3F1-3

Affinity Propagationによるコミュニティ抽出

Community Detection by Affinity Propagation

杉原 貴彦[*1] 村田 剛志[*1]
Takahiko Sugihara Tsuyoshi Murata

[*1]東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Community detection from networks is one of the important topics in link mining. We apply a clustering method called 'Affinity Propagation' to community detection. Affinity propagation performs clustering by message passing between all nodes. We propose several new similarity metrics of vertices to detect communities from simple networks and signed networks composed of positive and negative links. Using these similarities, our method detects communities with the same quality as the existing methods in a relatively short time. In addition, our method detects exemplars of the communities.

1. はじめに

近年，Web 技術の進歩に伴い，ネットワークからのコミュニティ抽出に関する研究が多くなされている．ネットワークを密に結合したノード集合であるコミュニティに分割することで，全体構造の把握や，トピックの抽出を行うことができる．Girvan・Newman[1] や，Clauset[2] らにより，ネットワークのリンク構造に注目し，コミュニティを抽出する手法が提案されている．

本研究では Frey らによって提案されたクラスタリング手法である Affinity Propagation(AP)[3] を用いてコミュニティ抽出を行う．AP に用いる類似度として Adamic/Adar[4] による類似度を提案し，4 つの実ネットワークからのコミュニティ抽出を行った．得られた結果をモジュラリティや計算時間によって評価した．その結果，AP は既存のコミュニティ抽出法と同等のモジュラリティを，比較的短い時間で得ることができた．

さらに，コミュニティ抽出の応用として，signed ネットワーク[*1]からのコミュニティ抽出を行った．signed ネットワークとは，友好関係や協力関係を表す正のエッジと，敵対関係を表す負のエッジの 2 種類のエッジを持つネットワークのことである．signed ネットワークにおけるノード間の類似度を定義し，signed ネットワークからのコミュニティ抽出を行った．AP を用いたコミュニティ抽出は，他手法による分割と類似した結果を得ることができるとともに，各コミュニティの代表ノード (exemplar) を得ることができた．

2. Affinity Propagation

本節では Frey らによって提案されたクラスタリング手法である Affinity Propagation(AP) について説明する．

2.1 Affinity Propagation の概要

AP はメッセージ交換によりクラスタリングを行う．メッセージとは，あるノードから他のノードへの評価値のことであり，数値が大きいほど評価が高いことを表す．AP では，各ノード間の類似度を示した similarity(2.1.1 節) をもとに算出される responsibility(2.1.2 節) と availability(2.1.3 節) という 2 種類のメッセージを用いる．メッセージ交換とは，これら 2 種類のメッセージを全てのノード間で交換し合うことを表す．

メッセージ交換により，どのノードがクラスタの中心である exemplar となるかを決定し，他のノードはどの exemplar のクラスタに所属するかを決定する．全てのノードが exemplar になる可能性を持ち，exemplar とそのクラスタのメンバー間の評価値が最も高くなるようにメッセージを交換していくので，AP ではクラスタ数やクラスタの形状は自動的に定まる．

2.1.1 similarity

similarity は各ノード間の類似度を表し，$s(i,j)$ と表される．この値が大きいほどノード i とノード j の類似度が高いことを表す．similarity は,「値が大きいほど類似度が高い」という条件を満たすならどのような指標を用いてもよい．また，similarity の $s(i,i)$ 成分は preference と呼ばれ，この値の大小により，クラスタ数が変化する．通常，preference には similarity の中央値を用いる．

2.1.2 responsibility

responsibility はクラスタメンバー i から exemplar 候補のノード k に送られるメッセージで，k がノード i の exemplar としてどれぐらい適切かを表す．メッセージは (1) 式のように表される．

$$r(i,k) \leftarrow s(i,k) - \max_{k' \mathrm{s.t.} k' \neq k} \{a(i,k') + s(i,k')\} \quad (1)$$

k 以外のノード k' からノード i への availability(後述) の値が大きいほど (1) 式の値は小さくなる．つまり，i が k 以外のノードから高い評価を得ている場合，i から k への responsibility の値は小さくなる．

