任意ノードの視点からのコミュニティ抽出法

Quantifining the Concept Definition of Community and Community Mining Algorithm from the Perspective of Arbitrary Node

高橋篤　　荒井幸代
Atsushi Takahashi　Sachiyo Arai

千葉大学大学院工学研究科
Graduate School of Engineering

A community has been generally discussed as a subgraph of which links are connected densely in a network. Since there is no quantitative criterion of a community, the existing methods of extracting a community have been evaluated by introducing a contents analysis in combination. Existing community mining methods need entire structure of the network. Therfore, it is defficult for existing community mining methods to target large scale network. However, network which need community mining is large scale. In this paper, we propose the quantitative definition and the criterion which take relative differences of connectivity between internal and external of the community into consideration. We show the proprieties of our proposed criterion through some experimental results, and discuss the structures of extracted community by using the existing methods. And, we suggest the explore community mining method which target large scale network.

1. はじめに

社会における人のつながりやインターネットを介した情報共有，また論文の被引用関係などを表すネットワークには，スケールフリー性やスモールワールドといった特徴を持つことが近年の複雑ネットワークの研究成果として注目されている．なかでもネットワークから，なんらかの意味を持つコミュニティ抽出法の研究が活発に行われている [Reichardt 06][Zhang 07][Tanigawa 07]．コミュニティは，一般に「ネットワーク内で密に繋がっている部分グラフ」と定義される [Flake 02]．しかし，概念の定義にとどまり，抽出結果の定量的評価は行われてこなかった．これは従来のコミュニティ抽出の対象が主として Web をはじめとした大規模なデータであったため，コミュニティ抽出結果は取り出されてきたコンテンツの内容で評価できるためと考えられる．

一方，意味のあるコミュニティを抽出する手法は情報検索に限らず，都市計画における道路網の整備や，新たな行政区分の実施においても重要である．これらのネットワークにおいて何をコンテンツとするかを特定することが難しく，抽出結果の評価には利用できない．また，既存の抽出方法は，ネットワーク全体の構造が既知であることが前提であるため現実的ではない．

本研究ではコミュニティの概念を定式化し，これに基づいたコミュニティの評価尺度を提案する．評価尺度はコミュニティとそのリンク構造に基づいて算出するため，ネットワーク全体の構造をあらかじめ必要としない．また，評価尺度を基にした新たな探索型のコミュニティ抽出法を提案する．

2. 問題設定と提案手法

本研究は WWW や大都市の交通網のような大規模ネットワークを対象とする．ネットワーク全体の構造を必要としないコミュニティ抽出法が要請される．

連絡先: 荒井幸代，千葉大学大学院工学研究科，〒 263-8522　千葉市稲毛区弥生町 1-33，043-290-3316，sachiyo@faculty.chiba-u.jp

2.1 コミュニティの定式化

「ネットワーク内で密に繋がっている部分グラフ」という概念定義に基づいて，コミュニティを定式化する [?][Flake 02]．本研究では密に繋がっている部分グラフを「部分グラフ内の連結度が部分グラフとその外部との連結度よりも密である部分グラフ」と捉えた，コミュニティの定式化を以下のように示す．

ネットワーク $G = (V, E)$ におけるノード集合 V の部分集合 $V_s \subseteq V$ を考える．ノード集合 V_s によって構成される部分グラフ $G_s = (V_s, E_s)$ が，以下の 2 つの条件を満たすとき G_s をコミュニティと呼ぶ．

