大規模社会ネットワークからのクラスタ構造の抽出 *

鶴見 敏行 †　　脇田 建 †

２００８年6月１０日

概要

Clauset，Newman，Moore はネットワークをボトムアップかつ貪欲に解析する手法（CNM 法）を提案し、50 万ノード程度までの社会ネットワーク解析を可能としたが、それ以上の規模については実用的な時間内での解析は困難であった。そこで CNM 法の合併の過程を観察した。その結果合併するクラスタサイズの不均衡が計算コストに大きく影響していることが明らかとなった。この観測から、合併するクラスタサイズの均衡が法の速度向上につながると考えた。本稿では、合併時のクラスタのサイズを考慮することにより合併比率を向上させる 3 種類の手法を提案する。提案手法の実験データセットとして、国内最大級のソーシャルネットワーキングサービスより 2006 年 10 月に取得した 550 万ユーザーの友人関係のネットワークを使用した。提案した 3 つの手法を用いたところ、CNM 法に比べ劇的なスケーラビリティの向上がみられた。もっとも速度向上がみられた手法では、100 万ノードに対して 5 分、400 万ノードに対しては 35 分程度で解析する事に成功した。また別の手法では、50 万ノードに対して 50 分（CNM 法より 7 倍早い）で解析でき、モジュール性の向上にも成功した。

1　はじめに

クラスタ構造の抽出は、複雑な社会ネットワークからその固有の性質を知るうえでの最初の重要なステップであり、今までに数々の研究がなされている。[5, 10, 3, 6, 7, 11, 13, 1, 2, 9]

われわれは、Clauset, Newman, Moore によって提案された手法 (CMM 法) [2] を実装し、Social Networking Service(SNS) より取得したネットワークのさまざまな部分集合に対しての解析を試みた。50 万ユーザー以下での解析は可能だった。しかし、さらに大きなネットワークでは実用的な時間での解析は不可能であった。

そこで CNM 法の合併の過程を観察した。その結果合併するクラスタサイズの不均衡が計算コストに大きく影響していることが明らかとなった。この観測から、合併するクラスタサイズの均衡が法の速度向上につながると考えた。

本稿では合併するクラスタの大きさについての不均衡を抑えるために、クラスタサイズの比である**合併比率**を定義し、合併するクラスタの選定の基準として従来用いられてきた**モジュール性**に合

---

* この論文は鶴見敏行さんが東京工業大学理学部情報科学科の学士論文研究を国際学会で発表した論文 [12] をもとに情報リテラシの実習教材に作り直したものです。

† 東京工業大学大学院数理・計算科学専攻

併比率を加味するヒューリスティックを提案する。この提案を用いて実験を行った結果、合併の不均衡を抑えることができ、数百万ノードのネットワークに対する解析も可能となった。

これ以後の本稿の構成は以下の通りである。2 節で関連研究について述べる。3 節で CNM 法を紹介し、その問題点を明らかにする。4 節では CNM 法の欠点を改善するヒューリスティクを提案し、5 節では提案方法を評価する。最後に 6 節で本稿をまとめる。

2 関連研究

クラスタを抽出する手法は大きく分けて二つある。一つは、グラフよりいくつかのノードを**種**として選び、それを含むクラスタを返す手法 [5, 3, 7, 11] である。もう一つは、ネットワーク全体をクラスタに分割する手法 [10, 6, 13, 2, 1, 9] である。

前者の手法は主にウェブの解析に使われている。ウェブの解析は、各ウェブページをノード、ハイパーリンクを**エッジ**とした有向グラフとしてウェブをとらえ、解析する。Kleinberg が提案した *HITS* [4, 5] は、ウェブの適当な部分グラフの中からハイパーリンクで密に結合された**オーソリティ**および**ハブ**を抽出する手法である。オーソリティとは、多くのページからハイパーリンクを張られている著名なページである。ハブとは、リンク集およびブックマークなど、多くの著名なページへハイパーリンクを張っているページである。HITS は、簡単な反復計算によって各ウェブページのオーソリティおよびハブのスコアを計算する。

