2013

Effects of the Network Structure on the Dynamics of Viral Marketing

Marek Opuszko
Friedrich-Schiller-University, Jena, Germany, marek.opuszko@uni-jena.de

Johannes Ruhland
Friedrich-Schiller-University, Jena, Germany, johannes.ruhland@uni-jena.de

Effects of the Network Structure on the Dynamics of Viral Marketing

Marek Opuszko and Johannes Ruhland

Friedrich-Schiller-University, Jena, Germany
{marek.opuszko,johannes.ruhland}@uni-jena.de

Abstract. We present an analysis of how the structure of a social network influences the diffusion of information in a viral marketing context. We performed diffusion simulations on a large number of real world and artificially generated network datasets. We analyze how the characteristics of a network and parameter settings like the selection of initial start nodes influences the diffusion. The results indicate that the network structure has a significant effect on the diffusion. Extreme cases show a difference in the diffusion of over 65%. Our investigation also proves that a viral marketing diffusion may be predicted without the knowledge of the whole network. We further provide useful recommendations for marketers which could be taken into consideration when marketing campaigns are conducted.

Keywords: viral marketing, information diffusion, social networks

1 Introduction

The success of virtual social communities on the web and the increasing resistance and avoidance of customers towards traditional forms of advertising [1] led marketers to turn to new forms of marketing such as social media marketing and viral marketing as a special type of electronic word-of-mouth marketing (WOM). Customers nowadays can easily share information including product information with their friends online, leading to a change in the information diffusion process within social peer groups. Solomon [2] states that WOM effects play an important role in the formation of judgments and attitudes towards innovations. Traditional marketing forms, on the other hand, are increasingly considered irrelevant by consumers [3]. Another important aspect of WOM and, in particular, viral marketing is the fact that the diffusion process may operate at very low costs, using an underlying social network and the participation of the network nodes themselves to actively spread an information artifact. With the increasing success of online communities, we can observe numerous examples of "viral" products, websites or user generated content. One of the most famous examples is the viral marketing campaign for the movie *The Blair Witch Pro-*

11th International Conference on Wirtschaftsinformatik,
27th February – 01st March 2013, Leipzig, Germany

ject. The monetary success of nearly 250 million US dollars exceeded the production costs of 60,000$ US by the factor of 4166[1].

Since viral marketing promises great effects at low costs, marketers and researchers showed interest in the understanding of the underlying processes. A main goal was to examine whether an information or innovation spreading dies out quickly or diffuses significantly into the population. Unfortunately these effects are difficult to observe in real world settings, as they usually appear spontaneous [4]. Following Bampo et al. [5], three main components determine the viral process: (1) the behavioral characteristics of the members of the network, (2) the seeding strategy and (3) the structure of the social network. Hinz et al. [6] added a fourth critical factor, (4) the attractiveness of the content supposed to be transported. Since factor (2), the seeding strategy, is the only parameter that is largely under the control of the campaign initiator, this issue has received much attention in different research fields. Seeding strategies usually aim to identify influential nodes as the initial set of nodes in a diffusion process.

Kempe and colleagues theoretically formalized this NP-Complete problem as the diffusion maximization problem [7]. Numerous strategies have been introduced and examined, mainly in the computer science and marketing community [6-10], still leaving some controversy and contradictory findings about the impact of seeding and viral marketing [6], [11]. According to a study among marketers in 2007, the major perceived problem of viral marketing is still the missing experience and a lack of measurability of the advertising effect [12]. Many factors of viral marketing diffusion still remain imprecise and vague. Leskovec et al. showed in their analysis that viral marketing is not as epidemic as hoped. They further stated that the topology of the addressed networks (factor (3)) should be analyzed in more detail [13]. Beside the conventional wisdom that the network structure influences the diffusion of information, its role has not been deeply investigated. Watts and Strogatz [14] showed that networks comprising high clustering and small path lengths, what they called "Small World Networks", facilitate an epidemic spread. Bampo et al. found with a simulation of a real viral marketing campaign in generated networks, that the structure of a social network has a significant impact on a viral marketing campaign performance [5]. Interestingly they found that scale-free networks, first introduced by Barabási and Albert [15], are very efficient for viral campaigns and small world networks generally temper the spread of information. Shakarian and Paulo investigated on viral marketing diffusion simulations in numerous networks and observed three different types of networks, each type showing a very distinct diffusion pattern [10]. To date, no research offers extensive investigations on what network characteristics exactly influence the information diffusion in social networks, and if, to what extent. Moreover, the majority of the research on seeding is based on investigations on very few network datasets only [6-8], [16]. The findings are therefore based on a specific network structure and may not be transferable to any other viral marketing application in networks with a different structure.

