ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

Project Report

A Contagion Based Model for the Spreading of Pirated Movies over the Internet

Simon Ghysbrecht & Pascal Zehnder

Zurich
December 2016

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.

Simon Ghysbrecht

Pascal Zehnder

Contents

1 Abstract

The spreading of pirated movies was modeled using a relatively simple contagion-based mechanism running on an Erdös-Rényi network. An interaction with movie viewings in theaters was introduced by constructing a similar model for the spreading of cinema viewers and combining them on a single lattice. In this way, attempts were made to simulate the influence of movie piracy on box-office revenue.

In a side project, we tried to assess how realistic it would be to support our model with analytical data, by setting up a data-collection system. This work could be used to further investigate on the real spreading of pirated files or the like.

2 Individual contributions

In the beginning, we both had to get acquainted with the subject, because were both new to the field. While Simon did a lot of research on the literature and was less familiar with coding MATLAB, Pascal was already more fluent with the program, and started experimenting with the implementation of some early models. In this way, we hoped to combine our knowledge so in the end, each of us could contribute to both the conceptual framework and the implementation in code. We then started to write a model from scratch, while both contributed ideas to the implementation. Simon did more on checking for errors in logic, Pascal did more on debugging of the actual code, and finally run the evaluations on a spare computer. At the end of the project Simon, who is more familiar with LaTeX, focused on writing the theoretical part of the report, while Pascal wrote the implementation. Pascal also did most of the plots, because he had immediate access to it, whilst sending gigabytes of data would not have made much sense. Simon then joined all the images, and after analyzing the results together, put them in the report. For the side project on data scraping Simon used the *scrapy* module in python to run batches of data acquirement, while Pascal did a long term gathering manually, and combined the data in plots using MATLAB. As a concluding remark we could say it was a pleasure to work with each other, as different knowledge lead to exchange while one also could focus on core areas and accept each others work and decisions.

3 Introduction and Motivations

In the last few decades, Internet piracy has known an incredible growth, reaching proportions by far exceeding any counterfeit type products before them. Illegal copies of digital files are being spread in huge numbers and all over the world. Most frequently shared is without any doubt entertainment material like music or movies.

Nonetheless, other types of products, ranging from software to scientific articles[1], are also being distributed in unauthorized ways and on a large scale.

As these illicit practices seem to have a considerable influence on the industries involved, quite some effort has been done to understand and combat Internet piracy, both on the demand and the supply side[2, 3, 4, 5, 6]. The problem is not a simple one, though, as it has multiple facets in different disciplines, and thus multiple possible approaches.

Whilst most people today use the Internet, only a small minority is familiar with its technical details. An investigation into any of its phenomena, however, will in some way be shaped by the underlying technological structures involved. Internet piracy is highly reliant on modern information-sharing technologies, so there is no way around the fact that it has strong *technical and technological facets*. There is also an obvious *economical facet*, as piracy affects a number of substantial industries. Moreover, Internet piracy is illegal in most countries in the world, which gives the situation an important *legal facet*. The fact that a huge number of people seem to perpetrate this criminal act could even give rise to an interesting *psychological* or *criminological facet*. In the end, however, illicit sharing of copyrighted material can be reduced to interactions between 'human actors', something we could call the *sociological facet*. It is this facet which will be central in our approach to model Internet piracy.

The idea is thus to simplify the concept of Internet piracy to a relatively simple social system where the human actors appear as social agents who distribute the pirated files amongst each other. In order to achieve this, the theory of random graphs can be used to construct the network of such agents, while the actual spreading of the pirated files can be simulated using an epidemic-type model.

As Internet piracy is often illegal, there is very little data available concerning the number of downloads and how this evolves in time. For this report, we came up with a method of data collection where information is gathered directly from the piracy portal sites. Nevertheless, backing the spreading of pirated files by empirical study hardly seemed realistic within the scope of this project. Therefore, we reasoned that, while theoretically investigating possible spreading mechanisms of pirated files alone might be valuable, it is probably more interesting to investigate some sort of interaction with the distribution of legal alternatives.

The focus in this work will be on movie piracy. The reason for this is three-fold:

1. Movies are one of the most frequently pirated file types.

2. There is abundant information available on the legal sales and distribution of movies, which will be advantageous for comparison to empirical or simulated data on movie piracy.

