Impact of the Network Structure on the SIR Model Spreading Phenomena in Online Networks

Marek Opuszko
Department of Business Informatics
Faculty of Economics and Business Administration
Friedrich-Schiller-University of Jena
Jena, Germany 07743
Email: marek.opuszko@uni-jena.de
Telephone: (0049) 3641–943314

Johannes Ruhland
Department of Business Informatics
Faculty of Economics and Business Administration
Friedrich-Schiller-University of Jena
Jena, Germany 07743
Email: j.ruhland@wiwi.uni-jena.de
Telephone: (0049) 3641–943310

***Abstract*—We present an analysis of how the spreading phenomena, which includes the spread of information, innovations, ideas, trends, etc. is influenced by the structure of the underlying (social) network. We conducted spreading simulations using the SIR (Susceptible-Infected-Recovered) model on a large number of real-world and artificially generated network datasets. The results show that the network characteristics have a significant effect on the SIR diffusion. The results reveal that the network structure affects the diffusion in terms of the achieved diffusion, the course of the diffusion and the predictability of the diffusion. We could also show that the effect of other parameters, such as the impact of seeding (network targeting), significantly differs in relationship to the underlying network structure. With these findings, we are able to provide a more differentiated picture on previous contradictory findings, especially in the field of seeding strategies. We further provide useful recommendations for future research to make results more generalizable and comparable.**

Keywords*—*online social networks, information diffusion

I. Introduction

The diffusion phenomena in social networks, such as the Internet or the World Wide Web, has attracted much attention in recent years, both in the research and the industrial sector. Developments like the Web 2.0 and subsequent technologies facilitated a peer-to-peer spread of information besides the traditional one-to-many broadcasters like TV or radio stations. The emergence of virtual social networks on the Internet offers new data and insights into this phenomenon, since it has been challenging to access detailed and large-scale data in the past. New forms of marketing emerged, such as ”viral marketing” as first coined by Steve Jurvetson in 1996 [1]. The metaphorical term *viral* directly leads the origin of the phenomenon in the research on the spread of epidemic diseases [2]. Although research helped to understand the spreading behavior of diseases like H1N1 [3], Severe acute respiratory syndrome (SARS) [4] and influenza A [5], the diffusion of violent topics in social media [6], or the success of product innovations [7], there are still unresolved issues regarding the influencing factors of a diffusion.

In an information diffusion or viral marketing scenario, 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 [8]. Since the seeding strategy, in other words the selection of the first network nodes, is largely under the control of the marketer, this issue received a lot of attention. In particular, the computer science and marketing community addressed this issue [9][10][11][12][13]. Kempe, Kleinberg, and Tardos, for instance, formulated the theoretical approach of the diffusion maximization problem [9]. Still, questions about the impact of seeding and the best seeding strategy exist [10][14]. Due to the previous focus on this issue the third aspect, the structure of the underlying network, only recently received consideration. One reason could be the progress made in social network analysis in the last years. Nevertheless, although efforts have been made, many questions remain open.

The majority of the research is based on investigations on only very few network datasets [9][12][15] and it is unclear if the findings are applicable to any other network. Watts and Strogatz further showed that some types of networks facilitate an epidemic spread [16]. Hence, a deeper understanding is needed, what characteristics of a network influence a diffusion.

In this paper, we examine an information diffusion process based on diffusion simulations using the SIR diffusion model on a large number of both real-world and artificially generated networks. We will introduce several metrics to describe network characteristics and evaluate the relation between the metrics and the diffusion results. This also includes the relation between several possible seeding methods to answer the question whether and how seeding methods interact with different types of networks. The paper is organized as follows. The next sections describes the latest state of research and highlights important research gaps. Section III explains the simulation analysis including the datasets and network characteristics used in this paper. Section IV presents the analysis and the results. Finally, the paper is closed with our conclusion in Section V.

