FA STAG E D S：快速近似图实体依赖发现

周广通*、Selasi Kwashie†、张一迪*、MichaelBewong‡、VincentM. Nofong§、Debo Cheng¶、何克庆‖ 和冯在文*

*华中农业大学信息学院，中国湖北武汉 电子邮箱：{*2019308210612, zhangyidi*}*@webmail.hzau.edu.cn*；*Zaiwen.Feng@mail.hzau.edu.cn*

†查尔斯特大学 *AI*与网络未来研究所，澳大利亚巴瑟斯特 电子邮箱：*skwashie@csu.edu.au*

‡查尔斯特大学计算、数学与工程学院，澳大利亚沃加沃加 电子邮箱：*mbewong@csu.edu.au*§矿业与技术大学计算机科学与工程系，加纳塔夸 电子邮箱：*vnofong@umat.edu.gh*

¶南澳大利亚大学 *STEM*，澳大利亚阿德莱德 电子邮箱：*chedy055@mymail.unisa.edu.au*‖武汉大学计算机科学学院，中国湖北武汉 电子邮箱：*hekeqing@whu.edu.cn*

arXiv:2304.02323v2 [cs.DB] 8 Apr 2023

摘要—**本文研究属性图中**近似**规则的发现。我们提出一种具有语义意义的误差度量，用于挖掘**近乎**满足的图实体依赖（GED），以容忍真实世界图中存在的错误和不一致。我们给出了 GED 满足性的一种新刻画，并设计了一种深度优先搜索策略，以高效遍历候选规则的搜索空间。此外，我们通过实验展示了所提出的** FastAGEDs**解决方案在三个真实世界图上的可行性和可扩展性。**

索引词—**图实体依赖、近似依赖、高效算法、图约束。**

I. 引言

近年来，研究人员提出了属性图的完整性约束（例如键 [1] 和函数依赖（FD） [2]），用于描述各种数据语义，并解决图数据质量和管理问题。图实体依赖（GED） [3], [4] 是一类新的基本约束，可统一属性图中的键和函数依赖。属性图 G 上的 GED φ 是一对$\varphi = (Q[\bar{u}], X \rightarrow Y)$，用于指定图模式 $Q[\bar{u}]$ 在 G的同态匹配上的依赖 $X \rightarrow Y$。直观而言，由于图没有模式，图模式 $Q[\bar{u}]$ 会识别出 G 中依赖$X \rightarrow Y$ 应当成立的实体集合。

GED 在数据质量与管理方面具有多种实际应用（参见 [3], [4] 了解更多细节）。例如，它们已被扩展用于：社交媒体网络中的事实核查 [5], 图与关系中的实体消解 [6], [7],一致性检查 [8], 等。

与其他数据完整性约束一样，自动发现 GED 对其推广应用和实际使用至关重要。事实上，关于发现问题的研究文献日益增多 [9]–[11]，涉及 GKeys [9], GFDs [10], 和 GEDs [11]。然而，现有研究关注的是发现完全成立的依赖关系。遗憾的是，现实世界数据中存在错误、例外和歧义，妨碍了语义上有意义且实用的图数据规则的发现。

因此，在本文中，我们研究自动挖掘几乎 成立的 GED——这是一个非常适合真实世界属性图的规则发现问题。主要贡献总结如下。首先，我们基于其语义为 GED 引入一种新的误差度量，从而提出并形式化近似 GED 发现问题。其次，我们提出一种新颖且高效的算法 FASTAGEDS，用于在大型图中发现近似图实体依赖。我们通过将不一致 与必要集 的概念扩展到图依赖关系，给出 GED 满足性的一种新刻画；并开发一种高效的深度优先搜索策略，以有效遍历候选 GED 空间，从而能够快速发现完全 GED 与近似图实体依赖。第三，我们在真实世界图上进行了大量实验，以证明我们的发现方案在大型图中的可行性与可扩展性。

A. 相关工作

近似数据依赖的发现并非数据库和数据挖掘领域的新课题，尤其是在关系数据场景中。关于近似函数依赖（FD）的发现，已有大量研究 [16]–[17], [21], [35]–[38] ，并提出了许多扩展形式（例如条件函数依赖 [39]–[41], 、基于距离的函数依赖 [15], [20], [31]–[34], 等）。总体目标是找出一个精简的有效依赖关系集合，使其在给定输入数据上（几乎）成立。然而，对于大多数依赖关系（包括 GED），这都是一项颇具挑战且往往难以处理的问题。对于图实体依赖，其困难在于，

*通讯作者：Zaiwen.Feng@mail.hzau.edu.cn

这是由四个因素造成的：a）存在作为拓扑约束的图模式；b）GED 的左侧/右侧集合包含三种可能的文字；c）GED 的蕴含分析和验证分析已被证明分别是 NP 完全和 co-NP 完全问题（详见 [3]）；d）数据可能包含错误和不一致之处。

II. 预备知识

本节介绍全文中使用的基本定义和概念，回顾 GED 的语法与语义，并讨论文献中依赖关系的近似满足。

A. 基本定义与概念

属性图、图模式及图模式匹配的定义遵循文献 [3], [4] 中的定义。我们使用字母表 **A**，**L**，**C** 分别表示属性、标签 和常量的可数无限全集。

图。我们考虑一个有向属性图 $G=(V,E,L,F_A)$，其中：(1) V是一个有限节点集；(2) E是一个有限边集，由 $E\subseteq V\times \mathbf{L}\times V$ 给出，其中 (v,l,v') 是一条从节点 v 指向节点 v'、标签为 $l\in \mathbf{L}$ 的边；(3) 每个节点 $v\in V$ 都有一个表示其身份的特殊属性 id，以及一个取自 L 的标签 $L(v)$；(4) 每个节点 v都有一个关联列表 $F_A(v)=[(A_1,c_1),...,(A_n,c_n)]$，其中包含属性-值对；这里 $c_i\in$C 是一个常量，$A_i\in$A 是 v 的一个属性，记作 $v.A_i=c_i$，且 $A_i\neq A_j$ 当且仅当 $i\neq j$。

