4C1-R-6-7

擬似クリークと負の制約を用いたグラフの構造変化検出

Detecting Structural Changes of Graph Based on Constrained Maximal k-Plex Search

大久保 好章*1　　原口 誠*1
Yoshiaki OKUBO　　Makoto HARAGUCHI

*1北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

This article discusses a problem of structural change detection given two graphs before and after some change, targeting vertex sets that are divergent pseudo-independent sets before the change and maximal k-plexes with smaller number of outgoing edges after the change. An effective search algorithm for them is presented as an extension of k-plex enumerator.

1. はじめに ー ターゲットとするもの

目まぐるしく変化する情報の世界においては，変化および変化の兆しを検出することは重要なタスクであると考えられる．顕著な変化を検出するシステムとしては，emerging pattern, contrast set，バースト解析など様々なアプローチがあるが，顕著ではない変化をも検出できる研究は未だ十分ではない．偶発的変化も変化としては極めて多数存在するからである．偶発的なものを排除し，ある程度の必然性を持つものに絞りこむことが重要である．

本稿では，ソーシャルネットワークに代表されるグラフ構造 [1] における変化検出を具体的なターゲットとし，イベント発生の前後，トピック，時間，カテゴリー等の文脈変化の前後における２つのグラフを比較し，潜在的だが決して偶発的ではない変化の候補を検出するタスクについて考察する．

無向グラフに対する頂点結合の必然性（もしくは逆に偶発性）の度合いを図る尺度として モデュラリティ [5] が提案され，グラフ分割・クラスタリング等において活発に使われている．変化前後のグラフにおける モデュラリティの差を直接評価する手法 [6] もあるが，抽出対象として (Acc) 変化前において結合の必然性がない，もしくは弱いが，(Nec) 変化後において結合の必然性がある，もしくは強いものを求める．Newman モデュラリティにおいては，頂点集合（クラスター）の必然性は，集合に属する頂点対の モデュラリティの総和で定められ，各対の モデュラリティは，(Neg) 高次数の頂点が結合していないほど小さな値（負）をとり，(Pos) 低次数の頂点が結合（隣接）しているほど高い値（正）をとる．(Neg) と (Pos) はそれぞれ，隣接関係の偶発性と必然性だと解する．この解釈に基づいて，本稿で求めたい頂点集合 X に対し，変化前後のグラフにおいて下記を要請する.,

変化前に対し負の制約 (Neg)： 多くの X 中の頂点は互いに非接続，かつ，頂点が持つエッジは殆どが外部への接続を与える（**発散的**）．

変化後に対し正の制約 (Pos)： 変化後において，多くの X 中の頂点は互いに隣接関係にあり，かつ，X は比較的低次数の頂点からなる．次数が小かつ相互に接続されることから，外部へ至るエッジ数は少数（**内部に閉じている**）.

研究会論文 [7] では，グラフの構造上の特徴として，変化前（後）において独立集合（クリーク）になることを要請した．本稿では，疑似的な独立集合と疑似クリークを許すものを与える．疑似クリークに関しては様々な定義が可能だが，ソーシャルネットワーク解析において一定の使われ方をしている k-plex [2] を本稿では考え，その全枚挙手法に基づいた解法を与える．

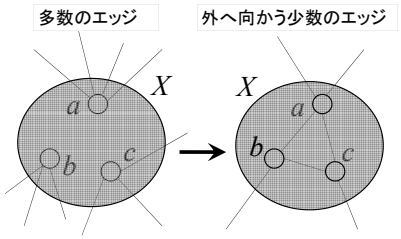

図 1: 発散的な (疑似) 独立集合から内向きに閉じた (疑似) クリークへの変化

2. 統合グラフとその評価

目標とする頂点集合 X に対し，最も厳密な場合は，変化前の補グラフと変化後のグラフにおいて共にクリークをなすことを要請する．ただし，例外的な場合を許容するために，X の少数の頂点対は,「変化前において非接続かつ変化後において接続」でない例外をある一定数だけ許す．このことを，統合グラフの k-plex として表現する．正および負の制約は，それぞれ，変化前後において X からその外部へ接続するエッジ数評価の下限値および上限値により実現する．特に，評価は，分枝限定法を用いたクリーク評価法と同様に，追加可能な候補頂点集合も含めた形で行う（暫定評価）.