2.1.3 availability

availability は exemplar 候補のノード k からクラスタメンバーとなる見込みのノード i に対して送られるメッセージで，ノード i が exemplar 候補 k のクラスタのメンバーとしてどれくらい適切かを示す．メッセージは (2)，(3) 式のように表される．

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \mathrm{s.t.} i' \notin i,k} \max\{0, r(i',k)\}\} \quad (2)$$

連絡先: 杉原 貴彦, 東京工業大学 大学院情報理工学研究科 計算工学専攻 村田研究室, 〒 152-8552 東京都目黒区大岡山 2-12-1 W8-59, sugihara@ai.cs.titech.ac.jp

*1 本稿では，正と負の 2 つのエッジからなるネットワークを “signed ネットワーク” と表し，エッジに種類がないものを “ネットワーク” と表している．

$$a(k,k) \leftarrow \sum_{i' \mathrm{s.t.} i' \neq k} \max\{0, r(i',k)\} \quad (3)$$

(2), (3) 式では，i 以外のノード i' からノード k への responsibility のうち，正であるものの和が用いられる．これは，ノード k が他のノードからどれだけ高評価を得ているかを表している．つまり，k が人気のあるノードであるほど，k から i への availability の値は大きくなる．

2.2 Affinity Propagation の手順

AP では responsibility と availability の反復計算を行い exemplar を求める．exemplar となるノードは，

$$r(k,k) + a(k,k) > 0 \quad (4)$$

を満たすノードである．反復計算は，

- 一定回数 exemplar となるノードが変化しなかった場合
- 反復回数が一定に達した場合 $(*)$

の 2 つのうち，どちらかの条件を満たした時点で終了する．exemplar 以外のノード i は，$r(i,k) + a(i,k)$ が最大となるノード k を exemplar とするクラスタに所属する．図 1 に AP の手順を示す．

Algorithm AP, AP(S)=C

S: Input, the similarity matrix of a network;
C: Output, the result vector;

1. Initialize R, A with 0; (R: responsibility, A: availability)
2. Update R by equation (1);
3. Update A by equation (2),(3);
4. Find exemplar k, where k satisfies (4)
5. If exemplars satisfy the condition (*)
   then go to 6, else go to 2;
6. Return C, where $C(i) = k$
   (k is the exemplar of the cluster that node i belongs)

図 1: AP の手順

3. 関連研究

3.1 モジュラリティ

ネットワークのコミュニティ抽出の際に有用な指標としてモジュラリティ [1] がある．コミュニティ内にエッジが多く，コミュニティ間にエッジが少ないような分割が良い分割であるという考えに基づき，モジュラリティはコミュニティの質を数値化するために定義された．モジュラリティ Q は以下の式で表される．

$$Q = \frac{1}{2m} \sum_{ij} (A(i,j) - \frac{d_i d_j}{2m}) \delta(c_i, c_j) \quad (5)$$

m はネットワークのエッジの総数，i,j はネットワークの任意の 2 ノード，A はネットワークの隣接行列，d_i はノード i の次数，c_i はノード i が含まれるコミュニティを表す．

3.2 モジュラリティを用いたコミュニティ抽出法

本節では，コミュニティ抽出法として，Girvan らによるエッジの媒介中心性を用いた分割的な手法である GN 法 [1] と，Clauset らによる貪欲法をベースとした最適化法である CNM 法 [2] を簡単に紹介する．

3.2.1 GN 法

GN 法はエッジの媒介中心性を用い，ネットワークを切断することでコミュニティを抽出する手法である．エッジの媒介中心性とは，あるエッジが，全ノード間の最短経路上に何回出現するかを表している．コミュニティ間に存在するエッジはこの媒介中心性が高くなりやすい．そこで，GN 法では，媒介中心性の高いエッジを取り除いていき，モジュラリティが極大となった時点での分割を結果とする．

3.2.2 CNM 法

CNM 法は各ノードのみからなるコミュニティを初期値として与え，それらを結合していくことでコミュニティを抽出する．モジュラリティを最も増加させるような結合を繰り返し，最終的なコミュニティを決定する．Clauset らはデータ構造を工夫することで，非常に大規模なネットワークのコミュニティ抽出を行っている．

