A-023

可変 j-核および可変 k-Plex 性を用いた孤立性擬似クリークの全列挙について
Isolated Psudo-Cliques with Variable j-Cores and Variable k-Plexes

ジェイ ホンジェ† 原口 誠†
Hongjie ZHAI Makoto HARAGUCHI

1. はじめに

本稿では，大規模なネットワークから意味のある塊を疑似クリークとして切り出す手法について論じる．多数の頂点からなる高密度の疑似クリークは確率的に生じにくく，したがって，疑似クリーク探索においては中規模もしくは小規模な頂点集合を目標とすることになる．この場合，疑似クリークの総数は急激に増加し，効率的な探索はより困難となる．また，総数の問題のみならず，疑似クリークの重なり現象も頻繁に生じ，他の疑似クリークとの識別性に欠く疑似クリークも一般には多数存在する．過去において様々な疑似クリークのモデル・定義が提案されてきたが，そのほとんどは疑似クリークのサイズや他のクリークとの識別性を陽には考慮せず，内部的な接続具合のみに着目する研究が多かった．本研究がその基礎としている k-Plex の場合もしかりである．ここで，k は擬似クリーク内で非接続数上限を定める定数である．サイズが一定以上になる場合は，密な内部接続が保証されるが，k の値によってはとても高密度とは言えない頂点集合が k-Plex となり，このことが，高速な疑似クリーク検出実現を困難なものにしてきた．そこで，疑似クリークのサイズと密度要請に基づいて k を関数（可変 k）として定めることを提案したい．このことにより，伝統的な k-Plex 枚挙器の問題点の解決を試みる．

次に，他者との識別性が高いとは，意味としては孤立性が高いことであり，先行研究としては孤立（疑似）クリーク [3] の研究がある．カット数同様に，他者との接続数を全体として抑え込む考えに基づいているが，一方，本稿においては擬似クリークの所属頂点がもつ接続能力（次数）に着目した新たな定義を与える．すなわち，擬似クリークを構成する頂点 x は，その次数の多くを擬似クリークの内部接続に消費するとの考えに基づき，頂点毎に内部接続下限を定まる関数 $j(x)$ [2] を導入し，非接続数上限制約を与える 可変 $k(x)$ との同時制約により，意味のある塊（極大 k-Plex）で接続数下限制約を満たす擬似クリークを高速に算出する手法を与える．ちなみに本稿で展開する内容は，１００万頂点を持つ規模のグラフに対してその高速性を実証済の定数 j-核性を持つ極大な定数 k-Plex [1] の拡張版である．ただし，関数 $j(x), k(x)$ を適切に定めるために，ターゲットとする擬似クリークのサイズレンジと密度要請に基づいたアプローチをとる点が本稿の特徴となっている．広いサイズレンジを設定する場合は，それに応じて擬似クリークの総数が爆発するので，現実的にはサイズレンジを「輪切り」にし，区分的に求める手法が現実的であると思われる．本稿の手法はそうした戦略のもとで動作させることを想定している．

†北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

2. 可変 k-Plex，可変 j-核，

多重辺を持たない無向グラフを考える．その頂点集合 X の誘導グラフにおける頂点 $x \in X$ の次数を $deg_X(x)$，全体での次数を $deg(x)$ と記す．X はその頂点 $x \in X$ の非接続数 $|X| - deg_X(x) \leq k(x)$ のとき （可変）k-Plex と呼ぶ．特に，非負整数値関数 $k(x)$ が定数のとき，定数 k-Plex という．定数 k-Plex のクラスは逆単調性を持つが，可変な場合も全く同様である．すなわち，k-Plex の部分集合は k-Plex である．次に，所与の非負整数値関数 $j(x)$ に対し，$deg_X(x) \geq j(x)$ が任意の $x \in X$ に対し成立するとき，X は j-核性という．頂点毎に接続数下限を設定する点に注意したい．$j(x)$ が定数のときは特に定数 j-核性と呼ぶ．定数 j-核性の場合と全く同様に，頂点集合 X の誘導グラフにおいて最大の j-核性集合 $core_j(X)$ が存在し，集合値関数として単調，つまり $X_1 \subseteq X_2 \Rightarrow core_j(X_1) \subseteq core_j(X_2)$ が成立する．$core_j(X)$ の構成は接続数下限制約を満たさない頂点が存在する限り，そうした頂点を除去する形で遂行される．いくつかの頂点除去に伴い，他の頂点の次数も減少するので，頂点除去プロセスは再帰的である．このように定義される $j(x)$ と $k(x)$ のもとで，探索目標とする頂点集合 X は j-核性の極大な連結 k-Plex である．j-核性でない極大 k-Plex も存在するが，識別性に欠く疑似クリークであり，今回の抽出目標からは排除する．

