2Q-01

属性付きグラフに対する効率的なコミュニティ問合せ処理

真次 彰平† 塩川 浩昭‡ 北川 博之‡

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

グラフ分析手法の一つであるコミュニティ問合せは，グラフが表現するエンティティの関係性の中からクエリに対して適切なノード集合(コミュニティ)を見つけ出す技術である．一般的にグラフはノードに属性値を持つ場合がある．属性付きグラフ [1]におけるコミュニティ問合せは与えられたクエリノードとクエリ属性に対し，クエリノード近傍のクエリ属性に関係のあるコミュニティを見つけ出すことが目的となる．

属性付きグラフに対するコミュニティ問合せ処理手法として LocATC [2]を提案されている．LocATC はコミュニティ問合せを Attribute Truss Community(ATC)問題として定式化し，ATC 問題を最適化するコミュニティを探索する手法を提案している．ATC 問題では，解となるコミュニティに頑健な truss 構造が一定割合以上含まれることを前提としている．しかし，実世界のグラフは多様な構造を持つことから，必ずしもこの前提が成立しない．ゆえに，LocATC により取得されるコミュニティは柔軟性を欠き，実データにおいて十分な精度を得られないという問題がある．

本研究では属性付きグラフにおけるコミュニティ問合せの精度向上に取り組む．従来手法 LocATC より柔軟なコミュニティ構造を取得するために，本研究は ATC 問題の構造的制約条件を緩和した緩和 ATC 問題を定義する．緩和 ATC 問題に対して，ビームサーチを用いたコミュニティ問合せ処理手法を提案し，従来技術よりも高精度なコミュニティ問合せを実現する．

2 前提知識

本研究の前提となる知識を概説する．属性付きグラフに対するコミュニティ問合せでは，簡潔なモデルとして連結なグラフ$G(V,E)$として与える．また，全ての属性の集合を$\mathcal{A}$とし，各ノード$v \in V$の持つ属性値の集合を $\mathrm{attr}(v) \subseteq \mathcal{A}$で表す．クエリ$Q=(v_q,W_q)$が与えられたとき，クエリノード$v_q$を含み，クエリ属性集合$W_q$に対して最も適切な部分グラフ$H$を検索することを考える．

Efficient Community Query Processing for Attributed Graph
Shohei Matsugu†, Hiroaki Shiokawa‡ and Hiroyuki Kitagawa‡
†College of Information Science, University of Tsukuba
‡Center for Computational Sciences, University of Tsukuba

2.1 ATC 問題

ATC 問題を定義するにあたり，部分グラフHを構造と属性の両観点から評価をする．まず構造を評価する指標として(k, d)-truss を定義する．

定義 1 [(k, d)-truss]．クエリノードv_qに対して部分グラフHが(k, d)-truss であるとは，$v_q \in H$である部分グラフHが k-truss かつ$dist_H(H,v_q) \geq d$を満たすことを言う．Hが k-truss であるとはH内の任意のエッジが k-2 個以上の三角形に属すことを言う．また$dist_H(H,v_q)$はv_qとHの任意のノード間における距離の最大値である．

次に属性を評価する指標として，Attribute Score Function を定義する．

定義 2 [Attribute Score Function]．部分グラフHとクエリ属性集合W_qが与えられたとき，Attribute Score Function $f(H,W_q)$を，$f(H,W_q)=\sum_{w\in W_q}\theta(H,w)\times score(H,w)$と定義する．ただし，$\theta(H,w)=\frac{|V_w\cap V(H)|}{|V(H)|}$，$score(H,w)=|V_w\cap V(H)|$である．

定義 1,2 より，ATC 問題を定式化する．

問題定義 1 [ATC 問題]．グラフ$G(V,E)$，クエリ$Q=(v_q,W_q)$，及びパラメータk,dが与えられたとき，$f(H,W_q)$が最大となる(k, d)-truss である部分グラフHを見つける．ただし，$f(H,W_q)=\sum_{w\in W_q}\frac{|V_w\cap V(H)|^2}{|V(H)|}$である．

先行研究 [2] では，ATC 問題を貪欲的に解く手法 LocATC を提案している．しかし，問題定義 1 に示したように，問合せ対象となるコミュニティは必ず(k, d)-truss を持つという強い構造的制約を持つ．ゆえに，LocATC は多様な構造を持ち得る実グラフに対して精度が低下する．

3 提案手法

本研究では属性付きグラフに対するコミュニティ問合せの精度向上を目指す．提案手法では ATC 問題の構造的制約条件を緩和することで，実グラフに対した柔軟なコミュニティ問合せを実現する．本節ではまず，ATC 問題の構造的制約を緩和した緩和 ATC 問題を 3.1 節で定義し，3.2 節にて緩和 ATC 問題の探索的解法を述べる．