3.3 singed ネットワークからのコミュニティ抽出

本節では正のエッジと負のエッジの 2 種類のエッジを持つ signed ネットワークについて考える．signed ネットワークでは，図 2 のように正のエッジを実線で，負のエッジを点線で表す．ネットワークでのコミュニティは，コミュニティ内にエッジが多く，コミュニティ間にエッジが少ないサブネットワークである．しかし，signed ネットワークではエッジの種類も考慮しなければならない．signed ネットワークにおけるコミュニティは，コミュニティ内に正のエッジが多く，コミュニティ間に負のエッジが多いサブネットワークである．コミュニティ抽出法としては，Yang[5] らによって，ランダムウォークによるコミュニティ抽出法が提案されている．

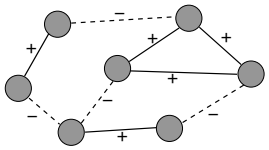

図 2: signed ネットワークの例

4. Affinity Propagation によるコミュニティ抽出

この節では，実際に Affinity Propagation(AP) をネットワークに適用することを考える．AP を用いるメリットとして，コミュニティ数が自動的に定まることや，コミュニティの中心となるノードがわかることがあげられる．AP をネットワークに適用する際，まず，入力となる similarity を定義しなくてはならない．

4.1 ネットワークにおける similarity

Liu らは，Shortest Path や，Jaccard 係数 (Jaccard coefficient) といったノード間類似度を用いた similarity を以下のように定義している [6].

- Shortest Path(SP)

$$s_{SP}(i,j) = -p(i,j) \quad (6)$$

式 (6) において，$p(i,j)$ はノード i からノード j までの経路長である．ノード i とノード j が近いほど類似度が高くなる．

- Jaccard 係数 (JC)

$$Jaccard(i,j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{7}$$

$$s_{Jaccard}(i,j) = \begin{cases} Jaccard(i,j) - 1 & if\ A(i,j) = 1 \\ Jaccard(i,j) - 2 & if\ A(i,j) = 0 \end{cases} \tag{8}$$

式 (7) において，N_i はノード i に隣接しているノードを表す．ノード i,j の両方に隣接しているノードが多いほど，類似度が高くなる．式 (8) では類似度を負にするため，ノード i,j 間のエッジの有無によって 1 や 2 を式 (7) の値から引いている．$A(i,j)$=1 はノード i,j 間にエッジが存在していることを表す．

Shortest Path では，隣接したノード間の類似度が詳しく表せず，Jaccard 係数では，値域が狭く，類似度が詳しく表せないと考え，本研究では Adamic/Adar[4] による similarity を提案する．Adamic/Adar はリンク予測などに用いられる類似度で，以下の式 (9) で表される．

- Adamic/Adar(AA)

$$AA(i,j) = \sum_{z \in N_i \cap N_j} \frac{1}{log(N_z)} \tag{9}$$

Adamic/Adar を用い，similarity を式 (10) のように定義する．max$AA(i,j)$ は，Adamic/Adar の最大値を表している．

$$s_{AA}(i,j) = \begin{cases} AA(i,j) - \max\limits_{ij} AA(i,j) - 1 & if\ A(i,j) = 1 \\ AA(i,j) - \max\limits_{ij} AA(i,j) - 2 & if\ A(i,j) = 0 \end{cases} \tag{10}$$

4.2 signed ネットワークにおける similarity

次に，AP によるコミュニティ抽出の応用として，signed ネットワークからのコミュニティ抽出を試みた．AP を signed ネットワークへ適用するため，signed ネットワークにおける similarity を定義する．Jaccard 係数をベースに，エッジの種類の組み合わせによって similarity を定義している．

$$New\_Jaccard(i,j) = \frac{CN_1 - CN_2}{|N_i \cup N_j|} \tag{11}$$

式 (11) において CN_1 はノード i,j に対し，同じ種類のエッジで接続されたノードの数を表す (図 3)．逆に CN_2 はノード i,j に対し，異なる種類のエッジで接続されたノードの数を表す (図 4)．直感的には，共通の友人や敵がいる場合，similarity は増加し，逆に一方にとっては友人だが，もう一方にとっては敵となるとなるようなノードが存在する場合，similarity は減少する．Jaccard 係数と同様に，similarity を負にするためノード i,j 間のエッジの有無や種類によって式 (11) の値から，1〜3 を引いている．

$$s_{signed}(i,j) = \begin{cases} New\_Jaccard(i,j) - 1 & if\ A(i,j) = 1 \\ New\_Jaccard(i,j) - 2 & if\ A(i,j) = 0 \\ New\_Jaccard(i,j) - 3 & if\ A(i,j) = -1 \end{cases} \tag{12}$$

