计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(10): 2208－2220, 2012

不确定图上期望最短距离的计算

李鸣鹏[1] 邹兆年[1] 高 宏[1] 赵正理[2]
[1](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)
[2](哈尔滨工业大学英才学院 哈尔滨 150001)
(limingpengconan@gmail.com)

Computing Expected Shortest Distance in Uncertain Graphs

Li Mingpeng[1], Zou Zhaonian[1], Gao Hong[1], and Zhao Zhengli[2]
[1](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)
[2](*School of Honors, Harbin Institute of Technology, Harbin* 150001)

Abstract This paper focuses on the shortest distance problem in uncertain graphs, which we call expected shortest distance problem. We analyze the complexity of this problem and prove that there is no polynomial time algorithm for it. To solve this problem, we utilize random sampling methods to acquire some possible worlds of the uncertain graph, then compute the shortest distance on each and estimate expected shortest distance with the average value of the finite ones. To improve efficiency, we propose two pruning techniques, which allow us to terminate a random sampling process faster. Furthermore, considering that different sampling orders of edges do not influence the result of sampling, but will determine the number of edges to be sampled in a sampling process, we propose two sampling orders for edges to reduce the number of edges sampled in each random sampling process. Then, we propose an approximation algorithm based on random sampling using antithetic variables which is an unbiased estimator for expected shortest distance and prove that it outperforms direct random sampling in both efficiency and sampling variance, while the latter one is the key criteria for evaluating the quality of an unbiased estimator. Our experiments on real uncertain graphs of protein-protein networks demonstrate the efficiency and accuracy of our algorithm.

Key words uncertain graph; expected shortest distance; random sampling; antithetic variables sampling; sampling variance

摘 要 研究了不确定图上的最短距离问题,提出了期望最短距离的概念,证明了该问题不存在多项式时间的算法.为了解决该问题,使用了随机采样技术获得不确定图的一些可能世界,在每个可能世界上计算有穷的最短距离,最后计算出平均值作为期望最短距离的估计值.为提高计算效率,使用了过滤条件来减少采样过程中采样的边数从而加快随机采样.在此基础上,提出了一种基于对称变量的、无偏的随机采样近似算法,并证明了与直接随机采样方法相比,该方法在不增加时间开销的同时能减小采样方差.通过真实数据上的实验表明,提出的算法在时间开销和采样方差上均明显好于直接随机采样方法.

收稿日期:2012-06-05;修回日期:2012-07-26
基金项目:国家"九七三"重点基础研究发展计划基金项目(2012CB316200);国家自然科学基金项目(61173023,61190115,61033015);中央高校基本科研业务费专项基金项目(HIT.NSRIF.201180)
通信作者:邹兆年(znzou@hit.edu.cn)

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

关键词 不确定图;期望最短距离;随机采样;对称变量采样;采样方差

中图法分类号 TP301.6

许多现实应用中的数据都可以抽象为不确定图,比如蛋白质交互网络[1]、社交网络[2]、无线传感器网络[3]、P2P网络[4]等.不确定图是指顶点确定、边以一定概率存在的图,且边的存在概率是相互独立的.一个确定图也可以看作是所有边的概率均为1的特殊不确定图.如图1(a)所示,g是一个无向不确定图,g中的每一条边都以一定概率存在,我们判断g中每一条边的是否存在后可以获得一个确定图G_1,如图1(b)所示,称之为不确定图的一个可能世界[5].一个不确定图蕴含了多个可能世界,而每一个可能世界都对应一个概率值,所有可能世界概率之和为1[5].

不确定图上的一个基本问题就是最短距离问题.与确定图上的最短距离相比,在不确定图上计算最短距离时有两点不同:1)不确定图上的边是以一定概率存在的;2)给定两个顶点s和t,每一条连接s和t路径都有可能成为连接s和t的最短路径.

文献[6]提出了概率语义下的最短路径问题,对于不确定图上给定的两个顶点,它们在不同的可能世界中有不同的最短路径,在图1(a)中的不确定图g中,s和t可以通过3条路径($P_1=e_1e_2$,$P_2=e_3e_4$,$P_3=e_5e_6e_7$)连接.表1描述了在g的不同可能世界中,s和t不同的最短路径及其相应的概率值,其中e_i表示边存在,$\bar{e}_i$表示边不存在,e_i^+表示边是否存在均可.文献[6]还给出了计算大于给定阈值的所有最短路径的算法.如果阈值为0.1,那么P_1,P_2,P_3均会被返回.

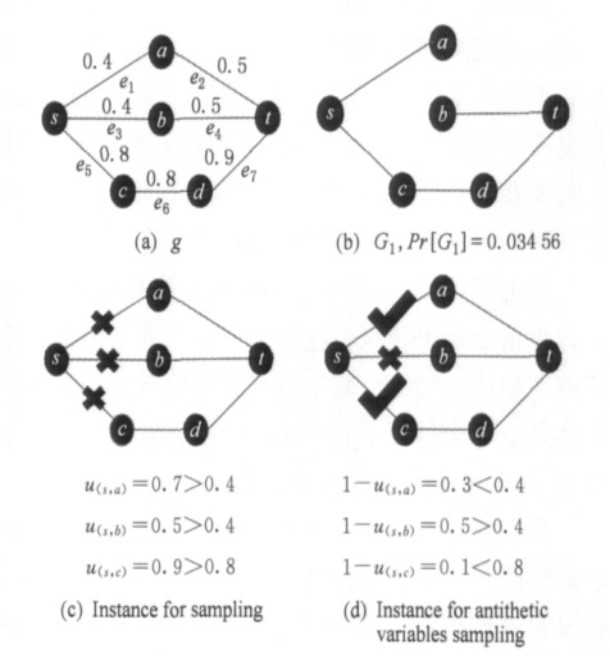


Table 1 Shortest Path and Probability of s and t in Possible Worlds of g

表1 g的不同可能世界中的s,t最短路径及其概率

Path	P_1	P_2	P_3
Possible Worlds	$e_1e_2e_3^+e_4^+e_5^+e_6^+e_7^+$ ($Pr=0.2$)	$e_1^+e_2^+e_3e_4e_5^+e_6^+e_7^+$ ($Pr=0.2$)	$e_1\bar{e}_2e_3\bar{e}_4e_5e_6e_7$ ($Pr=0.02304$) $e_1\bar{e}_2\bar{e}_3e_4e_5e_6e_7$ ($Pr=0.03456$) $e_1\bar{e}_2e_3\bar{e}_4e_5e_6e_7$ ($Pr=0.03456$) $\bar{e}_1e_2e_3\bar{e}_4e_5e_6e_7$ ($Pr=0.03456$) $\bar{e}_1e_2\bar{e}_3e_4e_5e_6e_7$ ($Pr=0.05184$) $\bar{e}_1e_2e_3\bar{e}_4e_5e_6e_7$ ($Pr=0.05184$) $\bar{e}_1\bar{e}_2e_3e_4e_5e_6e_7$ ($Pr=0.03456$) $\bar{e}_1\bar{e}_2e_3e_4e_5e_6e_7$ ($Pr=0.05184$) $\bar{e}_1\bar{e}_2e_3\bar{e}_4e_5e_6e_7$ ($Pr=0.05184$)
Probability	0.2	0.2	0.36864

概率语义下的最短路径问题可以用来获取无线传感网络中的最短路由.而当我们要考察两个传感器节点之间的传输质量时,不但要知道它们之间的最短路由,还需要计算出这些最短路由的平均传输长度.仍以图1(a)为例,s和t之间最短路由有$P_1(2,0.2)$,$P_2(2,0.2)$和$P_3(3,0.36864)$,其中$P_1(2,0.2)$表示路由的长度为2,存在概率为0.2.由上述信息我们发现,s和t之间更有可能存在一条长度为2的路由.但通过枚举g的所有可能世界,我们可以计算出s和t之间最短距离的平均值为2.51,这反映出在g的多数可能世界中,s和t之间存在着一条长度为3的最短路径.

此时概率语义下的最短路径已经无法适用,这是因为:1)概率语义下的最短路径仅仅返回了大于

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

给定阈值的最短路径,因此缺失了部分最短路径信息,在上面的例子中,如果阈值提高至 0.3,那么 P_1 和 P_2 均会丢失;2)对长度相同的最短路径,它们出现的可能世界会有重复.在计算平均路由长度时,我们需要计算"出现长度为 2 的路由"的概率,在上面的例子中,可表示为 $Pr(P_1 \cup P_2)$,这显然与 $Pr(P_1)+Pr(P_2)$不同.因为上例中的两条路径相互独立,我们有 $Pr(P_1 \cup P_2)=1-(1-0.2)\times(1-0.2)=0.36<0.2+0.2$.

