3E3-4

ソーシャルネットワーク上での影響を最大化するターゲットノード

Which Targets to Contact First to Maximize Influence over Social Network

斉藤 和巳 *1 Kazumi Saito　木村 昌弘 *2 Masahiro Kimura　大原 剛三 *3 Kouzou Ohara　元田 浩 *4 Hiroshi Motoda

*1静岡県立大学 University of Shizuoka　*2龍谷大学 Ryukoku University　*3青山学院大学 Aoyama Gakuin University　*4大阪大学 Osaka University

We address a new type of influence maximization problem which we call "target selection problem". This is different from the traditionally thought influence maximization problem, which can be called "source selection problem", where the problem is to find a set of K nodes that together maximizes their influence over a social network. The very basic assumption there is that all these K nodes can be the source nodes, i.e. can be activated. In "target selection problem" we maximize the influence of a new user as a source node by selecting K nodes in the network and adding a link to each of them. We show that this is the generalization of "source selection problem" and also satisfies the submodularity. The selected nodes are substantially different from those of "source selection problem" and use of the solution of "source selection problem" results in a very poor performance.

1. はじめに

Facebook, Digg, Twitter などのソーシャルメディアの出現により, 大規模なソーシャルネットワークが構築され, 情報, アイデア, 影響などの拡散に重要な役割を果たすようになり, 口コミやバイラルマーケティングの結果として大規模な拡散も起こる. このような現象は社会学, 心理学, 経済学, 計算機科学など多様な分野の研究者に注目されている [Kleinberg 08].

大規模なソーシャルネットワーク上での情報拡散に関して多くの多様な研究が展開されている [Newman 02, Leskovec 06, Bakshy 11]. 特に, 情報拡散を最大化するために, 情報源ノード集合の選択 [Kempe 03] やネットワークへのリンク追加 [Richardson 09, Sheldon 10] などの最適化問題に関して注目を集めている. これら研究で幅広く採用される情報拡散モデルは, IC (independent cascade) モデル [Kempe 03] や LT (linear threshold) モデル [Watts 07], または, これらの拡張版である [Kimura 09, Saito 09a, Gruhl 04, Saito 09b].

本稿では, 新たなタイプの影響最大化問題を扱う. 従来, この問題は, 与えられた情報拡散モデルとソーシャルネットワークに対し, 影響度を最大にする K 個のノード集合を求める問題として定義された. このとき, これら K 個のノード集合は確実に情報源ノードとして機能することが暗黙に仮定される. しかるに, このような仮定が常に妥当とは限らない. 例えば, twitter 上で自分のアイデアを広げることを考えれば, まず適当なフォロワーを獲得する必要があるが, これらファロワーが常に情報拡散に貢献するとは限らない. すなわち我々は, 新たなユーザ (ノード) から K 個のノード集合のそれぞれにリンクを追加し, この新ノードを情報源としたときの影響最大化問題を考える. 以降では, この問題をターゲット選択問題と呼び, IC と LT モデルでの特性を探求する.

ターゲット選択問題も, 従来の情報源選択問題と同様に, K 個のノード集合求めるので, NP-困難な組合せ最適化問題となる. ただし, 第 3. で示すように, その目的機関数はサブモジュラ関数となるので, 貪欲法で求める解の品質は最適解の 63% 程度となることを保証できる. さらに, 従来の情報源選択問題を効率良くために開発した手法を利用することができる. 具体的には, ボンドパーコレーション [Kimura 07], プルーニング [Kimura 09], 遅延評価 [Leskovec 07, Goyal 11], バーンアウト [Saito 09a], ヒューリスティックス [Chen 10a, Chen 10b], 信念伝搬 [Nguyen 12], 線形システム近似 [Yang 12] などである.

