Erasmus School of Economics

Master Thesis in Econometrics and Management Science

Specialisation: Econometrics

Diffusion over the network using latent variable model

Author:

Mateusz Kaźmierczak

Student number: 474508

Academic supervisor:

dr. Yutao Sun

Second assessor:

dr. Wendun Wang

Rotterdam, December 18, 2019

Abstract

This thesis investigates how the information spreads over the networks with the known structure dependent on the character of the connections. First, the dissimilarity scores between individuals within the networks are calculated. Then the probabilities of a strong contact are modelled using the EM algorithm. Finally, the obtained probabilities are applied as the transmission rates to the SI contagion model. Additionally, there are three artificial Erdős–Rényi networks with different sparsity created to show whether the sparsity of links plays an important role in the diffusion process. The approach proposed in the paper does not rely only on the fixed transmission rate independent from the nodes and vertices in the networks. It also tries to incorporate the individual-specific transmission rates, which vary per each connection. The thesis finds that the information speed and range may depend on where it starts. Firstly, the propagation of the information in the artificial networks shows that the sparsity can be an important factor in the range of the diffusion. Moreover, remote individuals can be not good starting points of the diffusion, whereas the initialization in the clique offers higher diffusion range on average. Individual-specific transmission rates may also ensure that the diffusion does not arrive to all of the individuals. Additionally, the individual-specific probability offers on average a smaller range of the diffusion.

Keywords: Network analysis, Diffusion process, Finite-mixture model, EM-algorithm, SI model

Contents

1 Introduction

A network is generally defined as a set of points joined together in groups by links. In a network, we may distinguish two types of components: points referred to as nodes or vertices and connections referred to as edges or links. Networks are an object of interest in many different sciences: physics, biology, social sciences or IT. In the real-life applications, one may find many networks composed of individuals linked to each other. The internet, social networks, railway or telecommunication systems are examples of a network. Additionally, in graph theory, the research on networks has also received considerable attention. For instance, the analysis of human network – a set of connected individuals – may offer many useful insights, which may enlarge the current knowledge in a specific domain. One may study the features of vertices, while another may focus on the nature of the connections between them. Moreover, the nature of networks allows for broader analysis, and one may research such issues as the patterns of connections within a system; the driving force of creating ties between points; the networks evolution over time or the effect of exogenous variables on the network.

Newman (2018) argues that we cannot understand the impact of the links in social networks on how people perceive the world, how they express their opinions or how the economy functions until the structure of the network is known. In general, networks do not fully represent the real structures, hence they should be treated merely as a simplification of real-life patterns.

Once the network structure is defined, we can study how an exogenous shock/event spreads over the network. Based on the nodes and edges, the speed of the diffusion may differ. The definition of the shock itself does not play an important role. It may be a political event, an economic shock, a treatment, a rumour or a disease. One can consider an individual who is willing to pass the shock on his neighbours/neighbouring nodes. We can distinguish two types of the individuals. It is either diffuser ,i.e. the one willing to spread the shock further, or uninfected individual who can receive/adopt/accept the shock or reject it. The diffuser may pass the shock only to the neighbours, i.e. neighbouring individuals. The uninfected neighbour may accept the information with a specific probability (also later referred to as the transmission rate). The probability of acceptance can be either

constant over the network or individual-specific, thus varying per each link. Once the neighbour accepts the shock, he or she decides whether to spread it further or not. As the recipient may also refuse to accept the information, the process of percolation continues until there is no neighbouring individual willing to accept the shock. In fact, the velocity and the adoption of an event depend on a given individual. The acceptance of a treatment may depend on the similarity between the neighbours and the connection they share between each other. Some people have influence on others' decision, while some of them are not able to spread the shock.

The paper sets multiple goals. The first is to construct a diffusion model which depends on the characteristics of individuals and the strengths of the connections. The second goal is to analyse how an exogenous shock spreads over a network with a known structure. Finally, we will consider how far the shock may go and which initializing agents offer the fastest spread of the information. The point of departure is to construct a probabilistic model which would deliver the acceptance probabilities based on the network structure and the characteristics of individuals. When the probability is known, we will simulate the diffusion using one of the contagion models.

To model the probabilities, one may consider the latent variable model where the unobserved variables are modelled by the observed coefficients. As the decision on acceptance is binary (either accept or reject), the finite-mixture model by Everitt and Hand (1981) is the most suitable choice. The mixture model assigns the value of the explanatory variable to one of two distributions/states. Once we have the complete data likelihood function, we are able to optimize it and obtain the probability of assignment to one of the states. Estimation of the optimal parameters can be done by the iterative maximization method, Expectation-Maximization (EM), first described by Neal and Hinton (1998). The EM algorithm allows to optimize the likelihood function using a two-stage procedure: Expectation and Maximization. The use of the standard Maximum Likelihood (ML) estimation tends to be biased in small samples, which the EM algorithm might tackle. The probabilities of state assignment are assumed to be equal to the rumour acceptance probability. Once we obtain the acceptance probabilities, we can simulate

the propagation of the rumour. As we have only two states of rumour acceptance, therefore one may consider the Susceptible-Infected (SI) model, where there are two states, susceptible and infected. Intuitively, the individual in the susceptible state has not received the information, whereas the infected individual is the one who accepted the rumour and is willing to spread it further. After the simulation we are able to perform the analysis and draw conclusions on that basis.

