Vol.2014-CLE-13 No.2
2014/5/16

個人に注目した社会ネットワークの分析に関する研究

菅田 貞治[1,a)]　伊東 樹希[1]　武藤 敦子[1,b)]　犬塚 信博[1,c)]

概要：本研究では出欠データから生成される友人ネットワークに注目し、時系列を伴って変化する局所的友人関係を分析した。まず、個人とその友人のみから構成される局所的ネットワークであるエゴセントリックネットワークの構造的特徴に基づいて個人を分類し、代表的なネットワークモデルと比較したところ友人ネットワークは特有の分類結果である事が示された。次にクラスタ間の時間的遷移を調べる事で学生は友人同士の結びつきが強固になる傾向が強い事が確認できた。また、友人生成に大きく関わった人物をネットワーク構造より推定する事で、その後の分類されるクラスタが予想できる可能性を示した。

キーワード：社会ネットワーク分析，エゴセントリックネットワーク，友人関係

1. はじめに

近年、社会ネットワーク分析の研究が盛んに行われており、その中でもグラフ部分構造を詳細に分析するエゴセントリックネットワーク分析がある。社会ネットワーク分析において、ネットワーク内の少数のまとまりに着目し、ネットワーク形成の原理を見出すネットワーク・センサスというモチーフを分析する手法が用いられてきた [1]。実際に現実社会のネットワークにおいて観測されるエゴセントリックネットワークは特徴的な構造を持つパターンが観測されることが明らかになっている [2], [3]。本研究では友人関係のネットワークに着目して各個人のエゴセントリックネットワークを抽出し、その特徴的な構造に基づいて個人を分類した。

友人関係についてのネットワークを把握する事は通常、アンケートなどで行うため手間を要するが近年はオンラインデータとして手軽に得られるようになり、また出欠データを用いることで友人を推定する手法が提案された [4]。現在は、この手法を利用した様々な研究が行われている [5], [6]。

また友人ネットワークは時間とともに絶えず変化している。そこで時系列に伴って変化する友人ネットワークの局所的構造の遷移について分析する。さらに友人が新しく出来る時、その要因となった人物を友人ネットワークを用いて推定する方法と、その評価値を定めた。この考え方を実際の友人ネットワークに用いた時の評価値の分布や分類との関係性を調べた。

このような分析を行う事で友人ネットワークから個人の役割・性格を把握しやすくなることが期待できる。これは多数の人間の管理を必要とする教育や人事などの関係者にとって役立つと考えられる。

本稿の構成は以下の通りである。第 2 章は本研究が対象としているエゴセントリックネットワークと友人推定手法について簡単に述べ、第 3 章から第 5 章では本稿の研究内容について説明する。第 3 章は個人の友人関係の分類、第 4 章は友人関係の時間推移に関する分析、第 5 章では友人関係の発生の要因の推定についてである。第 6 章では、まとめと今後の課題を述べる。

2. 友人関係とエゴセントリックネットワーク

2.1 エゴセントリックネットワーク

各行為者を頂点とし、行為者間の影響関係を辺で表した無向グラフを考える。社会ネットワーク分析では行為者の 1 人 1 人に注目するときに各行為者をエゴと呼ぶ。各エゴを中心としたローカルなネットワーク、つまりエゴと直接つながる行為者（オルター）の集合から誘導される部分グラフをエゴセントリックネットワーク（以下エゴネット）という。

例えばネットワーク全体 (図 1) からエゴを A としたエゴネット (図 2) やエゴを F としたエゴネット (図 3) が頂点ごとに抽出される。図 2 でのオルターは {B,C,H,I } 、図 3 でのオルターは {D,E,I,M,N} である。

[1] 名古屋工業大学
Nagoya Institute of Technology, Gokiso-cho, Showa 466–8555, Japan
[a)] sugata@nous.nitech.ac.jp
[b)] atsuko@nitech.ac.jp
[c)] inuzuka@nitech.ac.jp