而期望语义下的最短距离反映了不确定图上两个顶点之间最短距离的总体特征,文献[7]对概率语义和期望语义下的子图挖掘问题作了细致的对比,并说明了这两种语义各自适用的范畴.类似地,当我们需要获取两个顶点之间的平均最短距离时,期望语义下的最短距离是更好的选择.

综上所述,本文提出了期望最短距离的概念,就是计算给定两个顶点在连通的可能世界上的最短距离的期望值.事实上,两个顶点之间的最短距离可以看作是 g 的全体可能世界 $Imp(g)$上的一个概率分布,在 2.1 节中将会详细地介绍.

我们通过构造一个多项式的归约证明了计算期望最短距离一定不会比计算 d 可达问题[8]更容易,而 d 可达问题已经证明了是 $\#P$ 完全的[8-10],据此说明本文所提出的问题是不存在多项式时间算法的,除非 $P=NP$.

本文采用随机采样方法来解决不确定图上的期望最短距离计算问题,对于输入不确定图 g,我们按照其所有可能世界的概率分布进行独立、随机采样,获得 n 个可能世界,对于其中每个可能世界,我们计算 s 和 t 之间的最短路径长度,最后我们计算这 n 个可能世界上 s 和 t 之间有穷的最短路径长度的平均值作为 s 与 t 之间期望最短距离的估计值.我们提出了两个过滤条件来改进随机采样的计算效率,并在此基础上提出了基于对称变量的随机采样方法,就是在 n 次采样的奇数次采样时进行独立的随机采样,在偶数次采样时改为计算奇数次采样结果的对称变量.我们证明了本文提出的随机采样方法可以减小采样方差.

本文的主要工作及贡献如下:

1)提出了不确定图上的一个新的研究问题,期望最短距离,并给出了具体的形式化定义;

2)分析了所提出问题的复杂性,证明了期望最短距离问题不会比 d 可达问题容易解决,据此说明问题没有多项式时间算法,除非 $P=NP$;

3)采用随机采样的方法来计算问题的近似解,使用过滤条件来提高采样效率,又提出了基于对称变量的随机采样来减小采样方差,从理论上给出了准确性的保证;

4)通过实验验证了本文所提出的随机采样方法的有效性和准确性.

1 相关工作

不确定图上的可达问题已经有了较为广泛的研究[8-12],由于问题固有的复杂性,精确算法只能处理数据规模较小的情形.对于不确定图上的最短路径问题,一些确定图上的方法[13]由于不能处理边的不确定性,因此无法适用到不确定图上.文献[6]提出了概率语义下的最短路径问题,计算了不确定图的可能世界中概率超过给定阈值的所有最短路径,提出了一种基于访问候选路径集合的算法,并通过对同构子图的合并与剪枝技术提高了算法的效率.

期望最短距离反映了不确定图上两个顶点在所有连通的可能世界中的平均最短距离.概率语义下的最短路径问题由于只能求出一部分大于给定阈值最短路径,并且所求出的相同长度路径出现的可能世界有重复,因此不能用来计算期望最短距离.可达问题则仅仅关心两个顶点可达的概率.因此,上述工作和方法都不能适用于计算期望最短距离.本文采用了随机采样的方法来估计期望最短距离,使用了两个过滤条件加快随机采样,并通过引入对称变量减小了随机采样方法的方差.

2 问题定义及计算复杂性分析

2.1 不确定图

一个不确定的无向图可以表示为一个三元组,$G=(V,E,Pr)$,其中 V 是顶点集合,E 是边集合,$Pr:E\rightarrow(0,1]$是边上的概率函数.我们假定不确定图中边的存在是相互独立的.

一个不确定图蕴含了多个不同的确定图,记作 $Imp(g)$.对于每一个可能的确定图 $G=(V,E_G)$,我们称之为不确定图 g 的一个可能世界,G 的顶点集与不确定图 g 完全一致,边集合 E_G 满足 $E_G \subseteq E$.我们知道,E 共有 $2^{|E|}$ 个不同的子集,所以 $|Imp(g)|=2^{|E|}$.对于 g 的任何一个可能世界 G,我们可以计算其发生的概率,记作 $Pr[g \Rightarrow G]$,在不引起混淆时简记为 $Pr[G]$,如式(1)所示:

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$Pr[G]=\prod_{e\in E_G}p(e)\prod_{e\notin E_G}(1-p(e)). \quad (1)$$

在图 1 中,不确定图 g 共有 2^7 个可能世界,对于其中某一个可能世界 G_1,如图 1(b)所示,我们有 $Pr[G_1]=Pr((s,a))\times Pr((s,c))\times Pr((b,t))\times Pr((c,d))\times Pr((d,t))\times(1-Pr((s,b)))\times(1-Pr((a,t)))=0.034\,56$.

给定确定图 G 上两个顶点,它们可以通过许多条路径连接,其中路径上的顶点和边均无重复地称为简单路径. 我们使用长度最短的简单路径表示这两个顶点的最短距离. 对于不确定图 g,由于它包含了多个可能世界,并且在每一个可能世界中,给定两个顶点之间的最短距离都可能不同,因此最短距离是样本空间 $Imp(g)$ 上的一个随机变量,其概率分布如表 2 所示:

Table 2 Probability Distribution of Shortest Distance

表 2 最短距离的概率分布

Shortest Distance	Probability
d_1	$Pr(d_1)$
d_2	$Pr(d_2)$
⋮	⋮
d_m	$Pr(d_m)$
∞	$Pr(\infty)$

表 2 中 $d_1,d_2,\cdots,d_m$ 是 g 的 $2^{|E|}$ 个可能世界中所有不同的最短距离. 以图 1(a)为例,s 和 t 之间的最短距离有两个:2 和 3. 我们用 $dis_{s,t}(G)$ 表示 s 和 t 在可能世界 G 上的最短距离,$m=|\{dis_{s,t}(G)\,|\,G\in Imp(g)\}|$ 表示 s 和 t 在不确定图 g 所有可能世界中不同的最短距离的个数,$Pr(d_i\,|\,g)=\sum_{G\in Imp(g)\wedge dis_{s,t}(G)=d_i}Pr[G]$ 是 s 和 t 在所有最短距离为 d_i 的可能世界中的概率,简记为 $Pr(d_i)$.

值得注意的是,在 g 的一些可能世界中,顶点 s 和 t 不连通,此时我们认为 s 和 t 的最短距离 $dis_{s,t}(G)=\infty$,于是有 $Pr(\infty)=\sum_{G\in Imp(g)\wedge dis_{s,t}(G)=\infty}Pr[G]$.

定义 1. 不确定图 g 上两个顶点 s 和 t 之间期望最短距离就是在最短距离有穷的条件下,g 的所有可能世界中最短距离的期望,记作 $E_{s,t}(g)$,简记为 $E_{s,t}$:

$$E_{s,t}(g)=\sum_{dis_{s,t}(G)<\infty}(dis_{s,t}(G)\times Pr[G])\Big/\sum_{dis_{s,t}(G)<\infty}Pr[G]=\sum_{i=1}^{m}(d_i\times Pr(d_i))\Big/\sum_{i=1}^{m}Pr(d_i). \quad (2)$$

之所以要排除掉顶点 s 和 t 不连通的情况是因为此时它们的距离为 ∞,如果在计算期望值的时候将这种情况也考虑进去,那么 $E_{s,t}(g)=\infty$,这显然是没有任何意义的.

2.2 计算复杂性分析

在给出了问题的形式化定义后,我们首先需要考虑的是上述问题的计算复杂性,也就是问题是否存在多项式时间的算法. 对于网络可靠性问题和不确定图上的 d 可达问题,文献[8,14-15]证明了它们是 $\#P$ 完全的,也就是说上述问题不存在一个多项式时间的算法,除非 $P=NP$. 下面我们将在此基础上,说明期望最短距离问题也是不存在多项式时间的算法的.

定理 1. 期望最短距离问题不存在多项式算法,除非 $P=NP$.

证明. 用反证法. 假设本文的问题存在多项式时间算法 A. 那么对给定不确定图 g 以及顶点对 s 和 t,我们可以构造出计算 d 可达问题的算法 B,具体如下:

1) 首先求出 g 的最大可能世界 $\bar{G}=(V,E)$.

