誘導サブグラフによる重要ノード間の関係分析

Analysis of the relationship between the important nodes by induced subgraph

周萱[*1] 斉藤和巳[*1] 木村昌弘[*2] 元田浩[*3]
Xuan Zhou Kazumi Saito Masahiro Kimura Hiroshi Motoda

[*1]静岡県立大学 University of Shizuoka
[*2]龍谷大学 Ryukoku University
[*3]大阪大学 Osaka University

Social networks mediate the spread of various information including topics, ideas and innovations, and finding influential nodes is one of the most central problems in social network analysis. In this paper, we focus on widely-used fundamental probabilistic models of information diffusion through networks, such as the IC (Independent Cascade) model or the LT (Linear Threshold) model, and propose a method for analyzing the relationship between such influential nodes. More specifically, this method calculates the number of connected components of each induced subgraph constructed by selecting the top-k influential nodes or those easily influenced by other nodes, where the parameter k is changed from 1 to the number of nodes. In our experiments using large real networks with only colinks, we demonstrate that the IC model and the LT model show different characteristics, when we consider the top-k influential nodes or those easily influenced by other nodes.

1. はじめに

新商品の効果的な宣伝には，社会ネットワーク上での口コミ（word-of-mouth）による情報拡散は重要と考えられている．ここで社会ネットワークとは，例えばノードを人とし，その友人関係をリンクとして繋いだネットワークである．一方，このような情報拡散に対する基本的な確率モデルとして独立カスケードモデル（Independent Cascade Model）や線形閾値モデル（linear threshold model）などが広く研究されている [Kempe 03, Kimura 07]．このようなモデルを適用すれば，情報拡散影響度の高いノード（人物）を求めることができる．このとき，影響度が高い人達が互いに近い友人同士であるか，それとも殆ど友人関係にないかを調べることは，情報拡散問題において基本的なタスクの一つと考えるられる．

本稿では，このような影響度の高い重要ノード間の関係を調べる方法論として，ノードの重要度を土台に，誘導サブグラフの結合成分数に着目した分析法を提案する．実世界ネットワークの 2 つのデータセットを用いた実験結果では，情報が伝わる確率 p を変化させ，影響度における分析を行うとともに，情報を受け取りやすい被影響度についても同様な分析を行う．

2. 情報伝播の基本モデル

本稿では，有向グラフ $G = (V, E)$ により表現される社会ネットワーク上において，ある種の情報が伝播し広がっていく現象を論じる．その情報が伝わり，かつその情報を受け入れたノードを，“アクティブノード” と呼び，そうでないノードを “非アクティブノード” と呼ぶことにする．グラフ G 内のノードの総数を N とし，リンクの総数を L とする．ノード $v \in V$ の親ノード全体の集合を $\Gamma(v)$ とする．

Kempe らの研究 [Kempe 03] に従い，G 上の IC モデルと LT モデルを定義する．これらのモデルでは，情報拡散過程は離散時間 $t \geq 0$ で展開していく．また，ノードは非アクティブからアクティブに変化することはできるが，アクティブから非アクティブへは変化できないと仮定する．アクティブノードの初期集合 A が与えられたとき，A に属するノードは時刻 0 で初めてアクティブになったと見なし，その他のノードは時刻 0 では非アクティブであると見なす．

連絡先: 斉藤和巳，静岡県立大学経営情報学部，〒 422-8526 静岡市駿河区谷田 52 番 1 号，054-264-5436，k-saito@u-shizuoka-ken.ac.jp

2.1 Independent Cascade モデル

まず，IC モデルを定義する．本モデルでは，各有向リンク (u, v) に対して，実数値 $p_{u,v} \in [0, 1]$ を前もって指定しなければならない．ここに，$p_{u,v}$ はリンク (u, v) を通しての “伝播確率” と呼ばれる．本モデルの情報伝播過程は，アクティブノードの初期集合 A が与えられたとき，次のように進んでいく．ノード u は，時刻 t で初めてアクティブなったとしよう．このとき，u はその未だ非アクティブである子ノード v をアクティブにする唯一のチャンスを与えられ，その試行は確率 $p_{u,v}$ で成功する．そしてもし u が成功したならば，v は時刻 $t+1$ でアクティブとなる．ところで，v の複数の親ノードが時刻 t で初めてアクティブになった場合は，それらの親ノードが v をアクティブにする試行は任意の順序で順々に行われることになるが，これらの試行はすべて時刻 t で行われる．u が時刻 t で v をアクティブにするのに成功したか否かにかかわらず，時刻 $t+1$ 以降では，u はもはや v をアクティブにする試行を行うことはできない．非アクティブノードをアクティブにする新たな試行が不可能になったとき，本情報伝播過程は終了する．

