Semantic Proximity Search on Graphs  
with Metagraph-based Learning

# Abstract

鉴于无处不在的图形数据，如Web和社交网络，图形上的邻近搜索一直是一个活跃的研究课题。任务归结为测量图形上两个节点之间的接近度。尽管大多数早期的研究仅涉及同构或二分图，但许多真实世界的图与各种类型的对象是异构的，从而产生不同的邻近语义类。例如，在社交网络上，两个用户可以由于不同的原因而相近，例如是同学或家庭成员，其代表两种不同的邻近类别。因此，仅仅测量以前的作品所关注的“通用”接近形式就变得不够了。在本文中，我们将元图识别为一种新颖而有效的手段，用于表征所需的邻近类别的共同结构。随后，我们提出了一系列基于元图的接近度，并采用监督技术自动学习其家族内的正确接近形式以适应所需的类别。由于匹配（即找到实例）元图是昂贵的，我们提出了双阶段训练和基于对称的匹配的新方法来加速。最后，我们的实验表明我们的方法更加准确和有效。为了准确起见，我们在NDCG和MAP中的基线分别超过基线11％和16％。为了提高效率，双阶段培训将总体匹配成本降低了83％，基于对称性的匹配进一步降低了单个元图的成本52％。

# INTRODUCTION

Web和社交媒体的激增使得数据对象的集合越来越丰富。这些对象可以组织成图G =（V; E），其中节点V对对象进行建模，边E对它们的交互进行建模。这些图通常是异构的，包含不同类型的对象。考虑图1（a）中基于玩具社交网络的图表，该网络将各种用户及其属性互连。请注意，我们将每个用户和属性值视为一个节点，并且每个节点还与用户，学校或其他属性名称等类型相关联。我们称这个图形类型的对象图，在Sect.II中正式定义。

图表上的一个重要问题是邻近搜索。给定查询节点q，我们**如何测量其他节点与q的接近度，以便我们可以返回最接近q的节点**？然而，大多数早期的研究，包括Personalized PageRank [1]和SimRank [2]，**未能充分利用异构图中嵌入的丰富语义**。**具体地说，对于各种类型的互连对象，由于不同的潜在原因而存在不同的邻近语义类别，**如图1（b）所示。对于相同的查询节点（例如，Bob），**可能存在具有不同结果节点的多个接近类别**（例如，Alice作为家庭，而Tom作为同学）。因此，在不区分各种语义类别的情况下，仅仅测量“通用”形式的接近度是不够的。正如下面两种情景所设想的那样，这种差异化带来了无限可能性。

1. 基于圈子的朋友建议。在社交网络（例如，图1）上，通过圈子建议朋友极大地增强了用户体验。例如，谁和我对湖人的热情一样？ 谁是我的同学？ 每个案例都是一个圆圈或语义类，我们可以根据用户属性（如爱好和学校）区分它们。•
2. 上下文感知引文搜索。考虑连接论文，作者，期刊和关键词的引文图。通过上下文自动过滤论文可以提高生产力。例如，给定一篇论文作为查询，哪些引文解决了相同的核心问题？ 这些只是背景引用？ 每个案例都是一个上下文或语义类，我们可以根据关键字和期刊等纸质属性区分它们。

我们称之为搜索所需的语义类w.r.t.查询节点作为语义邻近搜索。从某种意义上讲，这是一个新的问题，以前对邻近搜索的研究[2]，[1]，[3]，[4]，[5]既不打算明确区分语义类，也不能有效地实现。除了邻近搜索之外，我们最接近的问题是社交圈学习[6]和图表上的关系剖析[7]。在语义接近方面，他们的圈子或关系也是语义导向的，但他们只能以无人监督的方式找到潜在的聚类，因此不会针对特定的兴趣类别。在搜索方面，他们需要进行冗长的优化才能获得全局配置，因此无法实时处理即席查询。

