

任意のネットワークを用いた情報拡散の統計的分析

臼井 翔平^{1,a)} 鳥海 不二夫¹

概要: 本研究では、情報拡散とネットワーク構造の関係を明確にする。近年、ソーシャルメディアの普及に伴い、友人関係ネットワーク上の情報拡散は様々な影響を社会現象に与えている。この情報拡散の拡散率を左右する最も大きな要因としてネットワークの構造が挙げられる。既存研究では、極一部の構造のネットワークを利用した分析が行われている。これに対して、本研究では様々なネットワーク構造を含むネットワークデータセットを用いる事によって、統計的な分析を行う。分析によって5つの構造特徴の内、情報拡散に影響を与える構造特徴とそうでない構造特徴を分類した。その結果、平均経路長 L 、クラスタ係数 C 、次数分布が情報拡散に影響を与え、次数相関 r は影響を与えていない事を明らかにした。平均経路長 L 、クラスタ係数 C が小さく、次数分布がべき分布ではないネットワークにおいて高い拡散率が得られた。

Influence analysis of information diffusion focusing on directed networks

SHOHEI USUI^{1,a)} FUJIO TORIUMI¹

1. はじめに

情報は個々人の相互作用によって、個人から個人へと伝播されていく [1]。ある個人が得た情報は個人から他者へと受け渡され、全体へと拡散されていく。この時、全く知らない誰かに情報を伝搬させるとは考えにくく、相互に関係する他者へと情報を伝える事になる。すなわち情報の拡散は、個々人が形成する人間関係のネットワーク上で行われる。

近年では、ソーシャルメディアの発展に伴い、人間関係ネットワークが目に見える形で具現化され、ネットワーク上での情報の伝達が明確になった。ソーシャルメディア上の多くのユーザは、ソーシャルメディアによって情報収集・発信をしているため、このソーシャルメディア上の情報拡散が注目を集めている。この情報収集・発信は、個人的な趣味趣向に関する情報だけに留まらず、政治や経済といった情報にまで至る。ソーシャルメディア上の情報拡散は、マーケティング [2]、政治的意見形成^{*1}等、様々な影響を現実の社会現象に与えている。

情報拡散の拡散率を左右する最も大きな要因としてネットワークの構造が挙げられる [3]。ネットワークの構造によって、情報が大きく拡散される事も、拡散されない事も考えられる。このネットワーク構造と情報拡散の関係については、多くの研究者が分析を重ねている [4], [5]。しかし、これらの分析で用いられているネットワークは、ネットワーク全体の構造からみれば一部のネットワークでしかない。このような一部のネットワークのみを用いても統計的に正しい結果は得られない。

これに対して、本研究では様々なネットワーク構造を含むネットワークデータセットを用いる事によって、統計的な分析を行う。本研究では、提案手法によるデータセットに含まれる各ネットワークの上で情報拡散モデルによる情報拡散シミュレーションを行う。これにより、各ネットワークの情報拡散能力を測定し、情報拡散能力とネットワークの構造特徴との関係について分析を行う。この分析により、情報拡散とネットワーク構造の関係を明確にする。

2. ネットワークデータセット

2.1 ネットワークデータセットの構築

様々なネットワークを含むデータセットを構築するため

¹ 東京大学大学院工学研究科
〒113-8654 東京都文京区本郷 7-3-1

^{a)} usui@crimson.q.t.u-tokyo.ac.jp

^{*1} 出典：クチコミ@参院選 2010 Twitter 分析レポート

に、貪欲なネットワーク生成手法を用いる [6]。まず、以下の特徴の範囲内から特徴点を抽出する。

- 平均経路長 L ($2 \leq L \leq 6$),
- クラスタ係数 C ($0 \leq C \leq 0.6$),
- 次数相関 r ($-0.5 \leq r \leq 0.5$),
- ベータ分布のパラメータ $\log \alpha$ ($0 \leq \log \alpha \leq 3$)
- ベータ分布のパラメータ $\log \beta$ ($0 \leq \log \beta \leq 4$)

なお、ベータ分布のパラメータ α 及び β は、ネットワークに対してベータ分布を最適化したパラメータである。

ベータ分布 $q(k)$ は以下の式で表される。

$$q(k) = \frac{x^\alpha(1-x)^\beta}{B(\alpha, \beta)}, \quad (1)$$

$$B(\alpha, \beta) = \int t^{\alpha-1}(1-t)^{\beta-1}, \quad (2)$$

なお、 α 及び β はフィッティングパラメータである。ここで、次数分布における x は以下の式で表される。

$$x = \frac{k}{k_{max} + 1}. \quad (3)$$

また、ベータ分布を確率分布とする (すなわち、 $\sum_k^{k_{max}} q(k) = 1$) ために、以下の処理を行う。

$$q'(k) = \frac{q(k)}{\sum_{k'}^{k'_{max}} q(k')}. \quad (4)$$

ベータ分布は α 及び β を調整する事によって、様々な分布の形状を表現する事が可能である。ネットワークの次数分布に対して α 及び β を、焼きなまし法 (Simulated Annealing:SA) [9] によって最適化する。最適化の際の目的関数として、ネットワークの次数分布 $p(k)$ とベータ分布 $q(k)$ の分布間二乗距離を、この特徴に関する距離として定義する。