連絡先: 斉藤 和巳, 静岡県立大学, 静岡市駿河区谷田 52-1, TEL&FAX: 054-264-5436, k-saito[at]u-shizuoka-ken.ac.jp

2. 情報拡散モデル

有向グラフ $G = (V, E)$ で表現されるネットワークを考える. ここで, V と E ($\subset V \times V$) はノード集合とリンク集合を表す. 以下では, 文献 [Kempe 03, Kimura 10] に従い, 情報源ノード集合から時刻 $t \geq 0$ で情報拡散する IC と LT モデルを定義する. なお, 各ノードが情報受信に成功した状態をアクティブと呼び, 情報を受けてない非アクティブな状態には戻れない SIR (Susceptible Infected Recovered) 設定とする.

IC モデルではパラメータとして各リンク (u, v) に拡散確率 $p_{u,v}$ を考える ($0 < p_{u,v} < 1$). いま, ノード u が時刻 t でアクティブになれば, u はその非アクティブな子ノード v を確率 $p_{u,v}$ でアクティブにするチャンスが一度だけ与えられる. u からの情報拡散が成功すれば, v は時刻 $t+1$ でアクティブとなる. もし v の複数の親ノードが時刻 t で同時にアクティブとなれば, このような情報拡散施行は任意の順番で施されるとする. 時刻 $t+1$ で新たにアクティブとなるノードなければ拡散は終了する.

LT モデルではパラメータとして各リンク (u, v) に重み $q_{u,v}$ を考える ($q_{u,v} > 0$). ただし, ノード v の親ノード集合 $B(v) = \{u \in V; (u, v) \in E\}$ に対し, $\sum_{u \in B(v)} q_{u,v} \leq 1$ の制約を持つ. また情報拡散に先立ち, 任意のノード $v \in V$ に閾値 θ_v が $[0, 1]$ での一様分布に従い付与される. 非アクティブなノード v はアクティブな親ノードの影響を受け, 時刻 t にて親ノードの影響重みの和が閾値 θ_v を初めて超えれば, つまり, $\sum_{u \in B_t(v)} q_{u,v} \geq \theta_v$ となれば, ノード v は時刻 $t+1$ でアクティブとなる. ここで, $B_t(v)$ は時刻 t 以前にアクティブな v の親ノード集合を表す. 時刻 $t+1$ で新たにアクティブとなるノードがなければ拡散は終了する.

時刻 $t = 0$ での初期情報源 (アクティブ) ノード集合 $W(\subset V)$ に対し, IC または LT の情報拡散モデルにより, 時刻 $t \geq 0$ でアクティブとなったノード数を $\varphi(W; G)$ とする. $\varphi(W; G)$ は確率変数となるので, その期待値 $\sigma(W; G)$ を定義でき, 以下では, $\sigma(W; G)$ を W のネットワーク G での影響度と呼ぶ.

3. ターゲット選択問題

まず, 影響最大化問題 [Kempe 03, Leskovec 07, Kimura 10, Chen 10b, Chen 10a] とも呼ばれる従来の情報源選択問題について述べる. 与えられたネットワーク $G = (V, E)$ と定数 K に対して, 影響度 $\sigma(W_K; G)$ を最大化する K 個のノード集合 $W_K(\subset V)$ を求める問題であり, 次式の最大化問題として定式化される.

$$\operatorname*{argmax}_{W_K \subset V} \sigma(W_K; G). \tag{1}$$

一方, 本論文で提案するターゲット選択問題では, ネットワーク $G = (V, E)$ と定数 K だけでなく, 外部情報源ノード $x \notin V$ と各リンク (x, v) に対して値 $\{r_{x,v} \mid v \in V\}$ が与えられる. $r_{x,v} \in [0, 1]$ は IC モデルなら拡散確率 $p_{x,v}$ に, LT モデルなら重み $q_{x,v}$ に対応する. つまり, K 個のノード集合 $W_K(\subset V)$ の選択により, ノード x から $w \in W_K$ への K 本のリンクを G に追加して拡張したネットワーク $G'(W_K)$ において, ノード x の影響度を最大化する問題であり, 次式の最大化問題として定式化される.