2.2 Linear Thresholod モデル

次に，LT モデルを定義する．本モデルでは，任意のノード $v \in V$ に対して，その親ノード u からの重み $w_{u,v}$ (> 0) を，$\sum_{u \in \Gamma(v)} w_{u,v} \leq 1$ となるように，前もって指定しなければならない．アクティブノードの初期集合 A が与えられ，各ノード v の閾値 θ_v が区間 $[0, 1]$ から一様ランダムに選ばれたとき，本モデルの情報伝播過程は次のように決定論的に進んでいく．時刻 t で非アクティブなノード v は，時刻 t でアクティブな親ノード u から，重み $w_{u,v}$ に従って影響を受ける．$\Gamma_t(v)$ を，時刻 t でアクティブであるノード v の親ノード全体の集合とする．もし，アクティブな親ノードからの重みの合計が閾値 θ_v 以上であれば，すなわち，$\sum_{u \in \Gamma_t(v)} w_{u,v} \geq \theta_v$ であれば，v は時刻 $t+1$ でアクティブとなる．非アクティブノードをアクティブにする新たな試行が不可能になったとき，本情報伝播過

程は終了する．

閾値 θ_v は，ある情報を親ノードが受け入れたとき，ノード v がそれを受け入れる傾向をモデル化していることに注意する．ところで，実世界ネットワークにおいて各ノードの閾値を前もって知ることは，一般には困難と考えられる．したがって，LT モデルではそれら閾値をランダムに選ぶことにしている．また，ターゲット集合の影響を推定する場合には，すべてのノードに対しすべての可能な閾値において平均化することにしている．それがゆえに，LT モデルを，$[0,1]^N$ 上の一様分布に随伴した確率モデルと見なすことにする．

3. 分析方法

ネットワーク構造が有向グラフ $G=(V,E)$ として与えられれば，前述した情報拡散モデルである IC モデルや LT モデルに従い，あるノード $v \in V$ のみを情報源としたときに，何個のノードに情報を拡散できるかの期待影響度を定義できる．詳細には，ノード $v \in V$ のみを情報源としたとき，別のノード $w \in V$ に情報拡散できる確率を $P_v(W=1)$ とする．ここで，W は確率変数で，ノード w がアクティブなら 1 を，非アクティブなら 0 となるものとする．なお，ノード v を情報源としたとき，あるノード w は，アクティブとなるか非アクティブのままかどちらかなので，両者の確率の和は 1 となる．すなわち，$P_v(W=0)+P_v(W=1)=1$ である．

上記の拡散確率 $P_v(W=1)$ を用いれば，ノード v の期待影響度 $\sigma(v)$ を以下で定義する．

$$\sigma(v)=\sum_{w\in V} P_v(W=1) \tag{1}$$

以下では，$\sigma(v)$ を単に影響度と呼ぶ．一方，影響度とは逆の立場で，あるノード w に対し，何個の情報源ノードからの情報拡散を受取ることができるかの期待被影響度も定義できる．このとき，ノード w の期待被影響度 $\tau(w)$ を以下で定義する．

$$\tau(w)=\sum_{v\in V} P_v(W=1) \tag{2}$$

以下では，$\tau(w)$ を単に被影響度と呼ぶ．なお，影響度や被影響度の推定には，ボンドパーコレーション法を適用することで，単純なシミュレーションでの推定と比較して，計算効率の大幅な向上が一般に実現できる [Kimura 07]．

情報拡散モデルにおいて，影響度や被影響度の高い重要ノード間の関係を調べるため，誘導サブグラフの概念を用いる．有向グラフ $G=(V,E)$ の誘導サブグラフ $G_0=(V_0,E_0)$ とは，ノード集合 V の部分集合 V_0 が与えられたとき，リンク集合を $E_0=E\cap(V_0\times V_0)$ で規定するグラフである．明らかに，与えられたノード集合 V_0 に対して，その誘導サブグラフが単一結合成分になるとは限らない．以下では，指標（影響度，または被影響度）の高い順にノード集合 V_0 を構築し，その誘導サブグラフの結合成分数を調べることで，重要ノード間の関係を調べる．具体的には，以下の手順となる．