II. Diffusion in Online Networks

The research on the diffusion of information in online networks has mainly derived from the studies of infectious diseases and epidemics [2][17]. Although there has been research on product growth models prior to the groundbreaking work made in the field of epidemics [18], the knowledge gained in the field of information diffusion has enriched research in other contexts, such as the diffusion of (product) innovations [19][7]. Especially the boost of computer networks, the progress made

ISBN: 978-1-61208-283-7

in social network analysis and the success of the Internet have caused a large spread of research on diffusion in networks in various contexts. Of particular interest is the research on electronic word-of-mouth marketing or viral marketing [20][21] and the the research on the effect and the optimization of seeding strategies [22][9][10][23]. The seeding strategy determines an initial set of (network) nodes, usually to maximize the spread of information or to boost the effect of word-of-mouth. However, in the past, contradictory findings have been made in terms of the effectiveness of seeding [10][9][12][14][24]. Aral, Muchnik, and Sundararajan recently argued that conventional wisdom about seeding strategies should be questioned as the effectiveness of seeding might be overestimated if other factors like network characteristics or node similarities (homophily) are not taken into account [24].

In the past, different models have been introduced to theoretically describe the spreading of diseases or information. Beside approaches like threshold models [25] based on the idea of adoption or cascade models [21], the SIR model, and in particular, the network SIR model can be considered as a reference model in terms of information diffusion due to its numerous successful applications [26][27]. The SIR acronym is derived from the three states or conditions a network node can occupy: *susceptible*, *infected* or *recovered*. In the context of information diffusion the *susceptible* state would reflect nodes that have not received a circulating information artifact. The *infected* nodes are nodes which have already received the artifact and are able to spread it further to neighbor nodes, and *recovered* nodes (sometimes also referred to as *removed* or *inactive*) may unsuccessfully receive the circulating information without any consequences. Due to the dichotomous nature of the possession of an information artifact, the peer-to-peer transmission, in other words an infection, can be modeled as a Bernoulli game. The outcome is 1 if the transmission was successful an 0 if not. Hence, there is an expected infection probability p of a single node reflecting the likelihood to infect other nodes.

Past research showed that the characteristics of many (social) networks facilitate an epidemic spread, which also accounts for information diffusion and viral marketing [16]. Bampo, Ewing, Mather, Stewart, and Wallace compared a real-world spreading in a large peer-to-peer network with a simulated SIR modeled spreading on generated random networks [8]. They created random graphs comprising the same topological properties like density and degree distribution as the real observed network. Their findings showed that the social network had a significant impact on the performance of a viral marketing campaign. They further stated that clustered networks are not very efficient and that small-world networks generally temper the spread of messages. 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 [13]. They grouped the networks based on the diffusion results into the groups *highly susceptible*, *susceptible*, and *diffusion hamper*. The results did, however, not provide an explanation what characteristics relate to a higher diffusion susceptibility. A more detailed examination was conducted by Opuszko and Ruhland [28], who investigated independent cascade and linear threshold diffusion simulation and showed that the diffusion strongly depends on the underlying network. They found that the seeding impact is not present in every network, especially not for cascade-like diffusions.

III. Method

To address the questions mentioned in the introduction, a simulation analysis was conducted using a set of 35 different network datasets (see next subsection for details). For every network dataset 200 *runs* of simulations with randomly chosen start parameters were conducted. For every *run* the following parameters were set randomly: the *number of start nodes* was set at random in the interval [1,50], the *seeding criterion* for choosing the start nodes (based on several network centralities, see below) was selected randomly and the nodes have been selected accordingly by calculating the metrics for all network nodes, the *infection probability* for all nodes was set randomly in the interval [0,1]. Every simulation run comprised 50 single simulations with the prior set parameters. This was done in order to assess the variation of a spreading under identical preconditions. We measured three simulation outcomes, the average number of *infected nodes*, in other words the achieved network diffusion, the average *number of circulations* and the *standard deviation of the diffusion* of one simulation run. The number of circulations reflects the discrete time steps a virus or a diffusion was present and circulating before it died out. The results of one simulation run were stored including all used parameters and the network metrics as one single case. All in all, 350,000 simulations were conducted. The resulting dataset comprised 7,000 cases for each simulation run including the network characteristics (see next subsections), the simulation parameters and the results.

