3N4-OS-28b-2

任意のネットワークを用いた情報拡散の統計的分析

Influence analysis of information diffusion focusing on directed networks

臼井 翔平 *1　　鳥海 不二夫 *1
Shohei Usui　　Fujio TORIUMI

*1 東京大学大学院工学系研究科
School of Engineering The University of Tokyo

In this paper, we find the most effective network for information diffusion. In the disaster situation, It is necesary to spread information widly. Network structure is one of the factor for deciding diffusion degree of information diffusion. We can discover network structure which maximize diffusion degree. However, the most effective structure has been not clear. Therefore, we try to find the most effective network to analyze feature which effect diffusion process with decision tree analysis. Although there are a lot of research of relation between network structure and infomation diffusion, almost all reseachers use basic network models. We cannot discuss unified argument from analysis with the network dataset generated by these basic network model because they only generate extremely part of the networks. In this research, we analyze with decision tree constructed by diffusion simulation on the network dataset which have large expression. IC model and LT model are used as information diffusion simulation. As a result, degree distribution bias and average shortest-path length effect simuration result in both IC model and LT model, and other features are not factor for maximizing diffusion degree. Therefore, low degree distribution bias and low average shortest-path length are important for effective information diffusion.

1. 序論

情報拡散の最大化問題は複雑ネットワーク科学における重要な課題の一つである [**?**]．この情報拡散は，人々が形成する人間関係のネットワーク上で行われるため，ネットワーク構造の影響を大きく受ける．しかし，現在ではどのようなネットワーク構造で拡散率が最大化するかはわかっていない．

ネットワーク構造と情報拡散の分析は多く行われている [**?, ?**]．しかし，多くの分析においては、基本的なネットワーク生成モデルで生成したネットワークが使われている．これらのネットワーク上で分析を行っても，ネットワークが取り得る特徴空間全体からすれば極一部の構造であり，択一的な議論は行えない．

本研究では，様々な構造のネットワークを含むデータセットを用い，決定木分析によって統計的に分析を行う事により，拡散現象に影響している特徴を明らかにし，情報拡散が最大となるネットワークを発見する．

2. 決定木の構築

本研究では決定木を用いることで，ネットワークの構造特徴が情報拡散に与える影響について分析を行う．本節では，決定木の目的変数とする Average Infected Degree(AID) 及び，説明変数とするネットワーク構造特徴の概説を行う．また，本研究で用いるデータセットについてもここで説明する．

2.1 Average Infected Degree(AID)

情報拡散は，広く感染症の伝染モデルを用いて表現される．本研究では，情報拡散モデルとして，IC model 及び LT モデル [**?**] の 2 つのモデルを用いる．それぞれのモデルの概説を行う．なお，本研究で用いるネットワークは無向ネットワーク $G(V, E)$(V は全ノード集合，E は全リンク集合) であるが，リンク $e(v, u)$ とリンク $e(u, v)$ は別のリンクとして扱う．本研究では，IC model，LT model について，それぞれ決定木を構築し分析を行う．

2.1.1 IC モデル

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，各リンク $e(v, u)$ 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．その際，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．この過程の $t = 0$ において，初期アクティブノード v_s を用意する．この時のアクティブノードの拡散率を $\sigma(v_s)$ とし，Average Infected Degree(AID)σ を以下のように定義する．

$$\sigma = \frac{1}{N} \sum_{v \in V} \sigma(v) \tag{1}$$

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 100 回 AID を算出し，平均をネットワークの IC model における拡散率とする．

2.1.2 LT モデル

これに対して，LT model は情報の受け手を主体とした情報拡散モデルである．LT model において，リンク $e(v, u)$ は影響力 w_{vu} をもつ．各重み w_{vu} は $\sum_u \in B(v) w_{vu} \leq 1$ を満たす．なお，$B_(v)$ はノード v の隣接ノード集合である．また，各ノード v は閾値 θ_v を持つ．時刻 t において，非アクティブノード v は $\sum_u \in B_t(v) w_{uv} \geq \theta$ となった時，アクティブとなる．ここで，$B_t(v)$ は時刻 t におけるアクティブな隣接ノード集合である．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．IC model の際と同様に，AID の平均を算出し，ネットワークの LT model における拡散率とする．

連絡先: 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
E-mail: usui@crimson.q.t.u-tokyo.ac.jp