图模式。图模式（记为 $Q[\bar{u}]$）是一个有向图 $Q[\bar{u}]=(V_Q,E_Q,L_Q)$，其中：(a) V_Q 和E_Q 分别表示图模式节点集和图模式边集；(b) L_Q 是一个标签函数，为每个节点 $v\in V_Q$ 和每条边 $e\in E_Q$分配一个标签；(c) $\bar{u}$ 是 V_Q中的所有节点，称为（图模式）变量。所有标签均取自 **L**，其中包括作为特殊标签的通配符“*”。若两个标签 $l,l'\in\mathbf{L}$ 满足以下条件之一，则称它们可以匹配，记为 $l\asymp l'$：(a)$l=l'$；或 (b) l 或 l' 为“*”。

图模式 $Q[\bar{u}]$ 在图 G 中的**匹配** 是一个从 Q 到 G 的同态 h，满足：(a) 对于每个节点 $v\in V_Q$，$L_Q(v)\asymp L(h(v))$；(b) 对于每条边 $e=(v,l,v')\in E_Q$，在 G中存在一条边 $e'=(h(v)\ l',h(v'))$，使得$l\asymp l'$。我们用 $H(\bar{u})^{\dagger}$表示 $Q[\bar{u}]$ 在 G 中的所有匹配构成的集合。图、图模式及其匹配示例见下文示例1。

示例 *1*（图、图模式与匹配）：图 1 (a) 展示了一个简单的图；而在图 1 中，(b)–(d) 是三个图模式的示例。我们在 G 中给出每个图模式及其对应同态匹配的语义如下：i) $Q_1[x,y]$ 描述了两个通用（标记为“*”）节点之间的“is a”关系 x 和 y。该模式在示例图中的匹配列表为 $H_1(x,\ y)=[\{1,0\},\{2,0\}]$。ii) $Q_2[x,y]$表示一个公司节点 x 与一个产品节点 y之间存在 create 关系，其在 G 中的匹配为：$H_2(x,\ y)=[\{1,4\}\{1,5\}\{2,6\},\{3,7\},\{3,8\},\{3,9\},\{11,12\},\{11,14\}]$。

†简单地 H 当上下文清晰时

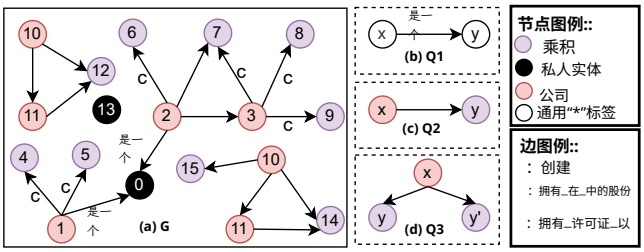

图 1. 示例：（a）图；（b）–（d）图模式

iii) $Q_3[x,y,y']$ 指定一个公司节点 x 与两个产品节点 y，y'之间存在 create 关系。因此，G中的匹配为 $H_3(x,\ y,\ y')=[\{1,4,5\},\{3,7,8\},\{3,7,9\},\{3,8,9\}]$。□

B. 图实体依赖（*GED*）

这里回顾图实体依赖（GED）的定义 [3], [4]。

GED 语法。一个 GED φ 是一个二元组（$Q[\bar{u}],X\to Y$），其中$Q[\bar{u}]$ 是一个图模式，X,Y 是 $\bar{u}$中的两个（可能为空的）文字集合。$\bar{u}$ 中的文字 w 是以下三种约束之一：（a）$x.A=c$，（b）$x.A=y.A'$，（c）$x.id=y.id$，其中 x，$y\in\bar{u}$是模式变量，$A,A'\in$**A**是非id 属性，$c\in$**C** 是一个常量。$Q[\bar{u}]$和 $X\to Y$ 分别称为 φ的模式/作用域 和依赖 。

图实体依赖语义。给定一个图实体依赖 $\varphi=(Q[\bar{u}],X\to Y)$，一个匹配 $h(\bar{u})$，属于 $Q[\bar{u}]$，位于 G 满足 文字 w，属于 $\bar{u}$，记作 $h(\bar{u})\models w$，如果：(a) 当 w 为 $x.A=c$ 时，则属性$h(x).A$ 存在，且 $h(x).A=c$；(b) 当 w 为 $x.A=y.A'$ 时，则属性 $h(x).A$ 和 $h(y).A'$ 存在，并且具有相同的值；以及 (c) 当 w 为 $x.id=y.id$ 时，则 $h(x).id=h(y).id$。

一个匹配 $h(\bar{u})$ 满足一组 X 文字，当 $h(\bar{u})$满足每个文字 $w\in X$ 时（对 $h(\bar{u})\models Y$ 也有类似定义）。我们记 $h(\bar{u})\models X\to Y$，如果 $h(\bar{x})\models X$ 蕴含 $h(\bar{x})\models Y$。

图 G 满足图实体依赖 φ（记作 $G\models\varphi$），如果对于 G 中 $Q[\bar{u}]$ 的所有匹配 $h(\bar{u})$，$h(\bar{x})\models X\to Y$。图 G满足一组GED（记作 $G\models\Sigma$），如果对于所有 $\varphi\in\Sigma$，$G\models\varphi$。

在示例2中，我们通过图1中的示例图和图模式说明GED的语义。

示例*2*（*GED* 的语义）：我们使用图 1(a)中的样本图，并采用图 1(b)–(d)中的图模式，来定义示例 GED：

1) φ_1：$(Q_1[x,y],y.A=y.A\to x.A=y.A)$– 此图实体依赖规定，对于模式 $Q_1[x,y]$ 在 G 中的任何匹配 $h(x,y)$（即，x 是_一个 y），如果节点 $h(y)$ 具有属性 A，那么 $h(x)$在 A 上必须与 $h(y)$ 具有相同的值。