3. Contrary to music files (often downloaded per album) or television series (often downloaded per season or collection of seasons), movies are generally downloaded separately.

As mentioned above, the interaction between movie piracy and more official alternatives is especially interesting. While other (legal) channels like DVD or Netflix are available, the focus will be on the interaction of movie piracy with movie theater releases. To that effect, the model should include a reciprocal influence between the spreading of the pirated files on the Internet and the number of movie viewers in theaters (which might further be referred to as 'cinema viewers').

Because there is a pretty straightforward relation between the amount of people seeing a movie in theaters and the box-office revenue of the movie, a good (albeit ambitious) research question would be:

To what effect does Internet movie piracy influence box-office revenues?

This will of course depend on certain circumstances. We have tried to include some of the real-life factors we thought might be influential in our theoretical model. It will be interesting to see how they affect our simulations.

4 Description of the Model

4.1 The General Model

The idea is to use an **epidemic-type model** to simulate the spreading of pirated movies over a network, where the nodes represent people who might be interested in seeing the movie, and connections represent possible contact between them. In order to make the model as physical as possible, each parameter of the network should in principle correspond to a practical element in movie piracy.

Because we are curious towards the interaction of movie piracy with cinema viewings, the evolution of people going to theaters has to be simulated as well. This can be done in an equivalent way as to the case of movie piracy, that is, using an epidemic-type model. The challenge will be how to correlate the two spreading mechanisms.

As a first approach, we decided to work on a **single network** on which both 'movie piracy' and 'cinema-going' spreads. This means both the nodes and the way they are connected are the same for the piracy and cinema spreading. One could easily think of ways to extend this model. It should be pretty straightforward, for example, to split up the connectivity for piracy and cinema-viewings by using separate adjacency matrices. One could also detach the networks all together. In this case, there would have to be composed a well-defined way of how the networks interact.

Furthermore, we set that each individual, represented by a single node, will only be able to be in **one of three possible states**:

0. individual did not see the movie yet
1. individual saw the movie in cinema
2. individual saw the movie by pirating it

It might be less worthwhile to go beyond this point, as it seems to us that watching a movie in the theater after having pirated it seems uncommon enough to ignore. Moreover, pirating the movie after going to the cinema does not directly influence cinema revenue, although it does indirectly as it might influence the spreading.

Concerning network topology, we chose to invoke the **Erdös-Rényi model**[7]. This gives us two important parameters: the size of the model corresponding to the number of nodes N, and the probability of having an edge between each pair of nodes p_{EG}. The expectation value of the average degree in this model is given by:

$$\langle k \rangle = (n-1)p_{\mathrm{EG}} \tag{1}$$

while the network is expected to be connected for p_{EG} above the sharp threshold:

$$p_{\mathrm{EG}} > \frac{\ln(N)}{N} \tag{2}$$

which will further be referred to as the *critical threshold*. The size of the network is related to the amount of people who are interested in and capable of watching the movie, or, in economical terms, the size of the market. The physical meaning of p_{EG} is more subtle, and will have to do with the importance of people talking about the movie towards the success of the film.

Initially, the generated Erdös-Rényi network is completely empty, that is, all nodes have value zero and thus no one is assumed to have seen the film. In order to create some initial cinema-goers, a raining function was implemented. Instead of using fixed seeders, the raining function was incorporated to be able to include external influences on the network, and might for example incorporate the effects of marketing before release of the movie.

For the spreading of pirated movies and cinema-goers on the network, a function was created which, for a fixed number of time steps, describes the spreading mechanism. We have chosen a system where, when a node is updated, the probability of being 'infected', depends on:

1. The number of neighbors who watched the movie, independent of whether they saw it in theater or by pirating

2. A *contact rate* β which can be related to the quality, budget or genre of the film

3. The amount of time since it has been released in theaters. We assumed the interest in seeing a movie *decays exponentially* with time.

Up to this point, there is no difference between piracy or cinema infection, as they depend on the same parameters. As a matter of fact, this distinction is only made AFTER infection. Basically, someone getting infected can be interpreted as that person deciding to watch the movie. We then introduce a new parameter: the *probability of cinema viewing once infected* p_c. The *probability of piracy once infected* is then $1 - p_c$. These parameters can be related to how easy it is to pirate a movie, what the general mentality of the people towards piracy is etc.

