Effect of In/Out-Degree Correlation on Influence Degree of Two Contrasting Information Diffusion Models

Kouzou Ohara[1], Kazumi Saito[2], Masahiro Kimura[3], and Hiroshi Motoda[4]

[1] Department of Integrated Information Technology, Aoyama Gakuin University
ohara@it.aoyama.ac.jp
[2] School of Administration and Informatics, University of Shizuoka
k-saito@u-shizuoka-ken.ac.jp
[3] Department of Electronics and Informatics, Ryukoku University
kimura@rins.ryukoku.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
motoda@ar.sanken.osaka-u.ac.jp

Abstract. How the information diffuses over a large social network depends on both the model employed to simulate the diffusion and the network structure over which the information diffuses. We analyzed both theoretically and empirically how the two contrasting most fundamental diffusion models, Independent Cascade (IC) and Linear Threshold (LT) behave differently or similarly over different network structures. We devised two rewiring structures, one preserving in/out-degree correlation and the other changing in/out-degree correlation while both preserving their in/out-degree distributions, and analyzed how co-link rate and in/out-degree correlation affect the influence degree of each diffusion model using two real world networks, each as the base network on which rewiring is imposed. The results of the theoretical analysis qualitatively explain the empirical results, and the findings help deepen the understanding of complex diffusion phenomena.

Keywords: Information diffusion, network structure, influence degree, node degree distribution

1 Introduction

The emergence of Social Media such as Facebook, Digg and Twitter has provided us with the opportunity to create large social networks, which are becoming an important medium for spreading information. Recently, substantial attention has been devoted to analyzing and mining social networks from the point of information diffusion [14, 15, 11, 19, 2, 1, 16]. One of the most well studied problems is the influence maximization problem, *i.e.*, the problem of finding a limited number of influential nodes that are effective for the spread of information. Many algorithms have been proposed to solve the problem using probabilistic information diffusion models on a network [8, 12, 5, 9, 6, 4]. In order to investigate diffusion phenomena using probabilistic models, it is indispensable to understand the behavioral differences among models, and provide an effective method for selecting the most appropriate model for a particular task we want to analyze.

There are two contrasting fundamental probabilistic models that have been widely used by many researchers. One is the *independent cascade (IC)* model [7, 8] and the other is the *linear threshold (LT)* model [18, 8]. The IC model takes a sender-centered approach such that each information sender independently influences its neighbors with some probability (*information push style model*). The LT model is a receiver-centered approach such that each information receiver adopts the information if and only if the number of its neighbors that have adopted the information exceeds some threshold, where the threshold is treated as a random variable (*information pull style model*). We analyze how the IC and the LT models differ from or similar to each other in terms of information diffusion for a wide range of social networks with different structures.

In this paper, we compare *influence degree* obtained by the IC and the LT models from the network structure perspective. Here, the influence degree of a node v under a probabilistic diffusion model in a network is defined to be the expected number of *active* nodes at the end of the information diffusion process that starts from the initial active node v, where nodes that have been influenced with the information are referred to as being active. First, we theoretically analyze the properties of the IC and the LT models on scale-free networks, and derive the following two properties: 1) as the in/out-degree correlation decreases, the influence degree decreases for the IC model but it does not change for the LT model and 2) as the co-link (bidirectional link) rate decreases, the influence degree increases for both the IC and the LT models, but the IC model is much less sensitive than the LT model. To verify these properties, we systematically generated a series of scale-free networks with varying in/out-degree correlation and co-link rate, applying two rewiring strategies, one preserving in/out-degree correlation and the other changing in/out-degree correlation while both preserving their in/out-degree distributions. We used two real world scale free networks as the bases to apply these strategies, and experimentally confirmed that the above two properties indeed hold.

2 Diffusion Models

Let $G = (V, E)$ be a directed network, where V and E ($\subset V \times V$) are the sets of all the nodes and links, respectively, and $|V| \leq |E|$ can be naturally assumed for commonly-seen social networks. We recall the definition of the IC and the LT models according to the literatures [8, 9]. In these models, the diffusion process proceeds from an initial active node in discrete time-step $t \geq 0$, and it is assumed that nodes can switch their states only from inactive to active (*i.e.*, the SIR setting).