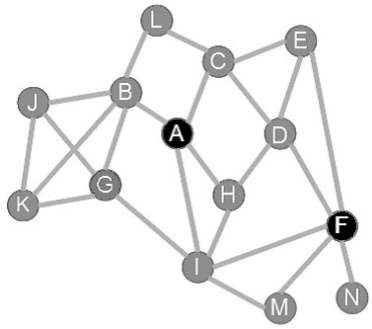

図 1 全体のグラフ

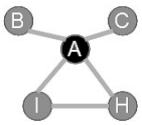

図 2 A のエゴネット

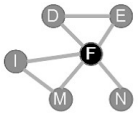

図 3 F のエゴネット

2.2 友人推定手法

本稿で扱う友人ネットワークは下村らが提案した友人スコア [4] を用いて生成した。友人スコアを用いる事により、客観的かつ十分な信頼度のある友人ネットワークが得られる。名古屋工業大学では平成 19 年度に出欠管理システムを導入した。本学の学生は各々の学生証に IC チップが内臓されており、出退席時に IC カードリーダから各学生の学生 ID と打刻時間を記録してサーバに蓄積・管理する。打刻データベースに蓄積された打刻データは打刻した学生を表す学生 ID、打刻した場所を表す端末 ID、打刻した時刻を表す打刻時刻の 3 つの属性を持つ。

友人スコアとは 2 人の学生ペアが友人である確率を表した物である。打刻データベースに蓄積された 2 人の打刻時間の差を算出し、その値によってその 2 人が友人関係であるのかどうかを推測する値である。この値が大きいほど共に行動している、つまり友人関係が強いという事である。詳しい友人スコアの算出方法は付録に記載した。この友人スコアの値が正である学生のペア間に辺を張る事によって友人ネットワークを生成する。

3. 構造に基づくエゴネットの分類

まずエゴセントリックネットワークの構造に基づいてエゴを分類する方法について提案し、前節の方法によって生成された友人ネットワークおよび各種社会ネットワークモデルに適用する。

3.1 トップダウンモデルでの分類

ネットワークの定量的指標を属性値とし、各エゴネットを決定木 (図 4) を用いてトップダウンに 8 つのクラスタに分類する。属性値の内、島はエゴネットからエゴを取り除いた際の連結成分、平均クラスタ係数は同値で分類されるエゴネットの平均値、グループはエゴを取り除いた際の連結成分の内、孤立ノードでないものを指す。ここでクラスター係数はエゴネットに現れた辺の数をオルターから 2 つ選ぶ組み合わせ数で割った値となる。

ここで図 4 について説明する。まずエゴを除いたエゴネットの次数が 0 ならば「孤立」、1 ～ 3 以下なら「友人少」というクラスタに分類する。どちらにも属さない場合、クラスター係数が 0 ならば「スター」、1 なら「完全グラフ」というクラスタに分類する。そして島 1 つで、かつクラスター係数が平均以下ならば「低クラスタ」、平均以上なら「高クラスタ」というクラスタに分類する。最後に頂点数 2 以上の島が 1 つのみで残りの頂点は全て次数 0 ならば「ゲートウェイ」、頂点数 2 以上の島が 2 つ以上なら「ハブ」というクラスタに分類する。なおゲートウェイとハブは本来、コミュニティをつなげる行為者を指す単語 [7] であるが、本稿ではエゴネットに合わせて定義した。

