac{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \wedge m_{g_i}(T,p_2,V_2) \,\}|}{|\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\}| \cdot |\{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_2,V_2) \,\}|} \end{aligned}$$

如果 p_1 未匹配 G 中的任何图，或 p_2 未匹配 G 中的任何图，则提升度$_G$ 的得分未定义（即 $\hat{}\hat{}\forall g \in G : \Omega^{nra}_{p_1,\ g} = \varnothing \vee \forall g \in G : \Omega^{nra}_{p_2,\ g} = \varnothing$，或者等价地，$\forall g \in G :\ P_g(E_{g,\ p_1,\ V_1;\ p_2}) = P_g(E_{g,\ p_2,\ V_2;\ p_1}) = 0$）。

$$\text{leverage}_G \ : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [-\tfrac{1}{4}, \tfrac{1}{4}] \cap \mathbb{Q},$$

$$(p_1, p_2, V_1, V_2) \longmapsto \hat{P}_G(E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1}) - \hat{P}_G(E_{G,p_2,V_2;p_1}) \cdot \hat{P}_G(E_{G,p_2,V_2;p_1})$$
$$= \frac{|\{\,(i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \wedge m_{g_i}(T,p_2,V_2)\,\}|}{|\mathcal{T}^n_{G,p_1,p_2}|} -$$
$$\frac{|\{\,(i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1)\,\}|}{|\mathcal{T}^n_{G,p_1,p_2}|} \cdot \frac{|\{\,(i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_2,V_2)\,\}|}{|\mathcal{T}^n_{G,p_1,p_2}|}$$

如果对于所有图 $g \in G$，都有 $\mathcal{T}^n_{g,\,p_1,\,p_2} = \varnothing$，则杠杆率$_G$ 的得分未定义；这意味着对于任何图，都不可能同时匹配这两个模式。

$$\text{conviction}_G : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow \mathbb{Q}_{>0} \cup \{\infty\},$$

$$(p_1, p_2, V_1, V_2) \longmapsto \frac{\hat{P}_G(E_{G,p_1,V_1;p_2}) \cdot (1 - \hat{P}_G(E_{G,p_2,V_2;p_1}))}{1 - \hat{P}_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_1,V_1;p_2})}$$

$$= \frac{\left(\sum_{(i,T)\in\mathcal{T}^n_{G,p_1,p_2}} \mathbb{1}_{\{m_{g_i}(T,p_1,V_1)\}}\right)^2 \cdot \left(|\mathcal{T}^n_{G,p_1,p_2}| - \sum_{(i,T)\in\mathcal{T}^n_{G,p_1,p_2}} \mathbb{1}_{\{m_{g_i}(T,p_2,V_2)\}}\right)}{|\mathcal{T}^n_{G,p_1,p_2}|^2 \cdot \left(\left[\sum_{(i,T)\in\mathcal{T}^n_{G,p_1,p_2}} \mathbb{1}_{\{m_{g_i}(T,p_1,V_1)\}}\right] - \left[\sum_{(i,T)\in\mathcal{T}^n_{G,p_1,p_2}} \mathbb{1}_{\{m_{g_i}(T,p_1,V_1)\wedge m_{g_i}(T,p_2,V_2)\}}\right]\right)}$$

当 $\hat{}\ P_G(E_{G,p_2,V_2;p_1} \mid E_{G,p_2,V_2;p_1}) = 1$（即 $\text{confidence}_G(p_1,p_2,V_1,V_2) = 1$）时，$\text{conviction}_G(p_1,p_2,V_1,V_2) \to \infty$：由于分母从正侧趋近于零，该分数趋于无穷大。我们将 $\text{conviction}_G(p_1,p_2,V_1,V_2)$ 定义为 ∞，适用于 $\text{confidence}_G(p_1,p_2,V_1,V_2) = 1$ 的情况。若 $\hat{}\ P_G(E_{G,p_2,V_2;p_1}) = 1$，则 conviction 未定义

请注意，这些度量指标是微平均。因此，较大的图对得分的影响可能强于较小的图。对于不希望出现这种情况的场景，我们引入宏平均指标，首先介绍宏平均相对支持度和宏平均置信度：

$$\begin{aligned}
\text{macro-relative-support}_G : \quad & (\mathcal{P}^n \times \mathcal{V}^n \times \mathcal{P}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\
& (p_i, V_i; p_j) \longmapsto \frac{1}{|G|} \sum_{g \in G} \hat{P}(E_{g,p_i,V_i;p_j}) \\
\text{macro-confidence}_G : \quad & (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,1] \cap \mathbb{Q}, \\
& (p_1, p_2, V_1, V_2) \longmapsto \tfrac{1}{|G|} \textstyle\sum_{g \in G} \text{confidence}_g(p_1, p_2, V_1, V_2)
\end{aligned}$$

如果存在一张图 $g \in G$ 未与 p_1 匹配（即，$\exists g \in G : \Omega^{nra}_{p_1,g} = \varnothing$），则宏观置信度$_G$ 的得分未定义。因此，给定一个由图组成的多重集 G，其中某些图未与 p_1匹配，我们需要约简 G，使约简后的集合中每张图都满足该条件。如果我们从一个由图组成的多重集 G 中推导出两条规则 $r_a = (p_a,1,p_a,2,V_a,1,V_a,2)$ 和 $r_b = (p_b,1,p_b,2,V_b,1,V_b,2)$，那么将 G 约简为满足 $\hat{}P(E_{g,p_a,1,V_a,1,p_a,2}) > 0$ 的那些图 g 所得的结果，其大小可能不同于将 G 约简为满足 $\hat{}P(E_{g,p_b,1,V_b,1,p_b,2}) > 0$的那些图 g 所得的结果。因此，如果 r_1 的提升度分数高于 r_2，规则 r_2 仍可能更有趣，因为对于 r_2，G 需要约简的程度小于对于 r_1所需的程度。因此，在实际应用中，仅按提升度分数对规则进行排序可能并不足够。

为了量化多重集 G 中满足某项指标所要求的特定图条件的图所占比例，我们将该指标的适用程度$_G$ 定义如下：

$$\text{degree-of-applicability}_G(c) = \frac{\sum_{g \in G} \mathbb{1}_{\{c\}}}{|G|}$$

其中 c 是一个布尔表达式。对于置信度，我们使用 $c = \hat{}P(E_{g,\,p_1,\,V_1,\,p_2}) > 0$，并在约简多重集 $G_c = \{\, g \in G \mid c \,\}$上计算宏平均置信度。

对一组基于图模式的关联规则进行排序时，一种简单的方法是采用两阶段排序：首先根据适用度得分按降序排序。若得分相同，则根据提升度分数等指标打破平局。这种两阶段排序方法也称为字典序排序。多准则决策（MCDM）领域提供了字典序排序之外的更多选择。我们将宏平均提升度定义如下：

$$\text{macro-lift}_G : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [0,\infty) \cap \mathbb{Q},$$

$$(p_1, p_2, V_1, V_2) \longmapsto \frac{1}{|G|} \sum_{g \in G} \mathrm{lift}_g(p_1, p_2, V_1, V_2)$$

如果存在某个图 $g \in G$，使得 p_1 或 p_2 不匹配（即 $\exists g \in G : \hat{} \hat{} \Omega^{nra}_{p_1,\ g} = \varnothing \vee \Omega^{nra}_{p_2,\ g} = \varnothing$），则宏提升度 $_G$ 的得分未定义。因此，我们可能需要使用 $c = P\ (E_{g,\ p_1,\ V_1,\ p_2})$ 约简 $G\ > 0 \wedge P\ (E_{g,\ p_2,\ V_2,\ p_1})$，并在约简多重集 G_c 上评估宏提升度。我们将宏平均杠杆率定义如下:

$$\text{macro-leverage}_G : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow [-\tfrac{1}{4}, \tfrac{1}{4}] \cap \mathbb{Q},$$

$$(p_1, p_2, V_1, V_2) \longmapsto \frac{1}{|G|} \sum_{g \in G} \mathrm{leverage}_g(p_1, p_2, V_1, V_2)$$

如果存在一个图 $g \in G$，使得 $\mathcal{T}^n_{g,p_1,p_2} = \varnothing$，那么宏观杠杆$_G$ 的得分未定义，这意味着两个模式无法同时在 g 上匹配（即 $\exists g \in G : \mathcal{T}^n_{g,p_1,p_2} = \varnothing$）。因此，我们可能需要用 $c = \mathcal{T}^n_{g,p_1,p_2} \neq \varnothing$ 约简 G，并在约简后的多重集 G_c 上评估宏观杠杆。我们将宏平均确信度定义如下:

$$\text{macro-conviction}_G : (\mathcal{P}^n \times \mathcal{P}^n \times \mathcal{V}^n \times \mathcal{V}^n) \longrightarrow \mathbb{Q}_{>0} \cup \{\infty\},$$

$$(p_1, p_2, V_1, V_2) \longmapsto \frac{1}{|G|} \sum_{g \in G} \mathrm{conviction}_g(p_1, p_2, V_1, V_2)$$

我们将确信度$_g$ 的得分定义为：当 $\hat{}\, P_G(E_{G,p_2,V_2};p_1 \mid E_{G,p_1,V_1};p_2) = 1$时，得分为 ∞。因此，当至少一个 $g \in G$ 的确信度为 ∞ 时，我们需要定义宏观确信度$_G$ 的得分。我们将 ∞ 视为一个值，并令其按如下方式表现: $\infty + \infty = \infty$; 对于任意 $x \in \mathbb{Q}$, $\infty + x = \infty$。另一种做法是，统计 G 中确信度$_g = \infty$ 的图的数量，然后在约简后的多重集 G_c 上评估宏观确信度，其中 $c = \hat{}P_g(E_{g,p_2,V_2};p_1 \mid E_{g,p_1,V_1};p_2) < 1 \wedge \hat{}P_g(E_{g,p_2,V_2};p_1) < 1$。

需要注意的是，对得分进行宏平均时，并不存在单一的概率空间——我们处理的是一族概率空间，而宏平均是对度量指标进行的外部聚合，并不对应于某个统一概率空间中的概率。因此，这些度量指标不再是单一测度 P 的纯函数。这会影响宏平均度量指标与其对应的微平均度量指标及基于项集的度量指标具有哪些共同特性，我们将在第4.5节中看到。

4.4 重新审视经典关联规则度量指标

原则上，我们可以将衡量基于图模式的关联规则性质的问题，重新表述为衡量基于项集的关联规则性质的问题——无论给定的是单个图，还是图袋及微平均度量指标，均可如此。该方法包括：针对给定的基于图模式的关联规则（T，p_1，V_1，V_2）以及图袋 G（其中可以只包含一个图），生成事务数据库 T。我们按如下方式生成图袋 T（使用袋并运算 $\uplus$）：

$$T = \biguplus_{g \in G} \ \biguplus_{X \in \mathcal{T}^n_{g,p_1,p_2}} [\{\, A \mid m_g(X, p_1, V_1)\,\} \cup \{\, B \mid m_g(X, p_2, V_2)\,\}]$$

因此，对于每个图 $g \in G$，对于每个项序列 $X \in \mathcal{T}^n_{g,p_1,p_2}$，我们向 T 中添加一个新事务 t，其中 $t \in \{\varnothing, \{A\}, \{B\}, \{A, B\}\}$。$A$ 和 B 是特殊符号。在图 g 和项序列 X 的上下文中，A 表示 $m_g(X, p_1, V_1)$ 为真，这意味着存在一个匹配 $\mu \in \Omega^{nra}_{p_1,g}$，使得 $\mu(V_1) = X$。B 表示 $m_g(X, p_2, V_2)$ 为真，这意味着存在一个匹配 $\mu \in \Omega^{nra}_{p_2,g}$，使得 $\mu(V_2) = X$。

由于对于每个 $g \in G$，我们恰好向多重集 T 中添加 $|\mathcal{T}^n_{g,p_1,p_2}|$ 个事务，因此 $|T| = |\mathcal{T}^n_{G,p_1,p_2}|$。此外，$T$ 中包含 A 的事务数等于 $\sum_{g \in G} |\{\, X \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(X, p_1, V_1)\,\}|$，$T$ 中包含 B 的事务数等于 $\sum_{g \in G} |\{\, X \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(X, p_2, V_2)\,\}|$，而 T 中同时包含 A 和 B 的事务数等于 $\sum_{g \in G} |\{\, X \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(X, p_1, V_1) \wedge m_g(X, p_2, V_2)\,\}|$。由此，我们构造了一个同态，使我们能够在经典事务数据库上评估基于项集的关联规则的经典度量指标（其中每个事务都是一个项集），而不必在事务数据库上评估基于图模式的关联规则的度量指标（其中每个事务都是一个图）。

给定基于图模式的关联规则（p_1，p_2，V_1，V_2）、图的多重集 G、两个分别对应模式 p_1 和 p_2 的特殊符号 A 和 B，以及生成的事务数据库 T，我们定义如下

事件之间的对应关系：

$$\begin{aligned}\hat{P}_G(E_{p_1,V_1;p_2}) &= \hat{P}_T(E_{\{A\}})\\ \hat{P}_G(E_{p_2,V_2;p_1}) &= \hat{P}_T(E_{\{B\}})\\ \hat{P}_G(E_{p_1,V_1;p_2} \cap E_{p_2,V_2;p_1}) &= \hat{P}_T(E_{\{A,B\}})\end{aligned}$$

由这些对应关系可以进一步得到以下对应关系：

$$\begin{aligned}\text{relative-support}_G(p_1,V_1;p_2) &= \hat{P}_T(E_{\{A\}})\\ \text{relative-support}_G(p_2,V_2;p_1) &= \hat{P}_T(E_{\{B\}})\\ \text{conf.}_G(p_1,p_2,V_1,V_2) = \hat{P}_G(E_{p_2,V_2;p_1} \mid E_{p_1,V_1;p_2}) &= \hat{P}_T(E_{\{B\}} \mid E_{\{A\}}) = \text{conf.}_T(\{A\} \Rightarrow \{B\})\\ \text{lift}_G(p_1,p_2,V_1,V_2) &= \text{lift}_T(\{A\} \Rightarrow \{B\})\\ \text{leverage}_G(p_1,p_2,V_1,V_2) &= \text{leverage}_T(\{A\} \Rightarrow \{B\})\\ \text{conviction}_G(p_1,p_2,V_1,V_2) &= \text{conviction}_T(\{A\} \Rightarrow \{B\})\end{aligned}$$

根据同态的存在性，我们可以得出结论：对于基于图模式的关联规则，微平均度量指标是基于项集的关联规则度量指标的泛化。

尽管从理论角度来看，这种同态很有意义，但利用该同态并实际生成基于项集的事务数据库并不切实际，原因有以下几点：i）集合 $\mathcal{T}^n_{g,\ p_1,\ p_2}$ 可能非常大，参见公式 1。预计通常大多数项序列都不对应任何匹配。因此，生成的数据库会很大，并且其中大多数事务将为空。此外，ii）对于每条规则，都需要生成另一个事务数据库；因此，在针对一组规则评估度量指标时，会产生很高的计算成本和内存成本。

上述事务数据库生成过程只能应用于微平均度量指标。对于宏平均度量指标，情况有所不同。在这种情况下，需要区分为不同图生成的事务。一种简单的方法是为每个图生成一个事务数据库，评估一个度量指标，然后对所得得分取平均值。

4.5 指标的特征

我们所说的指标特征，是指指标在特别有意义的情境下所获得的得分。在表 2 中，我们为比较和解释这些度量指标的得分时所涉及的相关情境引入了相应的名称。根据所处理的是基于项集的关联规则（项集设置）情境、基于图模式的关联规则与单个图的情境（单图设置）、基于图模式的关联规则与图袋且得分采用微平均的情境（图事务设置，微平均），还是基于图模式的关联规则与图袋且得分采用宏平均的情境（图事务设置，宏平均），这些情境的描述也有所不同。

表 2：描述和比较 ISAR 与 GPAR 度量指标得分时相关情形的名称和说明。

Name	情境描述
IDE	事件 E_A 和 E_B 完全相同，即 $E_A = E_B$。这意味着，对于每个事务 $t \in T$，当满足 $A \subseteq t$ 时，$B \subset t$ 也会满足，反之亦然（项集设置）。
	事件 $E_{g,\ p_1,\ V_1;\ p_2}$ 以及 $E_{g,\ p_2,\ V_2;\ p_1}$ 是相同的，即，$E_{g,p_1,V_1;p_2} = E_{g,p_2,V_2;\ p_1}$。That 表示，对于每个项序列 $T \in \mathcal{T}^n$，当满足 $m_g(T、\ p_1、\ V_1)$ 时，也满足 $m_g(T、\ p_2、\ V_2)$，反之亦然（单图设置）。
	事件 $E_{G,p_1,V_1;p_2}$ 和 $E_{G、\ p_2、\ V_2;\ p_1}$ 相同，即，$E_{G,p_1,V_1;p_2} = E_{G,p_2,V_2;\ p_1}$。这意味着，对于每一对 $(i,\ T) \in \mathcal{T}^n_{G,p_1,p_2}$，当 $m_{g_i}(T,\ p_1,\ V_1)$ 也满足 $m_{g_i}(T,\ p_2,\ V_2)$ 满足，反之亦然（图事务设置，微平均）。
	对于每个图 $g \in G$ 可以看出，$E_{g,\ p_1,\ V_1;\ p_2}$ 和 $E_{g,\ p_2,\ V_2;\ p_1}$ 相同（图事务环境，宏平均）。