2) 在 $\bar{G}$ 上使用 Dijkstra 算法求出 s 和 t 的最短路径长度 $d_{\min}$.

3) 新增 $d_{\min}$ 个顶点 $\{\tilde{v}_1,\tilde{v}_2,\cdots,\tilde{v}_{d_{\min}}\}$ 和 $d_{\min}+1$ 条边 $\{(s,\tilde{v}_1),(\tilde{v}_1,\tilde{v}_2),\cdots,(\tilde{v}_{d_{\min}-1},\tilde{v}_{d_{\min}}),(\tilde{v}_{d_{\min}},t)\}$ 至 g,且所有新增边的概率均为 1,得到不确定图 g_1. 在 g_1 上执行算法 A,可以求出 $E_{s,t}(g_1)$. 因为 g_1 中一定存在一条长度为 $d_{\min}+1$ 的路径,所以根据式(2)我们有 $E_{s,t}(g_1)=d_{\min}\times Pr(d_{\min}\,|\,g_1)+(d_{\min}+1)\times(1-Pr(d_{\min}\,|\,g_1))$. 于是可以求 $Pr(d_{\min}\,|\,g_1)$,而在上述构造过程中,易知 $Pr(d_{\min}\,|\,g_1)=Pr(d_{\min}\,|\,g)$. 所以我们通过向 g 中加入 $d_{\min}$ 个顶点和 $d_{\min}+1$ 条概率为 1 的边,利用算法 A 求出了 $Pr(d_{\min}\,|\,g)$.

4) 向 $Pr(d_{\min}+1\,|\,g)$ 中依次加入 $\{d_{\min}+1,d_{\min}+2,\cdots,|V|-1\}$ 个顶点和 $\{d_{\min}+2,d_{\min}+3,\cdots,|V|\}$ 条概率为 1 的边,得到不确定图 $\{g_2,g_3,\cdots,g_{|V|-d_{\min}}\}$. 在上述不确定图上依次执行算法 A,可以求出 $\{Pr(d_{\min}+1\,|\,g),Pr(d_{\min}+2\,|\,g),\cdots,Pr(|V|-1\,|\,g)\}$. 那么此时,对于任意给定的 d,s 和 t 在 g 中 d 可达的概率为 $\sum_{i=d_{\min}}^{d}Pr(i\,|\,g)$.

下面分析算法 B 的复杂性. ①的时间复杂性为 $O(|V|+|E|)$;②的时间复杂性为 $O(|E|+|V|\log|V|)$;③的时间复杂性为 $O(|V|)$与算法 A 的复杂度之和;

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同理④的时间复杂性为 $O(|V|^2)$ 与 $|V|$ 次执行算法 A 的复杂度之和. 算法 B 是一个多项式时间算法. 矛盾!

根据上述分析可知,我们的问题是不会比 d 可达问题更容易解决的,因此也是不存在多项式时间算法的,除非 $P=NP$. 由于不确定图上 d 可达问题的精确算法只能够应用到规模较小的图上(几十条边). 而在我们的问题中,如果想精确计算给定两点之间的期望最短距离,就需要枚举不确定图 g 全部的 $2^{|E|}$ 个可能世界,因此显然不能应用到规模较大的图上. 故本文主要探讨的是如何给出有效率、准确的近似算法.

3 基于对称变量的随机采样方法

解决不确定图上期望最短距离计算问题的一种最直观的方法如下:

1) 对输入不确定图 g 的可能世界按照其被 g 蕴含的概率进行独立、随机的 N 次采样. 由于 g 中的边相互独立,因此只需对 g 中每条边按其存在概率进行采样即可得到一个可能世界.

2) 在采样得到的每个可能世界上计算 s 与 t 之间的最短距离.

3) 计算 N 个可能世界样本中 s 与 t 之间的有穷最短距离的平均值作为 g 中 s 与 t 之间的期望最短距离的估计值.

该方法存在的主要问题是对边进行采样的次数过多,从而计算效率较低. 因此,为了提高期望最短距离的计算效率,本节提出了两个过滤条件来减少每次随机采样中需要采样的边数;并且为了使每次随机采样能更快地满足这些过滤条件,还提出了两种边采样顺序. 由于边的采样顺序对采样结果没有影响,因此该方法可以在不改变采样结果的前提下提高计算效率. 在此基础上,本节最后提出了一种基于对称变量的随机采样方法,并证明了该方法得到的结果是期望最短距离的无偏估计,并且能在不增加时间开销的前提下减小估计值的方差.

一种随机采样方法的准确性通常使用均方差来表示[16],当随机采样得到的估计值越接近真实值时,采样均方差越小. 为了保证随机采样结果的可信性,随机采样方法需要保证采样满足无偏性,也就是采样结果 $\hat{E}$ 的期望与真实值相等,即 $E(\hat{E})=E_{s,t}(g)$,而此时采样均方差恰好就是采样方差. 于是衡量一种随机采样方法的好坏就取决于采样方差的大小. 本文中提到的随机采样方法均满足无偏性,因此它们的准确性只体现在它们的采样方差上.

另外值得注意的是,对于给定的不确定图 g 和顶点 s 与 t,并不是所有的边都需要在随机采样时被采样到. 通过 $\bar{G}$ 我们可以枚举出 g 中所有连接 s 和 t 的路径,记作 $\mathrm{P}=\{P_1,P_2,\cdots,P_r\}$,对于没有在任意一条路径中出现的边我们不需对其采样. 用 E_s 表示需要随机采样的边集合,那么 E_S 满足 $E_S=\bigcup_{i=1}^{r}\bigcup_{e\in P_i} e$.

3.1 采样终止条件

在前面给出的最简单的采样算法中,在对 g 的可能世界进行随机采样时,需要对 g 中每条边进行一次采样;然而,实际上并不需要对 g 中每条边都进行一次采样. 下面给出两个采样终止条件,用于提前结束对边的采样,从而减少对边的采样次数.

设 E_1 是已经被采样过且被放入可能世界中的边构成的集合,E_2 是已经被采样过但未被放入可能世界中的边构成的集合,E_3 是尚未被采样的边构成集合. 设 P_2 是 P 中所有包含 E_2 中的边的路径构成的集合,$\mathrm{P}_1=\mathrm{P}/\mathrm{P}_2$. 以图 2(a)中的不确定图为例,设 $E_1=\{e_2\}$,$E_2=\{e_3\}$,$E_3=\{e_1,e_4\}$,则 $\mathrm{P}_2=\{e_1e_3e_4\}$,$\mathrm{P}_1=\{e_1e_2\}$.

很显然,如果 E_3 中的某条边 e 满足对任意 $P_i\in\mathrm{P}$,若 $e\in P_i$,都有 $P_i\in\mathrm{P}_2$,那么 e 在最终采样得到的可能世界中是否存在都不会影响在最终采样得到的可能世界中 s 与 t 之间的最短路径长度,因为 P 中包含 e 的所有路径都不出现在最终采样得到的可能世界中. 以图 2(a)中的不确定图为例,当 $E_1=\{e_2\}$,$E_2=\{e_3\}$ 时,对于 e_4,其所在路径 $e_1e_3e_4\in\mathrm{P}_2$,因此 e_4 是否存在不会影响 s 与 t 之间的最短路径长度.

于是,在对 g 的可能世界进行随机采样时,如果满足以下两个条件,那么本次随机采样过程可以终止,从而避免对 g 中每条边都进行一次随机采样.

条件 1. 若 $\mathrm{P}_2=\mathrm{P}$,则所有连接 s 和 t 的路径都不可能在最终采样得到的可能世界中,即在最终采样得到的可能世界中,s 和 t 一定是不连通的.

条件 2. 若 P_1 中最短的路径长度为 d,且 E_1 中的边恰能构成一条连接 s 和 t 的长度为 d 的路径,则在最终采样得到的可能世界中一定存在一条连接 s 和 t 的长度最短的路径,且该路径就是由 E_1 构成的可能世界中连接 s 和 t 的最短路径.

使用上述终止条件除了能够减少对边的采样次数外,还可以直接确定在最终得到的可能世界中 s

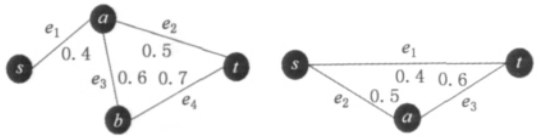
© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

和 t 之间的最短路径长度,从而无需在最终采样得到的可能世界上计算 s 与 t 之间的最短路径长度.

