

ノード属性を考慮した効率よいグラフのまとめ発見手法

寺田 愛花[†] 瀬々 潤[†]

[†] お茶の水女子大学理学部情報科学科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: terada@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp

あらまし 近年、様々な分野において膨大で複雑なグラフが解析されている。グラフのまとめを求めるために頻繁に用いられるのは、辺が高密度に張られた頂点集合を発見する手法である。これとは異なり、我々はグラフの頂点にラベルを付与し、辺をまとめることで、グラフを表すラベル間の関係を抽出し、要約を求める。このような要約を求めるために、まず接続関係が類似しているラベルを有する頂点でグループを構成し、次にグループ間の辺の接続関係による分割を行う手法を提案する。本手法で、辺の本数が44,209本、ラベルが103種類の遺伝子ネットワークを解析し、遺伝子の機能間の関係を抽出した。この結果は、一般的なグラフクラスタリング手法では求めることができないものであった。

キーワード グラフ, クラスタリング

Efficient Finding of Label Relationship Graphs from a Large Graph

Aika TERADA[†] and Jun SESE[†]

[†] Dept. of Computer Science, Ochanomizu Univ. 2-1-1 Otsuka, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: terada@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp

Abstract Analysis of a large and complicated graph has been attracted area because of wide range of applications. In contrast with most of the analysis finding dense parts of a graph, we here extract a graph depicting relations between labels on nodes connected by edges. Finding such the summarization, we introduce a novel method, which first aggregates the labels whose connecting labels are close to each other, then divides the groups by using edge connectivity. Our experimental result using a gene network consisting of 44,209 edges and 103 different labels shows that our result can find functional connection between genes, which cannot be found by a standard graph clustering method.

Key words graph, clustering

1. はじめに

実世界の様々な関係は、頂点が物を表し、関係がある物の間に辺を張るグラフで表現することができる。例えば、頂点を遺伝子、辺を相互作用とした遺伝子ネットワーク、Web ページを頂点、ハイパーリンクを辺とした Web ネットワーク、ユーザを頂点、友人関係を辺としたソーシャルネットワークなどである。近年、様々な分野でグラフが生まれているが、このようなグラフは、頂点数、辺の本数が膨大であり、複雑なため、単に頂点と辺を与えたグラフから情報を読み取ることは困難である。それゆえ、グラフのまとめを発見する技術が必要とされている。

本手法では、グループ間の関係を抽出することで、グラフ全体のまとめを求める。例えば、頂点を人とし、大学内の友人関係を表した図 1(a) から、図 1(b) という要約を求める。このグラフの頂点に書かれているのは、その人が所属するサークルで

ある。図 1(a) のグラフでは、バレーサークルに所属する人が二人ともテニスサークルに所属する人に辺を張り、テニスサークルに所属する人も二人ともバレーサークルに所属する人に辺を張ることから、図 1(b) の A で表されるバレーとテニスサークル間の強い関係がわかる。また、野球サークルに所属する二人と、サッカーサークルに所属する二人全員が、テニスサークルと吹奏楽サークルに友人がいることから、B の関係が分かる。

一般的なまとめの発見手法としてクラスタリングがある。グラフクラスタリングは、辺が高密度に張られている頂点集合を発見する手法か、疎な部分を分断する方法である [1]。しかしこの手法では、局所的に関係の強い頂点集合を求めることはできるが、その集合間の関係は抽出できず、グラフ全体の構造を見ることは難しい。例えば、図 1(a) をこの手法で分類すると、図 2 のクラスタが構成される。この結果から、各クラスタ内の人たちが仲がよいということは分かる。しかし、図 1(b) の A や

