On Modelling Social Propagation Phenomenon

Dariusz Król[1,2]

[1] Smart Technology Research Centre, Faculty of Science and Technology, Bournemouth University, United Kingdom
[2] Institute of Informatics, Wrocław University of Technology, Poland
Dariusz.Krol@pwr.wroc.pl

Abstract. Propagation phenomenon is an important problem that has been studied within varied research fields and application domains, leading to the development of propagation based models and techniques in social informatics. These models are briefly surveyed in this paper. This paper discusses common features and two selected scenarios of propagation mechanisms that frequently occur in social networks. In summary, a list of the most recent open issues on social propagation is presented.

Keywords: cascading behaviour, information diffusion, spreading model.

1 Introduction

Propagation phenomena have become a pervasive and significant feature of contemporary complex networks. By studying these phenomena, we can better understand the behaviour cascading in a system. Social networks based on e-mail contacts like *Hotmail*, *Yahoo* and social networking services like *Facebook*, *Twitter*, *YouTube*, for example, are generally modelled as complex structures whose nodes represent individuals, and whose links represent interaction, collaboration, or influence between them. They are designed for information sharing and above all for information spreading. As an obvious result, we can observe various propagation effects. Here we ask, what are the practical consequences of this day-to-day existence and interaction.

Propagation mechanisms could be very useful, depending on the different types of communication networks and their purpose. For instance in online networks, there are several popular methods frequently built on 'positive' spreading mechanism such as: capturing structures [1], detecting local communities [2], predicting future network structure and user's behaviours [3], discovering rumors [4], tracking collective blogging and argumentations [5], monitoring opinions [6], recommending trust and measuring influence of authorities [7], and finally maximizing immunization of the network [8]. On the other hand, propagation in the form of gossiping [9], misinformation [10], fraudulent reputation building or targeted cyber-attacks [11] could have extremely serious 'negative' effects. What is more, the in-depth understanding of propagation phenomenon could enable early identification of social crises, and provide methods to mitigate potential catastrophes.

N.T. Nguyen et al. (Eds.): ACIIDS 2014, Part II, LNAI 8398, pp. 227–236, 2014.
© Springer International Publishing Switzerland 2014

From a scientific point of view a lot of paramount questions w.r.t. social propagation have been raised and analysed. A few survey works on propagation mainly in online social networks have already been done [12, 13]. However, important questions about how to best fit a propagation model and efficiency indicators are still relevant. Due to lack of other relevant analyses our work provides some steps toward this direction. Though we do not present a full survey, the aim of this paper is an attempt to respond to realistic features of propagation issues that the social networks confront. The results aim to aid researchers and engineers involved in the development of new technologies to better understand propagation phenomena principles, its power and limitations.

From a business perspective, that might be useful for many services to aptly adopt the propagation mechanism to make the information (action) spreading faster and wider. Accurate pushing-pulling selected information to cooperators and business partners can speed-up communication and significantly lower the load costs. On the other hand, good knowledge of propagation phenomenon could be helpful to stop diffusing the undesirable things which can threaten individuals or organizations.

There are three major parts concerning modelling of social propagation: the major principles, an appropriate model, and finally a generic algorithm and its variants. In this paper, our goal is to find answer to the following problems: what properties of social propagation are essential for efficiency and robustness, and how a propagation process can directly correlate with time and underlaying structure. To the best of our knowledge, these questions have not been addressed by the research community at large.

2 Principles of Social Propagation Mechanisms

What are the major principles associated with propagation phenomena in social informatics?

The general idea behind social propagation is that individuals interact repeatedly under the reciprocal influence of other individuals, modelled as a *social influence score*, which may often generate a flood of action across the connections. Elementary social examples include rumor dissemination, forwarding memes, joining events or groups, hashtags re-tweeting, and purchasing products. We have investigated these situations in detail, and our main observations are summarized in Table 1.

