Signed ネットワークからのコミュニティ抽出

Community Detection from Signed Networks

杉原 貴彦
Takahiko Sugihara

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
sugihara@ai.cs.titech.ac.jp

劉 欣
Xin Liu

東京工業大学 大学院情報理工学研究科 計算工学専攻 / 科学技術振興機構 CREST
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology/JST CREST
tsinllew@ai.cs.titech.ac.jp

村田 剛志
Tsuyoshi Murata

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
murata@cs.titech.ac.jp, <http://www.ai.cs.titech.ac.jp/>

keywords: community detection, signed network, modularity.

Summary

Many real-world complex systems can be modeled as networks, and most of them exhibit community structures. Community detection from networks is one of the important topics in link mining. In order to evaluate the goodness of detected communities, Newman modularity is widely used. In real world, however, many complex systems can be modeled as signed networks composed of positive and negative edges. Community detection from signed networks is not an easy task, because the conventional detection methods for normal networks cannot be applied directly. In this paper, we extend Newman modularity for signed networks. We also propose a method for optimizing our modularity, which is an efficient hierarchical agglomeration algorithm for detecting communities from signed networks. Our method enables us to detect communities from large scale real-world signed networks which represent relationship between users on websites such as Wikipedia, Slashdot and Epinions.

1. はじめに

近年、Web 技術の進歩や SNS の普及により、ネットワークからのコミュニティ抽出に関する研究が盛んに行われている。コミュニティとは、エッジが密に結びついた部分ネットワークのことである。ネットワークから抽出するコミュニティの質を定量化する指標として、モジュラリティ [Newman 04] が Newman-Girvan らによって提案されている。ネットワークからのコミュニティ抽出では、このモジュラリティを最適化するアルゴリズム [Newman 04, Clauset 04] がしばしば用いられている。

本稿では signed ネットワークからのコミュニティ抽出を目標とする。Signed ネットワークとは正と負の 2 種類のエッジで構成されたネットワークのことである。正のエッジは友好関係や賛成関係を表し、負のエッジは敵対関係や反対関係を表している。社会学においては集団意思決定の様子を signed ネットワークとして表し、その安定性を分析する研究がなされている [猪原 11]。このような研究は signed ネットワークの構造理解に有効であると考えられる。

Signed ネットワークにおけるモジュラリティは Traag ら [Traag 09] や、Gomez ら [Gomez 09] により提案さ

れている。本稿では、Gomez らのモジュラリティにパラメータを加えてより柔軟にしたモジュラリティを新たに提案する。一方、signed ネットワークのモジュラリティ最適化法に関する研究は不十分であり、大規模な signed ネットワークからの高速なコミュニティ抽出はなされていない。本稿では通常のネットワークにおける高速な最適化手法として知られる CNM 法を signed ネットワーク用に拡張し、大規模 signed ネットワークからのコミュニティ抽出を行うことに成功した。

評価実験として、提案するモジュラリティ最適化手法を用いて人工 signed ネットワークと実 signed ネットワークからコミュニティ抽出を行った。正解となるコミュニティを持つ人工データを生成し実験を行った結果、ノイズが少なく提案手法に含まれるパラメータの値が大きい場合に正解コミュニティが抽出されることを確認した。また、ウェブサイト上でのユーザーの関係を表した実 signed ネットワークからもコミュニティ抽出を行った。その結果、提案手法はノード数が 60,332、全エッジ数が 647,570 の大規模な実 signed ネットワークからもコミュニティ抽出が可能であることを確認した。また、提案手法に含まれるパラメータの影響についても調べ、パラメータの値が

大きい場合は、正のエッジに関してより望ましいコミュニティが得られ、パラメータの値が小さい場合は、負のエッジに関してより望ましいコミュニティが得られることがわかった。

2. 関連研究

本節では通常のネットワークにおけるモジュラリティの説明と、モジュラリティを用いたコミュニティ抽出法について説明する。

2.1 モジュラリティ

ネットワークから抽出したコミュニティの良さを測る指標として、Newman モジュラリティ [Newman 04] がしばしば用いられる。Newman と Girvan は、ある部分ネットワークがコミュニティであるもってもらしさを「コミュニティ内部のエッジ数と、null モデルでのエッジ数の期待値の差」とし、それをモジュラリティによって定量化した。null モデルとは元のネットワークの次数を維持してエッジをランダムに張り替えたネットワークモデルである。モジュラリティ Q は以下の式で表される。

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \quad (1)$$

式 (1) において、 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_{ij} はネットワークの隣接行列の (i, j) 成分、 k_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。 δ はクロネッカーのデルタである。コミュニティには様々な定義が存在するが [Fortunato 10]、この Newman モジュラリティの値が大きくなるような部分ネットワークをコミュニティとするアプローチがしばしば用いられる。

2.2 モジュラリティ最適化法

モジュラリティはコミュニティ分割の質を定量的に評価する指標であり、コミュニティ抽出には、モジュラリティの値が大きい分割を探索するアルゴリズムが必要となる。そのようなアルゴリズムは一般に計算量が多いため、ヒューリスティックな手法が用いられることが多い。本節では、ネットワークのコミュニティ抽出に使用される代表的な 2 つの手法を簡単に紹介する。