後者の手法としてさまざまな提案がなされている。Kumar らによる *Trawling* [6] は特によく知られている。彼らは、2 億ページ以上の大規模なウェブのアーカイブに対してこの手法を適用し、10 万個を越えるクラスタの**コア**を発見した。コアとはオーソリティとハブから成るサイズの小さい仮定に基づいている。Clauset らにより提案された手法（CNM アルゴリズム）[2] は、本研究と関連が深いため、次節で説明する。

3 CNM 法

Newman と Girvan は、グラフ G に対するクラスタリング $\mathcal{C}$ が与えられたときに、そのクラスタリングの優劣を評価するための指標として**モジュール性** $Q(G,\mathcal{C})$ を定義し [8] 、モジュール性を指標として貪欲にクラスタを合併する手法を提案した。のちに Clauset，Newman，Moore らはそれにデータ構造上の改善を施した案を提案した (CNM 法) [2]。

CNM 法では、まずそれぞれのノードを独立したクラスタとして表す。そして、全クラスタ対 (c_i, c_j) について、それらを合併したときのモジュール性の上昇度 $(\Delta Q^{\mathcal{C}}_{c_i,c_j})$ を計算する。ここで $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ の定義は以下の通りである。

$$\Delta Q^{\mathcal{C}}_{c_i,c_j} = Q(G, \mathcal{C} \setminus \{c_i, c_j\} \cup \{c_i \cup c_j\}) - Q(G, \mathcal{C}).$$

CNM 法は、最大値を保持する $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ を反復的に合併し新しいクラスタとする。この反復は $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ が正である限り続く。クラスタの合併とともにクラスタ数は単調に減少するため、この手

$\mathcal{C} := \{\{v\} | v \in V\}$;
function $Join(c_i, c_j)$ {
  return $\mathcal{C} \setminus \{c_i, c_j\} \cup (c_i \cup c_j)$;
}
procedure updateDeltaQ() {
  $\forall c_i, c_j \in \mathcal{C}$.
    $\Delta Q^{\mathcal{C}}_{c_i,c_j} := Q(G, Join(c_i, c_j)) - Q(G, \mathcal{C})$;
}
while (**true**) {
  updateDeltaQ();
  $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ が最大となる $(c_i, c_j) \in \mathcal{C}^2$ を選択;
  if $(\max(\Delta Q^{\mathcal{C}}_{c_i,c_j} < 0)$ **break**;
  $\mathcal{C} := Join(c_i, c_j)$;
}

図 1　CNM 法 [2] の概略

法はいずれ停止する。

図 1 に、CNM 法の概要を与えた。updateDeltaQ の操作の費用が高いように思えるが、$\Delta Q^{\mathcal{C}}_{c_i,c_j}$ の再計算を要するのは c_i または c_j のいずれかが直前に合併した場合だけであり、その計算はその時点における値を用いて $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ の値から差分的に計算できる。このため、実際は効率的に計算することができる。また、$\Delta Q^{\mathcal{C}}_{c_i,c_j}$ の最大値を与えるクラスタ対もプライオリティ・キューを利用することで簡単に見つけることができる。

Clauset らは、この手法を用いて Amazon.com から得た数十万ノード規模のネットワークなどを解析した。

われわれは Clauset らの手法を Java によって実装した。データセットとして、わが国の代表的な SNS である mixi(http://mixi.jp/) の友人間関係（マイミクシィへの参照）のネットワークを用いた[*1]。実験環境として、Intel Xeon 2.80GHz，L2 cache = 2MB，メモリ= 4GB を用いた。

実験の結果 [2] で示唆される高効率に反して予想外に解析時間がかかり、1 週間で全体の 10% 程度の解析しかすることができなかった。

計算コストのボトルネックを調べるために、mixi ネットワークの様々な部分集合に対して解析を試みた。mixi はそれぞれのユーザーに対して、登録した順番に ID ナンバーを 1 から順番に与える。$U = \{1, 2, \ldots\}$ はユーザー ID の集合。$F \subset U \times U$ は友人関係の集合とすると、mixi ネットワークは、$G_{\mathrm{mixi}} = (U, F)$ のようなグラフで表せる。このグラフの部分集合を以下のように定義