A. Network Datasets

TABLE I. Networks used in the analysis

Name	Nodes	Edges
Dolphin social network [29]	62	159
Les Miserables character network [30]	77	254
Power grid network [16]	4,941	6,594
Student Network	471	926
Gnutella 2008 peer-to-peer network [31]	6,301	20,777
OCLinks social network [32]	1,899	20,297
NetScience coauthorship network [33]	1,461	2,742
Internet snapshot [34]	22,963	48,436
Hep-th coauthorship network [35]	7,610	15,751
Cond-mat 2003 coauthorship network [35]	30,460	120,029
Erdös collaboration graph [36]	6,927	11,850
Astrophysics coauthorship network [35]	16,046	121,251
Email messages network [37]	1,133	5,451
Jazz musician network [38]	198	2,742
PGP users giant component [39]	10,680	24,316
Barabási-Albert 1 [40]	60	177
Barabási Albert 2 [40]	80	237
Barabási Albert 3 [40]	1025	1024
Barabási Albert 4 [40]	1,400	2,798
Barabási Albert 5 [40]	5,241	15,714
Barabási Albert 6 [40]	20,000	39,998
Barabási Albert 7 [40]	30,000	119,996
Erdös-Rényi 1 [41]	868	1,040
Erdös-Rényi 2 [41]	914	12,683
Erdös-Rényi 3 [41]	1,000	14,902
Erdös-Rényi 4 [41]	1,000	25,362
Erdös-Rényi 5 [41]	6,290	19,996
Erdös-Rényi 6 [41]	6,917	23,767
Watts-Strogatz 1 [16]	60	177
Watts-Strogatz 2 [16]	80	240
Watts-Strogatz 3 [16]	60	177
Watts-Strogatz 4 [16]	1,400	2,800
Watts-Strogatz 5 [16]	6,927	14,994
Watts-Strogatz 6 [16]	20,000	40,000
Watts-Strogatz 7 [16]	30,000	120,000

Both real-world and artificially generated networks have been used as a source for the simulations. Table I shows the network datasets used for the SIR simulations. 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.

Three state-of-the-art algorithms have been used to generate the artificial networks: Erdös-Renyi game [41], Watts-Strogatz game [16] and Barabási-Albert game [40]. 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].

B. Network Characteristics

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 connections between the nodes.
- *Network density*, 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 that are sparsely connected to other groups. In contrast to components, these groups are not isolated from each other. We used the leading eigenvector community detection algorithm according to Newman [36].
- *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 [44] against the degree distribution of each network.

C. Seeding Criteria for the Selection the Seeding Nodes

There are several criteria for selecting the initial *infected* nodes. The most simple is 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 be most influential [45]. Kempe et al. initially formalized this problem as the influence maximization problem [9]. We will evaluate the following centrality measures as a criterion for the selection of the seeding nodes. All centrality metrics have been calculated according to Wasserman and Faust [46] as well as Newman [47]. We refer to those publications for further details on the calculations. It should be mentioned that the presented list of seeding criteria is far from complete. The methods used are based on network centralities and cannot reflect node behavioral aspects like adoption propensities or node homophily.

- *Degree Centrality*, one of the most common centrality measures. Degree centrality reflects the number of ties (also known as 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 the inverse farness of a node 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 [40]. 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.

IV. Analysis

A. Descriptive analysis

Fig. 1 shows the resulting SIR diffusion means of all networks in relation to the infection probability and boxplots for five different infection probability intervals. The plot highlights

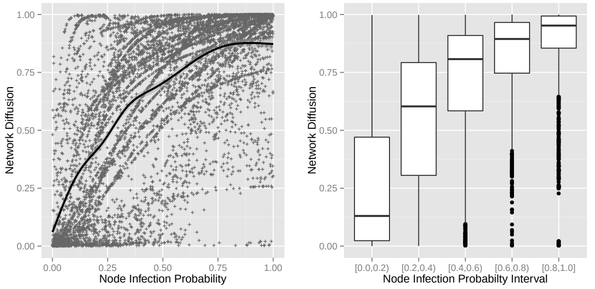

Fig. 1. Left Plot: Simulated diffusion of all networks depending on the infection probability of the network nodes. The figure includes a regression line (solid line). Right Plot: Boxplots of the network diffusion of different infection probability intervals