1. 各ノードを情報源としたときの影響度と被影響度を求める．
2. 指標（影響度，被影響度）の高い順にノードをソートする．
3. $k=1$ からノード数 $|V|$ まで以下の処理を行う．
   (a) 上位 k 個のノード間に張られるリンクを抽出し，誘導サブグラフを構築する．
   (b) その誘導サブグラフが何個の結合成分になるか調べ出力する．

一般に，任意の k において，結合成分数が小さければ，重要ノード群は互いに結合していると考えられる．一方，結合成分数が大きければ，重要ノード群は互いに離れて配置していると想定できる．

4. 評価実験結果

評価実験においては，実社会ネットワークの顕著な特徴を多くもつ大規模ネットワークの利用が望ましいと考えられる．本稿では，そのような実世界ネットワークの 2 つのデータセットを用いた実験結果を報告する．

4.1 ブログデータ

ある種の情報は，トラックバックを通してあるブログ著者から別のブログ著者へと伝播しうると考えられるので，ブログのトラックバックネットワークを用いて評価実験を行った．トラックバックネットワークデータは，「goo ブログ」(`http://blog.goo.ne.jp/usertheme/`) の「JR 福知山線脱線事故」というテーマからトラックバックを 10 段辿ることにより，2005 年 5 月に収集した．収集したネットワークは，12,047 ノードと 53,315 リンクをもつ連結有向グラフであるが，トラックバック作成には，基本として相互承認が必要なため，すべてが双方向リンクを持つようにリンクの追加を行った．これにより有向リンクの総数は 79,920 となった．本ネットワークは，たいていの大規模な実ネットワークと同様，次数分布も所謂べき乗則に従っていた．以降，本ネットワークデータをブログデータセットと呼ぶ．

一方，IC モデルにも LT モデルにも，前もって指定すべきパラメータがある．IC モデルでは，一様な確率 p を，任意の有向リンク (u,v) に対する伝播確率 $p_{u,v}$ に割り当てた．すなわち，$p_{u,v}=p$ とした．LT モデルにおいては，重みを次のように一様に設定した．任意のノード v に対して親ノード $u \in \Gamma(v)$ からの重み $w_{u,v}$ を，$w_{u,v}=1/|\Gamma(v)|$ で与えた．

図 1 には，IC モデルでの影響度に基づく評価結果を示す．図より，拡散確率 p が 0.01 と 0.02 の間で曲線の形状が大きく変わり，相転移のような現象がみられる．詳しくは，拡散確率が 0.02 より大きければ，結合成分数は殆ど 5 個以下程度である．一方，拡散確率が 0.01 より小さければ，上位ノード数が 1,000 程度のときに，結合成分数は 40 近くなっている．

図 2 には，IC モデルでの被影響度に基づく評価結果を示す．上位ノード数が 1,000 程度のときに，拡散確率が 0.005 の方が 0.01 より，結合成分数が若干多くなったりしているが，全体としてみれば，図 1 と類似した結果となった．つまり，情報発信でも，情報受取でもグラフは非常に類似していることが見て取れる．

一方，ＬＴモデルの場合には，情報発信と情報受取で大幅な違いがある．図 3 には，LT モデルでの影響度に基づく評価結果を示す．ノード数が 1,000 程度のときに，結合成分数は 50 近くになり，ノード数が 1,000 程度から 8,000 程度の間で，結合成分数は減少傾向にある．一方，ノード数が 8,000 程度以後は，結合成分数は一つになっている．

図 4 には，LT モデルでの被影響度に基づく評価結果を示す．ノード数が 3,000 程度まで，結合成分数が大きく増加している．一方，ノード数が 3,000 程度から 10,000 の間で，結合成分数はあまり変わっていない．ところが，ノード数が 10,000 程度以後で，結合成分数は急激に減少している．すなわち，図 3 と 4 には顕著な違いが見て取れる．