DIS 事件 E_A 和 E_B 互不相交，即 $E_A \cap E_B = \varnothing$。这意味着，对于每个事务$t \in T$，当满足 $A \subseteq t$ 时，就不满足 $B \subseteq t$，反之亦然（项集设置）。事件 $E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 互不相交，即 $E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1} = \varnothing$。这意味着，对于每个项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$，当满足 $m_g(T, p_1, V_1)$ 时，就不满足$m_g(T, p_2, V_2)$，反之亦然（单图设置）。事件 $E_{G,p_1,V_1;p_2}$ 和 $E_{G,p_2,V_2;p_1}$ 互不相交，即 $E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1} = \varnothing$。这意味着，对于每一对 $(i,T) \in \mathcal{T}^n_{G,p_1,p_2}$，当满足 $m_{g_i}(T, p_1, V_1)$ 时，也满足$m_{g_i}(T, p_2, V_2)$，反之亦然（图事务设置，微平均）。对于每个图 $g \in G$，$E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 互不相交（图事务设置，宏平均）。IND 事件 E_A 和 E_B 相互独立，即 $\hat{}P_T(E_A \cap E_B) = \hat{}P_T(E_A) \cdot \hat{}P_T(E_B)$。这意味着，对于每个事务 $t \in T$，当满足 $A \subseteq t$ 时，不会改变满足 $B \subseteq t$ 的概率，反之亦然（项集设置）。$\hat{}$ 事件 $E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 相互独立，即 $P_T(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) = \hat{}\hat{}P_T(E_{g,p_1,V_1;p_2}) \cdot P_T(E_{g,p_2,V_2;p_1})$。这意味着，对于每个项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$，当满足$m_g(T, p_1, V_1)$ 时，不会改变满足 $m_g(T, p_2, V_2)$ 的概率，反之亦然（单图设置）。$\hat{}$ 事件 $E_{G,p_1,V_1;p_2}$ 和 $E_{G,p_2,V_2;p_1}$ 相互独立，即 $P_T(E_{G,p_1,V_1;p_2} \cap E_{G,p_2,V_2;p_1}) = \hat{}P_T(E_{G,p_1,V_1;p_2}) \cdot \hat{}P_T(E_{G,p_2,V_2;p_1})$。这意味着，对于每一对 $(i,T) \in \mathcal{T}^n_{G,p_1,p_2}$，当满足$m_{g_i}(T, p_1, V_1)$ 时，不会改变满足 $m_{g_i}(T, p_2, V_2)$ 的概率，反之亦然（图事务设置，微平均）。对于每个图 $g \in G$，$E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 相互独立（图事务设置，宏平均）。POS 事件 E_A 和 E_B 正相关，即 $\hat{}\, P_T(E_A \cap E_B) > \hat{}P(E_A) \cdot \hat{}P(E_B)$。这意味着，对于每个事务 $t \in T$，当满足 $A \subseteq t$ 时，满足 $B \subseteq t$ 的概率会增加，反之亦然（项集设置）。$\hat{}$ 事件$E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 正相关，即 $P_g(E_{g,p_1,V_1;p_2} \curlywedge \hat{}E_{g,p_2,V_2;p_1}) > P_g(E_{g,p_1,V_1;p_2}) \cdot P_g(E_{g,p_2,V_2;p_1})$。这意味着，对于每个项序列$T \in \mathcal{T}^n_{g,p_1,p_2}$，当满足 $m_g(T, p_1, V_1)$ 时，满足$m_g(T, p_2, V_2)$ 的概率会增加，反之亦然（单图设置）。$\hat{}$ 事件 $E_{G,p_1,V_1;p_2}$ 和$E_{G,p_2,V_2;p_1}$ 正相关，即 $P_G(E_{G,p_1,V_1;p_2} \curlywedge \hat{}E_{G,p_2,V_2;p_1}) > P_G(E_{G,p_1,V_1;p_2}) \cdot P_G(E_{G,p_2,V_2;p_1})$。这意味着，对于每一对 $(i,T) \in \mathcal{T}^n_{G,p_1,p_2}$，当满足 $m_{g_i}(T, p_1, V_1)$ 时，满足$m_{g_i}(T, p_2, V_2)$ 的概率会增加，反之亦然（图事务设置，微平均）。对于每个图 $g \in G$，$E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 正相关（图事务设置，宏平均）。$\hat{}\,\hat{}\,\hat{}$ NEG 事件 E_A和 E_B负相关，即 $P_T(E_A \cap E_B) < P(E_A) \cdot P(E_B)$。这意味着，对于每个事务 $t \in T$，当满足 $A \subseteq t$ 时，满足 $B \subseteq t$ 的概率会降低，反之亦然（项集设置）。$\hat{}$ 事件 $E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 负相关，即 $P_g(E_{g,p_1,V_1;p_2} \curlywedge \hat{}E_{g,p_2,V_2;p_1}) < P_g(E_{g,p_1,V_1;p_2}) \cdot P_g(E_{g,p_2,V_2;p_1})$。这意味着，对于每个项序列 $T \in \mathcal{T}^n_{g,p_1,p_2}$，当满足 $m_g(T, p_1, V_1)$ 时，满足$m_g(T, p_2, V_2)$ 的概率会降低，反之亦然（单图设置）。$\hat{}$ 事件$E_{G,p_1,V_1;p_2}$ 和 $E_{G,p_2,V_2;p_1}$ 负相关，即 $P_G(E_{G,p_1,V_1;p_2} \curlywedge \hat{}E_{G,p_2,V_2;p_1}) < P_G(E_{G,p_1,V_1;p_2}) \cdot P_G(E_{G,p_2,V_2;p_1})$。这意味着，对于每一对 $(i,T) \in \mathcal{T}^n_{G,p_1,p_2}$，当满足 $m_{g_i}(T, p_1, V_1)$ 时，满足$m_{g_i}(T, p_2, V_2)$ 的概率会降低，反之亦然（图事务设置，微平均）。对于每个图 $g \in G$，$E_{g,p_1,V_1;p_2}$ 和 $E_{g,p_2,V_2;p_1}$ 负相关（图事务设置，宏平均）。

表3–6分别展示了基于项集的关联规则（记为 ISAR）、单图环境下基于图模式的关联（记为 GPAR（单图））、得分采用微平均的图事务环境下的基于图模式的关联（记为 GPAR（微观）），以及得分采用宏平均的图事务环境下的基于图模式的关联（记为 GPAR（宏平均））中，置信度、提升度、杠杆率和确信度这些度量指标的关键特征，分别针对

表3：置信度指标的主要特征，假设 $\hat{P}_T(E_B) > 0$、$\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}) > 0$、$\hat{P}_G(E_{G,\ p_1,\ V_1;\ p_2}) > 0$，或 $\forall g \in G : \hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}) > 0$。所有四种置信度度量的陪域均为 $[0,1] \cap \mathbb{Q}$。

Sit.	ISAR		GPAR（单图）	
IDE	$\Longrightarrow = 1$	(P1)	$\Longrightarrow = 1$	(P1)
DIS	$\Longleftrightarrow = 0$	(P2)	$\Longleftrightarrow = 0$	(P2)
IND	$\Longleftrightarrow = \hat{P}_T(E_B)$	(P3)	$\Longleftrightarrow = \hat{P}_g(E_{g,p_2,V_2;p_1})$	(P3)
POS	$\Longleftrightarrow > \hat{P}_T(E_B)$	(P4)	$\Longleftrightarrow > \hat{P}_g(E^{g,p_2}_{V_2;p_1})$	(P4)
NEG	$\Longleftrightarrow < \hat{P}_T(E_B)$	(P5)	$\Longleftrightarrow < \hat{P}_g(E^{g,p_2}_{g,p_2,V_2;p_1})$	(P5)
Sit.	**GPAR（微观）**		**GPAR（宏平均）**	
IDE	$\Longrightarrow = 1$	(P1)	$\Longrightarrow = 1$	(P6)
DIS	$\Longleftrightarrow = 0$	(P2)	$\Longleftrightarrow = 0$	(P7)
IND	$\Longleftrightarrow = \hat{P}_G(E_{G\ p_2,V_2;p_1})$	(P3)	$\Longrightarrow = \frac{1}{\lvert G\rvert}\sum_{g\in G}\hat{P}_G(E_{G\ p_2,V_2;p_1})$	(P8)
POS	$\Longleftrightarrow > \hat{P}_G(E_{G\ p_2,V_2;\ p_1})$	(P4)	$\Longrightarrow > \frac{1}{\lvert G\rvert}\sum_{g\in G}\hat{P}_G(E_{G\ p_2,V_2;\ p_1})$	(P9)
NEG	$\Longleftrightarrow < \hat{P}_G(E_{G,p_2,V_2;p_1})$	(P5)	$\Longrightarrow < \frac{1}{\lvert G\rvert}\sum_{g\in G}\hat{P}_G(E_{G,p_2,V_2;p_1})$	(P10)

表4：提升度量的关键特征，假设 $\hat{P}_T(E_A) > 0 \wedge \hat{P}_T(E_B) > 0$、$\hat{P}_g(E_{g,p_1,V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,V_2;p_1}) > 0$、$\hat{P}_G(E_{G,p_1,V_1;p_2}) > 0 \wedge \hat{P}_G(E_{G,p_2,V_2;p_1}) > 0$，或 $\forall g \in G : \hat{P}_g(E_{g,p_1,V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,V_2;p_1}) > 0$。所有四种提升度量的陪域均为 $[0,\infty) \cap \mathbb{Q}$。

Sit.	ISAR		GPAR（单图）	
IDE	$\Longrightarrow = 1/\hat{P}_T(E_B)$	(P11)	$\Longrightarrow = 1/\hat{P}_g(E_{g,p_1,V_1;p_2})$	(P11)
DIS	$\Longleftrightarrow = 0$	(P12)	$\Longleftrightarrow = 0$	(P12)
IND	$\Longleftrightarrow = 1$	(P13)	$\Longleftrightarrow = 1$	(P13)
POS	$\Longleftrightarrow > 1$	(P14)	$\Longleftrightarrow > 1$	(P14)
NEG	$\Longleftrightarrow < 1$	(P15)	$\Longleftrightarrow < 1$	(P15)
Sit.	**GPAR（微观）**		**GPAR（宏平均）**	
IDE	$\Longrightarrow = 1/\hat{P}_G(E_{G\ p_1,V_1;p_2})$	(P11)	$\Longrightarrow = \frac{1}{\lvert G\rvert}\sum_{g\in G}1/\hat{P}_G(E_{G\ p_1,V_1;p_2})$	(P16)
DIS	$\Longleftrightarrow = 0$	(P12)	$\Longleftrightarrow = 0$	(P17)
IND	$\Longleftrightarrow = 1$	(P13)	$\Longrightarrow = 1$	(P18)
POS	$\Longleftrightarrow > 1$	(P14)	$\Longrightarrow > 1$	(P19)
NEG	$\Longleftrightarrow < 1$	(P15)	$\Longrightarrow < 1$	(P20)

表中解释的 IDE、DIS、IND、正和负这几种情形（表 2）。对于每个得分，均提供相应命题的编号，其证明见附录。例如，表 3 表明，置信度$_G(p_1,p_2,V_1,V_2)$（即 GPAR（微观）语境下）的得分为 0，当且仅当（$\Longleftrightarrow$）出现 DIS 情形时（即 $E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1} = \varnothing$）（命题 P2）。此外，表 3 表明，置信度$_g(p_1,p_2,V_1,V_2)$（即 GPAR（单图）语境下）的得分在出现 IDE 情形时为 1（$\Longrightarrow$，但并非当且仅当）（即 $E_{g,p_1,V_1;p_2} = E_{g,p_2,V_2;p_1}$）（命题 P1）。

表 7 总结了度量指标的定义条件。我们从这些条件中删除冗余项。例如，我们写作 $\hat{P}_T(E_A) > 0$，而不是 $T \neq \varnothing \wedge \hat{P}_T(E_A) > 0$，因为由 $\hat{P}_T(E_A) > 0$ 可知 $T \neq \varnothing$。

从表3–6和表 7可以看出，这些度量指标在不同环境中的表现相似。在单图设置和图事务环境中，项集关联规则的度量指标与基于图模式的关联规则的度量指标具有相似特征，其中图事务环境中的得分采用微平均。这是因为这些特征源于我们拥有一个概率空间，且这些度量指标是根据概率定义的。然而，宏平均度量指标与其他度量指标之间存在一些差异。对于宏平均度量指标，我们常常（20种情况中有14种）无法根据得分推断出具体情况。例如，在 ISAR 环境中，如果提升度的得分为 1，我们就可以推断出 $\hat{P}_T(E_A)$ 和 $\hat{P}_T(E_B)$ 相互独立。然而，在 GPAR $\hat{}\ \hat{}$（宏平均）环境中，如果提升度的得分为 1，我们无法推断出 $P_G(E_{G,\ p_1,\ V_1;\ p_2})$ 和 $P_G(E_{G,\ p_2,\ V_2;\ p_1})$ 在每个 $g \in G$中都相互独立。

表 5：杠杆度量的主要特征，假设分别为 $T \neq \varnothing$, $\mathcal{T}^n_{g,\ p_1,\ p_2} \neq \varnothing$, $\mathcal{T}^n_{G,\ p_1,\ p_2} \neq \varnothing$，或 $\forall g \in G : \mathcal{T}^n_{g,\ p_1,\ p_2} \neq \varnothing$。所有四种杠杆度量的陪域均为 $[-\frac{1}{4}, \frac{1}{4}] \cap \mathbb{Q}$。

Sit.	ISAR		GPAR（单图）	
IDE	$\Longrightarrow = \hat{P}_T(E_A)(1 - \hat{P}_T(E_B))$	(P21)	$\Longrightarrow = \hat{P}_g(E_{g,p_1,V_1;p_2})(1 - \hat{P}_g(E_{g,p_2,V_2;p_1}))$	(P21)
DIS	$\Longleftrightarrow = -\hat{P}_T(E_A)\,\hat{P}_T(E_B)$	(P22)	$\Longleftrightarrow = -\hat{P}_g(E_{g,p_1,V_1;p_2})\hat{P}_g(E_{g,p_2,V_2;p_1})$	(P22)
IND	$\Longleftrightarrow = 0$	(P23)	$\Longleftrightarrow = 0$	(P23)
POS	$\Longleftrightarrow > 0$	(P24)	$\Longleftrightarrow > 0$	(P24)
NEG	$\Longleftrightarrow < 0$	(P25)	$\Longleftrightarrow < 0$	(P25)
Sit.	**GPAR（微观）**		**GPAR（宏平均）**	
IDE	$\Longrightarrow = \hat{P}_G(E_{G,\ p_1,\ V_1;\ p_2})(1 - \hat{P}_G(E_{G,p_2,V_2;p_1}))$	(P21)	$\Longrightarrow = \frac{1}{\lvert G\rvert}\sum_{g\in G}\hat{P}_g(E_{g,p_1,V_1;p_2})(1 - \hat{P}_g(E_{g,p_2,V_2;p_1}))$	(P26)
DIS	$\Longleftrightarrow = -\hat{P}_G(E_{G,p_1,V_1;p_2})\ \hat{P}_G(E_{G,p_2,V_2;p_1})$	(P22)	$\Longrightarrow = -\frac{1}{\lvert G\rvert}\sum_{g\in G}\hat{P}_g(E_{g,p_1,\ V_1;\ p_2})\hat{P}_g(E_{g,\ p_2,\ V_2;\ p_1})$	(P27)
IND	$\Longleftrightarrow = 0$	(P23)	$\Longrightarrow = 0$	(P28)
POS	$\Longleftrightarrow > 0$	(P24)	$\Longrightarrow > 0$	(P29)
NEG	$\Longleftrightarrow < 0$	(P25)	$\Longrightarrow < 0$	(P30)