[*1] 2005 年 11 月に取得した mixi より取得した約 100 万ユーザーからなるネットワーク

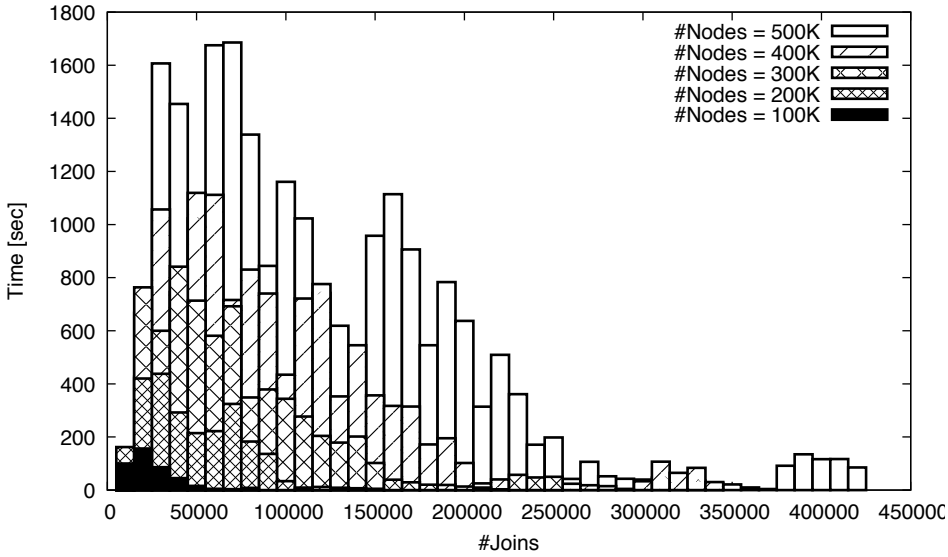

図 2　ネットワークの解析に要した時間の時系列遷移

する。

$$G^n_{\mathrm{mixi}} = (U(n), F \cap (U(n) \times U(n)))$$
$$\text{ただし } U(n) = \{u \in U | u \leq n\}$$

図 2 は mixi ネットワークの様々な部分集合 $G^N_{\mathrm{mixi}}(N = 100\mathrm{K}, 200\mathrm{K}, 300\mathrm{K}, 400\mathrm{K}, 500\mathrm{K})$ に対する、クラスタ抽出が完了するまでの解析に要した時間の時系列遷移である。横軸は累積の合併の回数、縦軸は 10,000 回の合併に要する時間を秒で表している。それぞれのデータセットでは、解析の前半で合併に要する時間がかかり、後半で劇的に減少していることが分かる。

[2] は CNM 法の計算量が $O(md \log n)$ と述べている。ここで n はノード数、m は全エッジ数、d はデンドログラムの高さである。また、疎な社会ネットワークにおいて、m を n、d を $\log n$ と近似でき、計算量は $O(n \log^2 n)$ となると予想している。一方、われわれの実験した実行時間の結果である図 2 のデータに対して $f(x) = ax^b$ を最小二乗近似した結果、$a = 1.51 \cdot 10^{-8} \pm 2.06 \cdot 10^{-8}$, $b = 2.13 \pm 0.104$ と 2 次のオーダーで上昇することがわかった。これより $d \approx \log n$ の性質が少なくとも mixi ネットワークでは現れないことがわかった。

われわれはデンドログラムがどのように成長するかを調べるため、合併履歴を観察したところ、少数の巨大クラスタが、多数の小さなクラスタと合併することにより急速に成長していることが観測でき、この現象のために不均衡なデンドログラムが作られていることが分かった。