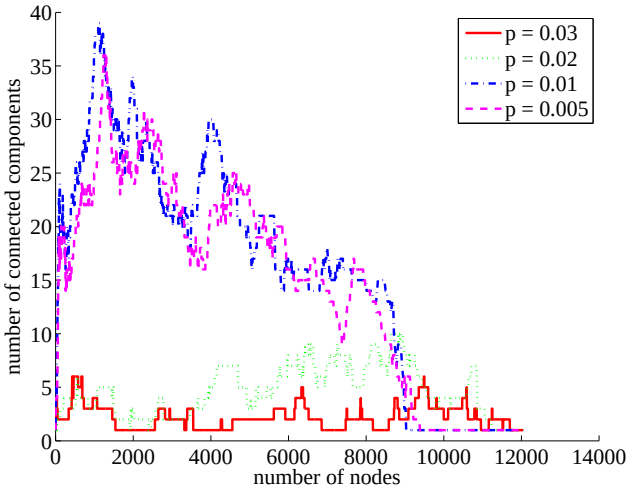

図 1: IC モデルでの影響度に基づく評価

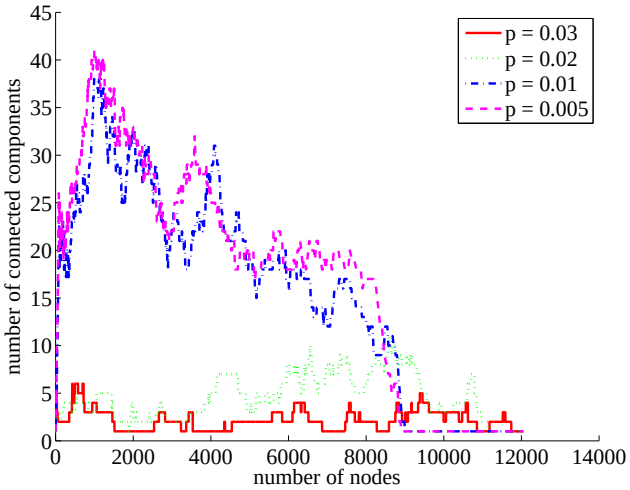

図 2: IC モデルでの被影響度に基づく評価

4.2 ウィキペディアデータ

「ウィキペディア」内の「人名一覧」から導かれる人物ネットワークを用いて評価実験を行った．具体的には,「人名一覧」に登場する人物において，ウィキペディア内の記事中に 6 回以上共起した 2 人の人物をリンクすることから得られる無向グラフの最大連結成分を抽出し，それら無向リンクを双方向リンクとみなすことにより有向グラフを構築した．以降，本ネットワークデータをウィキペディアデータセットと呼ぶ．ここに，ノード数は 9,481 であり，有向リンク数は 245,044 であった．

図 5 には，IC モデルでの影響度に基づく評価を示す．図より，ブログデータで顕著に現れた相転移のような現象は，比較的明確には現れなかったものの，拡散確率 p が 0.005 と 0.001 の間で，ある程度の曲線の形状の変化が見て取れる．詳しくは，拡散確率が 0.005 より大きければ，結合成分数は，ほとんど 4 個程度以下である．一方，拡散確率が 0.001 より小さければ，ノード数が 5,000 程度から 7,000 程度までで, 結合成分数は 10 近くになっている．

図 6 には，IC モデルでの被影響度に基づく評価を示す．ウィキペディアデータにおいても，情報発信でも，情報受取でもグラフは非常に類似していることが見て取れる．図より，拡散確率が 0.001 より小さければ，ノード数が 6,000 程度から 7,000 程度の間で，結合成分数は 12 くらいになるものの，全体としてみれば，図 5 とかなり類似してると言える．