条件 1． 部分グラフ G_s は連結である．

条件 2． $k_{G_s} > k'_{G_s}$

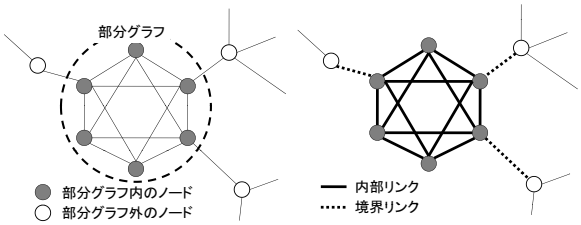
G：ネットワーク
V：ネットワーク内のノード集合
E：ネットワーク内のリンク集合
G_s：ネットワーク G 内の部分グラフ
V_s：部分グラフ G_s 内のノード集合
E_s：部分グラフ G_s 内のリンク集合
k_{G_s}：部分グラフ G_s 内のノードの平均次数 (コミュニティ内リンク密度)
k'_{G_s}：グラフ $G' = (V_s, E \setminus E_s)$ におけるノード $v_s \in V_s$ の平均次数 (コミュニティの境界リンク密度)

条件 1. はコミュニティが連結であることを示している．非連結の部分グラフは明らかに密に連結していないため，コミュニティではない．

条件 2. は部分グラフ内のリンクの連結度と部分グラフと外部とのリンクの連結度の比較を示す．左辺の k_{G_s} は部分グラフ内の連結度を示し，右辺の k'_{G_s} は部分グラフと部分グラフの外側とのリンクの連結度を示している．k_{G_s} の値が k'_{G_s} の値よりも大きい場合に，対象の部分グラフはコミュニティとなる．

図 1(i)(ii) を用いて定式化を説明する．図 1(i) の灰色の丸印が対象とする部分グラフ内のノードであり，白い丸印が部分グラフと接続しているノードである．ここでは部分グラフ内のリンクを内部リンク (図 1(ii) の太線のリンク)，部分グラフと部分グラフ外を連結するリンクを境界リンク (図 1(ii) の点線のリンク) と呼ぶ．

図 1(i) の部分グラフは連結であるため，条件 1. を満たす．また，この部分グラフの k_{G_s} の値は $24/6 = 4.0$，k'_{G_s} の値は $3/6 = 0.5$ となる．これより $k_{G_s} > k'_{G_s}$ になり，図 1(i) の部分グラフは条件 2. も満たす．この部分グラフは条件 1，2 の両方を満たすためコミュニティである．



(i) コミュニティの例　(ii) コミュニティに関係するリンクの名称

図 1: コミュニティに関する用語

2.2 評価尺度の定量化

2.1 節で示した定量化に基づいた評価尺度を示す．この評価尺度を以下ではコミュニティ濃度 Cd と呼び、式 (1) で定義する．

$Cd \geq 1$ を満たす部分グラフ G_s は周囲より密であることを示し，このときその部分グラフ G_s をコミュニティとみなす．

Cd 値が大きいとき，部分グラフ内の連結度が部分グラフとその外部との連結度よりも密であること，つまりコミュニティ内外の相対的な密度差が大きく，かつコミュニティ内部の密度が外部の密度よりも大きいことを意味する．

$$Cd = \frac{k_{G_s}}{k'_{G_s}} = \frac{|E_s| \times 2}{|V_s|} \Big/ \frac{(\text{境界リンク数})}{|V_s|} = \frac{|E_s| \times 2}{(\text{境界リンク数})} \quad (1)$$

2.3 提案手法

コミュニティ濃度 Cd に基づいた探索型コミュニティ抽出法 (Cd-based-Method:以下 CDB 法) を図 2 に示す．まず，コミュニティの探索の始点となるシードノードを選択し，これをコミュニティ内の初期ノードとする．次にコミュニティ内のノードと接続している各ノードをコミュニティに加えた場合のコミュニティ濃度 Cd を計算し，最もコミュニティ濃度 Cd 値が大きいノードをコミュニティ内に加える．終了条件を満たすと探索を終了する．

CDB 法はコミュニティ内のノードの加えられた順番とそのときのコミュニティ濃度 Cd が算出される．図 3 に横軸をコミュニティ内のノード数，縦軸をコミュニティ濃度 Cd としたグラフを示す．コミュニティ濃度 Cd の推移を計測し，図 3 中の丸印に囲まれた箇所のようにコミュニティ濃度が極大値を示したときに，等高線を引きシードノード中心のコミュニティ図を作成することができる．