2) φ_2: $(Q_2[x,y],\emptyset\to y.creator=x.name)$– 对于 Q_2 在 G 中的每个匹配 $h(x,y)$（即，公司 $h(x)$ 创建乘积 $h(y)$），那么 $h(y.creator)$ 和 $h(x.name)$ 必须具有相同的值。

3) φ_3：$(Q_3[x,y,y'],\ \emptyset\to y.creator=y'.creator)$——这表示，对于 G 中的任意匹配项 $h(x,\ y,\ y')$（即公司 $h(x)$

创建两个产品 $h(y,\ y')$），那么 $y,\ y'$ 必须在其属性上具有相同的值 $creator$。 □

C. 依赖关系的近似满足

在关系数据模型中，关于近似成立的依赖关系已有大量研究，因为它们在许多现实应用中具有实际意义。例如，已有多种度量被提出，用于量化两组属性/字面量集合之间存在的依赖程度，这些度量采用了信息论思想 [12]–[14], 概率方法 [18]和非概率方法 [21]。我们建议感兴趣的读者参阅 [22] ，以了解该主题近期的一项比较研究。

本工作旨在为图依赖关系（尤其是图实体依赖，GED）中依赖程度提供一种直观、计算高效且有效的度量。因此，我们基于图依赖关系的语义，扩展了 [21] 中的误差度量定义，用于近似图中的依赖关系。

GED的误差度量。考虑图 G中图模式 $Q[\bar{u}]$ 的匹配集合 $H(\bar{u})$。如果：$h \models X$，但 $h \not\models Y$，我们称匹配 $h \in H$违反了依赖 $X \to Y$ 在 $Q[\bar{u}]$ 上的约束。也就是说，直观地，依赖 $X \to Y$ 在 $Q[\bar{u}]$ 上的约束成立，当且仅当 H中不存在违反约束的匹配。因此，我们定义误差度量 e_3——类似于 g_3 中的 *[21]*——为消除以获得满足性所需删除的最少匹配数量。

令 E_3 表示需要从 H 中消除的匹配数量，以使 $X \to Y$ 成立，表示为:

$$E_3(X \to Y, H) = |H| - max\{|J| \mid J \subseteq H, J \models X \to Y\}.$$

图依赖关系的 e_3 误差度量如下公式 1 所示:

$$e_3(X \to Y, H) = E_3/(|H| - |adom(X)|), \qquad (1)$$

其中，$adom(X)$‡ 是 X上文字的活动域。我们指出，e_3 的分母不同于 g_3 ，并基于 [14]中的观察，即分子有上界——这与 [22]中修订后的 g'_3 定义一致。 e_3 的取值范围为 0 到 1；取值为 0 表示该依赖完全满足，取值为 1 则表示该依赖无效。介于 0 和 1 之间的任意取值都表示该依赖近似满足。

我们在定义 1 中如下定义相对于 e_3 的近似图实体依赖和误差阈值$\epsilon \in [0, 1]$ 。

定义 *1*（近似图实体依赖）：设 ϵ 为用户指定的误差阈值。若图实体依赖 σ：$(Q[\bar{u}],\ X \to Y)$ 在 G 中满足以下条件，则它是一个近似图实体依赖：$e_3(X \to Y,\ H(\bar{u})) \leq \epsilon$。 □

假设图2中的伪关系表表示模式 Q_3 在 G 中的匹配的属性值（来自图1）。在示例 3 中，我们基于 e_3 阐述近似图实体依赖的示例。

示例 *3*（衡量图依赖关系的误差）：考虑 GED σ_4：（$Q_3[x, y, y']$, $X_1 \to X_2$），以及 σ_5：（$Q[x, y, y']$, $X_3 \to X_1$），其中 X_1 是 {y.年 =y'.发行的}, X_2 是 {y.名称 = y'.名称}，且 X_3 是{x.名称 = EA}。我们计算 e_3（σ_4, H）= $\frac{4-1}{4-2}$ = 0.5 ——只需移除一个匹配项（h_1）和 $|adom$（X_1）| = 2；以及 e_3（σ_5, H）= $\frac{4-2}{4-1}$ = 0.667 ——需要移除两个匹配项（h_3, h_4）和 $|adom$（X_3）| = 1。 □

‡变量字面量的最大 $|adom(X)|$ 为 2，而常量字面量的最大值为 1。

III. 问题形式化

GED 的挖掘问题在 [11]中进行了研究。本节回顾有关 GED 发现的相关概念和定义，并将该问题扩展到近似满足场景。

给定一个属性图 G，一般的图实体依赖发现问题是寻找在 G上成立的 最小覆盖集GED。因此，直观地说，近似图依赖关系发现就是基于给定的误差度量和误差阈值，寻找几乎成立的最小覆盖集 GED。下面，我们将对该问题进行表述，并给出更正式的定义。

A. 持久、约简和最小图编辑距离依赖

在规则发现中，返回一组简洁且无冗余的规则至关重要。因此，与数据依赖发现文献 [11], [23]–[25], 一致，我们关注持久图模式上的非平凡 和非冗余 依赖关系集合。

持久化图模式。令 $M = \{m_1, \cdots, m_k\}$为模式 $Q[\bar{u}]$ 到图 G的同构集合；且 $D(v) = \{m_1(v) \cdots, m_k(v)\}$ 为包含 G中不同节点的集合，其函数 $m_1, \cdots, m_k$映射一个节点$v \in V$。