2.1 The Classical Approach to Propagation

Many propagation models, also called diffusion models [22], have been studied to date. To describe social and biological propagation, for instance, for spread of infectious diseases and computer viruses, the diffusion of innovations, political upheavals, and the dissemination of religious doctrine, the generalized contagion model was developed [23]. This non-graph based model exhibits behaviour that falls into one of two basic classes called respectively: epidemic threshold, and

Table 1. Principles of social propagation mechanisms

Topology-dependent. Propagation utilizes to the utmost the underlying structure of a network. Most social networks are characterized, among other qualities, with the *'small-world' phenomenon*, *triadic closure*, *assortative mixing* and *broad degree distributions*. Recently, more fascinating properties have been discovered, like over time *shrinking diameters*, *homophily tendency*, *aging effect*, *temporal dynamics* and the so-called *densification power law* [14], including advanced graph theoretic measures [15]. In the vast majority these features can significantly speed up the diffusion process.
Complexity. Social data are often interconnected, have overlapping communities and are coupled across time through processes. This evolving idea has been successfully implemented by *Google+ circles* and *Facebook smart lists*. When nodes in one network depend on nodes in another, a small disturbance in one network can cascade through the entire system often growing to sufficient size to collapse it [16].
Fast influence by ties. Propagation extends Granovetter's hypothesis that the majority of influence is collectively generated by weak ties, although strong ties are individually more influential [17]. Very often action spreads only partially overlapping the network, but extremely fast (in sublogarithmic time [18]).
Push-pull activity. For social propagation mechanisms, not only the topology of links, but also the communication activity is highly relevant [18]. The standard protocol is based on symmetric push-pull activity. It pushes the information in case it has, and pulls the information in case the neighbor has. A common form of positive and negative propagation is spreading of breaking news, rumors, fads, beliefs, sentiments and norms. While positive users' opinions promote an action, negative opinions suppress its adoption.
Survive or perish. Social informatics are ubiquitous and this applies also to propagation. Individuals tend to adopt the behavior of their peers, so first propagation happens locally in their neighborhood. Then, this behaviour might become global and survive or decay and finally perish as it crosses the network [19].
Epidemic-type and cascading-type. Propagation can be conceptualized either as an *epidemic-type dynamic*, where a node in an infected state may infect a neighbor independently of the status of the other nodes [20], or as a *cascading-type dynamic*, where the change may depend on the reinforcement caused by changes in other nodes [21].

critical mass. The first class is based on the idea of a threshold: an adoption depends on the fraction of neighbors which exceed a specific critical value. In the second class the population can be infected if the earliest outbreak size makes up a 'critical mass'. This contagion model can be identified with two seminal models for the spread of disease: the so-called $susceptible \rightarrow infective \rightarrow recovered$ (SIR) model and simplified $susceptible \rightarrow infective \rightarrow susceptible$ (SIS) model [24]. As a rule, all nodes in SIR model are in one of three states: susceptible (able to be infected), infected, or recovered (no longer able to infect or be infected). These three classes in specific rumor model correspond to ignorant, spreader, and stifler nodes. At each time step, nodes infected in the last time step can infect any of its neighbors who are in a susceptible state with a probability p. In this model, two transition rates $S \rightarrow I$ (the contact rate) and $I \rightarrow R$ (the rate of recovery) determine the cumulative number of infected nodes.

Many others model variations have been studied for social informatics. However, the two most widely employed models are: the Independent Cascade Model (ICM) and the Linear Threshold Model (LTM) [25]. In these graph-based models, each node is either active or inactive. An active node never becomes inactive again. The propagation process repeats until no more activations are possible. The spread represented in the ICM and LTM are very similar and proceeds in discrete steps as follows.

In the ICM, a process starts with an initial set of active nodes A_0, called later seed nodes. It is assumed that nodes can switch their states only from inactive to active, but not in the opposite direction. When a node i first becomes active in step t, it has a single chance of influencing each inactive neighbor j with probability $p(i, j)$. If node i succeeds, node j becomes active in step $t + 1$. If node j has incoming links from a few newly activated nodes, every propagation effort is randomly sequenced. When all the influence probabilities are equal to one, the ICM becomes equivalent to the deterministic model, in which every active node unconditionally activates all its neighbors.

In the LTM, each node i is described by a threshold θ_i from the range $[0, 1]$. This threshold represents the fraction of i's neighbors, denoted by $deg^{-}(i)$, that must be active in order for node i to become active. Here 'active' means that the node adopts the action and 'inactive' means otherwise. Given the thresholds and seed nodes A_0, the process unfolds deterministically in discrete steps. In step $t + 1$, all nodes that were active in step t remain active, and we activate any node j for which the total influence power of his active neighbors is at least $\theta_j : \sum p(i, j) \geqslant \theta_j$. Thus, the threshold θ_j represents the trend of a node j to adopt the action when his neighbors do.

