Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lecture with Computer Exercises:
Modelling and Simulating Social Systems with MATLAB

Project Report

Diffusion of Innovation and the Characteristics of Seeds

Sebastian Lechner, Adrian Oesch & Amrollah Seifoddini

Zurich
December 2014

Agreement for free-download

We hereby agree to make our source code for this project freely available for download from the web pages of the SOMS chair. Furthermore, we assure that all source code is written by ourselves and is not violating any copyright restrictions.

Sebastian Lechner Adrian Oesch Amrollah Seifoddini

Individual Contributions

Everybody was involved with everything but each of us focused on different parts more then others. Sebastian Lechner: writing, literature review. Adrian Oesch: coding, statistics. Amrollah Seifoddini: coding, networks.

Contents

1 Abstract

By applying a previously established innovation diffusion model to an agent-based simulation this paper attempts to find important seed characteristics for successful diffusion of innovation in a scale-free network. We were investigating node degree, eigenvector centrality, betweenness centrality, local clustering and added normally distributed variance through a personal preference value to each node. In addition we were also analyzing the predictive power of the seeds' neighbors. The main results suggest that degree, eigenvector centrality and betweenness centrality of the seeds are significant positively related to innovation diffusion success. The local clustering of the seeds seems to be negatively related to innovation diffusion success.

2 Introduction and Motivations

2.1 Research Objective

Very little is known about the programmability of innovation success. Innovation diffusion was introduced as research field in the 1960s by the work of Fourt and Woodlock (1960)[1], Rogers (1962)[2], and others (see review by Maede and Islam (2006)[3]). Since then the main body of research was contributed by marketing research, business administration, and social sciences. Its focus has been primarily on the *social* behavior of agents which adopt an innovation and on the stage of time when they do it.

The newly emerged science of networks has shed little light so far on the network topologies of successful innovation processes. To our knowledge very few factors are known to positively affect the diffusion of an innovation from a node to the neighboring nodes and the entire network. These factors include the degree of a node[4], its the eigenvector centrality[4], and the number neighboring nodes which are easily influenced[5]. Strikingly, there is no information in the literature on characteristic centrality measures of successful seeds, where seeds are the first agents in a network that carry an idea or innovation.

Consequently, *we want to know which network parameters can explain the difference between successful and non-successful seeds of innovation and why.* The possible applications are numerous: An understanding of the programmability of innovation success and idea spreading can help to increase the (initial) speed with which news messages are delivered to the public, to allow technological changes to occur more quickly, and to optimize marketing strategies for novel products. We will focus our work on the degree, betweenness centrality, eigenvector centrality, and local clustering coefficient, which, intuitively, should positively effect seeds in their attempt to penetrate the whole network with an innovation. Additionally, we want to take a look at the neighboring nodes of the seeds and investigate how these characteristics influence the success of innovation diffusion.

2.2 Conceptional Background

Networks of dynamical systems have been used to model a plethora of processes: Gene transcription and protein-protein interactions[6], evolutionary dynamics of language development[7], cooperation[8], and cancer[9], social systems[10], and others.

In one particular network problem scientists are trying to understand how ideas and technological innovations penetrate social systems once they emerge. Basically, one distinguishes two models of idea diffusion, a discrete and a continuous model. In social contagion models these two are closely related as the cumulative probability of the uptake of an idea shows a threshold-like behavior (see Fig. 1), which can be discretized easily.