次に、抽出された点を基に、Greedy なネットワーク生成手法を用いてネットワークを生成する。この時、生成されたネットワークと抽出された特徴点の距離が $D \geq 0.1$ である時、そのネットワークは破棄するものとする。これを繰り返し行う事によって、ネットワーク構造特徴空間上を網羅するようなネットワークデータセットを構築する事が出来る。

本研究では、ノード数 1000, リンク数 10000 のネットワークを 10000 個生成し、生成したネットワークに関して、全ての特徴量を計算し直した。さらに、類似したネットワーク構造を除外するために、半径 $D_F = 0.1$ の多次元球内にネットワークが高々 1 つとなるように無作為抽出した。その結果として、7028 個のネットワークを含むデータセット G_n を得た。fig:newArea2 に 3 次元特徴空間 (L, C, r) への射影した図を示す。

2.2 ネットワークの拡散率 δ

情報拡散とネットワーク構造の関係について分析するた

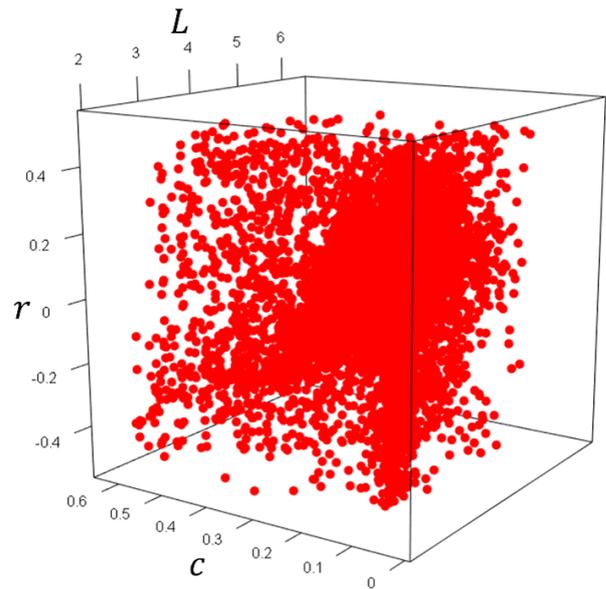


図 1 提案手法によるネットワーク群

めに、前節で示したネットワークデータセットの各ネットワークに対してそれぞれ情報の拡散率 δ を求める。拡散率 δ は、各ネットワーク上で独立カスケードモデル (IC モデル) [7] による情報拡散シミュレーションによって求める。

IC model は情報の送り手を主体とした情報拡散モデルである。IC model において、リンクで結ばれたノード v, u は情報拡散率 s_{vu} ($0 \leq s_{vu} \leq 1$) をもつ。ある時刻 t において、アクティブになったノード v は、リンクで結ばれた非アクティブなノード u をアクティブにする機会を一度だけ与えられる。すなわち、時刻 t においてアクティブになったノード v は時刻 $t+1$ に一度だけ情報の発信を行う。その際、ノード v が情報を発信する確率、つまり、ノード v の隣接ノード u がアクティブになる確率は s_{vu} となる。その施行が成功した場合、ノード u は時刻 $t+1$ においてアクティブとなる。この情報拡散過程は、アクティブなノードが増加しなくなった時点で終了する。

各モデルの $t=0$ において、初期アクティブノード v_s を用意する。この時、収束後のアクティブノードの割合を $\delta(v_s)$ とし、拡散率 δ を以下のように定義する。

$$\delta = \frac{1}{N} \sum \delta(v_s) \quad (5)$$

なお、 N はネットワークのノード数とする。各ネットワークで 10 回拡散率 δ を算出し、平均をそれぞれのネットワークのモデルにおける拡散率 δ とする。

3. 拡散現象の統計的分析

3.1 既存研究の検証

ここでは、まず既存のネットワーク構造と情報拡散との関係に関する知見について、本データセットにおける分析においても同様の事が言えるかどうかを検証する。Badhamら [8] は Badhama-Stocker (BS) アルゴリズムを用いる事に

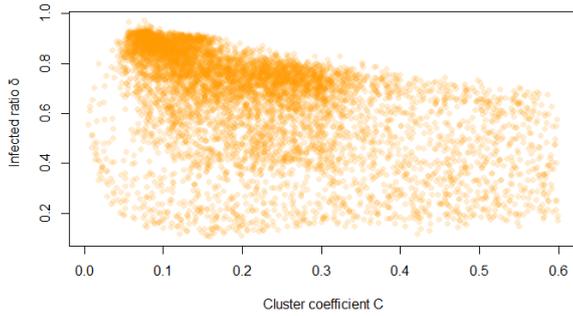


図 2 Infected degree vs Cluster coefficient

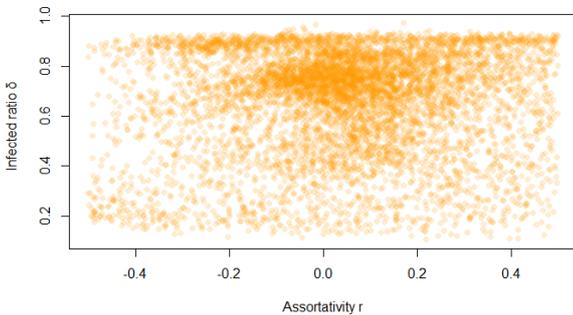


図 3 Infected degree vs Assortativity

よって、様々なクラスタ係数 C と次数相関 r を持つネットワークを生成し、その上で実験を行った。この結果より、次数相関 r が低く、かつクラスタ係数 C が低い方が拡散率が高い事を示している。