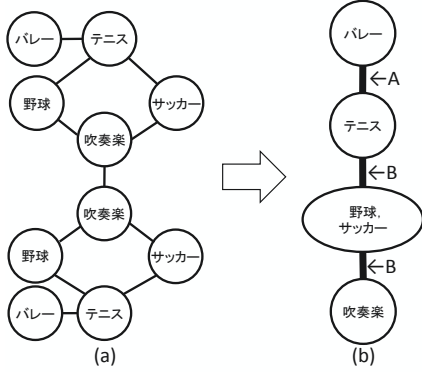


図 1 グラフの要約

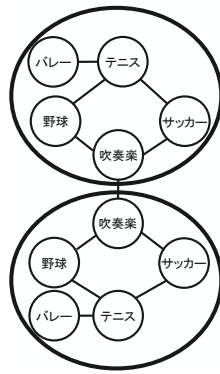


図 2 metis の分割結果

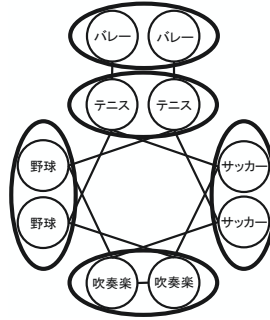


図 3 k -SNAP の結果

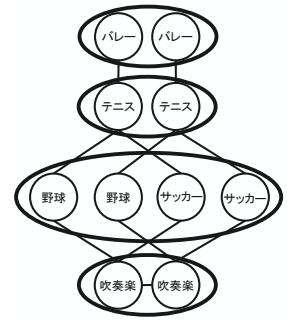


図 4 提案手法の結果

B のような、グループ間の関係を見ることはできない。

このようなグラフクラスタリングとは異なるアプローチで、接続する頂点が類似した辺をまとめることで、グラフ全体の要約を求める手法がある。

このようにグラフ全体の要約を求める手法として、 k -SNAP [2] がある。この手法は、頂点にラベルを付与したグラフの解析を行う。ラベルとは、頂点で表す物が持つ情報である。例えば、図 1(a) では、頂点で表している人が所属するサークルを、頂点のラベルで与えている。一般的なグラフクラスタリング [3] では、このラベルを有する頂点は扱えない。 k -SNAP はこのグラフを、頂点のラベルと、接続するグループが同一となるように分類する手法である。図 1(a) をこの手法で解析すると、図 3 というグループが構成される。この要約を見ることで、野球サークルはテニスサークルと吹奏楽サークルと関係が強いということや、サッカーサークルはテニスサークルと吹奏楽サークルと友達が多い、というようなサークル間の関係を見ることができる。

しかし k -SNAP は、同一グループ内の頂点が同一ラベルであるという制約があり、データ内のラベル数が k 以下でないといけないため、大学のサークルや遺伝子の機能など、ラベルの種類が多い場合、解析できないという問題点がある。例えば図 1 のグラフは、ラベルの種類が { バレー, テニス, 野球, サッカー, 吹奏楽 } の 5 種類であるため、このグラフを 4 つのグループに分類することはできない。

本手法では、 k -SNAP のラベルの制約を緩和し、 k 種類以上のラベルを有するグラフでも解析できるように拡張する。そのために、グラフの隣接構造を用いた頂点の併合と、ラベルの情報を用いた頂点の分割を行う。その結果、図 1(a) のグラフを図 4 のように分類する。 k -SNAP との違いは、野球サークルとサッカーサークルのような、異なるラベルを有する頂点が同一グループに属することである。この二つのサークルは、どちらもテニスと吹奏楽という同一のグループに対して辺を持っており、接続する頂点が類似している。本手法では、このように類似した辺を持つグループを同一グループとして分類することで、ラベル数が多いグラフを解析可能にするだけでなく、野球とサッカーのように、関係が類似したラベルを求めることができる。

本研究の提案手法を、頂点に遺伝子、辺に遺伝子ネットワーク、ラベルに遺伝子の機能を与えたネットワークを解析し、細胞内のネットワークの構造を抽出する。その結果を、クラスタ間の辺が疎になる分割をするグラフクラスタリング手法 metis [3] と、グラフの隣接構造を用いた頂点の併合のみ行った結果と比較する。

2. 提案手法

k -SNAP の手法は、ラベルの種類でグループを構成した後、グループ間の関係を抽出する分割を行う。しかしこの手法では、ラベルの種類が多いと分割できないという問題点がある。

この問題を改良するため、本手法ではラベルの種類でグループを構成した後、グループへ伸びる辺が類似するグループを併合することで、 k -SNAP のラベルの制約を緩和する。

本手法は頂点の併合と分割の 2 段階で構成される。

- (1) ラベルの種類が等しい頂点でグループを構成
- (2) $k/2$ 個まで類似するグループを併合
- (3) グループの個数が k 個になるまでグループ間の強い関係を抽出する分割をする

グラフ $G = (V, E)$ とする。 V は頂点集合、 E は辺の集合である。頂点にはラベルを与え、このラベル集合を $A = \{a_1, a_2, \dots, a_t\}$ と表す。頂点 v がラベル a_i を有する場合 $a_i(v) = 1$, それ以外の場合 0 とする。またグラフ G の分割を $\Phi = \{G_1, G_2, \dots, G_k\}$ と表記する。頂点 v は必ず一つのグループ G_i に属し、 $G_i \neq \phi$ ($\forall G_i \in \Phi$) かつ $G_i \cap G_j = \phi (i \neq j)$ である。頂点 v の属するグループを $\Phi(v)$ とする。

