2006－ICS－142（17）
2006／1／13

WWW ナビゲーション向けコミュニティ分割手法に関する一考察

安藤　潤　吉井　伸一郎

北海道大学大学院 情報科学研究科

近年、ネットワークトポロジーからコミュニティ構造を検出する多くのアルゴリズムが提案されており、複雑ネットワークの中のコミュニティ構造の解析が注目を集めている。これらのコミュニティ検出手法を、急速に成長を続けその重要性が認識されている WWW(World Wide Web) において応用することは、WWW 上でのナビゲーションに役立つと考えられる。本研究では、WWW ネットワークの例としてブログのネットワークを取り上げ、そのコミュニティ構造を複数のアルゴリズムにより抽出し、それらの結果を比較することによって WWW に適したコミュニティ分割手法について考察する。

Discussion about Community Extraction Methods for WWW Navigation

Jun ANDO　Shinichiro YOSHII

Graduate School of Information Science and Technology, Hokkaido University

Recently, many algorithms for detecting community structures in a network have been proposed, and community structures in complex networks have been attracting much attention. Applying these community detecting algorithms for the WWW (World Wide Web) is thought to be useful for navigating on the WWW. In our research, we focus on the network topology of weblogs and discuss about community detecting algorithms for WWW navigation by comparing the community structures which are extracted by several algorithms.

1. はじめに

様々な分野において、研究の対象となる系を要素と相互作用からなるネットワークとして捉え、解析することは効果的な手法であることが示されてきており、多くの研究者がそのような複雑ネットワークの研究に従事している [1, 2, 3]。特に、複雑ネットワークのコミュニティ構造の検出と解析が注目され、コミュニティ構造を発見する様々なアルゴリズムが提案されている [4, 5, 6, 7, 8]。

ここで、ウェブページを頂点、ウェブページ間のハイパーリンクを辺とみなすと、WWW は巨大なネットワーク構造を形成している。この WWW ネットワークは複雑ネットワークの代表的な例であり、その構造的特徴や性質が詳しく調査されている [9, 10, 11]。さらにウェブページの接続関係のみからウェブページを評価するランキングアルゴリズムが提案されており [12]、検索エンジンなどに応用されている。このように、現在も爆発的に成長を続ける WWW において、文書解析に依らない手法でウェブページの評価、分類などが可能になることは WWW ナビゲーションにとって重要であると考えられる。

WWW の中でも特にブログについて注目すると、ブログはそのトラックバックシステムにより双方向的なリンク構造を形成しており、ブログの著者であるブロガーの興味や関心を反映したネットワーク構造となっている。このため、ブログネットワークは WWW ネットワークより強いコミュニティ構造を有していると考えられる。Technorati[1]は、2005 年 12 月の時点で 2300 万のブログとそれらのブログの間の 16 億のリンクをトラッキングしているといい、なおもその数を増やしている。ブログに関する研究としては、ブログ空間上での情報の広がり方や、ネットワーク構造的な発展に関する研究がある [13, 14, 15]。

本研究では、WWW ネットワークの例としてブログのネットワーク構造に着目し、そのコミュニティ構造を抽出する。ブログのリンク構造を解析するため、自動的にデータを収集するクローラーを開発し、ブログネットワークのコミュニティ構造を 3 種類のアルゴリズムを用いて抽出する。これらの抽出されたコミュニティ構造の結果から、WWW ナビゲーションに適したコミュニティ分割について考察する。

[1] http://technorati.com/

2. ブログネットワークの解析

ブログを頂点、ブログ間の参照リンク、トラックバックリンクを辺とすると、ブログ空間はネットワークとして捉えることができる。ブログ間の参照リンクやトラックバックリンクはブロガーの興味や関心によって生成されると考えられ、ブログネットワークはブログ間の関連を表していると考えられる。ここではブログネットワークのコミュニティ構造の解析のため、データの収集を行い、解析対象の構造的、統計的特徴を調査する。

2.1 解析対象

本研究では国内大手のブログホスティングサイトのひとつである livedoor ブログ[2]を解析対象とした。livedoor ブログは非常に多くのユーザーに利用されており、ブログファン[3]によると、livedoor ブログにおいて 2005 年 11 月にブログを更新したユーザーの数はおよそ 27 万人であるとされている。