$$\operatorname*{argmax}_{W_K \subset V} f(W_K) = \operatorname*{argmax}_{W_K \subset V} \sigma(\{x\}; G'(W_K)), \tag{2}$$

ここで $G'(W_K) = (V \cup \{x\}, E \cup \{(x, w) | w \in W_K\})$ である. なお, LT モデルでは, ターゲットノード $w \in W_K$ の $v \in B(w)$ からの重み $q_{v,w}$ は, モデルの重み制約により, $(1 - r_{x,w})q_{v,w}$ に弱まるとする. このとき, ターゲット選択問題は情報源選択問題の自然な拡張となる. 何故なら, 各 $w \in W_K$ で $r_{x,w} = 1$ と設定すれば, ターゲットは確実にアクティブとなり, $\operatorname{argmax}_{W_K \subset V} f(W_K) = \operatorname{argmax}_{W_K \subset V} \sigma(W_K; G)$ となるからである.

影響度 σ はサブモジュラ関数となる [Kempe 03] ため, $\sigma(W' \cup \{v\}; G) - \sigma(W'; G) \geq \sigma(W \cup \{v\}; G) - \sigma(W; G)$ if $W' \subseteq W$ が成り立ち, $W_0 = \emptyset$ と初期化し, 既に選定したノード集合 W_{k-1} に $\sigma(W_{k-1} \cup \{v\}; G)$ を最大化するノード v を追加して W_k を求める再帰的な貪欲戦略により妥当な精度の近似解が求まる. ターゲット選択問題においても, 次式の関係より, f はサブモジュラ関数となることが分かる.

$$f(W_K) = \sum_{A \in 2^{W_K}} \sigma(A; G) \prod_{w \in A} r_{x,w} \prod_{w \in (W_K \setminus A)} (1 - r_{x,w}), \tag{3}$$

ここで 2^{W_K} は W_K のべき集合を表す. $r_{x,v} \in [0, 1]$ は IC モデルなら拡散確率 $p_{x,v}$, LT モデルなら重み $q_{x,v}$ に対応するので, 式 (3) は, 情報源ノード x からの可能な成功パターン A とそれが起こる確率を考慮して影響度を求めている. LT モデルでは, 元の重み $q_{v,w}$ は, モデルの重み制約により, $(1 - r_{x,w})q_{v,w}$ に弱めるので, 情報源ノード x がターゲットノード w への情報拡散が失敗したという条件の下で, ノード $v \in B(w)$ から w への情報拡散が $q_{v,w}$ に等しくなることに注意されたい. 式 (3) より, サブモジュラ関数 σ の線形和となるので, $f(W_K)$ サブモジュラ関数となる. この性質より, 情報源選択問題と同様に, 貪欲法でターゲット選択問題の妥当な近似解は, ボンドパーコレーション [Kimura 07], プルーニング [Kimura 09], バーンアウト [Saito 09a] などを適用すれば高速に結果が得られる.

4. 実験

大規模な現実のネットワークを用い, 影響度を尺度として, $G = (V, E)$ 上でのターゲット選択問題の性質などを実験評価する. 以下では, LT モデルでの実験結果のみを示す. 何故なら, LT モデルの方がターゲット選択において, より自然な問題設定と考えられるためである.

4.1 データと実験設定

実験には, 有向グラフとして表現される現実の 4 種のネットワークを用いた. 第 1 のネットワークは, Ameblo と呼ぶ日本のブログサイト “Ameaba” [*1] (詳細は [Fushimi 10] を参照) から収集した読者ネットワークである. Ameblo ネットワークは, ノード数が 56, 604, リンク数が 734, 737 である. 第 2 のネットワークは, Blog と呼ぶ文献 [Kimura 10] で使われたトラックバックのネットワークである. Blog ネットワークは, ノード数が 12, 047, リンク数が 53, 315 である. 第 3 のネットワークは, Cosme と呼ぶ日本のコスメ商品に関する口コミでのコミュニケーションサイト [*2] (詳細は [Ohara 12] を参照) から収集したファン関係ネットワークである. Cosme ネットワークは, ノード数が 45, 024, リンク数が 351, 299 である. 第 4 のネットワークは, Enron と呼ぶ Enron Email Dataset [Klimt 04] (詳細は [Ohara 12] を参照) から収集した電子メールの送受信ネットワークである. Enron ネットワークは, ノード数が 19, 603, リンク数が 210, 950 である.