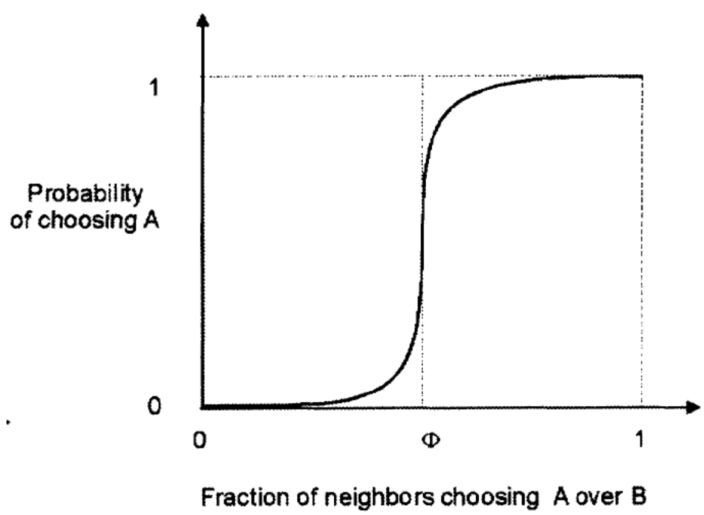

Figure 1: Models of social contagion show a sigmoidal behavior of the cumulative probability instead of a linear behavior as SIR-type models for example do. Figure was adapted from Watts (2004)[11].

In a discrete *threshold* model a node takes up an idea if the total utility gain u exceeds a certain discrete threshold θ. This model is used in our project. In another type of model the ideas diffuse according to a continuous function. Here, usually existing models from epistemological studies are applied[12]. Most epistemological models are based on the SIR model[13], the canonical model of biological contagion, where agents are classified as being either susceptible (S), infected (I), or as having recovered (R) from an infection. Importantly, SIR-type models are memory-less, *i.e.* that the probability of getting infected does *not* depend on the number of already infected agents. However, social contagion models usually assume that the diffusion of innovation is affected by the memory of past events.

3 Description of the Model

Our proposed threshold model for the diffusion of innovation is based on the model described by McCullen and co-workers (2013)[14]. It works in the following way: Assuming that the current state of innovation x_i of a node i is 0 if it has not (yet) accepted the innovation and 1 otherwise, one can define the following rule for the synchronous update of the state: The state of a node i at time $t+1$ is given by

$$x_i(t+1) = \begin{cases} 1 \text{ if } x_i(t) = 1 \\ 1 \text{ if } x_i(t) = 0 \text{ and } u_i(t) > \theta \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

where θ is the threshold and t is time. The utility u_i is composed of three components, the *personal* preference p_i, the *neighborhood* utility s_i, and the global *network* utility m_i, and is defined by

$$u_i(t) = \alpha p_i + \beta s_i + \gamma m_i \ , \tag{2}$$

where the weighting factors are $\alpha + \beta + \gamma = 1$. In our model, p_i is randomly drawn from a normal distribution with mean of 0.5 and a standard deviation of 0.1. s_i is the mean state of the neighboring nodes. So the more neighboring nodes have state 1 the higher is the neighborhood utility of a node, which in turn makes it more likely for node i to accept the innovation. m_i is the mean state of all nodes in the network. The mean state of the neighboring nodes is computed by

$$s_i(t) = \frac{1}{k_i} \sum_j^N A_{ij} x_j(t) \ , \tag{3}$$

which k_i being the degree of node i, A_{ij} the adjacency matrix, and x_j the innovation state of neighbors j. The global network utility is calculated by averaging over all nodes N,

$$m_i = \frac{1}{N} \sum_i^N x_i(t) \ . \tag{4}$$

First, we applied our algorithm on the "small world"-model by Watts and Strogatz[15] since it is one of the models used by McCullen and co-workers in their paper. However, we quickly realized that this model is not suitable for our analysis of seed characteristics since its degree distribution is either a delta function, a Poisson distribution, or something in between, depending whether the rewiring probability β is 0, 1, or something in between. The degree distribution is probably the most fundamental centrality measure. If it does not show a broad distribution, we are not able to observe a value of the degree specific for seeds.

Thus, we applied our model for the diffusion of innovation on an undirected scale-free network with preferential attachment as proposed by Barabasi and Albert[16]. Here, the degree distribution follows a power-law and the degree of the nodes is more heterogeneous and can reach very high values.