§1 GN 法

GN(Girvan-Newman) 法はエッジの媒介中心性を用いてコミュニティを抽出する手法である [Newman 04]。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。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。

§2 CNM 法

CNM(Clauset-Newman-Moore) 法は貪欲法をベースとした最適化手法である [Clauset 04]。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、隣接している 2 つのコミュニティを結合した場合のモジュラリティの増加量を計算する。増加量が最大となる結合を繰り返し、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数十万ノードの規模のネットワークに対するコミュニティ抽出を行なっている。

2.3 モジュラリティの拡張

Newman モジュラリティを拡張する試みがいくつかなされている。有向ネットワークのモジュラリティは Leicht と Newman によって提案され [Leicht 08]、重み付き有向ネットワークのモジュラリティは Arenas らによって提案されている [Arenas 07]。また、Newman モジュラリティは n 部ネットワークには適切でないため、2 部ネットワークへの拡張 [Barber 07, Guimera 07, Murata 09, Suzuki 09] や、3 部ネットワークへの拡張 [Neubauer 09, Murata 10] が提案されている。

3. Signed ネットワークからのコミュニティ抽出

Signed ネットワークとは、図 1 のように正と負の 2 種類のエッジからなるネットワークのことである。この図では実線が正のエッジを、点線が負のエッジを表している。正のエッジは好意や賛成を表し、負のエッジは敵意や反対を表している。実例として、Slashdot や Wikipedia といったウェブサイトにおけるユーザー間の関係を signed ネットワークとして表したネットワークデータが公開されている [Leskovec 10]。

3.1 Signed ネットワークにおけるコミュニティ

Signed ネットワークには負のエッジが存在しているため、通常のネットワークのコミュニティとは負のエッジを考慮しているかどうかという点で異なる。通常のネットワークを対象とした従来法は、内部に多くの負のエッジを含むようなコミュニティを抽出してしまう。

Signed ネットワークにおける望ましいコミュニティ構造では、コミュニティ内部で正のエッジが密で負のエッジ

図 1 Signed ネットワークの例

が疎であり、コミュニティ間で負のエッジが密であり正のエッジが疎であると考えられる．直感的には、友好関係にあるノードが同じコミュニティに多く存在し、敵対関係にあるノードが異なるコミュニティに多く存在しているほど良いコミュニティ構造ということになる．Signed ネットワークからのコミュニティ抽出による対立関係の理解によって、対立関係の原因となるトピックの理解や対立の緩和につながると考えられる [Bogdanov 10]．

3.2 Signed ネットワークにおけるモジュラリティ

Newman モジュラリティは、エッジの正負を考慮していないため、signed ネットワークに対して計算することができない．そのため、signed ネットワークのモジュラリティを定義する必要がある．Signed ネットワークのモジュラリティは Gomez ら [Gomez 09] や Traag ら [Traag 09] によって提案されている．Gomez らのモジュラリティは式 (2) で定義される．

$$Q_{\text{Gomez}} = \frac{m^+}{m^+ + m^-} Q^+ - \frac{m^-}{m^+ + m^-} Q^- \quad (2)$$

式 (2) において、 m^+ は正のエッジの総数、 m^- は負のエッジの総数をそれぞれ表している．式 (2) は正のモジュラリティ Q^+ に正のエッジの割合を掛けた値から、負のモジュラリティ Q^- に負のエッジの割合を掛けた値を引いている． Q^+ はコミュニティ内部で正のエッジがどれほど密であるかを表しており、式 (3) で表される．

$$Q^+ = \frac{1}{2m^+} \sum_{ij} (A_{ij}^+ - \frac{k_i^+ k_j^+}{2m^+}) \delta(c_i, c_j) \quad (3)$$

式 (3) において、 k_i^+ はノード i に接続している正のエッジの数を表す． A^+ は正のエッジの構造を表す隣接行列であり、以下を満たす行列である．

$$A_{ij}^+ = \begin{cases} 1 & (\text{ノード } i, j \text{ 間に正のエッジが存在する場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \quad (4)$$

同様に式 (2) の Q^- はコミュニティ内部で負のエッジがどれほど密であるかを表しており、式 (5) で表される．

$$Q^- = \frac{1}{2m^-} \sum_{ij} (A_{ij}^- - \frac{k_i^- k_j^-}{2m^-}) \delta(c_i, c_j) \quad (5)$$

式 (5) において、 k_i^- はノード i に接続している負のエッジの数を表す． A^- は負のエッジの構造を表す隣接行列であり、以下を満たす行列である．

$$A_{ij}^- = \begin{cases} 1 & (\text{ノード } i, j \text{ 間に負のエッジが存在する場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \quad (6)$$