Another problem is that for all the diffusion maximization approaches the whole network structure must be known ex ante, including every connection between any

[1] http://boxofficemojo.com/movies/?id=blairwitchproject.htm

two nodes. Obviously, this is not the case in many real world marketing applications. Usually only few companies, mainly the service providers themselves like Facebook or Google, have access to all information about the underlying network. Nevertheless, some information is available and assumptions about the network structure can be made. We know for instance that in 2011 an average Facebook user has 130 friends and posted 90 pieces of content every month[2]. We further know that a network like Facebook comprises more than 150 million users in the US[3] in 2012 and that the network in 2011 showed a degree of separation of only 3.74[4].

To address the open questions we will undertake a simulation analysis on a set of real-world and artificially generated networks using different types of seeding methods and two state-of-the-art diffusion models. We will introduce a set of metrics that have been calculated to characterize every network. These metrics will be combined with the diffusion simulation results to answer the question what network characteristics influence the message diffusion. This also includes the interaction between the seeding method and the network characteristics. We will further conduct a decision tree analysis to answer the question whether the diffusion can be predicted based on the characteristics about the network structure or not. We conclude our work with a discussion of the results, managerial implications and future work perspectives.

2 Theoretical Framework

The research of diffusion of information in large online communities mainly derived from the studies of epidemics [17]. In other contexts, early efforts have been made in the understanding of the adoption of innovations in the medical and agricultural sector [18-19] or the success of product innovations [20-22]. To describe the diffusion, several models have been introduced in the past. Our investigation will rely on two standard models in this research field, the *independent cascade model* and the *linear threshold model* which reflect a different transmission behavior of the network nodes. These models are well understood and have been used in various works studying the diffusion of information in social networks [7], [13], [23-24].

The linear threshold model, as first proposed by Granovetter [25], is based on the assumption of node specific thresholds. We speak of *active* nodes if a network node has adopted an innovation or received information, e.g. a marketing message. According to the model, a node is activated if the number of active nodes in the neighborhood reaches a threshold. A real world example would be the adoption of fashion or technology trends. In the context of virtual communities like Facebook, we can observe similar aspects. In Facebook, for instance, little messages like "*Your friend XYZ is now using …"* are presented on the starting screen after a user logs in, highlighting friends using apps or games. One could make the assumption that some users might

[2] http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/

[3] http://www.socialbakers.com/facebook-statistics/

[4] http://www.bbc.co.uk/news/technology-15844230. The degree of separation refers to the average number of friends in chain from one user to any other user in the network, excluding the end nodes.

also start using a certain app (usually a game) after a critical amount of friends did it before[5]. According to Kempe [7] the process works as follows: an inactive node v is activated if the total weight b_w of any of its active neighbors w reaches the threshold θ_v (in the interval [0,1]):

$$\sum_w b_w \geq \theta_v \tag{1}$$

The diffusion process unfolds as a deterministic process in discrete steps t, with all nodes active in t-1 still active in t, given an initial set of active nodes A_0 at the beginning of the diffusion process. Each node uniformly chooses a threshold θ_v or the threshold is hard-wired for every node. The process will stop, if no further activations have been made.

Based on the idea of interacting particle systems but later used in the context of marketing [26], the independent cascade model can also be used to model diffusion processes in networks. It starts with a set of initial nodes A_0 and unfolds in discrete time steps t. Unlike the linear threshold model, an active node v is given a single chance to activate any inactive neighbor node w in the time step succeeding its activation. Whether or not an activation attempt is successful, node v cannot make any further attempts to activate w in future steps. The activation is based on an activation probability $p_{v,w}$ defined for every node prior to the diffusion start. The process runs until no further activations are possible. In the context of virtual online communities the independent cascade model could be compared with posting, commenting or sharing content on a friends pin wall in online social networks like Facebook or sending information artifacts to friends per email. If a friend is "activated", the message might be forwarded to further people, etc. Moreover, platforms like Facebook reinforce this process by automatically[6] sharing every status update a user makes with every friend (neighbor) of the user. These status updates include pin wall posts, using the *like* or *recommend* function or using games and apps.

3 Method

To evaluate the influence of the network structure on the information diffusion we ran simulations on networks using the independent cascade and the linear threshold model with different parameter settings. The simulation process was conducted as follows. For every network dataset 200 runs of simulations with randomly chosen start parameters were conducted. For every run the number of start nodes was set at random in the interval [1,50]. For the independent cascade model the activation probability for all nodes was set randomly in the interval [0,1]. For the linear threshold model the node threshold has been set equivalently. Furthermore, the criterion for choosing the start nodes (described in the next subsection) was selected randomly and the nodes have been selected accordingly by calculating the metrics for all network nodes. For

[5] There are plenty of examples like the very popular farmville game on facebook, having millions of active users: http://mashable.com/2010/02/20/farmville-80-million-users/

[6] A user on Facebook may change this functionality by using individual privacy settings.

every run, 50 single simulations were accomplished. The resulting diffusion was averaged and the resulting diffusion mean and the diffusion variance were stored including all used parameters and the network metrics as one case. All in all 720,000 simulations were conducted. For each diffusion model a dataset with 7200 cases each comprising the network characteristics, the simulation parameters and the results was created as a base for the analysis.