本研究での結果を分析し、彼らの主張と比較する。fig:AIDvsC に、本データセットにおける、拡散率 δ とクラスタ係数 C の関係を示す。縦軸に拡散率 δ 、横軸にクラスタ係数 C をとり、各点はデータセットに含まれるネットワークである。クラスタ係数 C が低い構造の時に、多くのネットワークにおいて、高い拡散率 δ を獲得している事がみられる。この相関係数は -0.511 と非常に強い負の相関が示されている。すなわち、クラスタ係数が低い程、ネットワークの拡散率は高い事が示された。従って、クラスタ係数 C に関しては、本研究の分析と一致している事が確認できる。

続いて、fig:AIDvsr に拡散率 δ と次数相関 r の関係を示す。縦軸に拡散率 δ 、横軸に次数相関 r をとり、各点はデータセットに含まれるネットワークである。次数相関 r が高い時も低い時も同様に、拡散率 δ が高いネットワークと低いネットワークが共存している事がみられる。この相関係数は 0.0650 と非常に低く、相関は全くない事がわかる。

なぜこのような違いが生じたのかについて考察する。彼らが用いた BS アルゴリズムでは、次数相関と平均経路長が相関係数 0.723 で相関している。すなわち、次数相関が

表 1 BS モデルの各特微量と拡散率 δ との間の相関係数

	L	C	r	$\log\alpha$	$\log\beta$
$P_c(1.2)$	-0.800	-0.735	-0.634	0.100	0.243

ない時に平均経路長が最も短くなる。ところで、平均経路長と拡散率 δ が相関するのは自明である。経路が長ければその分ネットワークに存在する全てのノードに情報が届く確率は低くなるためである。すなわち、平均経路長が短い程拡散率が高くなる。Badham らの主張は次数相関がない時の方が拡散率が高いという知見であるが、次数相関がない時は平均経路長が短いため拡散率が高いと考えれば、我々の知見と合致する。

Badham らは分析する際に平均経路長の値については言及していない。従って、彼らの主張する、拡散率 δ と次数相関 r の関係は偽相関である可能性が高い。ここで、Badham らが用いた BS アルゴリズムの各特微量と拡散率 δ との間の相関係数を `tb:correlationBS` に示す。ここから、平均経路長 L 、クラスタ係数 C 、次数相関 r と拡散率 δ に高い負の相関が得られた。拡散率 δ と次数相関 r の関係から平均経路長 L の関係を取り除いた偏相関係数 $r_{r\delta\cdot L}$ を求める。

$$r_{r\delta\cdot L} = \frac{r_{r\delta} - r_{r\delta}r_{rL}}{\sqrt{1 - r_{r\delta}^2}\sqrt{1 - r_{rL}^2}} \quad (6)$$

偏相関係数は、 $r_{C\delta\cdot L} = -0.119$ となり、次数相関 r と拡散率 δ の間はずかにかにしか相関がない事がわかる。従って、次数相関と拡散現象に関係があるとは言えない。

3.2 拡散率 δ に影響を与える特徴の分析

3.2.1 拡散率 δ と各特微量間の相関係数

ここでは一つ一つの構造特徴と拡散率 δ の関係を見るために、今回採用した 5 つの特徴と拡散率 δ の相関を分析する。

`tb:MIC` に、各特徴と拡散率 δ との相関係数を示す。`tb:MIC` より、平均経路長 L 、クラスタ係数 C に負の相関があり、次数分布のパラメータ $\log\alpha$ に正の相関がある事が明確となった。また、次数分布のパラメータ $\log\beta$ と次数相関 r は大きな相関がない事がわかる。以降では、相関のあった特徴の内まだ分析していない平均経路長 L 及び次数分布のパラメータ $\log\alpha$ についてそれぞれ見ていく事とする。

3.2.2 拡散率 δ と平均経路長 L の関係についての分析

fig:AIDvsL に平均経路長 L と拡散率 δ の関係を示す。縦軸に拡散率 δ 、横軸に平均経路長 L をとり、各点はデータセットに含まれるネットワークである。まず、このグラフは右肩下がりになっている事から、平均経路長 L が高い