3.2 采样顺序

在对 g 的可能世界进行随机采样时,对 g 的边进行采样的顺序并不影响采样结果,然而不同的采样顺序却影响采样的时间效率.因此,为了尽快满足采样终止条件 1 和条件 2,下面提出两种采样顺序.

顺序 1. 按照边在 P 中所有路径上出现的次数从大到小的顺序进行采样,这是因为如果希望尽快满足采样终止条件 1,那么就需要每次尽可能多地排除 P 中的路径.例如,在图 2(a)的不确定图中,我们将按照 e_1,e_2,e_3,e_4 的顺序采样,此时,采样边数的期望值 $E[Sample1]=0.6\times1+0.4\times0.5\times2+0.4\times0.5\times0.4\times3+0.4\times0.5\times0.6\times4=1.72$.而当我们使用随机采样顺序采样时,通过对全部 4! 种采样顺序依次进行计算,我们可以求出此时采样边数的期望值 $E[Sample]=2.80$.



(a) Instance for sampling onder1 (b) Instance for sampling onder2

Fig. 2 Instances of uncertain graph for random sampling.

图 2 随机采样的不确定图实例

顺序 2. 按照边出现在 P 中的路径的最短长度从小到大的顺序进行采样,这是因为如果希望尽快满足采样终止条件 2,就需要找到当前所有可能路径中最短的,因此,我们会优先考虑 P_1 中的最短路径.例如,在图 2(b)的不确定图中,我们将按照 e_1, e_2,e_3 的顺序采样,此时,采样边数的期望值 $E[Sample2]=0.4\times1+0.6\times0.5\times2+0.6\times0.5\times3=1.9$,而通过依次计算全部 3! 种采样顺序下的采样边数,我们可求出随机采样顺序下采样边数的期望值 $E[Sample]=2.42$.

可见,使用采样顺序 1 和顺序 2 可以有效地减少一次随机采样中的需要采样的边数;而由于不同的采样顺序不会影响采样结果,因此两种采样顺序下随机采样的采样方差是相同的.

下面分析两种采样顺序的效率.对于拓扑结构更加复杂的不确定图,计算按照某一种采样顺序随机采样的期望采样次数需要通过枚举法实现,在最坏情况下,我们需要枚举 $|E|$ 条边可能构成的 $2^{|E|}$ 种组合.因此,通过直接计算两种采样顺序下随机采样的期望采样次数来比较两种采样顺序效率的好坏是不可行的.

通过图 2 中的两个例子我们可以发现,采样顺序 1 适用于 P 中的路径重叠较多的情况,采样顺序 2 适用于 P 中的路径重叠较少的情况.因此,我们自然地想到根据不确定图的拓扑结构分析两种采样顺序效率的好坏.但不幸的是即使对于相同拓扑结构的不确定图,当图中边的概率改变时,两种采样顺序的好坏也会随之改变.

在图 3 所示的不确定图中,从顶点 s 到 t 共有 4 条路径,其中边 e_1 和 e_6 分别在两条路径中出现,其余的边只在一条路径中出现.因此两种采样顺序分别为 $e_1,e_6,e_2,e_3,e_4,e_5,e_7,e_8,e_9,e_{10},e_{11},e_{12}$ 和 $e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8,e_9,e_{10},e_{11},e_{12}$.由于两种采样顺序的后 6 条边相同,因此我们只需要对前 6 条边进行比较.以采样顺序 1 为例:$E[Sample1]=0.5(1-p)\times2+0.25p\times4+0.5p\times0.75\times0.5\times5+(0.25p\times0.75+0.375p\times0.25)\times6=1+2.625p$.类似地,可以求出 $E[Sample2]=3+0.375p$.当 $p=0.89$ 时,$E[Sample1]=E[Sample2]$.当 p 的取值变化时,两种采样顺序期望采样次数的好坏也随之改变.因此,我们不能只根据不确定图的拓扑结构判断两种采样顺序效率的好坏.又因为通过枚举法给出两种采样顺序效率好坏的精确分析是不可行的,实验中,我们将按照两种采样顺序分别进行随机采样并对比采样效率.

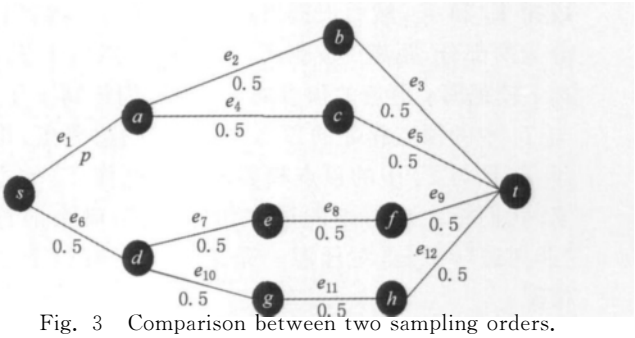


Fig. 3 Comparison between two sampling orders.

图 3 两种采样顺序对比

3.3 计算初始路径集合 P_0 的近似方法

第 3.1 节给出的采样终止条件面临着一个实际问题:在进行随机采样之前,需要枚举出顶点 s 与 t 之间的所有路径 P.文献[15]证明了该枚举问题是一个 $\#P$ 完全问题.因此,本文采用了一种启发式方法来枚举出一部分连接 s 和 t 的路径,称为初始路径集 P_0.然后,在随机采样过程中,将 P_0 看作 P,并根据 P_0 来判断本次采样是否满足采样终止条件.

我们可以采用直观的枚举法计算初始路径集合

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

P_0:①调用 Dijkstra 算法计算连接 s 和 t 的最短路径 P_{11},记 E_{11} 为空,将 P_{11} 加入队列和 P_0;②取出队首路径 P,依次去除 P 上每一条边 e 和与之对应的 E,计算连接 s 和 t 的最短路径 $\{P_{i1}, P_{i2}, \cdots, P_{i|P|}\}$,记 E_{ij} 为 $E \cup e$,然后去掉其中的重复路径和空路径,最后将其余路径按照长度从小到大的顺序加入队列合 P_0;③重复②直至 P_0 的大小不小于给定阈值.

上述枚举法由于会产生重复路径和空路径,效率较差;另一方面,通过枚举法得到的 P_0 不能保证覆盖 E_S 中全部的边,因此会影响采样终止条件的判定. 为保证采样终止条件可以被判断,初始路径集 P_0 应当覆盖 E_S 中所有的边. 下面介绍一种基于生成树的方法来计算满足上述条件的 P_0,其思想如下:首先,在 g 中找到一条从 s 到 t 的最短路径;然后,对 g 中每条未被覆盖的边找到一条包含该边的从 s 到 t 的路径. 重复上述过程直至 E_S 中所有的边都被覆盖为止.

因为上述计算 P_0 的方法是为了近似 P,从而计算出接近采样顺序 1 和顺序 2 的合理的采样顺序,因此 P_0 的大小和生成策略同样不会影响采样方差. 但不同的 P_0 对采样效率会产生影响,当 P_0 增大时 P_0 会更接近 P,采样顺序会更接近采样顺序 1 和顺序 2,采样效率会提高.

在本文中,计算一条从 s 到 t 且包含某条边的路径具体方法如下:分别以 s 和 t 为根建立两棵生成树 T_s 和 T_t,然后根据顶点到 s 和 t 的距离将它们分为两部分,距离 s 较距离 t 更近的顶点属于 T_s,距离 t 较距离 s 更近的顶点属于 T_t. 我们将属于 T_s 且与 T_t 中的顶点相邻的顶点称作 T_s 的边界点,将属于 T_t 且与 T_s 中的顶点相邻的顶点称作 T_t 的边界点. 我们将生成树中的边称作“树边”,剩余的称为“非树边”. 于是,对任意一条边 (u,v),有以下 3 种情况:

1) u 和 v 分别在 T_s 和 T_t 中;

2) u 和 v 在同一棵生成树中,且通过一条“树边”相连;

3) u 和 v 在同一棵生成树中,且通过一条“非树边”相连.

不失一般性,对于第 1 种情况,假设 u 在 T_s 中,v 在 T_t 中,因此只需在 T_s 中找到从 s 到 u 的只包含“树边”的路径 P,在 T_t 中找到从 v 到 t 的只包含“树边”的路径 P' 后,即可构造出从 s 到 t 的路径 $s \xrightarrow{P} u \rightarrow v \xrightarrow{P'} t$.