The IC model has a *diffusion probability* $p_{u,v}$ with $0 < p_{u,v} < 1$ for each link (u, v) as a parameter. Suppose that a node u first becomes active at time-step t, it is given a single chance to activate each currently inactive child node v, and succeeds with probability $p_{u,v}$. If u succeeds, then v will become active at time-step $t + 1$. If multiple parent nodes of v first become active at time-step t, then their activation trials are sequenced in an arbitrary order, but all performed at time-step t. Whether u succeeds or not, it cannot make any further trials to activate v in subsequent rounds. The process terminates if no more activations are possible.

The LT model has a *weight* $q_{u,v}$ (> 0) with $\sum_{u \in B(v)} q_{u,v} \leq 1$ for each link (u, v) as a parameter, where $B(v) = \{u \in V;\ (u, v) \in E\}$ is the set of parent nodes of node v. First,

for any node $v \in V$, a *threshold* θ_v is chosen uniformly at random from the interval $[0, 1]$. An inactive node v is influenced by its active parent nodes. If the total weight from v's active parent nodes at time-step t is no less than θ_v, *i.e.*, $\sum_{u \in B_t(v)} q_{u,v} \geq \theta_v$, then v will get activated at time-step $t+1$. Here, $B_t(v)$ is the set of all the parent nodes of v that are active at time-step t. The process terminates if no more activations are possible.

3 Analysis of Local Influence Degree

We first define local influence degree of node u, denoted by $\sigma_L(u)$, as the expected number of u's child nodes directly activated by u. For the IC model, $\sigma_L^{IC}(u)$ is given by $\sigma_L^{IC}(u) = \sum_{v \in F(u)} p_{u,v}$, where $F(u)$ stands for the set of u's child nodes defined by $F(u) = \{v \in V;\ (u, v) \in E\}$. For the LT model $\sigma_L^{LT}(u)$ is given by $\sigma_L^{LT}(u) = \sum_{v \in F(u)} q_{u,v}$ because each weight $q_{u,v}$ is regarded to be the probability that the threshold θ_v is chosen from the interval $[0, q_{u,v}]$. Then, we can calculate the average local influence degree over all nodes, denoted by $\bar{\sigma}_L(G)$. For the LT model, if we impose the condition $\sum_{u \in B(v)} q_{u,v} = 1$ for any node $v \in V$, we can prove $\bar{\sigma}_L^{LT}(G) = 1$ from the following relations.

$$\bar{\sigma}_L^{LT}(G) = \frac{1}{|V|} \sum_{u \in V} \sigma_L^{LT}(u) = \frac{1}{|V|} \sum_{u \in V} \sum_{v \in F(u)} q_{u,v} = \frac{1}{|V|} \sum_{(u,v) \in E} q_{u,v} = \frac{1}{|V|} \sum_{v \in V} \sum_{u \in B(v)} q_{u,v} = 1.$$

For the IC model, if we impose the uniform diffusion probability setting, i.e., $p_{u,v} = p$ for any link $(u, v) \in E$, which has been employed in many previous studies (*e.g.*, [8]), we can calculate $\bar{\sigma}_L^{IC}(G)$ as follows:

$$\bar{\sigma}_L^{IC}(G) = \frac{1}{|V|} \sum_{u \in V} \sigma_L^{IC}(u) = \frac{1}{|V|} \sum_{u \in V} \sum_{v \in F(u)} p_{u,v} = \frac{1}{|V|} \sum_{(u,v) \in E} p = \frac{|E|}{|V|} p,$$

where $\frac{|E|}{|V|}$ is equal to the average degree $d = \frac{1}{|V|} \sum_{u \in V} |B(u)| = \frac{1}{|V|} \sum_{u \in V} |F(u)| = \frac{|E|}{|V|}$, and is no less than 1 as we assume $|V| \leq |E|$. Thus, by setting the uniform diffusion probability to the inverse of average degree, i.e., $p = \frac{1}{d} = \frac{|V|}{|E|}$, we obtain $\bar{\sigma}_L^{IC}(G) = 1$. This makes the IC and LT models equivalent in terms of the average local influence degree. Hereafter, we impose these settings to evaluate the influence degree. Note that local influence degree of node u for the IC model becomes $\sigma_L^{IC}(u) = \sum_{v \in F(u)} p_{u,v} = p|F(u)|$.