虽然最终目标是处理任意语义类，但作为一个开始，我们只考虑本文中的对称类。也就是说，对于任何一对节点x和y，x与y的接近度总是与y与x的接近度相同。

## A.Insights和挑战

为了区分各种语义类，它给我们提出了一个自然的问题：什么类型的表示或结构可以表征一个类。理想情况下，我们需要一个不仅能够捕获不同语义的通用语言，而且还能够有效地在线处理以实现实时搜索。我们依赖于新颖的洞察力，即不同的语义类通常可以通过不同的说法共同的“结构”来表征。例如，在图1中，凯特和杰伊（同学）之间的接近可归因于他们的共同学校和专业如图2（a）中的结构（M1）所示。同样地，图2（b）和（c）示出了一些可能的结构，这些结构可以分别在不同程度上表征亲密朋友（M2和M3）和家庭（M4）。我们将这些常见的结构命名为meta-graph（在Sect.II中正式定义），因为它们将对象抽象为类型。也就是说，元图中的每个节点（由圆角矩形表示）描述对象的类型，而不是对象本身。直观地，两个节点“共享”类的更多特征元图更有可能满足该类接近度。除了捕获语义类之外，元图还可以实现在线邻近搜索。通过离线计算和索引元图，我们可以通过查找预先计算的元图来有效地支持任何查询。

注意，**已经提出了一种称为元路径**的较不通用的概念[4]，其仅考虑两个节点之间的公共路径结构。实际上，图2（b）中的元图M3也是一个元路径，它只捕获两个用户之间的公共地址。相比之下，**元图可以共同模拟多个共同属性**。考虑两个用户 - 雇主 - 用户和用户 - 爱好 - 用户的元路径。他们每个人都无法表达亲密朋友的亲近距离。然而，通过联合起来，我们获得了元图M2，它可以更好地描述亲密的朋友。换句话说，每个元图都是元路径的非线性组合，因此更具表现力。鉴于元图与元路径相比复杂性增加，使用和处理元图也更具挑战性，我们将在下面讨论。

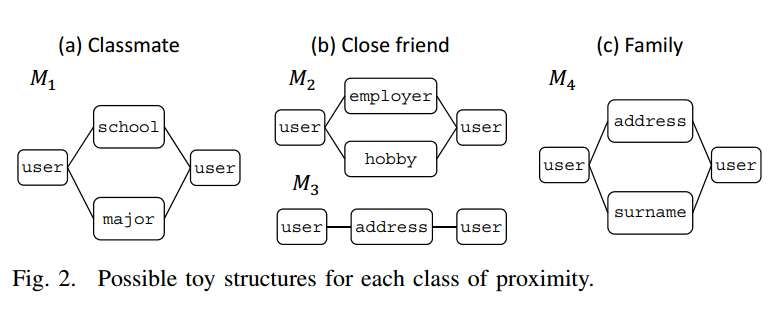
### 接近学习

首先，任意一类邻近的特征元图通常是未知的。虽然领域专家可以为某些特殊课程提供一些指导，但仅依靠人类智慧是不切实际的。此外，类可以由不同程度的多个元图表征。例如，亲密朋友可以是具有相同爱好的同事或仅仅是室友，分别对应于图2（b）中的M2和M3。也可以说M2更可能表示比M3更亲密的朋友。

为了更好地推广到不同的图和语义类，我们提出了一种监督方法，其基于一些示例查询和应答节点（例如，Kate作为查询，Alice和Jay作为答案）自动识别特征元图（例如，M2和M3）。在实践中，我们学习每个元图的权重，以量化它对所需类的特征。可以应用这些权重来回答相同类别邻近的未来查询。

为了实现邻近学习，大量的元图会产生进一步的挑战，特别是在大图上[8]，[9]。结果，处理大量元图会变得非常昂贵，特别是在大图上，更不用说基于这些元图学习了。

幸运的是，我们观察到并非所有元图对于给定的邻近类都是有用的。也就是说，许多元图没有表征，因此与期望的类无关（例如，**M2-M4对同学不是非常有用**）。因此，在这些不相关的元图上花费时间和计算资源是低效的。我们**提出了一种新的双阶段训练范式，它允许我们识别出最有希望的少数候选元图。**因此，我们可以丢弃所有其他元图并节省大量时间。



### 元图匹配

其次，有必要计算每个元图的实例，以便知道任何两个给定节点“共享”哪些元图。（实例和相关概念将在Sect.II中定义。）但是，计算元图的实例（也称为匹配元图）是非常昂贵的。它相当于解决NP-hard子图匹配问题[10]，[11]。此外，输入图上元图的实例数不符合向下闭包的特性，这排除了频繁子图挖掘的技术[12]，[8]，[9]。