不失一般性,对于第 2 种情况,假设 u 和 v 都在 T_s 中,且 u 是 v 的父亲节点,则首先在 T_s 中找到一条从 s 到 u 只包含“树边”的路径 P 以及从 v 到某个边界点 r 且只包含“树边”的路径 P';然后在 T_t 中找到从 r 到 t 且只包含“树边”的路径 P''. 因此,从 s 到 t 的路径为 $s \xrightarrow{P} u \rightarrow v \xrightarrow{P'} r \xrightarrow{P''} t$.

对于第 3 种情况,构造从 s 到 t 的路径的方法与第 2 种情况下的方法类似,此处不再赘述.

在图 4 的实例中,$T_s=\{s,a,b,c,u,v\}$,$T_t=\{t,d,e\}$,带下划线的顶点 a,c,d,e,v 是边界点,图 4 中实线表示的是树边,虚线表示的非树边. 此时 u,v 在同一生成树上且通过一条非树边相连,那么通过上述方法,我们可以构造出一条从 s 到 t 的路径:$(s,a)(a,u)\rightarrow(u,v)\rightarrow(v,e)\rightarrow(e,t)$.

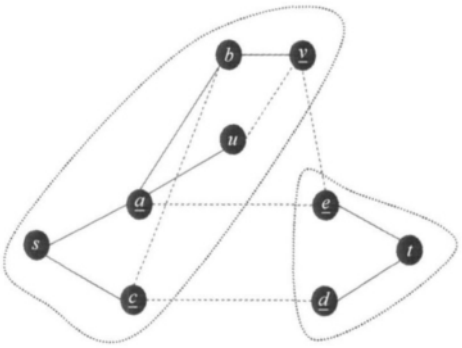


Fig. 4 Instance of spanning tree.

图 4 生成树实例

使用枚举法生成 P_0 的时间复杂度至少为 $O(|P_0|\times|E|+|P_0|\times|V|\log|V|)$;而使用生成树构造 P_0 的时间复杂度为 $O(|E|+|V|\log|V|+|P_0|\log|V|)$. 显然使用生成树构造 P_0 是更有效率的. 在实验部分,我们将进一步对比两种不同的初始路径集合 P_0 下的随机采样效率.

因为 $P_0 \neq P$,所以在随机采样结束时可能出现下面两种错误情况:

情况 1. 采样过程因满足终止条件 1 而终止,但 s 和 t 在采样得到的可能世界中仍然连通.

情况 2. 采样过程因满足终止条件 2 而终止,但在采样得到的可能世界中还存在一条不在 P_0 中且长度更短的路径.

当发生上述两种错误时,我们在采样结束后使用 Dijkstra 算法来计算采样得到的可能世界中从 s 到 t 的最短路径. 同时,我们还需要向 P_0 中加入新的路径,因此 P_0 在采样过程中是不断更新的.

由于初始路径集合在采样过程中是不断更新的,如果我们在每次随机采样结束时都检查是否出现了

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

错误,而每次检查都相当于调用了一次 Dijkstra 算法,那么每次随机采样都需要额外的 $O(|E|+|V|\log|V|)$的时间开销,这与直接随机采样的额外开销相当. 事实上,当初始路径集合 P_0 扩展到一定程度后,往往已经覆盖了 P 中全部的或大部分的路径,显然在这种情况下已经没有必要再做错误检查.

我们记每次随机采样出现错误的概率为 δ,那么连续 N 次采样均未出现错误的概率为$(1-\delta)^N$. 假设 $\delta=0.05$,$N=100$,则连续 100 次不发生错误的概率为 0.006,这是一个小概率事件. 于是,当连续 100 次不发生错误时,我们认为 P_0 不会再扩展,并在以后的随机采样中不再进行错误检查,而这样做导致错误发生的概率仅为 0.006. 这样的判断可以显著地降低错误检查的次数,并且不会给算法带来较大的误差.

3.4 基于对称变量的随机采样算法

对多维随机变量 $X=\{X_1,X_2,\cdots,X_l\}$,X_1,$X_2,\cdots,X_l$ 相互独立,其中 X_i 发生的概率为 Pr_i,我们可以利用均匀分布 $U_1,U_2,\cdots,U_l$,$U_i\sim(0,1)$来模拟 X_i,如下所示:

$$X_i=\begin{cases}1, & U_i<Pr_i;\\ 0, & U_i\geqslant Pr_i.\end{cases} \tag{3}$$

在文献[17]中,已经证明如果存在一个 X_1,$X_2,\cdots,X_l$ 上的单调函数 k(单调递增或递减),那么我们有:

$$Cov(k(U_1,U_2,\cdots,U_l),\ k(1-U_1,1-U_2,\cdots,1-U_l))\leqslant 0. \tag{4}$$

另一方面,我们知道对于两次相同的随机采样 Y_1,Y_2,$Var\left(\frac{Y_1+Y_2}{2}\right)=\frac{Var(Y_1)}{2}+\frac{Cov(Y_1,Y_2)}{2}$. 当 Y_1,Y_2 满足相互独立时,$Cov(Y_1,Y_2)=0$. 于是,我们利用 $Y_1'=k(U_1,U_2,\cdots,U_l)$,$Y_2'=k(1-U_1,1-U_2,\cdots,1-U_l)$来替换 Y_1,Y_2 达到减小采样方差的目标.

首先考虑上述方法是否可以应用到我们的问题上. 不确定图 g 的边是不确定的且相互独立的,因此 g 可以抽象为一个 $|E|$ 维的随机变量,而对一条边 e 的随机采样可以看作是确定采样得到的可能世界 G 中 e 是否存在. 因此一组变量$(U_1,U_2,\cdots,U_{|E|})$就对应了 g 的一个可能世界. 另外,s 和 t 在 g 的一个可能世界 G 中的最短路径长度一定是随着 G 的边集合的增大而减小的,故满足单调的性质. 因此,我们可以将上述方法应用到计算最短路径期望 $E_{s,t}(g)$上.

下面考虑上述方法的效率. 与两次独立的采样相比,使用一组对称变量相当于一次独立的随机采样和一次计算对称变量. 但是由于我们并不是对 E_S 中所有的边依次采样,而是在达到中的两个终止条件之一就终止一次采样,所以就会产生下面的问题:在一次随机采样过程中,我们在采样后得到采样结果$(u_1,u_2,\cdots,u_t)$,使采样终止,那么采样结果$(1-u_1,1-u_2,\cdots,1-u_t)$也一定能够使采样终止吗?如果不能,那么使用对称变量进行随机采样的方差还会好于独立采样的方差吗?

对于第 1 个问题,$(1-u_1,1-u_2,\cdots,1-u_t)$并不一定会使一次采样终止. 以图 1(c)为例,我们对不确定图 $E_{s,t}(g)$的边(s,a),(s,b),(s,c)作采样,由于 $u_{(s,a)}>Pr(s,a)$,根据式(4),$(s,a)\in E_2$,同理(s,b),$(s,c)\in E_2$,采样终止条件 1 满足,采样结束. 但是,如图 1(d)所示,我们发现计算对称变量$(1-u_{(s,a)},1-u_{(s,b)},1-u_{(s,c)})$并不能使一次采样结束. 对于第 2 个问题,我们仍然可以保证使用对称变量进行随机采样的方差会好于独立随机采样. 下面我们首先将证明独立随机采样和对称变量随机采样都是期望最短距离的无偏估计,然后再证明对称变量随机采样的方差好于独立随机采样的方差.

使用了终止条件和采样顺序的 n 次独立随机采样与 n 次直接随机采样相比,采样结果相同,采样边数减少,而采样结果可表示为

$$\hat{E}_B=\frac{\sum\limits_{dis_{s,t}(G_i)<\infty} dis_{s,t}(G_i)}{\sum\limits_{dis_{s,t}(G_i)<\infty} 1}. \tag{5}$$

定理 2. n 次独立随机采样结果的期望 $E(\hat{E}_B)=E_{s,t}(g)$.