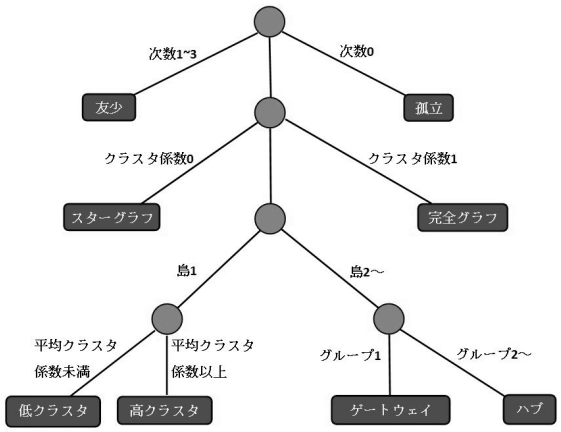

図 4 トップダウンモデルで用いる決定木

3.2 分類結果の分析

分類に用いたネットワークは 2012 年 7 月の名古屋工業大学 1 年生の 友人ネットワークと次の 3 つのネットワークモデルである。

- WS（ワッツ＝ストロガッツ）モデル
  (1) 頂点を円になるように並べて各頂点の近隣 2n 個（右 n 個、左 n 個）の頂点を辺で繋いだグラフ（レギュラーグラフ）を作る。
  (2) それらの辺を q の割合だけランダムに張り替える。
- BA（バラバシ＝アルバート）モデル
  (1) m+1 個の頂点から成る完全グラフを作る。
  (2) 新たに 1 個の頂点を追加し、その頂点から既に存在していた頂点の内、各頂点の次数に比例した確率で m 個だけ選んで辺を張る。
  (3) 所定の頂点数 n になるまで（2）を繰り返す。
- CNN（Connecting Nearest Neighbor）モデル

(1) 確率 p で新たに 1 個の頂点を追加し、その頂点から既に存在していた頂点からランダムに選んだ 1 個に辺を張る。また確率 1-p で、その時点でのネットワークのポテンシャルエッジの中からランダムに選んだ 1 個を実際の辺にする。なおポテンシャルエッジとは直接、辺で結ばれてはないが共通して辺が張られている頂点が 1 個以上、存在する 2 頂点の事である。

(2) 所定の頂点数 n になるまで（1）を繰り返す。

なお各ネットワークモデルは友人ネットワークと頂点・辺の数が同じになるように設定・調節した。具体的には WS モデルでは n=3、q=0.2、BA モデルでは m=3、CNN モデルでは p=0.3 と設定した。各モデルについては 100 回、ネットワークを生成・分類を行った時の平均とした。

分類結果を表 1 に示す。WS モデルは次数 6 のレギュラーグラフだったため「友少」に属するエゴは少なく、ランダムに辺が張り替えられた所は頂点数 1 の島になる事が多いため「ゲートウェイ」に属するエゴが非常に多くなった。BA モデルは次数が大きくなるほど、さらに次数が増える確率が高くなるため、ほとんどの頂点は最初の次数 3 から増えずに「友少」のクラスタに分類されたと考えられる。また BA モデルはクラスター性が低いため島が 1 つになる事が少なく、「低クラスタ」や「高クラスタ」に属するエゴが少なかった。CNN モデルはポテンシャルエッジに辺を張るため「スターグラフ」の数が非常に少なくなったが、次数が大きい頂点の周辺ほどポテンシャルエッジが多くなるため次数が小さい頂点の周辺のポテンシャルエッジが選ばれる確率が低くなり、「友少」に属するエゴが多くなったと考えられる。また各モデルは共通して「孤立」や「完全グラフ」に属するエゴが現れにくい事が解る。

一方、2012 年 7 月の友人ネットワークは「友少」や「ゲートウェイ」に属する学生が多かったが、全体的な偏りは各モデルと比べると小さく、「スターグラフ」を除いて、どのクラスタにも一定数の学生が分類されていた。以上より現実の友人ネットワークは各モデルとは大きく異なる分類結果になる事が分かった。

表 1 各ネットワークにおける分類の比率（%）

	WS	BA	CNN	友人ネット
孤立	0	0	1	6
友少	10	65	60	23
スターグラフ	6	6	1	1
完全グラフ	0	0	0	7
低クラスタ	8	1	10	10
高クラスタ	7	1	8	13
ゲートウェイ	60	23	18	33
ハブ	9	4	2	7

次にクラス間を繋いでいる辺を除去した時の友人ネットワークを再度クラスタリングし、除去する前の結果と比較した。表 2 に結果を示す。なお表の値は書かれている行のクラスタが除去前で列のクラスタが除去後であった学生の人数を表している。

表 2 クラス間の辺を除去した後のクラスタリングの変化（個数）

	A	B	C	D	E	F	G	H
A（孤立）	9	0	0	0	0	0	0	0
B（友少）	1	35	0	0	0	0	0	0
C（スター）	0	1	1	0	0	0	0	0
D（完全）	0	0	0	17	0	0	0	0
E（低クラ）	0	0	0	0	20	3	0	0
F（高クラ）	0	0	0	0	1	27	0	0
G（ゲート）	0	5	0	5	3	3	30	0
H（ハブ）	0	0	0	0	0	0	1	10