本手法は重み付きグラフ，重み付きではないグラフの双方に適用可能である．重み付きグラフの場合はリンクの重みをそのまま使用し，重み付きではないグラフではリンクの重みを 1 として使用する．

Cd-based-Method(V, E, N)
1: シードノードの選択
2: $Com[1] \leftarrow$ シードノード
3: $Cd[1] \leftarrow 0$
4: **for** $i \leftarrow 2$ to N **do**
5: $Com[i] \leftarrow$ コミュニティに含まれたときにコミュニティ濃度の値が最も大きくなるノード
6: $Cd[i] \leftarrow$ コミュニティ濃度の最大値
7: **end for**

図 2: コミュニティ濃度に基づいた抽出法のアルゴリズム

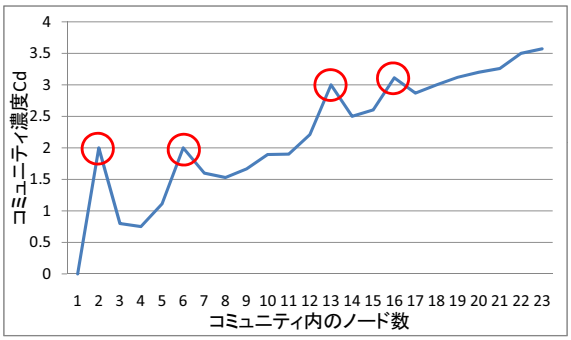


図 3: コミュニティ濃度 Cd の推移例

3. 実験結果

3.1 コミュニティ濃度に関する実験と結果

コミュニティ濃度 Cd の妥当性を既存の抽出法の比較実験を通して示す．

3.1.1 実験設定

提案した評価尺度の妥当性を示すために既存のコミュニティ抽出法を用いて実験を行う．比較対象とする抽出法は，Star and Diamond Clustering[Matsuo 02](以下，SDC と表記)，最大フローを用いた手法 [Flake 02]，betweenness を用いた手法 [Freeman 77] の 3 つである．

SDC はクラスター係数を尺度とし，クラスター係数が大きくなるようにリンクを削除する操作を繰り返し，最終的に残ったネットワークをコミュニティとする手法である．

最大フローを用いた手法は最小カットを見つけて，コミュニティを抽出する手法である．シードノードを選択し，シードノードから最大フローを用いて最小カットを探索して，最小カットによって切り取られた部分グラフをコミュニティとする手法である．

また，betweenness を用いた手法は betweenness の大きいリンクを除去してコミュニティを抽出する手法である．betweenness とは各ノード間の最短経路上のリンクに与える値であり，多くの最短経路上に位置するリンクの betweenness の値が大きくなる．betweenness の値が大きいリンクはコミュニティ間を連結するリンクであるため，除去することによりコミュニティを抽出することができる．

実験には千葉大学工学部都市環境システム学科で開講されている講義間の関連性を抽出したネットワーク [Nishijima 08] を用いる．実験に用いたネットワークは知識共創システムによって得られたネットワークであり，ノード数は 71，リンク数は 197 である．両手法によって抽出されたコミュニティ別に，コミュニティ濃度 Cd，クラスター係数，平均最短経路長を算出し，その数値の比較を通じて各手法を評価する．

3.1.2 結果と考察

3.1.1 節で述べた 3 つの手法で抽出されたコミュニティのうち，本稿では SDC による抽出結果だけを図 4 に紹介する．図中の太線で囲まれた部分グラフがコミュニティである．太線の外側にあるノードはコミュニティ内に含まれなかったノードである．