図 3 は $G^{500\mathrm{K}}_{\mathrm{mixi}}$ の解析プロセスにおける各合併ごとの**合併比率**を、片対数グラフにプロットした結果である。合併比率の定義を以下に示す。

$$\mathrm{ratio}(c_i, c_j) = \min(|c_i|/|c_j|, |c_j|/|c_i|)$$

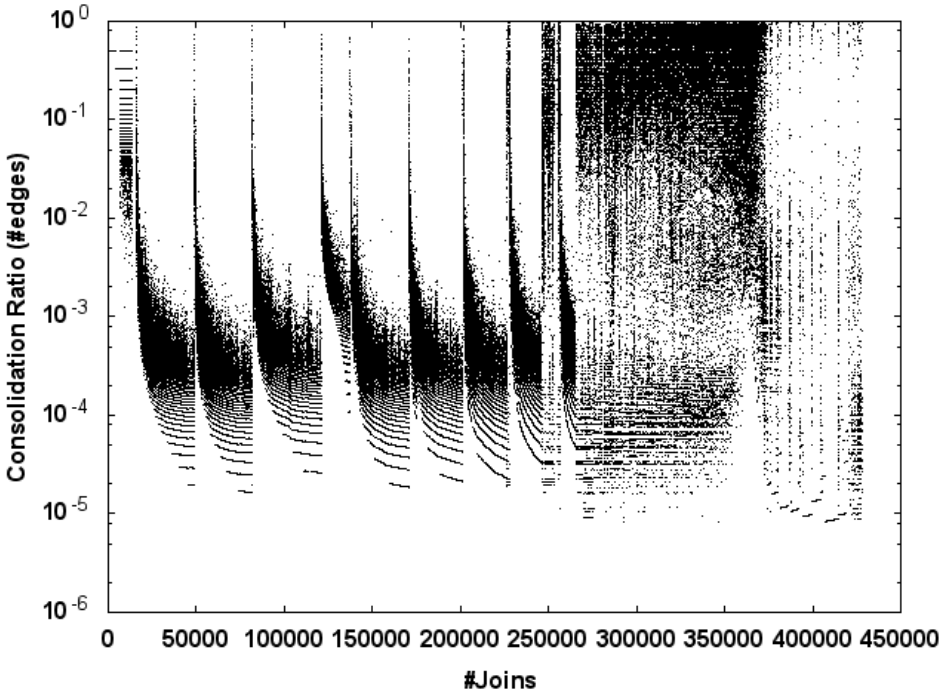

図 3　CNM 法におけるクラスターの合併比率の推移（片対数）

図 3 は n 回目の合併を、$c_k = Join(c_i, c_j)$ とすると、$(n, \mathrm{ratio}(c_i, c_j))$ にプロットしている。$|c|$ はクラスタのサイズである。この図より、解析の前半で、8 つの巨大なクラスタの成長を観測できた。われわれはこの偏って成長したクラスタが CNM 法を mixi ネットワークに適用したときの主な計算速度劣化の原因と考えた。

不均衡な合併はデンドログラムの高さを成長させ、$d \approx \log n$ ではなく、$d \approx n$ となり $O(n^2 \log n)$ となる。

4　提案手法

前節では CNM 法の問題点を指摘した。本節では CNM 法の問題点を改善し、速度を劇的に向上させる 3 つのヒューリスティックを提案する。

提案する手法を図 4 に掲げる。全体の構造は CNM 法とほぼ同じである。異なるのは合併するクラスタの対を決める戦略である。CNM 法では $\Delta Q^{\mathcal{C}}_{c_i,c_j}$ を用いていた。本手法ではそれに加えて合併比率 $(\mathrm{ratio}(c_i, c_j))$ を用いる。このヒューリスティックは、合併の均衡を保ち、平均したクラスタの成長を促すことにより、合併の不均衡による手法の性能の劣化を抑えるように設計した。

さて、本稿ではここまでクラスタサイズ $(|c_i|)$ を定義してこなかった。ここで、3 種類のクラスタサイズの定義を行うことにより、3 種類のヒューリスティック HE，HN，HE' を定義する。以下にそれぞれのヒューリスティックのクラスタサイズ $|c_i|$ の定義を示す。