ここで、 e_{ij}^+ , e_{ij}^- , a_i^+ , a_i^- を以下のように定義する．

$$e_{ij}^+ = \frac{1}{2m^+} \sum_{s \in c_i} \sum_{t \in c_j} A_{st}^+ \quad (7)$$

$$e_{ij}^- = \frac{1}{2m^-} \sum_{s \in c_i} \sum_{t \in c_j} A_{st}^- \quad (8)$$

$$a_i^+ = \frac{1}{2m^+} \sum_{s \in c_i, t \in V} A_{st}^+ \quad (9)$$

$$a_i^- = \frac{1}{2m^-} \sum_{s \in c_i, t \in V} A_{st}^- \quad (10)$$

式 (7),(8) はそれぞれ、コミュニティ i とコミュニティ j の間に存在する正のエッジの数を正のエッジの総数で割ったものと、両コミュニティ間に存在する負のエッジの数を負のエッジの総数で割ったものである．式 (9),(10) はそれぞれ、コミュニティ i に対して張られる全ての正のエッジの数を正のエッジの総数で割ったものと、コミュニティ i への負のエッジの数を負のエッジの総数で割ったものである．式 (7)~(10) を用いると Q^+ と Q^- は

$$Q^+ = \sum_i \{e_{ii}^+ - (a_i^+)^2\} \quad (11)$$

$$Q^- = \sum_i \{e_{ii}^- - (a_i^-)^2\} \quad (12)$$

と簡約化できる．

本稿では、Gomez らのモジュラリティにパラメータを導入し、より柔軟にしたモジュラリティを提案する．提案するモジュラリティは式 (13) で定義される．

$$Q_{\text{signed}} = \alpha Q^+ - (1 - \alpha) Q^- \quad (13)$$

式 (13) において、 α は正のエッジをどれだけ重要視するか、 $1 - \alpha$ は負のエッジをどれだけ重要視するかを表している． α の値にもよるが、 Q_{signed} の値は一般に 1 より小さく、分割によっては負の値も取りうる．この点は Newman モジュラリティと同様である．

例えば図 2 のような signed ネットワークを考えた場合、 α の値によって図 3 のような 4 パターンのコミュニティ分割が得られる． α の値が大きい場合、コミュニティ内の正のエッジを重要視するため、コミュニティ内に負のエッジが多い分割を許容する．逆に α の値が小さい場合、コミュニティ間の負のエッジを重要視するため、コミュニティ間に正のエッジが多い分割を許容する．一方、Gomez らによるモジュラリティではパラメータが固定であり、このような柔軟性を持たない．

抽出するコミュニティに求める特性によって α の値を調整する必要がある．例えば Traag らの論文で扱われている国家間の関係を表した signed ネットワーク [Traag 09] では、負のエッジが紛争を意味している．このような signed ネットワークでは負のエッジがコミュニティ内に含まれるのは望ましくないため、 α の値を小さくすべきだと考えられる．

また、Traag らのモジュラリティは (14) で定義される．

$$Q_{\text{Traag}} = Q_{\text{Traag}}^+ - Q_{\text{Traag}}^- \quad (14)$$

図2 簡単な signed ネットワーク

図3 4つの分割のパターン

式 (14) において Q_{Traag}^+ , Q_{Traag}^- は式 (15), 式 (16) で表される.

$$Q_{Traag}^+ = \sum_i \{e_{ii}^+ - \gamma^+ (a_i^+)^2\} \quad (15)$$

$$Q_{Traag}^- = \sum_i \{e_{ii}^- - \gamma^- (a_i^-)^2\} \quad (16)$$

γ^+ , γ^- は結果のコミュニティの大きさを調整するパラメータであり, γ^+ の値は小さいほどコミュニティが大きくなりやすく, 逆に γ^- の値は大きいほどコミュニティが大きくなりやすい.

正のモジュラリティから負のモジュラリティを引くという考え方は本稿で提案するモジュラリティと同様である. しかし, 正負のモジュラリティの重みが等しいという点や, パラメータの目的がエッジの正負の重要度の調整ではなく, コミュニティの大きさの調整である点で異なる.

3.3 最適化法

Signed ネットワークには負のエッジが存在しているため, 通常のネットワークにおけるモジュラリティ最適化法を直接適用することができない. 本稿では CNM 法 [Clauset 04] をもとに signed ネットワークにおけるモジュラリティ最適化法を提案する. 従来の CNM 法では通常のネットワークにおけるエッジのパターンに注目してモジュラリティの増加量を計算している. しかし, signed ネットワークでは通常のネットワークにはないパターンを考慮しなくてはならないため, CNM 法を直接適用する

ことはできない. 提案手法は, そのようなパターン (後述する式 (25) のようなパターン) や, 負のエッジのパターンも扱えるように CNM 法を拡張したものである.

提案手法ではまず, 一つのノードのみから構成されるコミュニティを初期状態とする. その後, 正負にかかわらずエッジで接続されたコミュニティペアに対し, そのペアを結合して一つのコミュニティとした場合の Q_{signed} の増加量 ΔQ_{signed} を計算し, ΔQ_{signed} が最大となるコミュニティペアを結合していく. 詳しい手順を図4に示す.

