4C1-R-6-3

ジニ係数による多重有向グラフとしての Favoritesネットワークの分析

Analyses of the Favorites Network as a multi directed graph with Gini index

加藤翔子[*1] 小出明弘[*1] 伏見卓恭[*1] 斉藤和巳[*1]
Shoko Kato Akihiro Koide Takayasu Fushimi Kazumi Saito

[*1]静岡県立大学
University of Shizuoka

To construct a new tendency model of users, we analyzed the Favorites system of Twitter from the point of view of multi directed graph with Gini index. Concretely, we explored degree distribution, correlations between indegree and outdegree, average of multiplicity, and Gini index of the Favorites network, random network, and Mentions network of Twitter. As results of the analyses, we observed existence of freaks in Gini index plot of the Favorites network, and structure of the network which has edges biased for particular users in compareing random networks.

1. はじめに

日本のような資本主義国家において，現実社会における消費者の嗜好や風潮，流行などの流れを把握し操作することは，経済的成功を収める為の重要な要因の一つである．従って，その動向を理解するために，実データとして収集し分析・モデリングを行うことも，重要視されている研究対象である．しかし，抽象的概念である嗜好を具体的数値へデータ変換することは容易なことではなく，そのため存在するデータの数も限られており，新たなユーザー嗜好の指標の発見が待たれている．

先立って発表した論文 [1] では，Twitter のツイートをお気に入り登録して保管できる Favorites 機能に着目し，新たなユーザー嗜好モデルの構築へ向けての基礎として，有向グラフや 2 部グラフの観点から分析したが，本論文では，これを多重有向グラフとしてとらえ，ユーザー毎の Favorites する頻度を，格差を評価する指標であるジニ係数などを用いて分析する．具体的には，隣接数分布，入出隣接数相関，多重度平均，ジニ係数の 4 つを Favorites ネットワークと同規模ランダムネットワークについて探究するとともに，同じく Twitter における機能の一つである Mentions 機能についても同様に分析し，結果を比較する．

Twitter [*1]は，急激なユーザー数の増加から，2007 年 3 月の一般公開以来，Huberman ら [2] や Kwak ら [3] など多くの研究者に注目され，様々な知見が得られているが，ユーザー嗜好モデルとしての研究，ひいてはジニ係数を用いた研究は既存の研究には見られず，従って本研究は意義のあるものと考える．

本論文の構成は以下となる．まず，本研究に用いた分析法について説明する．次に，Favorites 機能と Mentions 機能の概要や分析に用いたデータについて説明するとともに，評価実験による結果を報告する．最後に，本研究のまとめについて述べる．

なお，本研究で用いるネットワークを，本論文ではそれぞれ**「Favorites ネットワーク」「Mentions ネットワーク」**と呼ぶ．また，Favorites する・されるといったユーザーの行動をそれぞれ**「ふぁぼる・ふぁぼられる」**と呼び，ツイート内容に半角アットマークとそれに続くスクリーンネームを含むものを **@-message** と呼ぶ．

連絡先: 加藤翔子，静岡県立大学経営情報学部，〒 422-8526 静岡県静岡市駿河区谷田 52-1，054-264-5008，b09032@u-shizuoka-ken.ac.jp

∗1 http://twitter.com/

2. 分析法

Favorites ネットワークのノード（ユーザ）集合を $V = \{u, v, w, \cdots\}$ とする．いま，ある観測期間において，ノード u が v のツイートをふぁぼった回数を $m_{u,v}$ とし，ノード u から v へ $m_{u,v}$ 本のリンクを張ったとする多重有向グラフを考える．なお，Mentions ネットワークについても，ノード u から v への@-message 送信回数を $m_{u,v}$ とすれば同様に定式化できる．

いま，ノード v の入隣接ノード集合を $A(v) = \{u; m_{u,v} \neq 0\}$ とし，ノード v の出隣接ノード集合を $B(v) = \{w; m_{v,w} \neq 0\}$ とする．また，入隣接数が k のノード集合を $C(k) = \{v; |A(v)| = k\}$ とし，出隣接数が k のノード集合を $D(k) = \{v; |B(v)| = k\}$ とする．ここで，$|A(v)|$ などは集合 $A(v)$ の要素数を表す．まず，隣接数 k に対する入隣接数分布 $ia(k)$ と出隣接数分布 $oa(k)$ を次式で定義する．