抽出されたコミュニティの一部とノード数，クラスター係数，平均経路長，コミュニティ濃度 Cd を表 1 に示す．表中の a，b は抽出されたコミュニティを示す．

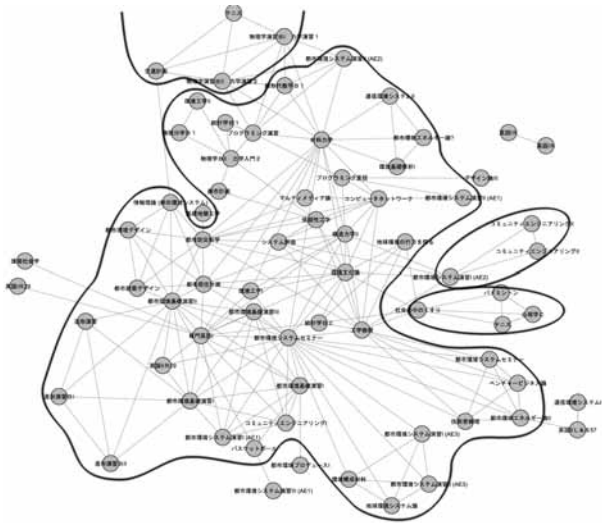


図 4: SDC によるコミュニティ抽出結果図

SDC 法によって抽出されたコミュニティの例を 2 つ (図 5(i)(ii)) 用いて，提案した Cd 値が従来のクラスター係数平均最短経路長よりも，部分グラフの「密なつながり」を定量的に示す評価尺度として妥当であることを以下に示す．表 1 にコミュニティa とコミュニティb におけるクラスター係数，平均最短経路長，コミュニティ濃度の各値を示す．

表 1: SDC の抽出結果

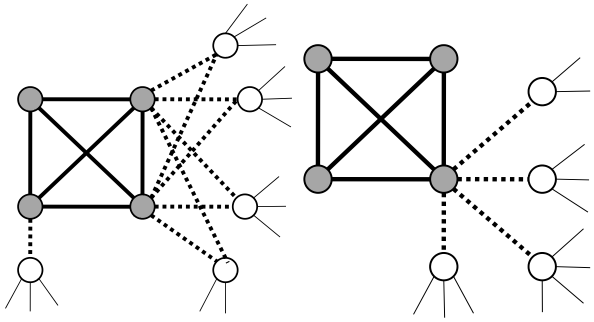
コミュニティ	a	b
ノード数	4	4
クラスター係数	1.00	1.00
平均最短経路長	1.00	1.00
Cd	1.33	3.00

コミュニティa と b の比較

コミュニティa と b はノード数，クラスター係数，平均最短経路長の値は等しいが，コミュニティ濃度 Cd は異なる．Cd だけが異なる理由を図 5(i)(ii) のコミュニティの構造を用いて説明する．

コミュニティa と b はともに K_4 の完全グラフであるため，内部リンク数は 6 本である．しかし，境界リンク数については，a は 9 本，b は 4 本である．内部リンク数が等しくても，境界リンク数が異なるために Cd 値に差が生じている．a と b では，b の方が境界リンク数が多いために Cd 値が小さい．

コミュニティは内部リンク数が等しい場合，境界リンク数の大小によってコミュニティ内部の連結度が密であるのか，外部の連結度が密であるのかが決まる．つまり，コミュニティ内外の相対的な密度差が境界リンク数によって決定される．コミュニティ内外の相対的な密度差が大きく，コミュニティ外部の連結度が密でない Cd 値は大きくなる．a と b では，a の方がコミュニティ外部の密度が小さいため，Cd 値が小さくなる．