図4のステップ (2) において ΔQ_{ij}^+ と ΔQ_{ij}^- はそれぞれ式 (17), (18) で表される.

$$\Delta Q_{ij}^+ = 2(e_{ij}^+ - a_i^+ a_j^+) \quad (17)$$

$$\Delta Q_{ij}^- = 2(e_{ij}^- - a_i^- a_j^-) \quad (18)$$

i, j が正のエッジで接続されている場合には $e_{ij}^+ = 1/2m^+$, $e_{ij}^- = 0$ を式 (17), (18) に代入し, 負のエッジで接続されている場合には $e_{ij}^- = 1/2m^-$, $e_{ij}^+ = 0$ を式 (17), (18) に代入する. 初期状態での行列 ΔQ^+ と ΔQ^- の (i, j) 成分は以下ようになる.

$$\Delta Q_{ij}^+ = \begin{cases} 1/m^+ - 2a_i^+ a_j^+ & (A_{ij}^+ = 1) \\ -2a_i^+ a_j^+ & (A_{ij}^+ = 1) \\ 0 & (\text{それ以外}) \end{cases} \quad (19)$$

$$\Delta Q_{ij}^- = \begin{cases} -2a_i^- a_j^- & (A_{ij}^- = 1) \\ 1/m^- - 2a_i^- a_j^- & (A_{ij}^- = 1) \\ 0 & (\text{それ以外}) \end{cases} \quad (20)$$

図4のステップ (3) において ΔQ_{signed} は式 (21) で表される.

$$\Delta Q_{signed} = \alpha \Delta Q^+ - (1 - \alpha) \Delta Q^- \quad (21)$$

- (1) 一つのノードのみから構成されるコミュニティを初期状態とする
- (2) エッジで接続されたコミュニティ i と j を結合した場合の $\Delta Q_{ij}^+, \Delta Q_{ij}^-$ を計算し, それらを行列 ΔQ^+ と ΔQ^- の (i, j) 成分とする
- (3) ΔQ_{signed} を計算し, その中で最大となるペアを結合し1つのコミュニティとする
- (4) 結合によって変化する $\Delta Q^+, \Delta Q^-$ を更新する
- (5) Q_{signed} が増加しなくなるまでステップ (3), ステップ (4) を繰り返す
- (6) Q_{signed} とその時点でのコミュニティ分割を結果として出力する

図4 Q_{signed} の最適化の手順

図5 エッジのパターン

式 (21) が最大となるコミュニティペア i, j を結合し 1 つのコミュニティ i' とする. ΔQ^+ , ΔQ^- の j 行, j 列を削除し, i 行, i 列を更新する.

図 4 のステップ (4) における具体的な更新手順について説明する. 本手法では式 (17), (18) のうち計算コストがかかる e_{ij}^+ や e_{ij}^- を計算せずに ΔQ^+ , ΔQ^- を更新することで計算時間を短縮する. 結合した 2 つのコミュニティ i, j とコミュニティ x との接続のパターンにより図 5 のような 4 つの場合分けをし, それぞれ適切な更新を行う.

まず, ΔQ^+ の更新手順を示す.

- (1) x が i, j の両方と正のエッジで接続されている場合
 $\Delta Q_{i'x}^+$ は以下のように変形できる.

$$\begin{aligned}\Delta Q_{i'x}^+ &= 2(e_{ix}^+ + e_{jx}^+ - (a_i^+ + a_j^+)a_x^+) \\ &= 2(e_{ix}^+ - a_i^+a_x^+) + 2(e_{jx}^+ - a_j^+a_x^+) \\ &= \Delta Q_{ix}^+ + \Delta Q_{jx}^+ \quad (22)\end{aligned}$$

この場合は, 単純な加算のみで求めることができる.

- (2) x が i のみと正のエッジで接続されている場合
 この場合は, $e_{jx}^+ = 0$ であるので以下のようになる.

$$\begin{aligned}\Delta Q_{i'x}^+ &= 2(e_{ix}^+ - (a_i^+ + a_j^+)a_x^+) \\ &= 2(e_{ix}^+ - a_i^+a_x^+) - 2a_j^+a_x^+ \\ &= \Delta Q_{ix}^+ - 2a_j^+a_x^+ \quad (23)\end{aligned}$$

- (3) x が j のみと正のエッジで接続されている場合
 この場合は, $e_{ix}^+ = 0$ であるので, 式 (23) と同様に変形し,

$$\Delta Q_{i'x}^+ = \Delta Q_{jx}^+ - 2a_i^+a_x^+ \quad (24)$$

と求められる.

- (4) x が i, j と正のエッジで接続されていない場合
 これは x と i, j とが負のエッジのみで接続されていて, 正のエッジが存在しない場合である. これは signed ネットワーク特有のパターンであり, 従来の CNM 法では考慮されていない. この場合は, $e_{ix}^+ = e_{jx}^+ = 0$ であるので

$$\Delta Q_{i'x}^+ = -2(a_i^+ + a_j^+)a_x^+ \quad (25)$$

と求められる.

次に, ΔQ^- の更新手順を示す. 正のエッジの場合と同様に 4 つのエッジのパターンを考える. それぞれのパターンにおける更新式は以下になっている.

- (1') x が i, j の両方と負のエッジで接続されている場合

$$\Delta Q_{i'x}^- = \Delta Q_{ix}^- + \Delta Q_{jx}^- \quad (26)$$

- (2') x が i のみと負のエッジで接続されている場合
 この場合は, $e_{jx}^- = 0$ であるので以下のようになる.

$$\Delta Q_{i'x}^- = \Delta Q_{ix}^- - 2a_j^-a_x^- \quad (27)$$

- (3') x が j のみと負のエッジで接続されている場合
 この場合は, $e_{ix}^- = 0$ であるので以下のようになる.

$$\Delta Q_{i'x}^- = \Delta Q_{jx}^- - 2a_i^-a_x^- \quad (28)$$

- (4') x が i, j と負のエッジで接続されていない場合
 これは x と i, j とが正のエッジのみで接続されていて, 負のエッジが存在しない場合である. この場合は, $e_{ix}^- = e_{jx}^- = 0$ であるので以下のようになる.

$$\Delta Q_{i'x}^- = -2(a_i^- + a_j^-)a_x^- \quad (29)$$

以上の更新を $\Delta Q_{signed} < 0$ となるまで繰り返していく. 更新が停止した時点でのモジュラリティ値とその時のコミュニティを結果として出力する.