$$ia(k) = |C(k)|, \tag{1}$$

$$oa(k) = |D(k)|. \tag{2}$$

また，入隣接数 k に対する出隣接数の相関 $ic(k)$ と出隣接数 k に対する入隣接数の相関 $ic(k)$ を次式で定義する．

$$ic(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} |B(v)|, \tag{3}$$

$$oc(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} |A(v)|. \tag{4}$$

一方，入隣接数 k のノード集合 $C(k)$ でのリンクの平均多重度 $im(k)$ と出隣接数 k のノード集合 $D(k)$ でのリンクの平均多重度 $om(k)$ を次式で定義する．

$$im(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} \frac{1}{k} \sum_{u \in A(v)} m_{u,v}, \tag{5}$$

$$om(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} \frac{1}{k} \sum_{w \in B(v)} m_{v,w}. \quad (6)$$

最後に，入隣接数 $k(>1)$ のノード集合 $C(k)$ での平均ジニ係数 $ig(k)$ と出隣接数 k のノード集合 $D(k)$ での平均ジニ係数 $og(k)$ を次式で定義する．

$$ig(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} \frac{\sum_{u \in A(v)} \sum_{x \in A(v)} |m_{u,v} - m_{x,v}|}{2(k-1) \sum_{u \in A(v)} m_{u,v}}, \quad (7)$$

$$og(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} \frac{\sum_{w \in B(v)} \sum_{x \in B(v)} |m_{v,w} - m_{v,x}|}{2(k-1) \sum_{w \in B(v)} m_{v,w}}. \quad (8)$$

上述した評価は隣接数 k でなく，次数に対しても同様に定義できる．ここで，ノード v の入次数と出次数は，$\sum_{u \in A(v)} m_{u,v}$ と $\sum_{w \in B(v)} m_{v,w}$ でそれぞれ定義される．なお，我々の実験では，隣接数でも次数でもほぼ同様な分析結果を得ているので，例えば，式 (7) や (8) のジニ係数は，隣接数 $k > 1$ を条件に簡潔に表現できるので，本稿では隣接数を土台とした分析結果について報告する．

さらに本分析では，Favorites や Mentions などの実ネットワークの特性を調べるために，これらを入リンク数や出リンク数を固定する条件で，リンクのペアを任意に入れ替えて構築するランダムグラフを考え，上述した指標での評価を試みる．なお，このようなリンク数不変での張り替えは，Newman の単純グラフでのランダムな張り替え法 [4] を，有向多重グラフに拡張したものである．

3. 評価実験

分析に用いたデータと実験結果，それに基づく考察を述べる．

3.1 ネットワークデータ

本論文で取り扱うデータは，Twitter の Favorites 機能と Mentions 機能を多重有向グラフ化したものである．両ネットワークの統計量を表 1 に示す．

3.1.1 Favorites ネットワーク

Twitter には自分や他人のツイートをお気に入り登録し，いつでも閲覧できる Favorites 機能が装備されている．類似機能としては，facebook [*2]のいいね！機能，はてなブックマーク [*3]などが挙げられる．この機能のユーザーをノード,「ふぁぼる・ふぁぼられる」の関係をリンクとしてネットワークを構築した．

データの取得期間は 2011 年 5 月 1 日から 2012 年 2 月 12 日までである．このネットワークは 189,717 ノードと 33,456,690 リンクをもつ有向ネットワークで構成される．また，自己リンクは除外している．

Favorites ネットワークのデータは Favotter [*4]の「今日の人気」ページから取得したものである．Favotter の仕様上，取得できるツイートは 5 人以上にふぁぼられたものに限られる．また，Favotter 上の表示でスクリーンネームに 2 バイト文字を有するノードが存在するが，このようなノードは，データを取得・分析するプログラムを作成する上で除外した．

3.1.2 Mentions ネットワーク

Twitter において,“ @Screen_name ”のように半角アットマークとそれに続くスクリーンネームを含むツイートを@-message, あるいは Mentions と呼ぶ．この機能のユーザーをノード,「@-message の送信・受信」の関係をリンクとしてネットワークを構築した．

データの取得期間は 2011 年 3 月 7 日から同月 23 日までである．このネットワークは 4,731,019 ノードと 193,913,339 リンクをもつ有向ネットワークで構成される．

Mentions ネットワークのデータは鳥海ら [5] のデータから“ @Screen_name ”を含むツイートを抽出したものである．

3.2 実験結果

本節では， 2. 節で説明した分析法を用い Favorites ネットワークと Mentions ネットワークを分析した結果とその考察を記す．

なお，以降の図では，入隣接数を横軸にとった場合は実際のプロット結果を青色，同規模ランダムネットワークでのプロット結果を緑色で示し，出隣接数を横軸にとった場合は実際のプロット結果を赤色，同規模ランダムネットワークでのプロット結果を緑色で示す．また，いずれも横軸は対数表示である．