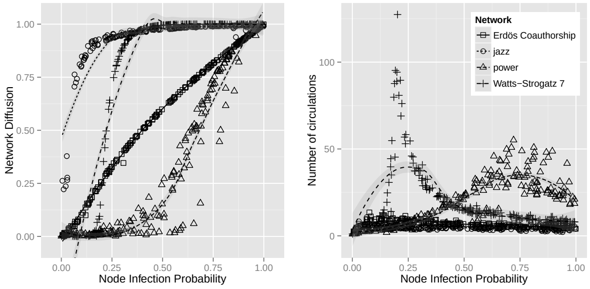

Fig. 2. Diffusion (left) and number of circulations (right) curves of four real-world networks given the infection probability. The figure includes smoothed regression lines.

a high variance in the diffusion depending on the underlying network. Interestingly, the standard deviation of the network diffusion for all simulation runs (comprising 50 individual simulations using a fixed parameter setting) is only 0.022, showing that within a similar parameter setting, the diffusion is quite foreseeable.

Aside from very high values of the infection probability, we can observe that the diffusion percentage spreads over the whole range from almost no diffusion to nearly total diffusion, especially in the middle and lower area of the infection probability. The boxplots in Fig. 1 depict the high variation and uncertainty in the diffusion. In the node infection probability interval of $[0.2, 0.4)$ the standard deviation is 0.306 having a mean of 0.543. When the networks are examined separately, the picture becomes more clear as seen in Fig. 2. Here we plotted the diffusion results and the number of circulations versus the node infection probability of four exemplary networks. As we can see, the curves significantly differ. As some diffusion curves seem to follow the typical S-shape (*power network*), some networks show an almost linear relation to the node infection probability (*Erdös Coauthorship Network*). We can further state a dramatic difference between some networks, especially in the node infection probability interval from 0.4 to 0.5. Comparing the *power* and the *jazz network* at a node infection probability of 0.4, we can observe a difference in the diffusion of almost 80%, having network diffusions around the mean of 0.08 for the *power network* and diffusions around the mean of 0.95 for the *jazz network*. This means that for identical SIR model settings, one network will usually reach diffusions of not even 10% of its network nodes, having other networks where almost the whole population (95%) is reached.

Similar results are highlighted by the graph showing the distribution of the number of circulations taken by a simulation. We can see characteristic differences for different networks, showing that some networks never reach the maximum number of circulations of 14 (*Erdös Coauthorship Network*), whereas other networks show a more than nine times higher maximum diffusion circulations of 127 (*Watt-Strogatz Network 7*). Moreover the plot shows that the maximum number of circulations is reached for totally different values of the node infection probability. The *Watt-Strogatz Network 7* reaches the maximum at avlues around 0.2 and the *power network* reaches the highest values at a node infection probability at around 0.75.

A first conclusion to draw from the figures is that the networks show tremendous variation in their diffusion behavior, both in terms of achieved diffusion (infected nodes) and the circulation time. To evaluate the influence of the network parameters on the diffusions we calculated the correlation between those parameters and the mean diffusion. The results revealed several significant correlations. It should be noted that all correlations reported hereinafter showed a p-Value of < 0.001, if not stated otherwise. Obviously, the infection probability plays an important role ($r = 0.644$). Furthermore, the *average closeness* ($r = 0.323$) had a moderate significant effect. The *network density* ($r = 0.264$), the *average eigenvector* ($r = 0.263$) and the *average degree* ($r = 0.230$) of a network correlated significantly with the network diffusion showing weak relationships. On the other hand, the network *diameter* ($r = -0.342$) and the *average path length* ($r = -0.302$) of the network showed moderate negative relationships.

The picture is similar regarding the number of circulations of a diffusion. We can identify various strong positive relationships. The *average path length* ($r = 0.588$) and the *diameter* ($r = 0.561$) of a network correlated with the number of circulations. The *average closeness* ($r = -0.421$) showed a strong, the *density* ($r = -0.317$) a moderate, the *average eigenvector* ($r = -0.260$) and the *average betweenness* ($r = -0.226$) showed weak negative relationships with the number of circulations.

When evaluating how strong the diffusion varies within one simulation run, represented by a diffusion run's standard deviation, we can observe moderate relationships with the *average betweenness* ($r = 0.343$) and the *density* ($r = 0.315$) of a network. This indicates that networks with high values for these metrics are usually less predictable in terms of the network diffusion for a given fixed parameter setting.