为了更有效地计算元图的实例，我们提出了一种新的子图匹配算法，该算法利用了元图的对称性。我们观察到许多有用的元图是对称的，如图2中的M1-M4。（我们将在Sect.II中进一步解释。）但是，现有方法在元图内的对称子结构上花费了大量的冗余计算。因此，我们提出了一种基于对称的匹配算法，该算法重新使用“对称”计算。因此，我们可以避免冗余并大幅提高元图匹配的效率。

## B. Contributions

基于上述见解，我们采用监督方法实现元图的语义邻近搜索。总而言之，我们做出以下贡献。

•问题。

我们确定了语义邻近搜索的新问题，并提出了元图的新概念来表示不同的语义类。（Sect.II）

•学习模式。

我们首先提出了一个基于元图的接近度系列，它可以量化不同的语义类。我们进一步开发了一个监督模型，以学习特征元图，以获得家庭中接近的正确“形式”。最后，我们介绍了一种新的双阶段训练范式，以避免处理绝大多数没有希望的元图。（Sect.III）

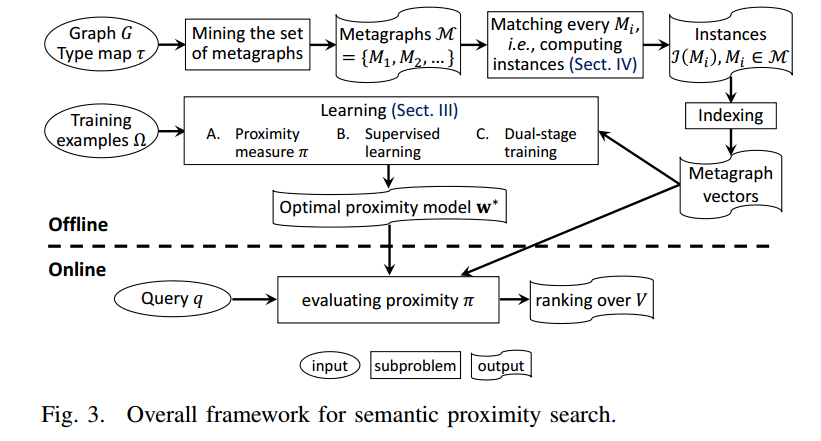
•匹配算法。

我们设计了一种有效的基于对称的元图匹配算法，该算法利用元图的对称性来避免冗余计算。（Sect.IV）

•实验。

我们在两个真实图表上的实验展示了我们方法的优越性。为了准确起见，我们在NDCG中的基线明显优于基线11％，在MAP中明显优于16％。为了提高效率，双阶段培训可以将元数据处理的总体成本降低83％，而基于对称的匹配可以将单个元数据的匹配时间缩短52％。（Sect.V）

# II. PRELIMINARIES AND OVERALL FRAMEWORK



## 离线阶段

它由三个主要的子问题组成。

1）**挖掘元图**，给定图G和类型映射函数τ，我们枚举元图集M.关于这个子问题有大量的文献[8]，[9]，与整个离线阶段相比，它的时间成本微不足道。因此，我们只是应用现有的最先进的方法GRAMI [9]。

2）**计算实例**，对于来自子问题1的输出的每个元图Mi 2 M，我们计算实例集I（Mi），以便我们可以进一步导出元图向量。**我们还将计算实例I（Mi）的过程称为匹配元图Mi.**我们在Sect IV中研究了这个有效元图匹配的子问题。（我们可以进一步索引实例，即根据方程1-2预先计算元图向量。元图向量是学习子问题输入的一部分，也是在线查询处理的一部分。）

3）最后，给定元数据向量和一组用于所需接近度的训练样例，我们需要学习最佳接近模型w \*。我们用Sect.III中的元图解决了这个学习的子问题。特别是，我们需要开发一系列基于元图的接近π来容纳任意类，一种监督方法，从训练样例Ω中学习家庭中的最佳π形式，以及减少双阶段训练方法 需要的元图数量。

## 在线阶段

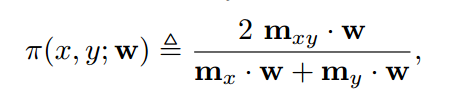
给定查询节点q预先计算的元图矢量，以及期望类的最佳模型w \*，我们可以计算q与其他节点v之间的接近度π。随后，我们按照π的降序对V中的节点进行排序。