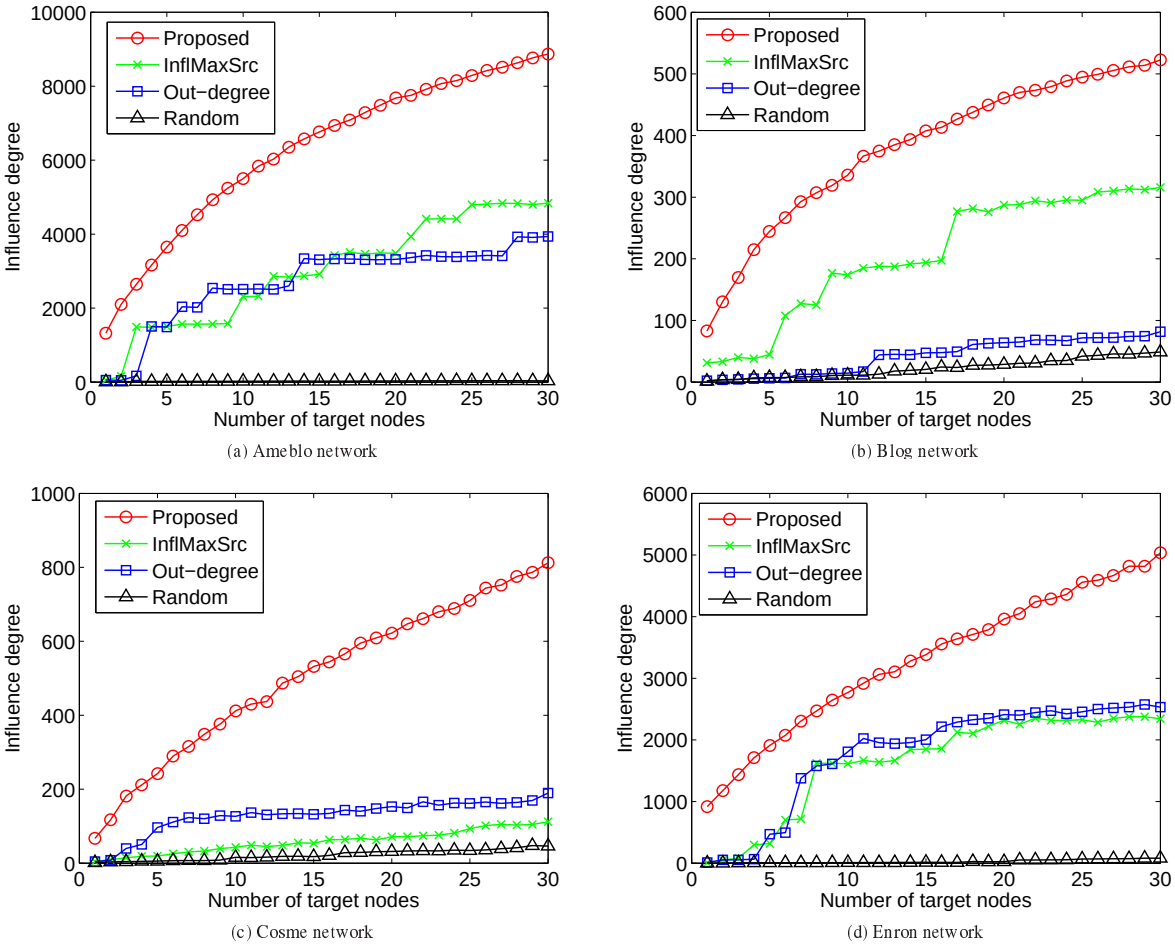
以降では, 式 (2) に基づき, 貪欲法でターゲット集合を求める提案解法を *Proposed* と呼ぶ. 次に, ターゲット選択問題の近似解法として 3 種の比較法について述べる. 第 1 の比較法は, *InflMaxSrc* と呼び, 元のネットワーク $G = (V, E)$ にて情報源選択問題の解として K ノードの集合を求める方法である. InflMaxSrc 法との比較により, ターゲット選択問題と情報源選択問題の違いを探求する. 第 2 の比較法は, *Out-degree* と呼び, 出次数の大きい順にソートして上位 K ノードの集合を求める方法である. Out-degree 法はソーシャルネットワークの中心性分析で標準的に用いられる解法である. 第 3 の比較法は, *Random* と呼び, 重複を許さず一様ランダムに K ノードの集合を求める方法である. Random 法によりベースラインの性能が分かる.