2.1 ラベルの種類でグループを構成

まず頂点をラベルの種類で分類する。そのために $u, v \in V$ に対し、 $\forall a_i \in A$ で $a_i(u) = a_i(v)$ ならば $\Phi(u) = \Phi(v)$, それ以外の場合、 $\Phi(u) \neq \Phi(v)$ となるようにグループを構成する。 t 個のクラスタが形成される。この分類により、ラベルが等しい頂点が全て同一グループに属するようにグループを構成する。

2.2 グループの併合

グループを $k/2$ 個まで併合する。併合するのは、グループへ伸びる辺が類似するグループである。この類似度で、グループの距離 $D(G_i, G_j)$ を定義し、この値が最小のグループを併合する。

グループ間の辺の伸び方の類似度はコサイン距離で計測する。しかし、グループへの辺の本数で類似度を定義すると、グループに属する頂点数の違いから距離に差が生じる。そのため、 $|E(G_i, G_j)|$ を G_i と G_j の間にある辺の本数とし、辺の密度 r_{ij} を定義する。

$$r_{ij} = \begin{cases} \frac{2|E(G_i, G_j)|}{|G_i|(|G_i|-1)} & (i = j) \\ \frac{|E(G_i, G_j)|}{|G_i||G_j|} & (i \neq j) \end{cases} \quad (1)$$

G_i から G_j へ伸びる辺の密度 r_{ij} を各要素としたベクトルを $\mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{in})$ とする。辺の伸び方が類似しているグループは、このベクトルの向きが類似している。そのため、本手法ではコサイン距離を用いてグループ間の類似度を計測する。グループ G_i と G_j の距離を次式で定義する。

$$D(G_i, G_j) = 1 - \frac{\mathbf{r}_i \cdot \mathbf{r}_j}{|\mathbf{r}_i||\mathbf{r}_j|} \quad (2)$$

この距離を全グループ間で算出し、値が最小のグループ G_i と G_j を併合する。この操作を繰り返し、グループ数が $k/2$ 個になるまで併合する。

2.3 グループ間の関係を抽出する分割

グループを併合した後、グループ間の関係を保ちつつ分割する。この手順は k -SNAP 法と同等である。この手法は、分割するグループ G_s とそれとは異なるグループ G_t を指標を定義して選択し、 G_s を G_t を利用して二分割にする。

G_i に属する頂点で、 G_j と辺を持つ頂点集合を $P_{G_j}(G_i) = \{u | u \in G_i, \exists v \in G_j, (u, v) \in E\}$ と定義する。グループ G_i が G_j に対して関係が強いとは、グループ内にこのような頂点が多いことである。そこで、グループ G_i と G_j の強さを次式で定義する。

$$p_{i,j} = \frac{|P_{G_j}(G_i)| + |P_{G_i}(G_j)|}{|G_i| + |G_j|} \quad (3)$$

我々が抽出したいグループは $p_{i,j}$ が 0 または 1 に近いクラスターである。そのため、グループ G_i の G_j に向けた乖離度を次式で定義する。

$$\delta_{G_j}(G_i) = \begin{cases} |G_i| - |P_{G_j}(G_i)| & (p_{i,j} > 0.5) \\ |P_{G_j}(G_i)| & (\text{上記以外}) \end{cases} \quad (4)$$

乖離度は値が小さい方が良い分割である。そのため、この値が最大のグループを分割する。

グループ G_i の指標を $CT(G_i) = \max_{G_j} \{\delta_{G_j}(G_i)\}$ とする。この値が最大のグループを G_s とし、 $G_t = \arg \max_{G_j} \{\delta_{G_j}(G_s)\}$ であるグループ G_t に対して辺を持つ頂点集合と持たない頂点集合に分割することで、グループ間の強い関係を抽出していく。これを、グループ数が k 個になるまで繰り返す。

3. 遺伝子ネットワークの解析

遺伝子ネットワークを解析し、機能間の関係を抽出する。解析するグラフは、頂点に遺伝子、辺に遺伝子ネットワーク、ラベルに遺伝子の機能を与える。その頂点数は 1,131 個、辺の本数は 44,209 本、ラベルは 103 個である。