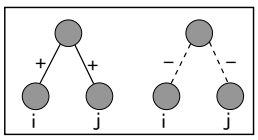

図 3: similarity が増加する組み合わせ

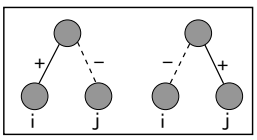

図 4: similarity が減少する組み合わせ

5. 実験

5.1 ネットワークでの実験

AP を用いて 4 つのネットワークからのコミュニティ抽出を行った．与える similarity には前節で紹介した，SP，JC，AA の 3 つを用いている．結果をモジュラリティや計算時間によって評価した．計算時間は，similarity の計算から AP の終了までを計測した．また，モジュラリティ最適化法である GN 法と CNM 法との比較を行った．実験環境は Core2Duo E7500, RAM 4GB のマシンを用い, R 2.10.1 で igraph のパッケージと Frey らのウェブサイト上で公開されている AP のパッケージを用いている．

5.2 結果

用いたネットワークの説明と，結果の表 (モジュラリティQ, コミュニティ数 C，計算時間，AP におけるメッセージ交換の反復回数) を以下に示す．表の AP(SP),AP(JC),AP(AA) は，それぞれ，Shortest Path，Jaccard 係数，Adamic/Adar による similarity を入力とする AP を表す．モジュラリティは，preference の値を 0 から 0.1 ずつ減らし，その中で得られた最大値を記録している．表中の太文字は 1 番良い値であることを表し，下線付き文字は 2 番目の値であることを表している．

- Zachary's Karate Club
  Zachary によって観察された，空手クラブにおける人間関係のネットワーク．ノード数は 34，エッジ数は 78.

メソッド	Q	C	時間	反復回数
AP(SP)	0.35692	2	0.18	115
AP(JC)	<u>0.38749</u>	3	0.22	157
AP(AA)	0.35996	2	<u>0.17</u>	111
GN 法	**0.40129**	5	0.28	-
CNM 法	0.38067	2	**0.10**	-

表 1: Zachary's Karate Club の結果

- The Bottlenose Dolphin Social Network
  ニュージーランドに生息するイルカの群れでの関係を表したネットワーク．ノード数は 62，エッジ数は 159.

メソッド	Q	C	時間	反復回数
AP(SP)	0.49438	4	0.35	136
AP(JC)	0.50225	5	<u>0.31</u>	115
AP(AA)	0.51092	4	0.33	130
GN 法	**0.51938**	5	0.37	-
CNM 法	<u>0.51459</u>	4	**0.11**	-

表 2: The Bottlenose Dolphin Social Network の結果

- Books on US Politics
  Amazon.com で販売されたアメリカの政治に関する書籍の関係を表したネットワーク．同じ購入者に買われた書籍間にエッジが張られる．ノード数は 105，エッジ数は 441.

メソッド	Q	C	時間	反復回数
AP(SP)	0.45708	4	0.83	128
AP(JC)	0.49746	3	<u>0.73</u>	133
AP(AA)	<u>0.51478</u>	4	0.84	149
GN 法	**0.51690**	5	1.5	-
CNM 法	0.50197	4	**0.11**	-

表 3: Books on US Politics の結果

- American College Football Teams
  アメリカのフットボールのチームの関係を表したネットワーク．試合を行った 2 チームの間にエッジが張られる．ノード数は 115，エッジ数は 616.

