# 不确定图上的高效 top-$k$ 近邻查询处理算法

张海杰 姜守旭 邹兆年
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 图的不确定性普遍存在,研究不确定图的高效查询处理具有重要意义.文中提出了不确定图上一种新型查询——近邻查询.给定一个查询标签集 $R$ 和距离约束 $\sigma$,在不确定图 $G$ 上进行近邻查询是要找到标签集包含 $R$ 并且任意两个顶点间距离不超过 $\sigma$ 的匹配顶点集.为解决该问题,文中首先提出了“可靠期望距离”,然后基于可靠期望距离建立了高效的近邻关系图索引,将不确定图上的近邻查询等价地转化为近邻关系图上的团查询问题,最后使用树搜索算法解决近邻关系图上的团查询问题.理论分析和实验结果表明文中提出的算法能够高效地完成不确定图上的 top-$k$ 近邻查询.

关键词 不确定图;近邻查询;可靠期望距离;近邻关系图
中图法分类号 TP311 **DOI** 号: 10.3724/SP.J.1016.2011.01885

## An Efficient Algorithm for top-$k$ Proximity Query on Uncertain Graphs

ZHANG Hai-Jie JIANG Shou-Xu ZOU Zhao-Nian
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

**Abstract** The uncertainty of graph data is widely existed in practice, researching on efficient query processing algorithms on uncertain graphs has significant meanings. This paper propose a new kind of query on uncertain graphs——proximity query. Given a query label set $R$ and a distance constrain $\sigma$, executing proximity query on an uncertain graph $G$ is to find some vertex subsets of $G$ that whose label set contains $R$ and for any two vertices in it, the distance between them can not exceed $\sigma$. To solve this issue, first, we propose a distance measure function named "Reliable Expectation Distance", and then design an efficient proximity relations graphs index and the proximity query on uncertain graphs is equally converted to the clique finding on proximity relations graphs. Finally, we propose a tree-based searching algorithm to finish the query processing. Theoretical and experimental results show that the proposed algorithm can efficiently retrieve the top-$k$ proximity query results.

**Keywords** uncertain graph; proximity query; reliable expectation distance; proximity relations graphs

## 1 引 言

近年来,现实世界中出现大量以图建模的数据,如化合物分子结构、移动自组织网络拓扑结构、社交网络等.并且由于数据的异构性、隐私保护、数据不完整、数据不精确等原因,图数据普遍存在不确定性.例如,在生物信息学领域,蛋白质交互(PPI)网

收稿日期: 2011-07-18;最终修改稿收到日期: 2011-08-19.本课题得到国家自然科学基金项目(61173023)以及中央高校基本科研业务费专项资金(HIT.NSRIF.201180)资助.张海杰,女,1986 年生,硕士研究生,主要研究方向为不确定图数据管理. E-mail: haijie0919@163.com.姜守旭,男,1968 年生,博士,教授,博士生导师,主要研究领域为大规模动态网络环境下的数据管理、对等计算与信任管理、数据库、无线传感器网络、容迟网络等.邹兆年,男,1979 年生,博士,讲师,主要研究方向为图数据挖掘.

络被表示成一个无向图,其中顶点代表蛋白质,边代表蛋白质之间的交互.由于蛋白质交互检测实验技术自身存在固有误差,实验测得的蛋白质交互网络存在不确定性[1].在移动自组织网络中,网络的拓扑结构被抽象成一个有向图,顶点代表通信节点,边代表节点之间的无线通信链路.由于节点的移动性,节点之间的无线通信链路是否正常工作是不确定的,因此移动自组织网络拓扑结构也存在不确定性[2].同样,在社交网络研究中,经常用图模型刻画用户之间的数据流通,但是匿名通信数据的干扰也会使社交网络带有不确定性[3].

传统图论中的图模型无法刻画图数据的不确定性.为此,我们拓展了不确定数据管理中常用的"可能世界"语义,提出了不确定图的"可能世界"语义模型[4].在该模型中,不确定图的每条边上带有一个(0,1]内的实数,表示该边实际存在的概率.该模型假定边的存在概率之间相互独立.依据边的存在概率对边进行独立随机选取,可以得到该不确定图的一种可能的确定图存在形式(即"可能世界"),称作"蕴含子图".理论证明,一个不确定图表达了其全部蕴含子图上的一个概率分布.

由于不确定图广泛存在且规模不断增加,因此对不确定图进行高效的查询处理具有非常重要的意义.现有的不确定图查询处理方面的研究几乎全部基于文献[4]中提出的不确定图语义模型,然而这方面研究结果目前还很少,仅限于概率$k$-近邻查询[5]、概率最短路径查询[6]、概率子图匹配查询[7].

本文研究不确定图上一种新型查询"近邻查询(proximity query)",其定义如下:给定一个不确定图$G$,$G$的每个顶点$v$上带有一个标签集合$L(v)$;同时,还给定一种$G$中顶点之间的概率距离度量函数$d$.不确定图$G$上的近邻查询是一个二元组$Q=\langle R,\sigma\rangle$,其中$R$是一个顶点标签集合,$\sigma\geq 0$.$Q$的查询结果是$G$中的"匹配顶点子集"构成的集族$F$,其中每个匹配顶点子集$W\in F$满足如下3个条件:

(1) $R\subseteq\bigcup_{v\in W}L(v)$;

(2) $W$的任意真子集均不满足条件1;

(3) $W$中任意两个顶点$u$和$v$在$G$中概率距离$d(u,v)$的最大值(记作$diam(W)$)不大于$\sigma$.

特别地,本文研究不确定图上的top-$k$近邻查询,即查询结果是近邻查询结果中前$k(k>0)$个具有最小$diam(W)$值的顶点子集$W$.

不确定图上的top-$k$近邻查询具有许多重要应用.例如,在生物信息学中,生物学家经常需要了解在蛋白质交互网络中具有某些功能的蛋白质在不同功能团中是如何连接的.我们已经知道蛋白质交互网络是一个不确定图(在该不确定图中,顶点标签代表蛋白质具有的功能),并且在该网络中,具有某些功能的蛋白质在不同功能团中的连结方式往往是不同的.如图1所示,具有$a$,$b$,$c$ 3种功能的蛋白质频繁地同时出现在不同的功能团中,然而其具体连接方式并不相同,如图1中虚线所示.在这种情况下,生物学家若想了解具有功能$a$,$b$,$c$的蛋白质之间是否存在紧密关联,传统子图匹配查询[8-9]显然是做不到的,因为生物学家必须先知道具有功能$a$,$b$,$c$的蛋白质之间的连接方式才能形式化给出子图匹配查询;而不确定图近邻查询并不要求知道具有功能$a$,$b$,$c$的蛋白质之间的具体连接方式,只需给出蛋白质之间的概率距离约束条件,因此更适用于生物学家的需求.

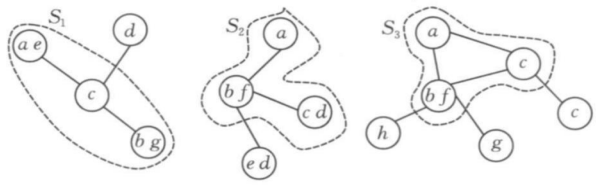


图1 蛋白质交互网络中的近邻查询

在社交网络分析中,Lappas等人[10]提出了社交网络中的团队构建问题.在该问题中,社交网络以确定图建模,图的顶点代表候选专家,顶点标签代表专家所掌握的领域知识,边代表专家之间的朋友关系,边上的权值代表专家之间的沟通代价(或熟悉程度).给定一个社交网络关系图$G$和一个任务需求$R$,团队构建问题是要组建一个专家团队,使得团队中的专家所掌握的领域知识能满足任务要求$R$,并且专家之间的沟通代价最小.由于社交网络中存在匿名通信数据,专家之间的朋友关系可能是不确定的,因此该社交网络需要用不确定图建模.在带有不确定性的社交网络上解决团队构建问题就需要应用不确定图近邻查询处理算法.

据我们所知,不确定图上的top-$k$近邻查询处理尚未被研究过,并且确定图上的相关研究结果也无法处理不确定图近邻查询.在文献[10]提出的解决团队构建问题的算法中,顶点间的距离采用加权最短路径长度;而在不确定图中,两个顶点在不同的蕴含子图中的加权最短路径长度不尽相同.另外,这些算法只返回代价最小的结果;而top-$k$近邻查询要求找到直径最小的$k$个结果.

另外,不确定图上的子图匹配查询算法无法处

理近邻查询问题, 因为子图匹配查询是基于结构的查询, 需要事先知道匹配顶点子集的具体结构, 而近邻查询不限制匹配顶点子集的结构, 其查询结果具有结构多样性.

