Wrocław University of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Applied Mathematics
Specialty:

Engineering Thesis

MODELLING THE EPIDEMICS IN SOCIAL NETWORKS

Alicja Nowakowska

keywords:

temporal social networks, epidemics, simulation

short summary:
In this paper the diffusion process of the Influenza virus is investigated on the basis of the computer simulation that is performed for the real data set of contacts between people. Then the simulation results are investigated on the 2 levels. The first one aims to investigate the chances of the spread of the disease over the population. The second is focused on the impact of the individual infection potential on the dynamics of the diffusion process. The attempt to explain the obtained results by the network characteristics is taken.

Supervisor	dr inż. Radosław Michalski	....	....
	Title/degree/name and surname	grade	signature

For the purposes of archival thesis qualified to:*
   a) category A (perpetual files)
   b) category BE 50 (subject to expertise after 50 years)
* delete as appropriate

stamp of the faculty

Wrocław, 2019

Contents

List of Figures

List of Tables

Introduction

Without no doubt, the proper understanding of the epidemic process plays a crucial role for the global health. To gain this essential knowledge one needs to analyze the way the disease spreads and find characteristics of its diffusion. The common approach applied to realize this challenging task is based on the construction of the network model of the population and then execution of the computer simulation on it. The combination of the modern network science, informatics and mathematics guarantees the adequate investigation of the results.

Problem Statement

The goal of this paper is the analysis of the epidemic process of the influenza virus throughout the computer simulation based on the real data set of contacts between people during specific interval of time. The taken approach consists of the construction of the temporal social network that represents the relations in the population and then simulation of the disease diffusion on it. The results are analyzed on two levels:

1. Global- the analysis is focused on the overall impact of the disease on the population. It's checked whether the disease has an epidemic potential when random person is infected. Moreover, factors that can influence on its dynamics are investigated.

2. Local- the analysis is focused on the impact of the specific person on the diffusion process when infected. This perspective permits the identification of the individual characteristics that make a person potentially dangerous for the population.

The first chapter delivers basic theoretical background on which the next chapters are based on. The second centers on the presentation of the simulation, its assumptions and the algorithm as well as on the obtained results. The third one focuses on the analysis of the factors that influence on the epidemic dynamics.

Chapter 1

Theoretical background

In this chapter the most important concepts for this study such as networks and its characteristics, basic reproductive number, SEIR model and Multiple linear regression will be introduced.

1.1 Networks

1.1.1 Basic information

Generally speaking, network is a mathematical object $G(V, E)$ that consists of nodes V and links E between them. Specifically, links represent interactions between the nodes, can be directed or undirected and can have weight or not. Furthermore, we can distinguish two basic network types:

- static- network that is stable over time,

- temporal- network that changes over time.

Both types are highly connected ideas, since static one is constructed through aggregation of the temporal one (Figure 1.1). However, temporal networks have a significant advantage for the epidemic modelling because they permit preservation of the real contact structure.
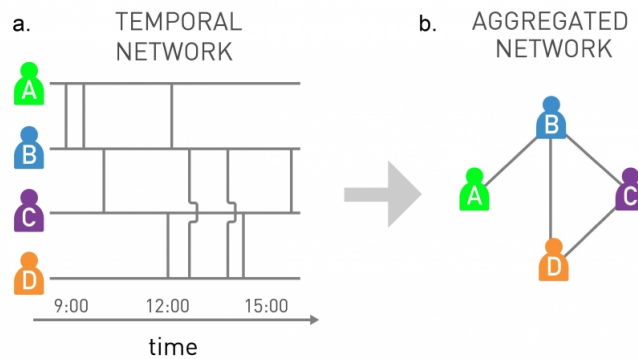
Figure 1.1: Relation between the static and temporal network[6]

1.1.2 Definitions

The following concepts are defined for the networks:

Definition 1.1 (Adjacency matrix)**.** The matrix representation of a network, denoted by A.

Definition 1.2 (Degree of a node)**.** Number of links that a node owns.

Definition 1.3 (Path)**.** An ordered list of links that connects two nodes.

Definition 1.4 (Shortest path)**.** Path between the node and u and v with the fewest number of links between them, denoted by $d(u, v)$[6].

Definition 1.5 (Clustering coefficient)**.** Shows how much the neighbors of a node are connected to each other. It's calculated for the node i using the formula:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{1.1}$$

where k_i is the number of neighbors of the node i and L_i represents the number of links between the neighbors of the node i[6].

Definition 1.6 (Centrality)**.** Degree of the importance of a node. Following definitions are measures of this characteristic.

Definition 1.7 (Degree centrality)**.** Degree of a node.

Definition 1.8 (Betweeness centrality)**.** Number of the shortest paths that go through the node divided by the total number of the shortest paths in the network[11].

Definition 1.9 (Random walk betweeness centrality)**.** Also called current flow betweeness centrality, considering the random walk, it's the expected number of pasess through the node[10].

Definition 1.10 (Closeness centrality)**.** Average length of the shortest path from the given node to every other node[8].

Definition 1.11 (Harmonic centrality)**.** Variation of the closeness centrality defined for the node i as[2]:

$$H_i = \sum_{u \neq v} \frac{1}{d(u, v)} \tag{1.2}$$

Definition 1.12 (Eigen vector centrality)**.** Having an equation

$$Ax = \lambda x, \tag{1.3}$$

where λ is the biggest eigen value of the matrix A, then the i-th element of a vector x is an eigen vector centrality measure of a node i[3].

One should also remember that in case of the networks whose links have the weight attribute, the centrality measures are calculated with the algorithms that respect that fact.

1.1.3 Scale-Free property

The next important topic for the network science is the Scale-Free property. Frequently, to obtain the valuable information about the network one is interested in the distribution of the degree of the node. It was observed that for many real networks such as social networks, World Wide Web or protein-protein interactions network, the certain special pattern in its structure exists. Specifically, in these networks the vast majority of the nodes have a rather similar low degree and very few nodes posses surprisingly high number of connections (they are called hubs). This network property is named Scale-Free property. For the networks that have it, the probability that a node has k number of links- p_k is inversely proportional to the k, what can be written as:

$$p_k \sim k^{-\gamma} \tag{1.4}$$

where γ is some specific number for the network called the degree exponent. Such kind of the distribution is named the Power Law distribution. If the logarithm is taken for the both sides the following linear relation is obtained:

$$\log p_k \sim -\gamma log k. \tag{1.5}$$

If one wants to check whether the network is characterized by the scale-free property, the graph of $\log p_k$ versus $\log k$ shall be plotted. If the linear relationship is observed then one may assume the hypothesis to be true[6]. Moreover, it's worth to remember that random networks do not posses this property.

1.2 Epidemiology

1.2.1 SEIR model

The most common concept for the epidemics modelling is the division of the society into groups depending on the state of the individuals. The most popular model is named SIR, and labels each person as one of the following:

- Susceptible- if they can get infected.

- Infectious- it they are infected.

- Recovered- if they were infected and now are immune to the disease.

The SEIR model that is used in this paper is its variation and assumes additionally the existence of the incubation period. It means that the Exposed group- ones who are infected but not yet infectious, is also incorporated. These compartment models can be represented in the form of differential equations, stochastic equations or as the computer simulations. They aim to predict and model the disease spread.

1.2.2 Basic reproductive number

The fundamental idea for the study of the epidemics is the basic reproductive number R_0, which is the average number of the secondary cases that the infected individual produces.

Considering the network that represents relation between people and assuming hat the individual will recover with probability equal to 1, it can be defined mathematically as[14]:

$$R_0 = \beta L k (1 + \frac{\sigma^2}{k^2}), \tag{1.6}$$

where

L - time that the node is infectious

β - probability of the disease transmission per unit time

k - average number of contacts or average degree of the node

σ^2 - variance of the distribution of the degree of the node or its number of contacts

Moreover, every disease has its own R_0 that depends on its characteristics. If the $R_0 > 1$ the disease will spread and persist in the group, whereas if $R_0 < 1$ it won't. One may also be interested in the equation for the probability of the disease transmission per contact (βL) which is easily obtained as:

$$\beta L = \frac{R_0}{k(1 + \frac{\sigma^2}{k^2})}. \tag{1.7}$$

1.3 Multiple Linear Regression

In this paper Multiple Linear Regression is a tool that helps to combine the network science and the epidemiological characteristics. Multiple Linear Regression can be described as a mathematical model that provides formula for the response variable in function of the explanatory variables. In other words, it aims to describe the relationship between them. It can be defined by the following equation[12]:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_p X_p + \epsilon \tag{1.8}$$

where:

- Y is a response variable,

- $X_1, X_2, \cdots, X_n$ are explanatory variables,

- $b_1, b_2, \cdots, b_n$ are coefficients of the relationship between Y and $X_1, X_2, \cdots, X_n$ respectively,

- A is an intercept,

- ϵ is an error.

This equation can be rewritten for the observation Y_i as:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + b_3 X_{i,3} + \cdots + b_n X_{i,p} + \epsilon_i, \tag{1.9}$$

where $i \in \{0, 1, \cdots, n\}$. However, when constructing a Multiple linear regression models on the basis of the real data one deals only with the estimation of b denoted as $\hat{b}$ and estimation of the error $\hat{\epsilon}$ also called residual.