表6：确信度指标的关键特征，分别假设 $\hat{P}_T(E_A) > 0 \wedge \hat{P}_T(E_B) < 1$, $\hat{P}_g(E_{g,p_1,V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,V_2;p_1}) < 1$, $\hat{P}_G(E_{G,p_1,V_1;p_2}) > 0 \wedge \hat{P}_G(E_{G,p_2,V_2;p_1}) < 1$，或 $\forall g \in G : \hat{P}_g(E_{g,p_1,V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,V_2;p_1}) < 1$。这四种确信度指标的陪域均为 $\mathbb{Q}_{>0} \cup \{\infty\}$。

Sit.	ISAR		GPAR（单图）	
IDE	$\Longrightarrow = \infty$	(P31)	$\Longrightarrow = \infty$	(P31)
DIS	$\Longleftrightarrow = 1 - \hat{P}_T(E_B)$	(P32)	$\Longleftrightarrow = 1 - \hat{P}_g(E_{g,p_2,V_2;p_1})$	(P32)
IND	$\Longleftrightarrow = 1$	(P33)	$\Longleftrightarrow = 1$	(P33)
POS	$\Longleftrightarrow > 1$	(P34)	$\Longleftrightarrow > 1$	(P34)
NEG	$\Longleftrightarrow < 1$	(P35)	$\Longleftrightarrow < 1$	(P35)
Sit.	**GPAR（微观）**		**GPAR（宏平均）**	
IDE	$\Longrightarrow = \infty$	(P31)	$\Longrightarrow = \infty$	(P36)
DIS	$\Longleftrightarrow = 1 - \hat{P}_G(E_{G,\ p_2,\ V_2;\ p_1})$	(P32)	$\Longrightarrow = \frac{1}{\lvert G\rvert}\sum_{g\in G}(1 - \hat{P}_g(E_{g,p_2,\ V_2;\ p_1}))$	(P37)
IND	$\Longleftrightarrow = 1$	(P33)	$\Longrightarrow = 1$	(P38)
POS	$\Longleftrightarrow > 1$	(P34)	$\Longrightarrow > 1$	(P39)
NEG	$\Longleftrightarrow < 1$	(P35)	$\Longrightarrow < 1$	(P40)

宏平均度量指标的另一个缺点是，如表 7 所示，得分有定义时所需满足的条件要严格得多，因为这些条件必须适用于 G 中的每个图。这些条件要求 G 是约简的，并且度量指标要在约简袋上进行评估，正如第4.3节所讨论的那样。

5 相关工作

相关工作来自三个领域：i）基于项集的关联规则概念的扩展，ii）关系数据库中函数依赖的扩展，以及 iii）霍恩规则。我们首先概述基于项集的关联规则的一般扩展，然后详细讨论最相关的概念。

5.1 关联规则的扩展

经典的关联规则概念——规则的形式为 $X \Rightarrow Y$，其中 X 和 Y 是项集，规则从事务数据库中挖掘，而事务是一个项集——已在多个方向上得到扩展。

- 时序关联规则 （参见 Zhou 和 Hirasawa [50]）包含时序信息，因此可以用于预测客户未来需求。例如，规则 $shoes\ (t = 0) \wedge cloth\ (t = 0)\ \Rightarrow$

表 7：度量指标得分有定义所需必要条件概览。

度量标准	条件
置信度$_T$	$\hat{P}_T(E_A) > 0$
置信度$_g$	$\hat{P}_g(E_{g,p_1,V_1;\ p_2}) > 0$
置信度$_G$	$\hat{P}_G(E_{G\ p_1,V_1;p_2}) > 0$
宏观置信度$_G$	$\forall g \in G : \hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}) > 0$
提升度$_T$	$\hat{P}_T(E_A) > 0 \wedge \hat{P}_T(E_B) > 0$
lift$_g$	$\hat{P}_g(E_{g,p_1,V_1;\ p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,V_2;\ p_1}) > 0$
提升度$_G$	$\hat{P}_G(E_{G\ p_1,V_1;p_2}) > 0 \wedge \hat{P}_G(E_{G\ p_2,V_2;p_1}) > 0$
宏提升度$_G$	$\forall g \in G : \hat{P}_g(E_{g,p_1,\ V_1;\ p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,\ V_2;p_1}) > 0$
杠杆率$_T$	$T \neq \varnothing$
杠杆率$_g$	$\mathcal{T}^n_{g,p_1,p_2} \neq \varnothing$
杠杆率$_G$	$\mathcal{T}^n_{G,p_1,p_2} \neq \varnothing$
宏观杠杆$_G$	$\forall g \in G : \mathcal{T}^n_{g,p_1,p_2} \neq \varnothing$
确信度$_T$	$\hat{P}_T(E_A) > 0 \wedge \hat{P}_T(E_B) < 1$
确信度$_g$	$\hat{P}_g(E_{g,p_1,\ V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,\ p_2,\ V_2;\ p_1}) < 1$
确信度$_G$	$\hat{P}_G(E_{G\ p_1,V_1;p_2}) > 0 \wedge \hat{P}_G(E_{G\ p_2,V_2;p_1}) < 1$
宏观确信度$_G$	$\forall g \in G : \hat{P}_g(E_{g,p_1,\ V_1;p_2}) > 0 \wedge \hat{P}_g(E_{g,p_2,\ V_2;p_1}) < 1$

$shoe\ oil$ ($t = 2$)表示客户通常会（在 $t = 0$）同时购买鞋和布料，并在两个月后（在 $t = 2$）购买鞋油。

• 一个相关概念是序列关联规则（参见 Srikant 等人、Fournier-Viger 等人 [45, 23]），其中事务是项集序列。例如，序列关联规则已应用于文本中的关系检测任务 [11]。举例来说，我们可以将“animals such as dogs”这样的句子转换为一个集合序列，每个集合都包含一个词元及其词性标注，即（$\{t{-}animals, p{-}NOUN\}$，$\{t{-}such, p{-}JJ\}$，$\{t{-}as, p{-}IN\}$，$\{t{-}dogs, p{-}NOUN\}$）。以下规则可检测出句子“animals such as dogs”中表达了下位词关系：（$\{p{-}NOUN\}$，$\{t{-}such, p{-}JJ\}$，$\{t{-}as, p{-}IN\}$，$\{p{-}NOUN\}$）$\Rightarrow hyponym$。

• 空间关联规则 （例如参见 Koperski 和 Han [30]）是形如 $X \Rightarrow Y$的规则，其中 X 和 Y是谓词集合，且某些谓词为空间谓词。例如，在以下规则中，出现了空间谓词 $within$、$adjacent$ 和 $closeTo$：$isA(X,\ city) \wedge within(X,\ BritishColumbia) \wedge adjacentTo(X,\ water) \Rightarrow closeTo(X,\ USA)$。这里，$X$是一个变量。该规则表示：如果存在某个事物，它是城市，位于不列颠哥伦比亚省之内，并且邻接水域，那么该事物就位于靠近美国的位置。

• 数值关联规则 （例如参见 Fukuda等人 [24, 29]）包含这样的陈述：在挖掘或应用规则时，需要对数值进行比较。例如，规则（$Balance \in [15821, 26264]$）$\Rightarrow(CardLoan = yes)$表示：如果客户的余额评分是数值区间[15821, 26264], 中的一个值，那么该客户很可能会使用信用卡贷款。

• 定量关联规则（例如，参见 Srikant 和 Agrawal [44]）与此相关，如以下规则示例所示：(Age: 30..39) $\wedge$($Married$: Yes) $\Rightarrow$($NumCars = 2$)，其中 Age和 $NumCars$是数值属性，$Married$ 是类别属性。在有关个人的数据库中，该规则表示，年龄在 30 至 39 岁且未婚的人通常拥有两辆车。Taboada 等人 [46] 的研究也与此相关，他们挖掘以下形式的规则：($A_1 > a_1$) $\wedge \ldots \wedge$ ($A_k > a_k$) $\Rightarrow$ ($A_m > a_m$) $\wedge \cdots \wedge$ ($A_n > a_n$)。其中一个挑战是找到合适的阈值。

• 而数值型和定量关联规则要求定义或确定清晰阈值，例如将量表划分为若干区间；模糊关联规则（参见 Hu 等人 [27]）则允许定义“高”和“低”等语言变量。以下是一个模糊关联规则的示例，摘自[4]: $Investment{-}in{-}construction = Very{-}high \Rightarrow Capital{-}of{-}transportation = Very{-}high$。该规则表示，建筑投资非常高往往会带来较高的交通资本。

模糊性的概念已与其他类型的关联规则相结合，例如形成了模糊时序关联规则（参见 Carinena [10]）。

• 与模糊关联规则类似，但并非基于模糊集隶属度的规则是渐进关联规则（例如，参见 Di-Jorio 等人 [16]）。与基于区间的规则（例如位于距市中心不到一公里处的房间价格在 *60* 到 *200* 美元之间）相比，渐进关联规则（如 $\{Pop^{\geq}, Dist^{\leq}\} \Rightarrow \{Price^{\geq}\}$）表达的是城镇越大、离市中心越近，价格就越高。

• 前面提到的关联规则类型关注的是提高规则的表达能力或改变规则的解释方式，而隐私保护关联规则（例如，参见 Saygin 等人 [41]）关注的问题则是：如何防止人们从事务数据库中学习到包含敏感数据的关联规则。

5.2 关系关联规则

针对更复杂事务的扩展涉及这样一种情形：给定一个被视为 Datalog 数据库的关系数据库。Dehaspe 和 Toivonen [14, 15] 提出了关系关联规则，之后 Goethals 和 Van den Bussche [25], 对其进行了扩展；这类规则以前件和后件的形式包含合取查询。例如，考虑以下两个用 Prolog 记号表示的合取查询：

$$Q_1(x, y) :- likes(x, \texttt{"Duvel"}), visits(x, y)$$
$$Q_2(x, y) :- likes(x, \texttt{"Duvel"}), visits(x, y), serves(y, \texttt{"Duvel"})$$

在第一个合取查询中，$Q_1(x, y)$ 是规则的头实体，而 $likes(x, \text{"Duvel"})$ $visits(x, y)$ 是规则的体。设 A、B、C 和 D 为常量，设 $likes(A, B)$ 为一个谓词，用来表示人物 A 喜欢饮料 B；设 $visits(A, C)$ 为一个谓词，用来表示人物 A 造访酒吧 C；设 $serves(C, D)$ 为一个谓词，用来表示酒吧 C 供应饮料 D。规则 $Q_1 \Rightarrow Q_2$ 可解释如下：如果一个喜欢饮料 Duvel 的人物 (x) 造访酒吧 (y)，那么该酒吧就供应饮料 Duvel。

我们可以将该关系关联规则表示为如下形式的基于图模式的关联规则 (p_1, p_2)：

$$r = (p_1, p_2, V)$$
$$p_1 = \{(v_1, likes, \text{"}Duvel\text{"}), (v_1, visits, v_2)\}$$
$$p_2 = \{(v_1, likes, \text{"}Duvel\text{"}), (v_1, visits, v_2), (v_2, serves, \text{"}Duvel\text{"})\}$$

前件合取查询和后件合取查询都可以视为图模式，区别在于变量不能出现在谓词位置。请注意，前件的主体包含于后件的主体中，这是这类规则的要求，但并非我们所引入的 GPAR 概念所要求的。其他差异与关系关联规则的求值语义有关。标准的合取查询求值语义与同态语义相当，而且合取查询不能包含形式为 $x \neq y$ or $x \neq A$ 的语句，其中 x 和 y 是变量，A 是常量；要实现类似无重复语义的求值，就需要添加此类语句。

合取查询的支持度定义为查询答案关系中的元组数量。对诸如 $Q_1(x, y) :- likes(x, \text{"Duvel"})$ $visits(x, y)$ 这样的合取查询进行求值时，变量 x 和 y 充当投影变量，所得关系是一组数对。扩展合取查询时，数对的数量不会增加，因此支持度度量是反单调的。

规则 $Q_1 \Rightarrow Q_2$ 的置信度定义为 Q_2 的答案关系中的元组数量除以 Q_1的答案关系中的元组数量。

我们可以将关系数据库表示为图 g。令 $\mathcal{T}_g^n$为由关系数据库中出现的常量组成的长度为 $n = |V|$ 的元组集合（不一定是无重复的）。使用我们的记号，其绝对支持度和置信度度量指标表示如下：

$$\text{absolute-support}_g(p, V) := |\{\, T \in \mathcal{T}_g^n \mid \exists \mu \in \Omega_{p,g}^{hom} : \mu(V) = T \,\}|$$
$$\text{confidence}_g(p_1, p_2, V) = \frac{|\{\, T \in \mathcal{T}_g^n \mid \exists \mu_1 \in \Omega_{p_1,g}^{hom} : \mu_1(V) = T \wedge \exists \mu_2 \in \Omega_{p_2,g}^{hom} : \mu_2(V) = T \,\}|}{|\{\, T \in \mathcal{T}_g^n \mid \exists \mu \in \Omega_{p_1,g}^{hom} : \mu(V) = T \,\}|}$$

这表明，他们的度量指标和置信度与我们的置信度度量指标之间存在一定程度的相似性，因为给定一个连接变量序列后，两个图模式之间对应关系的强度是根据共有项序列来评分的。

5.3 图实体依赖

由 Codd [13], 提出的函数依赖（*FD*）概念在关系数据库领域广为人知，函数依赖可确保数据一致性并指导模式规范化。

给定一个属性集合 $A = \{A_1, \ldots, A_n\}$，其定义域分别为 D_1，$\ldots D_n$，设 X，$Y \subseteq A$ 是两个属性集合，$R \subseteq D_1 \times \cdots \times D_n$ 是一个关系。函数依赖 $X \to Y$ 表示，对于任意两个元组 t_1，$t_2 \in R$，如果这两个元组在 X 中的属性上取值相同，那么它们在 Y 中的属性上也必须取值相同，即

$$\forall t_1, t_2 \in R : (t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y])$$

例如，设客户由电话号码国家代码（CC）、电话号码区号（AC）、电话号码（PN）、姓名（NM）、街道（STR）、城市（CT）和邮政编码（ZIP）来描述。因此，$R \subseteq D_{CC} \times D_{AC} \times \cdots \times D_{STR}$。函数依赖 $[CC, AC] \to [CT]$ 表示

$$\forall t_1, t_2 \in R : (t_1[CC, AC] = t_2[CC, AC] \Rightarrow t_1[CT] = t_2[CT]).$$

因此，电话号码国家代码和电话号码区号相同的客户居住在同一城市。Yao 等人 [49] 研究了从关系数据中挖掘函数依赖的问题。为了刻画更细粒度的数据关系，Bohannon 等人 [5], 将函数依赖的概念推广为条件函数依赖（*CFD*），允许对属性值指定条件:

$$\forall t_1, t_2 \in R : ((t_1[X] = t_2[X] \wedge C_1(t_1[X])) \Rightarrow (t_1[Y] = t_2[Y] \wedge C_2(t_1[Y])))$$

这里，C_1 和 C_2 是谓词，条件满足时其值为真，否则为假。例如，CFD$[CC =$"01" $AC =$"908" $PN] \to [STR, CT =$" MH" $ZIP]$ 表示

$$\begin{aligned}\forall t_1, t_2 \in R : ((& \\ & t_1[CC, AC, PN] = t_2[CC, AC, PN] \wedge \\ & t_1[CC] = \texttt{"01"} \wedge \\ & t_1[AC] = \texttt{"908"}) \Rightarrow (\\ & t_1[STR, CT, ZIP] = t_2[STR, CT, ZIP] \wedge \\ & t_1[CT] = \texttt{"MH"})).\end{aligned}$$

这里，$C_1 \equiv t_1[CC] =$"01"、$\wedge\, t_1[AC] =$"908"和 $C_2 \equiv t_1[CT] =$"MH"。因此，电话号码国家代码（即 01）、电话号码区号（即 908）以及电话号码相同的客户住在同一条街道上，居住在同一城市（即 MH），并且拥有相同的邮政编码。

Fan 等人 [17] 研究了从关系数据中挖掘条件函数依赖的问题。函数依赖和条件函数依赖虽然定义于关系数据的语境中，但近期的研究已将依赖的概念扩展至图结构数据。在此背景下，研究者提出了图函数依赖（*GFD*）[22, 19, 20], ，但其适用对象是类似于属性图的一类图 [6]。Fan 和 Lu [20], 将 GFD 扩展为图实体依赖（GED），后者涵盖了 GFD。

GED 用于检测不一致性，是一种一致性约束。给定一个无限的标签集合 Γ、一个无限的属性集合 Υ 以及一个无限的值集合 U，GED 定义于形式为 $G = (V, E, L, F_A)$ 的属性图，其中 V 是有限的节点标识符集合，$E \subseteq V \times \Gamma \times V$ 是有限的边集合；$(v, \iota, v') \in E$ 表示一条从节点 v 指向节点 v'、标签为 ι 的边；每个节点 $v \in V$ 都带有一个标签 $L(v)$；对于每个节点 $v \in V$，$F_A(v)$ 都是一个有限元组，其形式为 $(A_1 = a_1, \ldots, A_n = a_n)$，其中 $A_1, \ldots, A_n$ 是 Υ 中的属性，$a_1, \ldots, a_n$ 是 U 中的常量，且 $A_i \neq A_j$ 当 $i \neq j$ 时成立。我们用 $v.A_i$ 表示节点 v 的属性 A_i 的值。最后，每个节点 $v \in V$ 都有一个特殊属性 id，用于存储该节点的标识符。