2.2 データ収集

ブログネットワークのコミュニティ構造を解析するために、ブログ記事のタイトル、本文、日付、トラックバックリンクを HTML ドキュメントの解析により収集するクローラを開発した。開発したクローラを実行した結果、34,820 のブログ上をクローリングし、3,345,876 の記事を収集した。これらの収集した記事を解析した結果、4,056,471 の参照リンクと 1,018,595 のトラックバックリンクを発見した。図 1 はブログサイトあたりの記事数の分布である。この図からブログサイトあたりの記事数がベキ法則に従うことがわかった。さらに、記事あたりの参照リンク数およびトラックバックリンク数の分布を図 2 および図 3 に示す。トラックバックリンク数とは記事へトラックバック Ping が送信された回数である。どちらもベキ法則にしたがっていることがわかった。

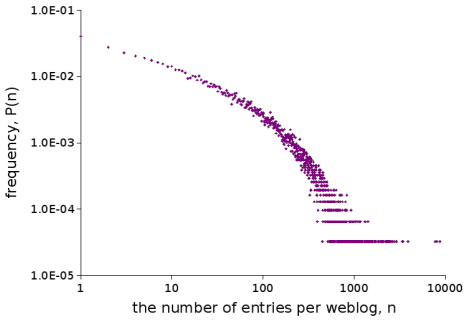

図 1: ブログサイトあたりの記事数の分布

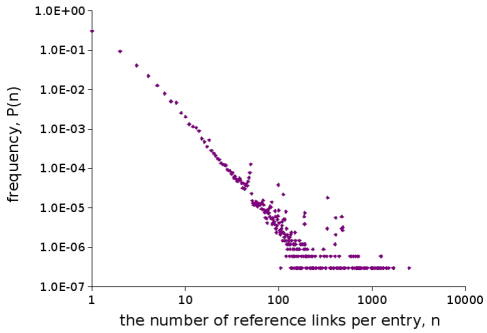

図 2: 記事あたりの参照リンク数の分布

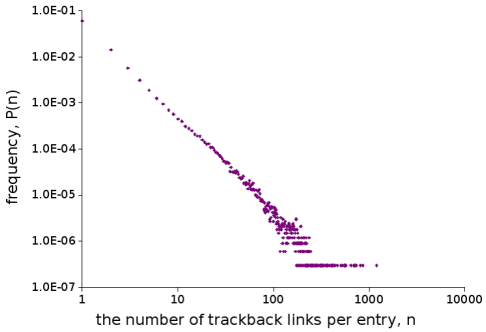

図 3: 記事あたりのトラックバックリンク数の分布

[2]http://blog.livedoor.com/
[3]http://www.blogfan.org/

2.3 ブログネットワークの構成

クローラにより収集したデータから頂点をブログ、辺をブログ間の参照リンクおよびトラックバックとするブログネットワークを構成する。ここで、より互いに関連の強いブログ同士は頻繁に参照リンクやトラックバックを用いてコミュニケーションを行っていると考えると、このコミュニケーションの回数によってブログネットワークの辺に重み付けすることができる。つまり、ブログネットワークは重み付き有向グラフとして表現することができる。こうすることにより、ブログ間の関連の強さに基づいたコミュニティの抽出を行うことができると考えられる。重み付きブログネットワークの辺の重みは以下の式で表される。

$$W_{ij} = \sum_{e_i \in i, e_j \in j} E_{e_i e_j} \tag{1}$$

ここで、e_i はブログ i の記事を表し、$E_{e_i e_j}$ は記事 e_i から記事 e_j への参照リンクまたはトラックバックの数を表している。提案されている多くのコミュニティ抽出アルゴリズムは辺の向きを考慮しないため、コミュニティ抽出の際は双方向の辺の重みを足し合わせることによって無向辺の重みとした。

2.4 ブログネットワークの特徴解析

ブログネットワークを構成した結果、頂点の数が 30,871、辺の数が 70,168 となる有向グラフを得た。このブログネットワークの次数分布を図 4、図 5 に示す。これらの図より、次数分布がベキ法則に従うことが示され、ブログネットワークがスケールフリー特性を持つことがわかった。このネットワークの頂点のうち、孤立した頂点、すなわち次数が 0 の頂点の数は 13,042 であった。また、このネットワークの最大弱連結成分 (Giant Weakly Connected Component) は。17,216 の頂点と 55,934 の無向辺からなっていた