提案手法では, Q_{signed} 全体ではなく, 結合によって変化する ΔQ_{signed} のみを計算することで計算量を減らしている. また, ΔQ_{signed} の計算においてもエッジのパターンに注目し, e_{ij}^+ や e_{ij}^- の計算を避けることで計算量を更に減らしている. これらの計算の工夫によって結果が変わることはなく, 計算量を大幅に減らすことができる.

4. 人工 signed ネットワークでの実験

提案した最適化手法を用いて人工 signed ネットワークからのコミュニティ抽出を行う. 以下, 実験はすべて Core 2 Duo, 2.93GHz, RAM 4GB の PC 上で行なっている.

4.1 人工 signed ネットワークの生成法

正解となるコミュニティが存在する人工 signed ネットワークを生成し, その正解とどれだけ近いコミュニティが抽出できるかを調べる. まず, 正のエッジだけで構成された 32 ノードのランダムネットワークを 4 つ生成する. これらのネットワーク内に正のエッジを 144 本ずつ張る. これら 4 つのネットワークを正解コミュニティとする. 次に各正解コミュニティ間に負のエッジを 96 本ずつランダムに張る. この時点で正のエッジは $144 \times 4 = 576$ 本, 負のエッジも $96 \times 6 = 576$ 本である. ノイズとして, 正解コミュニティ内の正のエッジのうちの $p_1 \times 100\%$ を負のエッジに変え, 正解コミュニティ間の負のエッジのうち $p_2 \times 100\%$ を正のエッジに変える. これは Danon の実験 [Danon 05]

図 6 人工 signed ネットワークの例 ($p_1 = 0.05, p_2 = 0.05$) . 赤実線が正のエッジ, 青点線が負のエッジを表す. ノードの色は所属する正解コミュニティに対応している.

で用いられた人工ネットワークを参考に生成した signed ネットワークであり, これを実験に用いる. 図 6 は D3.js (<http://d3js.org/>) の force layout を signed ネットワーク用に拡張したものを用いて, $p_1 = 0.05, p_2 = 0.05$ として生成した signed ネットワークを可視化した図である.

4.2 人工 signed ネットワークでの結果

まず, $p_1 = p_2 = 0.05$ とし, 提案手法によるコミュニティ抽出を行った. また, 計算時間比較のため, 提案手法以外に「結合した場合の Q_{signed} (式 (13)) を計算する手法 (手法 1)」「結合した場合の ΔQ_{signed} (式 (21)) をそのまま計算する手法 (手法 2)」による抽出も行った.

提案手法と手法 1, 手法 2 を用い, $\alpha = 0.5$ とした結果, 全ての手法で正解コミュニティを抽出することができた. 計算時間は表 1 のようになっている. 提案手法が高速にコミュニティを抽出できていることが確認できる.

次に, p_1, p_2 と, α の影響を調べるため, p_1, p_2 のどちらかの値を 0.05 で固定し, もう一方を 0.05 から 0.5 まで変化させた signed ネットワークをそれぞれ 10 個ずつ生成し実験を行った. $\alpha = 0.2, 0.5, 0.8$ としてコミュニティ抽出を行い, 正解とどれだけ近いコミュニティが抽出できるかを調べる. 精度については NMI (Normalized Mutual Information) により評価している. 正解コミュニティの集合 I と抽出されたコミュニティ J の NMI は式 (30) 値で表される. n はネットワークのノード数, n_i は正解コミュニティ i のノード数, n_j は抽出されたコミュニティ j のノード数, n_{ij} は正解コミュニティ i に所属するノードのうち抽出されたコミュニティ j に分類されたノード数をそれぞれ表している. 値が大きいほど精度が

表 1 計算時間

手法	計算時間 (秒)
手法 1	202.0
手法 2	2.85
提案手法	0.35

図 7 p_1 を変化させた場合の NMI

図 8 p_2 を変化させた場合の NMI

良いことを表し, 完全に一致した場合は値が 1 になる.

$$NMI(I, J) = \frac{-2 \sum_{i,j} n_{ij} \log_2(n_{ij}/n_i n_j)}{\sum_i n_i \log_2(n_i/n) + \sum_j n_j \log_2(n_j/n)} \quad (30)$$

図 7 が p_1 を変化させた場合の結果であり, 図 8 が p_2 を変化させた場合の結果である. それぞれの図において赤線, 緑線, 青線が提案したモジュラリティにおいて $\alpha = 0.8, 0.5, 0.2$ とした結果を表しており, 黒線が Gomez らのモジュラリティによる結果を表している.