3.1 Seeding Criteria for the Selection of Initial Start Nodes

Several criteria for selecting the initial active nodes A_0 exists. The most simple would be by randomly activating a set of nodes. A common method is to evaluate the centrality of every node in the network based on different centrality measures and to choose the most *central* nodes as those are supposed to by influential [23]. We will evaluate the following centrality measures as a criterion for choosing initial nodes based on a given graph $G = (V,E)$ comprising network nodes (vertices) V and edges E. All centrality metrics have been calculated according to Wasserman and Faust [27] as well as Newman [28]. We refer to those publications for further details on the calculations.

- *Degree Centrality*, one of the most common centrality measures. Degree centrality reflects the number of ties (aka neighbors or friends in the context of online social networks) of a node.
- *Betweenness Centrality*, this centrality reflects the probability of a node to lie on a shortest path between two randomly chosen nodes.
- *Closeness Centrality*, this reflects to the inverse farness of a node v to any other node in the network.
- *Eigenvector Centrality*, a natural extension of the degree centrality. The difference is that nodes also award "points" for the degree centrality of their neighbors. A node is central if it is connected to other *important* nodes.
- *Node Clustering Coefficient*, sometimes also referred to as transitivity. The clustering coefficient of a node is the relation of the number of pairs of neighbors that are connected to the number of pairs of neighbors. In online social networks this reflects to the connection among a users' friends.
- *PageRank Coefficient*, extension of the eigenvector centrality used by Google to rank the centrality of web pages [29]. The difference is that the centrality of a node is further divided by the out-degree of a node.
- *Random*, the initial start nodes are chosen randomly.

3.2 Network Metrics

To describe a network's structure and characteristics and to later evaluate diffusion predictors, we calculated several metrics to describe a network.

- *Number of network nodes,* usually the number of users.
- *Number of network edges,* the number of connection between the nodes.

- *Network density*, the density of a network is the relation between existing and possible edges.
- *Connected graph* (yes | no), a graph is connected if all nodes belong to one (giant) component and no individual clusters exist.
- *Average path length,* the average of all shortest paths between any two nodes of the network.
- *Number of components/clusters*, number of isolated components comprising at least two nodes.
- *Network/Graph diameter*, the diameter of a network is the length of the longest shortest path between two arbitrary nodes in the network.
- *Average node degree*, a normalized ([0,1]) metric of the degree centrality of all nodes.
- *Average node betweenness*, a normalized ([0,1]) metric of the betweenness centrality of all nodes.
- *Average node closeness*, a normalized ([0,1]) metric of the closeness centrality of all nodes.
- *Average node eigenvector*, a normalized ([0,1]) metric of the eigenvector centrality of all nodes.
- *Average clustering coefficient*, a normalized ([0,1]) coefficient of the clustering coefficient of all nodes.
- *Number of network communities*, communities are sub graphs or dense groups of nodes within a network which are sparsely connected to other groups. In opposite to components, these groups are not isolated from each other. We used the leading eigenvector community detection algorithm according to Newman [30].
- *Degree distribution power law fit*, since the degree distributions of network nodes often show a power law distribution, we fitted a power-law distribution with maximum likelihood methods as recommended by Newman against the degree distribution of each network [31].

3.3 Network Datasets

We used both real world and artificially generated networks as a source for the simulations. Table 1 shows the used networks for the simulation. The real world networks include some of the most common datasets used in social network analysis. One exception is the dataset Student Network. This network has been extracted in a former analysis. It comprises a Facebook friendship network of university freshmen after their first semester of study. It should be noted that prior to the analysis, all isolated nodes have been deleted from the graphs. To generate the artificial networks, three state-of-the-art algorithms have been used: Erdös-Renyi game [41], Watts-Strogatz game [14] and Barabási-Albert game [15]. The artificial networks have been generated in order to represent the characteristics of the real world datasets in terms of node and edge count. All calculations have been done using the igraph [42] package in the R software [43].