综合上述分析, 不确定图 top-$k$ 近邻查询向我们提出了如下挑战问题:

(1) 不确定图上的近邻查询基于不确定图顶点之间的概率距离度量函数, 对该函数既要考虑不确定图的结构也考虑不确定图的不确定性. 由文献[4], 由于一个不确定图有多种可能的存在形式(即蕴含子图), 相同的两个顶点在不同的蕴含子图中的距离可能不同, 所以如何度量不确定图顶点间的距离是一项挑战;

(2) 与传统的子图匹配查询不同, 近邻查询的查询结果具有结构多样性. 目前图数据上的查询多数是基于结构的查询, 并且不确定图本身的结构存在不确定性, 因此, 如何设计高效的近邻查询处理算法来解决这种结构多样性查询也是一项挑战.

为解决上述问题, 本文提出了不确定图上的可靠期望距离度量函数, 并基于可靠期望距离定义了近邻关系图. 本文证明了在不确定图上进行近邻查询等价于在对应的近邻关系图上进行团查询. 本文的主要贡献如下:

(1) 提出不确定图顶点间的可靠期望距离度量; 基于不确定图数据模型和可靠期望距离, 首次提出不确定图的 top-$k$ 近邻查询问题, 并形式化地定义了该问题;

(2) 基于可靠期望距离定义 $\sigma$-近邻关系图并设计高效的 $\sigma$-近邻关系图索引结构及其快速构建算法.

(3) 提出了一种高效的基于搜索树的算法来完成近邻查询处理, 其中利用一种两阶段预处理方法来降低搜索时间复杂度, 并利用分支界限方法高效地计算 top-$k$ 结果.

## 2　相关工作

子图匹配查询[8-9]是图查询领域的研究热点. 给定一个标签图 $G$, 一个连通的包含 $n$ 个顶点的查询图 $Q$, 子图匹配查询是从 $G$ 中找到 $n$ 个顶点, 这 $n$ 个顶点构成的子图同构于查询图 $Q$. 显然, 子图匹配查询是一种基于结构的查询.

模式匹配查询[11-12]是图查询领域另一种重要查询. 给定一个标签图 $G$, 一个连通的包含 $n$ 个顶点的查询图 $Q$ 以及距离约束 $\sigma$, 模式匹配查询是从 $G$ 中找到 $n$ 个顶点, 这 $n$ 个顶点与 $Q$ 的 $n$ 个顶点一一对应, 并且对应顶点的标签相同, 对于 $Q$ 中相邻的两个顶点, 其对应顶点在 $G$ 中距离不大于 $\sigma$.

在本文提到的近邻查询中, 查询 $Q$ 包括顶点标签集 $R$ 和距离约束 $\sigma$. 查询结果是图 $G$ 的顶点子集, 这个顶点子集的标签集包含 $R$, 并且其任意两个顶点间的距离不超过 $\sigma$. 因此, 近邻查询是一种基于距离的查询.

子图匹配查询和近邻查询是两种类型的查询, 利用子图匹配查询算法不能解决近邻查询问题. 近邻查询是一种特殊情况下的模式匹配查询. 当模式匹配查询的查询 $Q$ 是一个团时, 该模式匹配查询就等价于以 $Q$ 的顶点标签集作为输入的近邻查询. 因此, 可以使用模式匹配查询算法来解决近邻查询问题. 但是据我们所知, 目前尚没有不确定图上的模式匹配查询方面的工作, 而且确定图上的相关算法也无法直接应用到不确定图上, 因此有必要提出高效的不确定图上近邻查询处理算法.

## 3　问题定义

### 3.1　不确定图

本文沿用文献[4]提出的不确定图语义模型. 为使本文内容完整, 这里再次给出不确定图的相关定义.

定义 **1**(不确定图).　不确定图是一个五元组 $G=((V,E),\Sigma,L,P)$, 其中 $(V,E)$ 是一个无向图, $\Sigma$ 是一个标签集合, $L: V\rightarrow 2^{\Sigma}$ 是一个为顶点分配标签子集的函数. $P: E\rightarrow(0,1]$ 是一个为边分配存在概率值的函数, $P((u,v))$ 表示顶点 $u$ 和顶点 $v$ 之间的边存在概率. 特别地, $P((u,v))=1$ 表示边 $(u,v)$ 一定存在.

一个不确定图 $G$ 具有多种可能的存在形式, 每个存在形式被称为 $G$ 的一个蕴含子图. 形式化地说, 一个确定图 $G'=((V',E'),\Sigma',L')$ 是 $G=((V,E),\Sigma,L,P)$ 的一个蕴含子图当且仅当 $V'=V$, $E'\subseteq E$, $\Sigma'=\Sigma$, $L'=L$. 不确定图 $G$ 的所有蕴含子图构成的集合记为 $Imp(G)$. 本文假设不确定图中边的存在概率相互独立, 因此蕴含子图 $G'$ 被不确定图 $G$ 蕴涵的概率为

$$P(G\Rightarrow G')=\prod_{e\in E(G')}P(e)\prod_{e'\in E(G)\setminus E(G')}(1-P(e')) \quad (1)$$

文献[4]证明了 $G$ 的全部蕴含子图被 $G$ 蕴涵的概率和为 1, 即 $\sum_{G'\in Imp(G)}Pr(G\Rightarrow G')=1$.

图 2(a)是一个不确定图 $G$, 图 2(b)是 $G$ 的蕴含子图集 $Imp(G)$. 根据式(1)可以计算出 $G$ 的每个蕴含子图被 $G$ 蕴含的概率. 例如, $G_7$ 被 $G$ 蕴含的概率为 0.9×0.7×(1−0.4)=0.378.

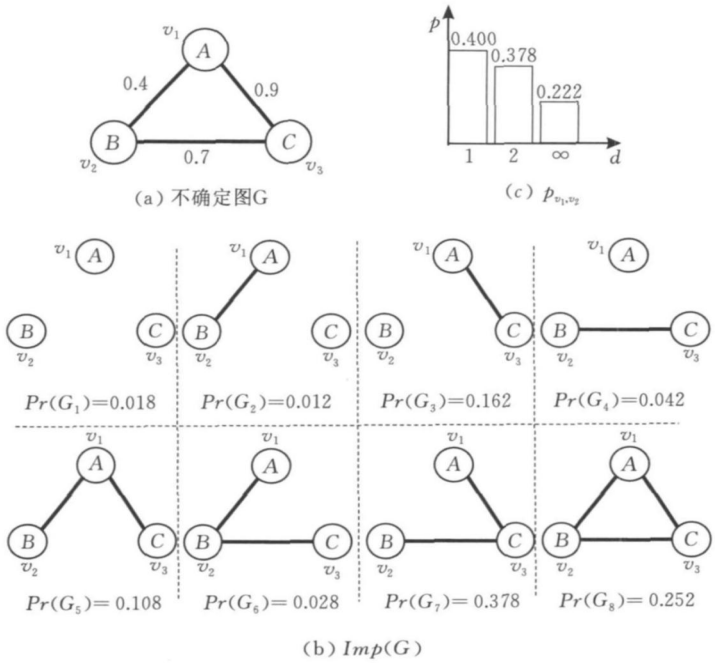


图 2 不确定图 $G$、$Imp(G)$ 以及顶点 $v_1$, $v_2$ 间距离的概率分布 $p_{v_1,v_2}$

### 3.2 不确定图顶点距离度量函数

在确定图中, 两个顶点间的距离用这两个顶点之间的最短路径长度表示. 然而, 在不确定图中, 两个顶点在不同蕴含子图中的最短路径长度不尽相同. 因此, 在定义不确定图上近邻查询之前, 需要先定义不确定图顶点距离度量函数. 本文基于文献[5]提出的不确定图顶点间的众数距离和期望可达距离, 提出一种新的不确定图顶点距离度量函数——可靠期望距离.

给定一个不确定图 $G$, $G$ 中顶点 $u$ 和 $v$ 距离等于 $d$ 的概率为

$$p_{u,v}(d)=\sum_{G'\in Imp(G),\, d_{G'}(u,v)=d} P(G\Rightarrow G') \qquad (2)$$

其中, $d_{G'}(u,v)$ 表示蕴含子图 $G'$ 中, 顶点 $u$, $v$ 之间的最短路径长度.

图 2(c)表示图 2(a)中 $G$ 的顶点 $v_1$ 和 $v_2$ 之间距离的概率分布. 由此, 文献[4]定义了如下两种距离度量函数:

(1) 众数距离. 不确定图 $G$ 中任意两个顶点 $u$ 和 $v$ 间的众数距离为 $d_J(u,v)=\arg\max_d\{p_{u,v}(d)\}$, 即 $u$ 与 $v$ 之间概率最大的距离值.