1.3.1   Model assumptions

To consider a model valid one needs to check whether the following assumptions are fulfilled[12]:

1. Linearity- Y has a linear relationship with the explanatory variables.

2. Homoscedasticity- the variance of the errors is constant.

3. Independence- errors are independent from one another.

4. Normality- errors have normal distribution $N(0, \sigma^2)$.

In practice, as already mentioned one does not know the errors but only the residuals and verifies the assumptions for them.

Furthermore, when choosing the explanatory variables one should avoid the selection of the correlated variables since the problem of the multicollinearity may appear. In consequence, it may be hard to distinguish which of them explains the best the model and it makes more complicated its proper optimization.

1.3.2   Model selection

It happens frequently that one needs to deal with the problem of the model selection. The general idea is to find the model with low complexity (minimization of the number of parameters) and at the same time good performance and adjustment. When facing this task statistical tests and measures that provide the information about the model characteristics should be used.

F-Test

When deciding whether the reduction of some variables is statistically justified the F-test may be applied. It's a statistical test that is used to test the null hypothesis that certain subset of the b coefficients is equal to 0. The size of this subset is denoted as q. The test statistic is of the form[12]:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p)} \tag{1.10}$$

where:

- $RSS_0 = \sum_{i=0}^{n}(\hat{Y}_i - m)^2$ and m is the mean of the all observed values Y_i,

- $RSS = \sum_{i=0}^{n}(\hat{Y}_i - Y_i)^2$.

If the F statistic belongs to the F-distribution with $(q, n-p)$ parameters, then there are no grounds to reject the null hypothesis. Alternatively, if the F statistic adopts high values comparing to those from the F- distribution, then the null hypothesis is rejected.

Global criterions

During the selection of the best regression model it turns out to be convenient to create global measures of the information that it provides. Two most common measures of this type are presented.

Definition 1.13 (Adjusted R^2)**.** It's defined by the formula[12]:

$$\text{adjusted } R^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} \tag{1.11}$$

where: $TSS = \sum_{i=0}^{n}(Y_i - m)^2$. The higher its value, the better model is obtained.

Definition 1.14 (Akaike Information Criterion)**.** It has the form[9]:

$$AIC = n\ln(2\pi) + n\ln(\frac{RSS}{n}) + n + 2p. \tag{1.12}$$

The lower its value, the better model is obtained. This criterion is also widely used when comparing hierarchical models. Particularly, if one deals with a complicated model with many variables and wants to find those that describe the best the response variable, then it's recommended to use for example the Backward method that is based on the AIC criterion. The procedure begins with calculation of the AIC of the model with all variables. Then p different models are created. Each model does not have one different explanatory variable but has all others. The model with the best AIC is chosen and the procedure is repeated and continues as long as one can obtain better AIC due to the variable elimination[4].

Chapter 2

Simulation

In this chapter the methodology of the epidemic simulation is presented. Both the algorithm for the calculation of the probability of the disease transmission per contact and the algorithm of disease spread is introduced. Next, the simulation is performed and its results are analysed.

2.1 Data

The data that is used for the simulation is a temporal network of contacts between 92 different people observed in the office building in France during ten days from June 24 to July 3, 2013, two of those days were days off. It has the tab- separated format with the rows of the form $[ID_1, ID_2, t]$, where ID_1 and ID_2 are the IDs of the persons in contact and each contact had a duration of 20 seconds from $t - 20s$ to t. The number of contacts equals 9827[5].

2.2 Assumptions

The disease that will be analyzed in this paper is Influenza which is contagious and transmitted via airborne route with the R_0 between 2 and 3[7]. Moreover, it's assumed that everyone in the population is adult and healthy and then it's given that the person is infectious during 7 days with 1 day of the incubation period[1]. The model used for the simulation is the compartment SEIR(Susceptible, Exposed, Infectious, Recovered) model. At the beginning of the simulation everyone is Susceptible. Each person represented by their unique ID is considered as the node and each contact as the link between the nodes. Regarding the following section Calculation of the probability of the disease transmission, the probability of the disease transmission per contact in the simulation is equal to 0.006.

2.3 Calculation of the probability of the disease transmission

2.3.1 Algorithm

To calculate the probability of the disease transmission for the considered network the 1.4 formula will be used. It's assumed that each node in the network posses the k_{node} attribute

which denotes the number of contacts of the node, at the beginning of the simulation k_{node}=0. The following pseudo-code describes the probability calculating procedure.

Data: the contact list, R_0
Result: the vector of the probabilities of the disease transmission per contact- v
Create the vector k_{nodes} that contains k_{node} value of every node;
Create an empty vector v;
for *every different moment t from the contact list* **do**
    for *every contact of the form [node₁, node₂, t] that took place in the t moment:*
    do
       | $k_{node_1}+ = 1$ and $k_{node_2}+ = 1$;
    end
    Calculate the variance σ^2 and average k of k_{nodes}.;
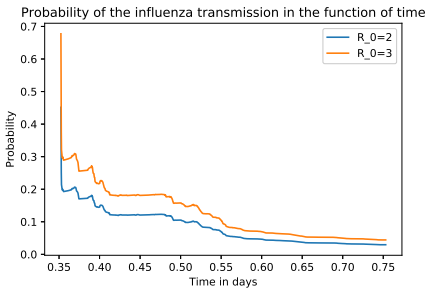   Calculate the $\beta L = \frac{R_0}{k(1+\frac{\sigma^2}{k^2})}$ and add it to the v.
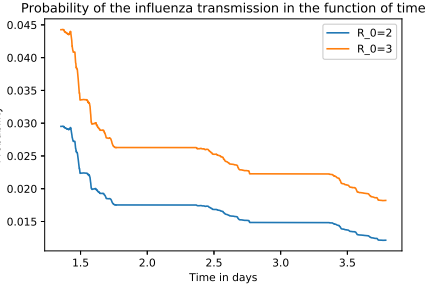**end**
Algorithm 1: Calculation of the probability of the disease transmission in time
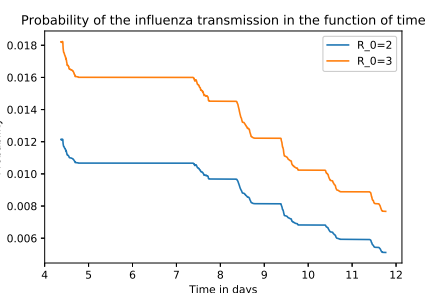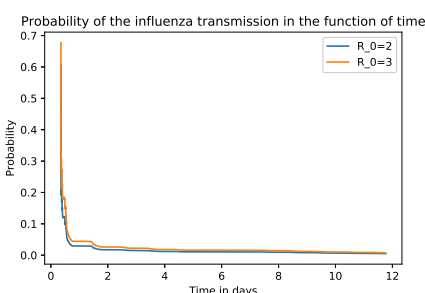
---

However, one must remember that the first obtained probabilities are incorrect since their value is more than 1 due to the few contacts considered at the beginning. To have a valuable result, the contacts list (or equivalently in the studied case, time of gathering the data) must be long enough.

2.3.2   Results

According to the algorithm, the probability of the disease transmission per contact in function of time for $R_0 = 2$ and $R_0 = 3$ was obtained (Table 2.1). The probability in the first minutes is not showed since it's incorrect.

Table 2.1: Graphs of the probability of the disease transmission per contact in time

Day	Graph	probability for $R_0 = 2$	probability for $R_0 = 3$
1st day		at the end of 1st day equals 0.03	at the end of 1st day equals 0.04
from 2-4		at the end of the 4th day equals 0.012	at the end of the 4th day equals 0.018
from 5-12		at the end of the last day equals 0.005	at the end of the last day equals 0.007
all		at the end of the last day equals 0.005	at the end of the last day equals 0.007

Moreover, the histograms of the obtained probabilities for both R_0 were created (Figure 2.1)
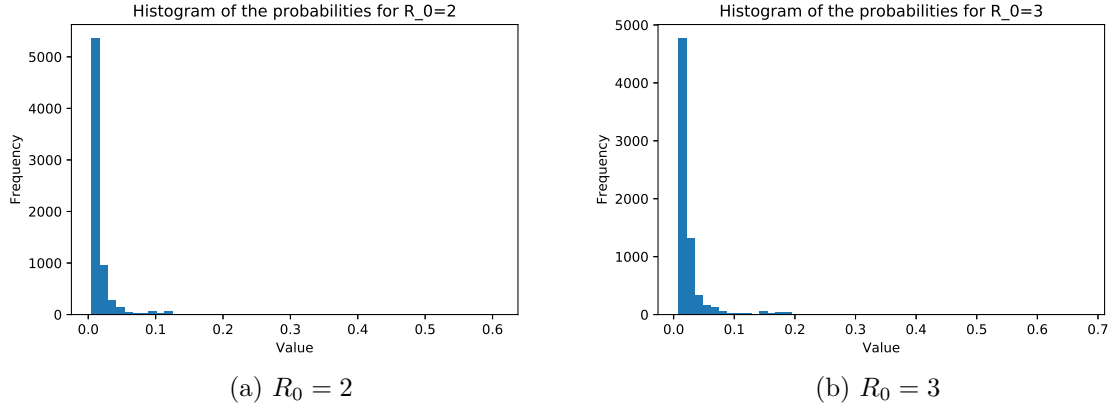

(a) $R_0 = 2$ (b) $R_0 = 3$

Figure 2.1: Histogram of the probabilities

2.3.3   Observations and conclusions

It can be noticed that in the both cases the probability of the disease transmission decreases over time. Taking a look at the Figure 2.2 one may see that the relation between the variance σ^2 and k has some slightly quadratic character. The consequence of this observation is that σ^2 grows faster in time than k and for that reason the denominator in the formula for the probability of the disease transmission per contact (1.4) increases in the function of time. However, taking into account the histograms, it's noted that the
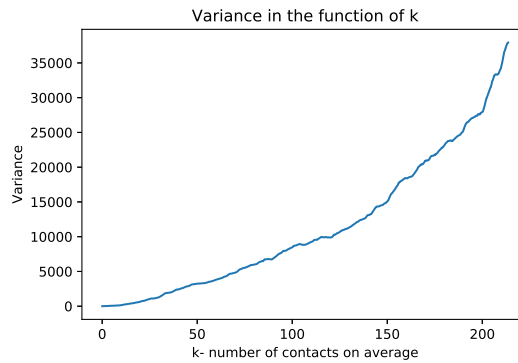
Figure 2.2: Variance in the function of k

majority of the calculated probabilities belongs to the first highest bin which corresponds to the probability 0.005 in case of the $R_0 = 2$ and 0.007 for $R_0 = 3$. Surprisingly, those results are very similar to the ones obtained in the other article where the probability of the disease transmission was calculated as 0.003[13]. In the following sections and chapters it will be assumed that the probability of the disease transmission per contact is equal to 0.006.