α の値が大きい場合は NMI の値が大きく, 正解とかなり近いコミュニティが抽出されている. 一方 α の値が小さい場合は, 正解コミュニティとは大きく異なるコミュニティが抽出されていることがわかる. これは α の値が小さい場合は正のエッジがコミュニティ内でどれだけ密であるかがあまり重要視されず, 正解コミュニティ内部での結合が α の値が高い時に比べて起こりにくいためだと考えられる.

Gomez らのモジュラリティについて, p_1 を変化させた場合は NMI の値が減少しているのに対し, p_2 を変化させた場合はあまり変化していない. これは, Gomez らのモジュラリティでは Q^+ , Q^- の重みとしてエッジの正負の割合を用いているためである. p_1 を変化させた場合は正のエッジの数が減り, 負のエッジが増えるため α の値が小さい結果に近づいていく. 逆に p_2 を変化させた場合は α の値が大きい結果に近づいていく. そのため, 図 7, 図 8 のような結果になったのだと考えられる.

p_1 が高い場合のような, 内部の負のエッジを許容するコミュニティを抽出する場合, Gomez らのモジュラリティ

では負のエッジの割合が増えてしまうため精度が悪くなると考えられる．それに対して，本稿のモジュラリティの α を高く設定した場合，内部の負のエッジを許容するコミュニティを高い精度で抽出することができる．

5. 実 signed ネットワークでの実験

5.1 用いる signed ネットワークについて

3 種類の実 signed ネットワークのデータを実験に用いた [Leskovec 10]．各データは Stanford Large Network Dataset(<http://snap.stanford.edu/data/index.html>) で公開されている．まず 1 つ目のデータは Wikipedia のページ管理者に立候補したユーザーに対する投票を表した signed ネットワークである．ノードは Wikipedia ユーザーを表しており，正のエッジは賛成，負のエッジは反対を表している．元々のネットワークは有向ネットワークだが，本稿の実験ではエッジの向きは無視している．また，次数が 1 のノードを削除し，すべてのノードの次数が 2 以上になるように手を加えている．ノード数は 4,786．正のエッジ数は 76,607，負のエッジ数は 21,849 である．

2 つ目のデータは Slashdot.com のユーザー間のタグを表した signed ネットワークである．Slashdot はユーザーが他のユーザーに対して "friend" と "foe" というタグを付けることができ，それぞれ，正のエッジと負のエッジで表される．Wikipedia と同様の前処理を行い，ノード数は 47,726．正のエッジ数は 328,873，負のエッジ数は 110,050 となっている．

最後のデータは Epinions.com のユーザー間の信頼度を表した signed ネットワークである．Epinions.com ではユーザーが投稿したレビューに対して "Very Helpful" から "Off Topic" までの範囲でレートをつけることができる．このレートをもとにエッジを張る．正のエッジはユーザーが他のユーザーのレビューを信頼していることを表し，負のエッジは他のユーザーのレビューを信頼していないことを表している．これも Wikipedia と同様の前処理を行い，ノード数は 60,332．正のエッジ数は 538,530，負のエッジ数は 109,040 となっている．

5.2 提案手法による結果

それぞれの signed ネットワークでのコミュニティ抽出の計算時間は，Wikipedia が平均約 70 秒，Slashdot が平均約 4,800 秒，Epinions が平均約 6,500 秒であった．従来研究では signed ネットワークのモジュラリティの定義にとどまり，その最適化手法についての検討が不十分であったため，せいぜい頂点数が数百の signed ネットワークしか扱えていなかった．数万ノード，数十万エッジの規模の signed ネットワークが扱えたのは CNM 法をベースにした本稿の最適化手法が有効であることを示している．

それぞれの signed ネットワークにおいて， α の値を 0.1 から 0.9 まで 0.1 ずつ増加させコミュニティ抽出を行っ

図 9 コミュニティ間の正のエッジの割合とコミュニティ内の負のエッジの割合 (Epinions)

図 10 最大のコミュニティに対して張られる負のエッジの，負のエッジ全体に対する割合

図 11 最大のコミュニティに対して張られる正のエッジの，正のエッジ全体に対する割合

た．図 9 は，Epinions における各 α での「コミュニティ間に存在する正のエッジの数を正のエッジの総数で割った値」を実線で，「コミュニティ内に存在する負のエッジの数を負のエッジの総数で割った値」を点線でそれぞれ表した結果である．また，図中の縦線上の結果は Gomez らのモジュラリティによる結果である．

図 9 において， α の値が小さい場合，すなわち，負のエッジを重要視したコミュニティ抽出を行った場合は，コミュニティ間の正のエッジは多いが，コミュニティ内の負のエッジは少ないコミュニティ構造となっている．一方， α の値が大きい場合，すなわち，正のエッジを重要視したコミュニティ抽出を行った場合は，コミュニティ内の負のエッジは多いが，コミュニティ間の正のエッジは少ないコミュニティ構造となっている．

図 10, 図 11 は各 signed ネットワークでの結果のうち最大のコミュニティに対して張られる正負のエッジ数について調べたものである．それぞれの図の横軸は α の値を表しており，図 10 の縦軸は最大コミュニティに対して張られる負のエッジの数を負のエッジの総数で割った値を表しており，図 11 の縦軸は最大コミュニティに対して張られる正のエッジの数を正のエッジの総数で割った値を表している．また，図中の円で囲まれている結果は Gomez らのモジュラリティによる結果である．