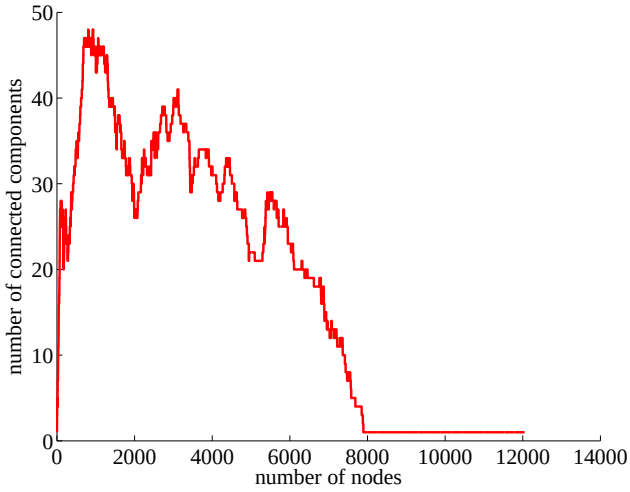

図 3: LT モデルでの影響度に基づく評価

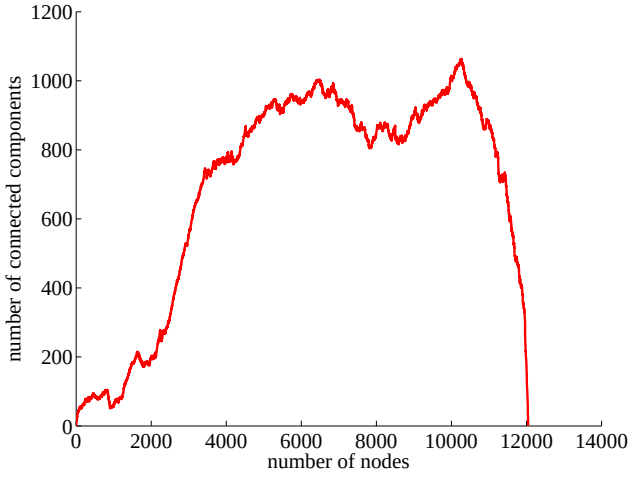

図 4: LT モデルでの被影響度に基づく評価

一方，ウィキペディアデータにおいても，LT モデルの場合には，情報発信と情報受取で大幅な違いがある．図 7 には，LT モデルでの影響度に基づく評価を示す．図より，ノード数が 6,000 程度前後において，結合成分数は 12 近くとなり，ノード数が 7,000 程度以降では，結合成分数は一つになっている．

図 8 には，LT モデルの被影響度に基づく評価を示す．図より，結合成分数が極大となる幾つかのピーが存在することが分かる．特に，ノード数が 6,000 程度時に，結合成分数は 350 程度で最大となっている．すなわち，図 7 と 8 には顕著な違いが見て取れる．したがって，全てのリンクが双方向リンクから構成されるネットワークでは，IC モデルの影響度と被影響度のグラフはほとんど変化しないが，LT モデルでは顕著な違いが起こることが示唆される．

