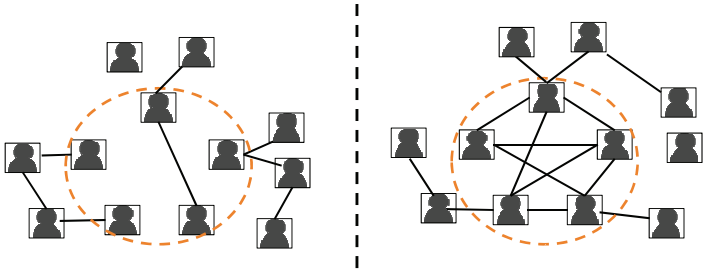
疑似独立集合制約と正規化カットを用いたグラフの構造比較

北海道大学 情報科学研究科
間澤直寛，Zhai HongJie，原口誠
電気通信大学 先進アルゴリズム研究ステーション
富田悦次

- グラフ(全体)の差異や距離ではなく，グラフの部分での差異を示す頂点集合．
- それらは複数あり，それゆえに，マイニングの問題と考える

Contrast vertex set:

2つのグラフの差異を示す頂点集合



Base graph ← 時間，場所， → target graph

anti-community

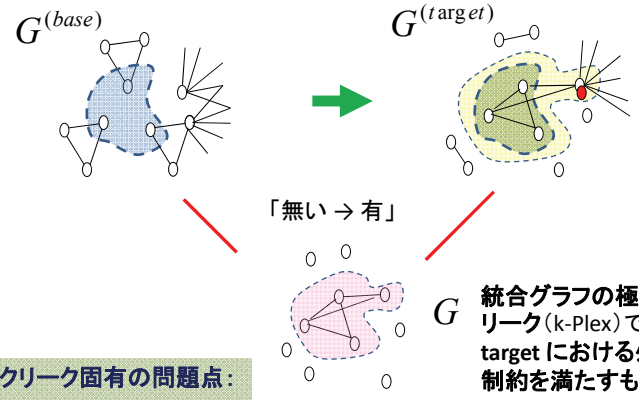
塊． 孤立性を要求 community

暗黙の仮定，ならびに目標

- 接続・非接続関係が全て同時に変わることは稀．
  多くの（特に非接続）関係は変わらない．
- 特に，ベースにおいて疎結合（非接続）な頂点集合で，
  ターゲットにおいて密結合のものを探す．
- ターゲットにおける community 性（外部との接続が少数）

関連研究 1：制約下での極大疑似クリーク探索

E.サラ et al.，MPS 87-2 (2012)
Y.Okubo et al., Discovery Science 2012.（和文版は JSAI2012 大会）



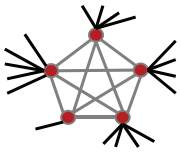
統合グラフの極大疑似クリーク（k-Plex）で，base, target における外部接続制約を満たすもの

② 疑似クリーク固有の問題点：

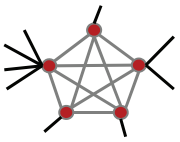
接続例外を許したクリーク
例外数の増加
⇒ 例外の組合せ爆発

① 孤立性がより高い部分を見落とす危険性

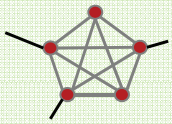
関連研究2： c-孤立(疑似)クリーク
Ito 2009, Komusiewicz 2009



最小c-孤立性：
外部次数c未満の頂点が存在



c-孤立性：
外部次数が平均c未満



最大c-孤立性：
各頂点の外部次数がc未満のクリーク

k=1 を除いて不十分 ⇒ 極大な最大c-孤立 *k*-plex 列挙 = **極大(*k*,*c*)-plex の列挙**

逆単調性が言えない ⇒ 組合せ論的探索の困難性
少なくとも自明ではない.