基于图像的最小支持§ (MNI) [26] 的 Q 在G 中，记作 $mni(Q, G)$，定义如下:

$$mni(Q, G) = min\{x \mid x = |D(v)|,\ \forall\ v \in V\}.$$

如果图模式 $Q[\bar{u}]$ 在图 G中满足 $mni(Q, G) \geq \tau$，其中 $\tau \in \mathbb{N}$ 是用户指定的最小 MNI 阈值，我们称其为 持久的 （即 频繁的）。

平凡 GED。如果满足以下条件之一，我们称 GED φ：$(Q[\bar{u}],\ X \to Y)$ 为平凡的：(a) X 中的文字集合无法满足（即 X 的计算结果为 false）；或 (b) Y 可由 X 推导得到（即 $\forall\ w \in Y$、 w 可通过等式运算符的传递性由 X 推导得到）。我们只关注挖掘非平凡 GED。

约简 GED。给定两个模式 $Q[\bar{u}] = (V_Q,\ E_Q,\ L_Q)$ 和 $Q'[u'] = (V'_{Q'},\ E'_{Q'},\ L'_{Q'})$，如果满足以下条件之一，则称 $Q[\bar{u}]$ 约简 $Q'[u']$，记作 $Q \ll Q'$：(a) $V_Q \subseteq V'_{Q'}$、 $E_Q \subseteq E'_{Q'}$；或 (b) L_Q 将 $L'_{Q'}$ 中的某些标签升级为通配符。也就是说， Q 是比 Q' 限制性更弱的拓扑约束。

给定两个 GED， $\varphi = (Q[\bar{u}], X \to w)$ 和 $\varphi' = (Q'[u'], X' \to w')$。如果满足以下条件，则称 φ 可约化 φ'，记作 $\varphi \ll \varphi'$：(a) $Q \ll Q'$；以及 (b) $X \subseteq X'$ 和 $w = w'$。

因此，如果：(a)$G \models \varphi$；并且对于任意 φ'，只要 $\varphi' \ll \varphi$、$G \not\models \varphi'$，我们就称图 G 中的图实体依赖 φ 是 **约简** 的。我们称图实体依赖 φ是**极小** 的，当且仅当它同时是非平凡的和约简的。

§鉴于该度量具有较高的计算效率和反单调性，我们采用了它

Pattern variables	x		y					y'				
Attributes	id	Name (A1)	id	Name (A2)	Genre (A3)	Year (A4)	Price (A5)	id	Name (A2)	Category (A3)	Released (A4)	Cost (A5)
Matches												
h1	1	GL	4	AF9	Racing	2018	$50	5	AF11	Racing	2018	$55
h2	3	EA	7	F20	Soccer	2019	$55	8	F20	Soccer	2019	$60
h3	3	EA	7	F20	Soccer	2019	$55	9	F21	Soccer	2020	-
h4	3	EA	8	F20	Soccer	2019	$60	9	F21	Soccer	2020	-

图 2. Q_3 in G 的匹配伪关系（参见图 1）

B. 问题定义

本文提出并研究了近似 GED 发现问题。

定义 *2*（近似 *GED* 发现问题）：给定一个属性图、 G、用户指定的 MNI 阈值 $\tau \in \mathbb{N}$ 和误差阈值 $e_3 \quad \epsilon \in [0,1]$。近似 GED 发现旨在找到 G 中所有极小 GED 的集合 Σ，使得对于每个 $\sigma: (Q[\bar{u}],\ X \to Y) \in \Sigma$，均有：a) $mni(Q,\ G) \geq \tau$，且 b) $e(X \to Y,\ H) \leq \epsilon$。 □

下一节将介绍我们的近似 GED 发现方案 FASTAGEDS。

IV. 快速近似 GED 挖掘

GED 描述了图中拓扑模式上的属性依赖。因此，我们首先考察图模式的发现（IV-A 小节），这些模式将作为属性依赖发现的“范围”（IV-B 小节）。我们在算法 1 中给出了该方案的简化伪代码。

A. 图模式（范围）发现

这里，我们简要讨论如何发现频繁图模式，并将其作为挖掘属性依赖的“作用范围”或松散模式。我们并未声称在这项任务上有新的贡献——这在数据挖掘和数据库领域文献中已是一个研究充分的主题 [27]–[30]。

在这项工作中，我们采用高效的基于 MNI 的技术 GRAMI [28], 来查找 τ-频繁图模式，将其作为我们依赖关系的作用范围（算法1 的第1行）；并在 [11] 中采用剪枝策略，以返回一组约简图模式（第2 行）。随后，我们在下文中基于每个约简图模式的匹配集合挖掘依赖关系。

B. 依赖挖掘

设 $\mathcal{Q}$ 为属性图 G 中的精简 τ-频繁图模式集合。我们基于用户指定的最低阈值 ϵ，为 G 中的每个模式 $Q \in \mathcal{Q}$ 挖掘近似图实体依赖的最小覆盖集 $\Sigma(Q, G, \epsilon)$，位于算法 1 的第 3–6 行。

新颖的 GED 刻画。我们提出了一种基于不一致集概念的新颖 GED刻画，用于满足性研究；该概念最初由 [35], [36]引入，用于关系函数依赖发现。

令 $H(Q[\bar{u}],\ G)$ 为输入图 G 中 $Q[\bar{u}] \in \mathcal{Q}$ 的同态匹配。我们引入以下概念：模式变量 $x,\ y \in \bar{u}$ 在 Q 的属性上的项/项集、映射到常量的变量以及标识符字面量。

算法 1 FASTAGEDS(G, τ, ϵ)

需要： 图 G，MNI 阈值 τ，误差阈值 ϵ
确保： 一个最小覆盖集 Σ 个GED，位于 G，使得 $\forall\, \sigma \in \Sigma$：$mni(\sigma.Q) \geq \tau \wedge e_3(\sigma) \leq \epsilon$。