2.1 統合グラフ

変化前後の（重み無し無向）グラフ $G_j = (V_j, \Gamma_j)$ $(j = 1, 2)$ を与える．j は変化前後を表す添字，頂点 $x \in V_j$ に対し $\Gamma_j(x)$ は G_j における x の隣接頂点集合を表す．特に，自己ループを認めない（$x \notin \Gamma_j(x)$）とする．目標となる頂点集合とは，

A: 原口 誠
北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
電話: 011-706-7106
E-mail: mh@ist.hokudai.ac.jp

G_j $(j = 1, 2)$ の共通の頂点集合 $X \subseteq V_1 \cap V_2$ で X 中の任意の頂点 x に対し，x と「G_1 において非接続 かつ G_2 において接続」でない頂点数が (x 自身も含めて) 所与の正整数 k 以下であるものをさす．G_1 の補グラフのエッジ集合を $\Gamma_1^c = (V_1 \times V_1) \setminus (\Gamma_1 \cup \{(x, x) | x \in V_1\})$ とすると，X は統合グラフ $G = (V = V_1 \cap V_2,\ \Gamma = \Gamma_1^c \cap \Gamma_2)$ における k-plex と一致する．すなわち，全ての $x \in X$ に対しその欠損エッジ数 $|\{y \in X - \{x\} \mid y \notin \Gamma(x)\}| \leq k - 1$．求める頂点集合に対する構造的要請はこの k-plex 性のみである．外部との接続を制約する正および負の制約は，統合グラフにおける性質ではなく，変化前後の元のグラフ G_j における制約として扱うことに特に注意する．

2.2 変化後の頂点集合評価

求める頂点集合 X は，変化後の G_2 において，多数が相互に結合しかつ X 以外への接続が（次数に照らして）少ないものであり，専門性の高い閉じたコミュニティと理解できる．外部への接続を下記で評価する．

定義 1 $X \subseteq V_1 \cap V_2$ の評価：G_2 において X 中の頂点から外部に出るエッジ率の最大値

$$E_2(X) = \max_{x \in X} \frac{outgoing_2(x|X)}{deg_2(x)}, \quad where$$
G_2 における x の次数：$deg_2(x) = |\Gamma_2(x)|$,
外部接続エッジ数： $outgoing_2(x|X) = |\Gamma_2(x) - X|$

クリーク同様に， k-plex の任意の部分集合は k-plex となり，それゆえに，本稿では極大 k-plex のみを評価の対象とし，これを次節で述べる漸増的なアルゴリズムによって求める．非極大な k-plex は，追加可能な頂点まで考慮に入れた下記で評価する（暫定評価）．

定義 2 統合グラフ G の k-plex X の暫定評価：

$$est\text{-}E_2(X) = \max_{x \in X} \frac{outgoing_2(x|X \cup Cand(X))}{deg_2(x)}, \quad where$$
X に追加可能な候補頂点集合
$Cand(X) = \{y \in V - X \mid\ X \cup \{y\}$ は $k\text{-}plex\ \}$.

候補集合は実際には，各頂点に欠損エッジ数のカウンターをつけ，上限値以下であるかをチェックして決める．

事実 1 *(1)* G の極大 k-plex X に対し，$E_2(X) = est\text{-}E_2(X)$.

(2) G の k-plex X_1, X_2 に対し，$X_1 \subseteq X_2 \Rightarrow est\text{-}E_2(X_1) \leq est\text{-}E_2(X_2)$

簡単に上記の事実を説明しておく．まず，G の極大 k-plex X に対し， X に追加可能な任意の頂点は候補集合 $Cand(X)$ に含まれる．よって (1) は明らか．次に，クリーク同様に，k-plex X への候補 $x \in Cand(X)$ の追加によって，追加後の候補頂点集合は単調に減少する．排除された追加前の候補頂点 y は，追加後の k-plex $X \cup \{x\}$ の外部に位置する頂点であり，こうした y への接続エッジ数が outgoing エッジとして足しこまれる．すなわち，暫定評価関数は単調増加である．

outgoing エッジ率上限 δ_2 を与え，単調性に基づく下記の枝刈りが可能となる．

事実 2 G の極大 k-plex M が G の k-plex C を含むとする．

$$\delta_2 < est\text{-}E_2(C) \Rightarrow \delta_2 < est\text{-}E_2(C) \leq est\text{-}E_2(M) = E_2(M)$$

