ETHlogo

**Lecture with Computer Exercises:**

**Modelling and Simulating Social Systems with MATLAB**

Project Report

|  |
| --- |
| Stochastic Simulation of Epidemic Outbreak in a large Network |

Jordi Christian, Schmid Yannick & Stücheli Pascal

Zürich

10. Mai 2012

**IMPORTANT**

**You MUST include the ETH declaration of originality here; it is available for download on the course website or at**

**http://www.ethz.ch/faculty/exams/plagiarism/index\_EN;**

**It can be printed as pdf and should be filled out in handwriting.**

**Agreement for free-download**

We hereby agree to make our source code of 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.

|  |  |
| --- | --- |
| Jordi Christian | Schmid Yannick |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| Stücheli Pascal |  |

**Table of contents**

[Abstract 4](#_Toc325636203)

[1 Individual contributions 5](#_Toc325636204)

[2 Introduction and Motivations 6](#_Toc325636205)

[3 Description of the Model 7](#_Toc325636206)

[3.1 The continuous disease model 7](#_Toc325636207)

[3.2 The stochastic disease model 7](#_Toc325636208)

[3.3 The network 8](#_Toc325636209)

[3.4 The transport 8](#_Toc325636210)

[4 Implementation 10](#_Toc325636211)

[4.1 Network generation 10](#_Toc325636212)

[4.2 Simulation of the infection 10](#_Toc325636213)

[4.3 Transport 10](#_Toc325636214)

[4.4 Experiments 10](#_Toc325636215)

[5 Simulation Results and Discussion 12](#_Toc325636216)

[6 Summary and Outlook 13](#_Toc325636217)

[Outlook 13](#_Toc325636218)

[References 14](#_Toc325636219)

# Abstract

# 1 Individual contributions

We initially divided the programming part which was easily feasible since there were three main tasks. Pascal Stücheli was responsible for the epidemic model and its implementation, Christian Jordi familiarized himself in with GEPHI and was responsible for the network generation. Yannick Schmid wrote the transport function. At any time we assisted each other because we had to make sure that the functions worked well together. Further we had to increase the overall performance of our simulation, mainly by making approximation or changing the structure of certain functions. At this point it was important to discuss the possibilities and limitations of the approximations and all of us contributed ideas and concerns. The analysis and interpretation of the results and also the writing of the report was done in a cooperative manner.

# 2 Introduction and Motivations

Last winter, the movie Contagion*[[1]](#footnote-1)* came to the cinemas. In this movie the fates of some characters are shown during the outbreak of a deadly viral infection disease. The first infected person in the movie was a travelling business woman, who carried the virus across two continents such that the disease could spread over the whole world very rapidly. After seeing this movie we asked ourselves how a realistic disease would behave if no remedy or cure could be found. Inspired by this movie, we thought about a large network representing the big cities on the world, which are either connected by air route or other means of public transportation.

Surveillance and simulation of infectious disease are necessary to predict possible outbreaks of epidemics. At the moment there are of course already several epidemic simulation tools. Some of them also contain stochastic effects like example given FluTE. Therefore we do not claim to invent something new, but we wanted to see if we would be able to implement an epidemic simulation in a large network with ordinary computational power. We want to emphasize that we tried to realize our own ideas to create a stochastic simulation of an epidemic.

The aim of this study is to follow the time evolution of discrete numbers of persons who are either susceptible or infected and the spreading of infected persons among a large scale-free network.

We wanted to investigate the following questions:

* How does the number of total infected in a network vary for a given time of simulation?
* How does the time vary among the different simulation until the whole population in the network has been infected?
* What are the variations for certain cities with different degrees? What are the differences for hubs and leaves in the network?
* Are there tendencies in the regional spreading like for example towards hubs?
* Can we gain predictive power from our result? For example, can we predict the time until a city becomes infected depending on the degree of the first infected city? Or depending on the distance?

Additionally we were also interest if we could handle such an amount of data. Actually we spent a lot of time improving the performance of our simulation and exploring approximations.

# 3 Description of the Model

3.1 The continuous disease model

Our model is basically a simplification of the SIR model proposed by Kermack and McKendrick[[2]](#footnote-2). The major difference is we did not consider the state R that represents removed people which became immune or died after being infected.