大部分は同じクラスタだったが「ゲートウェイ」だけ他のクラスタに遷移する学生が多く見られた。これはクラスが違う学生同士は基本的に会える機会が少ないが「ゲートウェイ」に属する学生がクラス間を繋いでいる事を示している。

4. エゴネットの時系列遷移の分析

前節では、ある時点でのエゴネットの分類を行った。本節では社会ネットワークの局所的な推移傾向を探るためにクラスタリングしたエゴネットの遷移図を得る手法を検討する。

4.1 提案手法

同じ頂点集合 (人の集合)N をもつ時系列 $t_1, t_2, ...$ に対応したグラフ系列 $G_1, G_2, ...$ を考える。これらに観測される互いに非同型のエゴネットすべての集合を EN とする。このとき EN の与えられた分割 (クラスタリング)$EN = EN_1 \cup EN_2 \cup ... \cup EN_m (EN_i \cap EN_j = \phi)$ に対し、対応するエゴネットの遷移図を考える。すなわち、遷移図は分割の各クラスタ $EN_1, ..., EN_m$ を頂点とする有向グラフであり、ある時刻 t での G_t に属する人のエゴネットがクラスタ EN_i に属し、G_{t+1} でその人のエゴネットが GN_j に属するとき、EN_i から EN_j へ遷移したと呼び、有向辺を持つ。有向辺は遷移の全体に対する割合に比例した確率をラベルとして持つこととする。この遷移図が本節の対象である。次にエゴネットの遷移図を得る 3 通りの手法および遷移図を定量的に評価する 2 つの指標について説明する。

4.1.1 遷移図のクラスタリングモデル

4.1.1.1 トップダウンモデル

3.1 節で説明した通り。

4.1.1.2 閉路ベクトルモデル

閉路ベクトルモデルはエゴ e のエゴネットを $(v_1, v_2, ...)$

に符号化する。v_i はエゴを始点かつ終点とする長さ i+1 の閉路の数である。これらのベクトル間のユークリッド距離を対応したエゴ間の類似度とし、EN に階層クラスタリングを適用する。構造的特徴に基づいた距離尺度は他にもあるが、計算量の都合上、本研究では閉路による距離を用いる。

例えば 2.1 節の図 2 と図 3 のエゴネット間の類似度は以下のように求める。A のエゴネットに含まれる閉路は長さ 2 が 4 本、長さ 3 が 1 本であり、F のエゴネットでは長さ 2 の閉路が 5 本、長さ 3 の閉路が 2 本である。ベクトルの最大長を 4 とした場合、A のエゴネットは (4,1,0,0) 、F のエゴネットは (5,2,0,0) となるため、これらのエゴネット間の距離は

$$\begin{aligned} dist &= \sqrt{(5-4)^2+(2-1)^2+(0-0)^2+(0-0)^2} \\ &= \sqrt{2} \end{aligned}$$

となる。

4.1.1.3 遷移優先モデル、クラスタ優先モデル

優先モデルは後述の状態遷移図の定量的評価指標であるクラスタ指標および遷移指標を最大値とする手法である。これらの指標は状態遷移図に対して、[0 1] の範囲で定まる指標であり、各指標値が最大となるようにクラスタリングを行う。

以下が優先モデルの手順である。

(1) 状態遷移図 TR の頂点集合 $EN_1, EN_2, ...$ の内、各指標値が最大となる 2 頂点対 EN_i, EN_j をグリーディに探索する。
(2) (1) で探索した EN_i, EN_j をクラスタリングし、生成された新しい状態遷移図を TR とする。
(3) 既定の頂点数になるまで (1)(2) の手順を繰り返す。