$\mathrm{ratio}(c_i, c_j) \equiv \min(s(c_i)/s(c_j), s(c_j)/s(c_i))$;
while (**true**) {
  updateDeltaQ();
  $\Delta Q^{\mathcal{C}}_{c_i,c_j}\mathrm{ratio}(c_i, c_j)$ が最大となる
    $(c_i, c_j) \in \mathcal{C}^2$ を選択;
  if $(\max(\Delta Q^{\mathcal{C}}_{c_i,c_j} < 0)$ **break**;
  $\mathcal{C} := Join(c_i, c_j)$;
}

図 4　提案手法の概略

HE　c_i の隣接クラスタ数
HN　c_i 内のノード数
HE'　HE と CNM 法を混同させた定義[*2]

5　評価

本節では、CNM 法と前節で紹介した 3 種類のヒューリスティック (HE、HE'、HN) を用いた手法の評価を行う。実験環境として、Intel Xeon 2.80GHz、メモリ 4GB を用い、Linux(Red Hat Linux version 2.6.16) 上で実験した。4 種類の手法の実装言語には Java(Java 5.0, Java HotSpot Server VM build 1.5.0_06 b-05) を用いた。また、Xeon は CPU を複数内蔵しているが、単一スレッドのみの実装を行い、並列化は行わなかった。

4 種類の手法を用い $G^n_{\mathrm{mixi}}, (n = 50\mathrm{K}, 100\mathrm{K}, \ldots 1,000\mathrm{K})$ を解析・比較した。解析時間を表 1、図 5 に示す。CNM 法が解析できた最大のデータセットは $G^{1,000\mathrm{K}}_{\mathrm{mixi}}$ であり 5.9 時間の解析時間を要した。最も解析時間が短かった手法は HN で、$G^{1,000\mathrm{K}}_{\mathrm{mixi}}$ に対して 5 分以内に解析することができた。HE と HE' は $G^{1,000\mathrm{K}}_{\mathrm{mixi}}$ に対してそれぞれ、36 分と 3 時間の解析時間だったが、HN よりも低速であるものの、データのサイズを考慮すれば十分に実用的である。

次にモジュール性の評価だが、モジュール性の向上を最優先とする CNM アルゴリズムと比較し、われわれの手法では合併比率も考慮したために、モジュール性が悪化することを予想していた。しかし、表 2 に見るように、HE' では CNM 法よりも若干であるがモジュール性が上回った。この結果より、HE' 法では $G^{500\mathrm{K}}_{\mathrm{mixi}}$ において、CNM 法の 7 倍の速度性能を持ち、モジュール性においても $8 - 11\%$ の向上が見られ、速度面・モジュール性の両面で CNM 法を上回る事ができた。