So far we focused on local influence degree of node $u \in V$ under the condition that the node u has become active. However, when considering the cascade of information diffusion, we need to consider the probability $r(u)$ that the node u is activated by its parent nodes. Namely, we consider cascading local influence degree defined by $\sigma_{CL}(u) = r(u)\sigma_L(u)$. As the simplest case, we employ the probability $r(u)$ that the node u is activated at the next time step by some active node selected uniformly at random from the node set V. For the IC model, $r^{IC}(u)$ is given by $r^{IC}(u) = \frac{1}{|V|} \sum_{s \in B(u)} p_{s,u} = \frac{p|B(u)|}{|V|}$, and for the LT model, $r^{LT}(u)$ is given by $r^{LT}(u) = \frac{1}{|V|} \sum_{s \in B(u)} q_{s,u} = \frac{1}{|V|}$. Thus we obtain the average cascading local influence degree $\bar{\sigma}_{CL}$ for the IC and LT models as follows:

$$\bar{\sigma}_{CL}^{IC}(G) = \frac{1}{|V|} \sum_{u \in V} r^{IC}(u)\sigma_L^{IC}(u) = \frac{p^2}{|V|^2} \sum_{u \in V} |B(u)||F(u)|, \tag{1}$$

$$\bar{\sigma}_{CL}^{LT}(G) = \frac{1}{|V|} \sum_{u \in V} r^{LT}(u)\sigma_L^{LT}(u) = \frac{1}{|V|^2} \sum_{u \in V} \sigma_L^{LT}(u) = \frac{1}{|V|}. \tag{2}$$

Therefore, by noting that the in/out-degree correlation $dc_{I/O}(G)$ is quantified by

$$dc_{I/O}(G) = \frac{\frac{1}{|V|}\sum_{u\in V}|B(u)||F(u)| - d^2}{\sqrt{\frac{1}{|V|}\sum_{u\in V}|B(u)|^2 - d^2}\sqrt{\frac{1}{|V|}\sum_{u\in V}|F(u)|^2 - d^2}},$$

and the denominator of $dc_{I/O}(G)$ is determined by the standard deviations of in/out-degree distributions, we can see that the average cascading local influence degree of the IC model is affected by the in/out-degree correlation $dc_{I/O}(G)$ when the standard deviations are fixed, as shown in Eq. (1), while that of the LT model is not affected, as shown in Eq. (2). Namely, we can conjecture that influence degree of the IC model also decreases when the in/out-degree correlation decreases.

Another important factor affecting influence degree is the co-link rate $cr(G)$ which is defined by $cr(G) = \frac{1}{|E|}\sum_{u\in V}|B(u)\cap F(u)|$. Evidently, for a bidirectional network G, we obtain $cr(G) = 1$ because $B(u) = F(u)$ for any $u \in V$. Assume a node $v \in B(u) \cap F(u)$; if v succeeds activating u, then the reverse link (u, v) never contributes to increasing an active node, conversely, if u succeeds activating v, then the reverse link (v, u) never does so. Thus, we conjecture that influence degree of the IC and LT model increases when the co-link rate $cr(G)$ decreases. However, there is a subtle difference between the IC and the LT models. Think of the network with co-link rate close to 1. Evidently the in/out-degree correlation is also close to 1. Assume that k parents of a node v which has a large degree $D = |F(v)| = |B(v)|$ get activated. The expected probability that the node v becomes activated is $1 - (1 - 1/d)^k$ for the IC model and k/D for the LT model where d is the average node degree. For the IC model the probability is large for a small number of k and insensitive to $|D|$. Thus, once it gets activated, the reverse k links which do not contribute further activation is small. On the other hand, for the LT model the node v is not activated unless k is large. Thus, once it gets activated, the reverse k links do not contribute further activation is also large. This implies that the IC model is less sensitive to the change of co-link rate than the LT model.

4 Experiments

To confirm our conjectures in Section 3, we conducted extensive experiments using both synthetic and real world large networks, rewiring their links according to the two strategies presented in this section. However, due to the page limitation, we show only the results for the two real world networks: one bidirectional and the other directional[1].

4.1 Rewiring Strategies

We devised two rewiring strategies. Both preserve the in/out-degree distribution. The first one rewires links of a given network G preserving the in/out-degrees of each node, which is equivalent to the method of generating randomized networks presented in [13]. We implemented this strategy by swapping the two destination nodes v and v' of links

[1] The networks we omitted here include synthetic networks generated by the BA model [3] and the CNN model [17], and four other networks derived from the real world data.