Both models have parameter attached to each directional link, i.e., propagation probability in the ICM and weight in the LTM. In the ICM-based virus-spreading model, the probability of being infected is proportional to the number of neighbors infected. Although both models appeared to be comparative, there are important differences. Intuitively, only a small number of nodes overlap for these models. Furthermore, it is more difficult for the LTM to transmit action to hub nodes than the ICM does.

2.2 Our Generic Approach

With respect to existing propagation models and aforementioned six principles, this directs us to the following generic model definition. A propagation $\wp(\Game, G)$ with action $\Game$ over the network is described by a directed connected graph $G = (N, L)$ where N is the node set and L is the set of directed links containing connected pairs of nodes (i, j) unfolds in discrete time-steps $t \geqslant 0$ and is defined as an ordered sequence of triplets (i, j, t). Each triplet corresponds to a single interaction event at time-step t between a pair (i, j), where $i, j = 1..N$. We assume that no individual adopts the same action more than once. In Twitter, for instance, user j adopts tweet posted by i at time t and re-tweets it, then the tweet becomes available to his followers. The total number of triplets is given by $|\wp(\Game, G)|$. At the empirical level, propagation with action over the network is a set of interaction events recorded every δ (the sampling resolution) during an interval Δt. This leads us to the notion of the propagation graph. The corresponding static propagation graph $G_{\wp}^{\Delta t}$ is obtained by aggregating all the same interactions events within Δt into a graph structure with links connecting in the direction of propagation. This propagation graph is now a directed acyclic graph (DAG) which may have disconnected components.

To facilitate the propagation dynamics, particularly its transitivity, in order to merge propagations along each path, to select one out of multiple propagations, or to combine propagations from all paths, a set of primitives based on path algebra [26] that effectively operate on a given graph should be formalized. For instance, for propagating trust presented as a triple $(belief, distrust, uncertainty)$ three associative operators: *concatenation*, *selection*, and *aggregation* are analysed. The description of these operators goes beyond the scope of this paper, however, the interested reader is referred to [27] for details.

Note that our basic model fulfils the requirements of both models: ICM and LTM. Contrary to the classic models, instead of considering sets of nodes, links and probabilities separately, the proposed model focuses on time-stamped propagation triplets, leading to a reduction of the complexity in propagation scenarios. Our proposal seems to be adjusted to any action-propagation model, including variations on compartmental models in epidemiology.

3 The Propagation Algorithm

How does the propagation process normally proceed?
Propagation $\wp(\mathcal{G}, G)$ in the form of triplets shows pathways for the transmission of infection, rumor, trust, or other quantities. Common algorithms for crawling or sampling social networks include, first of all, variants of breadth-first search and depth-first search. Using only social ties and forward paths we do not necessarily crawl an entire network. We only explore the connected component reachable from the set of seed nodes. The random walk algorithm and snowball sampling [19] for directed graph are the simplest implementations for propagation process by passing action from one node to its neighbor. There are numerous variations like following a randomly selected triadic node (friend, casual friend or grandparent). Another possible implementation is the preferential walk to follow someone from whom or through whom they have adopted actions using a back-propagation mechanism. It has been found in [22] that a combined strategy with triadic closure and coincided traffic-based shortcuts yield the best accuracy.

A general propagation algorithm has initialization including seed node(s) selection, propagation loop with threshold condition, output update and termination condition(s). By applying different social-based loop and influence conditions, we can consider various strategies. The choice of the strategy is generally dependent on the communication flow schema like (1) push and (2) pull. The first one accommodates the propagation along a sender-centered approach such as ICM, the second, receiver-centered approach such as LTM, which combines the propagation from different nodes.

Recall that a threshold value is associated with each node (i.e., social entity or person). This value represents the positive or negative influence of that individual. To get a better understanding we illustrate the key parts of propagation advancement with the aid of Algorithm 1. In the algorithm, the power to influence neighbors is modelled as a propagation probability denoted by $p(i, j)$. Note that this probability p could be time-varying t and algorithm α dependent as follows $p(i, j) = \alpha^t(i, j)$. In every step t, each node i belongs to one of the three sets: *waiting*, *active* and *inactive*. For readability, we omit the *inactive* set from our algorithm. The algorithm terminates, if the current time-step t reaches the time limit T, or there are no more *waiting* nodes, which mean that no more activations are possible.