证明. 在每次独立随机采样得到的可能世界 G 中,顶点 s 和 t 之间的最短距离可能是$\{d_1,d_2,\cdots,d_m,\infty\}$,且 $Pr(dis_{s,t}(G)=d_i)=Pr(d_i)$. 又因为当顶点 s 和 t 不连通时,一次随机采样结果将不会计入 n 次采样的平均值中,因此不会影响到 $\hat{E}_B$ 的取值. 因此,我们只需考虑顶点 s 和 t 之间连通的情况. 我们知道,当顶点 s 和 t 连通时,一次随机采样结果的期望 $E(\hat{E}_j)$满足:

$$E(\hat{E}_j)=\sum_{i=1}^{m} d_i\times\frac{Pr(dis_{s,t}(G)=d_i)}{1-Pr(\infty)}=E_{s,t}(g).$$

所以对于 n 次独立随机采样中有 K 次(假设为第 $i_1,i_2,\cdots,i_k$ 次)满足顶点 s 和 t 连通的情形,其中

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$K=\{0,1,2,\cdots,n\}$，我们有 $E(\hat{E}_B)=E(\sum_{j=1}^{K}\hat{E}_{ij}/K)=(E\sum_{j=1}^{K}\hat{E}_{ij})/K=E_{s,t}(g)$. 证毕.

定理 3. n 次对称变量随机采样结果的期望 $E(\hat{E}_A)=E_{s,t}(g)$.

证明. 由于在对称变量随机采样中，每一次随机采样与直接随机采样均满足同分布，因此满足期望相同. 所以 n 次对称变量随机采样结果的期望满足 $E(\hat{E}_A)=E_{s,t}(g)$. 证毕.

在 n 次独立随机采样中，由于只有当采样结果为有穷值时，采样结果才会被计入 n 次的平均值，因此，在计算采样方差时，我们只是计算其中 K 次满足顶点 s 和 t 连通的采样.

在对比 n 次独立随机采样和 n 次对称变量随机采样时，我们假设两种采样中分别有 K,K' 次满足顶点 s 和 t 连通. 事实上，由于每一次对称变量采样与独立随机采样均满足同分布，因此 K' 的期望与的 K 期望相同，即 $E[K]=E[K']=n\times(1-Pr(\infty))$. 此时 n 次独立随机采样的方差为 $\frac{1}{K}Var(Y_1)$，当 K 越大时采样方差越小. 为公平起见，我们在对比两种采样方法的方差时，将假设 n 次随机采样中使得顶点 s 和 t 连通的采样次数相同，即 $K=K'$.

注意到，K' 次对称变量采样中，一部分采样是作为一组对称变量成对出现的，而另一部分则是独立随机采样. 因此，我们只需要比较 K' 中 q 次对称变量随机采样与 q 次独立随机采样的方差即可.

定理 4. 使用对称变量进行随机采样 Y_1',Y_2' 的采样方差一定不会超过两次独立随机采样 Y_1,Y_2 的方差.

证明. 设第 1 次采样 Y_1' 中，我们对 $(u_1,u_2,\cdots,u_t)$ 采样后，采样终止，在第 2 次采样 Y_2' 中，我们首先计算 $(1-u_1,1-u_2,\cdots,1-u_t)$，随后又采样 s 次使得采样终止. 那么第 2 次采样可记作 $(1-u_1,1-u_2,\cdots,1-u_t,u_{t+1},u_{t+2},\cdots,u_{t+s})$. 由于第 1 次采样在 t 次后便终止，即其余的边是否存在对第 1 次的采样结果都不会产生影响，我们有 $k(u_1,u_2,\cdots,u_t)=k(u_1,u_2,\cdots,u_t,1-u_{t+1},1-u_{t+2},\cdots,1-u_{t+s})$. 于是我们构造了一组新的对称变量. 根据式(4)可知，这两次采样的协方差非正. 那么使用对称变量后两次采样的方差 $Var\left(\frac{Y_1'+Y_2'}{2}\right)\leqslant\frac{Var(Y_1')}{2}=\frac{Var(Y_1)}{2}=Var\left(\frac{Y_1+Y_2}{2}\right)$. 证毕.

定理 5. q 次对称变量随机采样结果的方差一定小于 q 次独立随机采样结果的方差.

证明. 对于 q 次独立随机采样 $Y_1,Y_2,\cdots,Y_q$，由于其中任意两次采样 Y_i,Y_j 的协方差均为 0，故有：$Var\left(\frac{1}{q}(Y_1+Y_2+\cdots+Y_q)\right)=\frac{1}{q}Var(Y_1)$. 对于 q 次对称变量随机采样，我们可将其分为 $q/2$ 个组，$\{Y_{11}',Y_{12}'\},\{Y_{21}',Y_{22}'\},\cdots,\{Y_{\frac{q}{2}1}',Y_{\frac{q}{2}2}'\}$，对于不同组内的任意两次随机采样，它们的协方差 $Cov(Y_{ik}',Y_{jk'}')=0,i\neq j,i,j=\{1,2,\cdots,\frac{q}{2}\},k,k'=\{1,2\}$. 那么：$Cov(Y_{i1}'+Y_{i2}',Y_{j1}'+Y_{j2}')=E[Y_{i1}'+Y_{i2}'][Y_{j1}'+Y_{j2}']-E[Y_{i1}'+Y_{i2}']E[Y_{j1}'+Y_{j2}']=\sum_{k=1}^{2}\sum_{l=1}^{2}(E[Y_{ik}'\times Y_{jl}']-E[Y_{ik}']\times E[Y_{jl}'])=0$. 于是我们有 $Var(\frac{1}{q}((Y_{11}'+Y_{12}')+(Y_{21}'+Y_{22}')+\cdots+(Y_{\frac{q}{2}1}'+Y_{\frac{q}{2}2}')))\leqslant\frac{1}{2q}Var(Y_1+Y_2)=\frac{1}{2q}\times 2Var(Y_1)=\frac{1}{q}Var(Y_1)$. 证毕.

对于直接随机采样，当我们增大采样次数 n 时，因为 $E[K]=n\times(1-Pr(\infty))$，所以 K 将有较大的概率增大，因此 n 次直接随机采样的方差 $\frac{1}{K}Var(Y_1)$ 会减小. 对于对称变量随机采样，当采样次数 n 增大时，同理，K' 将有较大的概率增大，此时其采样方差的下界为 $\frac{1}{K'}Var(Y_1)$，因此其方差也会以较大的概率减小.

算法 1 给出了一次随机采样过程中对一条边 e 随机采样后的具体操作：

算法 1. $Edge(g,s,t,e,u)$.

参数 g：不确定图；

参数 s,t：不确定图上两个顶点；

参数 e：需要采样的边；

参数 u：边 e 的随机采样值.

```
① if u<Pr(e) {e 出现在 E1 中} then
② E1=E1∪e;
③ if E1 contains a shortest path p{E1 包含当前最短路径 p} then
④ return p.len();
⑤ end if
⑥ end if
⑦ else {e 出现在 E2 中} then
⑧ E2=E2∪e;
```

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

⑨ if E_1 contains a shortest path p {E_2 包含当前最短路径 p} then
⑩ return $p.len()$;
⑪ end if
⑫ if E_2 contains a (s,t)-cut {E_2 中 s,t 不可达} then
⑬ return 0;
⑭ end if
⑮ end if
⑯ return ①;

首先,我们判断边 e 的随机采样值与 $Pr(e)$ 的大小关系(行①),根据结果的不同,将 e 加入 E_1 或 E_2(行②,⑧),在 E_1 或 E_2 更新后,我们需要判断此时是否已经满足了采样终止条件 1 和条件 2(行③、⑨、⑫). 当排除了所有可能路径时,返回 0,采样结束;而当找到了当前可能的最短路径时,返回此路径长度,采样结束;否则,返回①,继续采样下一条边.

算法 2 描述了一次对称变量随机采样的实现:

算法 2. $EA(g,Q,s,t,U)$.

参数 g: 不确定图;

参数 Q: 采样队列;

参数 s,t: 不确定图上两个顶点;

参数 U: 一次独立采样中对 Q 中一部分边的随机采样值.

① while $U\neq\varnothing$
② $e=Q.top()$;
③ $length=Edge(g,s,t,e,1-U(e))$
④ if $length\neq-1$
⑤ return $length$;
⑥ end if
⑦ end while
⑧ while $Q\neq\varnothing$
⑨ $e=Q.top()$;
⑩ 随机生成满足(0,1)均匀分布的 u;
⑪ $length=Edge(g,s,t,e,u)$;
⑫ if $length\neq-1$
⑬ return $length$;
⑭ end if
⑮ end while

算法 2 首先计算了上一次独立随机采样时所采样过边的对称变量,如果不能使采样终止,则继续按照采样队列的顺序进行随机采样直至终止.