2.4 Simulation algorithm and implementation

2.4.1 Implementation

The implementation of the simulation is based on the Networkx library, available in Python programming language, designed for the network investigation. Parting from the contact list, the network with the attributes S,E,I,R that denote number of people that belong to the Susceptible, Exposed, Infectious and Recovered group respectively, is created. The network consists of the *node* objects and each *node* has the following attributes:

- ID- unique ID of the person,

- *state*- one value from the set {'S','E','I','R'} that corresponds to the group that the node belongs to,

- $InfectionTime$- at the beginning of the simulation for all nodes its value is $False$, if the node is Exposed in the time T then $InfectionTime = T$.

2.4.2   Algorithm

The following algorithm describes the simulation procedure for the spread of the disease in the population.

Data: the contact list, p - probability of the disease transmission per contact, network with $S = 92$, $E = 0$, $I = 0$, $R = 0$, *nodes*- vector that consists of all nodes present in the network, *time* $= 0$

Result: S, E, I, R

Choose the *node* that will be exposed, $S- = 1$, $E+ = 1$, $node[state] = $ 'E' ;

for *every contact of the form [node$_1$, node$_2$, t]* **do**

    time=t;

    for *every node from the nodes* **do**

        if *node[state]=='E'* **then**

            if *the incubation period for the node has passed* **then**

              | $node[state]=$'I', $E- = 1$, $S+ = 1$

            end

        end

        if *node[state]=='I'* **then**

            if *the infection time for the node has passed* **then**

              | $node[state]=$'R', $I- = 1$, $R+ = 1$

            end

        end

    end

    if *node$_1$[state]= 'I' and node$_2$[state]='S'* **then**

        U- random variable from the uniform distribution with 0 and 1 as parameters;

        if *U<=p* **then**

            | $node_2$[state]='E', $node_2$[InfectionTime]=time, $E+ = 1$, $S- = 1$

        end

    end

    if *node$_2$[state]= 'I' and node$_1$[state]='S'* **then**

        U- random variable from the uniform distribution with 0 and 1 as parameters;

        if *U<=p* **then**

            | $node_1$[state]='E', $node_1$[InfectionTime]=time, $E+ = 1$, $S- = 1$

        end

    end

end

Algorithm 2: Simulation

Furthermore, the statement '**if** the infection time for the *node* has passed' can be rewritten as:

$$\textbf{if } node[InfectionTime] + 1 \text{ (length of the incubation period)} <= time,$$

and similarly '**if** the infection time for the *node* has passed' as:

$$\textbf{if } node[InfectionTime] + 1 + 7 \text{ (infection period)} <= time.$$

2.5 Selection of the exposed node

To perform the simulation one must choose the first exposed node and that can be done at random or not. Each approach reveals different epidemic and network characteristics.

2.5.1 Random

Random selection of the exposed node permits the observation of the overall epidemic behaviour. Particularly, by performing $N = 1000$ Monte Carlo repetition of the disease simulation with always randomly chosen the exposed node at the beginning, the general chances for the disease spread can be discovered.

   The first exposed node in every epidemic simulation was selected by the following procedure:

1. *A*- random integer from the set {0, 1, 2,...,92}.

2. The first exposed *node* is *nodes*[*A*].

   Later, to analyze the range of the disease spread, after each repetition the $M = 92 - S$ number was saved. It indicates how many people at the end of every simulation were out of the Susceptible group, or equivalently how many people got infected (Infectious and Recovered group) or were going to get infectious (Exposed group). Then, the histogram of the M values was created (Figure 2.3). It can be concluded that in 78 % of the cases the
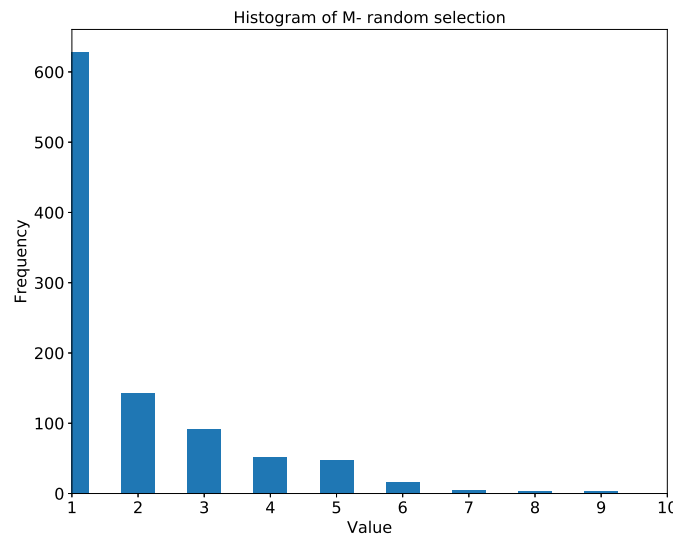

Figure 2.3: The histogram of M for the simulation with the random selection

disease does not spread or is transmitted to only one person. In the rest 22 % of the cases the disease spreads over population, and rarely has a potential to infect even up to 10 % of the population during the 12 days of observation.

2.5.2 Not random

Not random selection of the first exposed node permits the investigation of the infection potential of every node in the network. Specifically speaking, by performing many times

the epidemic simulation for the particular node, one may find out how many cases the node is capable to produce. In order to realize this task for each node the simulation was performed $N = 1000$ times and again the number $M = 92 - S$ was saved each time. Later, on this basis two statistics for each node were created:

- chance- probability that the specific node will produce more or equal to 2 cases (the disease will spread). In order to calculate this value, it's checked how many times $M \geq 3$ and then the obtained number is divided by N.

- mean- the average number M that occurs for the node after performing N simulation for it.

Both histogram of the mean and chance attribute of the nodes were created (Figure 2.4). As the results from the previous section indicated, only few nodes- around 10% have a

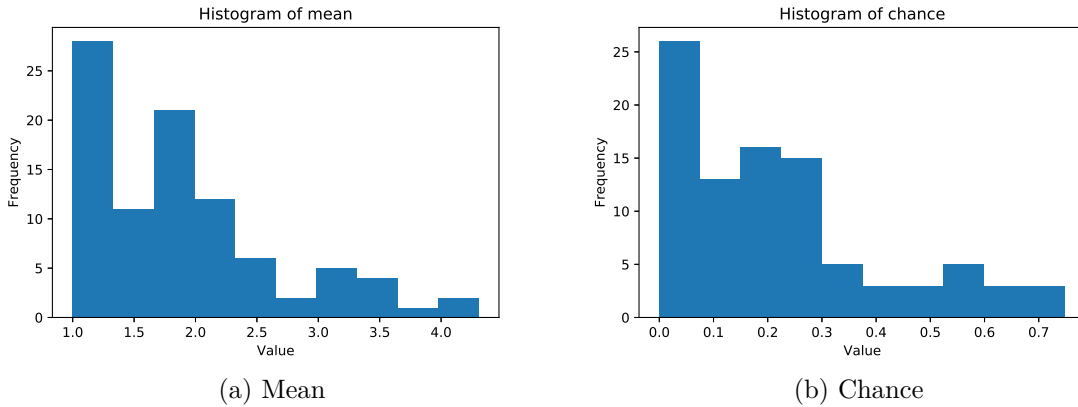(a) Mean   (b) Chance

Figure 2.4: Histograms of the mean and chance attributes

high infection potential and with the probability more than 0.5 provoke the spread of the infection. Similarly around 10% of the nodes have a high mean- over 3 (more than 2 secondary cases). That explains why so rarely the epidemic spreads over the network.