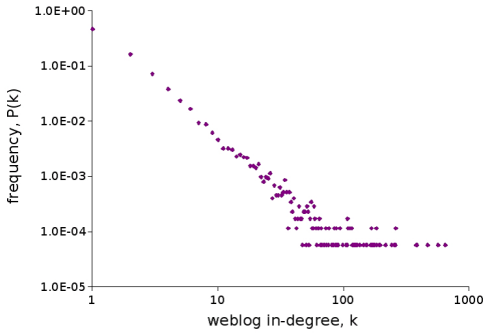

図 4: ブログネットワークの出次数分布

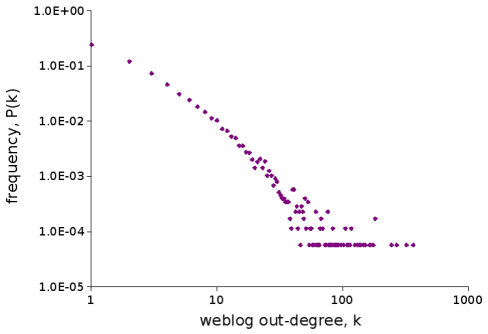

図 5: ブログネットワークの入次数分布

ネットワークの大域的な特徴を解析するための統計的指標として、平均最短経路長 (Average Shortest Path Length, L) およびクラスタリング係数 (Clustering Coefficient, C) がある。クラスタリング係数はネットワークのリンク構造が局所的にどの程度密であるかを示す指標であり、Watts によって提案された [1]。頂点 i のクラスタリング係数は以下の式で表される。

$$C_i = \frac{E_i}{k_i(k_i - 1)} \tag{2}$$

ここで、k_i は頂点 i の近隣の頂点の数、E_i は頂点 i の近隣間の辺の数である。ネットワーク全体のクラスタリング係数はすべての頂点のクラスタリング係数の平均から求められる。ブログネットワークを無向グラフとしたときの最大連結成分についてこれらの特徴量を解析した結果を表 1 に示す。この最大連結成分は 17,216 の頂点と 55,934 の無向辺からなっていた この結果をランダムネットワークの場合と比較する。解析対象と同等の頂点数と辺数を持つランダムネットワークの平均最短経路長およびクラスタリング係数は以下の式 (3)、式 (4) によって近似される。

$$L_{rand} \sim \frac{\log n}{\log \bar{k}} = \frac{\log 17216}{\log 6.95} = 5.0308\ldots\ , \tag{3}$$

$$C_{rand} \sim \frac{\bar{k}}{n} = \frac{6.95}{17216} = 4.03694\cdots \times 10^{-4}\ . \tag{4}$$

表 1: ブログネットワークの統計的特徴量

特徴量	値
平均次数 ($\bar{k}$)	6.95...
平均最短経路長 (L)	4.52813...
クラスタリング係数 (C)	0.10561...

比較の結果、ブログネットワークの平均最短経路長はランダムネットワークの平均最短経路長とほぼ等しく ($L \sim L_{rand}$)、ブログネットワークのクラスタリング係数はランダムネットワークのクラスタリング係数に対して十分大きい ($C \gg C_{rand}$) ことがわかった。すなわち、Watts によるスモールワールドの定義より、ブログネットワークはスモールワールド特性を持っていることがわかった。

3. ネットワークにおけるコミュニティ抽出手法

様々な複雑ネットワークのトポロジーにおけるコミュニティ構造の解析が近年注目を集めており、ネットワークトポロジーからコミュニティ構造を抽出する多くの手法が提案されている。ここで、ネットワーク構造におけるコミュニティとは簡単に互いに密に繋がった頂点の集合であるとする。Danon らは数多くのコミュニティ分割アルゴリズムにおけるその手法や計算量、正確性について比較した [16]。しかし、彼らの実験では解析対象とするネットワークはコミュニティ構造が埋め込まれた単純なランダムネットワークであるため、ブログネットワークのようなスケールフリー特性を持つネットワークにおいて同等の正確性が議論できるとは限らない。ここでは本研究で用いるコミュニティ抽出アルゴリズムについて説明する。