Table 1. Networks used in the analysis

Name	Nodes	Edges
Social network of Dolphin interactions [32]	62	159
Coappearance of Les Miserables (novel) characters [33]	77	254
Topology of US Western States power grid [14]	4,941	6,594
Facebook university freshmen (Student Network)	471	926
Snapshot of Gnutella peer-to-peer network 2008 [34]	6,301	20,777
Social network of the University of California (OCLinks) [35]	1,899	20,297
Coauthorship network of network scientists (NetScience) [36]	1,461	2,742
A snapshot of the structure of the Internet[7]	22,963	48,436
High-Energy Theory coauthorship network (Hep-th) [37]	7,610	15,751
Condensed matter E-Print coauthorship network 2003 [37]	30,460	120,029
Condensed matter E-Print coauthorship network 2005 [37]	39,577	175,639
Erdös collaboration graph[8]	6,927	11,850
Astrophysics E-Print coauthorship network [37]	16,046	121,251
Network of Email interchanges [38]	1,133	5,451
Network of Jazz musicians [39]	198	2,742
Network of users of the Pretty-Good-Privacy algorithm [40]	10,680	24,316
Network created according to Barabási-Albert 1 [15]	60	177
Network created according to Barabási Albert 2 [15]	80	237
Network created according to Barabási Albert 3 [15]	1,400	2,798
Network created according to Barabási Albert 4 [15]	20,000	39,998
Network created according to Barabási Albert 5 [15]	30,000	119,996
Network created according to Erdös-Rényi 1 [41]	868	1,040
Network created according to Erdös-Rényi 2 [41]	914	12,683
Network created according to Erdös-Rényi 3 [41]	1,000	14,902
Network created according to Erdös-Rényi 4 [41]	1,000	25,362
Network created according to Erdös-Rényi 5 [41]	6,290	19,996
Network created according to Erdös-Rényi 6 [41]	6,917	23,767
Network created according to Watts-Strogatz 1 [14]	60	177
Network created according to Watts-Strogatz 2 [14]	80	240
Network created according to Watts-Strogatz 3 [14]	60	177
Network created according to Watts-Strogatz 4 [14]	1,400	2,800
Network created according to Watts-Strogatz 5 [14]	20,000	40,000
Network created according to Watts-Strogatz 6 [14]	30,000	120,000
Network created according to Watts-Strogatz 7 [14]	6,927	14,994

[7] http://www-personal.umich.edu/~mejn/netdata/
[8] http://vlado.fmf.uni-lj.si/pub/networks/data/

4 Analysis

4.1 Descriptive Analysis.

Figure 1 shows the resulting diffusion means of all networks in relation to the activation probability and threshold. The plot highlights a high variance in the diffusion curve depending on the underlying network. As some diffusion curves seem to follow the typical S-shape using the independent cascade model, some networks show an almost linear relation to the activation probability. The linear threshold model shows similar picture. Here the general S-shape is very steep showing some kind of critical threshold or tipping point. Under a certain value the diffusion stays very low. Once this critical value is reached, the diffusion quickly reaches high values and a high network saturation.

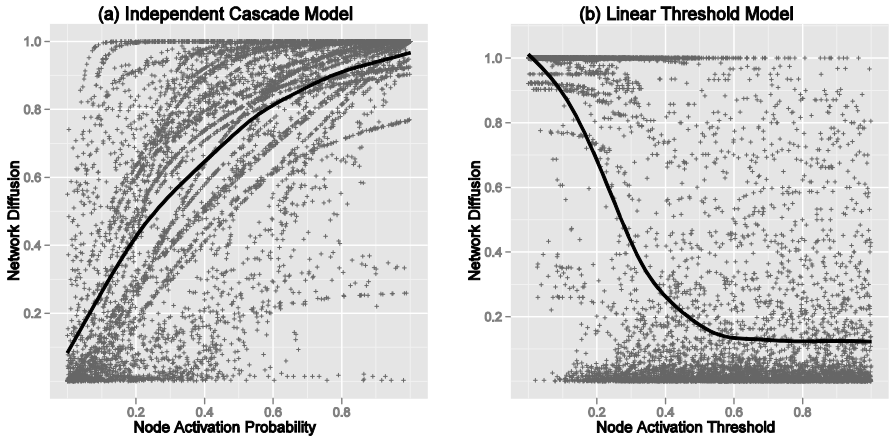

Fig. 1. Simulated diffusion of all networks depending on the activation probability of the network nodes. The figure includes a regression line (solid line).

To draw a more detailed picture, Figure 2 shows the diffusion curves of three exemplary networks, the cond-mat 2005 network, the power grid network and the Erdös collaboration graph network. The plot of the independent cascade model illustrates very clearly the effect of the underlying network. The Erdös graph shows an almost linear relationship whereas the cond-mat 2005 and the power grid network show dramatic differences. At a node activation probability value of 0.45 we can observe a very low mean diffusion of around 15% of the network nodes for the power grid network. On the other hand, the mean diffusion reaches values of 80% for the cond-mat 2005 network at the same value. Interestingly this does not account for the linear threshold model. Here the curve shapes are more similar. An interesting behavior shows the Erdös collaboration graph. As the only network, it reaches diffusion rates above 20% even at very high thresholds.