It is often the case that pirated versions of movies appear on the Internet only a certain time after theatrical release. Therefore, a last factor to include is *delay* δ, defined as the number of time steps at the beginning of the simulation during which piracy is completely suppressed.

4.2 How to Evaluate the Influence of Movie Piracy

There is a pretty much direct relation between the amount of people watching the movie in cinema's and the box-office revenues. Therefore, one can simply count the amount of people who went to the theaters after the final time step, and use this value as a direct measure of cinema revenue.

Determining the influence of movie piracy on this cinema revenue, however, is less simple. In this project, we approach the problem by constructing a reference system where only cinema viewing of movies is possible, but the rest of the parameters stay the same. The reference system has the same amount of nodes and the same connectivity. The difference, however, is that piracy infections are simply ignored. More specifically, nodes which get infected by piracy are immediately reset to zero. This basically boils down to assuming that people who'd normally pirate a movie won't watch it when piracy is not an option[1]. In the end, the number of cinema-goers in the network with piracy can be compared to the network without piracy. In light of this, we define the ratio:

$$\text{ratio} = \frac{n_{\text{cinema}}}{n_{\text{cinema\_without\_piracy}}} \tag{3}$$

[1]This is not completely true though, as they still might get cinema-infected at a later stage.

4.3 A More Realistic Reference System

Just assuming people who'd normally pirate a movie won't watch it when piracy is not an option hardly seems fair. Therefore, we propose a small modification to the reference model set out above, which we shall shortly investigate at the end of our discussion.

Instead of simply ignoring piracy infections, we introduce a probability $p_{p\to c}$ for a person who'd normally pirate a movie to go watch it in cinema if piracy becomes unavailable. This is formally the same as using different probabilities of cinema-going in the original versus the reference network. If the probability of cinema-going is p_c in the original network, the reference network will now adapt a value of $p_c+(1-p_c)\cdot p_{p\to c}$.

5 Implementation

The implementation of the model described above was done in MATLAB, and all files can be found on our GitHub page. The basic thoughts our explained in this section. In the end we are interested in the ratio between the amount of movie-viewers when piracy is activated compared to when it is not. The program is built in three main layers.

The first layer consists of a script called *Investigation of Parameters* (IOP), in which the investigation is set up and run. First, the seed is collected for the random number generator in order to be able to reproduce the calculations. The consequent lines are used to set up the parameters which are kept fixed. Due to the time-consuming nature of the simulations, sometimes running longer than one day, a wait bar has been added which shows the progress and the estimated time of the calculations. After this, a tensor with a size of (invest1, invest2, numberruns) is set up, where *invest1* and *invest2* are the number of investigated parameters, and *numberruns* is the number of runs which will be used for statistical analysis. This number has been chosen to be 50, based on one simulation with 500 runs. As can be seen in Fig 1, the results of this simulation clearly show that after 100 runs nothing significantly changes anymore, whereas between 30 and 100 runs the results are already a very close approximation to the results we get after 500 runs. Choosing an optimum between computational expense and accuracy, we will always perform 50 runs for producing data.

Consequently, two or three for-loops are entered. The first iterates over number-runs and aims at generating data to average over afterwards. The other two iterate over the parameters under investigation. In each loop the returned data from the *masterpirate* function are saved as a *.mat* file with a unique name for further or additional investigations.