3.2.1 隣接数分布

図 1a，図 1b，図 1c，図 1d には，式 (1) と (2) で示した，両ネットワークにおけるノード数の対数表示を縦軸にとった際の入隣接数と出隣接数の分布をそれぞれ示す．

図 1a，図 1b，図 1c，図 1d より，両ネットワークは，入出隣接数のいずれについてもスケールフリー性を有することがわかる．

また，Mentions ネットワークにおいて，ランダム張替後のプロット結果は右にシフトしているが，Favorites ネットワークにおいては，ランダム張替後もプロット結果はさほどシフトしない事がわかる．

この理由について 3.3 で考察する．

3.2.2 入隣接数相関・出隣接数相関

図 2a と図 2b には，式 (3) で示した，出隣接数の対数表示を縦軸にとった際の入隣接数の分布を示す．図 2c と図 2d には，式 (4) で示した，入隣接の対数表示を縦軸にとった際の出隣接数の分布を示す．

図 2a，図 2b，図 2c，図 2d より，両ネットワークは，入出隣接数のいずれについてもランダムネットワーク同様に正の相関がみられる．

この理由について 3.3 で考察する．

3.2.3 多重度平均

図 3a，図 3b，図 3c，図 3d には，式 (5) と (6) で示した，多重度平均の対数表示を縦軸にとった際の入隣接数，出隣接数の分布を示す．

図 3a，図 3b，図 3c，図 3d より，入出隣接数いずれについても，Mentions ネットワークでは，ある隣接数を境に平均多重度との相関が正から負の傾向へ遷移するが，Favorite ネットワークでは，多少のばらつきがあるにせよ正の相関を保つことがわかる．

また，Mentions ネットワークにおいて，ランダム張替後の平均多重度はいずれも小さく横ばいであるが，Favorites ネットワークにおいては，ランダム張替後も隣接数と平均多重度は正の相関関係にあることがわかる．

これらの理由について 3.3 で考察する．

∗2 http://www.facebook.com/
∗3 http://b.hatena.ne.jp/
∗4 http://favotter.net/

	ノード数	リンク数		平均隣接数	平均多重度	平均ジニ係数
Favorites	189,717	33,456,690	入リンク 出リンク	37.30329912	0.04740342 0.059854251	0.004645323 0.004338254
Mentions	4,731,019	193,913,339	入リンク 出リンク	10.77119707	0.000688349 0.001260317	0.000114739 0.000185724

表 1: ネットワーク全体の統計量

3.2.4 ジニ係数

図 4a，図 4b，図 4c，図 4d には，式 (7) と (8) で示した，ジニ係数の対数表示を縦軸にとった際の入隣接数と出隣接数の分布を示す．

図 4a，図 4b，図 4c，図 4d より，3.2.3 同様，入出隣接数いずれについても，Mentions ネットワークでは，ある隣接数を境にジニ係数との相関が正から負の傾向へ遷移するが，Favorites ネットワークでは，多少のばらつきがあるにせよ正の相関を保つことがわかる．

また，ランダム張替後のネットワークを比較すると，入出隣接数いずれについても，Favorites ネットワークは Mentions ネットワークよりも正の相関が強く出ることがわかる．

これらの理由について 3.3 で考察する．

3.3 分析結果の考察

3.2.1 より，先行論文 [1] にて確認した次数分布でのスケールフリー性を，隣接数分布においても確認した．従って，入隣接数においては，多くの支持を得られるユーザーは限られた数しか存在せず，一般的なユーザーが得られるふぁぼられ数は微々たるものであることが予想される．出隣接数においては，Favorite 機能に対する思い入れや熱心さがユーザーによって不均質であり，1 日にいくつもふぁぼるユーザーも少数存在するが，一般的なユーザーは一ヶ月間でも数える程度しか行わないことが予想される．

3.2.2 より，入隣接数と出隣接数の間に正の相関関係を確認した．これは，先行論文 [1] にて確認した，相互リンクを含むモチーフが統計的有意に頻出するという性質が影響していると考えられる．

3.2.3 より，Favorites ネットワークと Mentions ネットワークで，多重度との相関関係に差異を確認した．これは，Mentions ネットワークがユーザー間のコミュニケーションから成るネットワークであるため，隣接数が高くなると@-message の送受信に偏りが出にくいが，Favorites ネットワークはユーザーの嗜好を反映したネットワークであり，コミュニケーション的側面を持たないため，同一ユーザーに何度もふぁぼりやすく，そのため高隣接数においても多重度が大きいと考えられる．