1: $\mathcal{Q} \leftarrow GraMi(G, \tau)$ ▷ 挖掘频繁图模式
2: $\mathcal{Q}' \leftarrow reduce(\mathcal{Q})$ ▷ 消除冗余
3: **对于** $Q \in \mathcal{Q}'$ **执行**
4: $H(Q, G) \leftarrow match(Q, G)$ ▷查找匹配
5: $\Sigma_Q \leftarrow mineDep(H(Q, G)\ \epsilon)$ ▷ 查找近似的图实体依赖
6: $\Sigma = \Sigma \cup \Sigma_Q$
7: **返回** Σ

定义 *3*（项、项集与项集满足关系）：设项名 α 是图模式 Q 的模式变量 $x \in \bar{u}$ 的一个属性。项是由项名、模式变量 $u \subseteq \bar{u}$ 和常量 $c \in \mathbf{C}$ 组成的三元组，记作 $\alpha[u;\ c]$。项集 $X = \{\alpha_1[u_1;\ c_1],\ \cdots,\ \alpha_n[u_n;\ c_n]\ |\forall\ i,\ j \in [1,n],\ i \neq j\}$ 是由项名各不相同的项组成的集合。如果匹配 $h \in H$ 满足项集 X，当且仅当：$h(u_i) \models \alpha_i[u_i;\ c_i],\ \forall\ \alpha_i[u_i; c_i] \in X$。 □

$\alpha[u$的一个项；$c]$ 分别唯一映射到 $\bar{u}$ 的 Q 的常量、变量和标识符文字，如下所示：w 的 $\bar{u}$；

- 当 w 为 $x.A = c$时，则其项为 $\alpha[x;\ c]$
- 当 w 为 $x.A_1 = y.A_2$时，则其项为 $\alpha_{12}[x, y;\ _]$
- 当 w 为 x.标识符 $= y$.标识符时，则其项为标识符$[x, y;\ _]$。

因此，给定图模式 $Q[\bar{u}] \in \mathcal{Q}$，我们为模式变量 $x \in \bar{u}$ 的所有属性定义常量项，并在相同类型的模式变量对 $x,\ y \in \bar{u}$ 上定义变量项。

定义 *4*（分歧集与必要集）：给定匹配集合 H 的 $Q[\bar{u}]$在 G:

- 一个匹配的不一致集，$D(h)$，由 $h \in H$ 给出

$$D(h) = \{\alpha[u; c] \mid h \not\models \alpha[u; c]\}.$$

- 项的必要集，$\alpha[u; c]$，定义为

$$nec(\alpha[u; c]) = \{D(h) \setminus \alpha[u; c] \mid h \in H, \alpha[u; c] \in D(h)\}.$$

令 $\mathcal{I}$ 表示图模式 $Q[\bar{u}]$ 上所有项的集合，$P(\mathcal{I})$ 表示 $\mathcal{I}$ 的幂集。给定 $M \subseteq \mathcal{I},\ X \subseteq P(\mathcal{I})$，集合 M **覆盖项** X 当且仅当，对于所有 $Y \in X$，我们有 $Y \cap M \neq \emptyset$。此外，如果不存在能够覆盖 X 的 $M' \subset M$，则 M 是 X 的极小覆盖。

引理 *4.1*（*GED* 刻画）：给定项集 $X \subseteq \mathcal{I}$ 以及项 $\alpha[u; c] \notin X$ 在 $Q[\bar{u}]$ 上，$X \to \alpha[u; c]$ 在H 上成立，当且仅当：X 覆盖项 $nec(\alpha[u; c])$。 □

因此，我们将模式 $Q[\bar{u}]$ 的 GED 发现的极小覆盖问题归约为：对每个项 $\alpha[u; c] \in \mathcal{I}$，找出所有 $nec(\alpha[u; c])$ 的极小覆盖。

不一致集的构建。由前文可知，快速生成分歧集与必要集至关重要。我们基于 H的伪关系，构建匹配项 H 上的二元关系 $\mathcal{B}(\mathcal{I},\ H)$，其中匹配项 H 来自 $Q[\bar{u}]$中模式变量上的项集合 $\mathcal{I}$ 。

Pattern variables	x	y, y'				
Items	$A_1[x; EA]$	id[y, y'; _]	$A_2[y, y'; _]$	$A_3[y; Soccer]$	$A_3[y, y'; _]$	$A_4[y, y'; _]$
Matches						
h1	0	0	0	0	1	1
h2	1	0	1	1	1	1
h3	1	0	0	1	1	0
h4	1	0	0	1	1	0

图3. 项上的不一致表示例 Q_3

给定 $\mathcal{I}$ 和 H，对于每个 $h \in H$ 和 $\alpha[u;\ c] \in \mathcal{I}$

$$\mathcal{B}(h(\alpha[u;c])) := \begin{cases} 1, & \text{if } h \models \alpha[u;c] \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

我们通过一次扫描关系 $\mathcal{B}(\mathcal{I},\ H)$，为所有 $h \in H$ 构造不一致集 $D(h)$，并由此为所有项$\alpha[u$构造必要集 $nec(\alpha[u;c])$；$c] \in \mathcal{I}$

示例 *4*（不一致集与必要集）：使用运行示例，图2中的二元关系如图3所示，针对选定的一组项，使用公式2得出。由此得到的匹配项的不一致集如下¶:

$$D(h_1) = \{a_1, id, a_2, a_3^1\}; \qquad D(h_2) = \{id\};$$
$$D(h_3) = \{id, a_2, a_4\}; \qquad D(h_4) = \{id, a_2, a_4\}.$$

此外，这些项的必要集为:

$$D(id) = \{\{a_1, a_2, a_3^1\}, \{a_2, a_4\}\}; \qquad D(a_1) = \{\{id, a_2, a_3^1\}\};$$
$$D(a_2) = \{\{a_1, id, a_3^1\}, \{id, a_4\}\}; \qquad D(a_3^1) = \{\{a_1, id, a_2\}\}$$
$$D(a_4) = \{\{id, a_2\}\}.\square$$