Interestingly, the network size and the number of edges did not have a great effect on the diffusion. Highly connected and dense networks seem to lead to higher diffusion rates at comparable diffusion parameters. Another important factor appears to be the spatial dimension, represented by the average path length between two arbitrary nodes of the network and the network diameter. This directly relates to the *small world* property of networks, first discovered by Watts and Strogatz [16]. Small world networks are characterized by highly clustered nodes having small average path lengths. Many contemporary online social networks like Facebook show very small average path lengths and a small diameter. This is interesting as these networks may contain hundreds of millions of nodes. Since Milgram's first estimate of 6 degrees (edges) between any two people in the world [48], this number is estimated to be 3.74 in the Facebook online social network in 2011, comprising, at that time, over 721 million users [49]. We can expect shorter path lengths and shorter spatial dimensions of networks in the future. According to the results presented, we can expect higher diffusions under the same preconditions for future networks.

B. *Effect of generated random networks*

When conducting experiments, researchers or marketing campaigners sometimes make use of generated random networks to examine different spreading scenarios. We were interested whether these networks might show a significant difference in the diffusion compared to real-world networks. 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: $F(1, 6998) = 28.28, p < .001, \omega^2 = .003, d = 0.12)$ for the groups *real-world* ($\mu = 0.62, \sigma = 0.32$) and *random generated* ($\mu = 0.67, sd = 0.35$) networks. Another ANOVA of the effect on the circulation length in real-world and generated networks showed a significant difference: $F(1, 6998) = 10.712, p < .001, \omega^2 = .001, d = 0.07)$ for the groups *real-world* ($\mu = 9.28, \sigma = 6.38$) and *random generated* ($\mu = 10.19, sd = 14.14$) networks. Although the difference is significant, the effect according to ω^2 [50] or Cohen's d [51] effect size is rather minor. We can conclude that random networks tend to a slight overestimation in the diffusion and a small overestimation in the circulation length.

C. *Influence of the Number of Seeding Nodes*

To show whether the number of seeding 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 seeding nodes used. We can state that, although significant ($p < .001$), there is no overall effect ($r = 0.07$) of the number of seeding nodes on the diffusion.

When calculating the correlation for every network separately, the picture changes dramatically. We can observe networks that show a strong effect regarding the number of start nodes: *dolphins network* ($r = 0.45, p < .001$), *Watts-Strogatz network 1* ($r = 0.34, p < .001$). A first interpretation would be that small networks, in terms of the number of nodes, show a high correlation, as would be expected. Interestingly, there are other networks with only a few number of nodes that do not show this effect, e.g. the *jazz network* ($r = -0.03$). On the other hand, there are also quite big networks that show an effect like the *netscience network* ($r = 0.29, p < .001$).

Since the effect differs so strongly from network to network, we were interested if there might be a metric that could explain this behavior. Therefore we calculated several regressions of the effect with the network metrics. We could identify two significant relationships. The *average betweenness* ($p < 0.001, \beta = 0.78, R^2 = 0.59$) and the network *density* ($p < 0.01, \beta = 0.52, R^2 = 0.25$) of a network showed a strong relationship with the effect of the number of seeding nodes. Hence, if the network is characterized by bridging nodes with a high betweenness and a high density, this parameter is of higher importance.

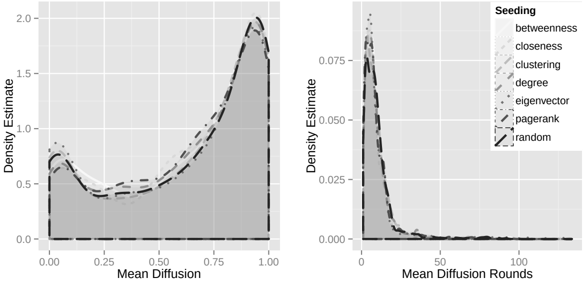

Fig. 3. Estimated densities of the mean diffusion and the number of circulations grouped by the seeding criterion