表 2 IC model の拡散率 δ と各特長の相関係数

L	C	r	$\log\alpha$	$\log\beta$
-0.502	-0.511	0.0650	0.398	-0.0637

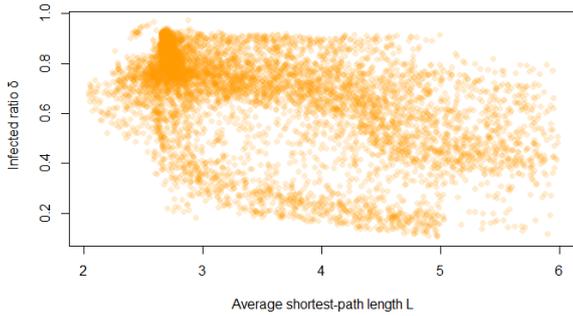


図 4 平均経路長 L と拡散率 δ の関係

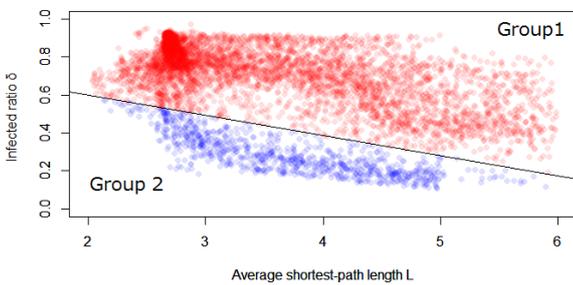


図 5 平均経路長 L と拡散率 δ の関係によるグループ分け

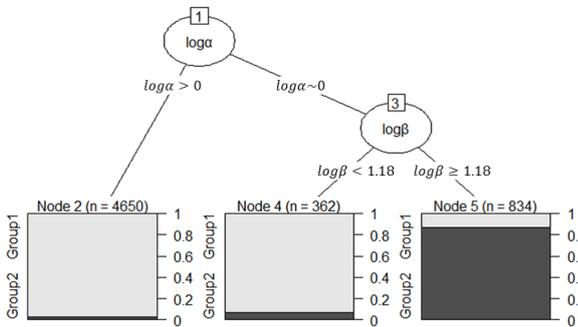


図 6 グループに関する決定木 (平均経路長 L に関して)

程, 拡散率 δ は小さい傾向にある事がわかる. この相関係数は -0.502 である.

次にグラフの中央付近から右側にほとんど点が存在していない領域がある事に着目する. この領域の上下で2つのグループに分かれるように直線を引く. この直線は, 直線からの正規化距離が 0.05 より小さい点の数が最も少なくなるように引いた. fig:Rplot37 に fig:AIDvsL を2つのグループに分けた図を示す. グラフの下部, 拡散率の低いグループ2では, 平均経路長 L と拡散率 δ は線形に相関しており, 相関係数は -0.806 である. この2つのグループに対して決定木を構築し, fig:Rplot38 に示す. この決定木は, それぞれの条件におけるグループ1とグループ2のネットワークの割合をそれぞれ示している. これより, $\log\alpha \approx 0$

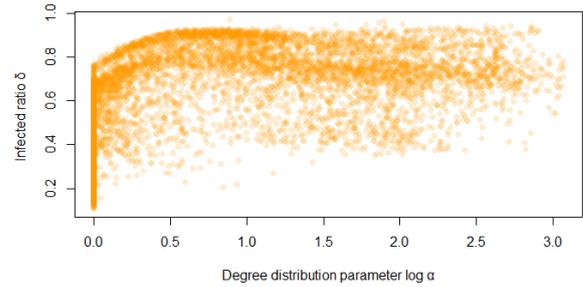


図 7 次数分布のパラメータ $\log\alpha$ と拡散率 δ の関係

かつ $\log\beta > 1.18$ であるネットワークの 87.5% がグループ2に所属している事がわかる. ところで, $\log\alpha$ が非常に小さく $\log\beta$ が大きい時, 次数分布はべき分布に近い形状を取る. 従って, 次数分布がべき分布になる時, ネットワークはグループ2に所属し, 拡散率 δ は小さくなる.

最後に, fig:Rplot37 におけるグループ1に着目する. グループ1では, 平均経路長が $L = 4.4$ を超えた辺りで拡散率 δ が急激に下がっているのがわかる. すなわち, $L > 4.4$ を閾値として情報が急激に広まらなくなっていく事が示されている. 具体的な数値に関しては, 第3.4節にて分析する.

3.2.3 拡散率 δ と次数分布のパラメータ $\log\alpha$ の関係についての分析

fig:AIDvsA に拡散率 δ と次数分布のパラメータ $\log\alpha$ の関係を示す. 縦軸に拡散率 δ , 横軸に次数分布のパラメータ $\log\alpha$ をとり, 各点はデータセットに含まれるネットワークである. これより, $\log\alpha \approx 0$ でなければ, 拡散率 δ は小さな値を取らない事がわかる. 実際に, tb:MIC より, 次数分布のパラメータ $\log\alpha$ と拡散率 δ の間には相関係数 0.398 の強い正の相関が確認されている. $\log\alpha$ が非常に小さい時, 次数分布はべき分布に近い形状を取る. すなわち, 次数分布がべき分布に近い形状を取った時は情報が拡散されにくくなると言える.

fig:AIDvsA について, 点の密度が濃い部分が2箇所見られる. 密度の濃い部分だけを見ると, $\log\alpha$ が高くなると急激に上昇し, $\log\alpha = 0.5$ 付近から徐々に下がっていく様子が見える. これを分析するために, 以下の式を用いて分析を行う.

$$\delta = \begin{cases} \frac{a}{1+b \cdot \exp(c \cdot \log\alpha)} & (\log\alpha \leq d\Gamma) \\ e \cdot \log\alpha + \frac{a}{1+b \cdot \exp(c \cdot d)} - e \cdot d & (\log\alpha > d\Gamma) \end{cases} \quad (7)$$