To evaluate the influence of the network parameters on the diffusions we calculated the correlation between those parameters and the mean diffusion. The results re-

vealed several significant correlations. Obviously the activation probability plays an important role (r = 0.69) as well as the node activation threshold (r = -0.60). The independent cascade model showed several significant correlations with the mean diffusion: *Average path length* (r = -0.34), *Network diameter* (r = -0.38), *Average node closeness* (r = -0.39), *Average node eigenvector* (r = -0.38), *Network density* (r = -0.31). Using the linear threshold model, the following significant correlations with the mean diffusion could be found: *Average node betweenness* (r = -0.37), *Network density* (r = -0.34). For both models we can state that a dense, highly connected network leads to higher diffusion rates. It is further important, that the network shows a small spatial extension. As mentioned in the introduction, we know that contemporary online social networks like Fakebook show very small average path lengths and a small diameter, although those networks contain hundreds of millions of nodes. This directly relates to the dynamics of small world networks of Watts and Strogatz, showing that diseases (in our context information) spread more easily in networks characterized as highly clustered yet having small path lengths [14].

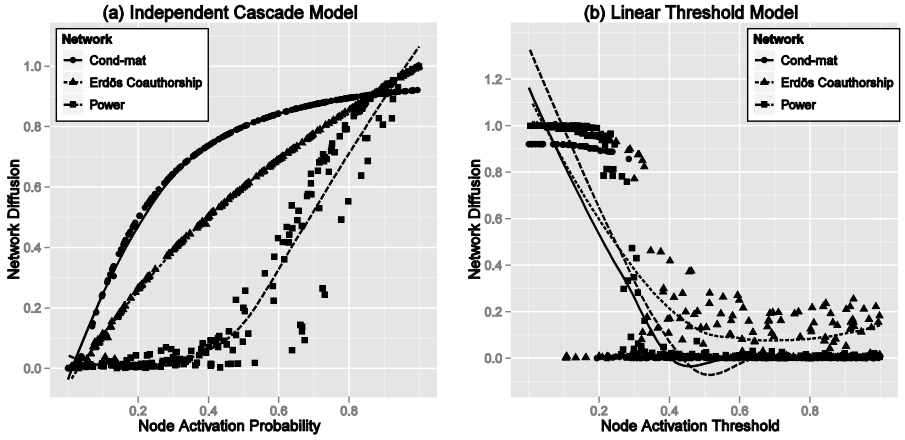

Fig. 2. Diffusion curves of three real world networks given the node activation probability (Independent Cascade Model) and the node activation threshold (Linear Threshold Model). The figure includes regression lines.

In many real world scenarios, marketers will not have a matching dataset to run simulations and create scenarios. Therefore artificially generated networks might be used, as we did in this paper. Since this is a convenient alternative, we were interested if generated networks show a significant effect on the diffusion. An ANOVA on the effect of the mean diffusion regarding an underlying real world or generated network was conducted. The ANOVA revealed a significant mean difference for both diffusion models: $F(1, 7198) = 84.56$, $p < .001$, $\omega = .01$) for independent cascade model, $F(1, 7198) = 14.48$, $p < .001$, $\omega = .002$) for linear threshold model. Although a significant difference exists, the effects are very low. Effect sizes like Kirk's ω or Cohens' d should always be calculated in order to determine the substantial effect. The gener-

ated networks tend to a slight overestimation of the diffusion. Marketers should bear this in mind when using generated networks.

4.2 Influence of Startnode Criterion - Independent Cascade Model

Figure 3 shows the estimated density from a kernel density estimator. The figure highlights a rather minor effect of this parameter on the diffusion. A conducted ANOVA showed a significant effect of the startnode criterion on the diffusion, $F(6, 7192) = 2.62$, $p < .01$, $\omega = .001$). Pagerank showed the highest overall diffusion mean of 0.7 ($\sigma = .34$) and betweenness ($\sigma = .36$) and eigenvector ($\sigma = .36$) the lowest mean with 0.65. However, the means hardly differ and the effects are very small. Moreover, as Figure 1, already showed, the standard deviation is very high. Only the difference between pagerank and betweenness as well as pagerank and eigenvector centrality showed small effects according to Cohen's *d*. Interestingly, the use of random startnodes vs. any other criterion showed very low effects to almost no effects. We can state that for the independent cascade model, the startnode criterion seems to have a minor importance having the pagerank criterion with the highest values.