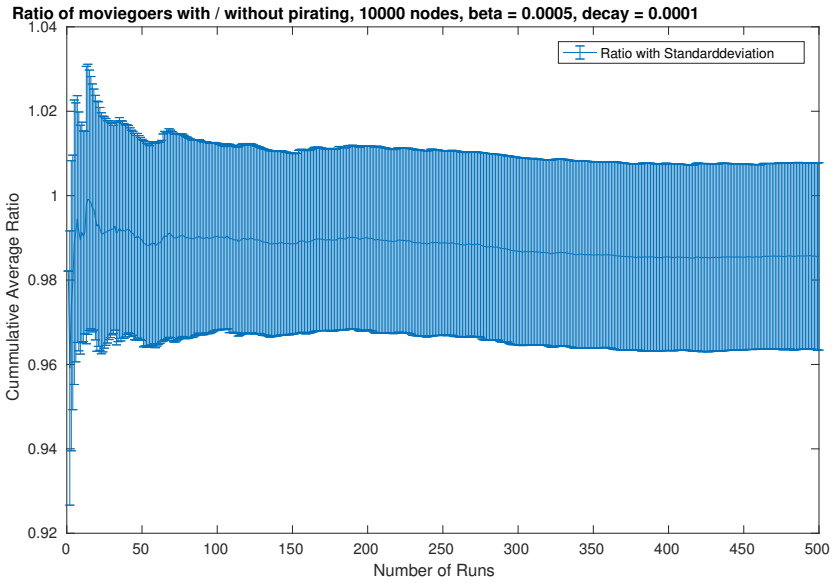


Figure 1: Cummulative average and standard deviation of ratio of movie goers for a network with 10000 nodes, a contact rate of $\beta = 0.0005$ and a decay factor of 0.0001

In the second layer, the function *masterpirate* takes all the inputed parameters and processes them. It calls *generate_EG* which provides the matrix for *creategraph* to set up a graph-type structure. Empty matrices for data storage are generated, then a while-loop is entered which runs the spreading. Spreading happens in a random manner. Each time step calls as many nodes as there are in the system, but in a random fashion, because of which some might get called more than ones, while other might not ever be called in that time step. After one time step the states are saved. Also, a decay factor gets updated which slows the spreading according to an exponential decay.

The spreading itself starts once the *raining*-function randomly infects some nodes. For our investigations two spreadings run simultaneously, one with and the other without pirating, set up in two functions named *spreading* and *spreading_no_piracy*, which constitute our third layer. Both functions will change the state of the contemplated node with a probability given by:

$$p = \left(1 - (1 - \beta)^{\text{neighs}}\right) * \text{decay} \tag{4}$$

In order to evaluate this, we generate a random number using the MATLAB *rand()* function. If the generated number is smaller than the probability p, the node is set to change its state. As explained in Section 4, this happens according to probability

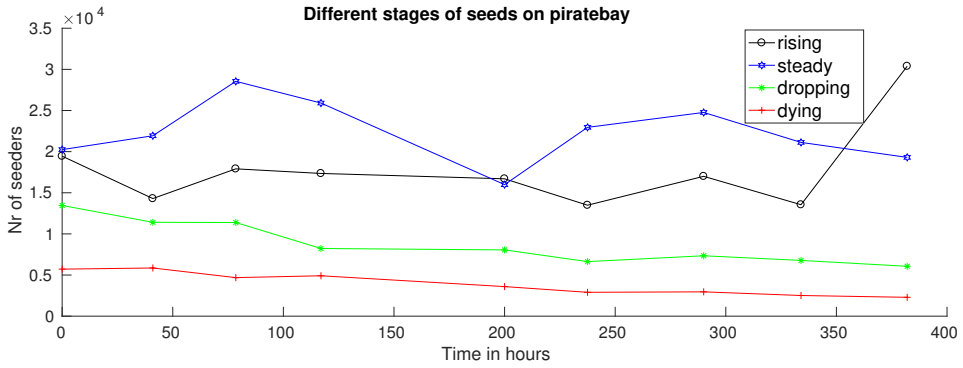


Figure 2: Empirical data from manually keeping track of the number of seeders for a few examples of pirated movies.

of going to cinema p_c or movie pirating $1 - p_c$. Again, evaluating the probabilities happens through use of the *rand()* function.

After finishing the simulations, the script averages all the stats computed, calculates a standard deviation and plots the data, after which a last save is performed.

6 Data-Scraping: a Side Project

At the start of our project we wanted to compare our simulations to the spreading of real movies, and therefore collected data on pirated movies. We did not have the time to compare any of those, nevertheless we wanted to show and shortly explain our processes and results on this. First, we did data collection by hand, and just selected the most seeded movies on *thepiratebay*[2] and followed those seeds for a few weeks. We focused only on the number of seeds. After some time we realized that we could see different stages of a lifetime of a seed. As can be seen in Fig.2, there were 4 phases, first one rising, then steady, dropping and dying. But what was not actually in those data was the initiation.