The contribution of this paper to the existing literature relies on the combination of two widely known aspects: the contagion model with the finite-mixture model. In academic literature, most of the compartmental models have the transmission rate set manually by the researcher. For instance, Watts (2002) conditions the transmission rate/acceptance probability on the number of infected neighbours. Therefore, estimating the diffusion rate conditional on the individuals' characteristics can be an addition to network theory. Moreover, from the empirical perspective, such a model may be successfully implemented in the network if the interactions between individuals are known. Once we quantify the interactions we are able to conduct the diffusion conditional on the interactions.

The research conducted in the paper can be relevant from both scientific and practical perspectives. Scientists might be interested in the performance of the diffusion process in various implementations of the model, while the business sector might be interested in properly positioning their product to the target audience, so that the information about the product could reach a larger number of customers.

The paper is constructed as follows. The chapter following the introduction provides an overview of the existing literature on the diffusion over networks. The next chapter explains the methodology behind the model. The sub-chapters describe the finite-mixture model and the contagion model separately. Chapter 4 describes the applied data sets, then Chapter 5 discusses the estimation of the parameters by the EM algorithm. Next section presents the application to empirical data. Chapter 7 concludes.

2 Literature review

The section provides a short review of the existing literature. The two first subsections in this chapter cover different approaches to network formation and diffusion process, while the last presents the existing econometric models with the latent variable. As these two fields are rarely treated together, they receive separate attention.

2.1 Network formation

For decades, network theory and contagion studies have mostly concentrated on the creation of networks. Many researchers propose various techniques of graph generation, albeit mostly on small-scale networks. A valid point is made by Erdős and Rényi (1959); they construct a mechanism of creating a random graph on a larger scale. Their way of creating the links is based on the complete randomness. Albert and Barabási (2002) question it by arguing that most of real networks are much more complex and not truly random. One of the studies on network configuration is done by Watts and Strogatz (1998). They observe that most individuals only make links within short distances. Moreover, people tend to form groups, which make networks more clustered than random graphs. For a fixed average degree k and a number of vertices N, they define a network which depends only on one parameter p called rewiring probability. They define network as a set of points arranged in a circle with N nodes connected to k nearest neighbours. If $p > 0$, we reconnect random node to another randomly chosen vertex with probability p. The higher p, the lower the clustering level and the more randomized the network. Watts and Strogatz (1998) apply the diffusion model to examine how contagion spreads in a such defined network. They show that the more random the network is i.e., the higher the rewiring probability is, the faster the infection spreads.

Barabási et al. (2016) point out that the assumption of a constant degree over all vertices might not be realistic. In fact, most of observed networks consist of hubs (nodes with a high degree) and of vertices scarcely connected to each other. Barabási proposes his own scale-free network model with a power law

degree distribution, which limits the number of possible edges (Albert and Barabási (2002)). In Figure 1, examples of Erdős–Rényi, small-world and scale-free graphs are presented.

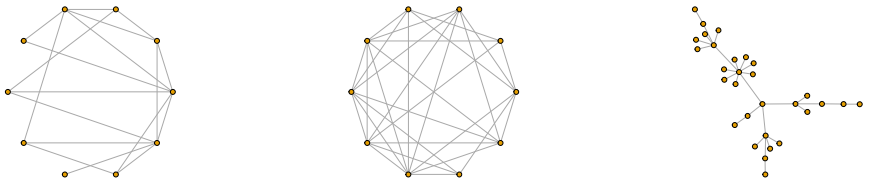


Figure 1: Examples of random graphs

2.2 Diffusion in network

The diffusion process in a social network receives a considerable attention in the literature. Early literature is mostly based on the epidemic model. The social sciences and epidemiology exhibit many similarities in terms of the information spread and therefore the study on the spread of a disease widely contributes to the study of social networks. In general, a disease spreads over the network once individuals make contacts. For instance, a contagious disease spreads through touch, while HIV is transmitted though sexual intercourse. An analogy occurs in the IT sector: a computer virus - a malicious software spreads between computers similarly to real infections. This is similar to social networks. People spread the rumour once they speak to each other. The transmission rate is reciprocal to human defence system. Hence, one may say that the acceptance of rumour resembles the immune system.