2.5.3   Nodes with the highest infection potential

In order to analyze what range the epidemic can have in the worst-case scenario the histograms of the M value for the nodes with the highest infection potential (with *chance* around 0.7) were created (Figure 2.5). It can be noticed that these nodes are capable to infect even up to 13 persons in 12 days. Moreover, it's possible that the contact structure has some importance for the simulation since in the case of the node with the $ID = 95$ it almost never happens that it'd infect 3, 5 or 9 people. The histograms for the other two nodes are fairly regular. The bigger value the lower probability of obtaining it.
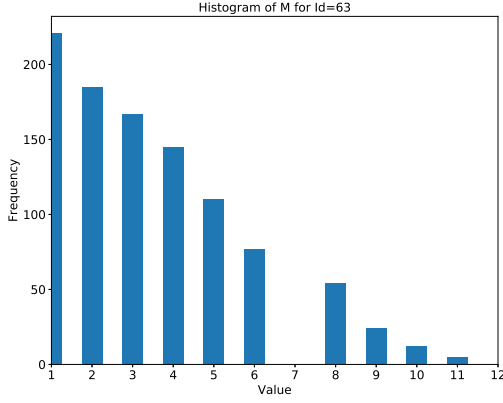
(a) *chance* = 0.75
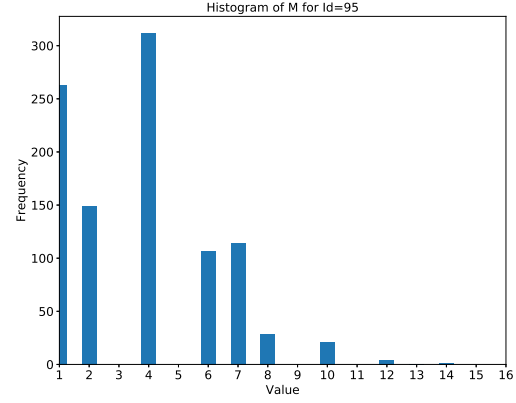
(b) *chance* = 0.74
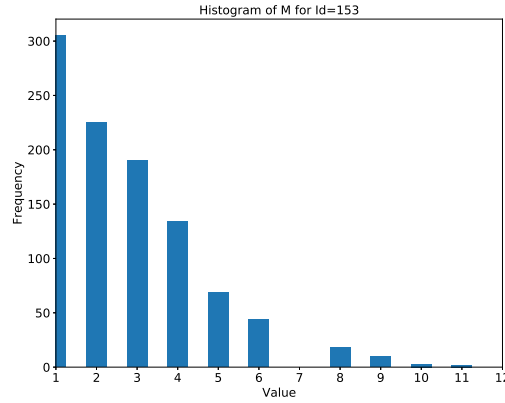
(c) *chance* = 0.7

Figure 2.5: Histograms of the M value for the nodes with the highest infection potential

2.6   SEIR graphs

Graphs that present the change of the number of people belonging to the compartment groups reveal the dynamics of the disease spread. For that reason, some example charts of this type for the nodes with the highest infection potential were produced (Figure 2.6). It can be easily noticed that the length of the experimental data does not permit the observation of the further spread of the disease, which would probably persist in the population for some time, in case of the example charts for the node 63 and 95. In the case of the example for the node 153 the all cycle of the epidemic spread is presented on the graph. Example graphs for the nodes with the low infection potential are not presented since they usually show only one or two people infected who do not pass the disease and due to that are less useful for the analysis of the disease spread.
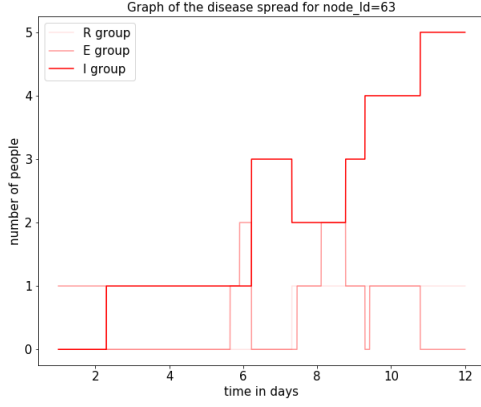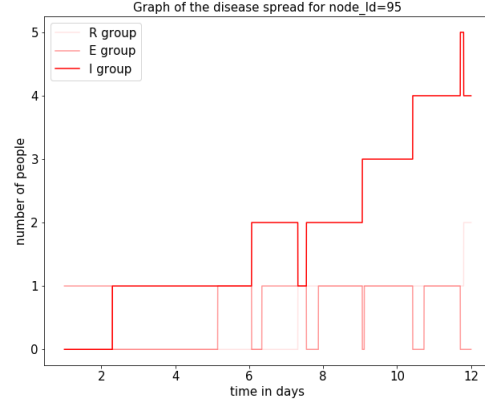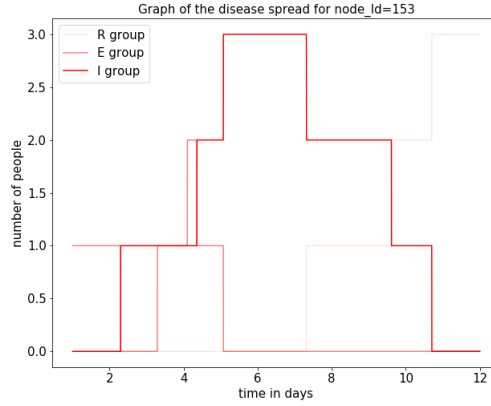
(a) *chance* = 0.75

(b) *chance* = 0.74

(c) *chance* = 0.7

Figure 2.6: Example SEIR graphs

2.7   Conclusion

Although the contact data is fairly short as for the analysis of the disease process, it can be concluded after this chapter that the spread of the disease over the population is quite an unlike event if we consider the random selection of the first exposed person. However, it's noticed that the selection of the first infected person plays an important role in this process. In other words, some nodes have a high infection potential and if infected can provoke the epidemic in the short interval of time. The reason for that may be that they have some special characteristics. The second observation about the dynamics of the disease diffusion is that the given contact structure may have an impact on the obtained results. The aim of the next chapter is the analysis of these hypothesis.

Chapter 3

Factors that influence the epidemic dynamics

In this chapter the characteristics that influence on the epidemic spread process on the both local and global level are investigated.

3.1 Impact on the global dynamics

According to the previous chapter, the factor that may have an impact on the disease spread process and its further analysis is the contact structure. In order to confirm or reject that hypothesis, the simulation results obtained for the original data contact structure shall be compared with its randomized modification.

3.1.1 The algorithm for the randomization

The randomized contact structure is generated by the following algorithm:

> **Data:** the contact list and network
> **Result:** new random contact list
> Create the empty vector l which will be the new contact list;
> **for** *every contact of the form [node$_1$, node$_2$, t]* **do**
> > new_1- the random node from the network's *nodes*;
> > new_2- the random node from the network's *nodes*;
> > add to l [new_1, new_2, t]
>
> **end**

Algorithm 3: Contact structure randomization

This procedure creates the new contact list where all persons have an equal chance for the contact. In other words there's no preferential attachment, the hubs won't exist and therefore the scale-free property is not preserved.

3.1.2 Contact structure comparison
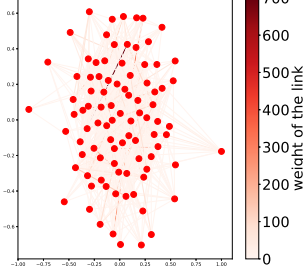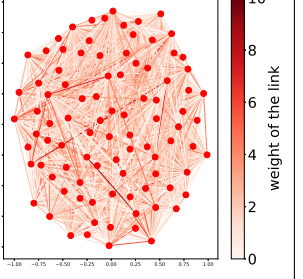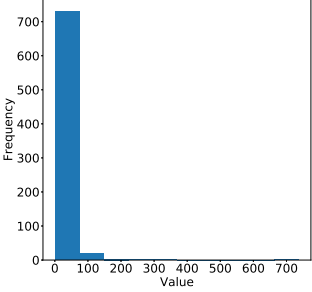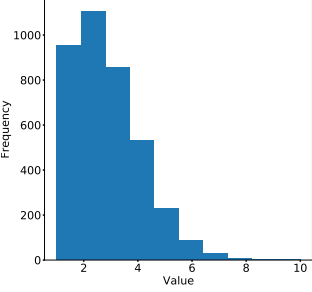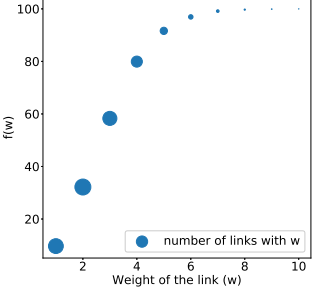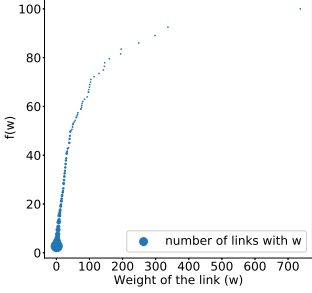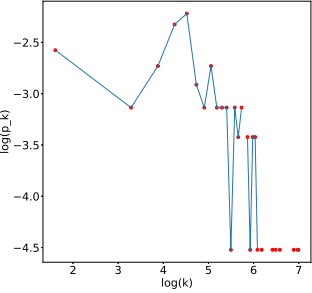
Before analyzing and comparing the simulation results for both contact structures, it's worth to firstly examine what is the difference between them. The best approach for the realization of this assignment is the creation of the static network on the basis of the contact list that represents the temporal one. Specifically, the nodes in the static network are all nodes that appeared in the temporal network (equivalently all persons that

appeared in the contact list) and links between nodes appear if the contact ever existed. However, in order to persist the contact structure in some way, all links have a weight attribute which denotes how many times there was a contact on it. Now, it's possible to compare the original and random contact structure what is done on 4 levels: visualization of both networks, analysis of the histogram of weights, analysis of the plots of the function $f(w)$ defined as :

$$f(w) = \frac{\text{number of contacts with weight} \leq w}{\text{total number of contacts} \cdot 100} \tag{3.1}$$

and by taking a look at the scale-free property where the overall number of contacts of a person at the end is now considered as the node's degree (Table 3.1).