After this result we tried to catch a glimpse of the rise of new seeds and selected a few top uploaders. We automated this process by building a python-based script using the *scrapy*[3] module. We then set up a scraper on the uploader page to collect data on the fifteen most recent files, scraping every quarter hour. The results of this small experiment can be seen in Fig. 3 .

Obviously, the method still requires some improvement. What further surprised us is the fact that the data on the number of seeds from the website itself only gets

[2]https://pirateproxy.vip/
[3]https://scrapy.org/

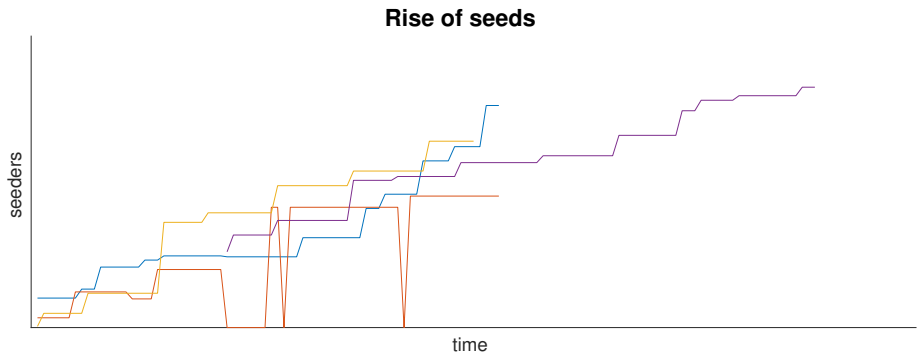


Figure 3: Some examples of empirical data using an automated scraper.

updated sporadically, which would make empirical analysis using this technique less interesting. We also sampled too little data to conclude any specific behavior.

7 Simulation Results and Discussion

The main goal of this section is to investigate in which way movie piracy influences cinema revenue within the framework of our model. This is done by performing simulations as explained in the previous chapter. Central to our approach is the ratio between the number of cinema-goers in the original network, which includes piracy, and the number of cinema-goers in the reference network, which excludes piracy.

The spreading of both the pirated movies and the cinema-goers is dependent on a number of parameters:

- the size N of the network
- the probability of connection p_{EG}
- the contact rate factor β
- the probability of cinema viewing p_c
- the delay δ

We will not be interested in investigating the effect of the decay factor. While the initial seeding by the raining function might be very influential, this aspect will be overlooked as well. Unless mentioned otherwise, we will take $N = 1000$, $\beta = 0.0002$, $p_c = 0.5$ and $\delta = 0$ as parameters.

In investigating the size, one expected effect might be *crowding* of the market by movie piracy spreading. If the amount of people interested in watching the movie is relatively small (that is, not that much larger) compared to the amount of people who end up pirating it, piracy will be expected to have a strong negative influence.

On the other hand, if the network is large and the amount of people pirating is small, the piracy might actually help the spreading, and end up having a positive effect. In more real-world terms, the piracy is said to have an information-spreading effect, or, less formally, to create mouth-to-mouth publicity.

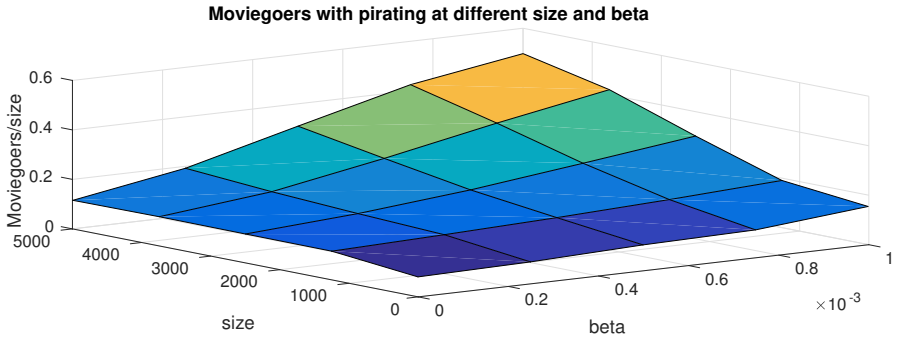
Comparing random graph networks of different sizes, however, is not a straightforward thing, as the connectivity, critical probability and average degree of the network all depend on the number of nodes. Moreover, graph size is relative to the contact rate factor and the maximum amount of time steps. For large enough contact rates and time steps, even large networks might get crowded!