今回の作戦

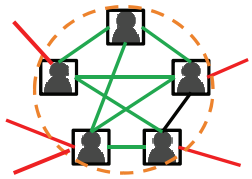
- 逆単調性に基づくクリーク探索の高速性を活かす：
  目標それ自体は，**接続例外を許した疑似クリーク**
  距離空間への埋め込み (**正規化カット最小化**)
  unit disk graph 風のグラフにおける**クリーク制約**.
- ベースでの疎結合性：**疑似独立集合制約**
  （補グラフにおける疑似クリーク）

  2つの制約（共に逆単調）を同時に満たす極大な頂点集合を列挙

◆ **埋め込み後の disk graph のクリークを，候補頂点を追加する形で形成する.**

◆ **その際，ベースでの制約を満たさない候補頂点をカット可能**

正規化カットと射影（埋め込み）



$$\mathrm{Ncut}(\mathbf{X},\overline{\mathbf{X}}) = \frac{\mathrm{cut}(\mathbf{X},\overline{\mathbf{X}})}{\mathrm{vol}(\mathbf{X})} = \frac{\sum_{i\in\mathbf{X}, j\in\overline{\mathbf{X}}} w_{ij}}{\sum_{i\in\mathbf{X}} d_i}.$$

$\sum_{1\le j\le k} Ncut(X_j, \overline{X}_j)$ の最小化問題の近似解

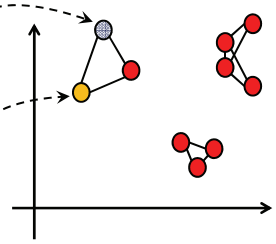
$$L_{sim} = D^{-1/2} L\, D^{-1/2} = D^{-1/2}(D-W)D^{-1/2}$$

$$V' L_{sim} V = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \qquad V = (\vec{v}_1 \quad \ldots \quad \vec{v}_n)$$

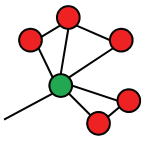
$$D^{-1/2}(\vec{v}_1 \quad \ldots \quad \vec{v}_k) = \begin{pmatrix} v_{11}/\sqrt{d_1} & & v_{1k}/\sqrt{d_1} \\ \vdots & & \vdots \\ v_{n1}/\sqrt{d_n} & & v_{nk}/\sqrt{d_n} \end{pmatrix}$$

一般化固有値問題の固有ベクトル

頂点の近接性と unit disk graph



射影距離の性質



狭い範囲内の配置

$$\mathrm{Ncut}(\mathbf{x},\overline{\mathbf{x}}) = \mathbf{x}'\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}\mathbf{x} = \frac{1}{2}\sum_{i,j} w_{ij}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)^2.$$

「端末」の次数は概ね小で，近場に配置した方が最小化に有利．「センター」は，**バランス条件**から少し離れて配置

自明な解 $\left(\sqrt{d_{i'}}\right)_i$ との直交条件 $\sum_i \sqrt{d_i} x_i = 0$

一般化固有値問題の解 $y = D^{-1/2}x$

$$y' L y = \frac{1}{2}\sum_{i,j} w_{ij}(y_i - y_j)^2 \qquad \sum_i d_i y_i = 0$$

固有ベクトルの距離の性質 (Mazawa's original slide)

一定の距離基準で辺を張ると、うまくクラスタリングできない場合がある。その例が、スターグラフである。

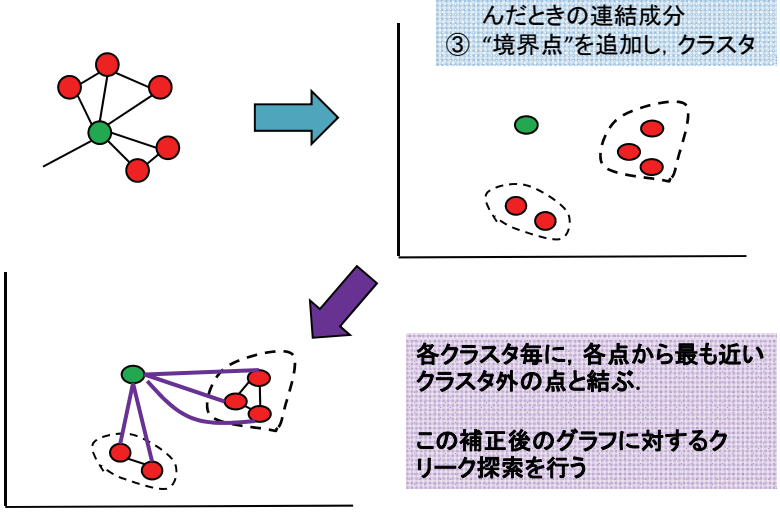
密度が高い部分と少し離れた値が出る。

ソーシャルネットワークは特定の人物に多くの人が集まることが多いのでスター構造になりやすい。

左の形状だと，フィルダーベクトルでは，スターも端末もごく近くに配置すると，センターの接続先と端末が近くになりまずい？
第2固有値だと，センターと端末の違いを認識している？
いずれにせよ，センターの外部との接続関係にも注意して，数値実験が必要．
バランス条件があるので，mazawa の主張はにわかには信じがたい．