Table 3.1: Contact structure comparison

Graph	Original contact structure	Random contact structure	Observations
Network visualization			The network for the random structure seems to have much more regular character (each node has similar number of links of the similar weight) than for the original one.
Histogram of weights			The histogram pattern is different. In case of the original contact structure the hubs seem to exist since the outliers take very high values comparing to the rest. The histogram for the random one has a regular character.
$f(w)$			In the both cases the links with the low weight are the most common ones. Nevertheless, it's noted that for the original one, the links with the very high weight (above 200) represent 20% of the all contacts.
Scale free property.		Does not posses, since all nodes have a similar degree and there are no hubs.	The graph for the original contact structure does not clearly indicate the existence of the scale free property (no linear relationship demonstrated), probably due to the rather small size of the network. However, it should have it since the majority of the nodes have the similar low degree and there are hubs.

It can be noted that both contact structure differ from one another significantly and in consequence produce different networks. Moreover, their layouts have an impact on the distribution of the chance attribute.

3.1.3   Comparison of the simulation results

The next step is the comparison of the simulation results for the both contact structure. Given the modified contact list the disease spread simulation with the random node selection was performed $N = 1000$ times as in the section 2.5.1. Later, the histograms of the M value were created and compared with the results for the original data (Figure 3.1). One may assume after analyzing them that the contact structure does not impact significantly on the results of the simulation, since the overall histogram construction is preserved. The most frequent event in the both cases is the infection of few individuals and it rarely happens that the range of the disease spread is wide. However, the difference exists, though it's quite slight, and can be noticed in the height of the bins. It's deduced that the random selection of the individuals for the contact list with the preservation of the time structure may be used for the analysis of the overall disease impact on the social network.

(a)   (b)
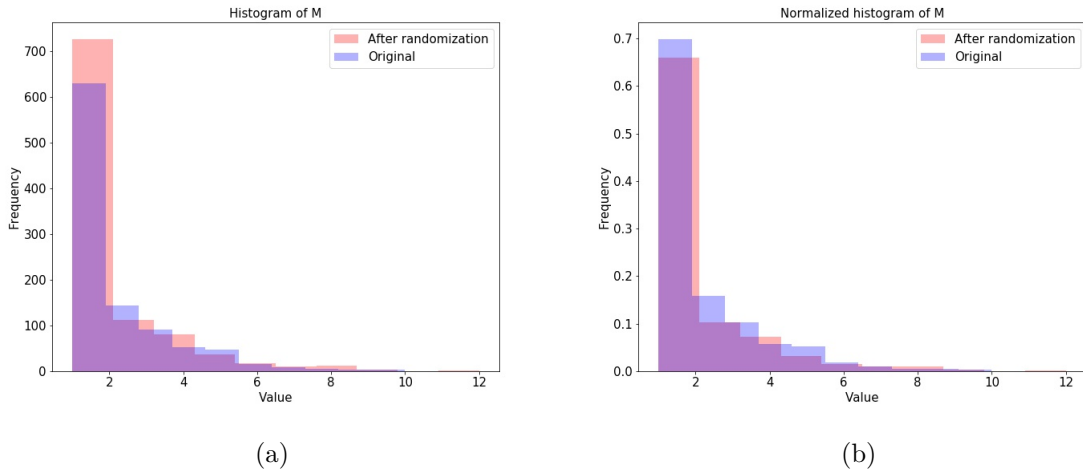
Figure 3.1: Histograms of the M value

3.1.4   Conclusion

After this section, it can be concluded that the contact structure proceeding from the data set significantly differs from the random one. Moreover, it seems that it demonstrates the scale-free property since it posses hubs and the majority of the links have a small number of contacts. Unfortunately it can't be seen on the proper graph because the network size is probably too small. Nevertheless the simulation results indicate that the overall epidemic diffusion can be well approximated by the randomized one if the time structure is preserved.

3.2 Impact on the individual

The second level of the investigation of the disease diffusion characteristics is the local one. Specifically speaking, the factors that influence the person's infectious potential are of the interest. It's assumed that the chance attribute of the node is the measure of its infection potential. Now, to find the characteristics that influence this value two approaches will be taken. First, the visual one and second, the creation of the mathematical formula for the infection potential.

3.2.1 Visual analysis

When looking for the nodes' characteristics that influence their infection potential it is first worth to visualize the network and look for some possible patterns. Although, the temporal network is a model that adjusts more to the reality than the static one, it does have some disadvantages due to its higher complexity. For that reason in order to analyze the network visually, it's easier to construct its simplified version- the static network with links weight, what was already done in the previous section. However, the large number of links obscures the identification of the possible location characteristics that impact on the infection potential. Therefore, the network with the reduced number of links (those that have weight more than 4) was created (Figure 3.2). This kind of simplification aims to highlight the important connections. When investigating the obtained graph, one may notice that nodes with the high infection potential frequently have connections of the big weight between each other. Similarly, nodes with the low infection potential, generally tend to have less connections with rather low weight. It seems that the majority of the nodes belong to the second group and only few represent the hubs characteristics. This pattern can be also noticed on the histograms of the number of contacts of the nodes presented on the Figure 3.3 i), j), k) and l) as well as on the distribution of the chance and mean attribute- a few number of nodes have a high infection potential and the majority does not. This is due to the fact that the observed network posses the scale-free property what is evident in its structure. These observations can indicate that both the number of contacts of the node and its location in the network influence on its infection potential. Equivalently, it can be said that the contact structure has an impact on the node's chance attribute what will continue to be investigated in the next sections.

3.3 Formula for the infection potential

Having in mind previous observations, the attempt to obtain the mathematical formula that would describe the chance attribute in the function of some variables is taken. To realize this task the Multiple linear regression model is constructed.

3.3.1 Model variables

The Multiple linear regression requires to firstly define the response variable (Y) which is the chance attribute in the analyzed case. Since, the centrality measures inform how important a node is for the network, due to its location, they are incorporated as the explanatory variables (X) in the model. Their values are calculated on the basis of the created static network with the all connections. Moreover, having in mind that the number
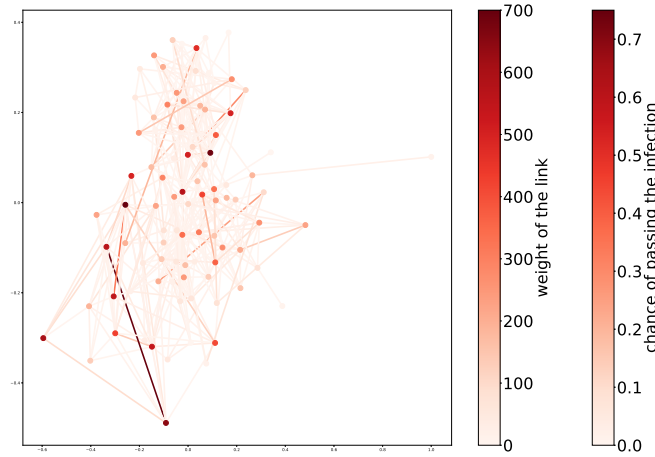
Figure 3.2: Network visualization at the end

of contacts may also influence on the infection potential of the node, k_{node} attribute's value in different interval of time is taken into account. Specifically:

- $k3$- denotes number of contacts of the node until the 3rd day,

- $k6$- denotes number of contacts between 3rd and 6th day,

- $k9$- denotes number of contacts between 6th and 9th day,

- *kend*- denotes number of contacts between 9th and 12th day.

In total it gives 12 possible explanatory variables to be considered. The distribution of their values is showed on the Figure 3.3. It's noted that the histograms for the centrality measures sometimes have a regular character as happens for the harmonic centrality (distribution similar to normal) and closeness centrality. The opposite occurs for the degree and information centrality. The first 3 histograms of the number of contacts are very similar to each other and represent the same pattern. On the last stage of the simulation the contact frequency seems to increase what impacts on the wider range of the outliers. Now it's possible to construct the Multiple linear regression model.