Input: graph G, threshold θ_i for node, seed nodes A_0, time limit T
Output: *triplets*

$active, triplets \leftarrow \varnothing$;
$waiting \leftarrow A_0$;
$t \leftarrow 0$;
while *termination condition not satisfied* **do**
  foreach *node* $i \in waiting$ **do**
    $t \leftarrow t + 1, waiting \leftarrow waiting \backslash \{i\}, active \leftarrow active \cup \{i\}$;
    `propagation loop` with `threshold condition`
  end
end

Algorithm 1. Generic propagation algorithm

In the algorithm, we use `propagation loop` with `threshold condition`, that determines how to select next node to activate it in the process to maximize the spread. In the push variant of propagation loop and condition (Algorithm 2), propagator one-to-many using deg^+ operator is required, e.g., sending photo to all my group members.

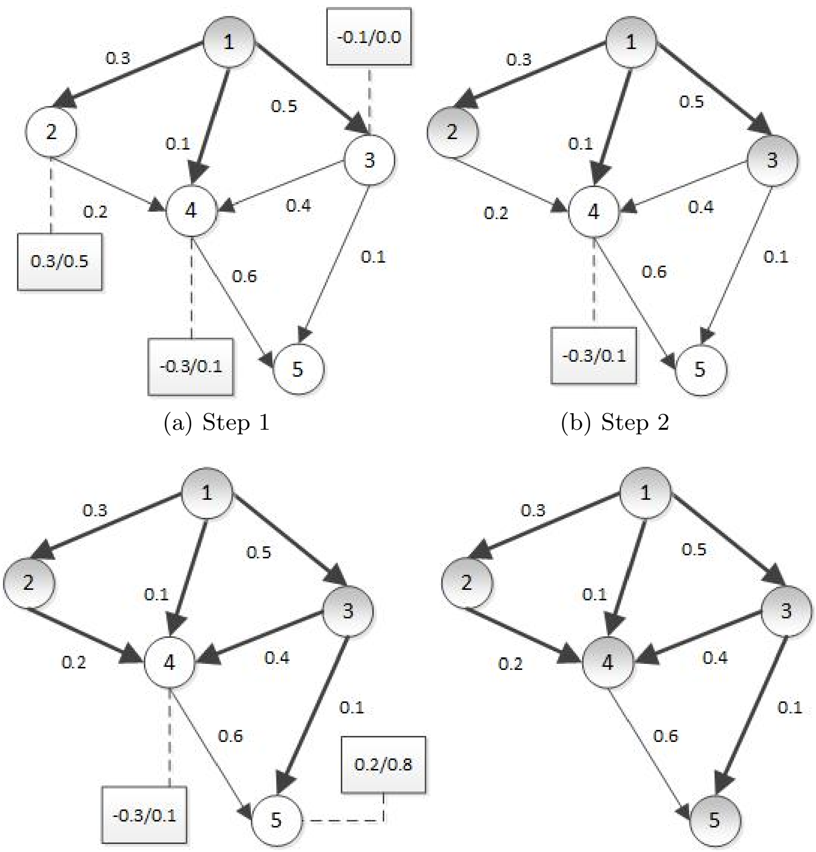
foreach *node* $j \in deg^+(i)$ **do**
  if $\alpha^t(i, j) \geqslant \theta_j \wedge j \notin active, waiting$ **then**
    $waiting \leftarrow waiting \cup \{j\}$;
    $triplets \leftarrow triplets \cup \{i, j, t\}$;
  end
end

Algorithm 2. Push variant of propagation loop with threshold condition

In the pull variant, see Algorithm 3, each node's tendency to become active increases monotonically as more of its neighbors become active. This time the propagator requires a many-to-one operator using deg^- which corresponds to a decision taken by an expert committee.

foreach *node* $j \in deg^{-}(i)$ **do**
  if $\sum_{k \in deg^{-}(j) \wedge k \in active} \alpha^{t}(k,j) \geqslant \theta_j \wedge j \notin active, waiting$ **then**
    $waiting \leftarrow waiting \cup \{j\}$;
    $triplets \leftarrow triplets \cup \{i,j,t\}$;
  end
end