4.1.2 遷移図の評価指標

4.1.2.1 クラスタ指標

クラスタ指標は構造的特徴が近いエゴネットパターンが同じクラスタ内にクラスタリングされているかを示す指標である。この指標を算出するために、本節では Maximum common subgraph 法 [8] という既存のグラフ類似性評価手法を用いる。2 つのエゴネットパターンのすべての組み合わせにおける類似度を算出し、クラスタ毎にクラスタ内のエゴネットパターンの類似度最遠値を算出する。クラスタ毎のこれらの値を対応したクラスタに含まれるエゴネットの数に合わせて加重平均した値がクラスタ指標となる。クラスタ指標を算出する数式は以下の通りである。

$$Cl = \frac{\sum_{i=1} Cn_i Dist(C_i)}{Cn}$$

Cn_i は i 番目のクラスタに含まれるエゴネットの個数、Cn はエゴネットの総数、$Dist(C_i)$ は i 番目のクラスタ C_i に含まれる全ての 2 対エゴネットパターン間の類似度の内、最も小さい値を示す。指標値の範囲は [0,1] であり、値が高いほどクラスタ内がまとまっているといえる。

4.1.2.2 遷移指標

遷移指標はクラスタ間の遷移がどれだけ単純化されているかを表す指標である。各クラスタ毎にそのクラスタを始点とする辺 (遷移) に対応する遷移確率を表すラベルの値からエントロピーを算出する。クラスタ毎のこれらの値を対応したクラスタに含まれるエゴネットの数に合わせて加重平均し、[0,1] の範囲に正規化する。遷移指標を算出する数式は以下の通りである。

$$En = -\frac{\sum_{i=1} Cn_i \sum_{j=1} p_{ij} log_2 p_{ij}}{Cn * log_2 C}$$

C はクラスタの個数、p_{ij} は i 番目のクラスタから j 番目のクラスタへの遷移確率を示す。この値が高いほどクラスタ間の遷移の傾向が理解しやすいといえる。

4.2 友人関係ネットワークへの適用実験

4.1 節で与えた手法を最終的なクラスタ数を 8 として 2007〜2013 年の 4、5、6、7 月に観測された 1 年生の友人ネットワークから生成された遷移図に適用した。表 3 は得た遷移図の各指標値である。表内の k の値は閉路ベクトルモデルにおける符号化ベクトル長を意味する。

また% 表記は遷移優先、クラスタ優先モデルにおける遷移、クラスタ指標を 100% とした時の割合を表す。図 5、6 に、それぞれトップダウンモデル、閉路ベクトルモデル (k=3) を適用した遷移図を示す。ここでは遷移確率 12.5% 以上の遷移と、エゴネットの総数に対するクラスタの割合を% 表記した。閉路ベクトルモデルではクラスタ内の頻度 1 位、2 位のエゴネットパターンを示した。

4.3 実験結果の考察

クラスタ指標はトップダウンモデルが最も高く、クラスタがよくまとまっている事を示している。遷移指標は指標優先モデルが最も高いが生成された実際のクラスタを確認すると 1 つのクラスタにほぼ全てのエゴネットが集中して、そのクラスタ内の自己ループにより遷移指標を高めているため、遷移の傾向を理解するには不適切である。また、両指標はトレードオフな関係にある事が分かる。

図 5 の遷移図から見出される傾向は 3.2 節での友人ネットワークの分類結果と同じく友少クラスタとゲートウェイクラスタに含まれるエゴネットの割合が大きい。遷移の傾向としてクラスタ係数が高くなる遷移の確率が高い、すなわちクラスタ係数があまり高くないエゴネットを持つ学生は友人同士の結びつきが強固になる傾向が強い事が分かる。図 6 の遷移図からは自己ループが高い遷移確率を持ち、また次数が近いクラスタ間に遷移がある事から、これらのエゴネットは友人関係が安定しており、構造変化が微小であ

ることが分かる。

表 3　モデル別指標値

	クラスタ指標	遷移指標
トップダウン	0.501 (95.2%)	0.300 (31.8%)
閉路 (k=3)	0.193 (36.7%)	0.658 (69.9%)
閉路 (k=4)	0.160 (30.4%)	0.793 (84.2%)
閉路 (k=5)	0.152 (28.9%)	0.843 (89.5%)
遷移優先	0.105 (20.0%)	0.942 (100%)
クラスタ優先	0.526 (100%)	0.216 (22.9%)