(a) Betweeness centrality

(b) Closeness centrality

(c) Harmonic centrality

(d) Information centrality

(e) Degree centrality

(f) Eigen vector centrality

(g) Random walk centrality

(h) Clustering coefficient

(i) $k3$

(j) $k6$
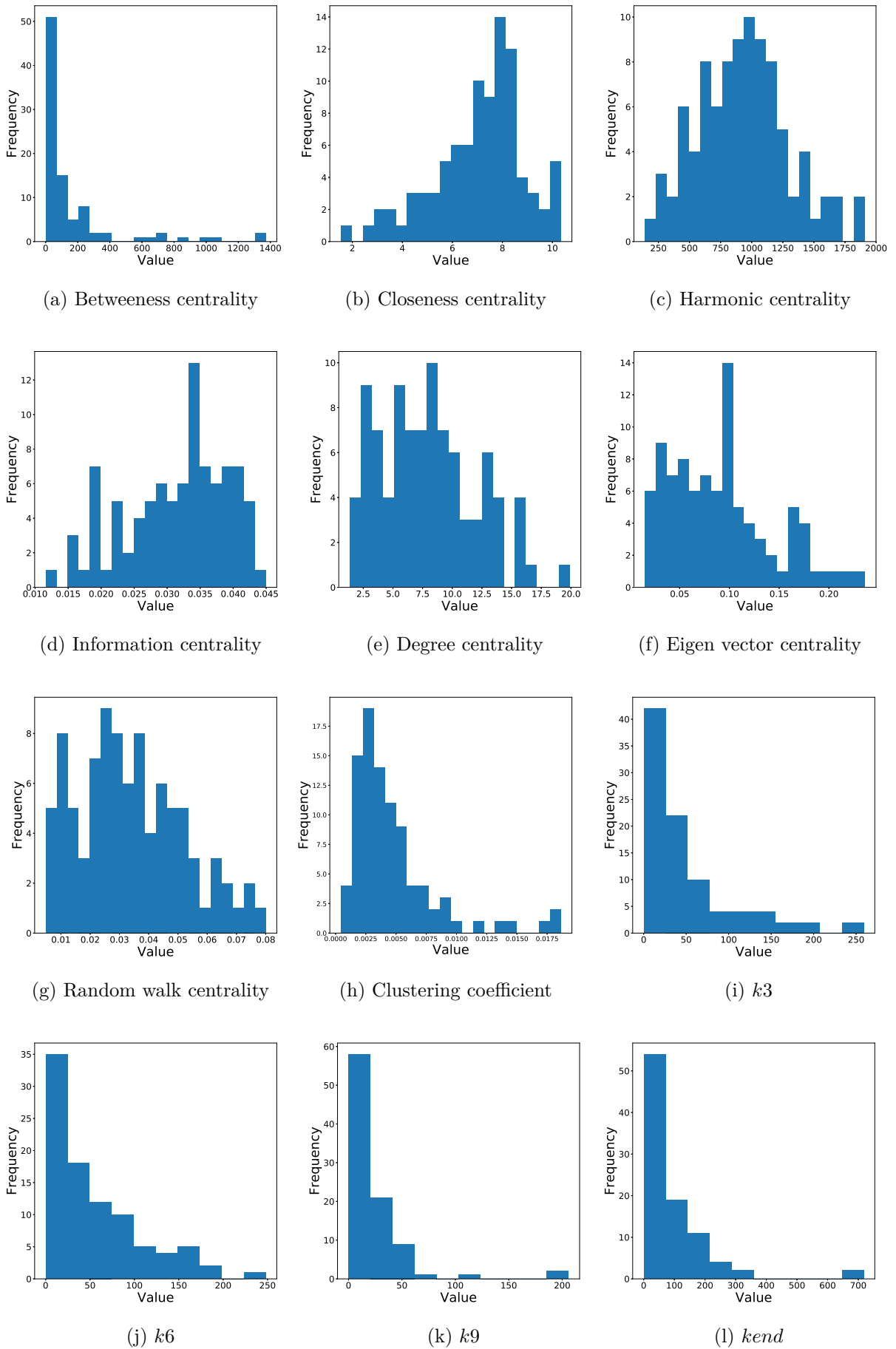
(k) $k9$

(l) $kend$

Figure 3.3: Value distribution histograms

3.3.2   Model with the multicollinearity problem

The first attempt to construct the Multiple linear regression model for the data, consisted of the incorporation of the all possible explanatory variables and then its optimization according to the Backward method based on the AIC criterion. The obtained model is showed on the Figure 3.2. The results indicate that the problem of the multiocllinearity of the explanatory variables may exist. It is due to the fact that the estimated values of the coefficients for the number of contacts at the end, closeness centrality and eigen vector centrality are under 0 and in the effect cause that the chance attribute instead of increase when they rise, it goes down. Probably, the incorporation of many correlated variables obscures the proper estimation of the coefficients and analysis of the model, although its global criterion have very good values: $AIC = -507$ and adjusted $R^2 = 0.89$.

Table 3.2: The best obtained model

	Estimate	Std.Error	p-value	$Pr(> |p|)$
Intercept	0.02	0.03	-0.5	0.6
betweeness centrality	0.00007	0.00003	2.6	0.1
clustering coefficient	11.6	2.8	4	0.0001
degree centrality	0.02	0.006	3.3	0.002
closeness centrality	-0.05	0.009	-5.6	$3.3 \cdot 10^{-7}$
eigen vector centrality	-0.96	0.005	-2	0.04
harmonic centrality	0.0004	0.00005	7.6	$3.5 \cdot 10^{-11}$
k6	0.001	0.0002	6	$4.4 \cdot 10^{-8}$
kend	-0.0005	0.00007	-6.8	$1.7 \cdot 10^{-9}$

To confirm the hypothesis of multicollinearity the matrix of the correlation of the explanatory variables was created (Table 3.4) .

Variable	betweeness centrality	clustering coefficient	degree centrality	closeness centrality	eigen vector centrality	random walk centrality	information centrality	harmonic centrality	k9	k3	k6	k_end
betweeness centrality	1	-0,02	0,3	0,57	0,23	0,32	0,29	0,53	0,35	0,31	0,22	0,41
clustering coefficient	-0,02	1	-0,38	0,29	-0,4	-0,35	-0,38	0,45	0,16	0,4	0,08	0,38
degree centrality	0,3	-0,38	1	0,41	0,96	0,97	0,95	0,33	0,29	0,3	0,07	0,18
closeness centrality	0,57	0,29	0,41	1	0,37	0,42	0,43	0,89	0,45	0,49	0,27	0,48
eigen vector centrality	0,23	-0,4	0,96	0,37	1	0,87	0,91	0,28	0,27	0,26	0,08	0,11
random walk centrality	0,32	-0,35	0,97	0,42	0,87	1	0,93	0,35	0,28	0,32	0,05	0,22
information centrality	0,29	-0,38	0,95	0,43	0,91	0,93	1	0,37	0,3	0,32	0,11	0,22
harmonic centrality	0,53	0,45	0,33	0,89	0,28	0,35	0,37	1	0,62	0,67	0,32	0,76
k9	0,35	0,16	0,29	0,45	0,27	0,28	0,3	0,62	1	0,45	0,08	0,42
k3	0,31	0,4	0,3	0,49	0,26	0,32	0,32	0,67	0,45	1	0,05	0,52
k6	0,22	0,08	0,07	0,27	0,08	0,05	0,11	0,32	0,08	0,05	1	0,18
k_end	0,41	0,38	0,18	0,48	0,11	0,22	0,22	0,76	0,42	0,52	0,18	1

Figure 3.4: Correlation matrix

It can be observed that some of the considered explanatory variables are in deed highly correlated to one another. Specifically, this problem exists for the degree centrality, eigen vector centrality, random walk centrality and information centrality and to eliminate the multicollinearity, only one of them should be incorporated at the same time. For the same reason, one should choose between harmonic centrality and closeness centrality and between harmonic centrality and number of contacts after 3, 6 and 9 days when including the explanatory variables. This conclusion will be taken into account in the following sections in order to create models without the multicollinearity.

3.3.3   Models with one explanatory variable

The next attempt to create a model for Y was based on Linear regression model with only one explanatory variable (Table 3.5). The reasons to investigate such models were that:

- without no doubt, the problem of the multicollinearity does not exist for them,

- it makes possible the identification of the influence of each variable on Y without the analysis of the possible interactions between the explanatory variables, as happened in the previous model.

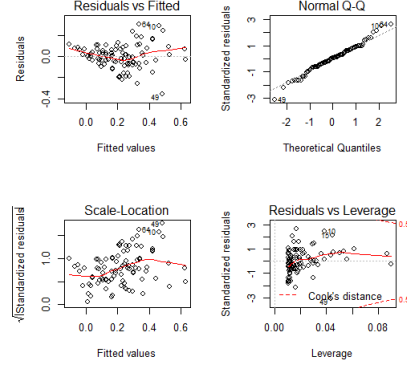Table 3.3: Linear regression models with one explanatory variable

Variable	Model Formula $Y = \beta_0 + \beta_1 X$	Standard Error β_0	Standard Error β_1	Adjusted R^2	F-test p-value
Betweeness centrality	$0.18 + 0.0003X$	0.02	0.00007	0.1588	0.00005
Clustering coefficient	$0.09 + 25X$	0.03	4.95	0.22	$1.281 \cdot 10^{-6}$
Degree Centrality	$0.01 + 0.014X$	0.04	0.005	0.09	0.002
Closeness centrality	$Y = -0.22 + 0.06X$	0.07	0.009	0.35	$3.08 \cdot 10^{-10}$
Eigen vector centrality	$Y = 0.16 + 0.95X$	0.04	0.4	0.05	0.014
Random Walk centrality	$Y = 0.1 + 3.5X$	0.04	1.05	0.1	0.001
Information centrality	$Y = -0.03 + 8X$	0.08	2.45	0.1	0.001
Harmonic centrality	$Y = -0.16 + 0.0004X$	0.03	0.00003	0.63	$2.2 \cdot 10^{-16}$
k3	$Y = 0.12 + 0.002X$	0.02	0.0003	0.4	$3.7 \cdot 10^{-12}$
k6	$0.05 + 0.003X$	0.01	0.0002	0.75	$2.2 \cdot 10^{-16}$
k9	$0.19 + 0.001X$	0.02	0.0006	0.04	0.02
kend	$Y = 0.14 + 0.0009X$	0.02	0.0001	0.32	$1.882 \cdot 10^{-9}$