4 实验结果

在实验部分,我们在真实数据上对算法的准确性和效率进行了考察,将不同随机采样方法计算的估计值和时间开销作了对比. 下面首先介绍实验中使用的随机采样方法;然后介绍了实验考察的几种测度,包括运行时间、相对误差和采样方差;随后介绍了实验环境、实验数据;最后给出了实验对比及分析.

在实验中,使用了 5 种采样方法估计期望最短距离:1) $\hat{E}_B$:直接随机采样,采样结束后使用 Dijkstra 算法来计算最短距离;2) $\hat{E}_1$:按照第 1 种采样顺序随机采样,返回期望最短距离 $E_{s,t}$ 的估计值,采样结束调用 Dijkstra 算法进行错误检查,如果连续 N 次未发生错误此后不再进行错误检查(实验中,我们对每一个不确定图 g 作 10 000 次采样,并设定 $N=100$,后面所提到的错误检查与 $\hat{E}_1$ 的错误检查相同);3) $\hat{E}_2$:按照第 2 种采样顺序进行随机采样计算 $E_{s,t}$ 的近似值,采样结束进行错误检查;4) $\hat{E}_{1A}$:在 $\hat{E}_1$ 的基础上引入对称变量进行随机采样,返回 $E_{s,t}$ 的估计值,采样结束时进行错误检查;5) $\hat{E}_{2A}$:在 $\hat{E}_2$ 的基础上引入对称变量给出 $E_{s,t}$ 的近似值. 采样结束时进行错误检查.

考察采样方法的准确性时,我们从相对误差和方差两个方面进行对比. 计算相对误差 $\varepsilon=\frac{|E_{s,t}-\hat{E}_{s,t}|}{E_{s,t}}$ 时,我们需要通过枚举出不确定图 g 的全部可能世界准确计算出期望最短距离 $E_{s,t}$,因此在实验中我们不能使用边数很大的数据;另一方面,如果实验数据很小,那么我们完全可以以较小的时间代价使用枚举法准确地计算除期望最短距离. 因此在实验中,我们使用了边数在 15~55 之间的连通子图. 随机采样方差 $\sigma=\frac{\sum_{\hat{E}_i<\infty}(E_i-\bar{E})^2}{K}$,其中 K 表示在全部的 n 次采样中共有 K 次满足 s 和 t 连通,$\bar{E}=\frac{\sum_{\hat{E}_i<\infty}\hat{E}_i}{K}$. 由于 $\hat{E}_B,\hat{E}_1,\hat{E}_2$ 3 种方法只是采样顺序不同,因此它们的采样结果是一致的,在考察准确性时我们只需要对 $\hat{E}_B,\hat{E}_{1A},\hat{E}_{2A}$ 作出对比即可.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

上述实验使用 C++实现，在双核 3.0 GHz Pendium® D CPU、3.25 GB 内存、Windows XP 系统的台式机上运行.

我们的全部实验在真实的不确定图上进行. 不确定图来自于欧洲分子生物学实验室的 STRING 数据库(http://string-db.org). 我们共使用了 3 个生物物种的蛋白质交互(PPI)网络，表 3 给出了实验中使用到的蛋白质交互网络的基本特性：

Table 3 Characteristics of PPIs in Experiments

表 3 实验中使用的不确定图(PPI)的基本特性

Species	Vertex No	Edge No	Average Probability Edges
E. coli	176	741	0.342
D. melanogaster	102	958	0.414
M. musculus	107	329	0.400

图 5(a)给出了不同采样方法的相对误差，我们将实验数据按照边数分为 4 组. 使用对称变量的两种采样方法的误差均好于直接采样，这表明相对误差的好坏与采样方差是一致的. 当边数增大时，3 种采样方法的相对误差都增大. 这是因为当边数增大时，连接 s 和 t 路径的数量会增加，在采样次数有限的情况下，误差会增大. 图 5(a)中顶点 s 和 t 的期望最短距离在 2～4 之间. 我们进一步考察了 s 和 t 的期望最短距离在 4～6 之间的情况，如图 5(b)所示，可以发现，此时 3 种方法的相对误差均有所增加，但相对趋势仍与图 5(a)保持一致. 这是因为 s 和 t 相距越远，不同长度的路径就会越多，因此相对误差会增大，但是 3 种采样方法的优劣性不会改变.

图 5(c)给出了 $\hat{E}_B$，$\hat{E}_{1A}$，$\hat{E}_{2A}$ 3 种采样方法的采样方差对比. 以直接采样的方差为基准值 1，其余采样方法的方差为 σ/σ_B. 正如我们分析的那样，使用了对称变量后的采样方差要好于直接采样方法的方差. 而使用不同采样顺序的采样方法的方差并没有相对优劣之分. 当边数增大时，使用对称变量的采样方法的优势变得更加显著. 图 5(c)中是 s 和 t 的期望最短距离在 2～4 之间的实验结果，图 5(d)中是期望最短距离在 4～6 之间的实验结果. 可见相对方差在不同的期望最短距离下没有固定的变化趋势，这说明 s 和 t 之间的距离对 3 种方法的影响程度是不确定的.

我们知道在未使用对称变量时，两种采样顺序的方差是相同的，因此 $\hat{E}_{1A}$ 和 $\hat{E}_{2A}$ 两种采样方法准确性上的差别是由于对称变量采样方法的协方差不同导致. 根据文献[17]的结果可知，当不确定图中边的概率越接近 0.5 且边的存在会影响顶点 s 和 t 的最短距离时，协方差越小此时采样方差和相对误差也越小. 因此不同采样顺序由于优先采样边的概率和

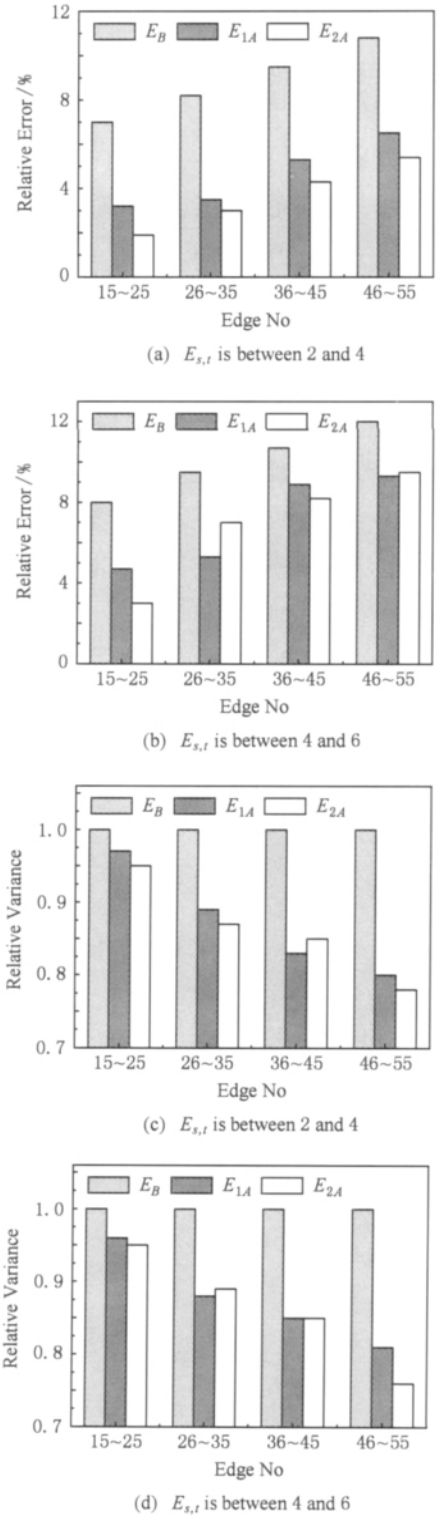


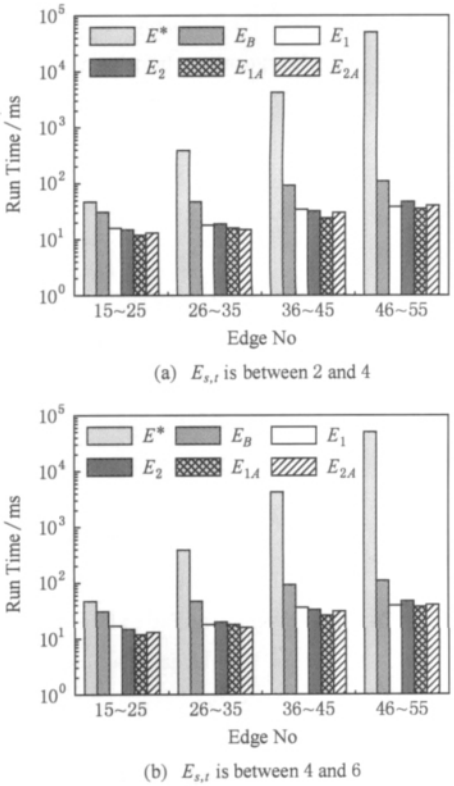
Fig. 5 Experimental results of accuracy.