提案法と 3 種の比較法の性能評価には, $f(W_K) = \sigma(\{x\}; G'(W_K))$ を用いる. ここで, W_K は各手法で求まる K ノードの集合を意味する. なお, 影響度 $\sigma(\{x\}; G'(W_K))$ については, 独立な 10, 000 回のシミュレーションによるアクティブとなったノード数の期待値として求めた. 最後に, LT モデルでの重み設定は, $q_{u,v} = 1/B(v)$ $(\forall u, v \in V)$ とした.

4.2 実験結果

図 1a, 1b, 1c, 1d に Ameblo, Blog, Cosme, 及び, Enron の各ネットワークでの結果を示す. 図では, ターゲットノード数 k に対して, 目的関数値 f (影響度) をプロットしている. ここで, サークル, クロス, スクエア, 及び, トライアングルの各マーカーは, 順番に, 提案法, InflMaxSrc 法, Out-degree 法, 及び, Random 法を表す. 実験結果より, 提案法の性能は 4 種のネットワークすべてにおいて, 比較法となる InflMaxSrc 法, Out-degree 法, 及び, Random 法の性能に優った. 特に, InflMaxSrc 法 や Out-degree 法と比較して, 提案法により, 少なくとも 2 倍程度も影響度が大きくなることが分かる. 詳細には, Random 法は常に最も低い性能となった. Out-degree 法の性能は, ネットワーク構造の特性に大きく依存すると想定され, InflMaxSrc 法の性能に優るケースも見られる. 従来の情報源選択問題において, InflMaxSrc 法の性能が, Out-degree 法 や Random 法の性能に常に優ることと対比できる (参照 [Kimura 10]). ターゲット選択問題では, 選択されたノード集合 W_* のすべてがアクティブとなる保証はなく, このことは, 情報源ノードとして望ましいノード集合がターゲットとして望ましいとは限らないことを意味する. 実際に, $k = 30$ において, 提案法の影響度 f は InflMaxSrc 法と比較して, 1.7 から 7.2 倍程度も優っている.

*1 http://www.ameba.jp/
*2 http://www.cosme.net/



(a) Ameblo network (b) Blog network

(c) Cosme network (d) Enron network

図 1: ターゲット選択問題での性能比較.

これら結果より, 提案法や 3 種の比較法で選択するノード集合は実質的に異なることが示唆される. この点を確認するために, F-尺度 $\mathcal{F}(k) = |W_k^* \cap W_k|/k$ による類似度を評価した. ここで, k はターゲット選択問題で選択するノード数, W_k^* と W_k は提案法で選択されたノード集合と 3 種の比較法のどれかで選択されたノード集合を表す. F-尺度が最大となるのは, Amebro ネットワークで $k = 3$ としたときの InflMaxSrc 法で $\mathcal{F}(k) = 0.33$ であった. 実際に, Cosme ネットワークでは, 実験範囲のすべての k において, 3 種の比較法との F-尺度は 0 となった. よって, 提案法や 3 種の比較法で選択するノード集合は実質的に異なることが確認できたと考える.