Taking a look at the value of the adjusted R^2 (one could also observe value of the *AIC* and the conclusions would be the same in this case) of the obtained models, it can be noticed that the best results were achieved when the chance attribute was described in terms of the number of contacts between 3rd and 6th day- $k6$ (although the variance of b_0 is quite big) and when the harmonic centrality was taken as the explanatory variable (estimation of the coefficients has lower variance but also the calculated adjusted R^2 is lower than for $k6$). In the both cases the p-value of the F-test indicates that the models
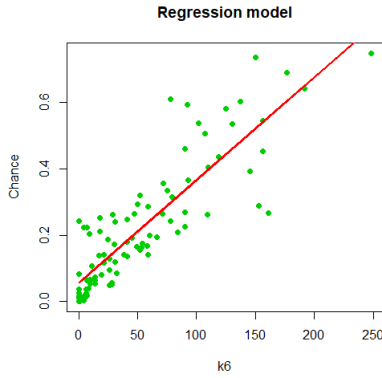
transmit more information than the simple mean model. To check whether the model is
valid the diagnostic graph of the linear regression assumptions were created for them(3.5).
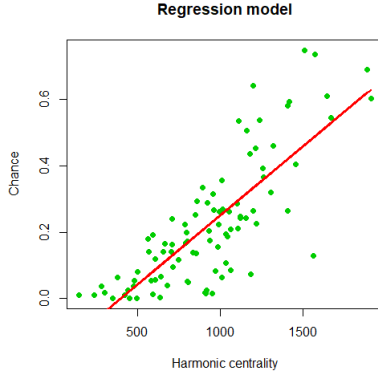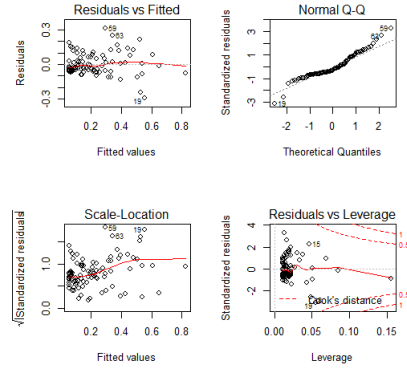

(a)   Model   for   the   harmonic
centrality- plot of the regression line

(b) Model diagnostics plots for the
harmonic centrality

(c) Model for the $k6$- plot of the
regression line

(d) Model diagnostics plots for the
$k6$

Figure 3.5: Selected models with one explanatory variable

It's noted that in the both cases the homcedasticity and normality assumptions are
violated. Therefore, the data transformations, that aim to eliminate this problem, were
done for the best model of one variable, that is with $k6$ as the explanatory variable. The
most accurate transformation turned out to be the root transformations of both X and
Y. The new model is of the form: $\sqrt{Y} = 0.119705 + 0.048279\sqrt{X}$. The adjusted R^2 value
and p-value of the F-test did not change and the $AIC = -429$. The diagnostic plots were
again created (Figure 3.6). One may observe that the variance of the residuals is now
rather stable and that the linearity is preserved. Regarding the normality, there are some
observations that do not match the normal distribution what is demonstrated on the qqplot.
However, the visual interpretation may not be sufficient and due to that the Shapiro-Wilk
normality test with the null hypothesis saying that data have the normal distribution, was
performed on the significance level 0.05. The obtained p-value equals 0.034 and therefore
the normality assumption is not fulfilled and the model is not considered valid.
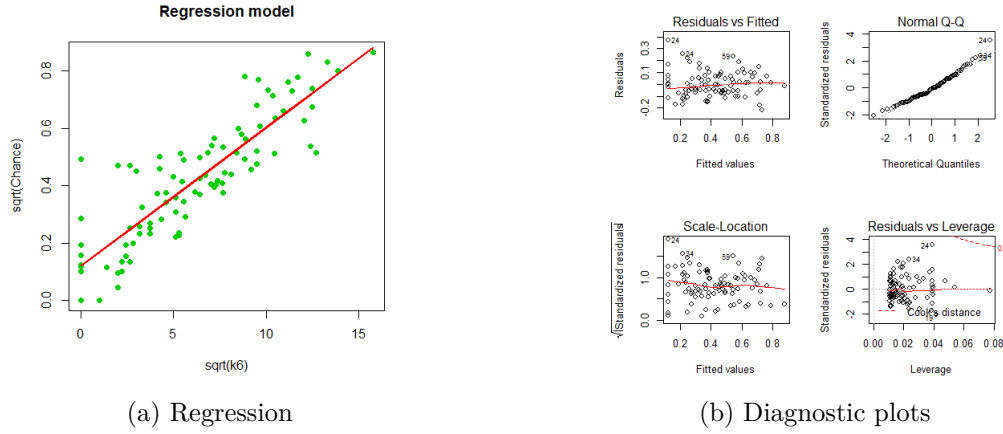
(a) Regression

(b) Diagnostic plots

Figure 3.6: Model for the k6 with the data transformations

3.3.4 More than one explanatory variable

Since the previous approach did not provide the desirable valid model, it's worth to try models with more than one explanatory variable since it's possible that the incorporation of more variables gives more accurate results. The taken strategy aims to avoid the inclusion of the correlated variables and as a result the variable selection is restricted. Created models are divided into two groups: those that incorporate the number of contacts and those that include the harmonic centrality. Then 4 different models for each group were constructed with only one of the variables: degree centrality, eigen vector centrality, random walk centrality and information centrality (Figure 3.7).
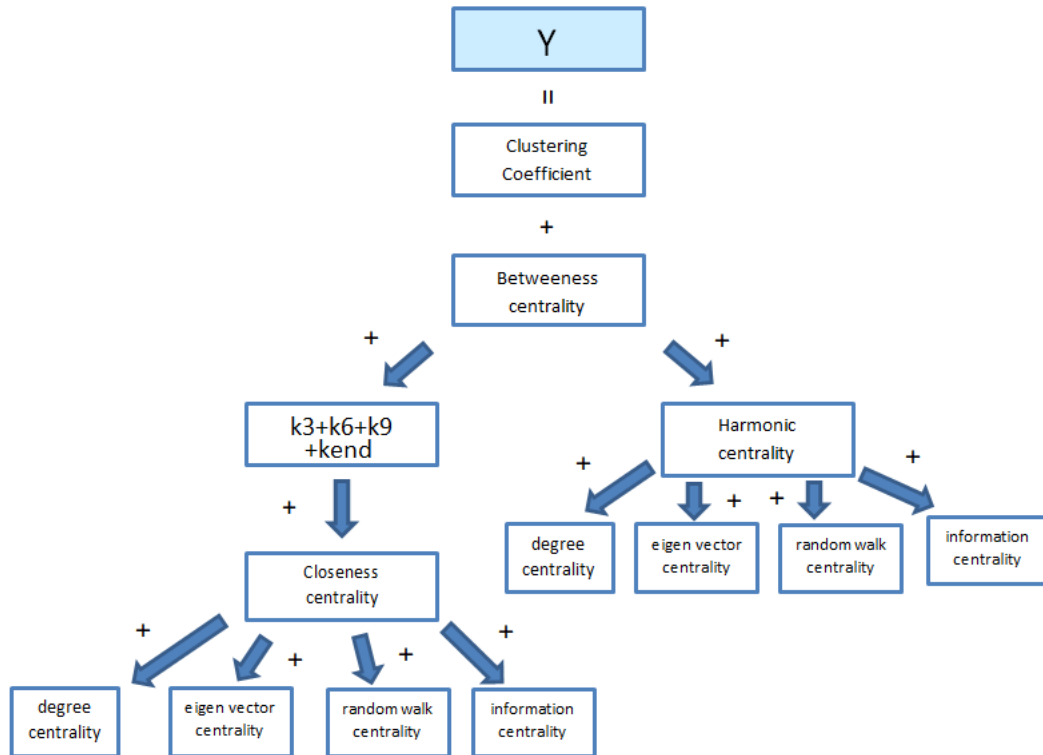

Figure 3.7: Variable selection schema

Each model was then optimized by the backward procedure and its AIC and R^2 were calculated (Table 3.4). The obtained results in the group are very similar to one another. Furthermore, models that include the number of contacts are characterized by the larger accuracy than those with the harmonic centrality. The best obtained model, that is the one with the random walk centrality and the number of contacts, will be analyzed.

Table 3.4: Comparison of the models with more than one explanatory variable

	degree centrality	eigen vector centrality	random walk centrality	information centrality
Models that include the harmonic centrality				
AIC	-402.83	-399.32	-402.84	-401.54
R^2	0.67	0.66	0.67	0.67
Models that include the number of contacts				
AIC	-499.34	-495.56	-500.75	-497.75
R^2	0.89	0.88	0.89	0.886

The best obtained model

The best obtained model that incorporates the number of contacts and the random walk centrality is shown on the Table 3.5. It can be noticed that all coefficients besides one are significant due to the low p-value of the F-test for them. Especially, the $k6$, $k3$ and the clustering coefficient seem to be very important. Similarly, the error of the estimation is rather small for each variable and the overall p-value of the F-test is smaller than $2.2 \cdot 10^{-16}$. Taking it all into account as well as the high value of the R^2 and low AIC, model can give the valuable description of the chance attribute.