(i) コミュニティa　(ii) コミュニティb

図 5: 抽出されたコミュニティの構造

以上より，コミュニティ濃度 Cd はコミュニティの内部と外部の相対的な密度差を反映できることがわかる．

3.2 CDB 法に関する実験と結果

3.2.1 実験設定

CDB 法の妥当性を示すために提案手法を重み付きグラフと重み付きではないグラフの 2 種類に対して実験を行う．実験の対象は千葉大学工学部都市環境システム学科で開講されている講義間の関連性をネットワークとして生成する知識共創システムを用いる [Nishijima 08]．重み付きではないグラフとはネットワーク内のリンクの重みをすべて 1 として扱うグラフのことをさす．重み付きのグラフとはネットワーク内のノードの関連性の重みをリンクの重みとして扱うグラフのことをさす．

3.2.2 結果と考察

CDB 法はシードノードによって抽出されるコミュニティ図は異なる．ここでは同一のシードノードを用いてリンクの重みを考慮した場合としない場合に抽出されるコミュニティ等高線図を観察する．各場合のコミュニティ濃度による等高線図を図 6，図 7 に示す．図中の黒丸のノードがシードノードである．等高線で囲まれた範囲に存在するノードがコミュニティ内のノードとなる．シードノード付近の領域がシードノードに最も関連性のあるコミュニティであり，外側の領域になるほどシードノードとの関連性が薄いコミュニティとなる．

重み付きではないグラフ

重み付きではない場合は，シードノードがクリークに属していても, そのクリークをコミュニティとして抽出するとは限らない．

これはクリーク内のノードはリンク数が多いため，クリーク内のノードをコミュニティ内に取り込むと一時的にではあるがコミュニティ濃度 Cd が下がるためである．提案手法は探索する際に 1 ホップ先の接続関係しか考慮しないため一時的にでも，コミュニティ濃度が減少するノードはコミュニティ内のノードになりにくい．探索の際に 2 ホップ先を見るなどの対処が必要である．