4.3 Influence of Startnode Criterion - Linear Threshold Model

As seen in Figure 3, the selection of a startnode criterion has a greater effect on the linear threshold model diffusion. An ANOVA also proves the visual impression with a significant effect and higher effect values, $F(6, 7193) = 27.26$, $p < .001$, $\omega = .021$). Still the overall effect is small to medium according to Kirk's ω. *Betweenness* showed the highest diffusion mean with 0.42 ($\sigma = .44$) followed by *closeness*, *degree* and *PageRank* with 0.39 ($\sigma = .44, .43, 43$). *Clustering* and *random* showed the lowest diffusion means with 0.25 and 0.26 ($\sigma = .39, .40$). Looking at pairs of groups we can observe differences between *clustering* and *betweenness* (Cohen's $d = -.41$), *random* and *betweenness* (Cohen's $d = -.38$), *clustering* and *closeness* (Cohen's $d = -.33$), *random* and *closeness* (Cohen's $d = -.30$), *degree* and *clustering* (Cohen's $d = -.33$), *eigenvector* and *clustering* (Cohen's $d = -.26$), *pagerank* and *clustering* (Cohen's $d = -.35$), *random* and *degree* (Cohen's $d = -.30$) and *random* and *pagarank* (Cohen's $d = -.32$). We can thus state that start startnode criterion has a larger effect using the linear threshold model. Here the *betweenness* centrality showed the highest overall diffusion mean and random performed significantly lower. The overall effect is, however, rather small. Furthermore, the standard deviation in every group was very high showing a high degree of dispersion.

To get some deeper insights on the effect of the startnode criterion, we calculated an ANOVA and the effect sizes for every network dataset separately. The results show a very diverse picture. For some datasets the startnode criterion using the independent cascade model had no effect at all (dolphins network, $\omega = 0.003$), whereas other networks showed a high sensitivity to this parameter (hep-th, $\omega = 0.10$). Even more evident is the picture using the linear threshold model. Again some networks showed no effect at all and others showed a very high effect (Watts-Strogatz Network 7, $\omega = 0.46$). We can conclude that the effect of the startnodes strongly differs from

network to network. To answer the question what characteristics may be related to this behavior, we calculated correlations of all network characteristics with the omega squared ω effect size a network showed on the selection of the startnodes. The results revealed that the *Number of components* and the *Average clustering coefficient* are positively correlated with the effect (r = 0.29 and r = 0.37) and *Degree distribution power law fit* is negatively correlated with the effect (r = -0.38) for the independent cascade model. This leads to the conclusion that highly clustered and fragmented networks show a high sensitivity to the selection of the initial start nodes. Concerning the linear threshold model, the results differ. Here we can observe a significant correlation with the *Number of network communities* (r = 0.50). Again, the conclusion is similar. If a network is highly fragmented, the selection of initial start nodes is more important.

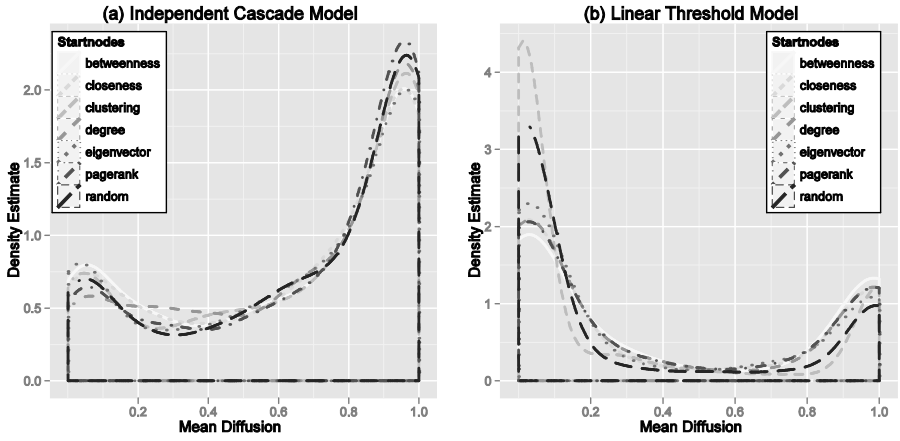

Fig. 3. Estimated densities of the mean diffusion grouped by the startnode criterion.

4.4 Influence of the Number of Initial Seeding Nodes

To prove whether the number of initial nodes has a significant influence on the resulting diffusion we calculated the overall correlation between the mean diffusion of a simulation run and the number of initial nodes used. We further calculated the same correlations for every network separately. Over all networks we can observe small but significant correlations between the mean diffusion and the number of initial start nodes, r = 0.05 for independent cascade model, r = 0.15 for the linear threshold model. Looking at the correlations for every network dataset separately, we can again state a very diverse picture. Using the independent cascade model, the correlations reach from no correlation at all to high significant correlations of r = 0.40 (Barabási-Albert 1). Similarly the linear threshold model shows even higher significant correlations up to r = 0.59 (Dolphins Network). To draw further conclusions we correlated the effect (r values) of the initial number of start nodes with the network characteristics of all networks. The results show that for the independent cascade model the *Average clustering coefficient* (r = 0.48), *Average node betweenness* (r = 0.74) and *Network densi-*

ty ($r = 0.68$) show significant correlations. The findings indicate that especially dense networks show a high effect regarding the number of initial start nodes. Marketers should therefore pay attention to the underlying density. Most real world networks, however, show a very low density. For those applications this parameter is of less importance.