In our model we set weighting factors to $\alpha = 0.3$, $\beta = 0.6$, and $\gamma = 0.1$. The reason is that we assume that, in reality, people which consider accepting an idea give the most value their social neighborhood (β) and the least value the general opinion of the network (γ).

Of course, in a real setting, very few ideas are successful in penetrating the whole social network. We set the utility threshold (θ) in such a way to obtain a success ratio close to 10%, which allows us to record enough data at a reasonable high success ratio.

Although McCullen and co-workers used 5% of the nodes as seeds, we decided to use only one seed, since it is unlikely that several persons have the very same idea at the same time in a social network. Also because for evaluating the charachteristics of seeds, we need to test all nodes as seed to capture the success factors for those who managed to spread their value in network.

4 Simulation Results and Discussion

To test our hypotheses we first plotted the results and checked visually if there might be a relation between the characteristics of the seed node and the success of an innovation diffusion simulation. As shown in Fig. 3 to 6 , seeds which cause a complete uptake of the innovation in the network show, on average, a higher degree distribution, eigenvector centrality, and betweenness centrality. Strikingly, the local clustering coefficient seems to negatively affect the successful penetration.

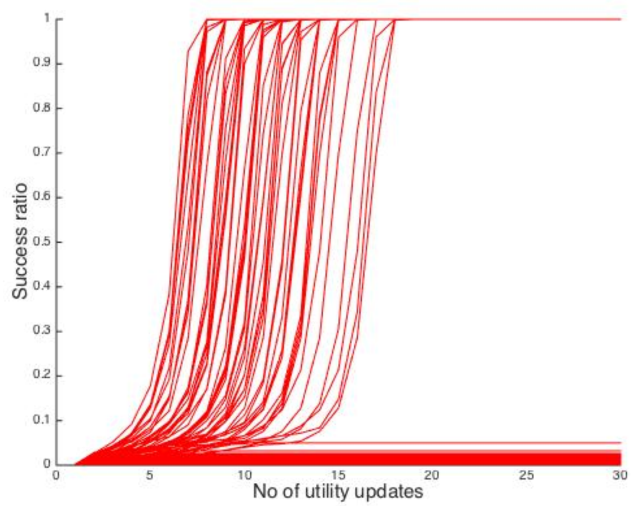

Figure 2: Each line represents an innovation diffusion simulation, with different nodes as seed. Our model shows an all-or-nothing effect, as network utility rises with more nodes in state 1.

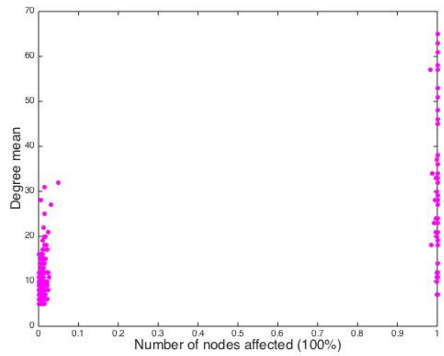

Figure 3: The impact of the mean degree of the seeds on the number of affected nodes. Each point represents one simulation run.

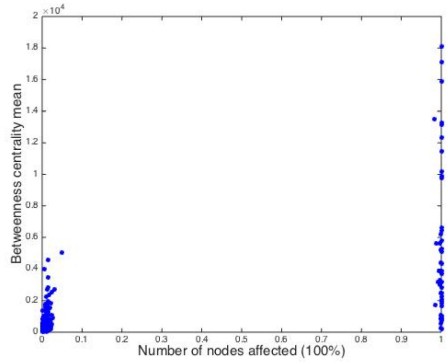

Figure 4: The impact of the mean betweenness centrality of the seeds on the number of affected nodes. Each point represents one simulation run.