(2) 期望可达距离. 不确定图 $G$ 中任意两个顶点 $u$ 和 $v$ 之间的期望可达距离为

$$d_{ER}(u,v)=\sum_{d|d<\infty} d\cdot\frac{p_{u,v}(d)}{1-p_{u,v}(\infty)},$$

即 $u$ 与 $v$ 之间距离的期望值(需要注意的是, 在计算期望可达距离时, 不考虑 $u$ 与 $v$ 之间距离为无穷的情况).

然而, 上述 2 种距离度量都有其局限性:

(1) 当顶点 $u$ 和 $v$ 在不确定图 $G$ 中的距离分布较平均时, 众数距离度量不具有一般性. 例如, 在顶点 $u$, $v$ 之间距离的概率分布中, $p_{u,v}(1)=p_{u,v}(2)=0.33$, $p_{u,v}(\infty)=0.34$, 此时 $u$ 和 $v$ 之间的众数距离为 $\infty$. 而实际上 $u$ 和 $v$ 之间的距离为无穷的概率只有 0.34, 为有穷的概率为 0.66;

(2) 为保证期望可达距离有穷, 计算时将 $d=\infty$ 的部分排除. 所以, 只要 $p_{u,v}(\infty)\neq 1$, $u$ 和 $v$ 之间的距离就始终有穷. 例如, 在顶点 $u$, $v$ 之间距离的概率分布中, $p_{u,v}(1)=0.05$, $p_{u,v}(\infty)=0.95$, 此时 $u$, $v$ 之间的期望可达距离为 1. 而实际上两个顶点之间距离有穷的概率仅为 0.05, 距离为无穷的概率是 0.95.

本文对上述 2 种不确定图顶点距离度量函数的不足进行修正, 提出可靠期望距离度量函数. 其定义如下: 不确定图 $G$ 中任意两个顶点 $u$ 和 $v$ 之间的可靠期望距离为

$$d_{RE}(u,v)=\begin{cases} d_{ER}(u,v), & \text{若 } p_{u,v}(\infty)<0.5 \\ \infty, & \text{否则} \end{cases} \qquad (3)$$

其含义是: 当 $p_{u,v}(\infty)<0.5$ 时, 期望可达距离

是一种非常可靠的距离度量, 于是采用期望可达距离来度量顶点间距离; 当 $p_{u,v}(\infty) \geq 0.5$ 时, 期望可达距离变得不可靠, 由于此时 $\infty$ 在所有可能距离值中概率最大, 根据众数距离定义, $u$, $v$ 之间的距离为 $\infty$.

### 3.3　近邻查询

定义 **2**(顶点集直径).　给定一个不确定图 $G=((V,E),\Sigma,L,P)$ 和 $G$ 的一个顶点子集 $W \subseteq V$, $W$ 在 $G$ 中的直径为 $diam(W)=\arg\max_{u,v\in W}\{d_{RE}(u,v)\}$.

由此, 我们给出不确定图近邻查询的形式化定义.

定义 **3**(不确定图上的近邻查询).　给定一个不确定图 $G=((V,E),\Sigma,L,P)$, $G$ 上的近邻查询是一个二元组 $Q=\langle R,\sigma\rangle$, 其中 $R\subseteq\Sigma$, $\sigma$ 为非负整数. $Q$ 的每个查询结果是 $G$ 的一个顶点子集 $W\subseteq V$, $W$ 满足如下性质:

(1) 标签集覆盖性.　$R\subseteq\bigcup_{v\in W}L(v)$;

(2) 极小性.　$W$ 的任意真子集都不满足性质 1;

(3) 紧密性.　$\mathrm{diam}(W)\leq\sigma$.

本文将 $Q$ 的每个查询结果称作 $Q$ 的匹配顶点子集.

对于用户给定的近邻查询, 满足条件的匹配顶点子集的数量可能非常大. 在这种情况下, 选择直径最小的前 $k$ 个匹配顶点子集便可满足应用需要.

由此, 不确定图上的 top-$k$ 近邻查询问题的定义如下:

输入: 不确定图 $G=((V,E),\Sigma,L,P)$, 查询 $Q=\langle R,\sigma,k\rangle$, 其中 $R\subseteq\Sigma$, $\sigma$ 为非负整数, $k$ 为正整数.

输出: $G$ 中直径最小的 $k$ 个匹配顶点子集. 若满足条件的匹配顶点子集的数量少于 $k$ 个, 则将其全部输出.

## 4　近邻查询处理

本节提出了一种高效处理不确定图上 top-$k$ 近邻查询的算法. 给定一个不确定图 $G=((V,E),\Sigma,L,P)$ 和一个 top-$k$ 近邻查询 $Q=\langle R,\sigma,k\rangle$, 算法包括以下 4 个步骤:

1. 计算 $G$ 中任意两个顶点 $u$ 和 $v$ 之间的可靠期望距离 $d_{RE}(u,v)$. 根据可靠期望距离定义可知, 为计算 $d_{RE}(u,v)$, 需先计算 $u$ 和 $v$ 之间距离的概率分布. 解决该问题的精确方法是穷举 $G$ 的所有蕴含子图, 然后在每个蕴含子图上计算 $u$ 和 $v$ 之间最短路径长度的概率分布. 然而, $G$ 的全部蕴含子图的数量为 $2^{|E|}$, 其中 $|E|$ 是 $G$ 的边数, 穷举方法效率显然非常低. 本文使用采样方法来近似计算 $u$ 和 $v$ 之间最短路径长度的概率分布, 在保证计算结果精度的前提下显著提高计算效率. 具体方法见 4.2.1 节.

2. 根据顶点间可靠期望距离以及距离约束 $\sigma$, 构建 $\alpha$-近邻关系图 $G^{\alpha}$. $G^{\alpha}$ 是一个确定图, 其构建方法如下: $G^{\alpha}$ 的顶点集和 $G$ 的顶点集相同, 并且对于 $G$ 中任意两个顶点 $u$ 和 $v$, 若 $d_{RE}(u,v)\leq\sigma$, 则在 $G^{\alpha}$ 中存在一条连接 $u$ 和 $v$ 的边. 本文 4.1 节将给出近邻关系图的形式化定义. 后文将证明 $G$ 中任意顶点子集 $W$ 的直径小于等于 $\sigma$ 当且仅当 $W$ 在 $G^{\alpha}$ 是一个团.

3. 根据查询 $Q$ 给出的标签集 $R$ 对 $G^{\alpha}$ 进行预处理. 通过分析可知, $G^{\alpha}$ 中可能出现在结果中的顶点 $v$ 具有如下性质:

3.1. $v$ 的标签集 $L(v)$ 与 $R$ 的交集不为空.

3.2. $G^{\alpha}$ 中与 $v$ 相邻的全部顶点的标签集与 $L(v)$ 的并集一定包含 $R$.

利用这两条性质, 对 $G^{\alpha}$ 中的顶点和边进行迭代删除, 直至将不可能出现在任何查询结果中的顶点及其相连的边全部删除. 记预处理后得到的图为 $G^{\delta}$.

4. 根据查询标签集 $R$, 在 $G^{\delta}$ 上找出 $k$ 个团, 使得每个团中所有顶点标签集的并集包含 $R$, 并且这些团在 $G$ 中的直径最小. 这 $k$ 个团就是不确定图 $G$ 近邻查询 $Q$ 的结果集. 本文采用树搜索策略来解决带标签覆盖的团查询问题, 并运用分支界限方法计算直径最小的 top-$k$ 结果.

对于两个查询 $Q_1=\langle R_1,\sigma_1,k_1\rangle$ 和 $Q_2=\langle R_2,\sigma_2,k_2\rangle$, 上述算法对 $G$ 中任意两个顶点计算得到相同的可靠期望距离; 并且, 若 $\sigma_1=\sigma_2$, 则上述算法为 $Q_1$ 和 $Q_2$ 计算得到相同的 $\alpha$-近邻关系图.

因此, 为避免重复计算, 对每个可能的距离 $\sigma$ 构造 $\alpha$-近邻关系图 $G^{\alpha}$, 并建立近邻关系图索引. 对于任意查询 $Q=\langle R,\sigma,k\rangle$, 算法可以直接在 $G^{\alpha}$ 执行查询处理.

本节余下内容将介绍索引结构、索引构造算法以及查询处理算法.