In a first approach, we compared ratios for different β's and different network sizes at critical probabilities of $p_{EG} = \ln(N)/N$. The number of cinema-goers with and without piracy and the number of movie piraters, all scaled by size, are given in function of β and N in Figs. 4a-4c. As can be expected, movie watchers in all cases increase with increasing value of β. A result which might be slightly more difficult to interpret, is the increasing amount of nodes occupied when the size gets bigger, even after scaling, and this for all three of the cases. This might be a display of the difficulty of comparing networks at different sizes. More specifically, comparing at critical probabilities for p_{EG} is probably not the way to go here.

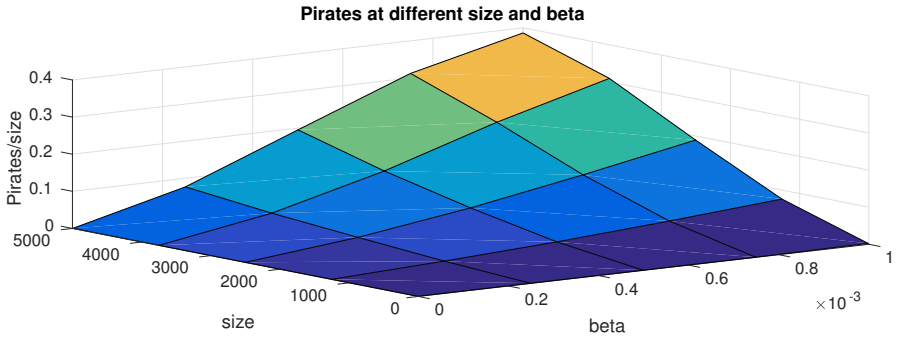
Instead, a better idea is probably to compare networks of different sizes for constant expectancy values of average degrees according to Eq. (1). Because of limited time and computational resources, however, we could not re-investigate the size-effects in the same way.

Subsequently, we investigated effects of the delay on piracy. Naturally, we expect a decrease of the amount of piraters as decay times get larger, as the piracy gets less time to spread, and appears after probabilities have decayed for a longer time. This can be seen in Fig. 4.

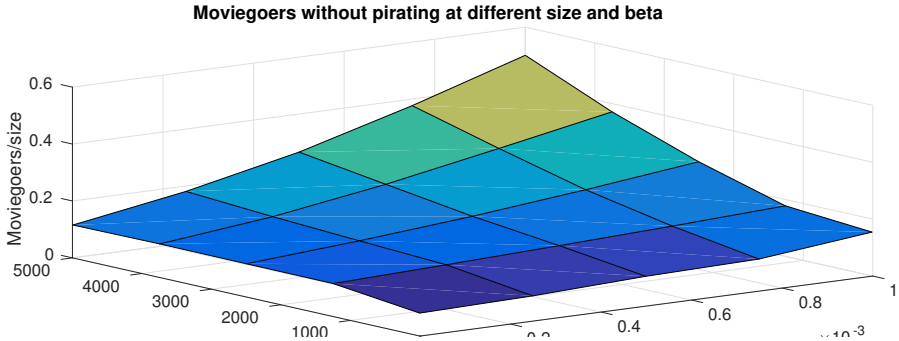
If we look instead at the movie-goers (Fig. 5) in function of delay, we see that delay does not directly seem to influence the amount of movie-goers. More generally, piracy seems not to influence movie-going at all. For this specific case, there thus seems no correlation between the piracy and the cinema-going network, which kind of defeats the purpose of our simulation. . A reason for this might be the unrealistic implementation of the reference system which simply spreads in the same way as the original system but simply ignores piracy. For a network which is big enough to avoid crowding effects, it indeed should not make much of a difference if some pirates appear, as long as the chance of cinema-going remains the same. The only effect is



(a) Number of cinema movie watchers in presence of pirating for different values of β and N normalized by size and for p_{EG} at critical probability. Colors represent standard deviations, and are generally larger when the actual numbers get larger.



(b) Number of pirated movie watchers for different values of β and N normalized by size and for p_{EG} at critical probability.