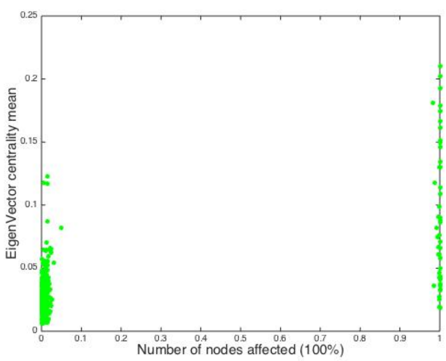

Figure 5: The impact of the mean eigenvector centrality of the seeds on the number of affected nodes. Each point represents one simulation run.

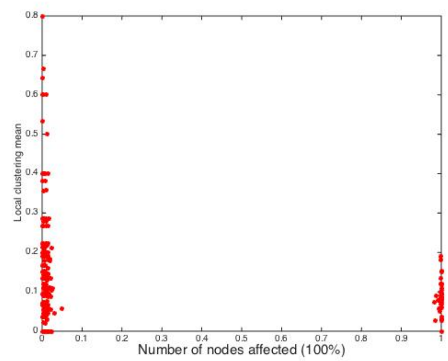

Figure 6: The impact of the mean local clustering coefficient of the seeds on the number of affected nodes. Each point represents one simulation run.

But as contagion models are complex, data can be messy and subtle effects might not always be visible by eye. Therefore we also applied statistical analyses to test relations. Because our model showed a binary all-or-nothing effect (see Fig. 2) we tried to predict a binary success variable through logistic linear regressions.

4.1 Overview of Seed Parameters and Success Rate

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
success	500	0.104	0.306	0	1
seed_perspref	500	0.504	0.103	0.143	0.831
seed_degree	500	9.944	8.999	5	65
seed_eigenc	500	0.033	0.030	0.006	0.210
seed_betweenness	500	885.040	2,139.630	35.768	18,131.290
seed_localc	500	0.125	0.138	0.000	0.800

Let us go through the table 1 row by row. An innovation diffusion simulation was counted as success when more than 95% of the nodes have changed state to 1. seed_perspref stands for the personal preference attributed to each node. One can clearly see the mean of 0.5 and the standard deviation of 0.1, the parameters we used to draw random values from this normal distribution. In the next row showing the degree of the seeds the power-law nature of the degree distribution is apparent, where the maximum value of 65 is much larger than one would expect for random graphs and other networks whose degree distribution do not follow a power-law. The minimum value of 5 comes from the parameters used to set up the scale-free network with preferential attachment, where we set the minimum edges every newly added nodes should have to 5. seed_eigenc stands for eigenvector centrality of each seed node, seed_localc for local clustering and seed_betweenness for betweenness centrality.

For further analyses we normalized the independent variables to simplify the comparison between them.

4.2 Multicollinearity

Table 2: Correlation Table of Independent Variables

	seed_perspref	seed_degree	seed_eigenc	seed_betweenness
seed_perspref				
seed_degree	-0.01			
seed_eigenc	-0.01	0.95***		
seed_betweenness	-0.02	0.97***	0.93***	
seed_localc	0.04	-0.11*	0.05	-0.09*
			Note:	*p<0.1; **p<0.05; ***p<0.01

The high correlations between degree, the eigenvector centrality, and the betweenness centrality have been determined experimentally[17] and we expect multicollinearity in the multiple logistic regression. Therefore the estimates of coefficients of multiple regression analyses must be viewed with some skepticism and future analyses would benefit from a more thorough investigation into this problem.

4.3 Logistic Regressions with Seed Characteristics

Table 3 and the following are structured the same way: For each parameter given in the very left column we performed a logistic regression with only one explanatory variable. The corresponding β_i-values are given in the columns further right. Additionally, we performed a logistic regression with all five centrality measures as explanatory variables and the β_1- to β_5-values are given in the very right column.