HN 法は、速度面において HE よりも若干の性能を有すが、モジュール性においては CNM 法の $21\% - 28\%$ 劣る結果となった。

図 7 に HN の $G^{500\mathrm{K}}_{\mathrm{mixi}}$ の解析プロセスにおける各合併ごとの合併比率を示す。この図からわかる

[*2] 詳細は [12] をご覧ください。

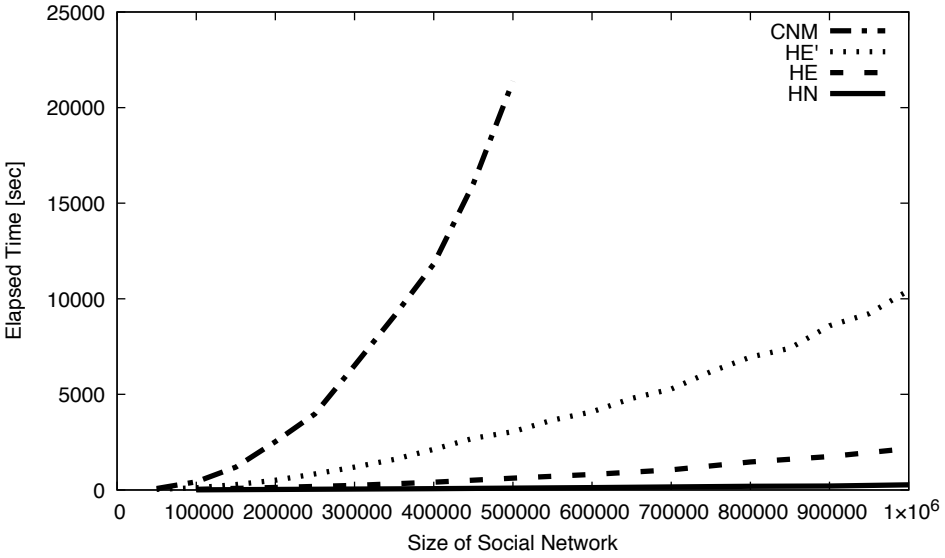

図 5　実行時間比較

表 1　実行時間（秒）

	200K	400K	600K	800K	1M
CNM	2,530	11,800	NA	NA	NA
HE	129	408	814	1470	2170
HE'	511	2,130	4,090	7,410	10,400
HN	25.7	70.0	123	190	268

表 2　モジュール性の比較

	100K	400K	700K	1M
CNM	0.496	0.478	NA	NA
HE	0.413	0.373	0.355	0.339
HE'	0.535	0.529	0.520	0.514
HN	0.421	0.382	0.358	0.351