搜索策略。给定项 $\alpha[u;\ c]$，其候选项集空间为 $P(\mathcal{I} \setminus \alpha[u;\ c])$，显然相对于大小 $|\mathcal{I}|$（即 $2^{|\mathcal{I}|-1}$）呈指数增长。因此，必要集 $nec(\alpha[u;c])$ 缩小了 $\alpha[u;\ c]$ 的候选左侧（LHS）。具体而言，要找出给定项 $\alpha[u;\ c]$ 的所有有效左部，只需找出 $nec'(\alpha[u;c])$ 的所有覆盖项集合，其中:

$$nec'(\alpha[u;c]) = \{D \in nec(\alpha[u;c]) \mid \nexists\ D' \subset D \in nec(\alpha[u;c])\}.$$

我们使用深度优先、从左到右策略，构造并遍历给定项 $\alpha[u;c]$ 及其极小必要集$nec'(\alpha[u;c])$的候选项-集格。格中的每个节点由两个元素组成：必要集（包含在“{}”中）以及其组成项（位于“[]”中）。我们按照组成项在必要项集中的出现顺序递减排列节点处的组成项，并按字典序打破平局。搜索树的根节点设置为 $nec'(\alpha[u;c])$。每个节点最多拥有其组成集中的唯一项数量个子节点，并且指向子节点的边以其对应的项标记。子节点的必要集是 $nec'(\alpha[u;c])$中尚未覆盖的项集集合。

存在三种叶节点的情况：a) 两个节点元素都为空；b) 仅剩余必要集是

¶为简洁起见，我们用小写字母表示项，例如 a_1 是$A_1[x; EA]$, a_3^1 是 $A_3[y; Soccer]$

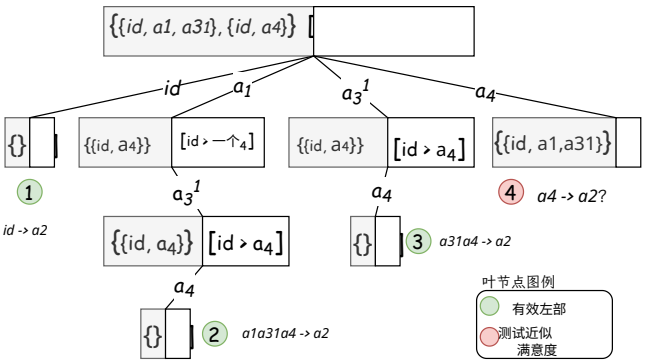

图4. FASTAGEDS 针对 $A_2[y, y$ 的 DFS 格；_]

表 I 数据集汇总

数据集	#节点	#边	#节点类型	#边类型
DBLP	300K	800K	3	3
IMDB	300K	1.3M	5	8
YAGO4	4.37M	13.3M	7764	83

为空；并且 c）只有组成项集合为空。这些情况产生的分支分别对应有效左部、非极小左部和无效左部。

示例 *5*（查找最小覆盖）：针对 $A_2[y, y;\ ]$（即 a_2）的候选左侧的深度优先搜索树如图 4 所示。叶节点 1，2，3 产生了完全满足条件的有效 GED，而叶节点 4 则产生了无效 GED。因此，我们根据指定的误差阈值，测试这些节点是否近似满足条件。例如，令 $\epsilon = 0.25$；我们计算$e_3($ $Q_3[x, y, y'],\ a_4 \to a_2) = \frac{1}{2}$。因此，叶节点 4不会产生近似 GED，除非 $\epsilon \geq 0.5$。

V. 实验

接下来，我们展示用于发现最小图编辑距离依赖的算法 FASTAGEDS 的实验研究。我们在不同的误差阈值和属性数量下进行了实验，以探究它们对可扩展性以及挖掘出的最小图编辑距离依赖数量的影响。

A. 实验设置

数据集。实验使用了三个真实世界的属性图：(a)*IMDB* [42] (b) *YAGO4* [43] (c)*DBLP* [44]。表 I 总结了这些数据集的详细信息。

实验。本文提出的所有算法均使用 Java 实现；实验在一台配备 128GB 内存、搭载 Linux 操作系统的计算机上运行，该计算机配有 2.20GHz 英特尔至强处理器。

B. 提案的可行性与可扩展性。

我们在所有实验中均设置 MNI 阈值 $\tau = 2$ 。

实验 a. 我们首先评估算法在三个数据集中挖掘（近似）GED 的时间性能，并考察不同错误率（即 $\epsilon = 0.0$，0。15，0。30）以及属性数量变化时的表现。结果见图 5 a) –