(c) Number of cinema movie watchers in absence of pirating for different values of β and N normalized by size and for p_{EG} at critical probability.

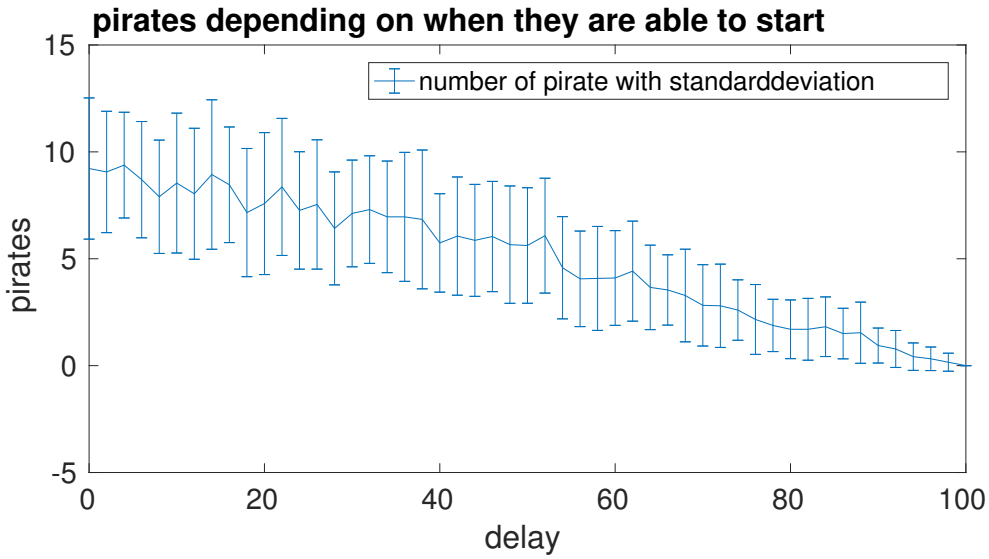


Figure 4: Absolute number of pirates for increasing times of delay.

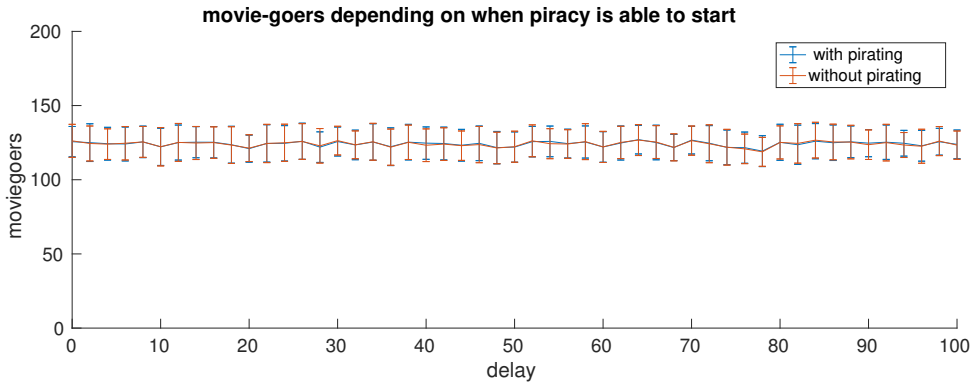


Figure 5: Absolute number of movie-goers for increasing times of delay.