Table 3.5: The best obtained model

	Estimate	Std.Error	p-value	$Pr(> |p|)$
Intercept	0.065	0.002	-3	0.003
k3	0.009	0.0001	6.7	$1.85 \cdot 10^{-9}$
k6	0.002	0.0002	12	$< 2 \cdot 10^{-16}$
k9	0.0008	0.0002	4	0.0001
random walk centrality	1.4	0.5	2.87	0.005
clustering coefficient	13	2.5	5.2	$1.5 \cdot 10^{-6}$
betweeness centrality	0.00004	0.00002	1.6	0.1

However, before considering a model useful, one needs to check whereas the models is valid. This is done via diagnostic plots (Figure 3.8).
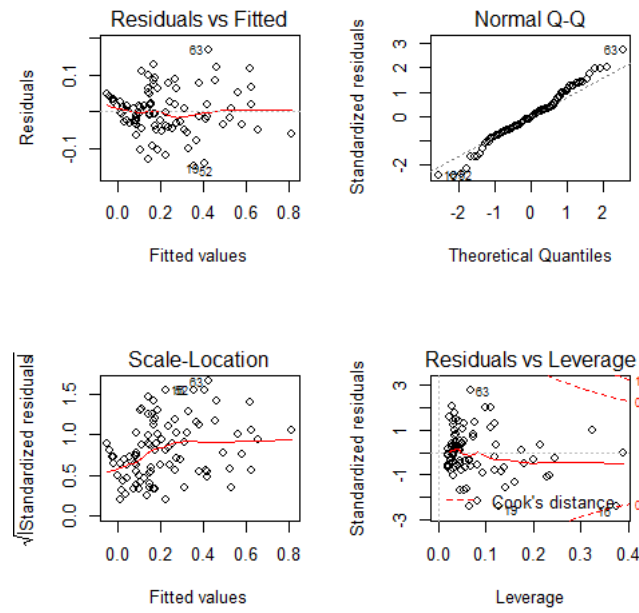
Figure 3.8: The best obtained model with the data transformations

It's observed that both the normality assumption as well as the homcedasticity is violated. In order to eliminate this issue, again the data root transformation is done on Y and on all explanatory variables. Nevertheless, the new model posses insignificant coefficient for the betweeness centrality (p-value 0.76) and when it's eliminated the AIC increases and takes the value -486.7 (smaller than before transformations). The coefficients of the new model without betweeness centrality and with data transformation are presented (Table 3.6). Obviously, their values differ from the model without transformations, however the significance of their values as well as the estimation variance is rather similar. Furthermore, the model's R^2 equals 0.897 (slightly bigger than before transformations) and the p-value of the F-test is smaller than $2.2 \cdot 10^{-6}$ (the same as before transformations).

Table 3.6: The best obtained model with the data transformations and without betweeness centrality

	Estimate	Std.Error	p-value	$Pr(> |p|)$
Intercept	-0.14	0.04	-3	0.003
k3	0.01	0.002	4.8	$5.63 \cdot 10^{-6}$
k6	0.032	0.002	12.6	$< 2 \cdot 10^{-16}$
k9	0.02	0.0026	6.7	$1.7 \cdot 10^{-9}$
random walk centrality	0.55	0.2	2.7	0.009
clustering coefficient	2.2	0.46	4.8	$6 \cdot 10^{-6}$

Now, the effect of the data transformation on the fulfillment of the regression assumptions can be investigated (Figure 3.9). It seems that the homocedasticity as well as the linearity is fulfilled since the residuals of the model are rather steadily scattered. The qqplot suggests that the residuals do not match exactly the normal distribution but at the same time that the difference is not very significant (only few observations stand out from the line). To confirm this observation, the Shapiro-Wilk normality test was performed.

The obtained p-value equals 0.72, what demonstrates, as suspected, that the normality assumption can't be rejected. The mean of the residuals takes the value $3.7 \cdot 10^{-18}$, so is very close to 0. The last to check is the independence of the residuals, the assumption that can be violated since the location of the node can influence on its chance attribute value (the node that has neighbours with the high infection potential can also have it). For that reason the autocorrelation graph for the residuals is also plotted and it's noted that residuals don't depend on one another. Having all that in mind, the model is considered valid.
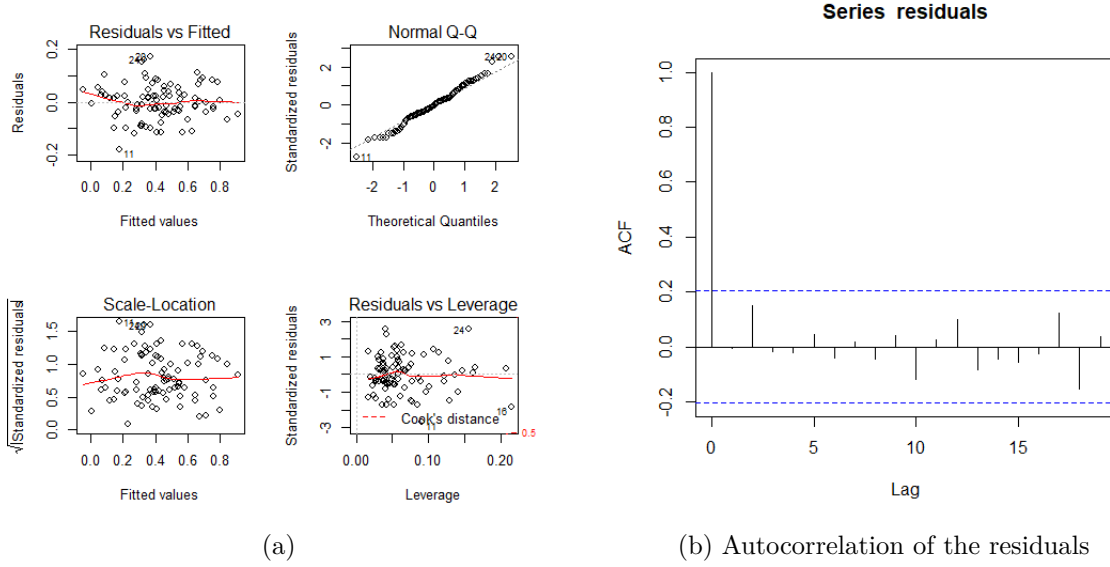

(a)   (b) Autocorrelation of the residuals

Figure 3.9: Diagnostic plots for the model with the data transformations and without betweeness centrality

3.3.5   Conclusion

It seems that the discovery of the good model for the infection potential of the person is a challenging task. It frequently happens that the global criterions have a promising values but the basic regression assumptions are not fulfilled and therefore models are not useful. The detection of the proper transformations may eliminate this problem but sometimes it's hard to encounter them. The obtained model seems to be good due to its low AIC and high R^2. Moreover, it's observed that the numbers of contacts in different time interval are important factors that influence on the infection potential of the individual, although the number of contacts on the last stage of the simulation is not necessary for the calculation. Probably, it is due to the fact the the first exposed person is infectious until the 9th day and the number of contacts until that day gives the overall estimation if the epidemic will spread or not. Similarly, as the visual analysis indicated the neighborhood of the node that is the belonging to the cluster or not impacts on its infection potential. In terms of the random walk centrality, even though it is incorporated in the model and is significant, the global criterions for it differed only slightly from the ones for the degree centrality, eigen vector centrality and information centrality. In summary, it was shown that the contact structure has an impact on local dynamics of the disease spread. However, one should remember that the network size is rather small and for that reason it could happen that the chosen model would not be the best one if the network was larger and therefore the used variables could be different.

Chapter 4

Discussion

In this paper it was shown that the disease characteristics such as its R_0, the incubation period or the time of the infection plays an important role for the diffusion process. The next taken observation is that the individual behaviour, that is the social relations that the person has, impacts on the disease transmission process, but the overall probability of the disease spread can be modelled via averaging and generalization of the contact structure.

However, the taken analysis has some disadvantages due to the fairly short time of the gathering data. The obtained results could be slightly different if it was longer, especially for the individuals capable to spread the disease, since the range of the epidemics potential could not be completely investigated. Similarly, the diffusion process was checked only for the influenza virus and on the next stages of this kind of work it would be worth to check the spread process of the other diseases that would have a different probability of transmission.

Bibliography

[1] Centers for disease control and prevention. `https://www.cdc.gov/flu/about/keyfacts.htm`.

[2] Python networkx package documentation. `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.harmonic_centrality.html#networkx.algorithms.centrality.harmonic_centrality`.

[3] Python networkx package documentation. `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html#networkx.algorithms.centrality.eigenvector_centrality`.

[4] R documentation. `https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/step`.

[5] Sociopatterns. `http://www.sociopatterns.org/datasets/contacts-in-a-workplace/`.

[6] BARABÁSI, A.-L. *Network Science*, 1 ed. Cambridge University Press, 2016.

[7] COPE1, R. C., ROSS1, J. V., CHILVER, M., STOCKS, N. P., MITCHELL, L. Characterising seasonal influenza epidemiology using primary care surveillance data. *Plos Computational Biology*, 1006377 (2018).

[8] GOLBECK, J. *Introduction to social media investigation.* 2015, ch. 21.

[9] GROSS, J. *Lecture notes in statistics.* Springer-Verlag Heidelberg, 2003, ch. Linear regression.

[10] ILKKA KIVIMÄKI, BERTRAND LEBICHOT, J. S. M. S. Two betweenness centrality measures based on randomized shortest paths. *Nature*, 19668 (2016).

[11] PEREZ, C., GERMON, R. *Automating Open Source Intelligence, Algorithms for OSINT.* 2016, ch. 7.

[12] RENCHER, A. C., SCHAALJE, G. B. *Linear models in statistics*, 2 ed. Wiley-Interscience a John Walley and Sons, Inc Publication, 2007.

[13] SALATHÉ, M., KAZANDJIEVA, M., LEE, J. W., LEVIS, P., FELDMAN, M. W., , JONES, J. H. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020-5 (2015).

[14] SMILKOV, D., HIDALGO, C. A., KOCAREV, L. Beyond network structure: How heterogeneous susceptibility modulates the spread of epidemics. *Nature*, 4795 (2014).