5 Network Diffusion Prediction

To exploit the findings for a marketer's decision making we conducted a classification analysis. Goal of the analysis is to investigate if the diffusion may be predicted if network characteristics (Diameter, Number of Nodes, Density, etc.) and diffusion parameters (Number of start nodes, startnode criterion, Assumed activation probability, etc.) are given. This gives a marketer the possibility to predict a diffusion (the success of a campaign) based on assumptions about the general network characteristics. A marketer can therefore easily create worst-case, best-case and intermediate scenarios to manage possible campaigns. On this account a C4.5 decision tree was created and evaluated using a 10 fold cross-validation. Since the decision tree only classifies categorical variables, we created a 10 bin class variable from the Mean Diffusion variable in the original dataset. This leads to ten diffusion classes, 1 for very low and 10 for very high diffusion. The decision tree is supposed to predict the correct diffusion class based on network characteristics given. As seen in Table 1, the first decision tree achieved notable results with an overall accuracy of 86.012% for the independent cascade model. False predictions are scattered close to the real class, especially for low and high diffusions.

Table 2. Confusion matrix of decision tree classification (independent cascade model).

Confusion Matrix N=7200										
Classified as										
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Real class**
832	67	14	2	3	1	2	4	0	0	**1** n=925
68	**200**	42	7	2	0	0	0	0	0	**2** n=319
10	44	**176**	37	12	2	0	2	0	0	**3** n=283
3	7	34	**163**	47	6	2	2	1	0	**4** n=265
3	1	3	39	**206**	25	13	4	2	1	**5** n=297
0	0	1	2	39	**264**	41	6	5	1	**6** n=359
0	0	1	2	10	38	**341**	57	9	0	**7** n=458
0	0	1	1	1	2	46	**478**	57	0	**8** n=586
0	0	0	1	0	2	8	45	**755**	62	**9** n=873
0	0	0	0	0	0	0	5	52	**2778**	**10** n=2835

The results of the linear threshold model show similar values and an overall prediction accuracy of 85.847%.

6 Conclusions

6.1 Practical Implications

The results of our analysis highlight the impact of the network structure on the diffusion process. This should be evaluated carefully when viral marketing is considered as a marketing tool. Dense networks and networks showing small average path lengths are very efficient for information diffusion whereas highly clustered networks are rather disadvantageous. If the network structure and the diffusion process are unknown, assumptions about the diffusion model need to be drawn since the diffusion varied considerably from the independent cascade to the linear threshold model. Bampo et al. showed that the network structure and the transmission behavior can be estimated during the first generations of a running campaign [5]. A campaign manager could use the estimated artificial network to forecast the campaign, considering the slight diffusion overestimation of such networks, as we have shown. We can further state that seeding matters, but strongly differs depending on the underlying diffusion model and the network structure. If the underlying network is fragmented and clustered, the seeding method is more important. This has already been indicated by Shakarian and Paulo who found that simulated diffusions in highly clustered networks required more initial seed nodes to reach a certain diffusion level [10]. The results also highlight that a campaign manager should rather concentrate on a strong growth than on massive seeding, since the number of initial seeds showed a minor effect.

6.2 Future Work

Our results also highlight that viral marketing is not a panacea to today's marketing. This directly confirms the findings of Leskovec et al. [44]. If the network structure is very disadvantageous, a campaign needs very high activation probability (or very low thresholds) to spread into the whole population. This might be unrealistic. Nevertheless, we can expect that virtual online networks will show even smaller average path lengths and a higher density in the future. Moreover, if currently still independent networks like Twitter and Facebook become more and more connected through aggregator services and mobile devices, the clustering is reduced, yet leading to a greater potential of electronic WOM, and therefore, viral marketing. Furthermore, the diffusion models used in this work might not be fully suited to describe spreading behavior. The behavioral characteristics of single nodes have not been taken into account. This should be addressed in the future to make the models more realistic. We further propose to conduct similar analyses on complex networks comprising different types of relations and multiple layers, as these networks might provide a more holistic reflection of the online communication.