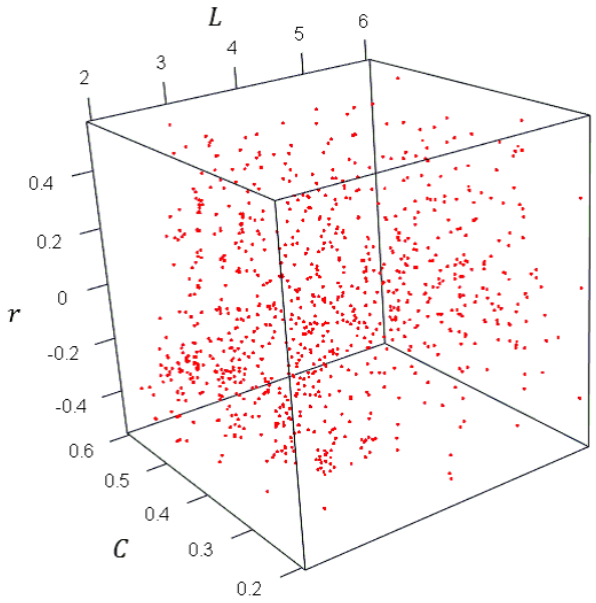


図 1: GGM の表現可能領域

2.2 ネットワーク構造特徴

本研究では，graphical game の分析のために以下の 5 つのネットワーク特徴量を用いる．

- 平均経路長 L [?]
- クラスタ係数 C [?]
- 次数相関 r [4]
- 次数分布のパラメータ α
- 次数分布のパラメータ β

次数分布のパラメータ α 及び β はネットワークの次数分布を β 分布により尤度最大化を行い，尤度が最も高い時のベータ分布のパラメータである．なお，本研究では次数分布による影響も考慮するため，べき分布以外の分布も用いる．そのため，様々な分布を取り得るベータ分布を用いた．

2.3 ネットワークデータセット

本研究では，ノード数 1000，リンク数 10000 のネットワークを 4700 個生成した．生成には，Greedy Growth Model [?] を用い，様々な特徴を持つネットワークを生成した．次数分布以外の特徴である 3 つについて，Figure 1 に 3 次元特徴空間にプロットした図を示す．この図より，本データセットが様々な特徴点であるネットワークを含んでいる事がわかる．このデータセットから特徴量と，IC モデル，LT モデルのそれぞれについての AID を算出し，それぞれのモデルに関する決定木を構築する．次節では，決定木を用いた分析の結果を解説する．

3. 決定木分析

本研究では Decision Tree を用いた統計的分析を行う事により，拡散現象への構造特徴の複合的な影響を分析する．まず，Decision Tree を構築するための目的変数と説明変数を解説し，その後使用するネットワークデータセットの解説を行い，Decision tree を用いた分析を行う．

表 1: IC model の ADI を最大化するネットワーク条件

	Highest	2nd highest	3rd highest
$\overline{\sigma}$	0.817	0.722	0.717
L	$L < 3.70$	$L \geq 3.70$	$L < 4.16$
C	-	-	-
r	-	-	-
α	$\alpha \geq 2.73$	$\alpha \geq 2.73$	$1.69 \leq \alpha < 2.73$
β	-	-	-

表 2: LT model の ADI を最大化するネットワーク条件

	Highest	2nd highest	3rd highest
$\overline{\sigma}$	0.767	0.657	0.656
L	$L < 4.41$	$L \geq 4.41$	$L < 4.37$
C	-	-	-
r	-	-	-
α	$\alpha \geq 1.94$	$\alpha \geq 1.94$	$1.14 \leq \alpha < 1.94$
β	-	-	-

3.1 IC model の決定木

決定木分析の結果を Figure 2 に示す．決定木の誤差率は 0.135 であり十分な精度であるといえる．

まず，決定木を見ると，次数分布の特徴 α 及びへ金経路長 L に関する分岐のみが存在している事がわかる．したがって，AID を左右しているのが，この 2 つの特徴のみであることが示唆される．各分岐を見てみると，次数分布の特徴 α は大きい方が，平均経路長 L は小さい方が AID の値が大きい事がみられる．

また AID が大きい 3 つの条件を Table 1 に示す．3 つの最大条件に共通する事として，α が大きいことが挙げられる．決定木の葉ノードの中ではこの 3 つのノードは α が最も大きい 3 ノードである．したがって，α が大きい事は AID が高くなるための必要条件となる．α が小さい時，次数分布の偏りは大きくなり，次数分布はべき分布に近い分布となる．Figure 3 に AID 最大の時の次数分布と最小の時の次数分布をそれぞれ示す．したがって，次数分布の偏りが小さい事が AID が高くなる条件である．