4.3 考察

IC モデルと LT モデルにおいて影響度と被影響度で違いの起こる直感的な理由について，ボンドパーコレーション過程の

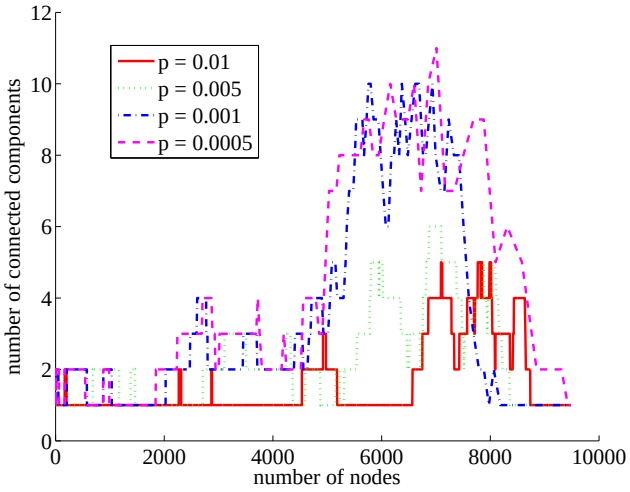

図 5: IC モデルでの影響度に基づく評価

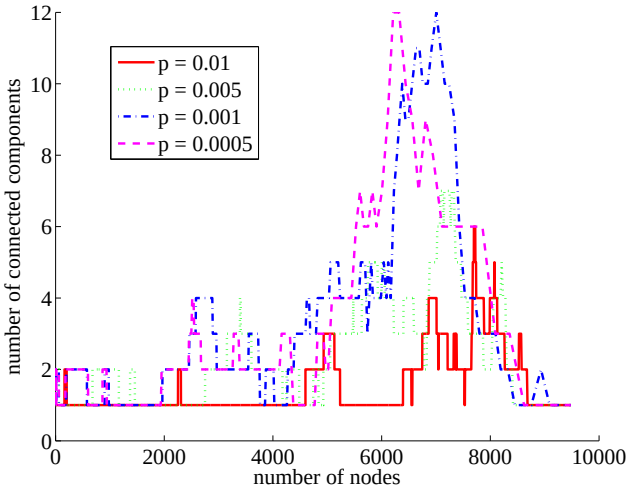

図 6: IC モデルでの被影響度に基づく評価

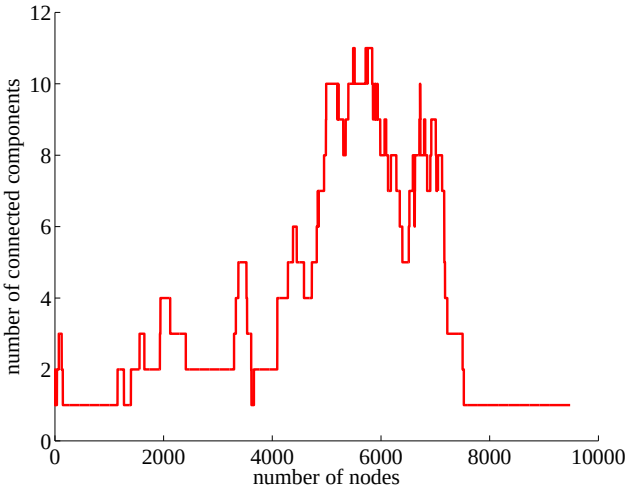

図 7: LT モデルでの影響度に基づく評価

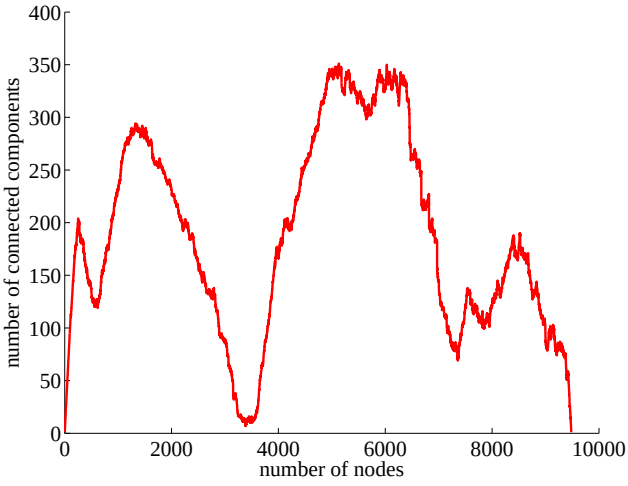

図 8: LT モデルでの被影響度に基づく評価

観点より考察する．いま，ネットワークのリンクが全てが双方向ならば，IC モデルではそれぞれに同じ確率で独立な試行を行うので，任意の 2 つのノードにおいて情報拡散の上流となるか下流となるか同じ確率になると考えられる．

一方，LT モデルではボンドパーコレーション過程において入リンクを一つ選ぶため [Kempe 03]，全てが双方向リンクでもアンバランスな状況になる．例えば，ノード v の次数が 1 で，ノード w の次数を 100 とし，これらが直接リンクしているとする．ノード v の次数は 1 なのでリンク (w, v) は確実に選ばれる．これに対して，リンク (v, w) が選ばれるのは，1/100 の確率であり，情報拡散において，上流となりやすいか下流となりやすいかでアンバランスが起きる．

5. おわりに

本稿では，影響度や被影響度の高い重要ノード間の関係を調べる方法論として，ノードの重要度を土台に，誘導サブグラフの結合成分数に着目した分析法を提案した．実世界ネットワークの二つのデータセット（ブログデータ，ウィキペディアデータ）を用いた実験では，全てのリンクが双方向リンクから構成されるネットワークにおいては，情報伝播影響度と被影響度に基づいた分析を行ったところ，情報発信でも，情報受取でも IC モデルでは同様な性質を示すのに対して，LT モデルの場合には，情報発信と情報受取で大幅に異なる性質を持つことが示唆される．このような違いを検出できることは，我々の提案する分析手法の有用性も示唆していると考えられる．今後の検討課題としては，多様な構造の有向ネットワークを用いた実証実験などをさらに進める予定である．

参考文献

[Kempe 03] D. Kempe, J. Kleinberg, and E. Tardos: Maximizing the spread of influence through a social network, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003).

[Kimura 07] M. Kimura, K. Saito, and R. Nakano: Extracting influential nodes for information diffusion on a social network. Proceedings of the 22nd AAAI Conference on Artificial Intelligence, pp. 1371–1376 (2007).