References

1. Hui, K.L., Lee, S.-Y.T., Png, I.P.L.: Consumer Privacy and Marketing Avoidance: A Static Model. Management Science 54 (6), 1094–1103 (2008)Solomon, M.R.: Consumer Behavior. Pearson Education (2006)Porter, L., Golan, G.J.: From Subservient Chickens to Brawny Men: A Comparison of Viral Advertising to Television Advertising. Journal of Interactive Advertising 6 (2), 26–33 (2006)
4. Sattelberger, F.: Erfolgsprognose bei Produktneueinführungen: Eine Untersuchung unter besonderer Berücksichtigung von Word-of-Mouth-Effekten. LIT, Münster (2010)
5. Bampo, M., Ewing, M.T., Mather, D.R., Stewart, D., Wallace, M.: The Effects of the Social Structure of Digital Networks on Viral Marketing Performance. Information Systems Research 19 (3), 273–290 (2008)
6. Hinz, O., Skiera, B., Barrot, C., Becker, J.: Seeding Strategies for Viral Marketing: An Empirical Comparison. Journal of Marketing 75 (6), 55–71 (2011)
7. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137-146. ACM Press, New York, USA (2003)
8. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Finding Influential Nodes in a Social Network from Information Diffusion Data. In: Liu, H. et al. (eds.): Social Computing and Behavioral Modeling. Springer US (2009)
9. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, pp. 1371–1376. AAAI Press (2007)
10. Shakarian, P., Paulo, D.: Large Social Networks can be Targeted for Viral Marketing with Small Seed Sets. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1–8. IEEE (2012)
11. Thompson, B.C.: Is the Tipping Point Toast?. California Management Review 43, 44–63 (2008)
12. GfK Marktforschung GmbH, Bereich Online Research: Alternative Werbeformen, Zentrale Ergebnisse, Welle 3, http://webguerillas.com/media/press/downloads/pdf/gfkstudie_2007.pdf
13. Leskovec, J., Adamic, L.A.: The Dynamics of Viral Marketing. TWEB 1, 1–46 (2008)
14. Watts, D.J., Strogatz, S.H.: Collective dynamics of "small-world" networks. Nature. 393 (4), 440–442 (1998)
15. Barabási, A., Albert, R.: Emergence of Scaling in Random Networks. Science 286 (5439), 509–512 (1999)
16. Sun, T., Chen, W., Liu, Z., Wang, Y., Sun, X., Zhang, M., Lin, C.-Y.: Participation Maximization Based on Social Influence in Online Discussion Forums. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media ICWSM'11 (2011)
17. Tillett, H.E.: Infectious Diseases of Humans: Dynamics and Control. Epidemiology and Infection 108 (1), p. 211 (1992)
18. Rogers, E.M.: Diffusion of Innovations. Free Press (1995)
19. Valente, T.W.: Network Models of the Diffusion of Innovations. Hampton Press (1995)
20. Bass, F.M.: A New Product Growth for Model Consumer Durables. Management Science 15 (5), 215–227 (1969)
21. Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to advance marketing theory development. Academy of Marketing Science Review (9) (2001)
22. Mahajan, V.: New-Product Diffusion Models (International Series in Quantitative Marketing). Springer (2000)

23. Junapudi, V., Udgata, G.K., Udgata, S.K.: Study of diffusion models in an academic social network. In: Janowski, T., Mohanty, H. (eds.): ICDCIT 2010. LNCS, Vol. 5966, pp. 267–278. Springer, Berlin Heidelberg (2010)
24. Kempe, D., Kleinberg, J., Tardos, É.: Influential Nodes in a Diffusion Model for Social Networks. In: ICALP 2005. LNCS, Vol. 3580, pp. 1127 – 1138. Springer (2005)
25. Granovetter, M.: Threshold models of collective behavior. American journal of sociology 83 (6), 1420–1443 (1978)
26. 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 (3), 211-223 (2001)
27. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
28. Newman, M.: Networks: An Introduction. Oxford University Press (2010)
29. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
30. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74 (3), 22 (2006)
31. Newman, M.E.J.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323–351 (2005)
32. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 56, 396–405 (2003)
33. Knuth, D.E.: The Stanford GraphBase: A Platform for Combinatorial Computing. Addison Wesley, Reading (1993)
34. Matei, R., Iamnitchi, A., Foster, P.: Mapping the Gnutella network. IEEE Internet Computing 6 (1), 50–57 (2002)
35. Opsahl, T., Panzarasa, P.: Clustering in weighted networks. Social Networks 31, 155–163 (2009)
36. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74 (3), 22 (2006)
37. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America 98, 404–9 (2001)
38. Guimer, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Physical Review E 68, (2003)
39. Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P., Guimerá, R.: Community analysis in social networks. The European Physical Journal B - Condensed Matter 38, 8 (2004)
40. Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., Arenas, A.: Models of social networks based on social distance attachment. Physical Review E 70 (2004)
41. Erdös, P., Rényi, A.: On random graphs. Publicationes Mathematicae (Debrecen). 6, 290–297 (1959)
42. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal. Complex Sy, 1695 (2006)
43. R Development Core Team: R: A Language and Environment for Statistical Computing, http://www.r-project.org (2010)
44. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Transactions on the Web 1, 5–es (2007)
45. Katz, E., Lazarsfeld, P.: Personal Influence: The Part Played by People in the Flow of Mass Communications. Transaction Publishers (2005)