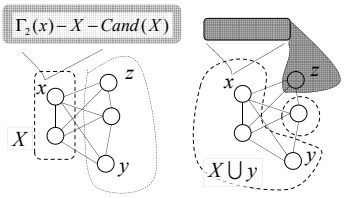

図 2: k-plex の G_2 評価

2.3 変化前の頂点集合評価

統合グラフ G での k-plex X は，各ノードに対し，他の X 内ノードへの欠損エッジ数は高々 $k - 1$ である．この制約から，各 $x \in X$ に対し，元の変化前の G_1 においては，x からは高々 $k-1$ 個のエッジしか X 内へ張れないことを意味する．したがって，x から G_1 において X の外部へ接続するエッジ数 $outgoing_1(x|X)$

$$deg_1(x) \geq outgoing_1(x|X) \geq deg_1(x) - k + 1$$

を満たす．所与の下限値 δ_1 に対し，$outgoing_1(x)$ を直接用いず，その近似値としての次数 $deg_1(x)$ を用いよう．実際， k-plex の先行研究では，k の数は比較的小さな値を想定しており，少なくとも変化前の評価においては次数が外部接続数の良い近似を与えると言える．実際の評価式は，変化後の評価同様に，候補頂点まで加味した形で与え，その単調性を実現する．

暫定評価定義 $est\text{-}E_1(X) = \max_{x \in X \cup Cand(X)} deg_1(x)$

単調性 $X_1 \subseteq X_2 \Rightarrow est\text{-}E_1(X_1) \geq est\text{-}E_1(X_2)$

枝刈規則 $\delta_1 > est\text{-}E_1(X) \Rightarrow$
X を含む G_1 の極大 k-plex M に対し，
$\delta_1 > E_1(M) = \max_{x \in M} outgoing_1(x|M)$

3. 構造変化頂点集合全列挙アルゴリズム

本節では，構造変化頂点集合を全列挙する方法について述べる．基本的には，極大クリーク全列挙手法 [4] を，統合グラフの極大 k-plex 生成のために用い，探索木を縦型に走査する．ただし，極大 k-plex の重複枚挙を防ぎ，無駄な探索枝を抑制するためのシンプルで効果的な制御規則を，極大クリーク全列挙法の拡張として新たに導入する．探索木としては，統合グラフの極大 k-plex 生成のための探索木を「なぞり」ながら，その部分木のみを結果的に生成する．

3.1 極大 k-plex 全列挙，特に Right 候補制御

極大 k-plex の全列挙としては 先行研究 [3] があるが，本稿と同じく，新規候補頂点を k-plex X に漸増的に追加し，極大 k-plex を得る手法である．すなわち，探索木の根は空頂点集合，探索木の葉は 極大 k-plex X ，根から葉へ至るパスは，k-plex 列 $\{X_n\}_n$ で，候補 $x_n \in Cand(X_n)$ に対し，$X_{n+1} = X_n \cup \{x_n\}$ である．無駄な枝を抑制する標準的な方法，例えば，固定順序のもとに頂点集合の重複枚挙を防ぐ方法を，任意の枝選択順序に拡張した Left 制御規則 などは本研究でも用いている．

さらに本稿では，特に，極大クリーク全列挙法 [4] における Right 候補制御を極大 k-plex 探索に拡張したものを新たに

提案する．Right 候補制御では，クリーク X の候補 $x,\ y \in Cand(X)$ で，y が x に隣接しているとき，$X \cup \{y\}$ を真に含む極大クリークは，x を含まない場合は x とは非隣接だが y とは接続された候補 $z \in Cand(X)$ を持つことに着目する．すなわち，$X \cup \{y\}$ を含む極大クリークは，x および $\{z \in Cand(X) \mid z \in \Gamma(y) - \Gamma(x)\}$ の要素を X に追加する探索パスで全て網羅できる．