この時, a, b, c, d, e, f はフィッティングパラメータであり, これらの値を, この曲線から正規化距離 0.05 の範囲にデータセットが多く存在するように設定した. これにより, 2本の曲線を引き, 各ネットワークをどちらの曲線に近いかによって, 2つのグループに分けた.

fig:Rplot40 に fig:AIDvsA を2つのグループに分けた図

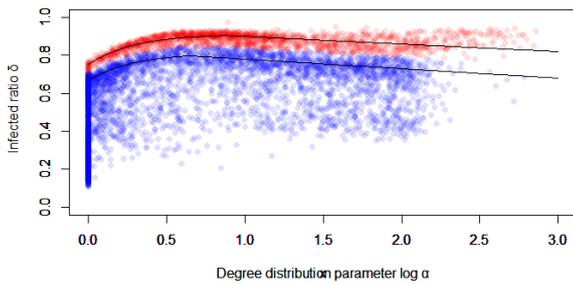


図 8 度数分布のパラメータ $\log\alpha$ と拡散率 δ の関係によるグループ分け

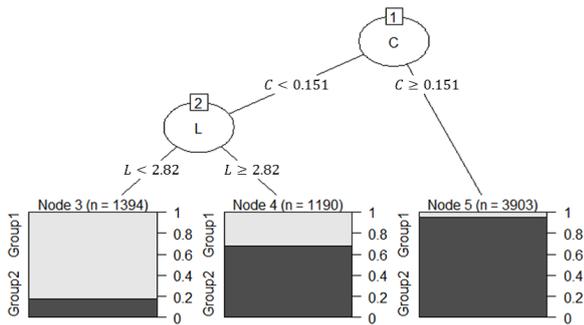


図 9 グループに関する決定木 (度数分布のパラメータ $\log\alpha$ に関して)

を示す。この2つのグループに対して決定木を構築し、fig:Rplot41に示す。この決定木は、それぞれの条件におけるグループ1とグループ2のネットワークの割合をそれぞれ示している。これより、 $C \geq 0.151$ であるネットワークの94.9%が拡散率の低いグループ2に所属している事がわかる。一方、 $C < 0.151$ かつ $L < 2.82$ であるネットワークの82.1%が拡散率が高いグループ1に所属している。

3.3 拡散率 δ の推定実験

前節では、4つの特徴が拡散率 δ に影響している事を示した。ここでは、これらの4特徴が拡散率 δ の決定にどの程度寄与しているのかを分析する。本研究では、5特徴を用いて分析を行っているが、そもそもこの5特徴で拡散率 δ が決定されない可能性もある。すなわち、5特徴以外に拡散率 δ に影響している特徴が存在する可能性である。これを検証するために、本研究では、構造特徴を用いて拡散率 δ を推定する事が可能かどうかを検証する。この推定が可能であれば、推定に用いた構造特徴を用いて拡散率 δ を決定する事が出来る事を意味する。なお、回帰分析に当たり、各特徴の値は標準化した。

3.3.1 実験設定

非線形回帰分析として、サポートベクター回帰 (Support Vector Regression:SVR) [10] を用いる。SVRはガウシアンカーネルを用い、二つのハイパーパラメータを焼きなま

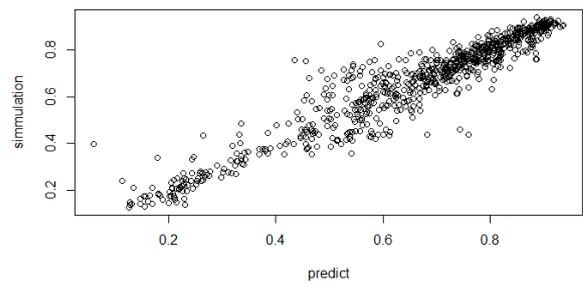


図 10 拡散率 δ の推定結果

表 3 説明変数毎の Infected degree δ の推定結果

	$L, C, r, \log\alpha$	$L, C, r, \log\beta$	$L, C, \log\alpha, \log\beta$
R^2	0.780	0.628	0.897
R^2	$L, r, \log\alpha, \log\beta$	$C, r, \log\alpha, \log\beta$	
	0.856	0.790	

し法 (Simulated Annealing:SA) [9] によって最適化する。さらに、評価関数として、推定値と実測値の決定係数 R^2 を与える。評価には8分割交差検定 (8-fold cross validation) を用いる。

3.3.2 推定結果

まず、本研究で用いた5つの構造特徴を用いて、拡散率 δ を推定可能であるかを検証する。推定結果の決定係数 R^2 は0.906であった。fig:Rplot08に推定値とシミュレーション値の関係を示す。縦軸にシミュレーション値を取り、横軸に推定値をとっている。本結果より、5つの構造特徴から拡散率 δ を十分推定できているといえる。従って、この5つの構造特徴を用いれば、拡散率 δ を決定できる。

次にそれぞれの構造特徴の寄与度を推定する。各特徴をそれぞれ抜いた4特徴から拡散率 δ を推定し、その時の下がり幅が大きい程、その特徴が与えている影響は大きい。tb:ICpredict1に各4特徴からの拡散率の推定結果の決定係数 R^2 を示す。tb:ICpredict1には、使用した4特徴とそれらの特徴からの推定結果の決定係数 R^2 が示されている。例えば、 $L, C, r, \log\alpha$ は平均経路長 L 、クラスタ係数 C 、次数相関 r 及び度数分布のパラメータ $\log\alpha$ を用いて推定実験を行った事を意味する。