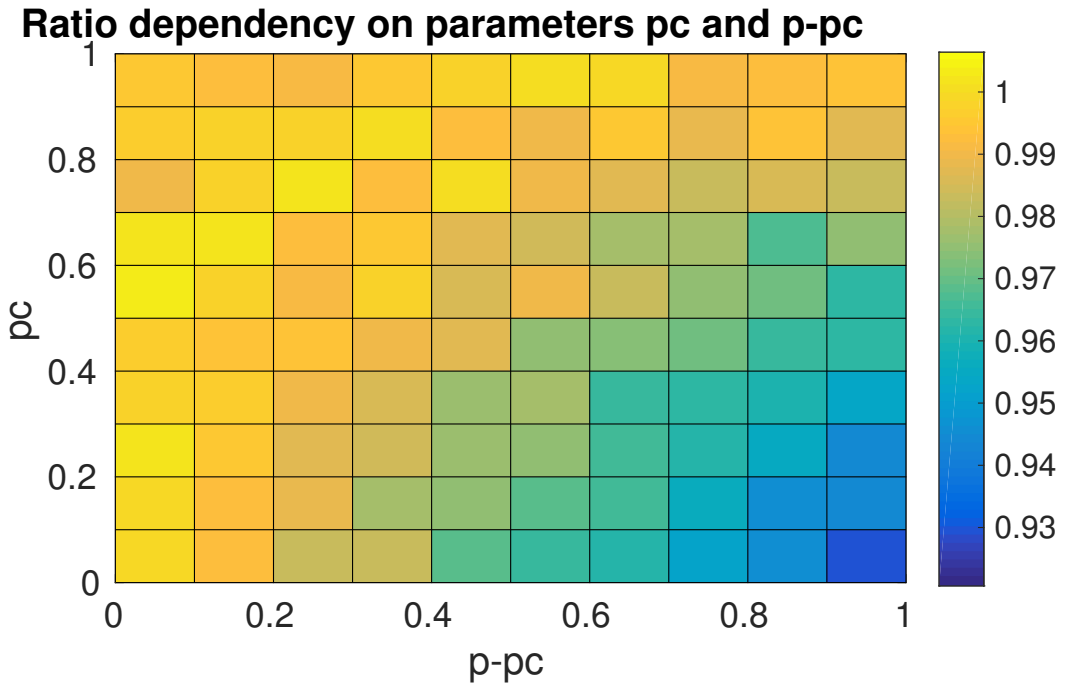


Figure 6

a slightly enhanced contamination effect, which might not be visible at all if β's are not very big.

Therefore, as discussed in Chapter 4.3, a new parameter $p_{p\to c}$ denoting the probability for a person who'd normally pirate a movie to go watch it in cinema if piracy becomes unavailable. This parameter influences spreading on the reference network only. A phase diagram of the ratio between cinema-goers with compared to without piracy in function of p_c and $p_{p\to c}$ is given in Fig. 6.

We clearly see an increasingly negative effect of piracy when $p_{p\to c}$ increases. Furthermore, this negative effect is smaller if p_c is smaller.

8 Conclusion

In the end, we can conclude that the interaction of the two networks did not show the desired results in our analysis. We therefore are forced to recognize that the interlinking of the systems might not be as straightforward as our model suggests. Rather, both systems have their own underlying structures, which might have a big impact on the both the spreading and interaction of the systems. In our model, we simplified the situation by introducing the same spreading patterns for both the

case of piracy and cinema spreading. In order to answer the question put forward at the beginning of our project, one might have to accept that the characteristics of the Internet on one side and the real world on the other, should be considered separately. Moreover, the Internet consists of such a huge and complicated structure, that finding the right parameters for a theoretical model or looking for data in an empirical analysis might be challenging. The spreading of movies in a cinema though should be more straight forward to simulate.

Concerning our own results, we can embrace the thesis, that the implementation of our code has been as easy and straightforward as possible, using the same files and functions for as many different steps as possible. In light of this, we tried our best to include as few parameters as possible. The subject showed to be more complex than anticipated, however, so we ended up with a few more than we started with.

References

[1] J. Bohannon, "Who's downloading pirated papers? everyone," *Science*, vol. 352, no. 6285, pp. 508–512, 2016.

[2] P. Belleflammea and M. Peitzb, "Digital piracy," 2014.

[3] V. Gehlen, A. Finamore, M. Mellia, and M. M. Munafo, "Uncovering the big players of the web," in *International Workshop on Traffic Monitoring and Analysis*, pp. 15–28, Springer, 2012.

[4] L. Ma, A. L. Montgomery, P. V. Singh, and M. D. Smith, "An empirical analysis of the impact of pre-release movie piracy on box office revenue," *Information Systems Research*, vol. 25, no. 3, pp. 590–603, 2014.

[5] C. Peukert, J. Claussen, and T. Kretschmer, "Piracy and box office movie revenues: Evidence from megaupload," *Available at SSRN 2176246*, 2015.

[6] A. Zentner, "Measuring the impact of file sharing on the movie industry: An empirical analysis using a panel of countries," *Available at SSRN 1792615*, 2010.

[7] P. Erdös and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.