### 4.1　索引结构

定义 **4**($\alpha$-近邻关系图).　给定不确定图 $G=((V,E),\Sigma,L,P)$ 和非负整数 $\sigma$, $G$ 的 $\alpha$-近邻关系图 $G^{\alpha}=((V^{\alpha},E^{\alpha}),\Sigma^{\alpha},L^{\alpha},W^{\alpha})$ 是一个确定加权图, 其中 $V^{\alpha}=V$, $\Sigma^{\alpha}=\Sigma$, $L^{\alpha}=L$, $E^{\alpha}=\{(u,v)\mid u,v\in V^{\alpha}$ 且在 $G$ 中 $d_{RE}(u,v)\leq\sigma\}$, $W^{\alpha}$ 是为 $E^{\alpha}$ 中的边分配权值的函数, 对于任意 $(u,v)\in E^{\alpha}$, $W^{\alpha}(u,v)=d_{RE}(u,v)$.

$\alpha$-近邻关系图是不确定图 top-$k$ 近邻查询处理算法使用的重要数据结构. 我们对 $\alpha$-近邻关系图有以下 3 个重要观察:

(1) 对不确定图 $G$ 上任意两个查询 $Q_1=\langle R_1,\sigma_1,k_1\rangle$ 和 $Q_2=\langle R_2,\sigma_2,k_2\rangle$, 若 $\sigma_1=\sigma_2$, 则 $G^{\sigma_1}=G^{\sigma_2}$;

(2) $G^{\alpha}$ 中任意边 $(u,v)$ 上的权值 $W^{\alpha}(u,v)$ 仅与不确定图 $G$ 相关而与查询 $Q$ 无关, 因为 $W^{\alpha}(u,v)=$

$d_{RE}(u, v)$;

(3) 若 $\sigma_1 \le \sigma_2$, 则 $G^{\sigma_1} \subseteq G^{\sigma_2}$.

根据观察(1), 可对 $\sigma = 1, 2, 3, \ldots$, 预先计算出 $G^{\sigma}$, 并将全部 $G^{\sigma}$ 按照 $\sigma$ 进行索引. 当处理查询 $Q = \langle R, \sigma, k \rangle$ 时, 直接根据 $\sigma$ 快速取回 $G^{\sigma}$, 并在 $G^{\sigma}$ 上执行查询处理.

根据观察(2), 可将所有 $G^{\sigma}(\sigma = 1, 2, 3, \ldots)$ 边上的权值仅存储一次, 从而降低空间复杂性. 我们使用表 Graph 存储 $G^{\sigma}$ 中边的权值.

根据观察(3), 可将 Graph 表中边的权值按递增的顺序排序, $G^{\sigma}$ 用 Graph 表中边的权值不大于 $\sigma$ 的所有记录表示.

具体的索引结构如图 3 所示, 包括两个部分:

第 1 部分是 Index 表, 具有 2 列: 第 1 列是距离约束 $\sigma$; 第 2 列是指向 Graph 表记录的指针, Graph 表的第 1 条记录到该指针所指记录之间的所有记录表示 $G^{\sigma}$. 第 2 部分是 Graph 表, 具有 3 列: 第 1 列和第 2 列是 $G^{\sigma}$ 的顶点; 第 3 列是 $G^{\sigma}$ 中边的权值.

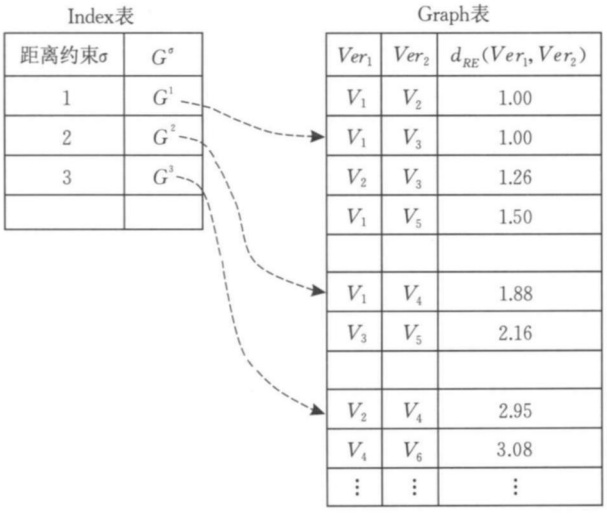


图 3 $\sigma$-近邻关系图索引结构

## 4.2 索引构造

### 4.2.1 索引构造第 1 阶段

$\sigma$-近邻关系图索引构建的第 1 阶段是计算不确定图 $G$ 中任意两个顶点间的可靠期望距离. 我们已经知道可靠期望距离 $d_{RE}(u, v)$ 的计算依赖 $G$ 上 $u$ 和 $v$ 之间距离的概率分布, 而通过穷举 $G$ 的全部蕴含子图的方法来计算 $u$ 和 $v$ 之间距离概率分布是极其低效的. 为了既保证计算结果精度又显著提高计算效率, 本文采用 Monte Carlo 方法近似计算可靠期望距离 $d_{RE}(u, v)$, 该算法的思想如下: 按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $G$ 的蕴含子图集合 $Imp(G)$ 进行 $r$ 次独立采样($r$ 的具体数值在后文给出), 得到 $r$ 个 $G$ 的蕴含子图 $G_1, G_2, \ldots, G_r$; 在每个采样得到的蕴含子图 $G_i$ 上, 计算任意两个顶点之间的最短路径长度; 然后, 对于 $G$ 中任意两个顶点 $u$ 和 $v$, 将 $G_1, G_2, \ldots, G_r$ 上 $u$ 和 $v$ 之间最短路径长度的分布作为 $G$ 上 $u$ 和 $v$ 之间距离概率分布的近似; 最后, 使用该近似距离分布, 计算近似可靠期望距离.

下面分析采样次数 $r$ 的取值.

引理 **1**. 对于一个不确定图 $G = ((V, E), \Sigma, L, P)$ 以及 $0 < \varepsilon, \delta < 1$, 设 $S = \{G_1, G_2, \ldots, G_r\}$ 是按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $G$ 的蕴含子图集合 $Imp(G)$ 进行 $r > 0$ 次独立采样得到的样本空间. 对 $G$ 的顶点 $u$ 和 $v$, 设

$$I_i = \begin{cases} 1, & \text{若 } u \text{ 和 } v \text{ 在抽样得到的蕴含子图 } G_i \text{ 中连通} \\ 0, & \text{否则} \end{cases} \tag{4}$$

得到, 若 $r \ge \frac{3}{\varepsilon^2 p(u, v)} \ln\left(\frac{2}{\delta}\right)$, 则

$$Pr\left(\left|\frac{1}{r}\sum I_i - p(u, v)\right| \ge \varepsilon p(u, v)\right) \le \delta,$$

其中, $p(u, v)$ 表示 $G$ 中 $u$ 和 $v$ 连通的概率.

证明. 引理 1 是 Chernoff Bound[13] 的简单应用. 其中 $I_1, I_2, \ldots, I_r$ 是独立同分布示性随机变量, 根据 Chernoff Bound 有

$$Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r} I_i - p(u, v)\right| \ge \varepsilon \cdot p(u, v)\right) \le 2\exp\left(-\frac{r \cdot p(u, v)\varepsilon^2}{3}\right).$$

若 $r$ 应满足 $r \ge \frac{3}{p(u, v)\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$, 有

$$Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r} I_i - p(u, v)\right| \ge \varepsilon \cdot p(u, v)\right) \le \delta.$$ 证毕.

引理 **2**. 对于一个不确定图 $G = ((V, E), \Sigma, L, P)$ 以及 $0 < \varepsilon, \delta < 1$, 设 $S = \{G_1, G_2, \ldots, G_r\}$ 是按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $G$ 的蕴含子图集合 $Imp(G)$ 进行 $r > 0$ 次独立采样得到的样本空间. 对 $G$ 的顶点 $u$ 和 $v$, 设

$$d_i = \begin{cases} u, v \text{ 在 } G_i \text{ 中最短路径长度}, & \text{若 } u, v \text{ 在 } G_i \text{ 中连通} \\ 0, & \text{否则} \end{cases} \tag{5}$$

得到, 若 $r \ge \frac{(n-1)^2}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$, 则

$$Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r} d_i - d_{ER}(u, v)\right| \ge \varepsilon\right) \le \delta,$$

其中 $n = |V|$.

证明. 引理 2 是 Hoeffding 不等式的简单应用. 根据式(5)可知 $Pr(d_i \in [0, n-1]) = 1$, 由 Hoef-

fding 不等式有

$$Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r}d_i - d_{ER}(u,v)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{r(n-1)^2}\right),$$

若 $r$ 满足 $r \geq \frac{(n-1)^2}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$, 则有

$$Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r}d_i - d_{ER}(u,v)\right| \geq \varepsilon\right) \leq \delta \quad \text{证毕.}$$

由引理 1 和 2, 我们得到如下定理.