この Right 制御規則は探索の初期段階で多くの枝を刈る効果を持ち，それゆえに，高速な極大クリーク全列挙を実現することが知られている．本稿における k-plex の場合は，少数のエッジの欠落を許すが，下記に記すように，Right 制御の考えは全く同様に適用できる．

k-plex における Right 候補：極大でない k-plex X を考える． $Cand(X)$ 中の候補 $u \in Cand(X)$ を一つ選び，

$$Right_u(X) = \Gamma(u) \cap \{y \in Cand(X) \mid X \subseteq \Gamma(y)\}.$$

u は探索木のノードとしての X に対しただ一つだけ選択され，X に対する Right 候補が決まる．u を含む 非 Right 候補を実際に新探索ノード展開のための枝として使ってよいかは，Left 候補制御に依存する．Left 候補とは，根 $X_0 = \phi$ から $X = X_n$ に至る探索パス上の X_j （$j \leq n$）において既に試行した，すなわち，枝として使われた候補を指し，非 Right かつ 非 Left な X の候補が実際に使用可能な枝となる．そうした枝が残っていない場合は，バックトラックし，親ノードの再試行を行う．

Right 候補制御の正当性の証明は，クリークの場合と同様に，

> k-plex X に対し，Right 候補のみを X に追加してできる k-plex は極大 k-plex ではない

を証明することなる．そのために，k-plex の内部表現，すなわち，頂点に持たせた欠損エッジ数のカウンターを用いて初めて正確な議論が可能だが，紙面の都合上，ここでは省略する．

3.2 構造変化検出アルゴリズム

本稿で求める頂点集合 M とは，

> M は 統合グラフにおける極大 k-plex で
> $E_1(M) \geq \delta_1$ (発散性) かつ $E_2(M) \leq \delta_2$ (内部閉鎖性)

の２つの性質を満たすものである．これを δ_1 δ_2-解と呼ぶ．統合グラフ G における 解 M に含まれる非極大 k-plex X は，暫定評価が持つ単調性により常に

$$est\text{-}E_2(X) \leq est\text{-}E_2(M) = E_2(M) \leq \delta_2.$$

よって，$estE_2$ 評価で M が生成されないことは生じない．また，$estE_1$ 評価でも，

$$est\text{-}E_1(X) \geq est\text{-}E_1(M) \geq E_1(M) \geq \delta_1.$$

よって，解 M が $estE_1$ 暫定評価 により途中で刈られることはない．ただし，暫定評価 $estE_1(M)$ と $E_1(M)$ は一般に異なるので，葉ノードでの $E_1(M) \geq \delta_1$ のチェックは必要である．

4. 実験

本稿での提案手法を C 言語で実装し，CPU: Intel® Core™-i3 M380 (2.53GHz)，主記憶：8GB の PC 上で，Twitter ユーザを頂点とする無向グラフを対象とした構造変化検出を試みた．

4.1 Twitter ユーザの関係グラフ

Twitter 検索 API を用いて，キーワード**民主党・自民党・公明党・社民党・共産党・みんなの党・国民新党・たちあがれ日本・新党大地・新党きづな・大阪維新の会**の何れかを含むツイートを収集し，それらツイートに関わったユーザ間の関係をグラフで表現する．具体的には，ユーザ A がユーザ B に対してリプライツイートを行った場合に，辺 (A, B) を張ることでユーザ間の関係グラフを作成する．ここでは，平成 24 年 4 月 4 日〜6 日の 3 日間に収集したツイート群から抽出したユーザ間の関係グラフを変化前グラフ G_1，平成 24 年 4 月 13 日〜15 日の 3 日間のツイート群からのそれを変化後グラフ G_2 とし，そこで観測される構造変化検出を試みる．G_1 の頂点総数は 4,150，辺総数は 3,405(密度 0.04%)，一方，G_2 の頂点総数は 3,755，辺総数は 3,069(密度 0.04%) であり，これらの統合グラフ G_{12} の頂点総数は 872，辺総数は 507(密度 0.13%) である．