メソッド	Q	C	時間	反復回数
AP(SP)	0.44203	9	1.73	176
AP(JC)	0.56299	10	0.99	141
AP(AA)	0.59925	11	1.01	138
GN 法	**0.60091**	10	2.87	-
CNM 法	0.56463	5	**0.11**	-

表 4: American College Football Teams の結果

Adamic/Adar による AP は Karate Club 以外のネットワークにおいて，Shortest Path や Jaccard 係数より高いモジュラリティを得ることができた．American College Football Teams や Books on US Politics において，Adamic/Adar による AP は，GN 法より短い計算時間で，CNM 法よりも高いモジュラリティを得ることができた．Adamic/Adar による AP は，比較的短い計算時間で，比較的高いモジュラリティを得ることができ，コミュニティ抽出に有用であると言える．

5.3 signed ネットワークでの実験

AP を用いて 2 つの sigend ネットワークからのコミュニティ抽出を行った．AP の入力には 4.2 節で定義した similarity を用いている．signed ネットワークの有向エッジは，全て無向エッジとみなして実験を行っている．2 つのノード間に正のエッジと負のエッジが両方存在する場合は，負の無向エッジが存在するものとみなしている．

- Gahuku-Gama Subtribes Network
  パプアニューギニアに存在する，16 の部族間の同盟関係や敵対関係を表した signed ネットワーク．正のエッジは同盟関係を表し，負のエッジは敵対関係を表している．AP を用いてコミュニティ抽出を行った結果，図 5 のように，円で囲まれた 3 つのコミュニティが得られた．preference には similarity の中央値を用いた．図 5 で名前が囲んであるノードは exemplar であることを表している．

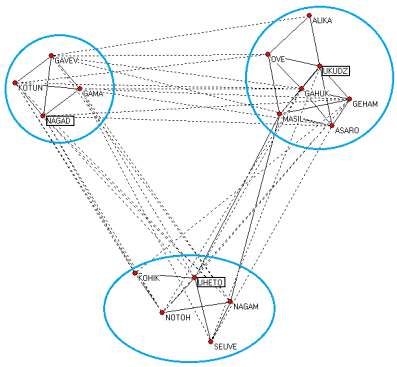

図 5: Gahuku-Gama Subtribes Network のコミュニティ

AP によるコミュニティ抽出の結果，Yang らによって報告されたコミュニティ [5] と全く同じものが得られた．

- Sampson Monastery Networks(friendship)
  Sampson によって作成された，New England の僧院における，18 人の僧侶の友人関係を表した signed ネットワーク．正のエッジは好意を，負のエッジは敵意を表している．実際のネットワークはエッジに 1〜3 の重みが付いているが，実験の際，重みは全て 1 とした．preference=-3 としたところ，以下の表 5 のようなコミュニティが得られた．表中の数字はノードの番号を表し，括弧でまとめられたノードがコミュニティである．下線付きの数字は，exemplar を表している．

AP による結果	(1, 2, 7, 8, 12, 14, 15, 16), (3, 13, 17, 18), (4, 5, 6, 9, 10, 11)
Bo Yang らによる結果	(1, 2, 7, 12, 14, 15, 16), (3, 17, 18), (4, 5, 6, 8, 9, 10, 11, 13)

表 5: Sampson Monastery Networks における結果

AP によるコミュニティ抽出の結果，Yang らによる結果 [5] と非常に近い結果を得ることができた．また，AP ではそれぞれのコミュニティの exemplar を得ることができた．

6. おわりに

本研究では AP を用いてネットワークと signed ネットワークからのコミュニティ抽出を行った．ネットワークにおけるコミュニティ抽出では，比較的短い時間で，モジュラリティ最適化法と同等のモジュラリティを得ることができた．また，signed ネットワークからのコミュニティ抽出では，他手法で報告されている分割と非常に近いコミュニティを得ることができた．

今後の課題としては，signed ネットワークのコミュニティの定量的な評価や，向きを考慮に入れた similarity の提案を行うことが挙げられる．

参考文献

[1] M. E. J. Newman and M. Girvan. Finding and evaluting community structure in networks. *Physical Review E*, 69(2), 2004.

[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phisical Review*, 70(6), 2004.

[3] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[4] L.A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2003.

[5] B. Yang, W. K. Cheung, and J. Liu. Community mining from signed social network. *IEEE Trans*, 19(10):1333–1348, 2007.

[6] Z. Liu, P. Li, Y. Zheng, and M. Sun. Community detection by affinity propagation. Technical Report No.001, Tsinghua University, 2008.