定理 **1**.　对于不确定图 $G=((V,E),\Sigma,L,P)$ 以及 $0<\varepsilon,\delta<1$, 设 $S=\{G_1,G_2,\cdots,G_r\}$ 是按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $G$ 的蕴含子图集合 $Imp(G)$ 进行 $r \geq \max\left\{\frac{6}{\varepsilon^2}, \frac{(n-1)^2}{2\varepsilon^2}\right\}\ln\left(\frac{2}{\delta}\right)$ 次独立采样得到的样本空间. 对 $G$ 的顶点 $u$ 和 $v$, 设 $p(u,v)$ 表示 $G$ 中 $u$ 和 $v$ 连通的概率, $p'(u,v)=\frac{1}{r}\sum_{i=1}^{r}I_i$, 其中 $I_i$ 如式(4)定义; 又设

$$d'_{RE}(u,v)=\begin{cases}\infty, & \text{若} p'(u,v)<0.5\\ \frac{1}{r\cdot p'(u,v)}\left(\sum_{i=1}^{r}d_i\right), & \text{否则}\end{cases},$$

其中 $d_i$ 的定义见式(5).

我们有如下结论:

(1) 若 $p(u,v)>0.5+\varepsilon$, 则 $Pr[p'(u,v)<0.5]<\delta/2$, 进而 $Pr[d'_{RE}(u,v)<\infty]<\delta/2$;

(2) 若 $0.5<p(u,v)<0.5+\varepsilon$, 则 $Pr[p'(u,v)<0.5]<\delta/2+(1-\delta)/2\cdot(\varepsilon-p(u,v)+0.5)/\varepsilon$, 进而 $Pr[d'_{RE}(u,v)<\infty]<\delta/2+(1-\delta)/2\cdot(\varepsilon-p(u,v)+0.5)/\varepsilon$;

(3) 若 $p(u,v)<0.5-\varepsilon$, 则 $Pr[p'(u,v)>0.5]<\delta/2$, 进而 $Pr[|d'_{RE}(u,v)-d_{RE}(u,v)|>\varepsilon]<\delta(1-\delta/2)$;

(4) 若 $0.5-\varepsilon<p(u,v)<0.5$, 则 $Pr[p'(u,v)>0.5]<\delta/2+(1-\delta)/2\cdot(\varepsilon-0.5+p(u,v))/\varepsilon$, 进而 $Pr[|d'_{RE}(u,v)-d_{RE}(u,v)|>\varepsilon]<\delta(\delta/2+(1-\delta)/2\cdot(\varepsilon-0.5+p(u,v))/\varepsilon)$.

根据上述定理, 得到如下计算不确定图中任意两个顶点之间可靠期望距离的算法.

算法 **1**.　近似可靠期望距离计算算法.

输入: 不确定图 $G=((V,E),\Sigma,L,P)$, 初始精度参数 $(\varepsilon_0,\delta_0)$, 初始距离约束 $\sigma_0$

输出: 不确定图 $G$ 的任意两个顶点 $u,v$ 之间的近似可靠期望距离 $d'_{RE}(u,v)$

1. 按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $Imp(G)$ 进行 $r \geq \max\left\{\frac{6}{\varepsilon_0^2}, \frac{(n-1)^2}{2\varepsilon_0^2}\right\}\ln\left(\frac{2}{\delta_0}\right)$ 次独立采样得到样本空间 $S=\{G_1,G_2,\cdots,G_r\}$, 其中每个 $G_i$ 是通过对 $G$ 的每条边 $e$ 按照其存在概率值 $P(e)$ 进行独立随机选取得到的.

2. 在 $S$ 中每个蕴含子图 $G_i$ 上, 使用 Johnson 算法[14] 计算 $G_i$ 中任意两个顶点之间的最短路径长度.

3. 对 $G$ 中任意顶点 $u$ 和 $v$, 令 $p'(u,v)$ 表示样本空间 $S$ 中 $u$ 和 $v$ 连通的概率. 由此, 计算 $u$ 和 $v$ 之间的近似可靠期望距离为

$$d_{RE}(u,v)=\begin{cases}\infty, & \text{若} p'(u,v)<0.5\\ \frac{1}{r\cdot p'(u,v)}\left(\sum_{i=1}^{r}d_i\right), & \text{否则}\end{cases}.$$

算法 1 第 1 步采样每个蕴含子图的时间复杂性为 $O(|E|)$, 故第 1 步的时间复杂性为 $O(r|E|)$. 第 2 步使用 Johnson 算法计算每个采样得到的蕴含子图 $G_i$ 中任意两个顶点之间的最短路径长度的时间复杂度为 $O(|V|^2\log|E|+|V||E|)$, 故第 2 步的时间复杂性为 $O(r(|V|^2\log|E|+|V||E|))$. 第 3 步计算 $G$ 中任意顶点 $u$ 和 $v$ 之间的近似可靠期望距离的时间复杂性为 $O(r|V|^2)$. 综上所述, 算法 1 的时间复杂度为 $O(r(|V|^2\log|E|+|V||E|))$.

4.2.2　索引构建第 2 阶段

$\alpha$-近邻关系图索引包括 Graph 表和 Index 表. 这两个表的建立过程如下: 对于 $G^{\alpha}$ 中任意两个顶点 $u$ 和 $v$, 若 $W^{\alpha}(u,v)\leq\sigma_0$, 则向 Graph 表中插入记录 $(u,v,W^{\alpha}(u,v))$; Graph 表建完后, 将 Graph 表的记录按照边的权值从小到大的顺序排序, 然后顺序扫描一遍 Graph 表, 对于每个距离 $d(d=1,2,\cdots,\sigma_0)$, 当首次遇到权值大于 $d$ 的记录 $t$ 时, 向 Index 表中插入记录 $(d,addr(t-1))$. 其中记录 $t-1$ 是 Graph 表中记录 $t$ 的前一条记录, $addr(t-1)$ 是记录 $t-1$ 的地址. Graph 表中从开始到记录 $t-1$ 之间的所有记录构成的集合就表示 $G^d$.

4.2.3　索引构造算法优化

引理 2 给出采样次数 $r \geq \frac{D^2}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$, 其中 $D$ 为样本空间中, $d_i$ (见式(5)) 的取值范围宽度. 引理 2 中 $D=n-1$, 当 $n$ 较大时, $r$ 将非常大, 此时索引构建效率将非常低. 通过选取合适的 $D$ 来降低 $r$, 可提高索引构建的效率. 为此我们提出如下 3 种优化策略:

(1) $D$ 等于近似直径. Aingworth 等人[15] 研究了近似计算图直径的算法 Appro-Diameter. 令 $D$ 等于近似直径会减少采样次数, 但 Appro-Diameter 算法时间复杂度高, 因此该方法不会明显提高索引构建的效率.

(2) $D$ 等于 $c\sigma$, 其中 $c(c\geq1)$ 为距离约束 $\sigma$ 的松弛系数. 通过实验我们得知: 在多数情况下, 当顶

点 $u$ 和 $v$ 在样本 $G_i$ 中的最短路径长度 $d_i$ 超过 $c\sigma$ 时，把 $d_i$ 当作无穷处理不会影响查询结果. 因为若 $d_{RE}(u,v)>\sigma$，只有当将超过 $c\sigma$ 部分当作无穷处理得到的近似可靠期望距离小于 $\sigma$ 时才会影响查询；若 $d_{RE}(u,v)\leq\sigma$，只有当将超过 $c\sigma$ 的部分当作无穷处理得到的近似可靠期望距离等于无穷时，才会影响查询. 大量实验表明，上述两种情况发生的概率非常低. 根据上述特性，在蕴含子图 $G_i$ 中，若 $u$ 和 $v$ 之间的最短路径长度大于 $c\sigma$，则认为 $u$，$v$ 不连通，否则用 $d_i$ 表示 $u$ 和 $v$ 之间的最短路径长度，$d_i$ 的取值范围为 $[0, c\sigma]$. 于是令 $D=c\sigma$，便有 $r\geq\frac{c^2\sigma^2}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$.

(3) 在具有小世界性质的图中，$D$ 近似等于 6[16]. 若不确定图 $G$ 的每个的蕴含子图都具有小世界性质，那么在图 $G$ 上利用算法 1 计算近似期望可达距离时，$r$ 满足 $r\geq\frac{18}{\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$.

在具体应用中，可以根据实际情况选择上述优化方法中的一种来加速计算.