DBSCAN とその補正

DBSCAN:
① 近傍内に多数の点を持つ内点(core pt)
② 近場の core pt を結線で結んだときの連結成分
③ "境界点"を追加し，クラスタ



各クラスタ毎に，各点から最も近いクラスタ外の点と結ぶ．

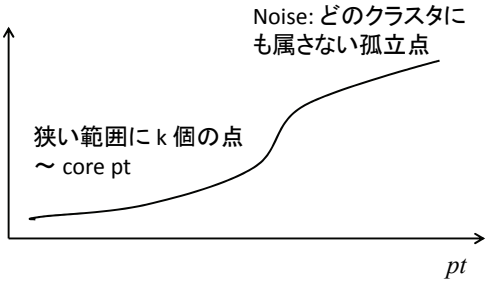
この補正後のグラフに対するクリーク探索を行う

minPts, ε の決め方（経験則）

- minPts = k から ε を決める

$k - dist(pt)$

K 個の点を含む最小半径

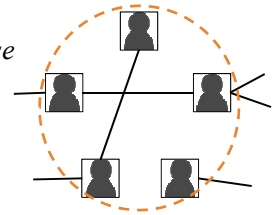


- minPts 大：希少な core pt.
  minPts 小：殆どが core pt
  core pt の数で適宜決める

疑似独立集合制約

補グラフにおいて k-plex

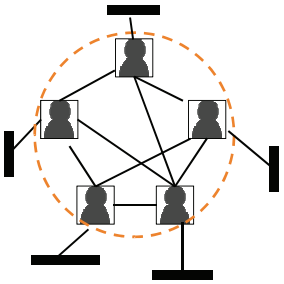
G_{base}



高々 (k-1) の結線.

逆単調性

$\overline{G}_{base}$



補グラフで 高々 (k-1) の欠落

実験データ

2006年9月の新聞記事において単語を頂点、単語同士の共起を辺として、頻度で重みを付ける。

全国版の某政党に関する記事から抽出した単語の関係グラフをBase。
北海道版からのそれをTargetとする。

頂点数：749
辺数(Base)：65463
辺数(Target)：87226
k=5

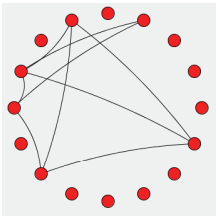
実験結果

得られた解の数は3041、解の最大サイズは25
K=5, minPts = 5, ε = 2*10^{-4}

この例では，k =4 でもOK

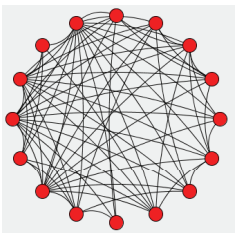
最も顕著な差異があった例

Base（全国版）



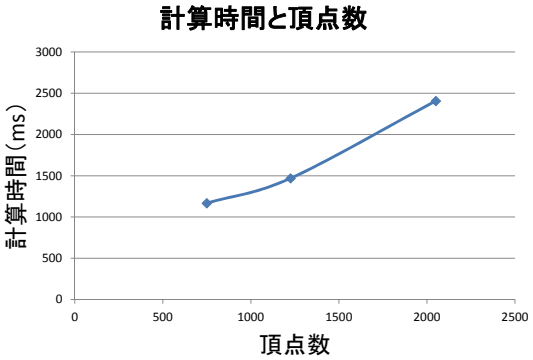
地方票
党友
落選議員名
健闘
etc.

Target（地方版）



2006年9月の某政党総裁選において、当選議員が全国的には圧勝であったが、北海道では他の議員も健闘していたという記事が見られた。

計算時間(固有値計算以外)



クリーク探索の高速性と，ダブル制約の効果

Target
平均次数:232.9132

距離空間射影
平均次数:5.9839

空間内に バランスをとって配置され，その結果，もとのグラフと比較すると疎になる

今後の課題

埋め込みの補正は一つの経験則的なパッチ

- ✓ DBSCAN 以外の Density-based clustering の検討
- ✓ 元のターゲットグラフの接続構造を保存する埋め込みの可能性

計算時間の問題点: ラプラシアンのスペクトル分解

- ✓ 近似高速化手法の導入