以下示例图表示：人物 Alice 与人物 Bob 已婚，且 Alice 的家庭住址是比勒费尔德:

$$\begin{aligned}\Gamma &= \{married To\} \\ \Upsilon &= \{id, name, homeAddress\} \\ U &= \{v_1, v_2, \texttt{"Alice"}, \texttt{"Bob"}, \texttt{"Bielefeld"}\} \\ G &= (V, E, L, F_A) \\ V &= \{v_1, v_2\} \\ E &= \{(v_1, married To, v_2)\}\end{aligned}$$

$$L = \{(v_1, person), (v_2, person)\} F_A(v_1) = \{(id, v_1), (name, \text{"Alice"}), (homeAddress, \text{"Bielefeld"})\} F_A(v_2) = \{(id, v_2), (name, \text{"Bob"})\}$$

图模式是一个图 $Q[\bar{x}] = (V_Q, E_Q, L_Q)$，其中 V_Q 是节点标识符的有限集合， $E_Q \subseteq V_Q \times (\Gamma \cup \{\text{"\_"}\}) \times V_Q$ 是边的有限集合，其中 $(v, \iota, v') \in E_Q$ 表示一条标记为 ι 的边，其中 $\iota = \text{"\_"}$ 表示通配符；L_Q 是一个函数，为每个节点分配一个来自 $\Gamma \cup \{\text{"\_"}\}$ 的标签。 $\bar{x}$ 是一个由互不相同的变量组成的列表，每个变量都指向 V_Q中的另一个节点。

在上述图的上下文中，下面给出了一个图模式的示例 $Q[\bar{x}]$。这里，我们将每个标识符 x_i 解释为一个变量。因此，在该图模式中，所有节点都是变量：

$$\begin{aligned} Q[\bar{x}] &= (V_Q, E_Q, L_Q) \\ \bar{x} &= (x_1, x_2) \\ V_Q &= \{x_1, x_2\} \\ E_Q &= \{(x_1, marriedTo, x_2)\} \\ L_Q &= \{(x_1, person), (x_2, person)\} \end{aligned}$$

最后，我们可以将图编辑依赖（GED）定义为 $Q[\bar{x}](X \Rightarrow Y)$，其中 $Q[\bar{x}]$ 是一个图模式，而 X 和 Y 是 $\bar{x}$的（可能为空的）文字项集合。 $\bar{x}$ 的文字项可以是常量文字、变量文字或标识符文字。常量文字的形式为$x.A = c$，其中 $x \in \bar{x}$ 是一个变量， $A \in \Upsilon$（具有 $A \neq id$）是一个属性，而 $c \in U$ 是一个常量。变量文字的形式为 $x.A = y.B$，其中 x， $y \in \bar{x}$ 是变量， A， $B \in \Upsilon$（具有 $A \neq id$和 $B \neq id$）是属性。标识符文字的形式为 $x.id = y.id$，其中 x， $y \in \bar{x}$ 是变量。

综合上述内容，对于示例图，我们可以定义如下图编辑依赖 ϕ:

$$\begin{aligned} \phi &= Q[\bar{x}](X \Rightarrow Y) \\ Q[\bar{x}] &= (V_Q, E_Q, L_Q) \\ \bar{x} &= (x_1, x_2) \\ V_Q &= \{x_1, x_2\} \\ E_Q &= \{(x_1, marriedTo, x_2)\} \\ L_Q &= \{(x_1, person), (x_2, person)\} \\ X &= \varnothing \\ Y &= \{x_1.homeAddress = x_2.homeAddress\} \end{aligned}$$

该图编辑依赖表达的是：如果两个人已婚，那么预期他们的家庭住址相同。在图上评估该图编辑依赖时，会发现不一致，即存在一对个人（v_1、 v_2），他们的姓名分别是 v_1Alice 和 v_2 Bob，其中 v_1 与 v_2 已婚，但家庭住址并不相同——事实上， v_2 的家庭住址缺失。尽管图编辑依赖并非用于扩展图，但仍可以 在有限范围内用于扩展属性图。此处，由于 v_2 的家庭住址缺失，可以通过将 v_1 的家庭住址复制给 v_2来纠正这种情况。虽然这一过程可以通过添加缺失的属性值来扩展图，但无法通过添加缺失的节点或边来扩展图。

5.4 图关联规则

Fan 等人提出的图关联规则（GAR） [18] 是图模式关联规则（GPAR）[21] 和图函数依赖（GFD） [22]概念的扩展。图关联规则可用于识别缺失信息和预测链接。给定一个无限的标签集合 Γ 、一个无限的属性集合 Υ 和一个无限的值集合 U，GAR 概念针对形式为（V, E, L, F_A）的图定义，与图实体依赖（见第 5.3 节）中的定义一致，不同之处在于不使用特殊的 id 属性： V是有限的节点标识符集合， $E \subseteq V \times \Gamma \times V$是有限的边集合，其中（$v, \iota, v'$）$\in E$表示一条从节点 v 指向节点 v'、标签为 ι的边；每个节点 $v \in V$都带有一个标签L（v）；对于每个节点 $v \in V$， F_A（v）都是一个有限元组，形式为（$A_1 = a_1, \ldots, A_n = a_n$），其中 $A_1, \ldots, A_n$ 是 Υ中的属性， $a_1, \ldots, a_n$ 是 U 中的常量，并且当 $i \neq j$时 $A_i \neq A_j$。我们用 $v.A_i$ 表示节点 v的属性A_i 的值。

图模式是一个图 $Q[\bar{x}] = (V_Q, E_Q, L_Q, \mu)$，其中 V_Q 是一个有限的节点标识符集合， $E_Q \subseteq V_Q \times (\Gamma \cup \{\text{"\_"}\}) \times V_Q$ 是一个有限的边集合，其中 $(v, \iota, v') \in E_Q$ 表示一条带有 ι 标签的边，其中 $\iota = \text{"\_"}$ 表示一个

通配符，而 L_Q 是一个函数，为每个节点分配一个来自 $\Gamma \cup \{"\_"\}$ 的标签。 $\bar{x}$ 是一个由互不相同的变量组成的列表，而 μ是一个双射，为 V_Q中的每个节点分配一个不同的变量。

图关联规则是一种形如 $Q[\bar{x}](X \to Y)$ 的表达式，其中 $Q[\bar{x}]$ 是图模式， $\bar{x}$ 是在 $Q[\bar{x}]$ 中出现的变量的非重复列表，X 和 Y 是 $Q[\bar{x}]$ 上的文字项集合。 $Q[\bar{x}]$ 上的文字项可以是属性文字项、边文字项、 ML 文字项、变量文字项或常量文字。此处我们忽略 ML 文字项。属性文字项的形式为 $x.A$，其中 x 是 $\bar{x}$ 中的变量， A 是 Υ 中的属性。边文字项的形式为 $\iota(x,\ y)$，其中 ι 是 Γ 中的标签， x 和 y 是 $\bar{x}$ 中的变量。变量文字项的形式为 $x.A = y.B$，其中 x 和 y 是 $\bar{x}$ 中的变量， A 和 B 是 Υ 中的属性。常量文字的形式为 $x.A = c$，其中 x 是 $\bar{x}$ 中的变量， A 是 Υ 中的属性， c 是 U 中的常量。

图模式按照同态语义进行评估。模式 $Q[\bar{x}]$ 与图 G 的匹配，是从 $Q = (V_Q, E_Q, L_Q, \mu)$ 到 $G = (V, E, L, F_A)$ 的映射 h，满足：对于每个节点 $u \in V_Q$， $L_Q(u) = L(h(u))$；对于每条边 $(u, \iota, u') \in E_Q$ 且带有 $\iota \neq "\_"$，$(h(u)\ \iota, h(u')) \in E$。对于每条边 $(u, "\_"\ u') \in E_Q$，都需要存在某个$\iota \in \Gamma$，使得 $(h(u)\ \iota, h(u')) \in E$。

Y中的边文字项使图关联规则（GAR）能够用于扩展图。下图表示：一位姓名为 Alice 的人物与一位姓名为 Bob 的人物已婚，且 Alice 居住在一座名为比勒费尔德的城市。

$$
\begin{aligned}
\Gamma &= \{person,\ city,\ marriedTo,\ livesIn\} \\
\Upsilon &= \{name\} \\
U &= \{\texttt{"Alice"},\ \texttt{"Bob"},\ \texttt{"Bielefeld"}\} \\
G &= (V,\ E,\ L,\ F_A) \\
V &= \{A,\ B,\ C\} \\
E &= \{(A,\ marriedTo,\ B),\ (A,\ livesIn,\ C)\} \\
L &= \{(A,\ person),\ (B,\ person),\ (C,\ city)\} \\
F_A(A) &= \{(name,\ \texttt{"Alice"})\} \\
F_A(B) &= \{(name,\ \texttt{"Bob"})\} \\
F_A(C) &= \{(name,\ \texttt{"Bielefeld"})\}
\end{aligned}
$$

以下 GAR 表示：如果人物 x_1 与人物 x_2 处于婚姻关系，并且人物 x_1 居住在城市 x_3，那么 x_2 居住在城市 x_3。

$$
\begin{aligned}
Q[\bar{x}] &= (V_Q,\ E_Q,\ L_Q,\ \mu) \\
\bar{x} &= (x_1,\ x_2,\ x_3) \\
V_Q &= \{A,\ B,\ C\} \\
\mu &= \{(x_1,\ A),\ (x_2,\ B),\ (x_3,\ C)\} \\
E_Q &= \{(x_1,\ marriedTo,\ x_2),\ (x_1,\ livesIn,\ x_3)\} \\
L_Q &= \{(x_1,\ person),\ (x_2,\ person),\ (x_3,\ city)\} \\
X &= \varnothing \\
Y &= \{(x_2,\ livesIn,\ x_3)\}
\end{aligned}
$$

因此，我们可以扩展该图，使其也表示 Bob 居住在比勒费尔德。

尽管 GAR 概念是对用于识别数据不一致性的概念的扩展，但图关联规则可以用于扩展图。本文的贡献之一是提出了基于图模式的关联规则这一概念，这些规则可用于扩展图。GAR 概念针对属性图定义，而 GPAR 概念针对有向标记多重图定义。GAR 中的模式依据同态语义进行评估，而 GPAR 中的模式依据无重复语义进行评估，因此可以更加关注图的拓扑结构。GAR 中的图模式不能在谓词位置包含变量，而 GPAR 中的图模式则可以；这不仅可以编写更通用的模式（例如，允许某条边具有任意标签），还可以表达这样的约束：两条边可以具有任意标签，只要两个标签相同或不同；也可以利用图中表示的关系信息，例如某个谓词被声明为对称，如图4所示。除结构上的差异外，本文还研究了单图设置和图事务环境，并提出和分析了用于评估模式得分以及衡量两个图模式之间相关性强度的度量指标。据我们所知，针对 GAR 概念尚未提出等价的度量指标。不过，对于 GAR 所依据的概念，文献 [21],中定义了支持度和置信度，但尚不清楚它们是否也适用于图关联规则。

5.5 基于图模式的关联规则

Wang等人提出了图模式关联规则这一概念 [47, 48], ，它不同于 Fan 等人 [21], 为社交网络分析提出的同名概念；后者的规则可用于推荐好友，但并不局限于社交网络。他们将一种属性图定义为有向节点标记图，其形式为元组 (V, E, L)，其中 V是节点标识符的集合， $E \subseteq V \times V$是有向边的集合， L 是一个函数，它将每个节点 $v \in V$映射到一个元组 $L(v)$，其形式为 $(A_1 = a_1, \ldots, A_n = a_n)$，其中 $A_1, \ldots, A_n$是互不相同的属性名称， $a_1, \ldots, a_n$是属性值。图模式 p定义为一个元组 (V_p, E_p, f)，其中 V_p是节点标识符的集合， $E_p \subseteq V_p \times V_p$是边的集合， f是一个函数，它将每个节点 $v \in V_p$映射到一个元组 $f(v)$，其形式为 $(A_1 = a_1, \ldots, A_m = a_m)$，其中 $A_1, \ldots, A_m$是互不相同的属性名称， $a_1, \ldots, a_m$是属性值。如果存在一个从 V_p到 V的单射 h，使得对于每个 $(v, v') \in h$，以及每个语句 $A_i = a_i \in L(v)$，都有 $A_i = a_i \in L(v')$，并且对于每个 $(v, v') \in E_p$，都有 $(h(v)\ h(v')) \in E$，那么图模式 (V_p, E_p, f) 就匹配图 (V, E, L)。因此，该图模式依据无重复节点语义进行评估，我们可以将 h 称为匹配。不过，由于边没有标签，无重复节点语义与无重复语义是一致的。图模式关联规则是形如 $Q_l \Rightarrow Q_r$ 的陈述，其中 Q_l 和 Q_r 是图模式。该规则表达的是：如果 Q_l 匹配图 G，那么 Q_r 也匹配 G。具体而言，如果存在一个从 Q_l 到 G的匹配 h_l，那么就存在一个从 Q_r 到 G 的匹配 h_r，使得对于每个 $u \in V_l \cap V_r$，都有 $h_l(v) = h_r(v)$——因此，可以通过匹配之间的共享变量对应关系来表达对应关系。

图模式的支持度采用基于最小像的支持度度量 [9] 定义。设图 $G = (V, E, L)$ 上图模式 $Q = (V_Q, E_Q, f)$ 的所有匹配集合记为 $\Omega_{Q,G}$。令 $img(u)$ 表示 V 中可通过某个匹配 $h \in \Omega_{Q,G}$ 映射到节点 $u \in V_Q$ 的节点集合，即 $img(u) = \{u' \in V \mid \exists h \in \Omega_{Q,G} : h(u) = u'\}$。给定图 G，图模式 Q 的支持度定义为 V_Q 中这些节点对应集合的最小基数，即 $supp(Q, G) = \min_{u \in V_Q} |img(u)|$。给定 GPAR $Q_l \Rightarrow Q_r$，其中包含 $Q_l = (V_l, E_l, f_l)$ 和 $Q_r = (V_r, E_r, f_r)$，构造图模式 Q_R 为 $(V_l, E_l \cup E_r, f_l)$。规则 $R = Q_l \Rightarrow Q_r$ 的置信度定义为 $supp(Q_R, G)/supp(Q_l, G)$。

与我们的方法不同之处在于，另一种类型的图被纳入考虑：节点带有属性-值对，多个节点可以具有完全相同的属性-值对集合，边没有标签，并且两个节点之间最多只能有一条边。与我们的方法相似之处在于所使用的求值语义，因为由于边没有标签，此处的*nrn*和*nra* 语义是一致的。另一个相似之处是，前件和后件都是图模式（不限制这些模式必须是路径形状，例如），并且匹配之间的对应关系可以通过共享变量表示。支持度的衡量方式存在显著差异；此外，置信度不仅基于显著不同的支持度度量，也不衡量基于共享变量的匹配之间对应关系的强度。最后，这些度量指标是针对单图设置而非图事务环境开发的，并且其定义并未使用概率。

5.6 路径关联规则

Sasaki 和 Karras [40] 提出了路径关联规则的概念，并对支持度、置信度和提升度这些度量指标进行了调整。为清晰起见，我们对其图定义做了稍微的调整。给定一个标签集合 $\mathcal{L}$ 和一个属性集合 $\mathcal{A}$，该概念针对形式为（$V, E, src, tgt, \lambda_V, \lambda_E$）的图定义，其中 V是节点标识符集合， E是边标识符集合，src 是将边标识符映射到该边源节点标识符的函数， tgt 是将边标识符映射到该边目标节点标识符的函数， λ_V是将每个节点标识符 v 映射到属性集合 $A(v) \subseteq \mathcal{A}$ 的函数，而 λ_E 是将每个边标识符 e映射到单个标签 $\lambda_E(e) \in \mathcal{L}$的函数。
路径是形如 $(v_0,\ e_0,\ v_1,\ \ldots\ldots,\ e_{n-1},\ v_n)$ 的序列，其中 $v_i \in V$ 是节点标识符， $e_i \in E$ 是边标识符。这里引入两种路径模式：简单路径模式和可达路径模式——本文重点讨论简单路径模式，并简称为路径模式。路径模式是形如 $(A_0,\ l_0,\ A_1,\ \ldots\ldots,\ l_{n-1},\ A_n)$ 的序列，其中 $A_i \subseteq \mathcal{A}$ 是属性集合， $l_i \in \mathcal{L}$ 是边标签。