In table 3 the analysis of the mean of the seed characteristics is given. Let us first focus on the simple logistic regressions. The betweenness centrality, the degree, and the eigenvector centrality positively and significantly correlate with the chance of a seed to successfully penetrate the network with an innovation. As expected the personal preference is of no importance, since the seeds already carry the innovation and thus the chances of success do not raise with an increase of the seed's utility u. Remarkably, the local clustering coefficient is a negative predictor of the success of the seeds. Overall, the degree is the best predictor of all centrality measures as can be seen by the largest log likelihood and the smallest Akaike's information criterion.

In the multiple logistic regression the only remaining significant predictor is the degree. The significance of the personal preference is an artifact that did not occur in other simulation runs using the same model. The model's prediction power does not increase substantially adding more parameters. Comparing model (1) and model (6) the log likelihood and even less the AIC, which punishes adding more parameters, do not turn much to the better.

Table 3: Results Seed Characteristics

	Dependent variable:					
	success					
	(1)	(2)	(3)	(4)	(5)	(6)
seed_degree	2.255***					3.434***
	(0.269)					(1.247)
seed_perspref		0.208				0.463**
		(0.147)				(0.225)
seed_betweenness			3.039***			−0.110
			(0.397)			(1.694)
seed_eigenc				1.910***		−1.049
				(0.255)		(0.738)
seed_localc					−0.421**	−0.289
					(0.193)	(0.483)
Constant	−3.004***	−2.171***	−2.661***	−2.796***	−2.214***	−3.268***
	(0.252)	(0.149)	(0.225)	(0.219)	(0.156)	(0.357)
Observations	500	500	500	500	500	500
Log Likelihood	−80.498	−165.890	−83.792	−96.150	−164.016	−75.784
Akaike Inf. Crit.	164.997	335.780	171.584	196.301	332.031	163.568
Note:						*p<0.1; **p<0.05; ***p<0.01

4.4 Logistic Regressions with the Centrality Measures of Neighboring Nodes

Table 4 comprises the logistic regression analysis of the mean of the centrality measures belonging to the neighbors of the seeds. It contains several surprising results.

In general, all centrality measures have an inverted sign compared to the centrality measures of the seeds and are worse predictors for the success than the mean degree of the seeds. Strikingly, the inverse relationship between the mean degree of the neighbors and the success of the seeds persists in the multiple logistic regression.

Importantly, the mean personal preference does not significantly affect the success of seeds. This is against the notion that any increase of the neighbor's utility should positively impact the diffusion of innovation. We explain in the next section why this needs not to be the case for the mean personal preference of the neighboring nodes of the seeds.

We hypothesized that for a seed it is sufficient if one of its neighbors has a high degree or another high centrality measure in order for the innovation to propagate. We thus performed a logistic regression only with the maximum values of the neighboring nodes for the centrality measures. The results are given in table 5. All centrality measures and also the personal preference positively correlate with the successful diffusion of an innovation in the single logistic regression. In the multiple logistic regression, however, the degree and the eigenvector centrality are non-significant on an α-level of 0.05.

Table 4: Results: Mean of Neighboring Nodes

	Dependent variable:					
	success					
	(1)	(2)	(3)	(4)	(5)	(6)
neighs_degree	-0.590^{***}					-6.834^{***}
	(0.175)					(1.533)
neighs_perspref		0.190				0.288
		(0.148)				(0.179)
neighs_betweenness			-0.469^{***}			2.926^{***}
			(0.174)			(0.966)
neighs_eigenc				-0.366^{**}		3.449^{***}
				(0.162)		(1.026)
neighs_localc					0.796^{***}	0.491^{***}
					(0.146)	(0.173)
Constant	-2.279^{***}	-2.168^{***}	-2.233^{***}	-2.204^{***}	-2.402^{***}	-2.827^{***}
	(0.164)	(0.148)	(0.158)	(0.153)	(0.176)	(0.237)
Observations	500	500	500	500	500	500
Log Likelihood	−160.441	−166.057	−162.782	−164.162	−151.131	−134.517
Akaike Inf. Crit.	324.882	336.115	329.563	332.325	306.263	281.035
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 5: Results Maximum Value of Neighboring Nodes