また，図 6 中の丸印で示した端点の 2 ノードは同時にコミュニティに取り込まれ，端点の 2 ノードだけでコミュニティを形成することはない．

重み付きグラフ

シードノードを含むクリーク内のノードをすべてコミュニティとして取り込むのではなく，クリーク外へのリンクの少な

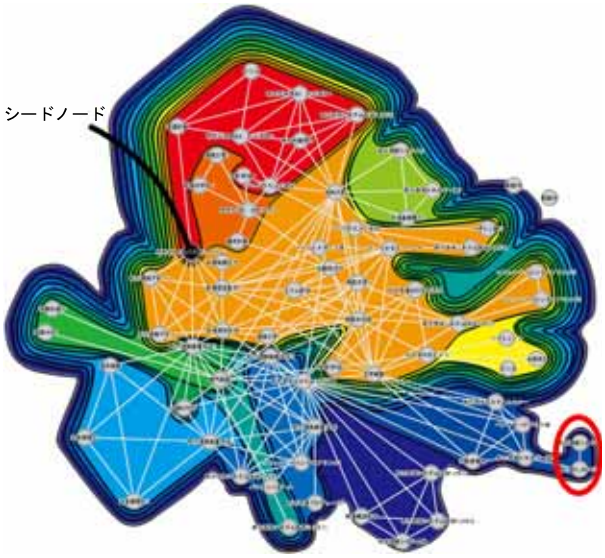


図 6: 重み付きではないグラフでの抽出図

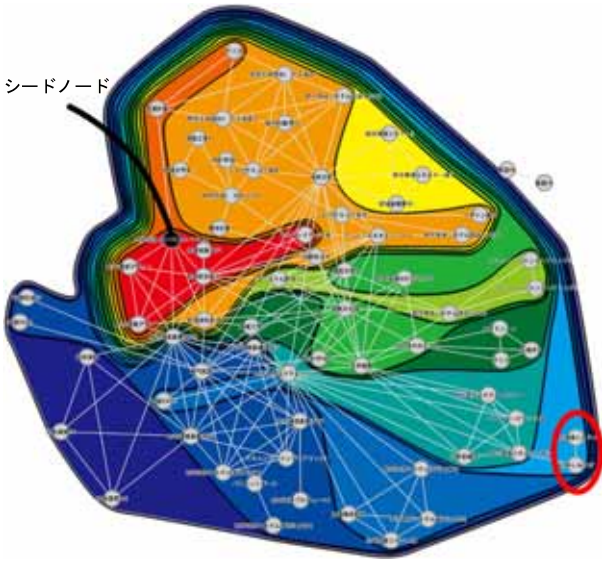


図 7: 重み付きグラフでの抽出図

いノードだけをコミュニティとして抽出することが可能である．

これは互いに関連性の大きなノード同士がクリークになるため，クリーク内のリンクは重みが大きくなる．これは一般的なネットワークの特性でもある．クリーク内のリンクは重みが大きいため，重みを考慮することによって，クリーク内のノードもコミュニティとして抽出することが可能になる．クリーク外へのリンクの多いノードが排除されたのは，クリーク外への重みがクリーク内への重みよりも大きいことによる．人間社会に例えればグループ内にいる人物が、グループ外にも多くの人脈を持っているためにグループから排除されることに相当する．また，重み付きではないグラフとは異なり，図 7 中の丸印のような端点の 2 ノードは，つながりが強ければ 2 ノードだけでコミュニティを形成し，他のコミュニティに容易に組み込まれないことなど興味深い特徴が観測された．

4. おわりに

コミュニティ濃度 Cd はコミュニティの内部の構造と外部の構造との相対的な密度差を明確に表現できる．

従来，つながりが密な部分グラフをコミュニティと呼び，コミュニティ内部の密度だけを評価し，外部の密度は考慮していなかった．しかし，外部との密度の方が内部の密度より大きいならば，相対的にはコミュニティ内部の連結度合は密とはいえない．そこで内外の密度の両方を考慮した相対的に密なコミュニティを抽出法を提案した．

相対的な密度差はコミュニティの定義の自然な拡張であり，既存の評価尺度である平均最短経路長やクラスター係数では評価に反映できなかったコミュニティの連結度を評価できることは，コミュニティ抽出法の評価尺度として有用である．

また，コミュニティ濃度を基礎としてコミュニティを探索する CDB 法の提案によって，相対的な密度差のあるコミュニティの特定を可能にしたことが本研究の貢献である．

参考文献

[Reichardt 06] J. Reichardt,and S. Bornholdt, “Statistical mechanics of community detection,” Physical Review E 74, 016110, 2006.

[Zhang 07] S. Zhang and R. Wang, X. Zhang, “Identification of overlapping community structure in complex networks using fuzzy c-means clustering,” Physica A 374, pp.483-490, 2007.

[Tanigawa 07] 谷川恭平, 土方嘉徳, 西田正吾, “コミュニティ発見のためのフレームワークの提案,” 第 10 回電子情報通信学会 Web インテリジェンスとインタラクション研究会 (電子情報通信学会 第二種研究会資料 IEICE SIG Notes WI2-2007-41), pp.13-18, 2007.

[Flake 02] G. W. Flake and S. Lawrence, C. L. Giles, F. Coetzee,“Self-Organization and Identification of Web Communities,” IEEE Computer, 35(3), pp.66-71, 2002.

[Matsuo 02] Yutaka Matsuo, “Clustering using Small World Structure,” Proc. 6th Int’l Conf. on Knowledge-based Intelligent Information Engineering Systems & Applied Technologies(KES2002), IOS Press/Ohmsha (ISSN:0922-6389), Crema, Italy, pp.1252-1256, 2002.

[Freeman 77] Linton C. Freeman, “A Set of Measures of Centrality Based on Betweenness,” Sociometry 40, pp.35-41, 1977.

[Nishijima 08] 西嶋寛, 荒井幸代, 檜垣泰彦, 土屋俊, “フォークソノミを用いた講義選択知識の抽出,” 電子情報通信学会技術研究報告, OIS2008-14, Vol.108, No.53, pp.79-84, 2008.