まず、 $L, C, \log\alpha, \log\beta$ での推定結果に着目する。推定結果の決定係数 R^2 は0.897を示し、これは5特徴での推定結果とほぼ変わらない。この推定結果から、次数相関 r は拡散率 δ に寄与していない事を示している。この結果は、前節の結果を裏付ける内容である。

続いて、 $L, C, r, \log\alpha$ での推定結果に着目する。推定結果の決定係数 R^2 は0.780となった。この推定結果より、度数分布のパラメータ $\log\beta$ は拡散率 δ の推定に強く寄与している事が示されている。

最後に、 $L, C, r, \log\beta$ での推定結果に着目する。推定結

果の決定係数 R^2 は 0.628 を示し、最小の結果となった。この推定結果より、 $\log\alpha$ が最も推定結果に寄与している事がわかった。

3.4 決定木分析

ここまでの分析で、単独の特徴ではなく、複数の特徴を同時に見る必要がある事が示唆された。ここでは、本研究において用いている5つの構造特徴を説明変数とし、拡散率 δ を目的変数として決定木を構築した。なお、決定木の複雑度パラメータ (Complexity: cp) は $cp = 0.01$ とし、10分割交差検定の誤差は 0.241 であり、十分に精度の高い決定木であるといえる。fig:decisionTree に決定木を示す。

まず、本決定木において着目する点は、次数相関による分岐が存在しない点である。すなわち、情報拡散 δ に対して、次数相関は影響を及ぼさない事が示されている。これはここまでの結果を裏付ける。

次に着目する点は、次数分布のパラメータ $\log\alpha$ である。次数分布のパラメータ $\log\alpha$ は一番根に近いところで分岐があり、 $\log\alpha > 0$ であるかどうかで分岐している。3.2.2 節において我々は、 $\log\alpha \approx 0$ でグループ分けできる事を示した。本決定木で得られた結果はこの示唆を裏付ける。

さらに、右の枝を辿っていくと、平均経路長 L についての分岐があり、 $L < 4.26$ であるかどうかで分岐している。fig:Rplot11 にこの分岐後、つまり、 $\log\alpha > 0$ である時の平均経路長 L と拡散率 δ を示す。 $L < 4.26$ である部分を赤い点、 $L > 4.26$ である部分を青い点として表す。この図を見れば、青い点のところ、急激に拡散率 δ が低下している様子わかる。従って、拡散率は $L = 4.26$ を超えると急激に減少する事がわかる。3.2.2 節において我々は、 $L = 4.4$ 付近で急激な変化を起こしている事を示唆した。本決定木で得られた結果はこの示唆を裏付けている。

さらに、右の枝をたどっていけば、クラスタ係数 C についての分岐があり、 $C > 0.162$ であるかどうかで分岐している。fig:Rplot10 にこの分岐後におけるクラスタ係数 C と拡散率 δ の関係、つまり、 $\log\alpha > 0$ かつ $L < 4.26$ である時のクラスタ係数 C と拡散率 δ を示す。 $C < 0.162$ である部分を赤い点、 $C > 0.162$ である部分を青い点として表す。この図から、青い点のところと赤い点のところで明らかな差があるのがわかる。3.2.3 節において我々は、 $C = 0.15$ 付近で変化を起こしている事を示唆した。本決定木で得られた結果はこの示唆を支持している。

最後に、次数分布のパラメータ $\log\beta$ について触れよう。本決定木において、いくつかの次数分布のパラメータ $\log\beta$ による分岐が存在する。3.2.1 節での相関による分析では、次数分布のパラメータ $\log\beta$ の影響は見られなかった。しかし、本決定木においては、次数分布のパラメータ $\log\beta$ は拡散率 δ に影響している事が示された。すなわち、相関分析では見えてこなかった関係を見つける事ができた。つま

表 4 ネットワークサンプルの特徴量

	L	C	r	$\log\alpha$	$\log\beta$	δ
network(a)	2.54	0.0787	-0.0433	1.86	3.01	0.956
network(b)	2.58	0.0560	-0.0154	1.90	3.03	0.967
network(c)	4.97	0.159	0.218	0.00	1.77	0.112
network(d)	4.89	0.155	0.397	0.00	1.67	0.106

りこれは、 $\log\beta$ は単独の特徴のみでは影響を与えていないが、複合的な影響を与えている。

$\log\alpha \approx 0$ における次数分布のパラメータ $\log\beta$ と拡散率 δ の関係について分析する。fig:Rplot12 に $\log\alpha \approx 0$ の時の次数分布のパラメータ $\log\beta$ と拡散率 δ の関係を示す。縦軸に拡散率 δ 、横軸に次数分布のパラメータ $\log\beta$ を取り、 $\log\beta < 1.25$ であるネットワークを赤い点、 $\log\beta > 1.25$ であるネットワークを青い点として表示している。これより、明らかに、次数分布のパラメータ $\log\beta$ と拡散率 δ は相関している。相関係数は -0.819 であり、非常に強い負の相関を示している。従って、次数分布のパラメータ $\log\alpha \approx 0$ である時、次数分布のパラメータ $\log\beta$ と拡散率 δ は強く相関する事がわかった。