3.1 Modularity の最大化に基づく手法

Newman らはネットワークのコミュニティ分割の質を評価する Modularity という指標を定義し、この Modularity を最適化することによりコミュニティ構造を抽出する手法を提案した [5]。ネットワークにおいてあるコミュニティ分割が与えられたとき、Modularity Q は以下の式で表される。

$$Q = \sum_i (e_{ii} - a_i^2) \tag{5}$$

ここで、$\mathbf{e}$ はコミュニティ i に属する頂点とコミュニティ j に属する頂点を接続する辺の数の割合 e_{ij} を要素とする $k \times k$ の対称行列であり、$\mathbf{a}$ はコミュニティ i に属する頂点に接続する辺の数の割合を表し、$a_i = \sum_j e_{ij}$ である。Modularity Q は、抽出されたコミュニティ内の辺の数がランダムネットワークの場合より多くない場合は 0 に近くなり、コミュニティ内の辺の数が多くなればなるほど 1 に近づく。Clauset らは Modularity を最大化する分割を探索することによりコミュニティ抽出を行う手法を提案した (以下この手法を Clauset 大域的アルゴリズムと呼ぶ)[17]。このアルゴリズムは疎なネットワークに対して非常に小さな計算量でコミュニティ分割を行うことができる。このアルゴリズムはネットーワーク全体を排他的にコミュニティへと分割するアルゴリズムである。このアルゴリズムはもともと重み無しネットワークのための分割アルゴリズムであるが、重み付きネットワークへの応用も可能である [18]。我々は重みがない場合とある場合について結果を比較した。

3.2 Local Modularity の最大化に基づく手法

Clauset は局所的なコミュニティ抽出の質を評価する Local Modularity という指標を提案し、ネットワークの局所的なリンク構造から Local Modularity を最大化するコミュニティ構造を抽出する手法を提案した (以下この手法を Clauset 局所的アルゴリズムと呼ぶ)[7]。この手法は、与えられたシードとなる頂点からスタートし、コミュニティ内部の辺の数がコミュニティ内の頂点から外の頂点へ接続する辺の数に対して大きくなるようなコミュニティの境界をグリーディ法により探索するアルゴリズムである。この手法の主な特徴は、局所的なコミュニティ構造の抽出においてネットワーク全体に関する知識が必要でないということと、ひとつのコミュニティを抽出するための時間計算量が $O(k^2 d)$ と小さいことである。ここで、k は探索する頂点の数、d は平均次数である。

3.3 最大流アルゴリズムに基づく手法

Flake らは最大流アルゴリズムに基づいてシードとして与えられた頂点の周辺のコミュニティ構造を抽出する手法を提案した。この手法では、シードとなる頂点とその d 次の近隣の頂点を含む部分グラフにおいて最大流アルゴリズムを適用し、シードの頂点から不飽和辺をたどって到達できる頂点をコミュニティのメンバーとする。このアルゴリズムを用いる欠点としては、スケールフリー特性を持つブログネットワークからある頂点の d 次の近隣を抽出すると、ハブの存在によりその近隣の数が非常に大きくなってしまう場合があり、このとき計算量が膨大になってしまうという点である。

4. コミュニティ抽出結果の比較

ブログネットワークの最大連結成分について上述の 3 つの手法によりコミュニティを抽出しその結果を主に抽出されたコミュニティの大きさについて比較する。

4.1 Clauset 大域的アルゴリズムによる抽出結果

ブログネットワークについて辺の重みを考慮しない場合と辺の重みを考慮する場合の両方においてそのコミュニティ構造を Clauset 大域的アルゴリズムによって抽出した。図 6 は抽出されたコミュニティの大きさの累積分布を示す。 重み無しネットワークにおけるコミュニティ抽出の結果、123 の多数の小さなコミュニティと少数の大きなコミュニティへ分割された。抽出されたコミュニティのうち、最大のコミュニティは 1316 の頂点を有していた。このコミュニティのようにネットワークの大きさに対して非常に大きなコミュニティでは、コミュニティ内のブログを調べてみると共通する話題を見つけることが困難であり、ブログの話題を反映した結果であるとは言いがたかった。一方、重み付きネットワークにおけるコミュニティ抽出の結果、385 のコミュニティに分割された。重み無しの場合と重み付きの場合を比較すると、重み付きにした場合の方が重み無しの場合に比べて非常に大きなコミュニティと非常に小さなコミュニティの数が減少し、コミュニティのサイズがより平均化されていることがわかる。これは、辺の重みを考慮することによって、より多くのコミュニケーションを行っていたコミュニティが大きなコミュニティから分離したり、小さなコミュニティ同士がその間の重みの大きい辺によって結合したことによると考えられる。図 7 は辺の重みを考慮しない場合に、リンク構造と交わされている話題の両方の面において明らかに 2 つのクラスタからなるコミュニティが抽出されたとき、辺の重みを考慮することによってこれらの 2 つクラスタが分離した例を示している。 以上のように、ネットワークの辺の重みを考慮することによりコミュニティ分割の結果が改善されることが期待される。