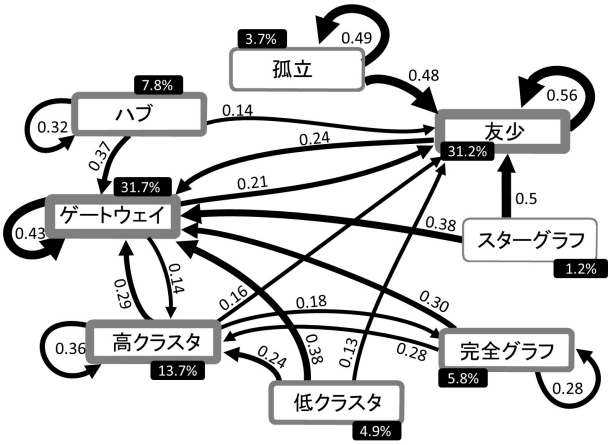

図 5　トップダウンモデル遷移図

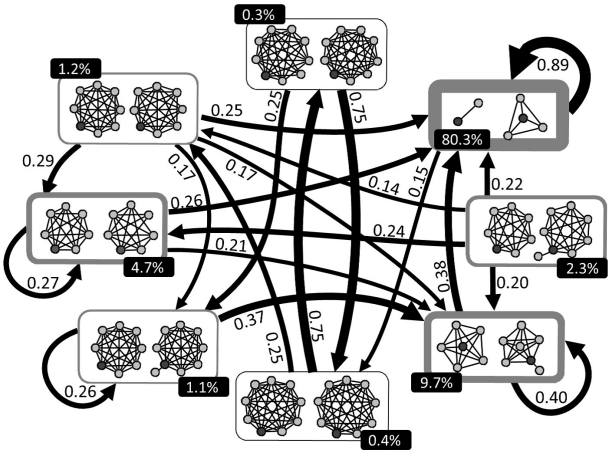

図 6　閉路ベクトルモデル (k=3) 遷移図

5. 関係発生の要因に関する分析

ここまではエゴネットを用いて個人の分類や友人関係の変化の傾向を調べた。本節では友人ネットワークの変化の中でも友人関係の発生に注目し、その要因となった人物を推定する方法を提案し、実際の友人ネットワークに対して適用する。さらに、その評価値と 3 節で行った分類との関係性について調べる。

5.1 キューピッドとパイオニア

友人ネットワークの変化において新しく関係が出来た 2 頂点 X と Y に関して、これらの両方に関係を持っていた頂点 Z が存在した場合、Z を関係発生の要因としキューピッドと定義する。Z が存在しなかった場合、X と Y を関係発生の要因としパイオニアと定義する。

図 7 に例を示す。B と G が友人になった時、その両方と既に友人だった A と H がキューピッドとなる。また C と D が友人になった時、共通の友人はいないため C と D がパイオニアとなる。

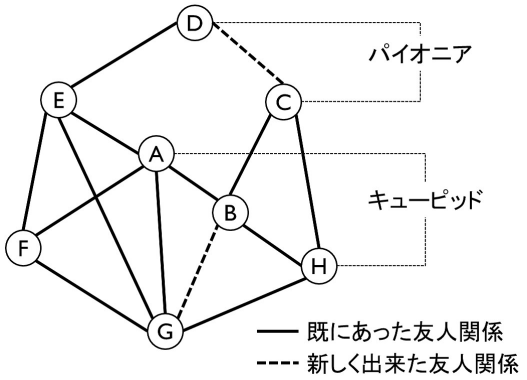

図 7　キューピッドとパイオニア

さらにキューピッド値（C 値）とパイオニア値（P 値) を定義する。算出方法は以下の通りである。

(1) 全ての頂点に C 値と P 値を与える。なお初期値は、どちらも 0 とする。
(2) ある関係が発生した 2 頂点の両方と関係を持っていた頂点の集合を F とする。$|F| > 0$ ならば、その時 F の要素であった頂点の C 値に $1/|F|$ 点を加算する。$|F| = 0$ ならば、関係が発生した 2 頂点の P 値に 1/2 点を加算する。
(3) (2) の操作を全ての発生した関係に対して行った最終的な C 値と P 値を、それぞれの頂点のキューピッド値、パイオニア値と定義する。