D. *Influence of Seeding Criterion*

Fig. 3 shows the estimated density from a kernel density estimator for the achieved diffusion and the number of circulations grouped by the seeding criterion. Interestingly, the figure highlights a rather minor effect of this parameter on the diffusion and the number of circulations. Surprisingly, an ANOVA showed that there was no significant overall effect of the seeding criterion. We should recall at this point that the seeding methods used in this paper are not a complete list of possible metrics. The results, however, confirm some previous findings, for instance Opuszko and Ruhland [28] who found that diffusions based on cascades show a rather minor sensitivity to the seeding method. Nevertheless, we were interested if the effect is network specific so we conducted an ANOVA for each network separately. The results showed that there are in fact some differences among the networks. The *netscience network* ($p < 0.001, \omega^2 = 0.103$), the *Barabasi-Albert network 3* ($p < 0.01, \omega^2 = 0.066$), the *hep-th network* ($p < 0.01, \omega^2 = 0.062$) showed a medium effect. Six other networks showed a weak effect. Still the number is quite low. We further conducted regressions to investigate if the effect is related to a network metric. Only the *number of components* showed a weak significant relationship ($p < 0.05, R^2 = 0.107$) with the effect.

When evaluating the effect of seeding on the number of circulations the overall results are similar as there was no significant relationship. The picture changes, however, if the relationship is calculated for each network separately. ANOVA calculations for every network showed a diverse picture. 19 networks showed at least a weak relationship. Some networks showed a very strong relationship with the effect: *Barabási-Albert 2 network* ($p < 0.001, \omega^2 = 0.303$), *Barabási-Albert 1 network* ($p < 0.001, \omega^2 = 0.201$), *jazz network* ($p < 0.001, \omega^2 = 0.184$). We can state that seeding has a much higher effect on the number of circulations than it had on the diffusion. Again we conducted several regressions to evaluate possible relationships of the effect to network metrics. The results reveal positive significant weak relationships with the *network density* ($p < 0.01, R^2 = 0.206$), the *average clustering coefficient* ($p < 0.05, R^2 = 0.152$) and the *average closeness* ($p < 0.05, R^2 = 0.106$). We can conclude that dense clustered and fragmented networks show a higher sensitivity to seeding in terms of the circulation time.

V. Conclusion

The results presented in this paper lead to two main outcomes: First, we have shown that the underlying network has a tremendous effect on the simulated SIR spreading. We have seen that networks show significantly different behavior under comparable parameters. This also includes the variation of the diffusion. Some networks show significantly more variation and uncertainty in the diffusion than others. As a consequence, we question if results based on one specific network are transferable to any other network. Therefore we recommend for future investigations that authors either present further network characteristics other than the network size, since we have seen that the size is not a profound network metric, or they make the dataset publicly available. In this way other researchers are able to assess the network characteristics. If that is not possible, perhaps due to copyright or secrecy issues, the researchers might present parameters to generate a random network reflecting the examined real-world network's characteristics. As we have shown, the use of random generated networks is suitable for analyses.

The second main insight is that seeding must be seen on a differentiated basis. Besides the fact that the overall effect in the SIR diffusion was rather minor, we could identify significant differences among the networks. This could possibly explain some of the previous contradictory findings [24][10][20]. We can further state that seeding might not affect cascade-like diffusions as much as it effects adoption-like diffusions. Opuszko and Ruhland [28] found in their experiments with linear-threshold models that the effect of seeding is higher in adoption models. However, we have to emphasize the fact that the seeding methods used in this paper are based on network centralities and might omit other node specific factors not related to a centrality metric. Aral and colleagues [24] argued that the omission of these characteristics is a factor in the ongoing disagreement about the effect of seeding. Based on our results we can state that even with a state-of-the-art simulation analysis, the picture is diverse and the results are related to the underlying network characteristics.

The results presented are not tied to the research in information diffusion. They relate to numerous research fields like viral marketing, the research on epidemics, the flow of innovations and ideas in collaboration networks etc.