如果对于所有 i， $A_i \subseteq A(v_i)$ 且 $l_i = l(e_i)$，则路径模式 $(A_0,\ l_0,\ A_1,\ \ldots\ldots,\ l_{n-1},\ A_n)$ 与路径 $(v_0,\ e_0,\ v_1,\ \ldots\ldots,\ e_{n-1},\ v_n)$ 匹配。给定图 G、路径 p 和节点标识符 v，$\mathcal{V}(p)$ 是这样一组节点标识符：对于其中每个标识符，都存在一条路径，该节点是路径的源节点，且该路径与路径模式匹配。
路径关联规则是一种如下形式的表达式： $p_x \Rightarrow p_y$ 其中 p_x 和 p_y 是路径模式。规则 $p_x \Rightarrow p_y$ 的绝对支持度定义为 $|\mathcal{V}(p_x) \cap \mathcal{V}(p_y)|$，相对支持度定义为 $|\mathcal{V}(p_x) \cap \mathcal{V}(p_y)|/|V|$，置信度定义为 $|\mathcal{V}(p_x) \cap \mathcal{V}(p_y)|/|\mathcal{V}(p_x)|$，提升度定义为$(|\mathcal{V}(p_x) \cap \mathcal{V}(p_y)| \cdot |V|)/(|\mathcal{V}(p_x)| \cdot |\mathcal{V}(p_y)|)$。尽管他们的图形式体系与我们的 DLM 形式体系有所不同，但这种情况类似于具有一个基于图模式的关联规则（p_1, p_2, V_1, V_2），其中 p_1 和 p_2 是线形图模式，不存在变量出现在谓词位置的情况，并且 $V_1 = V_2 = (v_1)$仅包含一个变量 v_1，它是两个模式的共同源头，且这些模式是在同态语义下进行评估的。

5.6.1 多关系关联规则（MRAR）

Ramezani、Saraee 和 Nematbakhsh 于 2014年 [38, 39, 37] 提出了多关系关联规则（MRAR）的概念。规则从单个有向标记多重图中挖掘而来，该图可以是 RDF 图，也可以是由关系数据库创建的图。

以下示例规则的含义是“居住在邻近气候类型为潮湿的城市的地点，且年龄小于 *20* 岁的人，健康状况良好：”

LiveIn(NearTo(气候类型(潮湿))), 年龄小于(20) -> 健康状况(Goo d)

前件是一组关系链（此处：LiveIn(NearTo(ClimateType(Humid))) 和 AgeLessThan(20) 是关系链），后件则是单条关系链（此处该关系链为 HealthCondition(Good)）。每条关系链都以一个实体结尾（此处为 Humid、20 和 Good），作者将其称为端点实体。我们可以将这样的关系链表示为图模式，并将上述规则表示为如下形式的基于图模式的关联规则（p_1，p_2，V_1，V_2）：

$$p_1 = \{(v_j, LiveIn, v_1)、(v_1, NearTo, v_2)、(v_2, ClimateType, Humid)、(v_j, AgeLessThan, 20)\} p_2 = \{(v_j, HealthCondition, Good)\} V_1 = (v_j) V_2 = (v_j)$$

请注意特殊变量 v_j（j 的用法，它表示 v_j 是连接变量），该变量出现在每条关系链的源端。作者定义了图模式（他们将其称为 L-大型项链）的相对支持度，以及规则的置信度。我们可以用如下记号形式化这些度量指标，其中 $\mathcal{T}_g$ 表示出现在图 g 中的节点集合：支持度是与端点通过关系链相连的实体数量，除以图中的实体数量。

$$support_g(p_1, p_2, (v_j)) = \frac{|\{\, t \in \mathcal{T}_g \mid \exists \mu_1 \in \Omega^{hom}_{p_1,g} : (v_j, t) \in \mu_1 \,\}|}{|\mathcal{T}_g|}$$

$$confidence_g(p_1, p_2, (v_j)) = \frac{|\{\, t \in \mathcal{T}_g \mid \exists \mu_1 \in \Omega^{hom}_{p_1,g} : (v_j, t) \in \mu_1 \wedge \exists \mu_2 \in \Omega^{hom}_{p_2,g} : (v_j, t) \in \mu_2 \,\}|}{|\{\, t \in \mathcal{T}_g \mid \exists \mu_1 \in \Omega^{hom}_{p_1,g} : (v_j, t) \in \mu_1 \,\}|}$$

MRAR 形式体系与本文介绍的 GPAR 形式体系之间的主要差异在于：MRAR 中的图模式比 GPAR 中的图模式受到更多限制：在 MRAR 中，谓词位置不允许出现变量，实体只能出现在三元组模式的宾语位置，关系链必须以端点实体结束，不能以变量结束。由于 MRAR 使用关系链时节点变量是隐含的，因此无法表达诸如“居住在父亲的出生国”之类的约束。因此，图模式是星状树。此外，MRAR 中只能有一个连接变量，并且 MRAR 模式是在同态语义下进行评估的。MRAR 规则的后件是单条关系链，尽管 MRAR 挖掘方法很可能可以轻松扩展，使后件由多条关系链组成。

MRAR 规则和 GPA 规则的评分方式不同：MRA 规则针对单个图进行评分，而 GPA 规则既可以针对单个图进行评分，也可以针对图袋进行评分。它们对支持度和置信度的定义可以采用概率空间的表述。因此，可以定义提升度、杠杆率和确信度等度量指标。

5.7 霍恩规则

根据 [31], 霍恩规则 R 是形如 $R = B_1 \wedge \cdots \wedge B_n \Rightarrow H$ 的陈述，其中 $B_1 \wedge \cdots \wedge B_n$ 是体原子的合取，H 是头原子，$\Rightarrow$ 表示逻辑蕴含。原子是形如 $r(X, Y)$ 的陈述，其中 r 是关系，X 和 Y 或者是常量，或者是变量。通常，在霍恩规则挖掘的语境中，还会施加进一步的约束，即关于规则的结构，要求规则必须是闭合且连通的。如果每个变量都至少出现在两个原子中，则霍恩规则是闭合的 。如果两个原子共享一个变量或常量，则称它们是连通的。如果规则中的每个原子都与该规则中的其他任一原子传递连通，则该规则是连通的 。对于规则实例化时的情形，还可以施加进一步的约束；实例化是指将规则体原子中的变量替换为图中的实体的过程。在此，对象标识 [42, 32] 约束要求每个变量都必须替换为不同的变量。

现有的霍恩规则挖掘工作 [34, 33, 32, 31, 26] 着重于从单个知识图谱（即资源描述框架图）的形式中，挖掘满足上述部分约束的霍恩规则——因此，仅考虑单图设置。

尽管霍恩规则可以看作由体图模式和头图模式构成，但两者之间存在一些差异：i）两种模式的谓词位置都不允许出现变量；ii）体图模式必须是连通的；iii）头图模式只能由单个三元组模式构成。此外，尽管对象标识比同态语义的约束更严格，但其约束仍弱于*nra* 语义。

正如我们所定义的，基于图模式的关联规则属于蕴含规则。然而，由于前面提到的差异，图模式关联规则并非霍恩规则。不过，每条霍恩规则都可以表示为一条图模式关联规则。最后，由于霍恩规则是在单图中挖掘的，因此用于评估霍恩规则的度量指标只考虑单图设置。

5.8 语义网规则形式体系

过去人们提出了许多规则形式体系，例如在语义网技术领域。就本文而言，将基于图模式的关联规则表示为以下形式的元组较为方便：(p_1，p_2，V_1，V_2)，或 (p_1，p_2)。

不过，我们可以依托现有标准，来描述和评估形式为 (p_1，p_2) 的基于图模式的关联规则中的图模式。本节将讨论如何使用***SPARQL 1.1*** 查询语言[12] 和语义网规则语言（***SWRL***）来表达基于图模式的关联规则，并确保其满足无重复语义。[13]

在第 2.2 节讨论的教育社会学示例基础上，我们定义一条规则，明确指出：对于与模式 p_2（如图 2 所示：某人物居住在其父亲的出生国，而该国不同于其母亲的出生国）匹配的人物，该人物属于某个类 $ClassX$。该规则的简化形式如下：

$$p_1 = \{(v_1, hF, v_2), (v_1, hM, v_3), (v_1, cor, v_4), (v_2, cob, v_4), (v_3, cob, v_5)\}$$
$$p_2 = \{(v_1, type, ClassX)\}$$

将规则表示为 SPARQL CONSTRUCT 查询时，我们需要将节点和边的标签转换为统一资源标识符。为便于阅读，我们在此使用紧凑的统一资源标识符，因此为其添加示例命名空间 ex。于是，cor（居住国）变为 ex:cor。然后，规则可表示为 SPARQL CONSTRUCT 查询：

```
PREFIX ex:  <http://example.org#>CONSTRUCT { ?v1 ex:type ex:ClassX . }
WHERE { ?v1 ex:hF ?v2 . ?v1 ex:hM ?v3 . ?v1 ex:cor ?v4 . ?v2 ex:cob ?v4 . ?v3
ex:cob ?v5 . FILTER ( ?v1 != ?v2&& ?v1 != ?v3 && ?v1 != ?v4 && ?v1 != ?v5 && ?
v2 != ?v3 && ?v2 != ?v4&& ?v2 != ?v5 && ?v3 != ?v4 && ?v3 != ?v5 && ?v4 != ?
v5 && ?v1 != ex:hM&& ?v1 != ex:hF && ?v1 != ex:cob && ?v1 != ex:cor && ?v2 !
= ex:hM&& ?v2 != ex:hF && ?v2 != ex:cob && ?v2 != ex:cor && ?v3 != ex:hM&& ?
v3 != ex:hF && ?v3 != ex:cob && ?v3 != ex:cor && ?v4 != ex:hM&& ?v4 !=
ex:hF && ?v4 != ex:cob && ?v4 != ex:cor && ?v5 != ex:hM&& ?v5 != ex:hF && ?
v5 != ex:cob && ?v5 != ex:cor ) }
```

FILTER 表达式确保图模式在*nra* 语义下进行求值。没有该筛选条件时，图模式则在*hom*语义下进行求值。此外，该规则可以表示为一条 SWRL 规则：

```
ex:hF(?v1, ?v2) ^
ex:hM(?v1, ?v3) ^
ex:cor(?v1, ?v4) ^
ex:cob(?v2, ?v4) ^
```

[12]https://www.w3.org/TR/sparql11-query/[13]
https://www.w3.org/Submission/SWRL/

ex:cob(?v3, ?v5) ~swrlb:notEqual(?v1, ?v2) ^ swrlb:notEqual(?v1, ?v3) ^ swrlb:notEqual(?v1, ?v4) ^ swrlb:notEqual(?v1, ?v5) ~swrlb:notEqual(?v2, ?v3) ^ swrlb:notEqual(?v2, ?v4) ~swrlb:notEqual(?v2, ?v5) ^ swrlb:notEqual(?v3, ?v4) ~swrlb:notEqual(?v3, ?v5) ^ swrlb:notEqual(?v4, ?v5) ~swrlb:notEqual(?v1, ex:hM) ^ swrlb:notEqual(?v1, ex:hF) ^ swrlb:notEqual(?v1, ex:cob) ^ swrlb:notEqual(?v1, ex:cor) ^ swrlb:notEqual(?v2, ex:hM) ^ swrlb:notEqual(?v2, ex:hF) ~swrlb:notEqual(?v2, ex:cob) ^ swrlb:notEqual(?v2, ex:cor) ~swrlb:notEqual(?v3, ex:hM) ^ swrlb:notEqual(?v3, ex:hF) ~swrlb:notEqual(?v3, ex:cob) ^ swrlb:notEqual(?v3, ex:cor) ~swrlb:notEqual(?v4, ex:hM) ^ swrlb:notEqual(?v4, ex:hF) ~swrlb:notEqual(?v4, ex:cob) ^ swrlb:notEqual(?v4, ex:cor) ~swrlb:notEqual(?v5, ex:hM) ^ swrlb:notEqual(?v5, ex:hF) ^ swrlb:notEqual(?v5, ex:cob) ^ swrlb:notEqual(?v5, ex:cor) -> 资源框架类型(?v1, ex:ClassX)

swrlb:notEqual 语句确保图模式是在*nra* 语义下求值的——在这里，它的作用与 SPARQL CONSTRUCT 查询中的筛选器类似。

但请注意，在基于图模式的关联规则中，我们允许变量出现在谓词位置。这在 SPARQL 查询和 N3 逻辑规则中是允许的，[14] 但在 SWRL、规则交换格式、[15] Datalog[12, 1], 和 Horn 规则中则不允许。因此，每条基于图模式的关联规则都可以用 SPARQL 表达，但并非每条规则都能用 SWRL、RIF 或 Datalog 表达。

6 总结与结论

有向标记多重图（DLM）的集合在许多领域都具有重要意义，例如社交网络、通信网络、引文网络、交通网络、贸易网络、生物相互作用网络、分子结构和化学反应网络。因此，适用于生成任务（即扩展给定图）或评估任务（即评估给定图的合理性）的方法也同样重要。这些任务可以通过基于规则的方法来实现；与基于人工神经网络的方法相比，这类方法具有可解释性，因此尤为引人关注。在概述经典的基于项集的关联规则及相关度量指标（置信度、提升度、杠杆率和确信度）、有向标记多重图以及图模式求值等概念，并特别关注无重复语义之后，我们正式介绍了基于图模式的关联规则（GPAR）这一概念，探讨了其解释方式，并展示了如何将其用于生成任务（例如链接预测和一般的图扩展）以及评估任务（例如合理性分析）。

我们定义了一个概率空间，并制定了置信度、提升度、杠杆率和确信度的概率版本，其中区分了单图设置与图事务环境；针对后者，我们引入了微平均度量和宏平均度量。我们展示了如何将 GPAR 度量指标的评估问题重新表述为基于项集的关联规则对应度量指标的评估问题。我们详细考察了这些度量指标，并描述了它们与经典基于项集的对应指标之间的关系，确定了关键性质得以保留或发生分歧的条件。

我们的分析表明，我们的框架能力超越了广泛的相关形式体系，包括图函数依赖、图实体依赖、关系关联规则、图关联规则、多关系关联规则、路径关联规则和霍恩规则；同时还表明，GPAR 可以表示为 SPARQL CONSTRUCT 查询。

未来的工作将聚焦于在不重复任何内容语义下高效挖掘频繁图模式，以及高效挖掘基于图模式的关联规则。

[14]https://www.w3.org/TeamSubmission/n3/[15]
https://www.w3.org/TR/rif-core/

致谢

本研究得到挪威研究理事会卓越中心 Integreat——挪威知识驱动机器学习中心的支持，项目编号为 332645。

参考文献

[1] Serge Abiteboul、Richard Hull 和 Victor Vianu。《数据库基础》，第 8 卷。Addison-Wesley Reading，1995 年。[2] Rakesh Agrawal 和 Ramakrishnan Srikant。大型数据库中关联规则挖掘的快速算法。载于超大型数据库国际会议第 *20* 届会议论文集，VLDB '94，第 487–499 页，美国加利福尼亚州旧金山，1994 年。Morgan Kaufmann 出版社有限公司。[3] Renzo Angles、Marcelo Arenas、Pablo Barceló、Aidan Hogan、Juan Reutter 和 Domagoj Vrgoˇc。图数据库现代查询语言基础。*ACM Comput.Surv.*，50(5)，2017 年 9 月。[4] Wai-Ho Au 和 Keith CC Chan。FARM：用于发现模糊关联规则的数据挖掘系统。载于*FUZZ- IEEE'99*。*1999* 年 *IEEE* 国际模糊系统会议论文集（目录号 *99CH36315*），第 3 卷，第 1217–1222 页。IEEE，1999 年。[5] Philip Bohannon、Wenfei Fan、Floris Geerts、Xibei Jia 和 Anastasios Kementsietsidis。用于数据清洗的条件函数依赖。载于*2007* 年 *IEEE* 第 *23* 届国际数据工程会议，第 746–755 页。IEEE，2006 年。[6] Angela Bonifati、Peter Furniss、Alastair Green、Russ Harmer、Eugenia Oshurko 和 Hannes Voigt。图数据库的模式验证与演化。载于概念建模国际会议，第 448–456 页。施普林格，2019 年。[7] Sergey Brin、Rajeev Motwani 和 Craig Silverstein。超越购物篮：将关联规则推广为相关性。*SIGMOD Rec.*，26(2)：265–276，1997 年 6 月。[8] Sergey Brin、Rajeev Motwani、Jeffrey D. Ullman 和 Shalom Tsur。购物篮数据的动态项集计数与蕴含规则。*SIGMOD Rec.*，26(2)：255–264，1997 年 6 月。[9] Björn Bringmann 和 Siegfried Nijssen。单个图中什么是频繁的？载于亚太知识发现与数据挖掘会议，第 858–863 页。施普林格，2008 年。[10] Purificacion Carinena。模糊时序关联规则：结合时序数据与定量数据以增强规则表达能力。*Wiley* 跨学科评论：数据挖掘与知识发现，4(1)：64–70，2014 年。[11] Peggy Cellier、Thierry Charnois、Marc Plantevit、Christophe Rigotti、Bruno Crémilleux、Olivier Gandrillon、Jiˇrí Kléma 和 Jean-Luc Manguin。从生物医学文本中挖掘序列模式，以发现基因相互作用及其上下文信息。《生物医学语义学杂志》，6(1)：27，2015 年。[12] Stefano Ceri、Georg Gottlob 和 Letizia Tanca。逻辑程序设计与数据库：综述。逻辑程序设计与数据库，第 1–15 页，1990 年。[13] Edgar F Codd。数据 库关系模型的进一步归一化。数据库系统，6(1972)：33–64，1972 年。[14] Luc Dehaspe 和 Hannu Toironen。关系关联规则的发现，第 189–208 页。Springer-Verlag，柏林、海德堡，2001 年。[15] Luc Dehaspe 和 Hannu Toivonen。频繁 Datalog 模式的发现。*Data Mining andknowledge discovery*，3(1)：7–36，1999 年。[16] Lisa Di-Jorio、Anne Laurent 和 Maguelonne Teisseire。从大型数据库中挖掘频繁渐进项集。载于智能数据分析国际研讨会，第 297–308 页。施普林格，2009 年。[17] Wenfei Fan、Floris Geerts、Jianzhong Li 和 Ming Xiong。条件函数依赖的发现。*IEEE*《知识与数据工程汇刊》，23(5)：683–698，2010 年。[18] Wenfei Fan、Ruochun Jin、Muyang Liu、Ping Lu、Chao Tian 和 Jingren Zhou。捕获图中的关联。*Proc.VLDBEndow.*，13(12)：1863–1876，2020 年 7 月。[19] Wenfei Fan 和 Ping Lu。图依赖。载于第 *36* 届 *ACM SIGMOD-SIGACT-SIGAI* 数据库系统原理研讨会论文集，PODS '17，第 403–416 页，美国纽约州纽约市，2017 年。计算机协会。[20] Wenfei Fan 和 Ping Lu。图依赖。*ACM Trans. Database Syst.*，44(2)，2019 年 2 月。