4.2 構造変化 3-Plex

$\delta_1 = 30$ および $\delta_2 = 0.95$ による外部接続度制約のもとでの構造変化 3-plex の例を図 3 に示す．

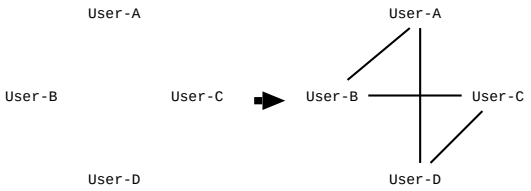

図 3: 構造変化 3-Plex

図 3 において，User-A と User-C はそれぞれ民主党の某衆議院議員，User-B は脱原発・反 TPP・消費税阻止等をプロフィールに掲げるユーザ，User-D は反原発・反 TPP を掲げるユーザである．これらユーザ間には，平成 24 年 4 月 4 日からの三日間において直接ツイートのやりとりはまったく無く，外部のユーザとのやりとりのみである．そこでのツイート内容は，例えば外部ユーザから User-A への『TPP 強行阻止の訴え』や，User-B から外部 (新聞社等を含む) への『消費税増税に関する反対意見表明』などがある．その約一週間後，これらユーザは直接ツイートのやりとりをしている．その主な内容は，User-B と User-D から民主党議員 User-A・User-C に対する直接の『民主党への苦言』等である．実際，4 月 13 日になされた大飯原発運転再開が妥当であるとの判断を巡って，社会的な議論が巻き起こっており，反原発を掲げる User-B や User-D は政権を担う民主党議員に直接訴えなければ気が済まなかったのかもしれない．この様に，本構造変化検出手法により，ある種のコミュニティ形成が始まる様子を捕えることが可能である．

本実験で用いたデータからは，比較的小さなユーザグループのみが構造変化頂点集合として検出された．その要因のひとつとして，変化前後の時間間隔が短かいことが考えられる．こうした小さなユーザグループの中には，これからさらに大きく成長する可能性のある萌芽的なコミュニティを含むことが期待でき，マイニングの観点からはより価値のあるグループになり得ると考えている．

ユーザ間の関係グラフを作成する際，今回はリプライツイートの有無のみに着目した無向辺を考えたが，これを有向辺として扱うことで現実の関係をより的確に反映したものとなろう．

さらに，リプライのやりとりの頻度を辺の重みとして反映させることも考えられる．これらの点を考慮した制約や評価尺度を導入することで，抽出対象となるグループにさらに明確な意味を持たせることができると思われる．

4.3 計算時間

本提案手法による構造変化検出の計算時間に影響するパラメータは k，δ_1 および δ_2 であるが，その中で最も支配的なものは k の値である．ここでは，外部接続度制約を外した場合，すなわち，$\delta_1 = 0, \delta_2 = 1.0$ とした場合の k の変化に伴う計算時間の挙動を，二つの DIMACS ベンチマークグラフ[*1] `c-fat200-1` と `c-fat200-2` において観察する．これは，統合グラフ G_{12} における極大 k-plex 列挙に相当しており，これらのうち，外部接続度制約を満たすものが実際の構造変化頂点集合となる．Right 候補制御を用いない素朴な極大 k-plex 列挙による計算時間と比較することで，本制御機構の有効性を確認する．

`c-fat200-1` は頂点数 200，辺密度 7.7% の (無向) グラフであり，極大クリーク数は 37，その最大サイズは 12 である．一方，`c-fat200-2` は頂点数 200，辺密度 16.3% のグラフで，極大クリーク数は 18，その最大サイズは 24 となっている．両グラフの性質の違いを一言で述べるならば，前者は比較的小さな極大クリークを多く含み，後者は比較的大きな極大クリークを少数含むものとなろう．それぞれのグラフに対する計算時間を図 4 に示す．