図 10, 図 11 において， α の値が小さい場合は，最大のコミュニティに対して負のエッジと正のエッジがともに多く張られている．つまり，敵対するノードが他のコミュニティに多く存在しているが，友人となるノードが他のコミュニティに所属する割合も多いことを表す．逆に α の値が大きい場合は，最大のコミュニティに対してどちらのエッジもあまり張られていない．つまり，友人となるノードが同じコミュニティに多く存在しているが，敵対するノードも同じコミュニティに多く所属していることを表す．

以上の図 9～図 11 より， α の値が小さい場合は，コミュニティ間に負のエッジが多く，負のエッジに関して望ましい分割ができていると考えられる．逆に， α の値が大きい場合は，コミュニティ間に正のエッジが少なく，正のエッジに関して望ましい分割ができていると考えられる．本稿のモジュラリティに導入したパラメータが有効に作用していることが確認できる．

5.3 他のモジュラリティとの比較

Traag らのモジュラリティを用いて同様にコミュニティ抽出を行った．図 12 は Epinions においてパラメータ γ^+ , γ^- を変化させた場合の最大コミュニティの大きさ(含まれるノード数)を表している．赤い線が $\gamma^+ = 0.1$ ，緑の線が $\gamma^+ = 0.5$ ，青い線が $\gamma^+ = 1.0$ での結果をそれぞれ表している．縦軸は大きさを表しており，横軸は γ^- の値を表している．Traag らの論文にあるように， γ^- の値が大きく， γ^+ の値が小さいほどコミュニティが大きくなることが確認できる．

また，図 13 は本稿のモジュラリティと Gomez らのモジュラリティを用いた場合の Epinions における最大コミュニティの大きさを表している．横軸は α を表しており，Gomez と表記されているのは Gomez らのモジュラリティによる結果であることを表す． α の値を変化させてもコミュニティの大きさはあまり変化していないことがわかる．

それぞれのモジュラリティの特徴として，本稿のモジュラリティは，パラメータ α の意味が「正のエッジの重要度」であり直感的に理解しやすく， α によってコミュニティの質を調整できる点で優れている．しかし，抽出されるコミュニティの大きさを調整することはできない．それに対して，Traag らのモジュラリティは γ^+ , γ^- とい

図 12 Traag らのモジュラリティによる結果での最大コミュニティの大きさ

図 13 Gomez らのモジュラリティと本稿でのモジュラリティによる結果での最大コミュニティの大きさ

うパラメータを持ち，結果のコミュニティの大きさを調整することができる．しかし，パラメータが2つあり複雑で，調整が難しい．Gomez らのモジュラリティはパラメータがないという点で優れるが，それ故結果の調整を行うことができない．

一般に，適切なパラメータの値を決定することは難しい．Traag らも彼らのパラメータ設定に際してそのように言及している [Traag 09]．適切な α の値は対象とするネットワークやコミュニティ構造に求められる特性によって異なると考えられる．そのため， α を変化させて試行錯誤を行う必要があり，試行錯誤を行う上で本稿で提案した高速なモジュラリティ最適化法は重要であると考えられる．

5.4 負を無視した抽出結果との比較

$\alpha = 1$ とした場合，つまり，負のエッジを無視した結果について説明する．Wikipedia のデータにおいて， $\alpha = 1$ と $\alpha = 0.5$ での結果について比較を行った．図 14 は， $\alpha = 1$ での最大コミュニティと $\alpha = 0.5$ での最大と 2 番目に大きいコミュニティについて表した図である． $\alpha = 1$ での最大コミュニティが 2 つに分裂し， $\alpha = 0.5$ での 2 つのコミュニティのメンバーとなっている．また，図 15, 図 16 は $\alpha = 1$ での最大コミュニティの一部を可視化したものである．ノードの色は $\alpha = 0.5$ での結果において所属しているコミュニティに対応している．図 15 は正と

図 14 $\alpha = 1$ の最大コミュニティがエッジの正負を考慮することによって分裂する様子

負の両方のエッジを表し、図 16 は負のエッジのみを表している。分裂した 2 つのグループの間には 9612 本 (全体の 12.5%) の正のエッジと 4432 本 (全体の 20.3%) の負のエッジが存在している。これらの中には多くの負のエッジが存在しており、敵対していた可能性が高い。しかし、 $\alpha = 1$ では一つのコミュニティとして抽出されているため、負のエッジを無視すると敵同士を同じコミュニティとしてしまうと考えられる。よって、エッジの正負を適切に考慮した本稿のようなコミュニティ抽出を行う必要がある。

6. おわりに

本稿では大規模 signed ネットワークからのコミュニティ抽出を行うため、既存のモジュラリティにパラメータを加えてより柔軟にしたモジュラリティと signed ネットワークにおけるモジュラリティ最適化法を提案した。提案した最適化法により signed ネットワークからコミュニティ抽出を行い、モジュラリティの式に含まれる α とコミュニティの負のエッジや正のエッジとの関係について調べた。数万ノード、数十万エッジの大規模な実 signed ネットワークからのコミュニティ抽出は今までに見受けられなかったが、提案手法によってこれが可能になった。筆者らは CNM の他に Fast Unfolding[Blondel 08] を拡張した手法によるコミュニティ抽出も行っており、CNM と比較して、計算時間は不安定であるが精度に関してはやや上回っているという結果を得ている。今回は向きを無視した signed ネットワークからのコミュニティ抽出を行ったが、今後は向きを扱えるようなモジュラリティとその最適化法について考えたい。