[21] Wenfei Fan、Xin Wang、Yinghui Wu 和 Jingbo Xu。带有图模式的关联规则。Proc. VLDB Endow.，8(12)：1502–1513，2015 年 8 月。[22] Wenfei Fan、Yinghui Wu 和 Jingbo Xu。图的函数依赖。载于 2016 年国际数据管理会议论文集，SIGMOD '16，第 1843–1857 页，美国纽约州纽约市，2016 年。计算机协会。[23] Philippe Fournier-Viger、Usef Faghihi、Roger Nkambou 和 Engelbert Mephu Nguifo。CMRules：挖掘多个序列共有的序列规则。Knowledge-Based Systems，25(1)：63–76，2012 年。[24] Takeshi Fukuda、Yasuhido Morimoto、Shinichi Morishita 和 Takeshi Tokuyama。数值属性优化关联规则的挖掘。载于第十五届 ACM SIGACT-SIGMOD-SIGART 数据库系统原理研讨会论文集，PODS '96，第 182–191 页，美国纽约州纽约市，1996 年。计算机协会。[25] Bart Goethals 和 Jan Van den Bussche。关系关联规则：获得 WARMeR。载于《模式检测与发现：ESF 探索性研讨会》，英国伦敦，2002 年 9 月 16–19 日，会议论文集，第 125–139 页。施普林格，2002 年。[26] Y Gu、Y Guan 和 P Missier。面向从知识图谱中学习实例化逻辑规则。arXiv 预印本 ArXiv:2003.06071，2020 年。[27] Yi-Chung Hu、Ruey-Shun Chen 和 Gwo-Hshiung Tzeng。使用模糊划分方法发现模糊关联规则。Knowledge-Based Systems，16(3)：137–147，2003 年。[28] Daniel Jurafsky 和 James H. Martin。《语音与语言处理：自然语言处理、计算语言学和语音识别导论，含语言模型》。第 3 版，2025 年。在线手稿发布于 2025 年 8 月 24 日。[29] Minakshi Kaushik、Rahul Sharma、Sijo Arakkal Peious、Mahtab Shahin、Sadok Ben Yahia 和 Dirk Draheim。数值关联规则挖掘方法的系统评估。SN Computer Science，2(5)：348，2021 年。[30] Krzysztof Koperski 和 Jiawei Han。在地理信息数据库中发现空间关联规则。载于国际空间数据库研讨会，第 47–66 页。施普林格，1995 年。[31] Jonathan Lajus、Luis Galárraga 和 Fabian Suchanek。使用 AMIE 3 进行快速且精确的规则挖掘。载于欧洲语义网会议，第 36–52 页。施普林格，2020 年。[32] Christian Meilicke、Melisachew Wudage Chekol、Manuel Fink 和 Heiner Stuckenschmidt。用于知识图谱补全的强化式任意时刻自底向上规则学习。arXiv 预印本 arXiv:2004.04412，2020 年。[33] Christian Meilicke、Melisachew Wudage Chekol、Daniel Ruffinelli 和 Heiner Stuckenschmidt。AnyBURL 简介。载于德奥人工智能联合会议（人工智能），第 244–248 页。施普林格，2019 年。[34] Christian Meilicke、Manuel Fink、Yanjie Wang、Daniel Ruffinelli、Rainer Gemulla 和 Heiner Stuckenschmidt。知识图谱补全的基于规则和嵌入的系统的细粒度评估。载于国际语义网会议，第 3–20 页。施普林格，2018 年。[35] Simon Ott、Christian Meilicke 和 Matthias Samwald。SAFRAN：一种可解释的基于规则的链接预测方法，性能超越嵌入模型，2021 年。[36] Gregory Piatetsky-Shapiro。强规则的发现、分析与呈现。Knowledge Discovery in Data-bases，第 229–248 页，1991 年。[37] Reza Ramezani、Mohammad Ali Nematbakhsh 和 Mohamad Saraee。在无需用户干预的情况下从语义网数据中挖掘关联规则。Journal of Computing & Security，7(1)，2020 年。[38] Reza Ramezani、Mohamad Saraee 和 Mohammad Ali Nematbakhsh。MRAR：多关系关联规则挖掘。Journal of Computing and Security，1(2)：133–158，2014 年。[39] Reza Ramezani、Mohammad Saraee 和 Mohammad Ali Nematbakhsh。在关联数据中发现关联规则：一种集中化方法。载于 2013 年第 21 届伊朗电气工程会议（ICEE），第 1–6 页。IEEE，2013 年。[40] Yuya Sasaki 和 Panagiotis Karras。在大型属性图中挖掘路径关联规则。载于第 33 届 ACM 信息与知识管理国际会议论文集，CIKM '24，第 1994–2003 页，美国纽约州纽约市，2024 年。计算机协会。[41] Yücel Saygin、Vassilios S Verykios 和 Ahmed K Elmagarmid。隐私保护关联规则挖掘。载于第十二届数据工程研究问题国际研讨会论文集：电子商务/电子商务系统工程，RIDE-2EC 2002，第 151–158 页。IEEE，2002 年。

[42] Giovanni Semeraro、Floriana Esposito、Donato Malerba、Clifford Brunk 和 Michael Pazzani。学习逻辑程序时避免非终止：以 FOIL 和 FOCL 为例。在逻辑元编程国际研讨会，第 183–198 页。施普林格，1994年。[43] Rahul Sharma、Minakshi Kaushik、Sijo Arakkal Peious、Alexandre Bazin、Syed Attique Shah、Iztok Fister、Sadok Ben Yahia 和 Dirk Draheim。关联规则挖掘、联机分析处理和统计推理的统一新框架。*IEEE Access*，10：12792–12813，2022年。[44] Ramakrishnan Srikant 和 Rakesh Agrawal。在大型关系表中挖掘定量关联规则。*SIGMOD Rec.*，25(2)：1–12，1996年6月。[45] Ramakrishnan Srikant 和 Rakesh Agrawal。挖掘序列模式：泛化与性能改进。在数据库技术扩展国际会议，第 1–17 页。施普林格，1996年。[46] Karla Taboada、Kaoru Shimada、Shingo Mabu、Kotaro Hirasawa 和 Jinglu Hu。使用遗传网络编程挖掘连续属性的关联规则。在第九届遗传与进化计算年会，GECCO '07，第 1758 页，纽约，美国，2007年。计算机协会。[47] Xin Wang 和 Yang Xu。挖掘图模式关联规则。在数据库与专家系统应用国际会议，第 223–235 页。施普林格，2018年。[48] Xin Wang、Yang Xu 和 Huayi Zhan。通过图模式扩展关联规则。*Expert Systems withApplications*，141：112897，2020年。[49] Hong Yao 和 Howard J Hamilton。从数据中挖掘函数依赖。*Data Miningand KnowledgeDiscovery*，16(2)：197–219，2008年。[50] Huiyu Zhou 和 Kotaro Hirasawa。在推荐系统中演化时序关联规则。*NeuralComputingand Applications*，31(7)：2605–2619，2019年。

A 证明

我们重新回顾四种情境中的主要定义，这些定义是证明各项度量指标特性所必需的。请注意，对于宏平均指标的情境，我们直接沿用为 GPAR 单图上下文定义的内容，因为得分经过宏平均后，我们处理的是一族概率空间。

情境 ISAR：基于项集的关联规则

定义汇编：

$$
\begin{aligned}
r &= A \Rightarrow B \\
\Omega &= T \\
E_A &= \{\, t \in T \mid A \subseteq t \,\} \\
E_B &= \{\, t \in T \mid B \subseteq t \,\} \\
E_A \cap E_B &= \{\, t \in T \mid (A \cup B) \subseteq t \,\} \\
\hat{P}_T(E_A) &= \frac{|E_A|}{|T|} \\
\hat{P}_T(E_B) &= \frac{|E_B|}{|T|} \\
\hat{P}_T(E_A \cap E_B) &= \frac{|E_A \cap E_B|}{|T|} \\
\hat{P}_T(E_B \mid E_A) &= \frac{|E_A \cap E_B|}{|E_A|}
\end{aligned}
$$

情境 GPAR（单图）：基于图模式的关联规则，单图设置

定义汇编：

$$
r = (p_1, p_2, V_1, V_2)
$$

$$
\begin{aligned}
\Omega &= \mathcal{T}^n_{g,p_1,p_2} \\
E_{g,p_1,V_1;p_2} &= \{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_1,V_1) \,\} \\
E_{g,p_2,V_2;p_1} &= \{\, T \in \mathcal{T}^n_{g,p_1,p_2} \mid m_g(T,p_2,V_2) \,\} \\
\hat{P}_g(E_{g,p_1,V_1;p_2}) &= \frac{|E_{g,p_1,V_1;p_2}|}{|\mathcal{T}^n_{g,p_1,p_2}|} \\
\hat{P}_T(E_{g,p_2,V_2;p_1}) &= \frac{|E_{g,p_2,V_2;p_1}|}{|\mathcal{T}^n_{g,p_1,p_2}|} \\
\hat{P}_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) &= \frac{|E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}|}{|\mathcal{T}^n_{g,p_1,p_2}|} \\
\hat{P}_g(E_{g,p_2,V_2;p_1} \mid E_{g,p_1,V_1;p_2}) &= \frac{|E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}|}{|E_{g,p_1,V_1;p_2}|}
\end{aligned}
$$

上下文 GPAR（微观）：基于图模式的关联规则、图事务环境、微平均定义汇编:

$$
\begin{aligned}
r &= (p_1, p_2, V_1, V_2) \\
\Omega &= \mathcal{T}^n_{G,p_1,p_2} \\
E_{G,p_1,V_1;p_2} &= \{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_1,V_1) \,\} \\
E_{G,p_2,V_2;p_1} &= \{\, (i,T) \in \mathcal{T}^n_{G,p_1,p_2} \mid m_{g_i}(T,p_2,V_2) \,\} \\
\hat{P}_G(E_{G,p_1,V_1;p_2}) &= \frac{|E_{G,p_1,V_1;p_2}|}{|\mathcal{T}^n_{G,p_1,p_2}|} \\
\hat{P}_G(E_{G,p_2,V_2;p_1}) &= \frac{|E_{G,p_2,V_2;p_1}|}{|\mathcal{T}^n_{G,p_1,p_2}|}
\end{aligned}
$$

泛化

为了将三种样本空间中的任意一种抽象到一般环境中，我们采用以下定义:

$$
\begin{aligned}
E_1 &= \{\, x \in \Omega \mid cond_1(x) \,\} \\
E_2 &= \{\, x \in \Omega \mid cond_2(x) \,\} \\
\hat{P}_\Omega(E_1) &= \frac{|E_1|}{|\Omega|} \\
\hat{P}_\Omega(E_2) &= \frac{|E_2|}{|\Omega|}
\end{aligned}
$$

其中 $cond_1$ 和 $cond_2$ 是布尔表达式，即 $A \subseteq t$ 和 $B \subseteq t$、$m_g(T,\ p_1,\ V_1)$ 与 $m_g(T,\ p_2,\ V_2)$，或者$m_{g_i}(T,\ p_1,\ V_1)$ 与 $m_{g_i}(T,\ p_2,\ V_2)$，分别如此。我们将感兴趣的情形定义如下:

$$
\begin{aligned}
\texttt{IDE} &\Leftrightarrow E_1 = E_2 \\
\texttt{DIS} &\Leftrightarrow E_1 \cap E_2 = \varnothing \\
\texttt{IND} &\Leftrightarrow \hat{P}_\Omega(E_1 \cap E_2) = \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) \\
\texttt{POS} &\Leftrightarrow \hat{P}_\Omega(E_1 \cap E_2) > \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) \\
\texttt{NEG} &\Leftrightarrow \hat{P}_\Omega(E_1 \cap E_2) < \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)
\end{aligned}
$$

现在可以将度量指标定义为其广义形式:

$$
\begin{aligned}
\text{confidence}_\Omega(r) &= \hat{P}_\Omega(E_2 \mid E_1) = \frac{\hat{P}_\Omega(E_1 \cap E_2)}{\hat{P}_\Omega(E_1)} = \frac{|E_1 \cap E_2|}{|E_1|} \\
\text{lift}_\Omega(r) &= \frac{\hat{P}_\Omega(E_1 \cap E_2)}{\hat{P}_\Omega(E_2)} = \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|}
\end{aligned}
$$

$$\begin{aligned}\text{leverage}_\Omega(r) &= \hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) \\ \text{conviction}_\Omega(r) &= \frac{1 - \hat{P}_\Omega(E_2)}{1 - \hat{P}_\Omega(E_2 \mid E_1)} = \frac{|E_1| \cdot (|\Omega| - |E_2|)}{|\Omega| \cdot (|E_1| - |E_1 \cap E_2|)}\end{aligned}$$

我们将确信度$_\Omega(r)$定义为 ∞，条件是置信度$_\Omega(r) = 1$。

A.1 关于置信度特征的证明

A.1.1 上下文：ISAR、GPAR（单一）和 GPAR（微观）

命题 1. 假设 $\hat{P}_\Omega(E_1) > 0$，则 *IDE* $\iff$ 置信度$_\Omega(r) = 1$。

证明。 ($\Rightarrow$) 假设 $E_1 = E_2$ 成立。那么置信度$_\Omega(r) = 1$ 成立，因为置信度$_\Omega(r) = \frac{|E_1 \cap E_2|}{|E_1|} \overset{\text{IDE}}{=} \frac{|E_1|}{|E_1|} = 1$。

($\Leftarrow$) 假设置信度$_\Omega(r) = 1$ 成立。那么，一般而言，$E_1 = E_2$ 并不成立，因为 $\frac{|E_1 \cap E_2|}{|E_1|} = 1 \iff |E_1 \cap E_2| = |E_1| \iff E_1 \subseteq E_2 \not\Leftrightarrow E_1 = E_2$.