Algorithm 3. Pull variant of propagation loop with threshold condition



(a) Step 1 (b) Step 2

(c) Step 3 (d) Step 4

Fig. 1. Preference-pull variant of propagation algorithm

In order to show the diversity of threshold conditions we employ one more example. The preference variant of condition could incorporate the factor positive/negative influence $f(i,m)$ of node i corresponding to propagated item m, where $f(i,m) \in [-1,1]$. This idea of using preferences is motivating by bundling multiple items in viral marketing. Then, in the pull case, a node i is activated when the sum of the preference scores on the propagated items $\sum_{m=1}^{k} f(j,m)$ and the influence scores from a non-empty set of neighbour active nodes, is greater or equal to the threshold θ_i. In addition, the influence factor f can be changed by using a temporal decay function as in [19]. Additionally, we can immunize selected nodes by adding them to *inactive* set.

In order to have a better understanding of the propagation algorithm, we capture the dynamics of it's pull variant with preference in a step-by-step fashion. As shown in Figure 1(a), the seed node 1 propagates action to all neighbours 2, 3 and 4. Since the sum of influence $p(1,2) = 0.3$ and preference value $f(2, item) = 0.3$ exceeds the threshold value $\theta_2 = 0.5$, according to the associated rectangle [0.3/0.5], node 2 becomes an active node. Similarly, node 3 becomes active. In contrast, node 4 remains inactive because its reference value gains $f(4, item) = -0.3$, see Figure 1(b). In this particular case, the sum of influence and preference values does not meet the threshold value for node 4. Afterwards, in the next time step, see Figure 1(c), nodes 2 and 3 (which are active) propagate action to their neighbours 4 and 5. As a result, node 4 receives the accumulated influences of three nodes and now despite a negative preference value $f(4, item) = -0.3$ meets the threshold value $\theta_4 = 0.1$. For the last node 5 the activation from node 3 was sufficient, see Figure 1(d). Since all available nodes became active, the propagation process terminates.

4 Summary

We have touched upon a few topics of social propagation, briefly explaining the model, and the algorithm, while referring to articles in parts where more details can be found. Thanks to its simplicity, our generic model is potentially useful in a wide range of scenarios. For a model and an algorithm, not only a theoretical analysis but also numerical simulations are required to demonstrate how accurate they are in fitting real issues, but these are beyond the scope of this paper, and will be addressed in future research.

Social propagation constitutes a large area of research that is rather young but rapidly evolving. Therefore, our analysis is by no means complete. Despite the considerable amount of ongoing research, we are still far from a satisfactory adherence to reality and proper utilization of propagation in the majority of cases. Multiple important questions are still open, like understanding the tipping point of epidemics, predicting who-wins among competing viruses/products, developing effective algorithms for immunization, and building more realistic propagations models while analyzing numerous real datasets. In this regard, the vast number of propagation techniques, which still remain unexplored, need to be thoroughly investigated in order to improve propagation capabilities and enhance system's efficiency.

One interesting but to date unsolved problem is how to learn on the fly the time-varying elements of propagation by mining the present and archived log of past propagations. The simple and insufficient algorithm for capturing the influence probabilities among the nodes and the prediction time by which an influenced node will perform an action after its neighbors have performed the action is presented in [28].

Another open issue is the low efficiency of greedy algorithms to compute the influence maximization problem to large networks. Even with recent optimizations, it still takes several hours. Improving the greedy algorithm is difficult, so this leads to a second possibility - the quest for appropriate heuristic.

Social propagation occurs in various forms. Some of them, like social influence mining, community detection, locating and repairing faults, finding effectors and maximizing influence, constitute the key components that enable useful insights into network behaviour and developing future services.

We hope that this paper provides several advanced points for engineers and scientists to use the propagation methods in more effective manner. However, it is possible to exploit appropriately this potential only after making further investigations on real data what we plan soon.

Acknowledgments. This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme under grant FP7-PEOPLE-2010-IEF-274375-EPP.