	Dependent variable:					
	success					
	(1)	(2)	(3)	(4)	(5)	(6)
neighsmax_degree	1.322^{***}					-3.434^{*}
	(0.298)					(1.853)
neighsmax_perspref		1.215^{***}				1.243^{***}
		(0.185)				(0.243)
neighsmax_betweenness			1.179^{***}			2.190^{**}
			(0.235)			(1.065)
neighsmax_eigenc				1.465^{***}		2.291^{*}
				(0.329)		(1.302)
neighsmax_localc					1.166^{***}	0.822^{***}
					(0.145)	(0.161)
Constant	-2.663^{***}	-2.663^{***}	-2.611^{***}	-2.744^{***}	-2.713^{***}	-3.583^{***}
	(0.243)	(0.209)	(0.221)	(0.261)	(0.210)	(0.351)
Observations	500	500	500	500	500	500
Log Likelihood	−149.235	−139.638	−147.970	−147.932	−128.118	−101.682
Akaike Inf. Crit.	302.470	283.275	299.940	299.864	260.237	215.365
Note:						*p<0.1; **p<0.05; ***p<0.01

4.5 Model Comparison

We analyzed whether the incorporation of the parameters of the neighboring nodes add any value to our initial model with only the seed parameters. As can be seen in table 6 the parameters of neighboring nodes add little statistical value. The best combination is (3): the seed parameters combined with the mean of the neighboring parameters, which is only slightly better than model (1). Overall, the degree of the seeds is the best predictor from all analyzed parameters as was discussed in section 4.3.

Table 6: Results Model Comparison with different Variable Sources

	Dependent variable:			
	success			
	(1)	(2)	(3)	(4)
seed_params	x			
seed_params & neighsmax_params		x		
seed_params & neighs_params			x	
seed_params & neighsmax_params & neighs_params				x
Observations	500	500	500	500
Log Likelihood	−75.681	−68.753	−67.338	−63.118
Akaike Inf. Crit.	163.362	159.507	156.677	158.236
Note:		*p<0.1; **p<0.05; ***p<0.01		

4.6 Discussion

From all analyzed centrality measures of seeds and their neighboring nodes the best predictor for the successful diffusion of innovation is the mean degree of the seeds, although on a generally low level of predictability. This is, of course, what one would expect given the central role of the degree distribution for the network topology. For comparison, the mean degree of the neighboring nodes is more interesting: it exhibits a negative correlation with the innovation success. The meaning of this finding, however, is not accessible to us.

The mean local clustering coefficient of the seeds correlates negatively with the innovation diffusion, whereas the mean and maximum local clustering coefficient of the neighboring nodes correlate positively. In our view a local dense cluster of nodes that contains only one seed node is highly stable and might resist the innovation. Thus, high clustering might cause adoption once the innovation penetrates the cluster, however, it might also weaken adoption by making the initial diffusion from inside more difficult.[18] The phenomena of inert local clusters with seeds in the centre could explain why successful seeds did show on average a lower clustering coefficient in our simulation. Following this line of reasoning, a neighbor of a seed with a high clustering coefficient resides inside a local cluster and should enable the spreading of the innovation from the seed to the other nodes of the local cluster. Consequentially, in our simulations the positive β-values for the mean and the maximum local clustering coefficient of neighboring nodes were highly significant both in the single and multiple logistic regression.

Apart from the centrality measures we also looked at the personal utility gain, dubbed the personal preference, a node receives from accepting the innovation. Although the personal preference does not play a role as mean value of the neighboring nodes, its maximum of the neighbors positively correlates with the diffusion success. We suggest that, on average, a single neighbor of the seed with a high personal preference suffices to start the process of innovation diffusion. Of course this is no guarantee that the whole network will be affected by the innovation.

This leads us to the last point: Structural characteristics of seeds only allow for marginal explanation in variance of success, so they are not as important to innovation diffusion as previously stated. Other factors as global network parameters such as the susceptibility of nodes in general might play a more important role. Seeds might be indeed just the kick-off and not as relevant when looking at innovation diffusion as a whole.