□

命题 2. 假设 $\hat{P}_\Omega(E_1) > 0$，则 *DIS* $\iff$ 置信度$_\Omega(r) = 0$。

证明. ($\Rightarrow$) 假设 $E_1 \cap E_2 = \varnothing$ 成立。那么，置信度$_\Omega(r) = \frac{|E_1 \cap E_2|}{|E_1|} \overset{\text{DIS}}{=} \frac{|\varnothing|}{|E_1|} = 0$。

($\Leftarrow$) 假设置信度$_\Omega(r) = 0$ 成立，则 $E_1 \cap E_2 = \varnothing$ 成立，因为如果 $\frac{|E_1 \cap E_2|}{|E_1|} = 0$，那么只有在 $E_1 \cap E_2 = \varnothing$ 时，该分数才会变为 0。

□

命题 3. 假设 $\hat{P}_\Omega(E_1) > 0$，则 独立性 $\iff$ 置信度$_\Omega(r) = \hat{P}_\Omega(E_2)$。

证明. 通过一系列双向逻辑等价关系进行证明:

$$\begin{aligned}&\text{confidence}_\Omega(r) = \hat{P}_\Omega(E_2) \\ \iff &\frac{|E_1 \cap E_2|}{|E_1|} = \frac{|E_2|}{|\Omega|} \\ \iff &|E_1 \cap E_2| = \frac{|E_1| \cdot |E_2|}{|\Omega|} \quad \text{(divide by } |\Omega| \text{ which is non-zero because } \hat{P}_\Omega(E_1) > 0) \\ \iff &\frac{|E_1 \cap E_2|}{|\Omega|} = \frac{|E_1|}{|\Omega|} \cdot \frac{|E_2|}{|\Omega|} \\ \iff &\hat{P}_\Omega(E_1 \cap E_2) = \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) \\ \iff &\hat{P}_\Omega(E_1) \text{ and } \hat{P}_\Omega(E_2) \text{ are independent.}\end{aligned}$$

□

命题 4. 假设 $\hat{P}_\Omega(E_1) > 0$，则 正 $\iff$ 置信度$_\Omega(r) > \hat{P}_\Omega(E_2)$。

证明。 省略——可用与命题 3 相同的方法证明。唯一需要更改的是将$=$ 替换为 $>$。

□

命题 5. 假设$\hat{P}_\Omega(E_1) > 0$，则负 $\iff$ 置信度$_\Omega(r) < \hat{P}_\Omega(E_2)$。

证明。 略去——可用与命题 3 相同的方法证明。唯一需要更改的是将$=$ 替换为 $<$。

□

A.1.2 上下文：GPAR（宏）

$$\text{macro-confidence}_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} \text{confidence}_g(p_1, p_2, V_1, V_2)$$

命题 6. 假设对于每个 $g \in G$，$\hat{} P_\Omega(E_{g,p_1,V_1;p_2}) > 0$，则 *IDE* $\Longrightarrow$ 宏观置信度$_G$(p1, p2, V1, V2) = 1.

证明。 ($\Rightarrow$) 假设对于每个 $g \in G$，E_1 与 E_2 相同，则宏观置信度$_G(p_1, p_2, V_1, V_2) = 1$ 成立，因为

宏观置信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G}$ 置信度$_g(p_1, p_2, V_1, V_2)$ IDE $\overset{Prop.1}{=} \frac{1}{|G|} \sum_{g \in G} 1 = 1$.

($\Leftarrow$) 假设宏观置信度$_\Omega(p_1, p_2, V_1, V_2) = 1$ 成立。那么，$E_1 = E_2$ 对每个 $g \in G$ 一般而言并不成立，因为如果 n 中的值之和等于 n，那么每个加数都等于 1，这在 $E_1 \subseteq E_2$ 时成立。 □

命题7. 假设 $\hat{} P_\Omega(E_{g,\ p_1,\ V_1;\ p_2}) > 0$ 对每个 $g \in G$ 都成立，则 *DIS* $\Longleftrightarrow$ 宏观置信度$_G(p_1, p_2, V_1, V_2) = 0$。

证明。 ($\Rightarrow$) 假设对于每个 $g \in G$，E_1 与 E_2 不相交，那么
宏观置信度$_G(p_1, p_2, V_1, V_2) = 0$ 成立，因为
宏观置信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G}$ 置信度$_g(p_1, p_2, V_1, V_2)$ DIS $\overset{Prop.2}{=} \frac{1}{|G|} \sum_{g \in G} 0 = 0$。
($\Leftarrow$) 假设宏观置信度$_G(p_1, p_2, V_1, V_2) = 0$ 成立，那么对于每个满足 E_1 与 E_2 不相交的 $g \in G$，若 $[0, 1]$ 中的 n 值之和等于 0，则每一项都等于 0，因此，对于每个 $g \in G$，E_1 与 E_2 不相交。
□

命题 8. 假设对每个 $g \in G$，$\hat{} P_\Omega(E_{g,p_1,V_1;p_2}) > 0$，
则 独立性 $\Longrightarrow$ 宏观置信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} \hat{P}_g(E_{g,p_2,V_2;p_1})$.

证明. ($\Rightarrow$) 假设对所有 $g \in G$，均有 $\hat{} P_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) = \hat{} P_g(E_{g,p_1,V_1;p_2}) \cdot \hat{} P_g(E_{g,p_2,V_2;p_1})$，则宏观置信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} \hat{} P_g(E_{g,p_2,V_2;p_1})$，因为宏观置信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G}$ 置信度$_g(p_1, p_2, V_1, V_2)$ 独立性 $\overset{Prop.3}{=} \frac{1}{|G|} \sum_{g \in G} \hat{} P_g(E_{g,p_2,V_2;p_1})$。
—
($\Leftarrow$) 逆命题并不成立，因为不同的加数集合可能得到相同的和 . □

命题 9. 假设对于每个 $g \in G$，$\hat{} P_\Omega(E_{g,p_1,V_1;p_2}) > 0$，则 *POS* $\Longrightarrow$ 宏观置信度$_G(p_1, p_2, V_1, V_2) > \frac{1}{|G|} \sum_{g \in G} \hat{} P_g(E_{g,p_2,V_2;p_1})$。 —

证明。 略——可使用命题 4，以与命题 8 相同的方式证明。唯一需要做的改动是将 $=$ 替换为 $>$。
□

命题 10. 假设对于每个 $g \in G$，$\hat{} P_\Omega(E_{g,p_1,V_1;p_2}) > 0$，则 负 $\Longrightarrow$ 宏观置信度$_G(p_1, p_2, V_1, V_2) < \frac{1}{|G|} \sum_{g \in G} \hat{} P_g(E_{g,p_2,V_2;p_1})$。 —

证明。 省略——可使用命题 5，以与命题 8 相同的方式证明。唯一需要修改的是将 $=$ 替换为 $<$。
□

A.2 关于提升度特征的证明

A.2.1 上下文：ISAR、GPAR（单一）和 GPAR（微观）

Pro位置 11。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) > 0$，则 *IDE* $\Longrightarrow$ 提升度$_\Omega(r) = 1/\hat{} P_\Omega(E_{\quad 2})$.

证明。 (⇒) 假设 $E_1 = E_2$ 成立，则 $\text{lift}_\Omega(r) = 1/\hat{}P_\Omega(E_2)$ 成立，因为 $\text{lift}_\Omega(r) = \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|} \overset{\text{IDE}}{=} \frac{|E_1| \cdot |\Omega|}{|E_1| \cdot |E_2|} = 1/\hat{}P_\Omega(E_2)$。

(⇐) 假设 $\text{lift}_\Omega(r) = 1/\hat{}P_\Omega(E_2)$ 成立，则 $E_1 = E_2$ 通常不成立，因为

$$\text{lift}_\Omega(r) = 1/\hat{}P_\Omega(E_2) \iff \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|} = \frac{|\Omega|}{|E_2|} \text{ (除以 } |\Omega|, \text{ 其非零是因为 } \hat{}P_\Omega(E_1) > 0) \iff \frac{1}{|E_1|} \cdot \frac{|E_1 \cap E_2|}{|E_2|} = \frac{1}{|E_2|} \iff |E_1 \cap E_2| = |E_1| \iff E_1 \subseteq E_2 \not\Longleftrightarrow E_1 = E_2.$$

□

命题 12。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) > 0$，则 *DIS* $\iff$ 提升度$_\Omega(r) = 0$。

证明。 (⇒) 假设 $E_1 \cap E_2 = \varnothing$ 成立，则提升度$_\Omega(r) = 0$ 成立，因为 提升度$_\Omega(r) = \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|} \overset{\text{DIS}}{=} \frac{|\varnothing| \cdot |\Omega|}{|E_1| \cdot |E_2|} = 0$。

(⇐) 假设提升度$_\Omega(r) = 0$ 成立，则 $E_1 \cap E_2 = \varnothing$ 成立，因为

$$\text{lift}_\Omega(r) = 0 \iff \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|} = 0.$$

该分数只有在 $E_1 \cap E_2 = \varnothing$ 时才可能为 0，因为 $\Omega \neq \varnothing$，因为 $\hat{} P_\Omega(E_1) > 0$。 □

命题 13。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) > 0$，则 *IND* $\iff$ 提升度$_\Omega(r) = 1$。

证明。 通过一系列双向逻辑等价关系进行证明：

$$\text{提升度}_\Omega(r) = 1 \iff \frac{|E_1 \cap E_2| \cdot |\Omega|}{|E_1| \cdot |E_2|} = 1 \text{ (除以 } |\Omega|^2, \text{ 其非零是因为 } \hat{}P_\Omega(E_1) > 0) \iff |E_1 \cap E_2| = |E_1| \cdot |E_2| \cdot |\Omega| \iff \hat{}P_\Omega(E_1 \cap E_2) = \hat{}P_\Omega(E_1) \cdot \hat{}P_\Omega(E_2)。$$

□

命题14。 假设 $\hat{} P_\Omega(E_1) > 0$ 且 $\hat{} P_\Omega(E_2) > 0$，则 *IND* $\iff$ 提升度$_\Omega(r) > 1$。

证明。 略——可用与命题 13 相同的方法证明。唯一需要改动的是将 $=$ 替换为 $>$。 □

命题15。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) > 0$，则独立性 $\iff$ 提升度$_\Omega(r) < 1$。

证明。 省略——可用与命题 13 相同的方法证明。唯一需要更改的是将 $=$ 替换为 $<$。 □

A.2.2 上下文：GPAR（宏）

$$\text{macro-lift}_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} \text{lift}_g(p_1, p_2, V_1, V_2)$$

命题*16.* 假设对于每个 $g \in G$，$\hat{} P_g(E_{g,p_1,V_1;p_2}) > 0$ 且 $\hat{} P_g(E_{g,p_2,V_2;p_1}) > 0$，则 *IDE* $\Longrightarrow$宏提升度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} 1/\hat{} P_g(E_{g,p_2,V_2;p_1})$。

证明。 ($\Rightarrow$) 假设 $E_{g,\ p_1,\ V_1;\ p_2} = E_{g,\ p_2,\ V_2;\ p_1}$ 对每个 $g \in G$ 均成立，则 $\text{macro-lift}_G(p_1,\ p_2,\ V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} 1/\hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，因为

$\text{macro-lift}_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G} \text{lift}_g(p_1, p_2, V_1, V_2) \overset{\text{IDE}, Prop.11}{=} \frac{1}{|G|} \sum_{g \in G} \hat{P}_g(E_{g,p_2,V_2;p_1})$.
($\Leftarrow$) 假设 $\text{macro-lift}_G(p_1,\ p_2,\ V_1,\ V_2) = \sum_{g \in G} 1/\hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，则 $E_{g,\ p_1,\ V_1;\ p_2} = E_{g,\ p_2,\ V_2;\ p_1}$ 不成立，因为相同的总和可以由不同的加数集合得到。 □

命题17。 假设 $\hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2}) > 0$ 和 $\hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1}) > 0$ 对每个 $g \in G$ 均成立，则 *DIS* $\Longleftrightarrow$宏提升度$_G(p_1, p_2, V_1, V_2) = 0$。

证明。 ($\Rightarrow$) 假设对每个 $g \in G$，$E_{g,\ p_1,\ V_1;\ p_2} \cap E_{g,\ p_2,\ V_2;\ p_1} = \varnothing$ 均成立，则宏提升度$_G(p_1,\ p_2,\ V_1,\ V_2) = 0$ 成立，因为宏提升度$_G(p_1,\ p_2,\ V_1,\ V_2) = \frac{1}{|G|} \sum_{g \in G} \text{lift}_g(p_1, p_2, V_1, V_2) \overset{\text{DIS}, Prop.12}{=} \frac{1}{|G|} \sum_{g \in G} 0 = 0.$
($\Leftarrow$) 假设宏提升度$_G(p_1,\ p_2,\ V_1,\ V_2) = 0$ 成立，则$E_{g,\ p_1,\ V_1;\ p_2} \cap E_{g,\ p_2,\ V_2;\ p_1} = \varnothing$ 成立，因为如果n 在 $[0, 1]$ 中的值之和等于 0，那么每个加数都等于 0，因此，对每个$g \in G$，$E_{g,\ p_1,\ V_1;\ p_2}$ 与$E_{g,\ p_2,\ V_2;\ p_1}$ 不相交。 □

命题18。 假设对每个 $g \in G$，$\hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2}) > 0$ 和$\hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1}) > 0$ 均成立，则 *IND* $\Longrightarrow$宏提升度$_G(p_1, p_2, V_1, V_2) = 1$。

证明。 ($\Rightarrow$) 假设对于每个 $g \in G$，$\hat{} P_g(E_{g,p_1,V_1;p_2 \cap E_{g,p_2,V_2;p_1}}) = \hat{} P_g(E_{g,p_1,V_1;p_2}) \cdot \hat{} P_g(E_{g,p_2,V_2;p_1})$ 成立，则宏提升度$_G(p_1, p_2, V_1, V_2) = 1$ 成立，因为提升度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|} \sum_{g \in G}$ 提升度$_g(p_1, p_2, V_1, V_2)$ 独立性 $\overset{Prop.13}{=} \frac{1}{|G|} \sum_{g \in G} 1 = \overline{1^\circ}$
($\Leftarrow$) 假设宏提升度$_G(p_1, p_2, V_1, V_2) = 1$ 成立，则对于每个 $g \in G$，$\hat{} P_g(E_{g,p_1,V_1;p_2 \cap E_{g,p_2,V_2;p_1}}) = \hat{} P_g(E_{g,p_1,V_1;p_2}) \cdot \hat{} P_g(E_{g,p_2,V_2;p_1})$ 不成立，因为提升度得分是区间 $[0, \infty)$中的值，并且相同的总和可以由不同的加数集合得到。 □

命题19。 假设对于每个 $g \in G$，$\hat{} P_g(E_{g,p_1,V_1;p_2}) > 0$ 且 $\hat{} P_g(E_{g,p_2,V_2;p_1}) > 0$ 成立，则 *POS* $\Longrightarrow$宏提升度$_G(p_1, p_2, V_1, V_2) > 1$。

证明。 略——可依据命题18，以命题14为基础用相同方法证明。唯一需要更改的是将 $=$ 替换为 $>$。 □

命题20。 假设对于每个 $g \in G$，$\hat{} P_g(E_{g,p_1,V_1;p_2}) > 0$ 且 $\hat{} P_g(E_{g,p_2,V_2;p_1}) > 0$ 成立，则 *IND* $\Longrightarrow$宏提升度$_G(p_1, p_2, V_1, V_2) < 1$。

证明。 略——可依据命题18，以命题15为基础用相同方法证明。唯一需要更改的是将 $=$ 替换为 $<$。 □

A.3 关于杠杆率特征的证明

A.3.1 上下文：ISAR、GPAR（单一）和 GPAR（微观）

Proposition 21. *IDE* $\Longrightarrow \text{leverage}_\Omega(r) = \hat{P}_\Omega(E_1)(1 - \hat{P}_\Omega(E_2))$.

证明。 (⇒) 假设 $E_1 = E_2$ 成立，则 leverage$_\Omega(r) = \hat{P}_\Omega(E_1)(1 - \hat{P}_\Omega(E_2))$ 成立，因为 leverage$_\Omega(r) = \hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ IDE $= \hat{P}_\Omega(E_1) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$。
(⇐) 假设 leverage$_\Omega(r) = \hat{P}_\Omega(E_1)(1 - \hat{P}_\Omega(E_2))$ 成立，则 $E_1 = E_2$ 不成立，因为

$$
\begin{aligned}
&\text{leverage}_\Omega(r) = \hat{P}_\Omega(E_1)(1 - \hat{P}_\Omega(E_2)) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) = \hat{P}_\Omega(E_1)(1 - \hat{P}_\Omega(E_2)) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) = \hat{P}_\Omega(E_1) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) = \hat{P}_\Omega(E_1) \\
\iff &E_1 \subseteq E_2 \not\Longleftrightarrow E_1 = E_2.
\end{aligned}
$$

□

命题 22。 *DIS* $\iff$ 杠杆率$_\Omega(r) = -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$。

证明。 (⇒) 假设 $E_1 \cap E_2 = \varnothing$ 成立，则 杠杆率$_\Omega(r) = -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ 成立，因为 杠杆率$_\Omega(r) = \hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ DIS $= -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$。(⇐) 假设 杠杆率$_\Omega(r) = -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ 成立，则 $E_1 \cap E_2 = \varnothing$ 成立，因为

$$
\begin{aligned}
&\text{leverage}_\Omega(r) = -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_1) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) = -\hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_1) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) = 0 \\
\iff &E_1 \cap E_2 = \varnothing.
\end{aligned}
$$

□

命题 23。独立性 $\iff$ 杠杆率$_\Omega(r) = 0$。

证明。 (⇒) 假设 $\hat{P}_\Omega(E_1 \cap E_2) = \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ 成立，则杠杆率$_\Omega(r) = 0$ 成立，因为杠杆率$_\Omega(r) = \hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ 独立性 $= \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) = 0$.
(⇐) 假设杠杆率$_\Omega(r) = 0$ 成立，则 $\hat{P}_\Omega(E_1 \cap E_2) = \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2)$ 成立，因为

杠杆率$_\Omega(r) = 0 \iff \hat{P}_\Omega(E_1 \cap E_2) - \hat{P}_\Omega(E_1) \cdot \hat{P}_\Omega(E_2) = 0 \iff P_\Omega(E_1 \cap E_2) = P_\Omega(E_1) \cdot P_\Omega(E_2)$。

□

命题 24. 正 $\iff$ 杠杆率$_\Omega(r) > 0$.