To derive the disease model a city with N inhabitants is considered. The inhabitants are either susceptible for the disease or they already carry the disease and can therefore infect susceptible persons. In the equations below susceptible persons are labelled S and infected persons I. At a constant birth rate B healthy (and therefore susceptible) children are born while dyeing is a first order process with a death rate d. The amount of newly infected persons per time unit depends on the availability of susceptible persons and their meetings with infected persons. The contact rate β is the rate at which susceptible persons become infected after a meeting with an infected person.

For further simplification we assume that the number of newborn balances the number of dyeing people per time unit in order to neglect the birth and death in the city.

3.2 The stochastic disease model

Since the goal of this study is to follow discrete numbers of infected persons and to integrate noise, we obviously ended up with a stochastic model. For a proper stochastic simulation spatial homogeneity is assumed. This means the population in a city is well mixed and there are no spatial accumulations of either susceptible or infected persons.

In the SI model there is only one transition possible.

Where s denotes a susceptible person j or k and i denotes an infected person j or k. The propensity a is the probability per time unit for a successful infection. The propensity depends on the meetings between susceptible and infected persons as well as on the rate of successful infections after meetings.

Hypergeometrische Verteilung erklären...

Ansteckung bei Zusammentreffen von einem S und einem I -> mean +/- std \* error (resp. randn).

3.3 The network

Instead of modelling an existing city network an artificial scale free network was created. A scale free network was chosen, because it can be used to simulate airtraffic [watanabe] and we intended the formation of hubs.

The network was created around a seed of five nodes. Each new node i introduced into the network has a probability of

to connect with target node j. In this way nodes with a high degree kj are more likely to be selected and a scale free network is created. [mamo]

In a first attempt new nodes were allowed to form one connection when they were introduced. We decided that our network would look more realistic if loops were possible. Hence new nodes were allowed to form two bonds.

Picture Picture

It is well known that the population of cities can be described by a power law (Zipf’s Law). Prof. Guido Spars et al [] determined the Zipf parameters of german cities with high accuracy. We assigned a population to each node according to its rank in the network and the value predicted by Zipf’s Law

Our final network consisted out 10 000 cities and had a population of 93 Millions

Zipf’s law, Scale-free.

3.4 The transport

People from a city x are able to travel to city y if the nodes x and y are connected by an edge. However, we want the total population of a city to be constant. Therefore there the same number of persons travel from city x to city y as vice versa while the composition of the travellers (susceptible or infected) can vary. If a fixed number of total travellers per simulation day are assumed – like for example seats in airplanes – the portion of infected travellers is hyper geometrically distributed. Txy corresponds to the total travellers on the edge between city x and city y and is equivalent to the number of drawings without laying back out of a population with Sx,y + Ix,y people and Ix,y as the number of infected persons[[3]](#footnote-3).

|  |
| --- |
|  |
| **Fig. 1** Sketch of two cities x and y where S and I correspond to the number of susceptible or infected persons. Edges are indicated as arrows. T denotes the total number of travellers along an edge and is the same in both directions of the edge. |

The number of total travellers along an edge should obviously not exceed the population of either city connected by the specific edge. It must also be ensured that cities with multiple edges do not have higher traveller fluxes than populations. Also a factor depending on the degrees k of the connected cities compared to the average degree in the network was included to contribute more transports between cities with many connections – intuitively that is like cities with more airports and stations have more transport capacity. Finally we included a factor to reduce the transport of infected people assuming that people who feel sick travel less.

Actually, it would have been nicer if the degree factor would be divided by two. However, since the transport rate is further reduced[[4]](#footnote-4), this flaw is corrected.

# 4 Implementation

*4.1 Network generation*

Jordi ☺. Kannst du ev. Die edges- und die cities-Matrix erklaeren, aber vor allem beschreiben was die Funktionen machen. (Du darfst auch neue section machen)

4.2 Simulation of the infection

As the name indicates, the function Simulate\_Infection computes the number of infections per city and per time step. Therefore the number of meetings between infected and healthy persons is calculated according to an arbitrarily chosen mean and standard deviation of daily meetings with other persons. Further, a Gaussian distributed error was included. As long as the number of infected people in a city is below a certain threshold the infection is simulated in a stochastic manner. The number of meetings which result in an infection is a hyper geometrically distributed number, as mentioned in section 3.1. Because MATLAB is very slow in drawing hyper geometrically distributed random numbers, a binomial distributed random number was used as an approximation. The drawback of the binomial distribution in this situation is that a person can be infected more than one time. However, by using small time steps, the probability two infect two times the same persons becomes rather low. As soon as the number of infected people exceeds the threshold, the simulation works in a deterministic manner and is equal to the expected value.