3.2.4 より，Favorites ネットワークにおける入出隣接数とジニ係数は正の相関関係にあることを確認した．これは，ふぁぼり数が高いユーザーは特定ユーザーあるいは特定ユーザー群に偏ってふぁぼる傾向があり，ふぁぼられ数が高いユーザーは特定ユーザーあるいは特定ユーザー群に偏ってふぁぼられる傾向がある，ということを示唆しており，従って熱狂的ファンの存在と，高入隣接数ユーザーは熱狂的ファンの存在によってその隣接数を獲得していることを明らかにしている．

また，一見同じようなプロット結果だが，3.2.3 と 3.2.4 を合わせて考察すると，平均多重度が大きいほど特定ユーザー群に偏ったふぁぼり・ふぁぼられをしていることが確認でき，従って本研究におけるジニ係数の応用は意義のある結果となった．

また，3.2.1，3.2.3，3.2.4 より，ランダム張替後の Favorites ネットワークにおいて，隣接数分布は右にシフトしづらく，平均多重度とジニ係数が入隣接数出隣接数いずれも正の相関関係にあることがわかるが，これは，ランダム張替後も特定のユーザーに偏ってリンクが生成されることを示唆しており，従って Favorites ネットワークはリンクに偏りが生まれやすい構造を持つことがわかる．

4. おわりに

新たなユーザー嗜好モデルの構築への基盤として，Favorites ネットワークを多重有向グラフとして捉え，格差を評価するジニ係数などを用いて分析し，Mentions ネットワークと結果を比較し観察した．

その結果，隣接数分布におけるスケールフリー性を，入隣接数と出隣接数における正の相関関係を，平均多重度プロットにより高入出隣接数における平均多重度が大きいことを，ジニ係数プロットにより熱狂的ファンユーザーの存在を，平均多重度が大きいほどジニ係数も高くなることを，各ランダムネットワークの比較により Favorites ネットワークの特定ユーザーにリンクが偏りやすいという構造を，それぞれ確認した．

今後は，さらに多様な分析法を実施し，Favorites モデルの提案に向けて研究を進めていく予定である．

謝辞

本研究は，科学研究費補助金基盤研究 (C) (No. 22500133) の補助を受けた．

参考文献

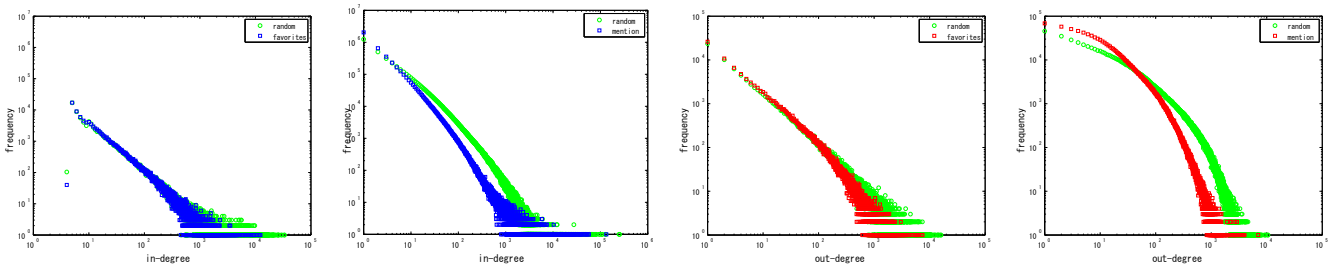
[1] 加藤翔子，伏見卓恭，斉藤和巳, “Twitter の Favorite 機能を用いたユーザー嗜好分析”，第 4 回データ工学と情報マネジメントに関するフォーラム，2012.

[2] B.A.Huberman, D.M.Romero and F.Wu, “Social networks that matter: Twitter under the microscope”, First Monday, Volume 14. Number 1. January 5 2009.

[3] H.Kwak, C.Lee, H.Park, and S.Moon, “What is Twitter, a social network or a news media?” In Proceedings of the 19th international conference on World wide web, pp.591-600. ACM, 2010.