3. $j(x)$, $k(x)$ の定義例

大規模スパースグラフを想定したとき，ターゲットとする疑似クリークに対しどの程度の密度要求を課すかは一般に微妙であり，実際問題としては密度パラメータ ζ をチューニングしながら実験を繰り返すのが常であると思う．本研究においては非接続数上限と接続数下限によって疑似クリークを制約するので，密度パラメータと接続数パラメータを仲介する第 3 のパラメータとして頂点集合サイズ（頂点数）を暗に想定している．これはサイズが大になるに伴い，非接続数上限と接続数下限は増加するからである．

サイズ下限 n_1 およびサイズ上限 n_2，さらに密度パラメータ ζ に対し，目標とする頂点集合 X は下記を満たさなければならない：

$$n_1 \leq |X| \leq n_2,\ deg_X(x) \geq |X|\zeta \geq n_1\zeta \qquad (1)$$

これに加え，各点ごとの孤立性条件 $deg_X(x) \geq deg(x)\tau$ を要請する．τ は頂点 x が持つリソース $deg(x)$ のうち，少なくとも $\tau deg(x)$ を疑似クリークのために使用すること意味している．これと式 1 の連言により j の定義を得る．

$$j(x) = \max\{n_1\zeta, deg(x)\tau\}$$

少なくとも $j(x)$ 個の内部接続を持つので，欠損数は高々 $|X|-j(x)$ である．この単純な事実に基づき，$k(x)$ を以下で定めても良い．

$k(x) = \max\{n_2 - j(x), 1\}$，但し，$n_2 - j(x) \geq |X| - j(x)$.

以下の節では，$j(x), k(x)$ の定義法に依存しない一般的な疑似クリーク探索法について述べるが，上記の定義を採用した場合，出力疑似クリークは必ずしも密度要求を満たすとは限らない．言うまでもなく，こうした出力は簡単な密度テストを施すことにより棄却する．言い換えれば，厳密解を最初から求めるのではなく，ある程度荒いが高速に漏れなく求めておき，事後テスト（単に次数チェック）を行えば良い．

4. j-核性極大 k-Plex

j-核性 k-Plex を探索するボトムアップ手法について述べる．基本的には，暫定的な k-Plex X に追加可能な $y \notin X$（候補頂点）を加え，j-核性の極大な k-Plex を求める．これは，k-Plex の逆単調性に基づいて極大擬似クリークを形成する標準的なプロセスであるが，j-核性にはなりえない暫定頂点集合 X を早期に枝刈る点が重要である．ここで，X の候補とは追加後の $X \cup \{y\}$（Xy と略記）が k-Plex であり，かつ，要素の追加により j-核性となる可能性のあるものを指す．j-核を求める集合値関数 $core_j$ は単調だが，j-核性は逆単調ではないので，形成プロセスの中途段階ではあくまでも拡大可能性に基づく判断しかできないことに注意したい．

まず，目的の疑似クリークは j-核性なので，全体グラフの最大 j-核 $core_j(V)$ の部分として出現する．よって，$core_j(V)$ の誘導グラフを改めて入力グラフとし，その頂点からなる単集合が暫定 k-Plex X の初期値となる．