证明。 略——可用与命题 23 相同的方法证明。唯一需要改变的是将 $=$ 替换为 $>$。

□

命题 25. 负 $\iff$ 杠杆率$_\Omega(r) < 0$.

证明。 略去——可用与命题 23 相同的方法证明。唯一需要修改的是将 $=$ 替换为 $<$。

□

A.3.2 上下文：GPAR（宏）

$$\text{macro-leverage}_G(p_1,p_2,V_1,V_2)=\frac{1}{|G|}\sum_{g\in G}\text{leverage}_g(p_1,p_2,V_1,V_2)$$

命题 26。 $IDE \Longrightarrow$ 宏观杠杆$_G(p_1,p_2,V_1,V_2) = \frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{g,p_1,V_1;p_2})(1-\hat{P}_g(E_{g,p_1,V_1;p_2}))$.

证明。(⇒) 假设对于每个 $g\in G$， $E_1=E_2$ 都成立，则
宏观杠杆$_G(p_1,p_2,V_1,V_2) = \frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{g,p_1,V_1;p_2})(1-\hat{P}_g(E_{g,p_1,V_1;p_2}))$ 成立，因为
宏观杠杆$_G(p_1,p_2,V_1,V_2) = \frac{1}{|G|}\sum_{g\in G}$ 杠杆率$_g(p_1,p_2,V_1,V_2) \overset{\text{IDE}\ Prop.21}{=} \frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{g,p_1,V_1;p_2})(1-\hat{P}_g(E_{g,p_2,V_2;p_1}))$。

(⇐) 假设宏观杠杆率$_G(r) = \frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2})(1-\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}))$成立，那么 $E_1=E_2$ 对每个 $g\in G$ 都不成立，因为相同的和可以由不同的加数集合得到。 □

命题 27。 $DIS \Longrightarrow$ 宏观杠杆$._G(p_1,p_2,V_1,V_2) = -\frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{p_1,\ V_1;\ p_2})\cdot\hat{P}_\Omega(E_{p_2,\ V_2;\ p_1})$。

证明。(⇒) 假设对于每个满足 $g\in G$的情形， E_1 和 E_2 都不相交，那么宏观杠杆率$_G(p_1,p_2,V_1,V_2) = -\frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{p_1,V_1;p_2})\cdot\hat{P}_\Omega(E_{p_2,V_2;p_1})$成立，因为宏观杠杆率$\frac{1}{|G|}\sum_{g\in G}$ $_g$ p_1 p_2 V_1 V_2 (, , V_1,V_2) 杠杆率 (, , ,) DIS $\overset{Prop.22}{=} \frac{1}{|G|}\sum_{g\in G}-\hat{P}_g(E_{g,p_1,V_1;p_2})\cdot\hat{P}_g(E_{g,p_2,V_2;p_1})$。

(⇐) 假设宏观杠杆率$_G(p_1,\ p_2,\ V_1,\ V_2) = -\frac{1}{|G|}\sum_{g\in G}\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2})\cdot\hat{P}_\Omega(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，那么对于每个 $g\in G$， E_1 和E_2 都不相交；而 $g\in G$ 不成立，因为相同的和可以由不同的加数集合得到。 □

命题28。独立性 $\Longrightarrow$ 宏观杠杆$_G=0$。

证明。(⇒) 假设对每个 $g\in G$，$\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}\cap E_{g,\ p_2,\ V_2;\ p_1}) = \hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2})\cdot\hat{P}_g(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，那么宏观杠杆率$_G=0$成立，因为
宏观杠杆率$_G(p_1,\ p_2,\ V_1,\ V_2) = \frac{1}{|G|}\sum_{g\in G}$ 杠杆率$_g(p_1,\ p_2,\ V_1,\ V_2) \overset{\text{独立性}\ Prop.23}{=} \frac{1}{|G|}\sum_{g\in G}0=0$。

(⇐) 假设宏观杠杆$_G(p_1,\ p_2,\ V_1,\ V_2) = 0$ 成立，则 $\hat{P}_g(E_{g,\ p_1,\ V_1;\ p_2}\cap E_{g,\ p_2,\ V_2;\ p_1}) = \hat{P}_g(E_{g,p_1,V_1;p_2})\cdot\hat{P}_g(E_{g,\ p_2,\ V_2;\ p_1})$ 对每个 $g\in G$ 均不成立，因为杠杆分数是区间$[-\frac{1}{4},\frac{1}{4}$中的值，并且相同的总和可以由不同的加数集合得到。 □

命题29。正 $\Longrightarrow$ 宏观杠杆$_G>0$。

证明。 省略——可按照命题18中的相同方式，基于命题24进行证明。唯一需要修改的是将 $=$ 替换为 $<$。 □

命题30。负 $\Longrightarrow$ 宏观杠杆$_G<0$。

证明。 省略——可根据命题 25，以与命题18相同的方式证明。唯一需要更改的是将 $=$ 替换为 $<$。 □

A.4 关于确信度特征的证明

A.4.1 语境：ISAR、GPAR（单一）和 GPAR（微观）

Pro位置 31。假设 $\hat{P}_\Omega(E_1)>0$ 和 $\hat{P}_\Omega(E_2)<1$，则 $IDE \Longrightarrow$ 确信度$_\Omega(r)=\infty$.

证明。 (⇒) 假设 $E_1 = E_2$ 成立，则确信度$_\Omega(r) = \infty$ 成立，因为如果 $E_1 = E_2$，那么置信度$_\Omega(r) = 1$，进而根据定义，确信度$_\Omega(r) = \infty$。
(⇐) 假设确信度$_\Omega(r) = \infty$成立，则 $E_1 = E_2$ 不成立，因为确信度$_\Omega(r)$ 只有在置信度$_\Omega(r) = 1$时才可能为 ∞，而这要求 $E_1 \subseteq E_2$。 □

命题 32。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) < 1$，则 *DIS* $\iff$ 确信度$_\Omega(r) = 1 - \hat{} P_\Omega(E_2)$。

证明。 (⇒) 假设 $E_1 \cap E_2 = \varnothing$ 成立，则确信度$_\Omega(r) = 1 - \hat{} P_\Omega(E_2)$ 成立，因为 $\hat{} \hat{}$ 确信度$_\Omega(r) = \frac{1 - P_\Omega(E_2)}{1 - \hat{} P_\Omega(E_2 \mid E_1)} \overset{\text{DIS}}{=} \frac{1 - P_\Omega(E_2)}{1 - 0} = 1 - \hat{} P_\Omega(E_2)$。

(⇐) 假设确信度$_\Omega(r) = \hat{} P_\Omega(E_1) \cdot (1 - \hat{} P_\Omega(E_2))$ 成立，则 $E_1 \cap E_2 = \varnothing$ 成立，因为

$$
\begin{aligned}
&\text{conviction}_\Omega(r) = 1 - \hat{P}_\Omega(E_2) \\
\iff &\frac{1 - \hat{P}_\Omega(E_2)}{1 - \hat{P}_\Omega(E_2 \mid E_1)} = 1 - \hat{P}_\Omega(E_2) \quad \text{(divide by } 1 - \hat{P}_\Omega(E_2) \text{ which is non-zero)} \\
\iff &\frac{1}{1 - \hat{P}_\Omega(E_2 \mid E_1)} = 1 \\
\iff &\frac{1}{\frac{\hat{P}_\Omega(E_1)}{\hat{P}_\Omega(E_1)} - \frac{\hat{P}_\Omega(E_1 \cap E_2)}{\hat{P}_\Omega(E_1)}} = 1 \\
\iff &\frac{\hat{P}_\Omega(E_1)}{\hat{P}_\Omega(E_1) - \hat{P}_\Omega(E_1 \cap E_2)} = 1 \\
\iff &\hat{P}_\Omega(E_1) = \hat{P}_\Omega(E_1) - \hat{P}_\Omega(E_1 \cap E_2) \\
\iff &\hat{P}_\Omega(E_1 \cap E_2) = 0 \Longrightarrow E_1 \cap E_2 = \varnothing.
\end{aligned}
$$

□

命题 33。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) < 1$成立，则
独立性 $\iff$ 确信度$_\Omega(r) = 1$。

证明。 (⇒) 假设 $\hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2} \cap E_{g,\ p_2,\ V_2;\ p_1}) = \hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2}) \cdot \hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，则确信度$_\Omega(r) = 1$成立，因为

$$\text{conviction}_\Omega(r) = \frac{1 - \hat{P}_\Omega(E_2)}{1 - \hat{P}_\Omega(E_2 \mid E_1)} \overset{\text{IND}}{=} \frac{1 - \hat{P}_\Omega(E_2)}{1 - \hat{P}_\Omega(E_2)} = 1.$$

(⇐) 假设确信度$_\Omega(r) = 1$ 成立，则 $\hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2} \cap E_{g,\ p_2,\ V_2;\ p_1}) = \hat{} P_g(E_{g,\ p_1,\ V_1;\ p_2}) \cdot \hat{} P_g(E_{g,\ p_2,\ V_2;\ p_1})$ 成立，因为

$$
\begin{aligned}
&\text{conviction}_\Omega(r) = 1 \\
\iff &\frac{1 - \hat{P}_\Omega(E_2)}{1 - \hat{P}_\Omega(E_2 \mid E_1)} = 1 \\
\iff &1 - \hat{P}_\Omega(E_2) = 1 - \hat{P}_\Omega(E_2 \mid E_1) \\
\iff &\hat{P}_\Omega(E_2 \mid E_1) = \hat{P}_\Omega(E_2).
\end{aligned}
$$

□

命题34。 假设 $\hat{} P_\Omega(E_1) > 0$ 和 $\hat{} P_\Omega(E_2) < 1$，则
正 $\iff$ 确信度$_\Omega(r) > 1$。

证明。 省略——可用与命题 33 相同的方法证明。唯一需要更改的是将 $=$ 替换为 $<$。

□

命题 35。 假设$\hat{}\ P_\Omega(E_1) > 0$ 且$\hat{}\ P_\Omega(E_2) < 1$，则负 $\Longleftrightarrow$ 确信度$_\Omega(r) < 1$。

证明。 省略——可用与命题 33 相同的方法证明。唯一需要更改的是将 $=$ 替换为 $<$。

□

A.4.2 上下文：GPAR（宏）

$$\text{macro-conviction}_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|}\sum_{g\in G}\text{conviction}_g(p_1, p_2, V_1, V_2)$$

命题 36。 假设对于每个 $g \in G$，$\hat{}\ P_\Omega(E_1) > 0$ 和$\hat{}\ P_\Omega(E_2) < 1$ 成立，则 *IDE* $\Longrightarrow$宏观确信度$_G(p_1, p_2, V_1, V_2) = \infty$。

证明。（$\Rightarrow$）假设对于每个 $g \in G$，$\hat{}\ P_\Omega(E_1) > 0$ 且$\hat{}\ P_\Omega(E_2) < 1$ 成立，则宏观确信度$_G(p_1, p_2, V_1, V_2) = \infty$ 成立，因为
宏观确信度$._G(p_1,\ p_2,\ V_1,\ V_2) \underset{|G|}{=} \sum_{g\in G}$确信度$_g(p_1,\ p_2,\ V_1,\ V_2)$ $\overset{\text{IDE}}{=} \overset{Prop.31}{=} \frac{1}{|G|}\sum_{g\in G}\infty = \infty$。

（$\Leftarrow$）假设确信度$_G(p_1,\ p_2,\ V_1,\ V_2) = \infty$ 成立，那么对于每个 $g \in G$，$\hat{}\ P_\Omega(E_1) > 0$ 和$\hat{}\ P_\Omega(E_2) < 1$ 都不成立，因为只要至少一个加数为 ∞，总和就足以计算为 ∞。 □

命题 37. 假设对于每个 $g \in G$，$\hat{}\ P_\Omega(E_1) > 0$ 和 $\hat{}\ P_\Omega(E_2) < 1$ 成立，则 *DIS* $\Longrightarrow$ 宏观确信度$_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|}\sum\overline{_{g\in G}}(1 - \hat{}P_g(E_{g,\ p_2,\ V_2;\ p_1}))$。

证明。（$\Rightarrow$）假设对于每个 $g \in G$，$E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1} = \varnothing$ 成立，那么宏观确信度$._G(p_1, p_2, V_1, V_2) = \frac{1}{|G|}\sum_{g\in G}(1 - \hat{}P_g(E_{g,p_2,V_2;p_1}))$ 成立，因为
宏观确信度$._G(p_1, p_2, V_1, V_2) = \frac{1}{|G|}\sum_{g\in G}$确信度$\cdot_g(p_1, p_2, V_1, V_2)$ DIS $\overset{Prop.32}{=}$
$\frac{1}{|G|}\sum_{g\in G}(1 - \hat{}P_g(E_{g,p_1,V_1;p_2}))$。

($\Leftarrow$) 假设宏观确信度$_G(p_1,\ p_2,\ V_1,\ V_2) = \frac{1}{|G|}\sum_{g\in G}(1 - \hat{}P_g(E_{g,\ p_2,\ V_2;\ p_1}))$ 成立，则 $E_{g,\ p_1,\ V_1;\ p_2} \cap E_{g,\ p_2,\ V_2;\ p_1} = \varnothing$ 对每个 $g \in G$ 都不成立，因为相同的和可以由不同的加数集合得到。 n

□

命题 38。 假设$\hat{}\ P_\Omega(E_1) > 0$ 和$\hat{}\ P_\Omega(E_2) < 1$ 对每个 $g \in G$ 均成立，则独立性 $\Longrightarrow$宏观确信度$_G(p_1, p_2, V_1, V_2) = 1$。

证明。($\Rightarrow$) 假设$\hat{}\ P_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) = \hat{}P_g(E_{g,p_1,V_1;p_2}) \cdot \hat{}P_g(E_{g,p_2,V_2;p_1})$ 成立，则宏观确信度$_G(p_1, p_2, V_1, V_2) = 1$ 成立，因为

$$\text{macro-conv.}_G(p_1, p_2, V_1, V_2) = \frac{1}{|G|}\sum_{g\in G}\text{conv.}_g(p_1, p_2, V_1, V_2) \overset{\text{IND},Prop.33}{=} \frac{1}{|G|}\sum_{g\in G}1 = 1.$$

($\Leftarrow$) 假设宏观确信度$_G(p_1, p_2, V_1, V_2) = 1$ 成立，则$\hat{}\ P_g(E_{g,p_1,V_1;p_2} \cap E_{g,p_2,V_2;p_1}) = \hat{}P_g(E_{g,p_1,V_1;p_2}) \cdot \hat{}P_g(E_{g,p_2,V_2;p_1})$ 成立，因为确信度分数是区间 $\mathbb{Q}_> \cap \{\infty\}$ 中的值，并且同一个和可以由不同的加数集合得到。

□

命题 39. 假设对于每个 $g \in G$，$\hat{}\ P_\Omega(E_1) > 0$ 和$\hat{}\ P_\Omega(E_2) < 1$，则 *POS* $\Longrightarrow$宏观确信度$_G(p_1, p_2, V_1, V_2) > 1$。

证明。 略——可以依据命题 34，以与命题 38 相同的方式证明。唯一需要更改的是将 $=$ 替换为 $<$。

□

命题 40. 假设对于每个 $g \in G$，都有$\hat{}\ P_\Omega(E_1) > 0$ 和$\hat{}\ P_\Omega(E_2) < 1$，则 *NEG* $\Longrightarrow$宏观确信度$_G(p_1, p_2, V_1, V_2) < 1$。

证明. 省略——可基于命题 35，以与命题 38 相同的方式证明。唯一需要更改的是将 $=$ 替换为 $<$。

□