次に，AID が最大となる場合と 2 番目に大きい場合を比較すると，最大となる場合は平均経路長 L が小さい事がわかる．また，3 番目に大きい場合も平均経路長 L が小さいという事がみられる．したがって，平均経路長 L が小さいと AID は高い値である事は明らかである．しかし，2 番目に高い条件を代表とするように，平均経路長が大きい値であっても AID が高い値をとる事はありうる．したがって，平均経路長 L は AID が高いための必要条件ではないが，平均経路長 L が小さい方が AID が高い値をとりやすい事がわかった．

3.2 LT model の決定木

決定木分析の結果を Figure 4 に示す，決定木の誤差率は 0.182 であり十分な精度であるといえる．LT model の場合は IC model と比較して，次数分布の特徴 α と平均経路長 L の他にも，クラスタ係数 C や，次数分布の特徴 β が分岐として表れている．したがって，LT モデルによる拡散現象は，クラスタ係数 C 及び，次数分布の特徴 β の値からも影響を受けている事がわかる．しかし，次数相関 r は現れていないため，次数相関 r は AID には影響していない事が示唆された．

また，クラスタ係数 C による分岐は，$0.835 \leq \alpha < 1.14$ の

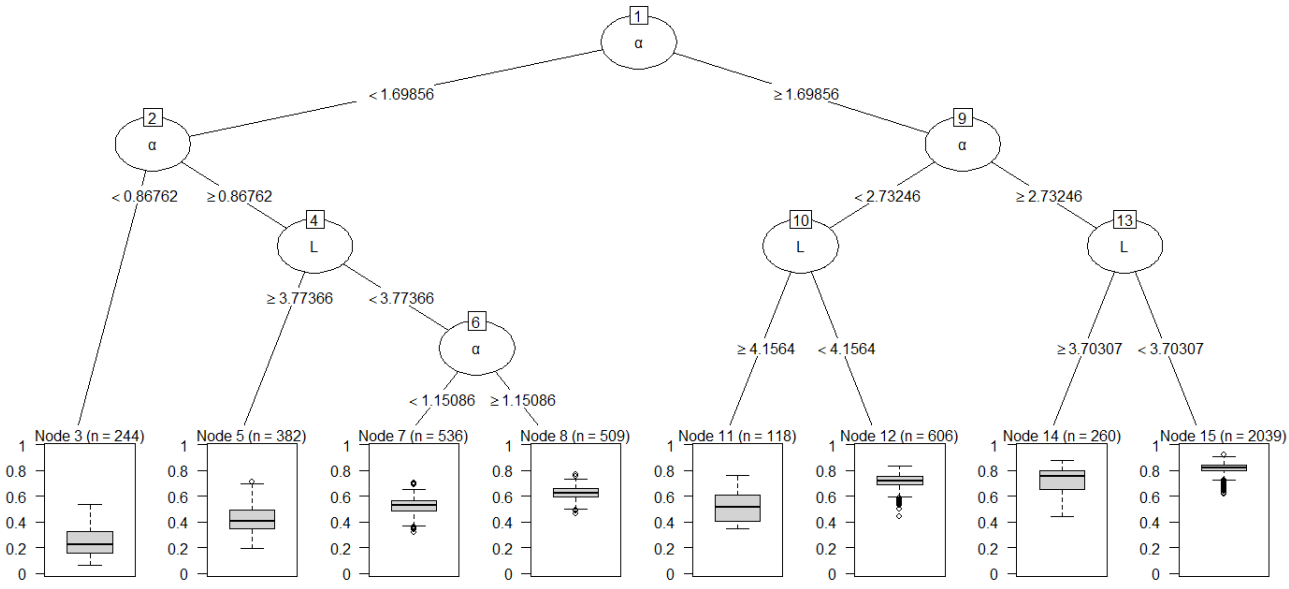


図 2: IC model に関する決定木

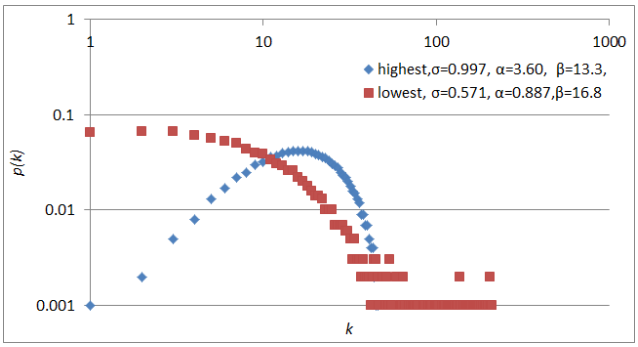


図 3: 最大 AID(blue:$\sigma = 0.997$) と最小 AID(red:$\sigma = 0.571$) の時の次数分布

時，クラスタ係数 $C < 0.571$ の方が AID が高いという分析結果が出ている．クラスタ係数 $C = 0.571$ という値を持つネットワークは非常に高いクラスタ性を有している． Figure 5 にクラスタ係数 $C \geq 0.571$ であるネットワークを示す． このように，クラスタ係数が非常に高い時，ネットワークはいくつかの塊に分裂する可能性が高い．$C \geq 0.571$ であるネットワークを louvain 法を用いてモジュラリティを求めたところ，であった．この時，情報拡散はある塊で拡散され，他の塊では拡散され難い．したがって，AID が低い値となったと考えられる．

次に，次数分布の特徴 β は $0.835 \leq \alpha < 1.14 \bigwedge C < 0.571$ の時，$\beta < 18.2$ の方が AID が高い．それぞれ，$\beta < 18.2$ と $\beta \geq 18.2$ であるネットワークの次数分布を Figure 6 に示す．β の値が大きいとロングテイルになる事が見られる．したがって，大きなハブノードが存在しない方が AID が高い事がわかる．