ところで、 β が大きい時のベータ分布の形状についてみる。fig:beta-distribution2 に $\log\alpha = 0$ の時のベータ分布の形状変化を示す。 β が大きくなっていくと、次数分布の傾きが大きくなっていく。すなわち、大きなハブノードができる事となる。これは、べき指数を増やしていった時と同じである。従って、べき乗が強くなるほど、情報が伝わりにくくなる事が示唆された。

本節では決定木を構築して拡散率と各特徴に関して分析を行い、特徴同士の複合的な影響を見る事が出来た。

3.5 ネットワークサンプル

最後に本データセットにおいて、最大の拡散率 δ と最小の拡散率 δ を持つネットワーク (a)(b) 及び、ネットワーク (c)(d) の各特徴量と拡散率 δ を tb:featureIC に示す。これらのネットワークは、4.3.2 節の推定実験において高い推定結果を得ており、この特徴は本分析結果として高い拡散率 δ を持つと判定されるものである。

最大の拡散率 δ を持つネットワーク (a) 及び (b) は fig:decisionTree の決定木においては最も右の条件に合致する。すなわち、 $\log\alpha > 0$ かつ $L < 4.26$ かつ $C < 0.162$ である。最大の拡散率 δ を持つネットワークは、ネットワーク (a)(b) 共にランダムネットワークに近いネットワークとなっている。逆に、最小の拡散率 δ を持つネットワーク (c) 及び (d) は、fig:decisionTree の決定木においては最も右の条件に合致する。すなわち、 $\log\alpha \approx 0$ かつ $\log\beta > 1.25$ かつ $C > 2.67$ である。最小の拡散率 δ を持つネットワークは、ネットワーク (c)(d) 共にスケールフリー性の強いネットワークとなっているのがわかる。

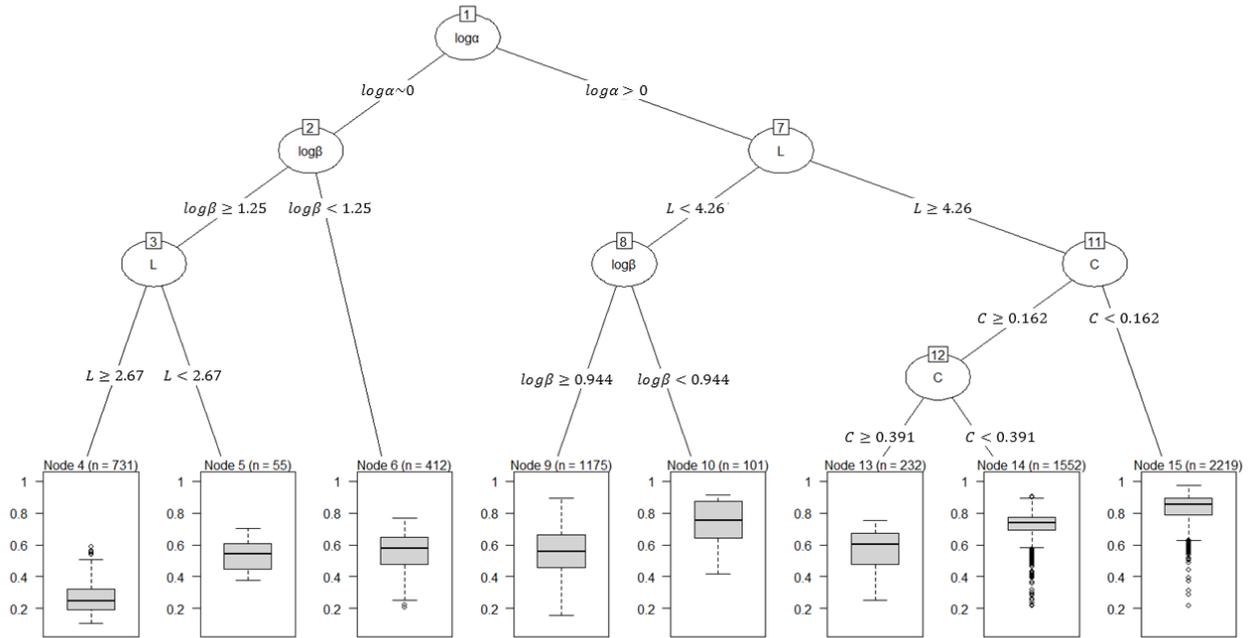


図 11 決定木

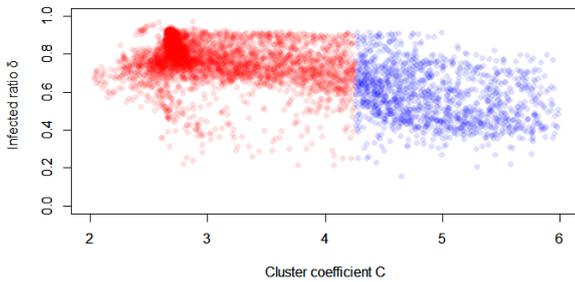


図 12 $\log \alpha > 0$ における平均経路長 L と拡散率 δ の関係 (赤 : $L < 4.26$, 青: $L > 4.26$)