5 Summary and Outlook

We applied a threshold model on an undirected scale-free network to answer the question which characteristic parameters of seeds of innovation and of their neighboring nodes may allow to forecast the success of an innovation diffusion. The results for the seeds are straightforward: Here, the degree, the betweenness centrality, and the eigenvector centrality show a positive correlation with the diffusion success, whereas the local clustering coefficient is negatively correlated. The picture for the neighboring nodes, however, is more complex. What clearly can be said is that the diffusion success profits significantly from a large and positive local clustering coefficient. In general, the centrality measures for the seeds and their neighbors show a statistically low ability to predict the diffusion success, which suggests that global network parameters and parameters of other nodes may be more important predictors.

Real social networks are mostly directed and the connection between the nodes are usually weighted. As a next step we thus suggest to apply the threshold model on an empirical network with directed and weighted edges. Principally, any extension of this model to other empirical or theoretical networks should prove itself as valuable.

The threshold model we used had been developed in 2013, but due to the fast progress in the network sciences we expect that soon improved threshold models for the diffusion of innovation will be suggested. Our simulation can be easily adapted to incorporate any other diffusion model and we are looking forward to see it being used with models that allow a more detailed analysis of the seed parameters and their effect on the diffusion of innovation.

6 References

References

[1] LA Fourt and JW Woodlock. Early Prediction of Market Success for New Grocery Products. *Journal of Marketing*, 25(2):31–38, 1960.

[2] EM Rogers. *Diffusion of Innovations*. New York: The Free Press, 1962.

[3] N Meade and T Islam. Modelling and Forecasting the Diffusion of Innovation – A 25-Year Review. *International Journal of Forecasting*, 22(3):519–545, 2006.

[4] SA Delre et al. Will It Spread or Not? The Effects of Social Influences and Network Topology on Innovation Diffusion. *Journal of Product Innovation Management*, 27:267–282, 2010.

[5] DJ Watts and PS Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458, 2007.

[6] A-L Barabasi and ZN Oltvai. Network Biology: Understanding the Cell's Functional Organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[7] MA Nowak and DC Krakauer. The Evolution of Language. *PNAS*, 96(14): 8028–8033, 1999.

[8] MA Nowak. Five Rules for the Evolution of Cooperation. *Science*, 314(5805): 1560–1563, 2006.

[9] I Bozic et al. Evolutionary Dynamics of Cancer in Response to Targeted Combination Therapy. *eLife*, 2, 2013.

[10] SP Borgatti et al. Network Analysis in the Social Sciences. *Science*, 323:892–895, 2009.

[11] DJ Watts. The 'New' Science of Networks. *Annual Review of Sociology*, 30(1): 243–270, 2004.

[12] LMA Bettencourt et al. The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models. *Physica A: Statistical Mechanics and its Applications*, 364:513–536, 2006.

[13] WO Kermack and AG McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A.*, 115(772): 700–721, 1927.

[14] NJ McCullen et al. Multiparameter Models of Innovation Diffusion on Complex Networks. *SIAM Journal on Applied Dynamical Systems*, 12(1):515–532, 2013.

[15] DJ Watts and SH Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, 1998.

[16] AL Barabasi and R Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[17] TW Valente et al. How Correlated Are Network Centrality Measures? *Connections (Toronto, Ont.)*, 28(1):16–26, 2008.

[18] D Acemoglu, A Ozdaglar, and E Yildiz. Diffusion of Innovations in Social Networks. *Decision and Control and European Control Conference CDC-ECC, 50th IEEE Conference*, pages 2329–2334, 2011.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Diffusion of Innovation and the Characteristics of Seeds

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Lechner	Sebastian
Oesch	Adrian
Seifoddini	Amrollah

With my signature I confirm that
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, 12.12.2014	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.