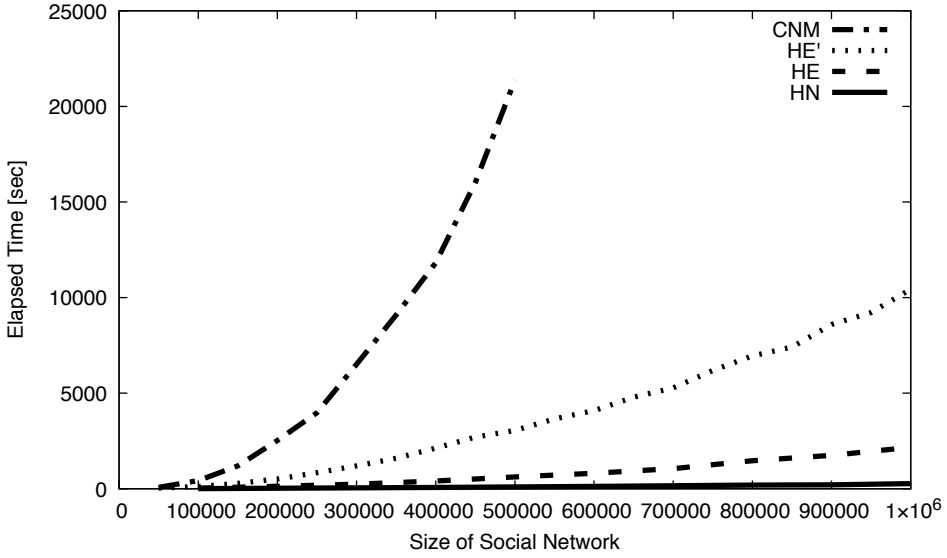

図 6　スケーラビリティ

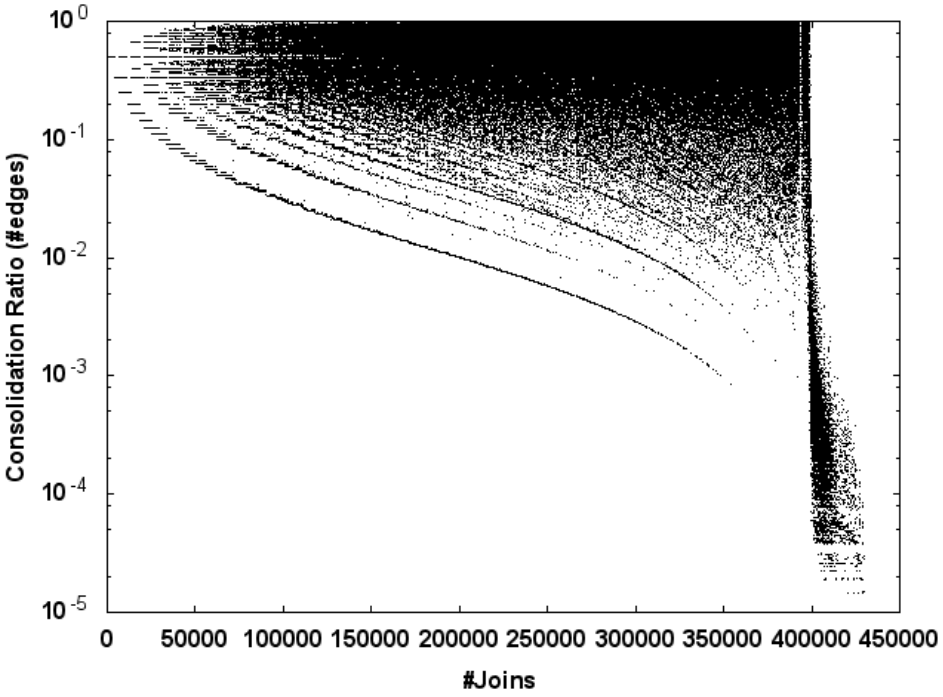

図 7　HN におけるクラスターの合併比率の推移（片対数）

ように、CNM 法と比べ劇的に合併比率の偏りを抑えることができた。

われわれはさらに 2006 年 10 月に取得した約 550 万ユーザーからなる mixi ネットワークに提案するヒューリスティクを適用した。図 6 はそのときの実行時間である。この図より、HE と HN はほぼ線形で実行時間が増加していることがわかる。一方 HE' は、他の 2 つよりも遅いが、それでもなお 550 万ノードに対しても実用的な時間での解析が可能であった。われわれの手法のスケーラビリティは、物理メモリサイズにより制約されており、HE 法と HN 法は物理メモリ不足のため 550 万ノードが実行できなかった。

6 まとめ

本研究では従来の Clauset, Newman, Moore が提案した手法のボトルネックが合併の不均衡にあることを明らかにした。そして、3 種類のヒューリスティックを提案し、ボトルネックを取り除くことによって 500 万ノード以上のネットワークの解析も可能にした。

今後の課題としては、データ構造の改善によりより大規模なネットワークに対する解析が可能になると考える。また、並列化を施すことにより、数億ノード規模のネットワークに対する解析にも対応できると考えられる。

謝辞

本研究の一部は文部科学省科学研究費助成金（18300041 号）の援助を受けています。本稿の草稿について貴重なコメントをいただいた越田港さんに感謝します。

参考文献

[1] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma. Block-level link analysis. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pp. 440–447, New York, NY, USA, 2004. ACM.

[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, Vol. 70, p. 066111 (6 pages), 2004.

[3] Jeffrey Dean and Monika R Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, Vol. 31, No. 11–16, pp. 1467 – 1479, 1999.

[4] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring Web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space—structure in Hypermedia Systems: Links, Objects, Time and Space—structure in Hypermedia Systems*, HYPERTEXT '98, pp. 225–234, New

York, NY, USA, 1998. ACM.

[5] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, Vol. 46, No. 5, pp. 604–632, September 1999.

[6] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for emerging cyber-communities. *Computer networks*, Vol. 31, No. 11, pp. 1481–1493, May 1999.

[7] Joel C. Miller, Gregory Rae, Fred Schaefer, Lesley A. Ward, Thomas LoFaro, and Ayman Farahat. Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 444–445, New York, NY, USA, 2001. ACM.

[8] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, Vol. 69, p. 026113 (16 pages), 2004.

[9] Mikael Onsjö and Osamu Watanabe. A simple message passing algorithm for graph partitioning problems. In Tetsuo Asano, editor, *in Proceedings of 17th International symposium, ISAAC 2006*, Vol. 4288 of *LNCS*, pp. 507–516, Kolkata, India, December 2006. Springer.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[11] Masashi Toyoda and Masaru Kitsuregawa. Creating a Web community chart for navigating related communities. In *Proceedings of the 12th ACM Conference on Hypertext and hypermedia*, HYPERTEXT '01, pp. 103–112, New York, NY, USA, 2001. ACM.

[12] Ken Wakita and Toshiyuki Tsurumi. Finding community structure in mega-scale social networks:[extended abstract]. In *Proceedings of the 16th international conference on World Wide Web*, pp. 1275–1276. ACM, 2007.

[13] Fang Wu and Bernardo A Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp. 331–338, 2004.