3.1 緩和 ATC 問題の定義

問題定義 1 に示した通り，ATC 問題は truss 構造を多く含むコミュニティしか見つけることができない．しかし，これは実グラフにおいて，精度を大幅に低下させる要因となる．本研究ではkを可変とした(k, d)-truss，すなわち(*, d)-truss を考え，緩和 ATC 問題を次のように定義する．

定義 3[(*, d)-truss]．クエリノードv_qが与えられたとき，部分グラフHが(*, d)-truss であるとは，$v_q \in H$を満たし，かつHが$dist_H(H, v_q) \geq d$を満たすことを言う．

問題定義 2[緩和 ATC 問題]．グラフ$G(V, E)$，クエリ$Q = (v_q, W_q)$，及びパラメータdが与えられたとき，$f(H, W_q)$が最大となる(*, d)-truss である部分グラフHを見つける．

3.2 ビームサーチを用いた問合せ処理手法

問題定義 2 で示した緩和 ATC 問題の解法を提案する．本稿ではビームサーチを用いた緩和 ATC 問題の解法を提案する．

ビームサーチは，幅優先探索のようにある状態から遷移先の状態をキューにプッシュすることで，逐次的に探索木上を走査する探索手法である．そこに，ビーム幅Bというパラメータを加え，遷移先の状態数がBを超えた場合，暫定評価の最も良いB個のみを探索し，残りの状態は枝刈りする．これにより，探索木の各レベル(深さ)において多様な状態遷移の可能性を保持しつつ，全探索と比較して高速に探索を終える．

提案手法では，まずビームサーチの初期状態としてクエリノードのみを要素とした部分グラフを入力する．次に優先度付きキューからビーム幅Bに応じて最良のB個だけ部分グラフをポップし，それぞれについてその部分グラフとその隣接ノードからなる(k, d)-truss を$2 \leq k \leq k_{\max}$の範囲で計算し，見つかったそれぞれの(k, d)-truss を個別に優先度付きキューにプッシュする．ただし，$k_{\max}$はG中の最も大きな(k, d)-truss におけるkの値とする．これを繰り返し，その過程で最も Attribute Score Function が高かった部分グラフを解として出力する．

4 評価実験

本節では提案手法の精度を評価する．本実験では$B = 10$および$B = 30$と設定した提案手法ならびに従来手法 LocATC [2]を比較する．

データセット：先行研究 [2]で利用されていた Facebook データセットからノード ID 0, 107, および 348 を中心とした Ego ネットワークを用いる．

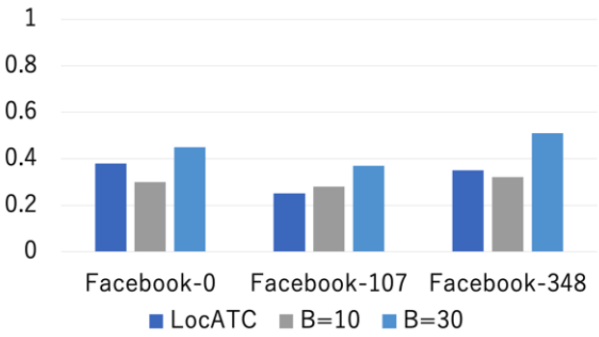

図 1. F1-Score の比較

クエリ：先行研究[2]に基づき，ランダムなクエリ$Q = \{v_q, W_q\}$を 100 回与える．属性集合はクエリノードが含まれるコミュニティで最頻出な属性値と，その他のコミュニティに最も現れない属性値の 2 つを与える．

評価指標：本実験では，各データセットで提供されている真のコミュニティと各手法が検出したコミュニティの間の精度を F1-Score を用いて評価する．

実験結果：実験結果を図 1 に示す．図 1 より，提案手法はビーム幅を十分に大きくすることで LocATC と比較してより高い精度でコミュニティを推定できることがわかる．この理由はビーム幅の増加に伴い，より多様な状態を保持しながら探索したからであると考えられる．以上により，提案手法の有効性を実験的に確認した．

5 まとめ

本稿では ATC 問題の構造的制約条件を緩和した緩和 ATC 問題を新たに定義し，ビームサーチを用いた緩和 ATC 問題に対する素朴な解法を提案した．我々の実験において提案手法は従来手法である LocATC と比較してより正確に真のコミュニティに近いコミュニティを算出でき，実データに対して有効であることを示した．今後の課題として，提案手法の高速化が挙げられる．

謝辞

本研究の一部は，科研費若手研究 (18K18057) ならびに JST ACT-I の支援を受けたものである．

参考文献

[1] Y. Zhou, et al., "Graph Clustering Based on Structural/Attribute Similarities," PVLDB, 2009.

[2] X. Huang et al., "Attribute-Driven Community Search," PVLDB, 2017.