FarCand:「遠すぎる候補」を持つ場合：
初期単集合を含め，X のサイズが小さい場合，候補頂点は X から距離２以上のものも含まれてくる．そうした頂点 $y \notin X$ のうち，距離（最短パス長）が $k(y)-|X|+1$ を超える候補 y は，X と最短パス上で $k(y)+1$ 以上の非接続頂点（y 自信も含む）を持つ．つまり，Xy を含む k-Plex は存在しないことがわかる．よって，候補からこうした「遠すぎる候補」y を削除し，さらに残った候補と X の和集合に対し j-核 Z を求める．$X - Z \neq \phi$，すなわち，j-核計算により暫定 k-Plex の一部の頂点が除去される場合は，X を含み，かつ j-核性の極大 k-Plex は存在しないことを意味し，X を直ちに棄却できる．$X \subseteq Z$ の場合は，j-核性を持つ極大 k-Plex に成長できる可能性があり，X に対して $Z - X$ の頂点のうち，X から距離１のものをさらに追加するプロセスを繰り返す．

NonFarCand: 全ての候補が暫定頂点集合と隣接：
暫定 k-Plex に距離１の候補を追加するプロセスが進行すると，全ての候補 y の X から距離が１，つまり，直接接続された状態になる．この場合は，「遠すぎる候補を排除した上での j-核計算」は不要である．つまり，暫定 k-Plex X とその（距離１内の）候補 y に対し，Xyz が k-Plex となる候補（これも距離１）に絞りこみ，そうした候補と Xy の合併集合に対する j-核 Z を計算する．Xy の一部の頂点が j-核から排除されれば，Xy は棄却され，$Xy \subseteq Z$ ならば，$Z - Xy$ が Xy の新たな候補集合となる．

特に，NonFarCand の暫定頂点集合がさらに下記の条件を満たす場合は，j-核の計算は不要で，単に通常の極大 k-Plex 探索が可能となる．

> 全ての候補 y が距離１内でかつ $|X| \geq j(y) + k(y)$ が成立すれば，X は j-核性である．

探索停止条件：候補集合が空であること．特に，X が極大であるときのみ，X を j-核性極大 k-Plex として出力する．

通常の 極大 k-Plex 探索における候補集合からさらに絞り込まれた候補集合を本手法では求めており，候補集合が空でも，k-Plex としては極大でない場合もある．よって，X と直接隣接した頂点 y に対し Xy が k-Plex であるかを検査し，そうした y が存在すれば非極大，存在しなければ目的の j-核性極大 k-Plex だとわかる．

探索の完全性：Z が j-核性を持つ極大 k-Plex の場合は，Z の形成プロセスにおける j-核計算で求まる最大 j-核は必ず Z を含み，よって，Z の形成プロセスにおける任意の暫定頂点集合が棄却されることはない．

5. おわりに

本稿では，j-核および k-Plex の定数パラメータを各頂点毎に関数として設定することにより，密結合で外部との識別性に優れた疑似クリークを効率良く（探索の初期段階で無駄なことをしない）求める手法を提案した．ここで，関数 $j(x)$, $k(x)$ ともに，ターゲットとする疑似クリークのサイズ，密度，識別性に直接関与するパラメータから設計できる点も従来の j-核や k-Plex 探索手法とは異なっている．今回は紙面の都合で述べなかったが，サイズ下限制約から分枝限定枝刈，探索最終フェイズにおける右候補枝刈等，いくつかのさらなる技法が可能である．実装は定数 j-核・定数 k-Plex 版 [1] に準じて行えばよく，その実験結果は発表時に報告したい．

参考文献

[1] Hongjie Zhai, Makoto Haraguchi, Yoshiaki Okubo and Etsuji Tomita: A Fast and Complete Enumeration of Pseudo-Cliques for Large Graphs, PAKDD 2016 (Part I), Springer-LNAI 9651, pp. 423 - 435, 2016.

[2] Yoshiaki Okubo and Makoto Haraguchi: Enumerating Maximal Isolated Cliques Based on Vertex-Dependent Connection Lower Bound, MLDM 2016, Springer LNAI, 2016 (to appear).

[3] Ito, H. and Iwama, K.: Enumeration of Isolated Cliques and Pseudo-Cliques, ACM Transactions on Algorithms, 5(4), Article 40, 2009