7. 謝辞

本研究は科研費 22300049 の助成を受けたものである。

◇ 参考文献 ◇

[Arenas 07] Arenas, A., Duch, J., Fernandez, A., and Gomez, S.: Size reduction of complex networks preserving modularity, *New Journal*

of Physics, Vol. 9, No. 176, pp. 1–15 (2007)

[Barber 07] Barber, M. J.: Modularity and community detection in bipartite networks, *Physical Review E*, Vol. 76, No. 066102, pp. 1–9 (2007)

[Blondel 08] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, Vol. P10008, No. 10, pp. 1–12 (2008)

[Bogdanov 10] Bogdanov, P., Larusso, N., and Singh, A.: Towards Community Discovery in Signed Collaborative Interaction Networks, in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pp. 288–295 (2010)

[Clauset 04] Clauset, A., Newman, M. E. J., and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 066111, pp. 1–6 (2004)

[Danon 05] Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A.: Comparing community structure identification, *Journal of Statistical Mechanics: Theory and Experiment*, pp. 1–10 (2005)

[Fortunato 10] Fortunato, S.: Community detection in graphs, *Physics Reports*, Vol. 486, No. 3-5, pp. 75–174 (2010)

[Gomez 09] Gomez, S., Jensen, P., and Arenas, A.: Analysis of community structure in networks of correlated data, *Physical Review E*, Vol. 80, No. 016114, pp. 1–5 (2009)

[Guimera 07] Guimera, R., Sales-Pardo, M., and Amaral, L. A. N.: Module identification in bipartite and directed networks, *Physical Review*, Vol. 76, No. 036102, pp. 1–8 (2007)

[Leicht 08] Leicht, E. A. and Newman, M. E. J.: Community Structure in Directed Networks, *Physical Review*, Vol. 100, No. 118703, pp. 1–4 (2008)

図 15 $\alpha = 1$ の最大コミュニティを可視化した図。赤実線が正のエッジ、青点線が負のエッジを表す。ノードの色は $\alpha = 0.5$ で所属しているコミュニティに対応している。

図 16 図 15 のうち、負のエッジのみを可視化した図。白いノードと黒いノードの間に多くの負のエッジが存在している。

- [Leskovec 10] Leskovec, J., Huttenlocher, D., and Kleinberg, J.: Predicting Positive and Negative Links in Online Social Networks, *WWW'2010*, pp. 641–650 (2010)
- [Murata 09] Murata, T.: Modularities for Bipartite Networks, in *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pp. 245–250 (2009)
- [Murata 10] Murata, T.: Modularities for Heterogeneous Networks, in *Proceedings of the 21th ACM Conference on Hypertext and Hypermedia*, pp. 129–134 (2010)
- [Neubauer 09] Neubauer, N. and Obermayer, K.: Towards Community Detection in k-Partite k-Uniform Hypergraphs, in *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs* (2009)
- [Newman 04] Newman, M. E. J. and Girvan, M.: Finding and Evaluating Community Structure in Networks, *Physical Review E*, Vol. 69, No. 026113, pp. 1–15 (2004)
- [Suzuki 09] Suzuki, K. and Wakita, K.: Extracting Multi-facet Community Structure from Bipartite Networks, in *Proceedings of the International Symposium on Social Intelligence and Networking (SIN09)*, pp. 312–319 (2009)
- [Traag 09] Traag, V. and Bruggeman, J.: Community detection in networks with positive and negative links, *Physical Review E*, Vol. 80, No. 036115, pp. 1–9 (2009)
- [猪原 11] 猪原健弘: ニューカムのバランス概念を用いた社会ネットワークの分析, 経営情報学会 2011 年春季全国研究発表大会 (2011)

〔担当委員: 鹿島 久嗣〕

2012 年 7 月 6 日 受理

著 者 紹 介

杉原 貴彦

2011 年東京工業大学工学部情報工学科卒業。同年同大学院情報理工学研究科修士課程進学。2013 年 3 月同大学院修士課程修了予定。及び同年 4 月より楽天株式会社に入社予定。複雑ネットワークの解析に関する研究に従事。

劉 欣

2007 年武漢大学大学院修士課程修了。2011 年東京工業大学大学院情報理工学研究科博士課程修了。博士(工学)。現在は武漢理工大学准教授及び東京工業大学特別研究員(JST CREST)。ウェブマイニングやソーシャルネットワーク解析に関する研究に従事。

村田 剛志(正会員)

1990 年東京大学理学部情報科学科卒業。1992 年同大学院修士課程修了。東京工業大学工学部助手、群馬大学工学部助手、同講師、国立情報学研究所助教授、科学技術振興事業団さきがけ研究 21 研究員(兼任)を経て、2005 年より東京工業大学大学院情報理工学研究科助教授(現在は准教授)。博士(工学)。Web マイニング、リンク分析に関する研究に従事。情報処理学会、日本ソフトウェア科学会、AAAI、ACM、IEEE 各会員。