4.3 Transport

According to section 3.4 the total number of travellers along every edge is calculated before the actual simulation. Then for every city and time step the number of infected travellers is simulated. The transport function uses the edges matrix and in order to avoid constraints from calculating always in the same order along the edges matrix, the rows of the matrix are shuffled in every time step. Again, we have hyper geometrical distribution for the number of infected travellers. So, the mean and the standard deviation of the hyper geometrical distribution are computed. This is done with the MATLAB function hygestat. A Gaussian distributed error is included to introduce noise in the transport. Finally, the number of infected persons in the cities is updated due the change through transport.

4.4 Experiments

Three simulations were performed to characterize the epidemic outbreak in the scale-free network. To start each simulation, the first infected person – the seed - was set randomly into a city. The parameters like infection probability and mean of daily meetings were chosen empirically and tested in smaller networks. Implementation of real parameter values would be interesting for further experiments and is discussed in the section 6.

In the first experiment 80 days were simulated and the number of cities with at least ten percent infected citizens was taken as output. Additionally, the ratio of total infected persons over the whole network population was computed. The aim was to compare the time courses of the epidemic in multiple simulations and to measure the temporal variations.

The second experiment was performed to test the hypothesis that the epidemic spreads faster when started big cities with high degrees than in small cities. Since, we assumed a high influence of the nodes connected to the seed, the sum of the degrees of the direct neighbours was considered too. In every simulation, the time until at least twenty cities had at least one infected person was measured.

In the third experiment, in addition to the seed, a random target city was chosen. In order to find the number of edges in the shortest path connecting the seed with the target city, which is the distance, we used the MATLAB Graph theory tool box. The function graphshortestpath solves the shortest path problem. Then, the1 time until at least one infected person appeared in the target city was measured.

# 5 Simulation Results and Discussion

|  |
| --- |
|  |
| **Fig. 2** This figure provides the correlation of the distance and the time needed for the disease to reach the random target city. The distance is the shortest path from the seed (random city where the first infected person was set) to a random target city and corresponds exactly to the number of edges, since we used un-weighted edges in our network. The error bars and the mean are shown in blue. The black line is the result of the linear regression analysis of the time measurements city (parameters of lin. regression). |

|  |
| --- |
|  |
| **Fig. 3** The mean of the number of cities with at least ten percent infected citizens after simulating 80 days is shown by the black line. In addition, the mean of the ratio of the total number of infected persons over the total network population is represented by the blue line. The means were calculated from the data of 147 simulations. |

|  |
| --- |
|  |
| **Fig. 2** This figure provides the correlation of the distance and the time needed for the disease to reach the random target city. The distance is the shortest path from the seed (random city where the first infected person was set) to a random target city and corresponds exactly to the number of edges, since we used un-weighted edges in our network. The error bars and the mean are shown in blue. The black line is the result of the linear regression analysis of the time measurements city (parameters of lin. regression). |

Expectancy range until infection of 1000 people = 18.2029 +/- 5.1615

# 6 Summary and Outlook

Outlook

In our experiment a completely fictional network was used, so obviously, the next step would be to implement real data and an existing network. For the network it would make sense to take a continent or even the whole world as long as only the important cities are considered, such that the performance does not suffer too much. It would be interesting to weight the edges according to geographic distances and to include real transportation data. Furthermore, the integration of either a healing rate or an immunisation would extend the simulation and make it more realistic.

We assume that infection simulation will gain further importance for the understanding of the spreading in network. Either one can think of diseases like H1N1 or sexually transmitted diseases like HIV. But also the spreading of information or malware in computer networks could be simulated with programs, such as ours.

# References

1. Contagion. Dir. Steven Soderbergh. Perf. Matt Damon, Kate Winslet, Jude Law. Warner Bros. Pictures, 2011. [↑](#footnote-ref-1)
2. Kermack, W.O. and McKendrick, A.G. "A Contribution to the Mathematical Theory of Epidemics." Proc. Roy. Soc. Lond. A **115**, 700-721, 1927. [↑](#footnote-ref-2)
3. The colon of Sx,y and Ix,y means „or“. [↑](#footnote-ref-3)
4. The factor is not shown in the equation above, but was included in the simulation (see the FINAL\_SIMULATION.m files). [↑](#footnote-ref-4)