# III. LEARNING PROXIMITY WITH METAGRAPHS

在本节中，我们建议基于元图学习最佳接近度。我们首先定义一系列邻近度量，它们可以灵活地满足不同的语义类。接下来，给出一些期望类的训练样例，我们开发了一种监督方法来选择近似族中的最优模型。最后，我们提出了一种新的双阶段训练范式，它不需要我们计算所有元图的实例。

## Metagraph-based Proximity

给定具有某些特征元图的类，良好的邻近度量必须考虑两个方面。首先，如果x和y共享许多特征元图，则x和y更可能满足所需的类。其次，如果x（或y）不加区分地出现许多元图，则x和y可能只是偶然出现以共享许多特征元图。结合这两个方面，我们提出了基于元图的接近度量π如下：MGP



MGP需要一系列具有参数w的邻近度量。我们将w解释为元图的特征权重向量，其可以变化以迎合不同类别的邻近度。考虑图2中的玩具示例。一个好的w可以是（0.9; 0; 0; 0）T 同学，（0; 0.6; 0.4; 0）T为亲密朋友，和（0; 0; 0; 0.8） T为家庭。因此，在接近族中，类的最优模型完全由其最佳权重向量w \*指定，我们的目标是自动学习。

有趣的是，MGP满足一些理想的性质，如定理1中所总结的。其中，对称性，self-maximum和尺度不变性可以很容易地从等式3得出。部分传递性意味着如果一个节点x接近y和z，则y和z也趋于彼此接近。这是社交网络上的普遍现象，其中朋友的朋友比随机的人更可能是朋友。由于空间限制，我们省略了它的证明。

## Supervised Learning

对于期望的邻近类别，我们假设一些训练示例Ω作为监督。就像成对学习排名[13]一样，每个例子都是三元组（q; x; y），使得节点x在节点y w.r.t.query节点q之前排序。也就是说，x与q的接近度应大于y。这些社交网络的例子通常可以通过用户研究来收集[6]，[7]。像Facebook和Google+这样的平台也允许用户直接标记他们的连接类。

### 目标函数

给定训练样例Ω，我们可以通过最大化对数似然来找到最佳权重w \*。直观地，当x与q的接近度越来越大于y时，它更可能观察到一个例子（q; x; y）。换句话说，例子的概率P（q; x; y; w）倾向于随着x和y与q，π（q; x; w） - π（q; y;W）。特别是，我们在方程4中定义概率。我们采用sigmoid函数将邻近差异转换为概率值，这是一种常见的做法[13]。注意，μ2（0; 1）是用于控制分布形状的缩放变量。

## Dual-Stage Training

如Sect.I所示，即使只有几种类型的对象，也存在大量的元图M.虽然计算单个元图的实例已经非常简单（将在Sect.IV中解决），但计算所有元图的实例变得过于苛刻。为了缩短整体匹配时间，我们提出了一种新的双阶段训练范式。与之前的成本感知学习范式[14]，[15]，[16]，[17]不同，其目的是降低在线测试期间的特征提取成本，我们的目标是减少离线的特征提取（即元图匹配）成本 训练。正如我们将在实验中看到的，元图匹配在整个离线阶段占主导地位，而如果元图预先匹配离线，在线测试可以非常快。

### Seed metagraphs

如果不计算任何元图的实例，就没有任何线索可以找到有希望的候选人。但是，一次性从M中选择所有候选K是不可行的。相反，我们首先将少量种子元图K0识别为初始候选者并计算它们的实例I（M），这可以引导我们更多的候选者。种子必须符合以下标准。

# VI. RELATED WORK

## Meta-structures

虽然已经提出了不那么普遍的元路径概念[4]，但metapaths在捕获节点之间的交互方面比元路径更具表现力和有效性，如Sect.I中所述。鉴于metapaths的复杂性和多样性的增加，我们无法以与元路径相同的方式处理元图。首先，基于元路径的PathSim [4]依赖于手动选择有用的元路径。考虑到更多的metapaths和任意类别的接近，这变得困难。因此，我们提出了一种有监督的学习方法。虽然另一项工作[24]也采用了元路径学习，但它只是针对不同的聚类任务而设计的。其次，元图比元路径更难以匹配。因此，我们开发了双阶段训练和基于对称的匹配，以提高效率。