References

[1] A. M. Kaikati and J. G. Kaikati, "Stealth marketing: How to reach consumers surreptitiously," *California Management Review*, vol. 46, no. 4, pp. 6–23, 2004. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1394975

[2] H. E. Tillett, "Infectious Diseases of Humans; Dynamics and Control," *Epidemiology and Infection*, vol. 108, no. 1, p. 211, 1992. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1753-6405.1992.tb00056.x/abstract

[3] L. Y. L. Yulian, "Investigation of Prediction and Establishment of SIR Model for H1N1 Epidemic Disease," in *Bioinformatics and Biomedical Engineering iCBBE 2010 4th International Conference on*. IEEE, 2010, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5517654&tag=1

[4] Q. Chen, "Application of SIR model in forecasting and analyzing for SARS," *Beijing da xue xue bao Yi xue ban Journal of Peking University Health sciences*, vol. 35 Suppl, no. Suppl, pp. 75–80, 2003.

[5] R. Casagrandi, L. Bolzoni, S. A. Levin, and V. Andreasen, "The sirc model and influenza a," *Mathematical Biosciences*, vol. 200, no. 2, pp. 152 – 169, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0025556405002464

[6] J. Woo, J. Son, and H. Chen, "An SIR model for violent topic diffusion in social media," in *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2011, pp. 15–19.

[7] V. Mahajan, *New-Product Diffusion Models (International Series in Quantitative Marketing)*. Springer, 2000.

[8] M. Bampo, M. T. Ewing, D. R. Mather, D. Stewart, and M. Wallace, "The Effects of the Social Structure of Digital Networks on Viral Marketing Performance," *Information Systems Research*, vol. 19, no. 3, pp. 273–290, 2008. [Online]. Available: http://isr.journal.informs.org/cgi/doi/10.1287/isre.1070.0152

[9] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. New York, New York, USA: ACM Press, Aug. 2003, p. 137. [Online]. Available: http://dl.acm.org/citation.cfm?id=956750.956769

[10] O. Hinz, B. Skiera, C. Barrot, and J. Becker, "Seeding Strategies for Viral Marketing: An Empirical Comparison," *Journal of Marketing*, vol. 75, pp. 55–71, 2011.

[11] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, Jul. 2007, pp. 1371–1376. [Online]. Available: http://dl.acm.org/citation.cfm?id=1619797.1619865

[12] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Finding Influential Nodes in a Social Network from Information Diffusion Data," *Social Computing and Behavioral Modeling*, pp. 1–8, 2009. [Online]. Available: http://www.springerlink.com/content/p54n387114265075/

[13] P. Shakarian and D. Paulo, "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)*, no. 2, 2012, pp. 1–8. [Online]. Available: http://arxiv.org/abs/1205.4431

[14] B. C. Thompson, "Is the Tipping Point Toast ?" *California Management Review*, vol. 43, no. 4, pp. 44–63, 2008. [Online]. Available: http://m6d.com/wp-content/themes/m6d/documents/is-the-tipping-point-toast.pdf

[15] T. Sun, W. Chen, Z. Liu, Y. Wang, X. Sun, M. Zhang, and C.-Y. Lin, "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.

[16] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–2, Jun. 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[17] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control (Oxford Science Publications)*. Oxford University Press, USA, 1992.

[18] F. M. Bass, "A New Product Growth for Model Consumer Durables," *Management Science*, vol. 15, no. 5, pp. 215–227, Dec. 1969. [Online]. Available: http://dl.acm.org/citation.cfm?id=1245920.1245934

[19] E. M. Rogers, *Diffusion of Innovations*. Free Press, 1995.

[20] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web*, vol. 1, no. 1, pp. 5–es, May 2007. [Online]. Available: http://arxiv.org/abs/physics/0509039

[21] J. Goldenberg, B. Libai, and Muller, "Using complex systems analysis to advance marketing theory development," *Academy of Marketing Science Review*, vol. 9, 2001.

