7L-05

層化抽出法を用いた1対全ノードに対するs-t信頼性推定の高速化

柳澤 隼也† 塩川 浩昭‡

† 筑波大学情報学群情報科学類 ‡ 筑波大学計算科学研究センター

1 はじめに

s-t信頼性は不確実グラフにおいて2ノード間の接続確率を評価する重要な指標のひとつである．s-t信頼性の計算は♯P困難であるため，近似解を計算する手法が提案されている．しかしながら，不確実グラフにおいてtop-k検索やクラスタリングなどの分析処理を実行するためには，1対全ノードに対するs-t信頼性計算を繰り返し実行する必要がある．この処理は，これまで提案されてきた近似計算手法を用いた場合でも，効率的な処理が難しい．

そこで本研究では，1対全ノードに対する高速なs-t信頼性推定手法を提案する．提案手法では幅優先探索とサンプリングに基づくs-t信頼性推定手法を統合する．これにより，少ない計算回数で高精度に1対全ノードに対するs-t信頼性を推定する．

2 不確実グラフとs-t信頼性

不確実グラフは$\mathcal{G} = (V, E, p)$で与えられる．VとEはそれぞれノード集合と有向エッジ集合である．pは確率関数で$p: E \to (0,1]$である．不確実グラフ$\mathcal{G}$は，結果が$\mathcal{G}$の部分グラフである確率空間とみなすことができ，各エッジの存在は，他のエッジの存在とは独立に確率$p(e)$で決定する．

各部分グラフは可能世界と呼ばれ，$\mathcal{G}$における可能世界を$G_i = (V, E_i)$と定義する．ここでE_iはG_iに存在するエッジの集合で$E_i \subseteq E$である．また便宜上，G_iが$\mathcal{G}$の可能世界である時，$G_i \sqsubseteq \mathcal{G}$と表す．$m$個のエッジを持つ不確実グラフ$\mathcal{G}$を考えた時，$\mathcal{G}$における可能世界は$2^m$通り生成しうる．可能世界$G_i$の生成確率を$\Pr(G_i)$で表すと確率$\Pr(G_i)$は以下で定義できる．

$$\Pr(G_i) = \prod_{e \in E_i} p(e) \prod_{e \in E \setminus E_i} (1 - p(e)) \quad (1)$$

s-t信頼性は不確実グラフ$\mathcal{G}$において始点ノード$s \in V$から終点ノード$t \in V$が到達可能である確率を計算したものである．ある可能世界Gについて指示関数I_Gを定義し，始点ノードsから終点ノードtが到達可能であるとき$I_G(s,t)$は1となり，そうでないときは0となるものとする．このとき不確実グラフに$\mathcal{G}$おける始点ノードs・終点ノードt間のs-t信頼性$R(s,t)$は次のように計算する．

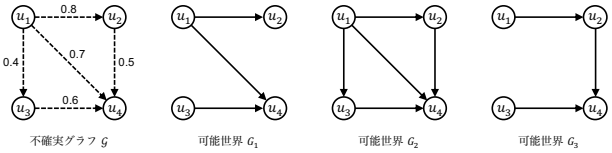

図 1 不確実グラフ

$$R(s,t) = \sum_{G \sqsubseteq \mathcal{G}} I_G(s,t) \cdot \Pr(G) \quad (2)$$

しかしながら，s-t信頼性の計算は♯P完全問題である．したがって，一般には可能世界Gをサンプリングすることで近似的にs-t信頼性を計算する．

3 関連研究

最も標準的な手法はモンテカルロ法を用いた近似アルゴリズム[3] (以降，MC) である．式(1)に示す確率の分布で$\mathcal{G}$からランダムにサンプリングしたK個の可能世界$G_1, G_2, \ldots, G_K$を用いた時，s-t信頼性の推定値$\hat{R}(s,t)$は次のように計算する．

$$\hat{R}(s,t) = \frac{1}{K} \sum_{i=0}^{K} I_{G_i}(s,t) \cdot \Pr(G_i) \quad (3)$$

1対全ノードに対するs-t信頼性の推定では，全てのノードvがノードsから到達回数を保持する変数$Reach[v]$を持つ．各可能世界でノードsから到達可能なvについて$Reach[v]$をインクリメントし，K個の可能世界の探索終了時，ノードsからノードvへのs-t信頼性の推定値は$\hat{R}(s,v) = Reach[v]/K$によって求めることができる．

Zhu [2] らはMCでの各可能世界探索時のトラバース処理に大きな重複があることに着眼し，可能世界をビット表現で持つことで一回の幅優先探索で1対全ノードに対するs-t信頼性を推定することができる手法BFS Sharing (以降，BFSS) および事前サンプリング手法を提案した．しかし，MCおよびBFSSでは推定精度を確保するためには多くのサンプル数を必要とするという問題点がある．

Fast reliability estimation for one-to-all nodes using stratified sampling

Junya Yanagisawa †, Hiroaki Shiokawa ‡

† College of Information Science, University of Tsukuba

‡ Center for Computational Sciences, University of Tsukuba

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

4 提案手法

本研究では，BFSS の事前サンプリングとして不確実グラフにおける層化抽出法 [1] を統合したアルゴリズムを提案する．提案手法では少ないサンプル数で分散の少ない高精度な推定値を高速に求めることを可能とする．提案手法は事前サンプリングと推定部から成る．