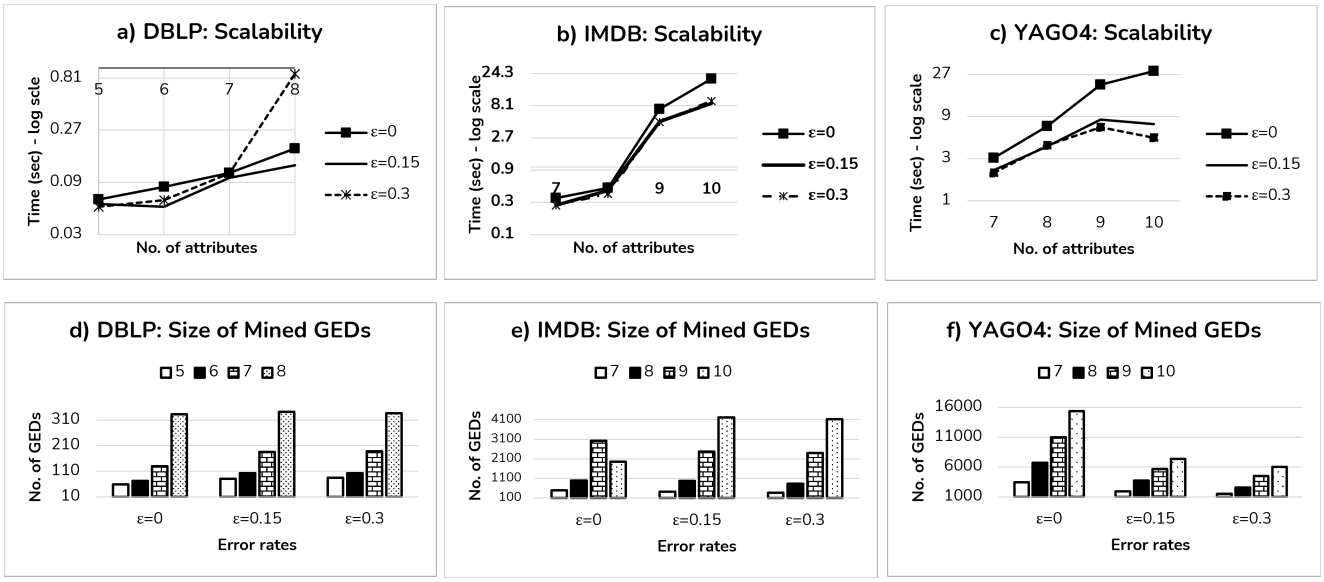

图 5. FASTAGEDS：实验结果

c)，分别对应*DBLP*、*IMDB* 和 *YAGO4* 数据集。结果表明，无论设定的最大误差是多少，随着属性数量增加，耗时都会显著增加。此外，与其他数据集相比，在 DBLP 数据中发现 GED 的效率最高，因为其频繁模式产生的匹配数量最少

实验-b.接下来，我们报告在属性数量不同的情况下，误差阈值与挖掘出的GED 数量之间的相关性（参见图5d）– f））。在这组实验中，我们设置了一系列阈值来开展实验，阈值范围为 0 到 0.3。总体而言，对于给定的错误率，挖掘出的规则数量会随着属性数量的增加而增加。然而，在固定属性数量下，错误率增加所产生的影响因数据集而异。除*YAGO4* 数据外，其他数据集中的情况较为复杂；在 YAGO4数据中，可以观察到明显的下降趋势。*YAGO4* 数据集中的这一现象可以用其错误性来解释——它比其他数据集更加脏。实际上，错误率增加原则上会返回更多规则，但后续的蕴含分析会剪枝掉更多冗余规则。因此，最终得到的规则集更加简洁。

VI. 结论

本文提出并研究了 GED 的近似发现问题。我们基于 GED 的语义提出了一种误差度量，并通过必要集和覆盖集的概念刻画其满足情况。我们开发的算法 FASTAGEDS 采用深度优先搜索策略遍历 GED 的候选空间，以寻找完整和近似 GED。最后，我们通过在三个真实世界图上进行实证评估，证明 FASTAGEDS 具有良好的有效性和可扩展性。

参考文献

[1] Fan, W., Fan, Z., Tian, C. 和 Dong, X.L., 2015年. 图的键. 《VLDB基金会会刊》, 8(12), pp.1590-1601. [2] Fan, W., Wu, Y. 和 Xu, J., 2016年6月. 图的函数依赖. 载于2016年数据管理国际会议论文集 (pp. 1843-1857). [3] Fan, W. 和 Lu, P., 2017年5月. 图的依赖关系. 载于第36届ACM SIGMOD-SIGACT-SIGAI PODS研讨会 (pp. 403-416). ACM. [4] Fan, W. 和 Lu, P., 2019年. 图的依赖关系. 《ACM数据库系统汇刊》(TODS), 44(2), pp.1-40. [5] Lin, P., Song, Q. 和 Wu, Y., 2018年. 基于本体子图模式的知识图谱事实核查. 《数据科学与工程》, 3(4), pp.341-358. [6] Kwashie S, Liu L, Liu J, Stumptner M, Li J, Yang L. Certus：一种基于图差分依赖（GDD）的有效实体消解方法. 《VLDB基金会会刊》. 2019年2月1日;12(6):653-66. [7] Ma, H., Alipourlangouri, M., Wu, Y., Chiang, F. 和 Pi, J., 2019年. 属性图中基于本体的实体匹配. 《VLDB基金会会刊》, 12(10), pp.1195-1207. [8] Fan, W., Lu, P., Tian, C. 和 Zhou, J., 2019年. 推导图的确定性修复. 《VLDB基金会会刊》, 12(7), pp.752-765. [9] Alipourlangouri, M. 和 Chiang, F., 2018年. Keyminer：发现图的键. 载于VLDB研讨会 TD-LSG. [10] Wenfei Fan, Chunming Hu, Xueli Liu, 和 Ping Lu. 2020年. 发现图函数依赖. 《ACM数据库系统汇刊》 45, 3, 第15篇 (2020年9月), 42页. [11] Liu, D., Kwashie, S., Zhang, Y., Zhou, G., Bewong, M., Wu, X., Guo, X., He, K. 和 Feng, Z., 2023年. 一种发现图实体依赖（GED）的高效方法. arXiv预印本 arXiv:2301.06264. [12] Reimherr, M. 和 Nicolae, D.L., 2013年. 论依赖的量化：开发可解释度量的框架. [13] Kramer, S. 和 Pfahringer, B., 1996年. 强部分决定的高效搜索. 载于知识发现与数据挖掘国际会议论文集 (pp. 371-378). [14] Giannella, C. 和 Robertson, E., 2004年. 关于函数依赖的近似度量. 《信息系统》, 29(6), pp.483-507. [15] Kwashie, S., Liu, J., Li, J. 和 Ye, F., 2014年7月. 挖掘差分依赖：一种子空间聚类方法. 载于澳大拉西亚数据库会议 (pp. 50-61). Springer, Cham. [16] Mandros, P., Boley, M. 和 Vreeken, J., 2017年8月. 发现可靠的近似函数依赖. 载于第23届ACM SIGKDD知识发现与数据挖掘国际会议论文集 (pp. 355-363).