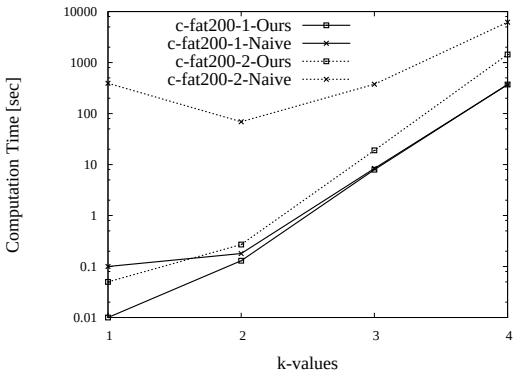

図 4: 計算時間 (外部接続度制約なし)

k の増加に伴い計算時間が指数関数的に増大することがわかる．前章で議論した Right 候補制御は，`c-fat200-1` に対しては，その効果があまり確認できなかった．これは，`c-fat200-1` では，各 k における k-Plex の最大サイズが，いずれも極大クリークと同じ 12 と比較的小さく，探索の各段階で分枝数がさほど大きくならず，Right 候補制御の効果が現われにくかったことを意味している．一方，`c-fat200-2` においては，素朴なアルゴリズムと比較して，Right 候補制御が極めて有効に機能していることがわかる．しかし同時に，より大きな k に対応するにはその抑制効果をさらに高める工夫が必要であることも示唆される．

5. まとめと今後の展望

前節までで問題の所在と解法および実験結果について述べた．本節では今後の課題等について述べる．

構造変化検出のために，クリークを用いる [7] や k-plex を用いる本研究は，ともに，変化後のグラフにおいて頂点集合から外部へのエッジが少数であること，および，変化前のグラフにおいて外部へのエッジが多数であることを制約として用いた．一方，構造制約として，クリークもしくは k-plex の意味での疑似クリークを形成する条件を課したが，前者の外部との接続制約の一部は，候補集合の定義の中に組み込むことが可能である．そのことにより，潜在的な枝数をさらに減らせ，効率化を期待できる．

また，極大 k-plex 枚挙法としては，Right 候補制御が本質的だと述べた．本稿で与えた Right 候補の定義は，クリークの場合のほぼ直接的な拡張であり k-plex であることを十分反映したものにはなっていない．これに関しても，本研究で定めた Right 候補を完全に含む形でのより大きな Right 候補集合を既に得ており，別の機会に発表したい．

上記で述べたことは，変化検出の基礎として，クリークや疑似クリーク列挙法に関するテクニカルな話題である．一方，変化そのものに対する考察も必要不可欠であろう．本研究のアイデアの源泉は文献 [6] で述べたように，

一度に全ての関係（接続関係）が変化することは稀である

という経験則を暗に想定している．ただし，場合によっては，あるいは，グラフの作り方・制限の仕方によっては，グラフ全体で大規模な変化が生じることもありえるだろう．そうした場合，本稿で言う極大 k-plex の数は膨大なものになることもないとは言えない．これに対しては，同じく分枝限定手法の枠組みで，k-plex の別の単調評価に基づくトップN法による最適化が自明に可能である．このように，ドラスティックな変化データに対する手法の検証も残っている．

参考文献

[1] B. Furht (ed.), Handbook of Social Network Technologies and Applications, Springer, 2010.

[2] S. B. Seidman and B. L. Foster, A Graph Theoretic Generalization of the Clique Concept, Journal of Mathematical Sociology, 6, pp. 139 - 154, 1978.

[3] B. Wu and X. Pei, A Parallel Algorithm for Enumerating All the Maximal k-Plexes, Proc. of the PAKDD 2007 Workshops, LNAI-4819, pp. 476 - 483, 2007.

[4] E. Tomita, A. Tanaka and H. Takahashi, The Worst-Case Time Complexity for Generating All Maximal Cliques and Computational Experiments, Theoretical Computer Science, 363(1), pp. 28 - 42, Elsevier, 2006.

[5] M. E. J. Newman, Finding Community Structure in Networks Using the Eigenvectors of Matrices, Physical Revivew E, 74(3), 036104, American Physical Society, 2006.

[6] M. Haraguchi, Things That Change, Things That Do Not Change, Proc. of GCOE-NGIT 2012, pp. 19 - 21, 2012.

[7] エラウィンディ サラ・原口 誠・大久保 好章・富田 悦次, クリーク全列挙に基づく構造変化検出アルゴリズム，情報処理学会研究報告，Vol. 2011-MPS-087 No. 32, 2012.

*1 `ftp://dimacs.rutgers.edu/pub/challenge/graph/benchmarks/clique/`