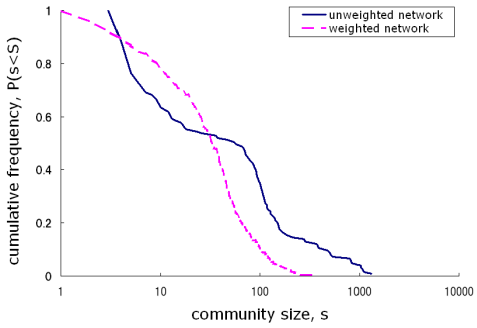

図 6: Modularity の最大化に基づく手法によるコミュニティサイズの分布

4.2 Clauset 局所的アルゴリズムによる抽出結果

探索する頂点の数を 200 に設定し、Clauset 局所的アルゴリズムをブログネットワークに適用して各頂点をシードとしてコミュニティを抽出した。図 8 は抽出されたコミュニティの大きさの分布を表している。 この分布を見ると、25 までの大きさのコミュニティが多く抽出され、それ以上の大きさについては特にピークがみられない。抽出されたコミュニティの中を調べると、大きさが 50 くらいまでのコミュニティではある特定の話題が扱われているとみなせるコミュニティが数多く抽出された。一方、抽出された大きなコミュニティでは複数の話題が存在したり、特定の話題を見つけられなかったりした。このため、WWW ナビゲーションへ応用する際にはコミュニティの最大サイズを 50 程度に設定すると良いと考えられる。

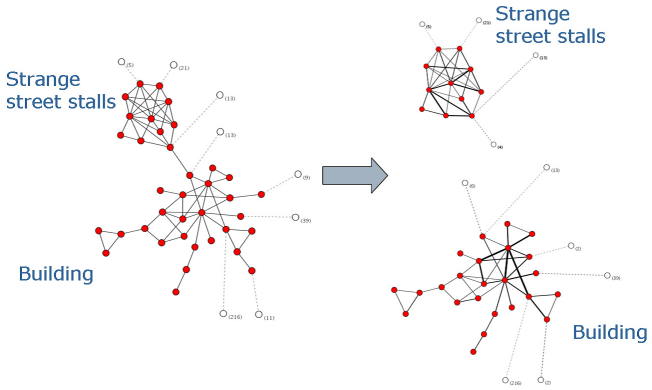

図 7: 辺の重みを考慮することによってコミュニティ抽出結果が改善する例

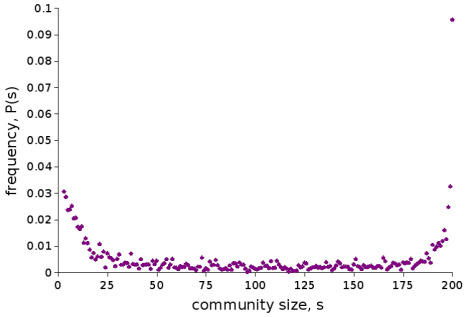

図 8: Clauset 局所的アルゴリズムに基づく手法により抽出されたコミュニティのサイズ分布

4.3 Max-flow アルゴリズムに基づく手法

Max-flow アルゴリズムに基づく手法によるコミュニティ抽出において、設定可能なパラメータは次の 3 つである。ひとつめは与えられた頂点から何次の近隣まで探索するかを決める深さ d、ふたつめは計算の打ち切りの基準となる最低メンバー数、みっつめは Max-flow アルゴリズムを繰り返し適用する回数である。本研究では $d = 2$、最低メンバー数を 30、繰り返し回数の上限を 10 回と設定した。さらに計算量を抑えるために、d 次の近隣を得る際に得る頂点数の上限を 200 と設定した。以上の設定でコミュニティ抽出を行った結果、コミュニティサイズの分布は図 9 のようになった。 この結果、コミュニティの大きさに最低の 30 と 150 のあたりにピークがあることがわかった。この分布の形については詳しく解析していないため、なぜこのような分布になったのかは不明であるが、何らかの意味があると考えられる。また、この手法によって抽出されたコミュニティの特徴として、次数の大きな頂点が抽出された多くのコミュニティに頻繁に出現しているということが挙げられる。