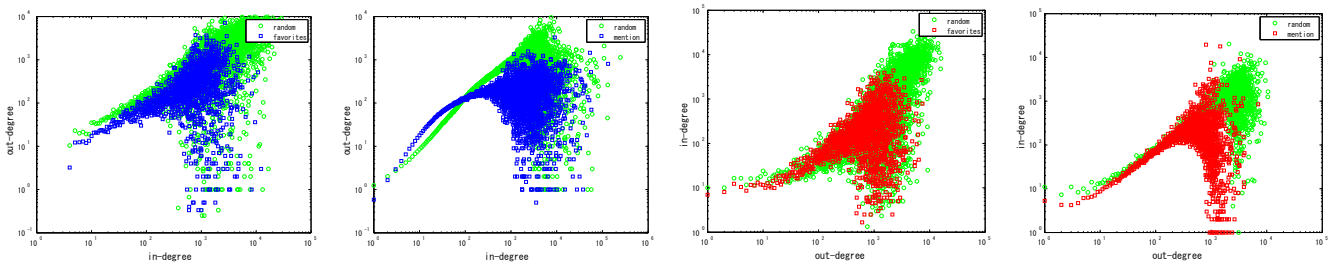
[4] M. E. J. Newman,“The structure and function of complex networks” , SIAM Review, Vol.45, pp.167–256, 2003.

[5] 鳥海不二夫，篠田孝祐，栗原聡，榊剛史，風間一洋，野田五十樹，“震災がもたらしたソーシャルメディアの変化”，第 7 回ネットワークが創発する知能研究会，2011.



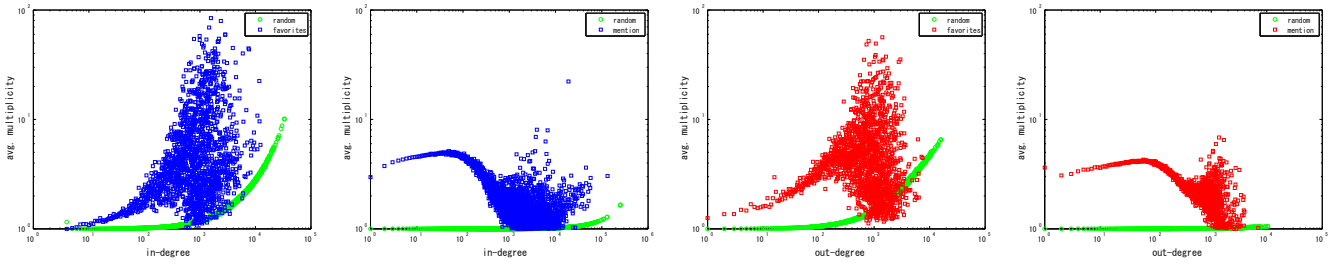
(a) Favorites 入隣接数. (b) Mentions 入隣接数. (c) Favorites 出隣接数. (d) Mentions 出隣接数.

図 1: ネットワークの隣接数分布比較



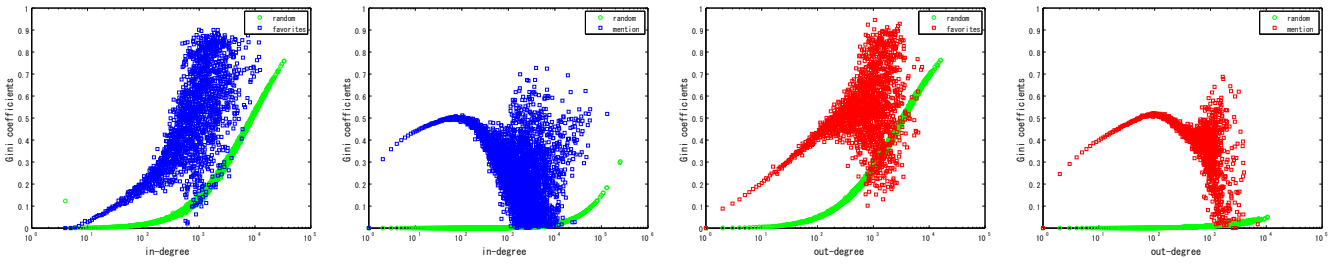
(a) Favorites 入出隣接数. (b) Mentions 入出隣接数. (c) Favorites 出入隣接数. (d) Mentions 出入隣接数.

図 2: ネットワークの入隣接数と出隣接数の相関



(a) Favorites 入隣接数. (b) Mentions 入隣接数. (c) Favorites 出隣接数. (d) Mentions 出隣接数.

図 3: ネットワークの多重度平均比較



(a) Favorites 入隣接数. (b) Mentions 入隣接数. (c) Favorites 出隣接数. (d) Mentions 出隣接数.

図 4: ネットワークのジニ係数比較