5. おわりに

本論文では, ターゲット選択問題と呼ぶ, 新たなタイプの影響最大化問題を提案した. 従来の影響最大化問題では, 選択したノード集合は情報源として機能し確実にアクティブになると暗黙に仮定されていた. よって, この問題を情報源選択問題と呼び, 例えば, 1,000 人のユーザに新商品をプロモートするようなバイラルマーケティングの最も単純なモデルと考えた. これに対し, 少し異なる視点を当て, より自然で現実的な設定としたのがターゲット選択問題である. 今後は, 現実の情報拡散データを用いた評価などを進める予定である.

謝辞

本研究は, 科学研究費補助金基盤研究 (C) (23500194) の支援を受けた.

参考文献

[Bakshy 11] Bakshy, E., Hofman, J., Mason, W., Watts, D.: Everyone's an influencer: Quantifying influences on twitter. In: Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM2011). pp. 65–74 (2011)

[Chen 10a] Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010). pp. 1029–1038 (2010)

[Chen 10b] Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010). pp. 88–97 (2010)

[Fushimi 10] Fushimi, T., Saito, K., Kimura, M., Motoda, H., Ohara, K.: Finding relation between pagerank and voter model. In: Proceedings of the 11th International Workshop

on Knowledge Management and Acquisition for Smart Systems and Services (PKAW 2012). pp. 208–222 (2010)

[Goyal 11] Goyal, A., Lu, W., Lakshmanan, L.: Influence spread in large-scale social networks - a belief propagation approach. In: Proceedings of the 20th International World Wide Web Conference (WWW2011). pp. 47–48 (2011)

[Gruhl 04] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations 6, 43–52 (2004)

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

[Kimura 09] Kimura, M., Saito, K., Motoda, H.: Efficient estimation of influence functions fot sis model on social networks. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09) (2009)

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

[Kimura 10] Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery 20, 70–97 (2010)

[Kleinberg 08] Kleinberg, J.: The convergence of social and technological networks. Communications of ACM 51(11), 66–72 (2008)

[Klimt 04] Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). pp. 217–226 (2004)

[Leskovec 06] Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. In: Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06). pp. 228–237 (2006)

[Leskovec 07] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). pp. 420–429 (2007)

[Newman 02] Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)

[Nguyen 12] Nguyen, H., Zheng, R.: Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012). pp. 515–530. LNAI 7524 (2012)

[Ohara 12] Ohara, K., Saito, K., Kimura, M., Motoda, H.: Effect of in/out-degree correlation on influence degree of two contrasting information diffusion models. In: Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP 2012). pp. 131–138 (2012)

[Richardson 09] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002). pp. 61–70 (2002)

[Saito 09a] Saito, K., Kimura, M., Motoda, H.: Discovering influential nodes for sis models in social networks. In: Proceedings of the Twelfth International Conference of Discovery Science (DS2009). pp. 302–316. Springer, LNAI 5808 (2009)

[Saito 09b] Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proceedings of the 1st Asian Conference on Machine Learning (ACML2009). pp. 322–337. LNAI 5828 (2009)

[Sheldon 10] Sheldon, D., Dilkina, B., Elmachtoub, A., Finseth, R., Sabharwal, A., Conrad, J., Gomes, C., Shmoys, D., Allen, W., Amundsen, O., Vaughan, W.: Maximizing the spread of cascades using network design. In: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10). pp. 517–526. AUAI Press, Corvallis, Oregon (2010)

[Watts 07] Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)

[Yang 12] Yang, Y., Chen, E., Liu, Q., Xiang, B., Xu, T., Shad, S.: On approximation of real-world influence spread. In: Proceedings of the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012). pp. 548–564. LNAI 7524a (2012)