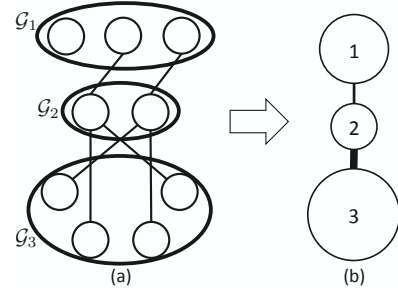


図 5 結果のグラフの表し方

このネットワークを k -SNAP で解析すると、ラベルの制約から 103 個のクラスターが構成されるため、46 個に分類することはできない。

そこで本研究の提案手法と比較するため、クラスター間の辺が疎になる分割をするグラフクラスタリング手法 metis [3] と、ラベルの種類でグループを構成した後、46 個まで併合のみ行う手法による解析も行った。これらの結果を、ラベルのばらつき、グループ間の乖離度、グループ間の辺の本数という三つの観点から比較する。

3.1 結果の表し方

本研究では、解析結果をグラフで表す。結果のグラフの頂点 i はグループ G_i であり、多くの頂点が属するグループほど与える頂点は大きい。また、関係が強いグループを抽出するために、 $p_{i,j} > 0.5$ のグループ間に辺を張る。 $p_{i,j}$ の値が大きいほど太い辺で表している。例えば、図 5(a) の分類結果は図 5(b) のグラフで表される。図 5(a) の G_2 は頂点数が 2 個、 G_3 はその倍の 4 個の頂点が属している。よって図 5(b) の G_3 の頂点は、 G_2 の倍の大きさと与えられる。また、グループ G_2 と G_3 の強さは $p_{2,3} = 1.0$ であり最大値なので、図 5(b) の辺 (2,3) は太い辺が与えられている。これに対して、 G_1 と G_2 は $p_{1,2} = 0.8 > 0.5$ のため、図 5(b) のグラフで辺 (1,2) は張られるが、 $p_{1,2} < p_{2,3}$ であるため、辺 (2,3) よりも細い辺が与えられる。

本手法の解析結果が図 6、metis の結果が図 7、ラベルの種類でグループを構成した後、併合のみ行った結果が図 8 である。

3.2 機能 (ラベル) のばらつきの比較

特定の機能を持つ遺伝子が、一つのグループに集中しているかを調べる。機能を持つ遺伝子が少ないものに注目しても、その分布を考察することはできない。そのため、多くの遺伝子が持つ機能 mRNA_processing-Reactome と、機能 Ribosomal-proteins に着目し、各手法のばらつきを比較する。

まず、機能 mRNA_processing-Reactome に着目する。この機能を持つ遺伝子が主に属するグループは、各結果の図のグループ A である。図 7 の metis の結果では、この機能を持つ遺伝子が複数のグループに分散し、他の機能との関係を抽出できない。図 6 の提案手法と図 8 の併合のみ行った結果を比較すると、どちらの結果も機能 mRNA_processing-Reactome を持つ遺伝子がほぼ一つのグループに属しているが、併合のみ行った結果では、グループが孤立点であり、他の機能を持つ遺伝子グループとの関係を見ることができない。本手法の結果では、

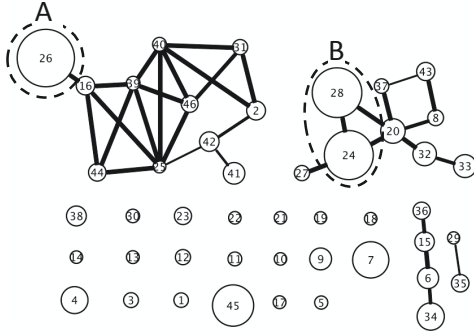


図 6 提案手法の分割結果

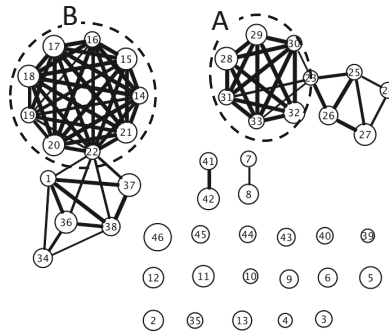


図 7 metis の結果

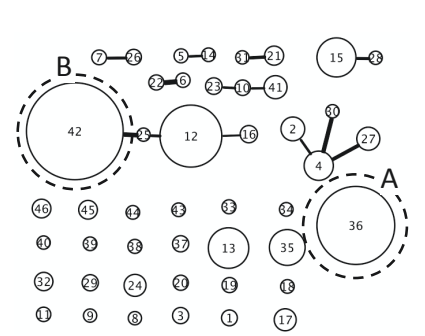


図 8 併合のみ行った結果

A が G_{16} へ辺を持ち、機能 mRNA_processing_Reactome が、 G_{16} に属する遺伝子が持つ機能と関連があることが分かる。実際にその機能を調べると、 G_{16} の機能は DNA から転写する機能であり、 G_{26} の機能 mRNA_processing_Reactome の前段階であった。