AID が大きい 3 つの条件を Table 2 に示す．この 3 つの条件では，次数分布の特徴 α 及び平均経路長 L のみ表れている．つまり，クラスタ係数 C 及び，次数分布の特徴 β は AID を最大化する時には関係していない事がわかる．IC モデルの場合と同様に，次数分布の特徴 α が大きい事が AID が高いための必要条件となり，平均経路長 L が小さい事も重要であるという結果となった．この事から，拡散現象にはこの 2 つの特徴が最も重要である事が示唆される．しかし，これはあくまでも 2 つのモデルに共通しているだけであるため，より多くの拡散モデルによる検証が必要となる．

また，IC モデルの AID 最大化条件 (Table 1) と LT モデルの AID 最大化条件 (Table 2) を比較する．次数分布の特徴 α に関して，IC モデルの条件は LT モデルの条件よりも高い事がわかる．対して，平均経路長 L に関して，IC モデルの条件は LT モデルの条件よりも低い事がわかる．すなわち，IC モデルの AID 最大化条件の方が，LT モデルの条件よりも厳しい条件であるといえる．逆に，LT モデルでは，何らかの別の特徴による影響が出ている可能性も示唆される．これにを明らかにするには，ネットワーク特徴量を増やし，決定木の誤差を少なくする事が必要となるが，今後の課題とする．

4. 結論

本研究では，情報拡散を最も効率的に行えるネットワーク構造を発見するために，決定木分析を行い，情報拡散に影響を与える構造特徴を明らかにした．情報拡散の拡散率には IC model 及び LT model を用いたソーシャルシミュレーションにより算出した．この拡散率から Average Infected Degree(AID) を定義し，この値を目的変数として，決定木を構築した．この決定木は，IC model と LT model のそれぞれについて構築し，分析を行った．分析により，IC model の AID には次数分布に関する特徴量 α 及び平均経路長 L のみが影響している事を明らかになった．α が大きいとき，次数分布の偏りが小さいので，次数分布の偏りが小さく，平均経路長が短い時に高い AID となる事がわかった．また，LT model における AID が最大化する条件は IC model の場合とほぼ変わらない結果となった．したがって，この 2 つが情報拡散にとって重要なネットワーク構造特徴である事がわかった．