図 7 では A と H のキューピッド値に 1/2 点が加算され、C と D のパイオニア値に 1/2 点が加算される。つまりキューピッド値は大きいほど自分の周りの友人同士を友人にする力が強い事を表し、パイオニア値は大きいほど共通の友人が居ない人物を友人にする力が強い事を表す。

5.2 評価値の分析

1 年生の 2012 年 4〜7 月の各月の友人ネットワークを用いて、3 回の変化での各学生のキューピッド値（C 値）とパイオニア値（P 値）を調べた。表 4 に 2013 年 1 月の友人ネットワークを分類した時の各クラスタ内に属する学生の 2012 年 4〜7 月の各評価値の平均を示す。この表からクラスタによって評価値に差が出ている事が分かる。例えば

キューピッド値が大きい学生は将来、「完全グラフ」や「高クラスタ」に、パイオニア値が大きい学生は将来、「スター」や「ハブ」に属しやすいなどと推測できる。本結果は学生の評価値によって、その後の分類されるクラスタが予想できる可能性を示した。

表 4 分類と評価値

2013 年 1 月時点での	2012 年 4〜7 月の評価値	
クラスタリング結果	キューピッド値	パイオニア値
孤立	0.52	0.50
友少	0 .45	1.30
スター	0.00	2.00
完全グラフ	1.28	0.94
島クラスタ	0.88	0.85
高クラスタ	1.42	0.79
ゲートウェイ	1.20	1.36
ハブ	1.09	1.59

6. おわりに

本研究では実際の友人ネットワークと各種ネットワークモデルに対してエゴネットの構造に基づいた分類を行って結果を比較した。その結果、実際の友人ネットワークは、どのネットワークモデルとも異なる特有の分類結果になる事や「ゲートウェイ」はクラス間を繋いでいる事が多い事が分かった。さらに時間と共に変化するエゴネットのクラスタ間の推移の分析を行った。その結果、クラスタ係数があまり高くないエゴネットを持つ学生は友人同士の結びつきが強固になる傾向が強い事が分かった。また関係発生の要因となった頂点をキューピッドまたはパイオニアとし、それらを数値化した評価値を用いて分類との関係性を分析した。その結果、学生の評価値によって、その後の分類されるクラスタが予想できる可能性を示した。

今後の課題としては 1 か月ごとでの友人ネットの変化ではなく、より短い期間での変化で調べる事や閉路ベクトルモデルでのエゴネットのクラスタリングをネットワークモデルにも適用し、比較する事などが挙げられる。他にも評価指標の改良による有効なボトムアップ手法の提案、及び異なる友人ネットワークにおける遷移傾向の分析、評価値と社会的属性との関係性の調査などが挙げられる。

謝辞

本研究の一部は、名古屋工業大学情報基盤センターより支援を受けたものであり、ここに深く感謝致します。

参考文献

[1] Stanley Wasserman and Katherine Faust: *Social Network Analysis: Methods and Applications*. Cambridge University Press (1994).
[2] 竹内晨, 犬塚信博: エゴセントリックネットワークのパターンマイニング, 情処全大講演論文集 pp.653-655 (2011).
[3] N. Inuzuka, S. Takeuchi, H. Matsushima: *Pattern Mining on Ego-Centric Networks of Friendship Networks*. in Lecture Notes in Computer Science, vol. 6884 (L. C. Jain et al (eds.) Knowledge-Based and Intelligent Information and Engineering Systems), Springer, pp.89-97 (2011).
[4] N.Inuzuka, T.Nakano, K.Shimomura: *Friendship Analysis Using Attendance Records to University Lecture Classes* , IASK International Conference Teaching and Learning pp.478-486 (2008).
[5] H.Matsushima, S.Kadosaka, S.Yamamoto and N.Inuzuka: *Analysis of Friendship Network Using Attendance Records to Lecture Classes*, 30th Sunbelt Conf., (Tech. Rep., Inuzuka labo.) (2010).
[6] N. Inuzuka, T. Kondo, and S. Yamamoto: *Analysis of Asymmetric Friendship among Students from Class Attendance Records*, Studies in Computational Intelligence, vol. 199 (K. Nakamatsu et al (eds.), New Advances in Intelligent Decision Technologies), pp.393-403 (2009).
[7] Bin-Hui Chou, Einoshin Suzuki: *Discovering Community-Oriented Roles of Nodes in a Social Network*, Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science Volume 6263 pp.52-64 (2010).
[8] Horst Bunke, Kim shearer: *A graph distance metric based on the maximal common subgraph*, Pattern Recognition Letters 19 pp.255-259 (1998).