別の機能 Ribosomal_Proteins に着目する。その結果は各図のグループ B である。この機能でも、図 7 の metis の結果は遺伝子が複数のグループに分散していることが分かる。図 6 の提案手法と図 8 の併合のみ行った結果の比較をすると、提案手法の結果では、同じ Ribosomal_Proteins を持つ遺伝子でも、 G_{27} が持つタンパク質への翻訳機能と直接的に関わっている G_{24} と、間接的に関わっている G_{28} があることが分かる。一方で、図 8 の併合のみ行った結果では、このような違いを観測することができない。

このことから、遺伝子の機能間の関係を最も良く表しているのは、図 6 の本手法である。

3.3 グループ間の乖離度の比較

三種類の手法の結果を乖離度で比較するために、各手法で次式を計測した。

$$\frac{\sum_{G_i, G_j \in \Phi} (\delta_{G_j}(G_i) + \delta_{G_i}(G_j))}{k} \quad (5)$$

これは、全グループ間の乖離度 $\delta_{G_j}(G_i)$ の平均であり、この値は小さい方が良い分割であることを示している。

各手法のこの値は、提案手法が 15.84, metis が 24.76, 併合のみ行った結果が 21.06 である。最小値は本手法であることから、これが最もグループ間の密な関係が抽出できることがわかる。

3.4 辺の本数の比較

グループ間の辺の本数で、三種類の手法を比較する。解析したグラフ G の分割を $\Phi = \{G_1, G_2, \dots, G_{46}\}$ 、また、結果のグラフ G' の分割を $\Phi = \{G'_1, G'_2, \dots, G'_{46}\}$ とし、比較のために次式を算出した。

$$\frac{\sum_{G'_i, G'_j \in \Phi} E(G'_i, G'_j)}{\sum_{G_i, G_j \in \Phi} E(G_i, G_j)} \quad (6)$$

この値は結果のグラフ G' の辺が表す、解析したグラフ G の辺の本数の平均である。例えば、図 5(a) では、 $|E(G_1, G_2)| = 2$, $|E(G_2, G_3)| = 4$ であることから、図 5(b) が表す平均の辺の本

数は 3 である。グループ間の関係が強いほど、平均の辺の本数は多い。

各手法のこの値は、本手法が 586.93, metis が 436.86, 併合のみ行った結果が 90.61 である。本手法が最も値が大きく、グループ間の強い関係を表していることが分かる。

3.5 結果の考察

遺伝子ネットワークで、本研究の提案手法と、辺が疎である分割をするグラフクラスタリング手法、ラベルによる分類後、併合のみ行う三種類の手法を比較した。

機能のばらつき、グループ間の乖離度、グループ間の辺の本数の観点から比較した結果、どの比較においても本手法が最良の結果であった。

4. まとめと今後の課題

接続する頂点が類似した辺をまとめることで、ラベル付きグラフ全体の要約を求めることを目的とした手法を提案した。このような要約を発見する k -SNAP には、グループ内の頂点が同一のラベルであるという制約があり、ラベルの種類が多いと解析できないという問題点があった。そのため、グループを併合することでこの制約を緩和し、ラベルの種類が多い場合でも実行できるよう拡張した。また、本手法と、グラフクラスタリング手法 metis、個数が k 個になるまでグループを併合する手法で遺伝子ネットワークを解析し、比較した結果、本手法が最も遺伝子の機能間の関係を抽出できており、グループ間の関係も強かった。

今後の課題は、解析したデータが遺伝子ネットワークのみであるため、Web、ソーシャルネットワークなど他のデータも解析することである。また、グループの併合の終了条件を $k/2$ 個としたが、この終了条件を分割の指標を使って求めるように改良したい。そのためには、分割の良さを表す指標を定義する必要がある。また、この指標を用いて併合と分割を繰り返し行い、より良いクラスタリングの結果を求めるように拡張する。

文 献

- [1] 林幸雄. ネットワーク科学の工具箱. 近代科学社, 2007.
- [2] Yuan Yuan Tian, Richard A. Hankins, and Jignesh M. Patel. Efficient aggregation for graph summarization. In *SIGMOD*, 2008.
- [3] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J.SCI.COMPUT.*, 1998.