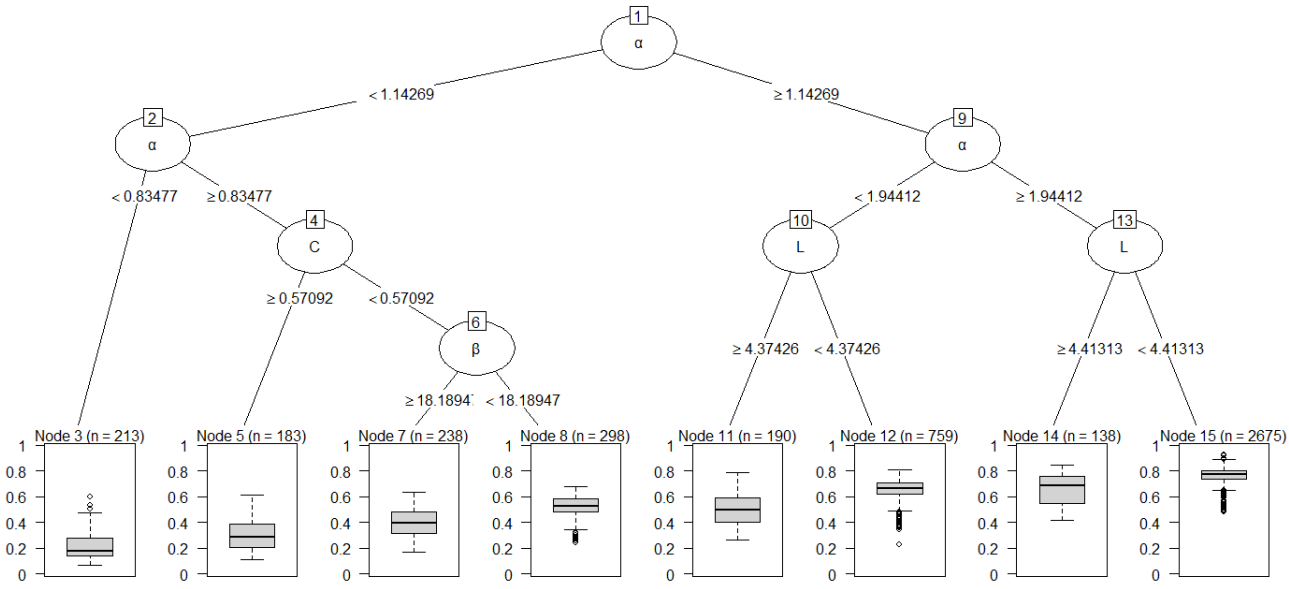
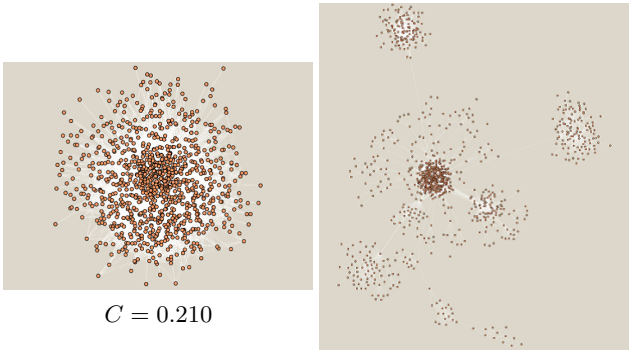


図 4: LT model に関する決定木



$C = 0.210$

$C = 0.600$

図 5: ネットワーク例

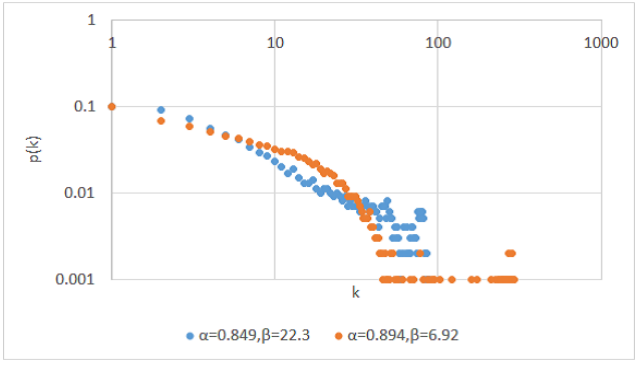


図 6: $\beta < 18.2$ 及び $\beta \geq 18.2$ の次数分布例

今回使った 2 つのモデルは情報拡散に関する基本的なモデルである．より複雑なモデルを用いて分析を行う事を今後の課題として挙げる．また，今回使った特徴は 5 つの基本的な特徴量である．そこで，今回使った以外の特徴についても分析を進める必要がある．

参考文献

[1] Fujio Toriumi, Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, and Itsuki Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pp. 1025–1028, 2013.

[2] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press, 1994.

[3] Gunther Schmidt. *Relational Mathematics*, Vol. 132 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2011.

[4] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, Vol. 67, No. 2, p. 026126, February 2003.

[5] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[6] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[7] Shohei Usui, Fujio Toriumi, Takatsugu Hirayama, and Kenji Mase. Analysis of influential features for information diffusion. In *SocialCom 2013*, Washington, D.C., September 2013.