The modelling of the information spread starts with simulating real epidemic behaviour. For instance, McKendrick (1925) applies the logistic growth equation to the bacterial growth data. His approach (Kermack and McKendrick (1927)) laid foundation for the so-called Kermack-McKendrick theory, which predicts the distribution of infected individuals in the population over time. Their model assumes that each individual has an equal chance per time unit to meet each other completely at random. It lacks realistic representation as people do not make

contact at random and they meet only a small fraction of the total population. Contrary to the Kermack-McKendrick theory, in real networks, a potential disease does not spread at random but only via edges. Bailey et al. (1975) extends the field by introducing new classes of contagion models. Apart from the simple two-stage Susceptible-Infected (SI) model, he proposes three- and four- states models, such as Susceptible-Infected-Susceptible (SIS), Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Recovered-Susceptible (SIRS). Accordingly, the individual may either recover from a disease (SIR), go back to initial susceptible stage (SIS) or recover and be exposed to the disease again (SIRS). The family of compartmental models offers a straightforward solution to model the diffusion over the network. For instance, Lamberson (2016) shows that epidemic models offer easily interpretative parameters and, from the mathematical perspective, they are applicable to many different configurations. To establish a broader framework of the contagion models, infection denotes the acceptance of the event which will be later called an adoption. Similarly the transmission rate simply means the probability of accepting the event.

As an alternative to the disease models, Watts (2002) proposes threshold models. The acceptance or rejection of the event depends on the threshold level defined by the researcher. For instance, an individual accepts the rumour/shock if the fraction of the affected neighbours over all neighbours has exceeded the threshold level set by the researcher. Otherwise, the individual will deny spreading it further. Additionally, Campbell (2013) proposes an extension of the threshold models, where the event is defined as a product with a fixed price. Its adoption depends on how much the individual values the product. If we define the utility of adoption, then the individual adopts the shock only if his utility of the adoption is higher than the fixed price of the event. However, the diffusion process is split into two ways. Firstly, the agents are informed about the event and if their utility exceeds the price, then they adopt it and also spread it further. On the other hand, if the price is higher than the utility, then the agent neither accepts nor spreads.

Moreover, Campbell (2013) shows that cliques (densely connected group of individuals) slow down the spread of information. Intuitively, when two vertices are part of a triangle, not only are they connected directly, but they also have a

mutual neighbour through which the event could be spread. Thus, it enables the event to be delivered twice to the same recipient, what makes one connection in the clique redundant. Therefore, in the cliqueless network, the information spreads faster and reaches a larger number of vertices (only if the network is not sparse).

Another contribution to the field of percolation in networks was done by Lamberson (2010). His model assumes that social learning depends on prior beliefs that an individual has. The prior belief is calculated based on Bayes' rule. The initial belief represented as a payoff is compared to the final posterior payoff the individual has once the information is delivered to him. The belief of the individual is drawn from some constant distribution and then the individual compares his belief to the payoffs of the neighbours, who spread the shock. However, the current prior belief is not compared to the current payoff of the shock transmitted from the neighbours, but to the payoff from adopting the neighbours from the previous time point. The following process can be explained as social learning. The individual updates his beliefs if his neighbours react positively enough (measured by payoff) to the upcoming shock/event.

2.3 Latent variable models

The latent variable model corresponds to the statistical model in which the observable variables are interrelated with an unobserved (latent) variable. This class of models assumes that the observable variables are conditionally independent of each other. However, the latent variable may explain the relationship of the observable variables. Such an assumption is called local independence. A significant contribution to the latent variable statistics is made by Everett (1984). He categorizes the models based on the types of variables: either categorical or continuous for both observable and latent variable. In total, there are four categories of statistical models with the latent variables. In the literature one may find several different methods under each category. Firstly, the item response theory has developed many techniques of relating categorical observable variables to the categorical latent variable. The item response models are used to measure unobserved traits like attitude or behaviour. They receive considerable attention in sociology and psychology, e.g. in psychological or aptitude tests. Rasch (1960) contributes by

proposing a model of continuous latent and observable variables, named after him, where he assumes that the probabilities of respondents answering correctly the question asked in a survey can be approximated by the logistic function of distance (on the linear scale) between the respondent's ability and the difficulty of a given question.

Contrarily, when dealing with continuous observable and continuous latent variables, one may use factor analysis. Mardia et al. (1979) propose factor analysis as a mathematical model which tries to explain the correlation between the observables and the underlying factors. In the model, the observable variables depend on the latent variables (factors) and a random error. Such specification is particularly applicable to areas where the expected variable is not precisely measurable. Examples here include IQ in psychology ,which cannot be directly quantified, financial or economic factors, which do not influence the market directly. The estimation of the factor model can be done by principal components or maximum likelihood. Furthermore, based on the scree plot or likelihood ratio test, we are able to decide what number of factors offers the largest information gain.

