

# Cypher查询等价性证明

汤磊, 窦文生, 郑莹莹, 许利杰, 王伟, 魏峻, 黄涛

## Proving Cypher Query Equivalence

The 41st IEEE International Conference on Data Engineering (ICDE 2025)

联系方式: 汤磊, tanglei20@otcaix.iscas.ac.cn

### Cypher查询语言

Cypher 是一种面向属性图数据库的声明式查询语言。通过图模式匹配的方式描述查询目标。

图模式

```
MATCH (writer :Person) -[:Write]-> (book :Book) <-[:Read]- (reader)
```

谓词

```
WHERE book.name = 'Harry Potter'
```

查询结果

```
RETURN book.name, reader.name
```

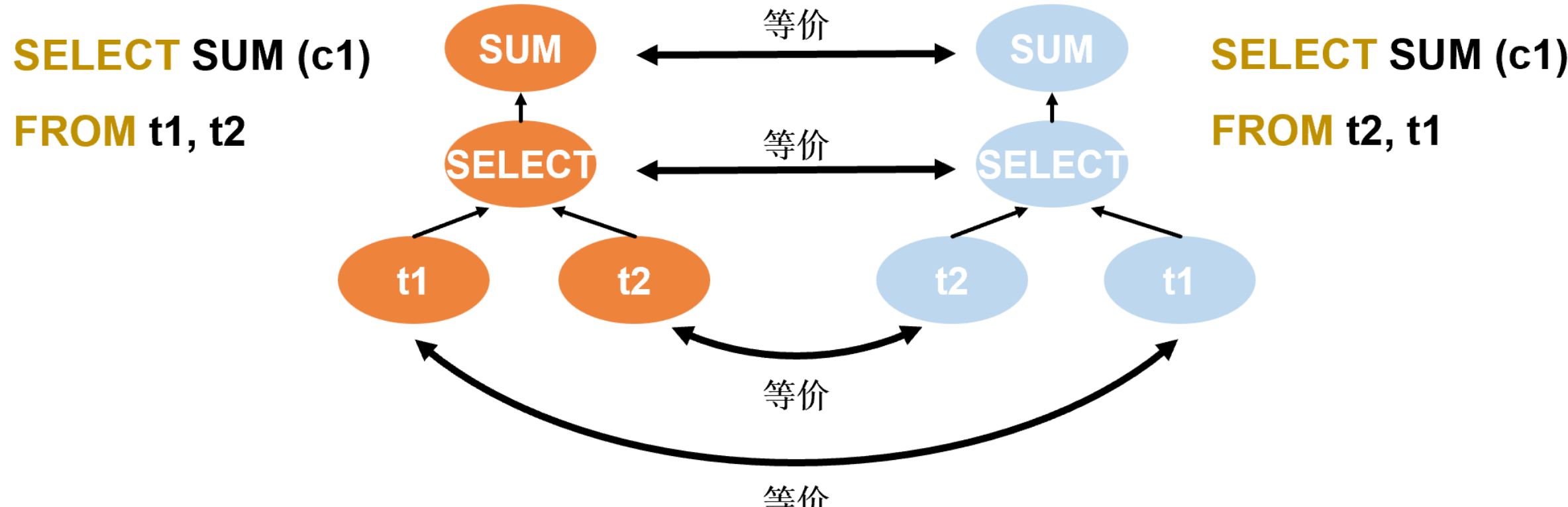
### 查询等价性问题

查询等价性问题是数据库研究领域的基础问题。两个图查询语义等价, 需要满足以下条件

- ◆ 两条查询的结果集包含相同的元组
- ◆ 任意元组在两条查询的结果集中的多重性 (出现次数) 相等
- ◆ 如果两条查询使用了 **ORDER BY** 排序, 则它们的结果集元组顺序一致

### 现有工作的不足

#### □ 现有工作



- ◆ 现有方法将SQL查询建模为形式化表达, 通过对比其结构或使用SMT求解器证明其等价性

#### □ 图查询与SQL查询的差异

图查询	SQL查询
属性图模型	关系型模型
图模式匹配	表元组计算

- ◆ 现有方法建模元组和数据表的对应关系, 不适用于属性图
- ◆ 现有的方法使用键组合关系元组, 无法正确建模Cypher查询语义

### 属性图和Cypher查询建模

#### □ 基于半环语义的代数表达式

$$Node(e) = \begin{cases} 1 & e \text{ 为节点} \\ 0 & \text{其他} \end{cases} \quad Lab(e, label_i) = \begin{cases} 1 & e \text{ 具有标签 } label_i \\ 0 & \text{其他} \end{cases} \quad Rel(e) = \begin{cases} 1 & e \text{ 为边} \\ 0 & \text{其他} \end{cases} \quad [\sigma] = \begin{cases} 1 & \sigma \text{ 为 true} \\ 0 & \sigma \text{ 为 false} \end{cases}$$

#### □ Cypher查询建模

```
MATCH (n1 :Person) -[ r1 :Read ]-> (n2 :Book) WHERE n2.name = 'Harry Potter' RETURN n1.name, n2.name
```

$$\sum_{n1, r1, n2} Node(n1) \times Lab(n1, Person) \times Node(n2) \times Lab(n2, Book) \times Rel(r1) \times Lab(r1, Read) \times [out(r1) = n1] \times [in(r1) = n2] \times [n2.name = 'Harry Potter']$$

#### □ SMT约束求解

$Q_1$  代数表达式:  $g_1(t) = \sum_{e_1, e_2, e_3} [t.c_1 = e_1] \times Node(e_1) \dots$

$Q_2$  代数表达式:  $g_2(t) = \sum_{e_1, e_2, e_3} [t.c_1 = e_3] \times Node(e_3) \dots$

$\exists t. g_1(t) \neq g_2(t)?$

### 实验验证

□ 通过翻译SQL等价数据集和重写真实Cypher查询, 构造了一个包含148对等价Cypher查询的数据集, 并基于此数据集进行实验验证

□ 有效性: 成功证明138/148对Cypher查询等价性

□ 效率: 平均证明延迟38ms

查询来源	等价查询对总数	成功证明数
SQL等价数据集	80	73
Cypher测试基准	45	42
Cypher开源工具	23	23
总数	148	138