图 5 算法准确性实验结果

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

拓扑结构不同导致准确性不同.因为我们使用的实验数据是根据真实数据随机生成的连通子图,所以实验中两种采样顺序下的采样方差并没有相对好坏的区分.

图 6(a)和 6(b)给出了各种采样方法的运行时间对比.E^* 表示通过枚举计算期望最短距离的方法.可以发现当边数很小时,E^* 的时间代价是可以接受的,而边数增加时 E^* 的时间代价呈指数型增长.使用了采样队列的采样方法因为无需对所有的边采样,时间开销与直接随机采样相比下降 50%以上;在使用对称变量后,由于不会带来额外开销,运行时间又比未使用对称变量时有所下降.与相对误差、采样方差一样,两种采样顺序在运行时间上也没有明显的优劣之分.s 和 t 之间的距离对运行时间的影响也是微乎其微的.



(a) $E_{s,t}$ is between 2 and 4

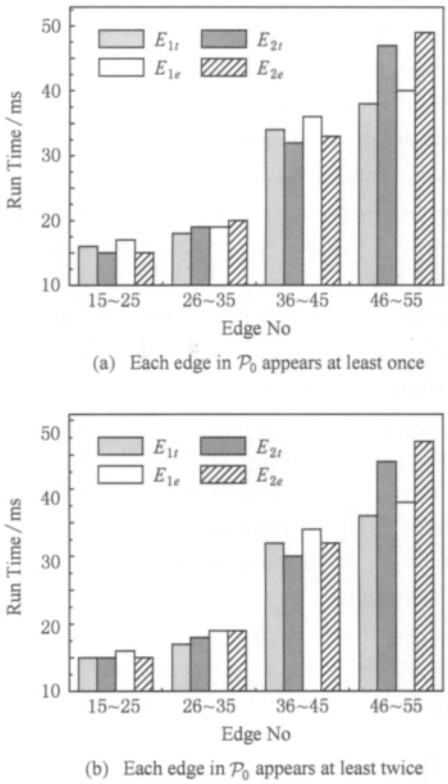
(b) $E_{s,t}$ is between 4 and 6

Fig. 6 Experimental results of run time.

图 6 算法运行时间实验结果

图 7 对比了不同的初始路径集合 P_0 下随机采样方法的运行时间.因为初始路径集合的不同实际就是采样顺序的不同,而在图 5 的实验中不同采样顺序下对称变量采样的方差没有相对好坏的区分,所以我们只对 $\hat{E}_1$ 和 $\hat{E}_2$ 两种采样方法进行了考察.我们仍然按照边数大小进行了 4 组实验.因为 s 和 t 之间不同的距离对运行时间的影响很小,所以不再进行对比;又因为 P_0 的大小会影响采样方法的效率,所以我们对比了不同大小的 P_0 下随机采样方法的运行时间.

在计算 P_0 时,我们首先使用生成树的方法计算 P_0,然后根据 P_0 的大小使用枚举法计算出另一个初始路径集合 P_0,如图 7(a)所示;为了得到不同大小的 P_0,我们在第 2 次使用生成树法计算 P_0 时,将循环终止条件改为"每条边都被覆盖至少两次",从而得到更大的初始路径集合,如图 7(b)所示.



(a) Each edge in $\mathcal{P}_0$ appears at least once

(b) Each edge in $\mathcal{P}_0$ appears at least twice

Fig. 7 Experimental results of run time under different initial path set.

图 7 不同初始路径集合下算法运行时间

图 7(a)和 7(b)中的结果表明,生成树法计算得到的初始路径集合 P_0 下的随机采样方法($\hat{E}_{1t}$ 和 $\hat{E}_{2t}$)的运行时间要好于枚举法得到的 P_0 下的采样方法($\hat{E}_{1e}$ 和 $\hat{E}_{2e}$)的运行时间.而当 P_0 增大时,采样方法的运行时间会随之减少,这与我们的分析是一致的.

5 结 论

本文研究了不确定图上的期望最短距离问题.我们给出了问题的定义,分析了问题的时间复杂度,

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

并使用了随机采样的方法计算问题的近似解;我们还使用了过滤条件来提高随机采样的效率,并通过对称变量的思想改进了随机采样方法的采样方差;最后我们通过真实数据上的实验对比验证了所提出方法的有效性和准确性.在今后的工作中,我们将进一步从理论上探讨不同的采样顺序对随机采样方法效率和准确性的影响以及初始路径集合大小对随机采样方法效率的影响.

参 考 文 献

[1] Asthana S, King O K, Gibbons F D, et al. Predicting protein complex membership using probabilistic network reliability [J]. Genome Research, 2004, 14(6): 1170-1175

[2] Swamynathan G, Wilson C, Boe B, et al. Do social networks improve E-commerce?: A study on social marketplaces [C] // Proc of the 1st Workshop on Online Social Networks. New York: ACM, 2008: 1-6

[3] Ghosh J, Ngo H Q, Yoon S, et al. On a routing problem within probabilistic graphs and its application to intermittently connected networks [C] //Proc of INFOCOM'07. Piscataway, NJ: IEEE, 2007: 1721-1729

[4] Pandurangan G, Raghavan P, Upfal E. Building low-diameter peer-to-peer networks [J]. IEEE Journal on Selected Areas in Communications, 2003, 21(6): 995-1002

[5] Aggarwal C C. Advances in Database System [M]. Berlin: Springer, 2009

[6] Yuan Y, Chen L, Wang G. Efficiently answering probability threshold-based shortest path queries over uncertain graphs [C] // Proc of DASFFA'10. Berlin: Springer, 2010: 155-170

[7] Li J, Zou Z, Gao H. Mining frequent subgraphs over uncertain graph databases under probabilistic semantics [EB/OL]. (2012-02-28) [2012-06-20]. http://www.springerlink.com/content/6666pl933v743g38/

[8] Jin R, Liu L, Ding B, et al. Distance-constraint reachability computation in uncertain graphs [J]. PVLDB, 2011, 4(9): 551-562

[9] Colbourn C J. The Combinatorics of Network Reliability [M]. New York: Oxford University Press, 1987

[10] Yuan Ye, Wang Guoren. Answering probabilistic reachability queries over uncertain graphs [J]. Chinese Journal of Computers, 2010, 33(8): 1378-1386 (in Chinese)
(袁野,王国仁. 面向不确定图的概率可达查询[J]. 计算机学报, 2010, 33(8): 1378-1386)

[11] Floyd R. Algorithm 97: Shortest path [J]. Communications of the ACM, 1962, 5(6): 345

[12] Yildirim H, Chaoji V, Zaki M J. Grail: Scalable reachability index for large graphs [J]. PVLDB, 2010, 3(1): 276-284

[13] Jin R, Xiang Y, Ruan N, et al. 3-hop: A high-compression indexing scheme for reachability query [C] //Proc of the 35th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 813-826

[14] Valiant L G. The complexity of enumeration and reliability problems [J]. SIAM Journal on Computing, 1979, 8(3): 410-421

[15] Ball M O. Computational complexity of network reliability analysis: An overview [J]. IEEE Trans on Reliability, 1986, 35(3): 230-239

[16] Thompson S. Sampling [M]. 2nd ed. New York: Wiley, 2002

[17] Ross S M. Introduction to Probability Models [M]. New York: Academic Press, 2007

Li Mingpeng, born in 1989. PhD candidate. His current research interests include querying and mining uncertain graph data.

Zou Zhaonian, born in 1979. PhD, lecturer, member of China Computer Federation. His main research filed is graph data mining.

Gao Hong, born in 1966. PhD, professor and PhD supervisor, senior member of China Computer Federation. Her research interests include data mining and wireless sensor networks.

Zhao Zhengli, born in 1992. Undergraduate. His research include graph querying, artificial intelligence.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net