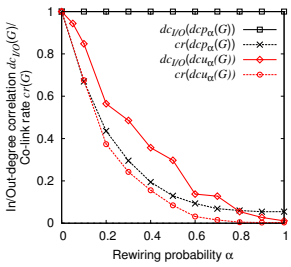
$e = (u, v)$ and $e' = (u', v')$ from two starting nodes u and u'. The links are chosen uniformly at random. Obviously, this never changes $dc_{I/O}(G)$, but does change $cr(G)$. We refer to this rewiring strategy as the DCP (in/out-Degree Correlation Preserved) method, and denote the network G rewired by this method by $dcp_\alpha(G)$, where α is the link rewiring probability, *i.e.*, v of e and v' of e' are swapped with the probability α. The larger α is, the smaller $cr(G)$ is. Thus, the DCP method allows us to investigate how the co-link rate affects the influence degree of the IC and the LT models. The second one rewires links changing the in/out-degree correlation. This is to confirm our conjecture that the in/out-degree correlation affects the influence degrees of the IC model. We implemented this by swapping $E_I(v)$, all the incoming links to a node v, and $E_I(v')$, all the incoming links to a node v' with a probability α. Nodes v and v' are randomly chosen. Namely, $E_I(v)$ becomes $\{(u, v); u \in B(v')\}$, and $E_I(v')$ becomes $\{(s, v'); s \in B(v)\}$ after swapping. This method changes the in-degree of chosen nodes without changing their out-degree while preserving the in/out-degree distributions of the network G. We refer to this method as the DCU (in/out-Degree Correlation Unpreserved) method, and denote the network G rewired by the DCU method with a link rewiring probability α by $dcu_\alpha(G)$. The larger α is, the smaller the in/out-degree correlation is.

4.2 Datasets and Network Structure

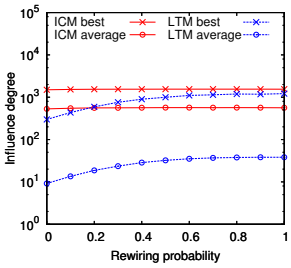
In this section, we explain the two real world networks for which we present the experimental results. The first one is a bidirectional network derived from the Enron Email Dataset [10]. We regarded each email address as a node, and constructed a bidirectional link between two email addresses u and v only if u sent an email to v and received an email from v. After that, we extracted the maximal strongly connected component. We refer to this bidirectional network as the Enron network, which has $4,254$ nodes and $44,314$ directed links. The second one is a directional network derived from a Japanese word-of-mouth communication site for cosmetics, "@cosme"[2], where each user page is associated with *fan links*. A fan link from user u to user v is generated if user v registers user u as his/her favorite user. We extracted a fan network from @cosme by tracing up to ten steps in the fan links starting from a randomly chosen user in December 2009. The resulting network has $45,024$ nodes and $351,299$ directed links. We refer to this network as the Cosme network.

For these networks, we investigated the influence degree $\sigma(v)$ of each node v of the networks $dcp_\alpha(G)$ and $dcu_\alpha(G)$ under the IC and the LT models, varying α from 0.0 to 1.0 by 0.1. Note that $dcp_{0.0}(G) = dcu_{0.0}(G) = G$. The influence degree $\sigma(v)$ was estimated by the empirical mean of the number of active nodes obtained from 10,000 independent runs of information diffusion based on the bond percolation technique [9]. According to the discussion in Section 3, we set a unique value $p = 1/d$ to every $p_{u,v}$ for the IC model. Namely, p was set to 0.10 for the Enron network, and 0.13 for the Cosme network.

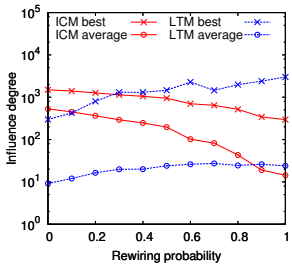
[2] `http://www.cosme.net/`



(a) The change of In/Out-degree correlation and Co-link rate.

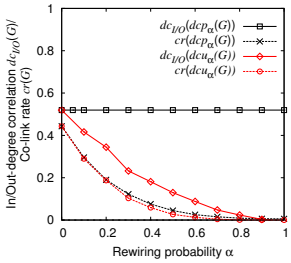


(b) The best and the average influence degrees with the DCP method.

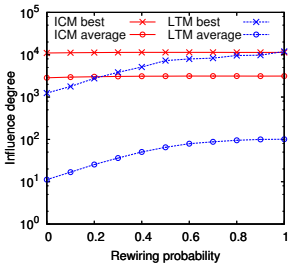


(c) The best and the average influence degrees with the DCU method.