4.2.4 索引维护

索引构建完成后，若用户给出新的精度要求 $\varepsilon'$ 和 $\delta'$，索引维护算法将首先计算满足该精度要求所需采样次数 $r'$. 设构建索引时算法 1 使用的采样次数为 $r$，如果 $r\geq r'$，则索引仍然有效；如果 $r<r'$，则说明采样次数不足，不能满足精度要求 $\varepsilon'$，$\delta'$，此时需要更新索引. 具体方法如下：对不确定图 $G$ 的任意顶点 $u$ 和 $v$，设初始样本空间上的近似期望可达距离为 $d'_{ER}(u,v)$，连通概率为 $p'(u,v)$. 按照 $G$ 的蕴含子图被 $G$ 蕴含的概率对 $Imp(G)$ 进行 $r-r'$ 次独立采样得到补充样本空间. 设补充样本空间上的近似期望可达距离为 $d''_{ER}(u,v)$，连通概率为 $p''(u,v)$. 因此，更新后的近似连通概率以及近似期望可达距离分别为

$$\tilde{p}(u,v)=\frac{p'(u,v)\cdot r+p''(u,v)\cdot(r'-r)}{r'},$$

$$\tilde{d}_{ER}(u,v)=\frac{r\cdot p'(u,v)\cdot d'_{ER}(u,v)+(r'-r)\cdot p''(u,v)\cdot d''_{ER}(u,v)}{p(u,v)}.$$

最后，使用索引构建算法 1 的第 3 步来更新计算 $G$ 中任意两个顶点 $u$ 和 $v$ 之间的可靠期望距离，接着使用 4.2.2 节介绍的方法更新 Graph 表及 Index 表.

**4.3** 查询处理算法

给定查询 $Q=\langle R,\sigma,k\rangle$，查询处理算法首先从 $\alpha$-近邻关系图索引中取出与 $\sigma$ 对应的 $\alpha$-近邻关系图 $G^\alpha$. 对于 $G^\alpha$，有如下定理.

**定理 2.** 给定不确定图 $G=((V,E),\Sigma,L,P)$ 和 $G$ 的 $\alpha$-近邻关系图 $G^\alpha$，$G$ 中任意顶点子集 $W$ 的直径 $diam(W)$ 小于等于 $\sigma$ 当且仅当 $W$ 在 $G^\alpha$ 是一个团.

证明. 首先证明充分性. 若 $W$ 在 $G^\alpha$ 中是一个团，则 $W$ 中任意两个顶点之间的可靠期望距离都小于等于 $\sigma$，于是 $diam(W)\leq\sigma$. 然后，证明必要性. 若 $diam(W)\leq\sigma$，则 $W$ 中任意两个顶点之间的可靠期望距离都不超过 $\sigma$，于是在 $G^\alpha$ 中，$W$ 中任意两个顶点之间都存在一条边，即 $W$ 在 $G^\alpha$ 中是一个团.

证毕.

根据定理 2 易得，在不确定图 $G$ 上执行 top-$k$ 近邻查询 $Q=\langle R,\sigma,k\rangle$ 等价于在 $G^\alpha$ 上计算顶点标签集能够覆盖 $R$ 的直径最小的前 $k$ 个团. 为解决后一问题，本文首先对 $G^\alpha$ 进行预处理，以缩小 $G^\alpha$ 的规模，提高计算效率；然后，在预处理后的 $G^\alpha$ 上进行树搜索，从而计算得到 $G^\alpha$ 上顶点标签集能够覆盖 $R$ 的直径最小的前 $k$ 个团. 第 4.3.1 节提出预处理方法；第 4.3.2 节给出树搜索算法.

4.3.1 预处理算法

**定义 5**($\alpha$-可达标签集). 设 $v$ 是不确定图 $G$ 的 $\alpha$-近邻关系图 $G^\alpha=((V^\alpha,E^\alpha),\Sigma^\alpha,L^\alpha,W^\alpha)$ 中的顶点，$v$ 的 $\alpha$-可达标签集为

$$reachable(v)=L(v)\cup\left(\bigcup_{(v,u)\in E_\alpha}L(u)\right).$$

设不确定图 $G$ 的顶点子集 $W$ 是查询 $Q$ 的一个匹配顶点子集，$W$ 中的顶点满足下面定理 3 给出的性质.

**定理 3.** 设不确定图 $G$ 的顶点子集 $W$ 是查询 $Q$ 的一个匹配顶点子集，对任意 $v\in W$，有 $R\subseteq reachable(v)$ 且 $L(v)\cap R\neq\emptyset$.

证明. 由定理 2，$W$ 在 $G^\alpha$ 中是一个团，于是对任意 $v\in W$，有 $reachable(v)=\bigcup_{v\in W}L(v)$. 再由匹配顶点子集的标签覆盖性，有 $R\subseteq\bigcup_{v\in W}L(v)=reachable(v)$. 由 $W$ 的极小性可知，对任意 $v\in W$，有 $L(v)\cap R\neq\emptyset$.

证毕.

由定理 3，可直接得到如下预处理规则：若 $\alpha$-近邻关系图 $G^\alpha$ 中的顶点 $v$ 满足 $L(v)\cap R=\emptyset$ 或 $R\not\subseteq reachable(v)$，则 $v$ 一定不会出现在任何查询结果中，可将 $v$ 从 $G^\alpha$ 删除.

基于上述预处理规则，本文提出一种两阶段预处理算法.

第 1 阶段，将 $G^\alpha$ 中满足性质 $L(v)\cap R=\emptyset$ 的

顶点全部删除, 同时删除与这些顶点相连的边;

第 2 阶段. 计算 $G^\sigma$ 中剩余顶点的 $\sigma$-可达标签集, 将满足性质 $R \not\subset reachable(v)$ 的顶点全部删除, 同时删除与这些顶点相连的边; 然后更新被删除顶点相邻顶点的 $\sigma$-可达标签集. 接下来继续迭代执行上述过程, 直至将不可能出现在任何查询结果中的顶点及其相连的边全部删除.

算法 **2**.　两阶段预处理算法.

输入: $\sigma$-邻近关系图 $G^\sigma$, 查询 $Q=\langle R, \sigma, k\rangle$

输出: 过滤掉 $G^\sigma$中不可能出现在结果集中顶点及其相连边后得到的确定加权图

1. 初始化待判定队列 $K$ 为 $G^\sigma$的顶点集 $V^\sigma$
2. For $K$ 中的每个顶点 $v$
3.　If $L(v) \cap R = \emptyset$
4.　　从 $K$ 中删除顶点 $v$, 同时从 $G^\sigma$删除 $v$ 及其相连的边
5. While $K$ 不为空
6.　For $K$ 中每个顶点 $v$
7.　　计算 $v$ 的$\sigma$-可达标签集 $reachable(v)$;
8.　　If $R \not\subset reachable(v)$
9.　　将 $v$ 添加到过滤顶点队列 $K'$
10.　　将 $v$ 从 $K$ 中移除
11.　For $K'$ 中每个顶点 $v$
12.　　从 $G^\sigma$中删除顶点 $v$ 以及相连的边
13.　　将与 $v$ 相邻的顶点添加到 $K$
14.　Return $G^\sigma$

算法 2 给出了两阶段预处理算法的伪代码描述. 第 1 行将待判定队列 $K$ 初始化为 $G^\sigma$的顶点集 $V^\sigma$; 第 2~ 4 行执行第 1 阶段过滤. 将不包含任何查询标签的顶点及其相连边删除; 第 5~ 13 行执行第 2 阶段过滤; 最后, 第 14 行返回预处理后的 $G^\sigma$. 实验表明, 该预处理算法可以有效降低查询处理算法的时间复杂性.

### 4.3.2　带剪枝的树搜索算法

根据定理 2, 我们将不确定图 $G$ 上执行 top-$k$ 近邻查询 $Q=\langle R, \sigma, k\rangle$ 的问题等价地转化为在预处理后的 $G^\sigma$上计算顶点标签集能够覆盖 $R$ 的直径最小的前 $k$ 个团的问题. 为解决该问题, 本文提出了一种树搜索算法.

给定 $G$ 的 $\sigma$-近邻关系图 $G^\sigma=((V^\sigma, E^\sigma), \Sigma^\sigma, L^\sigma, W^\sigma)$, 用 $\prec$ 表示 $V^\sigma$上一种全序关系, 如顶点标号的大小关系. $G^\sigma$中所有团可以被组织到一棵搜索树中, 该树具有如下性质:

(1) 树的每个节点唯一表示 $G^\sigma$中一个团;

(2) 树的根节点表示平凡团, 即空集;

(3) 树的非叶节点 $C$ 的儿子节点表示另一个团 $C'$, 满足

① $C \subset C'$;

② $|C'| = |C| + 1$;

③ 顶点 $v \in C' \setminus C$ 与 $C$ 中任一顶点 $u$ 都满足 $u \prec v$.