付　　録

A.1 友人スコア生成手法

下村らの出欠データからの友人関係の推測では、ある 2 人の学生間の打刻差時間のデータの列 T に対して、その 2 人が友人であるという事象 f の事後確率 $P(f \mid T)$ は、ベイズの定理を用いて以下の通り得られる。

$$P(f \mid T) = \frac{P(f) \cdot P(T \mid f)}{P(T)}$$

ここで、打刻差データ T 中の各要素 t について発生確率 $P(t)$ は、各々独立であると仮定する。また、事象 f に対する条件付き確率 $P(t \mid f)$ も各要素で独立であるとする。このとき、次の通りに変形することができる。

$$P(f \mid T) = P(f) \prod_{t \in T} \frac{P(t \mid f)}{P(t)} \tag{A.1}$$

次に、打刻差時間 t で打刻する全ての打刻対のうち、友人であるものが行った打刻対の割合を考える。この値を r_t としたとき、友人ペアに限定した打刻差 t の総データ数を打刻差 t の総データ数で割れば良いので次のように表せる。

$$r_t = \frac{X_f \cdot m_f \cdot P(t \mid f)}{X \cdot m \cdot P(t)}$$

ここで X は学生ペアの数、X_f は友人ペアの数、m は 1 組の学生ペアから発生する打刻データの平均件数、m_f は友人ペアに限定したとき 1 組の学生ペアから発生する打刻差データの平均件数である。

すると、次の式が得られる。

$$P(t \mid f) = \frac{X \cdot m \cdot P(t) \cdot r_t}{X_f \cdot m_f}$$

$X_f = X \cdot P(f)$ であるため、

$$P(t \mid f) = \frac{m \cdot P(t) \cdot r_t}{P(f) \cdot m_f}$$

となる。これを式 (A.1) に代入する。

$$\begin{aligned} P(f \mid T) &= P(f) \prod_{t \in T} \frac{m \cdot r_t}{m_f \cdot P(f)} \\ &= P(f)^{(1-n)} \left(\frac{m}{m_f}\right)^n \prod_{t \in T} r_{tf} \end{aligned}$$

ここで n は T に含まれるデータ数である。

また、友人でない確率 $P(\overline{f} \mid T)$ も同様に、

$$P(\overline{f} \mid T) = P(\overline{f}) \prod_{t \in T} \frac{P(t \mid \overline{f})}{P(t)} \tag{A.2}$$

である。また、各打刻差時間での友人以外のペアの割合は友人でないペアの打刻差 t の総データ数を打刻差 t の総デー タ数で割れば良いので次の通り表せる。

$$1 - r_t = \frac{X_0 \cdot m_0 \cdot P(t \mid \overline{f})}{X \cdot m \cdot P(t)} \tag{A.3}$$

ここで X_0 は友人以外のペアのみの数、m_0 は友人ペア以外に限定したとき 1 組の学生ペアから発生する打刻データの平均件数である。

式 (A.3) と $X_0 = X \cdot P(\overline{f})$ より $P(t \mid \overline{f})$ を求め、式 (A.2) に代入して次を得る。

$$P(\overline{f} \mid T) = P(\overline{f})^{(1-n)} \left(\frac{m}{m_0}\right)^n \prod_{t \in T} (1 - r_{tf})$$

$P(f \mid T)$ にロジット関数を用いた式 (A.4) を友人スコアとする。

$$\begin{aligned} logit P(f \mid T) &= log \frac{P(f \mid T)}{1 - P(f \mid T)} \\ &= log(P(f \mid T)) - log(P(\overline{f} \mid T)) \end{aligned} \tag{A.4}$$

式は (A.4) より求めた友人スコアが正であれば友人であると判断する。