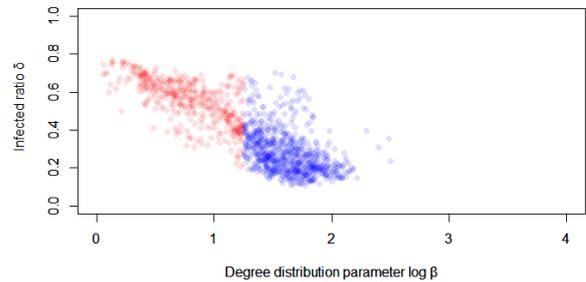


図 14 $\log \alpha > 0$ における次数分布のパラメータ $\log \beta$ と拡散率 δ の関係 (赤 : $\log \beta < 1.25$, 青: $\log \beta > 1.25$)

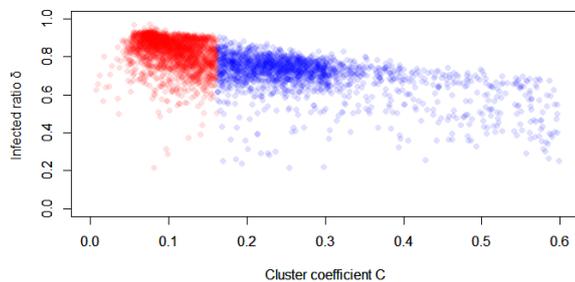


図 13 $\log \alpha > 0$ かつ $L < 4.26$ におけるクラスタ係数 C と拡散率 δ の関係 (赤 : $C < 0.162$, 青: $C > 0.162$)

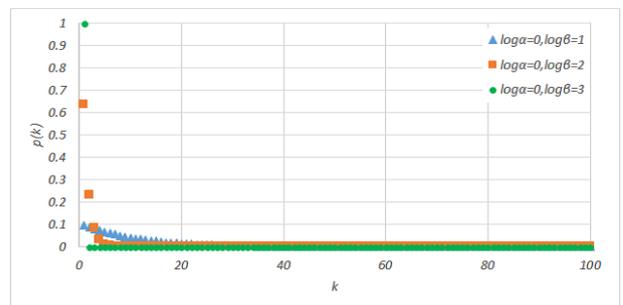


図 15 $\log \alpha = 0$ の時のベータ分布の形状変化

4. 結論

本章では、前章にて提案したネットワーク構造特徴空間

を網羅するようなネットワークデータセットを用いて、情報拡散とネットワーク構造の関係について分析を行った。分析によって5つの構造特徴の内、情報拡散に影響を与える構造特徴とそうでない構造特徴を分類した。

分析には相関分析、回帰分析、決定木分析の3つの手法

を用いた。その結果、平均経路長 L 、クラスタ係数 C 、次数分布のパラメータ $\log\alpha$ 及び $\log\beta$ が情報拡散に影響を与え、次数相関 r は影響を与えていない事を明らかにした。拡散しやすいネットワークの条件は以下の通りである。

- 平均経路長 L が小さい
- クラスタ係数 C が小さい
- 次数分布のパラメータ $\log\alpha$ が大きい
- 次数分布のパラメータ $\log\beta$ が小さい

この結果は、3つの分析で一貫しており、様々な角度から支持する結果を導いた。

本研究ではネットワーク構造特徴として、5つの特徴量のみを扱った。この5つの構造特徴以外の要素を発見し、これを含めたネットワーク構造特徴空間を埋めるようなネットワークデータセットへと拡張する必要がある。また、本研究では、ネットワークの構造特徴として、ネットワークの全体としての特徴のみを扱った。しかし、ネットワークのより細かい部分に目を向け、ミクロスケールの構造特徴を用いた分析も今後の課題として挙げる。

参考文献

- [1] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA, 2014.
- [2] 鶴見 裕之, 増田 純也, 中山 厚徳. 商品に関する twitter 上のコミュニケーションと販売実績の関連性分析 (特集: 声・質的データを経営・マーケティングに活かす). *オペレーションズ・リサーチ: 経営の科学*, Vol. 58, No. 8, pp. 436–441, aug 2013.
- [3] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Why rumors spread so quickly in social networks. *Commun. ACM*, Vol. 55, No. 6, pp. 70–75, June 2012.
- [4] Damián H. Zanette. Critical behavior of propagation on small-world networks. *Phys. Rev. E*, Vol. 64, p. 050901, Oct 2001.
- [5] Jie Zhou, Zonghua Liu, and Baowen Li. Influence of network structure on rumor propagation. *Physics Letters A*, Vol. 368, No. 6, pp. 458 – 463, 2007.
- [6] 白井 翔平, 鳥海 不二夫. 情報拡散に影響するネットワーク構造特徴. *人工知能学会論文誌*, Vol. 30, No. 1, pp. 195 – 203, 2015.
- [7] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pp. 137–146, New York, NY, USA, 2003. ACM.
- [8] Jennifer Badham and Rob Stocker. The impact of network clustering and assortativity on epidemic behaviour. *Theoretical Population Biology*, Vol. 77, No. 1, pp. 71 – 75, 2010.
- [9] Scott Kirkpatrick, Daniel C. Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *SCIENCE*, Vol. 220, No. 4598, pp. 671–680, 1983.
- [10] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, Vol. 14, No. 3, pp. 199–222, 2004.