通过设置合理的顶点之间的全序关系 $\prec$, 可进一步降低搜索空间. 本文采用 Rarest First[10]思想, 先找出 $G^\sigma$中出现次数最少的标签 $L_{Rarest}$, 然后令 $G^\sigma$中包含标签 $L_{Rarest}$的顶点 $u$ 与任意不包含标签 $L_{Rarest}$的顶点 $v$ 满足关系 $u \prec v$.

按照 Rarest First 原则建立全序关系后, 开始建立搜索树, 具体算法过程如下: 首先将当前搜索节点 $S$ 置为搜索树的根节点. 计算 $S$ 的候选扩展顶点集合 $cand(S)$. $cand(S)$满足对任意顶点 $v \in cand(S)$, 有(1) 对任意 $u \in S$, 有 $u \prec v$; (2) $v$ 与 $S$ 中所有顶点都相邻; (3) $L(v) \cap (L(S) \setminus R) \neq \emptyset$, 其中 $L(S)$表示 $S$ 中所有顶点上的标签集的并集. 特别的是, 当 $S=\emptyset$ 时, $cand(S)$为包含标签 $L_{Rarest}$的顶点集. 然后, 为每个顶点 $v \in cand(S)$, 建立新节点 $v \cup S$, 易知每个新节点都是 $S$ 的儿子节点. 接下来, 依次将每个新构造的节点作为当前搜索节点 $S$, 迭代执行上述扩展操作, 直到待扩展节点的候选扩展顶点集为空或者当前节点满足查询要求时, 结束扩展. 算法 3 给出了带标签覆盖的团搜索算法的伪代码描述, 其子过程 $TSC$ 描述了建立搜索树的过程.

算法 **3**.　带标签覆盖的团查询算法.

输入: $\sigma$-近邻关系图 $G^\sigma$, 标签集 $R$, 整数 $k$

输出: 直径最小的前 $k$ 个标签覆盖 $R$ 的团

1. 初始化结果集 $F$ 为空集
2. $F = TSC(\emptyset)$
3. Ruturn $F$ 中直径最小的前 $k$ 个顶点集.

// 建立搜索树的子过程 $TSC(vector\ S)$ 描述如下

// 输入: 团 $S$

// 输出: 标签集包含 $R$ 的团

4. If $S$ 的标签集合包含 $R$
5.　Return $S$
6. Else
7.　计算 $S$ 的扩展候选集合 $cand(S)$
8.　If $cand(S)$ 不为空
9.　For $cand(S)$ 中的每个顶点 $v$
10.　$TSC(v \cup S)$

在搜索树的基础上运用分支界限方法计算直径最小的前 $k$ 个团可以尽早剪枝. 使用一个尺寸为 $k$ 的最大堆 $H_{top\text{-}k}$来存储当前搜索结果中直径最小的

前 $k$ 个团, 其中堆的根节点表示 $H_{top\text{-}k}$ 中半径最大的团 $S$, $S$ 在 $G$ 中的直径为 $diam(S)$. 当树搜索算法建立新节点 $S'$ 时, 比较 $diam(S')$ 与 $diam(S)$ 大小. 若前者大, 停止继续扩展该节点, 即剪枝; 否则, 验证 $S'$ 的标签集合能否覆盖 $R$, 若能, 则用 $S'$ 替换掉 $H_{top\text{-}k}$ 中的根节点数据, 并重新维护最大堆 $H_{top\text{-}k}$; 否则, 继续扩展该节点. 算法结束时, $H_{top\text{-}k}$ 中存储的 $k$ 个团即为半径最小的 $k$ 个标签覆盖 $R$ 的团. 最后, 将 $H_{top\text{-}k}$ 中 $k$ 个团全部输出. 至此, 完成了不确定图上的 top-$k$ 近邻查询.

# 5 实验结果

## 5.1 实验配置

我们进行了大量实验来验证本文提出的索引构建算法的执行效率及其可扩展性以及查询处理算法的性能. 由于目前尚没有处理不确定图上的 top-$k$ 近邻查询的算法, 因此我们主要考察本文提出的方法在不同规模的数据集上以及不同参数条件下的效率. 实验中使用的数据取自 BioGRID 数据库[17] 以及通过修改真实图数据得到的合成不确定图数据. 表 1 给出了实验中使用到的不确定图数据的特性. 其中 $Rat$、$ThaleCress$、$Worm$ 和 $Human$ 是取自 BioGRID 数据库的真实数据; $G_1$, $G_2$, $G_3$, $G_4$ 为合成数据, 它们是从 Human 中随机抽取不同数量的顶点得到的子图.

实验运行环境为具有 Intel Pentium 4 3.2GHz CPU、512MB 内存、运行 Ubuntu 10.04 的 PC 机. 算法采用 C++ 语言编写, 并使用 Boost Graph Library 1.44.0.

表 1 实验使用的不确定图数据的规模

| 不确定图名称 | 顶点数 | 边数 |
|---|---|---|
| $Rat$ | 129 | 90 |
| $ThaleCress$ | 529 | 597 |
| $Worm$ | 827 | 961 |
| $Human$ | 4121 | 10008 |
| $G_1$ | 500 | 823 |
| $G_2$ | 1000 | 1607 |
| $G_3$ | 1500 | 6958 |
| $G_4$ | 2000 | 16589 |

实验中使用 4.3.2 节提到的第 2 种索引构建优化策略, 采样次数 $r$ 满足 $r=\max\left\{\frac{6}{\varepsilon^2}, \frac{c^2\sigma^2}{2\varepsilon^2}\right\}\ln\left(\frac{2}{\delta}\right)$, 其中距离约束 $\sigma \le 4$, 并以小世界理论为参考, 设置距离约束的松弛系数 $c \in [1.5, 6]$.

## 5.2 索引构建算法性能分析

我们首先考察算法 1 中采样次数 $r$ 对索引构建时间的影响. 图 4 给出在 $Rat$、$ThaleCress$ 和 $Worm$ 3 个不确定图上, 当采样次数从 1 增加到 150 时, 索引构建时间的变化情况, 其中 $\sigma=3$, $c=2$. 从图 4 可以看出, 索引构建时间随采样次数的增加线性增长, 并且在采样次数相同的情况下, 索引构建时间随不确定图尺寸的增大而增加.

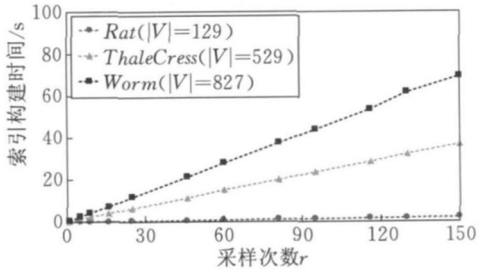


图 4 不同采样次数下的索引构建时间

用户给定精度要求参数 $\varepsilon$ 和 $\delta$, 设顶点 $u$ 和 $v$ 之间精确可靠期望距离为 $d_{RE}(u, v)$, 近似可靠期望距离为 $\hat{d}_{RE}(u, v)$, 定义指示函数 $I(u, v)$:

$$I(u,v)=\begin{cases}1, & \text{若} |\hat{d}_{RE}(u,v)-d_{RE}(u,v)| \le \varepsilon \\ 0, & \text{否则}\end{cases}.$$

由此我们定义近似距离相对于精确距离的精确度

$$Precision=\frac{2}{n(n-1)}\sum_{u,v\in V} I(u,v).$$

图 5 给出在 $Rat$、$ThaleCress$ 和 $Worm$ 3 个不确定图上, 当 $1-\delta$ 从 0.1 增加到 0.9 时, 近似可靠期望距离的精确度变化情况, 其中 $\sigma=3$, $c=2$, $\varepsilon=0.5$. 横坐标表示由 $\varepsilon$, $\delta$, $\sigma$ 确定的采样次数, $\delta$ 从 0.9 减小到 0.1. 从图 5 可以看出, 在 $Rat$、$ThaleCress$ 和 $Worm$ 3 个不确定图上计算得到的近似可靠期望距离的精确度均高于 $1-\delta$, 并且随着 $1-\delta$ 的增加, 近似可靠期望距离的精确度达到较高水平后保持稳定.