4.1 事前サンプリング部

事前サンプリング部では BFSS に用いるエッジのビットベクトルの構築を行う．

提案手法は確率空間 $\mathcal{G}$ を $r+1$ 個の確率部分空間 $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_r$ に分割する．始点ノード s から幅優先探索で r 本のエッジの表 1 に示すようなパタンでサンプリングすることで重複しない確率部分空間に分割する．表 1 における 1 および 0 はサンプリングするおよびしないことを示す．$*$ は部分空間がエッジをサンプリングするか決定していないことを示す．提案手法ではこの展開を再帰的に行い，それぞれの部分空間の確率に比例してサンプル数を割り当てる．この再帰的処理は閾値 θ によって制御され，割り当てられたサンプル数が θ より小さい時，可能世界を割り当てられたサンプル数だけ生成する．

4.2 推定部

推定部では事前サンプリング部で得られた K 個の可能世界を用い，1 対全ノードに対する s-t 信頼性を BFSS を用いて同時に計算する．

まず，BFSS は不確実グラフ $\mathcal{G}$ の各エッジ e に K ビットベクトル $\mathcal{B}_e$ を割り当てる．エッジ $e \in E_i$ の場合，i 番目のビット $\mathcal{B}_e[i]$ は 1 であり，それ以外では $\mathcal{B}_e[i]$ は 0 である．同様に，各ノード $v \in V$ には，ノード s からノード v の到達可能性を保持するために K ビットベクトル $\mathcal{B}_v$ を持つ．i 番目のビット $\mathcal{B}_v[i]$ は、ノード v が可能世界 G_i のノード s から到達可能な場合にのみ 1 であり，それ以外の場合は $\mathcal{B}_v[i]$ は 0 である．そのため，全ての i について $\mathcal{B}_s[i] = 1$ かつ $\mathcal{B}_v[i] = 1\ (v \neq s)$ に設定しておく．

次に BFSS は，ノード s から幅優先検索(BFS)を行う．BFS はノード u からノード v に到達すると，ビット単位の AND 演算によって $\mathcal{B}_v$ を更新する．つまり $e = (u, v)$ とした時，$\mathcal{B}_v = \mathcal{B}_u\ AND\ \mathcal{B}_e$ で計算できる．BFSS は，ノードのすべてのビットベクトルが更新されるまで BFS を行う．

最後に s-t 信頼性を推定する．$\|\mathcal{B}_v\|_1$ は $\mathcal{B}_v$ における 1 の数を意味するので，ノード s からノード v への s-t 信頼性の推定値は $\hat{R}(s, v) = \|\mathcal{B}_v\|_1 / K$ によって求めることができる．

5 評価実験

提案手法を Last.fm から求めたソーシャルグラフを用いて評価した．ノード数とエッジ数はそれぞれ 6,899，23,696 である．ランダムに取得した 2 ホップ離れた 2 ノードを 100 セット用意し，収束基準 $\rho_K = \frac{V_K}{R_K} < 0.001$ で収束とし，収束するまでサンプル数を 250 ずつ増やす．V_K は各セットでの 100 回の実行結果による不偏分散の平均値で，R_K は各セットでの 100 回の実行結果による平均信頼性の平均値である．本実験は収束時の実行時間で比較を行った．パラメータは $r = 50, \theta = 5$ を用いた．実験は Intel Xeon 3.5GHz , 128GB RAM の Linux サーバを用いた．実験結果を表 2,3 に示す．

表 1. 確率空間のサンプリングパターン

部分空間	e_1	e_2	e_3	$\cdots$	e_r	e_{r+1}	$\cdots$	e_m
$\mathcal{G}_0$	0	0	0	$\cdots$	0	$*$	$\cdots$	$*$
$\mathcal{G}_1$	1	$*$	$*$	$\cdots$	$*$	$*$	$\cdots$	$*$
$\mathcal{G}_2$	0	1	$*$	$\cdots$	$*$	$*$	$\cdots$	$*$
$\mathcal{G}_3$	0	0	1	$\cdots$	$*$	$*$	$\cdots$	$*$
$\vdots$				$\ddots$				
$\mathcal{G}_r$	0	0	0	$\cdots$	0	$*$	$\cdots$	$*$

表 2 より提案手法が最も高速であることがわかる．online はサンプリングを実行時に行った場合の結果である．online の場合でも提案手法は既存手法より高速である．表 2 の事前サンプリングについても，収束に必要なサンプル数が少ないため高速であることがわかる．

表 2. 実行時間の比較

手法	サンプル数	実行時間 (ms)
MC	1000	2396.552
BFSS	1000	199.245
提案手法	250	68.254
BFSS (online)	1000	441.689
提案手法 (online)	250	172.769

表 3. 事前サンプリングの比較

手法	サンプル数	実行時間 (ms)
BFSS	1000	714.715
提案手法	250	333.090

まとめ

BFSS の事前サンプリングとして不確実グラフにおける層化抽出法を用いることで 1 対全ノードに対する s-t 信頼性推定を高速化する手法を提案した．実データを用いた評価実験により，提案手法が既存手法より高速に計算できることが示された．

参考文献

[1] R. Li et al., “Recursive Stratfied Sampling: A New Framework for Query Evaluation on Uncertain Graphs.” IEEE TKDE, 28(2):468–482, 2016.

[2] R. Zhu, Z. Zou, and J. Li. “Top-k Reliability Search on Uncertain Graphs. ” In Proc. ICDM, 2015

[3] G. S. Fishman. “A Comparison of Four Monte Carlo Methods for Estimating the Probability of s-t Connectedness.” IEEE Tran. Rel., 1986.

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.