5. 考察

ネットワークのコミュニティ構造を抽出するアルゴリズムは多数提案されているが、どのようなアルゴリズムが WWW ナビゲーションに適しているかについては議論されていない。このため、本研究ではいくつかのコミュニティ分割手法をブログネットワークに適用しその結果を比較した。その結果、まずひとつにブログネットワークをコミュニケーションの頻度で重み付けすることによってブログの内容をよく反映したコミュニティが抽出されることが明らかになった。これは、コミュニケーションの頻度がブログ間の関連性を表しているためであると考えられる。また、Clauset 大域的アルゴリズムでは、ネットワークの重みを考慮することによって多少改善されたが、コミュニティの大きさに偏りが見られた。この中の非常に大きなコミュニティではブログの内容を反映したコミュニティが抽出されたとは言い難かった。本研究では抽出されたコミュニティの評価は人の判断によって行ったが、一般的にブログのネットワーク構造からコミュニ

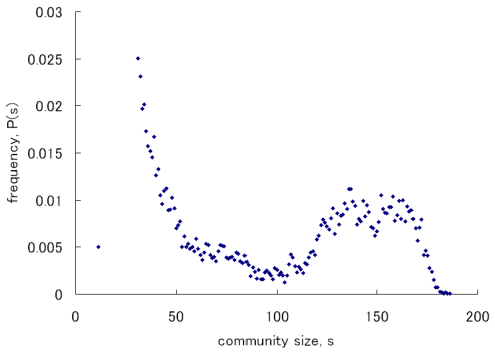

図 9: Max-flow アルゴリズムに基づく手法によるコミュニティサイズの分布

ティを抽出した際、あらかじめ分類情報などがないため、その抽出結果について評価することが困難である。このため、コミュニティ抽出による結果を評価する手法が必要である。また、これらの手法を WWW ナビゲーションへ応用する際に、その精度の他に重要になると考えられるのが計算量の問題であり、即時性を求めるとより高速な手法が要求されると考えられる。

6. おわりに

本研究では、ブログのネットワーク構造についてその構造的特徴および統計的特徴を調査し、3 種類のコミュニティ分割手法によってどのようなコミュニティ構造が抽出されるか比較した。今後の課題としては、ブログネットワークのコミュニティ分割についてその結果を評価手法の提案がある。

参考文献

[1] D. J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, 1999.

[2] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of networks : From biological nets to the Internet and WWW*. Oxford University press, Oxford, 2003.

[3] Alberto-Laszlo Barabasi. *Linked: The New Science of Networks*. 2002.

[4] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, Vol. 35(5), pp. 66–71, 2001.

[5] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, Vol. 69, p. 026113, 2004.

[6] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, Vol. 69, p. 066133, 2004.

[7] Aaron Clauset. Finding local community structure in networks. *Phys. Rev. E*, Vol. 72, No. 026132, 2005.

[8] Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M. Gleiser, and Roger Guimera. Community analysis in social networks. *European Physics Journal B*, Vol. 38, No. 2, pp. 373–380, 2004.

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

[10] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 509, 1999.

[11] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *In Proceedings of the 9th international WWW Conference*, pp. 309–320, 2000.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *technical report*, 1998.

[13] Eytan Adar, Li Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. *In Proceedings of the 13th International WWW Conference*, 2004.

[14] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Commun. ACM*, Vol. 47, No. 12, pp. 35–39, 2004.

[15] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. *In Proceedings of the 12th International WWW Conference*, pp. 568–576, 2003.

[16] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *J. Stat. Mech.*, 2005.

[17] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, Vol. 70, No. 06111, 2004.

[18] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, Vol. 70, No. 056131, 2004.