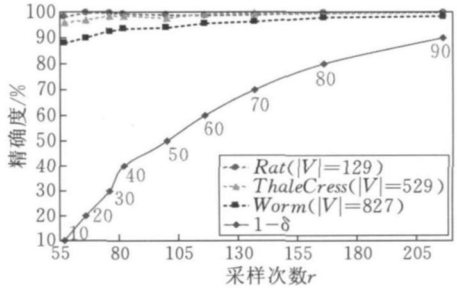


图 5 不同 $\delta$ 下的近似距离精确度

## 5.3 查询处理算法性能分析

本节考察近邻查询 $Q=\langle R, \sigma, k\rangle$ 中的 $R$, $\sigma$ 和 $k$

对本文提出的算法性能的影响, 此外还考察不确定图的规模对算法性能的影响.

首先, 考察查询标签集 $R$ 的尺寸以及距离约束 σ 对查询处理时间的影响. 图 6 给出在不确定图 *Worm* 上, 当距离约束从 1 增加到 4 时, 不同尺寸标签集的查询处理时间变化. 其中 *Query_R3*, *Query_R5*, *Query_R7*, *Query_R9* 分别对应查询标签集尺寸等于 3, 5, 7, 9 的查询, 查询处理时间是随机生成的尺寸相同的 20 组查询的平均查询处理时间, 并且 $k$= 20. 从图 6 可以看出, 当查询标签集的尺寸相同时, 随着距离约束 σ 的增加, 查询处理时间呈指数增长; 当距离约束 σ 相同时, 查询处理时间随查询标签集尺寸的增大而增加.

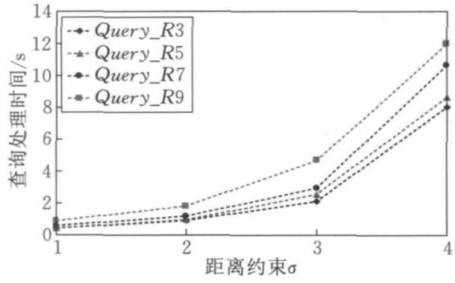


图 6　不同尺寸查询标签集以及不同距离约束 σ 的查询处理时间

其次, 考察结果数量 $k$ 以及图数据的规模对查询处理时间的影响. 图 7 给出在合成不确定图 $G_1$, $G_2$, $G_3$, $G_4$ 上, 当整数 $k$ 从 20 增加到 100 时, 查询处理时间的变化. 其中 $|R|$ = 5, σ= 2, 查询处理时间是随机生成的 20 组查询的平均处理时间. 从图 7 中不难看出, 在同一个图上进行查询时, $k$ 的大小对查询处理时间的影响很小; 并且在 $k$ 相同的情况下, 随着不确定图规模的扩大, 查询处理时间呈指数增长.

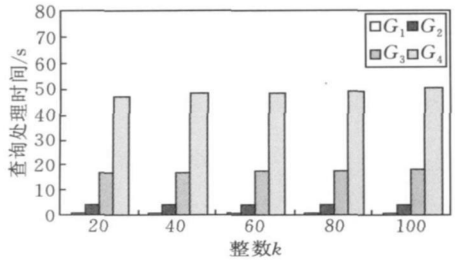


图 7　不同的整数 $k$ 以及不同规模图数据的查询处理时间

## 6　结　论

本文研究了不确定图上的 top-$k$ 近邻查询问题, 首次形式化定义了不确定图 top-$k$ 近邻查询问题, 设计了有效支持该查询处理的 σ-近邻关系图索引结构, 提出了基于 Monte Carlo 随机算法的不确定图顶点间可靠期望距离计算算法, 并提出了基于此算法的 σ-近邻关系图索引构建算法. 在 σ-近邻关系图索引的基础上, 将不确定图 top-$k$ 邻近查询问题转化为计算 σ-近邻关系图上带标签覆盖的 top-$k$ 团问题. 本文提出了一种两阶段预处理算法来有效减小 σ-近邻关系图的规模, 还提出了基于分支界限法的树搜索算法来快速计算 σ-近邻关系图上带标签覆盖的 top-$k$ 团. 本文进行了大量实验来考察索引构建算法和查询处理算法的性能. 实验结果表明本文提出的算法能有效处理不确定图 top-$k$ 近邻查询.

## 参　考　文　献

[1] Asthana S, King O D, Gibbons F D, Roth F P. Predicting protein complex membership using probabilistic network reliability. Genome Research, 2004, 14: 1170-1175

[2] Adar E, Re C. Managing uncertainty in social networks. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2007, 30(2): 15-22

[3] Biswas S, Morris R. Exor: Opportunistic multi-hop routing for wireless networks. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 133-144

[4] Zou Zhao-Nian, Li Jian-Zhong, Gao Hong, Zhang Shuo. Mining frequent subgraph patterns from uncertain graph data. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1203-1218

[5] Potamias M, Bonchi F, Gionis A, Kollios G. $k$-Nearest neighbors in uncertain graphs. The VLDB Endowment, 2010, 3(1-2): 997-1008

[6] Yuan Ye, Chen Lei, Wang Guo-Ren. Efficiently answering probability threshold-based shortest path queries over uncertain graphs//Proceedings of the IEEE DASFAA. Tsukuba, Japan, 2010: 155-170

[7] Zhang Shuo, Gao Hong, Li Jian-Zhong, Zou Zhao-Nian. Efficient query processing on uncertain graph databases. Chinese Journal of Computers, 2009, 32(10): 2066-2079(in Chinese)
(张硕, 高宏, 李建中, 邹兆年. 不确定图数据库中的高效查询处理. 计算机学报, 2009, 32(10): 2066-2079)

[8] Zou Lei, Chen Lei, Lu Yan-Sheng. Top-$k$ subgraph matching query in a large graph//Proceedings of the ACM 1st Ph.D. Workshop in CIKM. Lisbon, Portugal, 2007: 139-146

[9] Wang Hong-Zhi, Li Jian-Zhong, Luo Ji-Zhou, Gao Hong. Hash-based subgraph query processing method for graph-structured XML documents. The VLDB Endowment, 2008, 1(1): 478-489

[10] Lappas T, Liu Kun, Terzi E. Finding a team of experts in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, NY, USA: ACM, 2009: 467-476

[11] Cheng Jie-Feng, Jeffrey Xu Yu, Ding Bo-Lin, Wang Hai-Xun. Fast graph pattern matching//Proceedings of the IEEE 24th International Conference on Data Engineering. Cancun, Mexico, 2008: 913-922

[12] Zou Lei, Chen Lei, Tamer Ozsu M. Distance-join: Pattern match query in a large graph database. The VLDB Endowment, 2009, 2(1): 886-897

[13] Mitzenmacher M et al. Probability and Computing: Randomized Algorithms and Probabilistic Analysis. England: Cambridge University Press, 2005

[14] Johnson Donald B. Efficient algorithm shortest paths in sparse networks. Journal of ACM, 1977, 24(1): 1-13

[15] Aingworth D, Chekuri C, Motwani R. Fast estimation of diameter and shortest paths (without matrix multiplication)// Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'96). PA, USA, 1996

[16] Elmacioglu E, Lee D. On six degrees of separation in DBLP-DB and more. ACM SIGMOD Record, 2005, 34(2): 33-40

[17] Zou Zhao-Nian, Li Jian-Zhong, Gao Hong, Zhang Shuo. Finding top-$k$ maximal cliques in an uncertain graph//Proceedings of the IEEE 26th International Conference on Data Engineering. California, USA, 2010: 649-652

**ZHANG Hai-Jie**, born in 1986, M.S. candidate. Her research interests focus on uncertain query and mining.

**JIANG Shou-Xu**, born in 1968, Ph. D., professor, Ph. D. supervisor. His research interests include data management in large-scale dynamic networks, peer-to-peer computing and trust management, databases, wireless sensor networks, delay tolerant networking, etc.

**ZOU Zhao-Nian**, born in 1979, Ph. D., lecturer. His research interests focus on graph data mining.

**Background**

In recently years, a large number of data are modeled by graphs in real world, e. g. molecular structure of compounds, topological structure of mobile ad-hoc networks, the structure of social networks, etc. But in the practice applications, because of expression of distributed and heterogeneous environment, privacy protection, incomplete data and input errors, the uncertainty of graphs is widespread. Therefore, it is of significant to study the efficient query processing algorithms on uncertain graphs. Nowadays, the research on uncertain graphs is one of the research focuses, including query processing and important pattern mining. However, almost all the existing works consider the exact structure of graphs, e.g. sub-graph query and sub-graph pattern mining. But in the scenario such that team formation, users may do not know the exact structure of the answers beforehand, it's necessary to propose a new kind of query to adapt these scenarios. This paper propose proximity query. The proximity query on uncertain graphs does not limit the structure of the answers but it's required that the vertices in the answers be close enough. Theoretical and experimental results show that the proposed algorithm can efficiently retrieve the top-k proximity query results.