[22] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*. New York, New York, USA: ACM Press, Aug. 2001, pp. 57–66. [Online]. Available: http://dl.acm.org/citation.cfm?id=502512.502525

[23] J. Yang, C. Yao, W. Ma, and G. Chen, "A study of the spreading scheme for viral marketing based on a complex network

model," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 4, pp. 859–870, Feb. 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0378437109009042

[24] S. Aral, L. Muchnik, and A. Sundararajan, "Engineering Social Contagions: Optimal Network Seeding in the Presence of Homophily ," *Forthcoming in Network Science*, February 18 2013. [Online]. Available: http://dx.doi.org/10.2139/ssrn.1770982

[25] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978. [Online]. Available: http://www.jstor.org/stable/10.2307/2778111

[26] C. Moore and M. Newman, "Epidemics and percolation in small-world networks," *Physical Review E*, vol. 61, no. 5, pp. 5678–5682, May 2000. [Online]. Available: http://arxiv.org/abs/cond-mat/9911492/

[27] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[28] M. Opuszko and J. Ruhland, "Effects of the network structure on the dynamics of viral marketing," in *Proceedings of the 11th International Conference on Wirtschaftsinformatik. Internationale Tagung Wirtschaftsinformatik (WI-2013), 11. February 27 - March 1, Leipzig, Germany*, R. Alt and B. Franczyk, Eds. Leipzig, Germany: Universitt Leipzig, Leipzig, 2 2013, pp. 1509–1524.

[29] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 56, pp. 396–405, 2003.

[30] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*. Reading: Addison Wesley.

[31] R. Matei, A. Iamnitchi, and P. Foster, "Mapping the Gnutella network," *IEEE Internet Computing*, vol. 6, no. 1, pp. 50–57, 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=613352.613670

[32] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, May 2009. [Online]. Available: http://dx.doi.org/10.1016/j.socnet.2009.02.002

[33] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, p. 22, Sep. 2006. [Online]. Available: http://arxiv.org/abs/physics/0605087

[34] M. Newman, "Network data," [retrieved: 11 2012], http://www-personal.umich.edu/~mejn/netdata/, 2011.

[35] M. E. Newman, "The structure of scientific collaboration networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–9, Jan. 2001. [Online]. Available: http://arxiv.org/abs/cond-mat/0007214/

[36] V. Batagelj and A. Mrvar, "Pajek datasets," [retrieved: 11 2012], http://vlado.fmf.uni-lj.si/pub/networks/data/, 2006.

[37] R. Guimer, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review*, vol. E 68, no. 065103, 2003.

[38] A. Arenas, L. Danon, A. Díaz-Guilera, P. Gleiser, and R. Guimerá, "Community analysis in social networks," *The European Physical Journal B - Condensed Matter*, vol. 38, no. 2, p. 8, 2004.

[39] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical Review E*, vol. 70, no. 5, Nov. 2004.

[40] A. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. [Online]. Available: http://www.sciencemag.org/content/286/5439/509.abstract

[41] P. Erdös and A. Rényi, "On random graphs," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959. [Online]. Available: http://www.renyi.hu/~p_erdos/Erdos.html\#1959-11

[42] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Sy, p. 1695, 2006. [Online]. Available: [retrieved:112012],http://igraph.sf.net

[43] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010. [Online]. Available: [retrieved:112012]http://www.r-project.org

[44] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, May 2005. [Online]. Available: http://dx.doi.org/10.1080/00107510500052444

[45] V. Junapudi, G. K. Udgata, and S. K. Udgata, "Study of diffusion models in an academic social network," in *Distributed Computing and Internet Technology*, ser. Lecture Notes in Computer Science, T. Janowski and H. Mohanty, Eds., vol. 5966. Berlin, Heidelberg: Springer Berlin Heidelberg, Feb. 2010, pp. 267–278. [Online]. Available: http://dl.acm.org/citation.cfm?id=2127870.2127905

[46] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, ser. Structural analysis in the social sciences, 8, M. Granovetter, Ed. Cambridge University Press, 1994, vol. 8, no. 1.

[47] M. Newman, *Networks: An Introduction*, M. E. J. Newman, Ed. Oxford University Press, 2010.

[48] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 60–67, 1967. [Online]. Available: http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf

[49] B. B. C. BBC, "Facebook users average 3.74 degrees of separation," BBC [retrieved: 05 2013], http://www.bbc.co.uk/news/technology-15844230, 2011.

[50] S. Olejnik and J. Algina, "Generalized eta and omega squared statistics: measures of effect size for some common research designs." *Psychological Methods*, vol. 8, no. 4, pp. 434–447, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14664681

[51] J. Cohen, "A power primer." *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159, 1992. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3153748&tool=pmcentrez&rendertype=abstract