[17] Huhtala, Y., K¨arkk¨ainen, J., Porkka, P. 和 Toivonen, H., 1999年. TANE：一种发现函数依赖与近似依赖的高效算法 .《计算机杂志》 , 42(2), pp.100-111. [18] Piatetsky-Shapiro, G., 1992年7月. 概率数据依赖. 载于机器发现研讨会（苏格兰阿伯丁）. [19] Piatetsky-Shapiro, G. 和 Matheus, C., 1993年7月. 度量大型数据库中的数据依赖. 载于数据库知识发现研讨会 (pp. 162-173). [20] Jixue Liu, Selasi Kwashie, Jiuyong Li, Feiyue Ye, 和 Millist Vincent. 近似差分依赖的发现. cs.DB. 2013年. [21] Kivinen, J. 和 Mannila, H., 1995年. 从关系中近似推断函数依赖.《理论计算机科学》, 149(1), pp.129-149. [22] Weytjens, S., 2021年. 近似函数依赖：度量比较及一种以相关性为重点的发现工具. [23] Fan, W., Hu, C., Liu, X. 和 Lu, P., 2020年. 发现图函数依赖.《ACM数据库系统汇刊》(TODS), 45(3), pp. 1 42 - Liu, J., [24] Li, J., Liu, C. 和 Chen, Y., 2010年. 从数据中发现依赖关系——综述.《IEEE知识与数据工程汇刊》, 24(2), pp.251-264. [25] Caruccio, L., Deufemia, V. 和 Polese, G., 2020年. 从数据中挖掘松弛函数依赖.《数据挖掘与知识发现》, 34(2), pp.443-477. [26] Bringmann, B. 和 Nijssen, S., 2008年. 单个图中什么是频繁的？. 载于《知识发现与数据挖掘进展：第12届亚太会议，PAKDD 2008 日本大阪，2008年5月20-23日, 论文集12》 (pp. 858-863). Springer Berlin Heidelberg. [27] Khan, A., Yan, X. 和 Wu, K.L., 2010年6月. 面向大型图中的邻近模式挖掘. 载于2010年ACM SIGMOD数据管理国际会议论文集 (pp. 867-878). [28] Elseidy, M., Abdelhamid, E., Skiadopoulos, S. 和 Kalnis, P., 2014年. Grami：单个大型图中的频繁子图与模式挖掘.《VLDB基金会会刊》, 7(7), pp.517-528. [29] Aberger, C.R., Lamb, A., Tu, S., N¨otzli, A., Olukotun, K. 和 R´e, C., 2017年. Emptyheaded：一种用于图处理的关系引擎.《ACM数据库系统汇刊》(TODS), 42(4), pp.1-44. [30] Teixeira, C.H., Fonseca, A.J., Serafini, M., Siganos, G., Zaki, M.J. 和 Aboulnaga, A., 2015年10月. Arabesque：一种分布式图挖掘系统. 载于第25届操作系统原理研讨会论文集 (pp. 425-440). [31] Koudas, N., Saha, A., Srivastava, D. 和 Venkatasubramanian, S., 2009年3月. 度量函数依赖. 载于2009年IEEE第25届国际数据工程会议 (pp. 1275-1278). IEEE. [32] Caruccio, L., Deufemia, V. 和 Polese, G., 2020年. 从数据中挖掘松弛函数依赖.《数据挖掘与知识发现》, 34(2), pp.443-477. [33] Kwashie, S., Liu, J., Li, J. 和 Ye, F., 2014年7月. 挖掘差分依赖：一种子空间聚类方法. 载于澳大拉西亚数据库会议 (pp. 50-61). Springer, Cham. [34] Kwashie, S., Liu, J., Li, J. 和 Ye, F., 2015年9月. 条件差分依赖（CDDs）. 载于东欧数据库与信息系统进展会议 (pp. 3-17). Springer, Cham.[35] Mannila, H. 和 R¨aih¨a, K.J., 1987年9月. 依赖推断. 载于第13届超大型数据库国际会议论文集 (pp. 155-158). [36] Mannila, H. 和 R¨aih¨a, K.J., 1994年. 从关系推断函数依赖的算法.《数据与知识工程》, 12(1), pp.83-99. [37] Bleifuß, T., B¨ulow, S., Frohnhofen, J., Risch, J., Wiese, G., Kruse, S., Papenbrock, T. 和 Naumann, F., 2016年10月. 面向大型数据集的函数依赖近似发现. 载于第25届ACM信息与知识管理国际会议论文集 (pp. 1803-1812). [38] Kruse, S. 和 Naumann, F., 2018年. 近似依赖的高效发现.《VLDB基金会会刊》, 11(7), pp.759- 772.[39] Liu, J., Li, J., Liu, C. 和 Chen, Y., 2010年. 从数据中发现依赖关系——综述.《IEEE知识与数据工程汇刊》, 24(2), pp.251-264. [40] Golab, L., Karloff, H., Korn, F., Srivastava, D. 和 Yu, B., 2008年. 论为条件函数依赖生成近优表格.《VLDB基金会会刊》, 1(1), pp.376-390.

[41] Medina, R. 和 Nourine, L.， 2009年5月。《函数依赖、条件函数依赖和关联规则的统一层次结构》。载于 ICFCA（第9卷，第98–113页）。 [42] IMDB数据集。http://www.imdb.com/interfaces。 [43] YAGO4数据集。https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/。 [44] DBLP数据集。https://dblp.uni-trier.de/xml/。