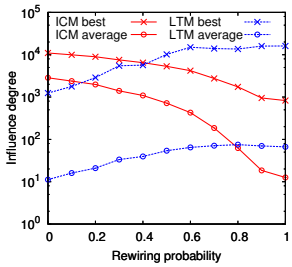
Fig. 1: Experimental results for the Enron network.



(a) The change of In/Out-degree correlation and Co-link rate.



(b) The best and the average influence degrees with the DCP method.



(c) The best and the average influence degrees with the DCU method.

Fig. 2: Experimental results for the Cosme network.

4.3 Experimental Results

Figures 1a and 2a show how the in/out-degree correlation $dc_{I/O}(G)$ and the co-link rate $cr(G)$ of a given network G change with the two rewiring methods, DCP and DCU, for the Enron and the Cosme networks, respectively. We see that both methods work just as we intended: $cr(G)$ decreases in a similar fashion for both the DCP and the DCU methods, as the rewiring probability α becomes larger, while $dc_{I/O}(G)$ does not change with the DCP method, but it does decrease similarly to $cr(G)$ with the DCU method. Note that both $dc_{I/O}(G)$ and $cr(G)$ of the Enron network are 1.0 for $\alpha = 0.0$ because it is bidirectional.

Figure 1b illustrates how the DCP method affects the best and the average influence degrees over all the nodes of the Enron network. As we expected, both influence degrees of the LT model become larger as the rewiring probability becomes larger, and the co-link rate becomes smaller. The influence degrees of the IC model does not seem to increase, but indeed they slightly increase within the range of $\alpha = 0.0$ to 0.6 where the co-link rate drastically decreased. This qualitatively supports the analysis in Section 3.

The same tendencies can be found in the result for the Cosme network as shown in Fig. 2b. We also observed the same tendencies for the other networks we omitted here.

Figures. 1c and 2c show how the DCU method affects the best and the average influence degrees of the IC and the LT models. Both $dc_{I/O}(G)$ and $cr(G)$ decrease with α. This imposes two conflicting factors for the IC model, but the effect of $dc_{I/O}(G)$ surpasses and the influence degrees of the IC model decrease for both the Enron and the Cosme networks. On the other hand, the influence degrees of the LT model are affected by only $cr(G)$. Thus, they increase in the same way as in Figs. 1b and 2b. The same observation is obtained for the other networks. This also qualitatively supports the analysis in Section 3.

5 Conclusion

Understanding how information diffuses over a large social network is important to do any kind of social network analysis, but it is difficult because actual diffusion depends on both the diffusion model employed and the properties of the network structure over which the information diffuses. Independent Cascade (IC) and Linear Threshold (LT) models have been used widely by many researchers. Both are probabilistic models but have contrasting properties, *i.e.*, information push (IC) and information pull (LT). Social networks have common characteristics. The most important one would be the scale free property. There can be many structures that hold this property. We devised two rewiring strategies that can systematically transform one network structure to another structure preserving the scale free property, one preserving in/out-degree correlation (DCP method) and the other changing in/out-degree correlation (DCU method). Each strategy was successively applied with different probabilities to two real world social networks, generating a series of networks, each with a gradually changing structure. We chose co-link rate and in/out-degree correlation as the two parameters that characterize the network structure, and investigated how these parameters affects the influence degree of the two models (IC and LT). The major new findings are 1) the IC model is sensitive to in/out-degree correlation and the influence degree is positively correlated to it, whereas the LT model is insensitive to it and 2) Both the IC and the LT models are negatively correlated to co-link rate, but its dependency is much less sensitive in the IC model. These properties can be qualitatively derived by the theoretical analysis and verified by the extensive experiments using the above networks as well as others not reported in this paper. These findings are useful in deepening our understanding of the complex information diffusion phenomena over a social network.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-11-4111, and JSPS Grant-in-Aid for Scientific Research (No. 23700181).

References

1. 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)
2. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM conference on Electronic Commerce. pp. 325–334 (2009)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
4. 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)
5. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009). pp. 199–208 (2009)
6. 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)
7. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12, 211–223 (2001)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. Melo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science 298, 824–827 (2002)
14. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)
15. 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)
16. Romero, D., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th International World Wide Web Conference (WWW2011). pp. 695–704 (2011)
17. Vázquez, A.: Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Physical Review 67(5), 056104 (2003)
18. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA 99, 5766–5771 (2002)
19. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)