References

1. Fu, X., Wang, C., Wang, Z., Ming, Z.: Threshold random walkers for community structure detection in complex networks. Journal of Software 8(2) (2013)
2. Barbieri, N., Bonchi, F., Manco, G.: Cascade-based community detection. In: Proceedings of the 6th International Conference on Web Search and Data Mining, WSDM 2013, pp. 33–42. ACM, New York (2013)
3. Kim, H., Tang, J., Anderson, R., Mascolo, C.: Centrality prediction in dynamic human contact networks. Comput. Netw. 56(3), 983–996 (2012)
4. Doerr, B., Fouz, M., Friedrich, T.: Why rumors spread so quickly in social networks. Commun. ACM 55(6), 70–75 (2012)
5. Zhao, L., Guan, X., Yuan, R.: Modeling collective blogging dynamics of popular incidental topics. Knowledge and Information Systems 31(2), 371–387 (2012)
6. Garg, P., King, I., Lyu, M.R.: Information propagation in social rating networks. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012, pp. 2279–2282. ACM, New York (2012)
7. Chen, Y.C., Peng, W.C., Lee, S.Y.: Efficient algorithms for influence maximization in social networks. Knowledge and Information Systems 33(3), 577–601 (2012)
8. Gao, C., Liu, J., Zhong, N.: Network immunization and virus propagation in email networks: experimental evaluation and analysis. Knowledge and Information Systems 27(2), 253–279 (2011)
9. Liu, D., Chen, X.: Rumor propagation in online social networks like twitter – a simulation study. In: 2011 Third International Conference on Multimedia Information Networking and Security (MINES), pp. 278–282 (2011)
10. Nguyen, N.P., Yan, G., Thai, M.T., Eidenbenz, S.: Containment of misinformation spread in online social networks. In: Proceedings of the 3rd Annual ACM Web Science Conference, WebSci 2012, pp. 213–222. ACM, New York (2012)

11. Piraveenan, M., Uddin, S., Chung, K.: Measuring topological robustness of networks under sustained targeted attacks. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 38–45 (2012)
12. Bonchi, F.: Influence propagation in social networks: A data mining perspective. IEEE Intelligent Informatics Bulletin 12(1), 8–16 (2011)
13. Sun, J., Tang, J.: A survey of models and algorithms for social influence analysis. In: Aggarwal, C.C. (ed.) Social Network Data Analytics, pp. 177–214. Springer US (2011)
14. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th Int. Conf. on Knowledge Discovery and Data Mining, KDD 2005, pp. 177–187. ACM (2005)
15. Fay, D., Haddadi, H., Thomason, A., Moore, A.W., Mortier, R., Jamakovic, A., Uhlig, S., Rio, M.: Weighted spectral distribution for internet topology analysis: theory and applications. IEEE/ACM Trans. Netw. 18(1), 164–176 (2010)
16. Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., Havlin, S.: Catastrophic cascade of failures in interdependent networks. Nature 464(7291), 1025–1028 (2010)
17. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
18. Doerr, B., Fouz, M., Friedrich, T.: Social networks spread rumors in sublogarithmic time. In: Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, STOC 2011, pp. 21–30. ACM, New York (2011)
19. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. 81, 591–646 (2009)
20. Shah, D., Zaman, T.: Detecting sources of computer viruses in networks: theory and experiment. SIGMETRICS Perform. Eval. Rev. 38(1), 203–214 (2010)
21. Borge-Holthoefer, J., Banos, R.A., Gonzalez-Bailon, S., Moreno, Y.: Cascading behaviour in complex socio-technical networks. Journal of Complex Networks 1(1), 3–24 (2013)
22. Weng, L., Ratkiewicz, J., Perra, N., Goncalves, B., Castillo, C., Bonchi, F., Schifanella, R., Menczer, F., Flammini, A.: The role of information diffusion in the evolution of social networks. In: Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining, KDD 2013. ACM (2013)
23. Dodds, P.S., Watts, D.J.: A generalized model of social and biological contagion. Journal of Theoretical Biology 232(4), 587–604 (2005)
24. Jacquez, J.A., O'Neill, P.: Reproduction numbers and thresholds in stochastic epidemic models. Math. Biosciences 107(2), 161–186 (1991)
25. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining, KDD 2003, pp. 137–146. ACM (2003)
26. Rodriguez, M.A., Neubauer, P.: A path algebra for multi-relational graphs. In: Proceedings of the 27th International Conference on Data Engineering Workshops, ICDEW 2011, pp. 128–131. IEEE Computer Society, Washington, DC (2011)
27. Hang, C.W., Wang, Y., Singh, M.P.: Operators for propagating trust and their evaluation in social networks. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009, pp. 1025–1032 (2009)
28. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, WSDM 2010, pp. 241–250. ACM, New York (2010)