Accordingly, when one has continuous observable variables and is interested in estimating the categorical latent variables, the literature offers many different configurations. Everitt and Hand (1981) widely contribute to the latent class models. They create the class of mixture models in which the continuous observable variable was assigned to one of the latent sub-populations. The most well-known model is the finite-mixture model where, based on the outcome of the variable, the observation is categorized to one of the latent states. To obtain the parameters of the model one may use two different approaches, either the frequentist with EM algorithm or the Bayesian approach (Frühwirth-Schnatter (2006)). As rumour propagation has two states (accept rumour or reject), the finite-mixture model is a good example of the diffusion process in latent class modelling.

The finite-mixture model has also been widely implemented in time series (Hamilton (1990)). One may condition the current state on the previous realization of the observable variables. Hence, one may use the Markov switching model described by Kim (1994), where additionally he implements the first-order Markov property to the existing two-states finite-mixture model. The imposed

Markov property implies that the current state of the variable (at time t) depends on one period ago. Therefore, one needs to estimate transition probabilities between the states, thus improving the fit of the model.

The models with the continuous observable and categorical latent variables have been implemented in the panel data. For instance, Bonhomme and Manresa (2015) describe grouped fixed effect (GFE) model which allows the data to be arbitrarily correlated with latent group-specific variables. Unlike the finite-mixture model, the state (categorical latent variable) can be time-varying. Moreover, the number of states is unrestricted and can be estimated from the data. The class of GFE models offers many applications in micro- and macroeconomics. For example, Acemoglu et al. (2008) implements the GFE in political sciences, where he finds no positive association between democracy and income when modelling country's fixed effects with the GFE model.

The models with latent categories receive considerable attention in text mining. Blei et al. (2003) describe the Latent Dirichlet Allocation (LDA) model, which can extract several themes that are unobserved in the text. The main role of the LDA is to decompose the text into latent topic probabilities in the document and word probabilities in each topic.

3 Methodology

This section discusses the methods for obtaining probabilities based on the similarities between individuals. Subsection 3.1 defines the dissimilarity score. Next, the finite-mixture model is presented and several characteristics are reviewed. Subsection 3.3 discusses the EM-algorithm used for obtaining the model coefficients and the transition probabilities. The last subsection describes the application of the obtained parameters to the diffusion model.

3.1 Network structure and dissimilarity score

One may define a network as $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}, \mathcal{E}$ are sets of vertices and edges respectively, i.e.

$$\mathcal{V} = \big\{D_i : i = 1 \dots N\big\} \tag{1}$$

$$\mathcal{E} = \big\{L_{i,j} : L_{i,j} \in \mathcal{I}\big\}, \tag{2}$$

where D_i denotes vertex, $L_{i,j}$ denotes the link between nodes D_i,D_j and $\mathcal{I}$ is a subset of $\big\{1 \dots N\big\}$, where N is the number of individuals. If there is no link between i and j, then $L_{i,j} \notin \mathcal{E}$. Accordingly, we can represent $\mathcal{E}$ by a symmetrical adjacency matrix denoted by $\boldsymbol{A}$ of the size $N \times N$. The symmetry of $\boldsymbol{A}$ means that, in this paper, we consider only an undirected network. $\boldsymbol{A}$ consists of entries i, j defined as follows:

$$A_{i,j} = \begin{cases} 1 & L_{i,j} \in \mathcal{E} \text{ (nodes are connected)} \\ 0 & L_{i,j} \notin \mathcal{E} \text{ (nodes are not connected).} \end{cases} \tag{3}$$

Let K denote the total number of non-zero elements above the main diagonal (or below, due to symmetry) in $\mathbf{A}$, which equals the total number of edges.

After observing the network, the next step is to properly design the percolation process of the treatment in such a created network. The starting point is to define the variable that might explain the velocity and the probability of the rumour spread. To simplify the exposition, we reindex $L_{i,j}$ by k for $k = 1, \dots, k$. Note that $L_{i,j}$ is independent of time since a temporal network is not considered in this

paper.

Let us define $\boldsymbol{X}$ as the matrix of characteristics of a size $N \times P$, where P is the number of characteristics. Each individual has his own attributes. Some of the individuals are more similar to each other in terms of features, some of them lie far from each other on the *similarity space*. Let us also define a variable u_k which is a dissimilarity score of the connection k between two corresponding individuals. One may define it as a homophily score: the more similar the agents, the lower the value of dissimilarity score. Hence, the dissimilarity may be interpreted as a distance in characteristics between the individuals. One may use the Euclidean distance for calculating remoteness of the agents. Then, the dissimilarity score of the link is defined as follows:

$$\begin{aligned} U &= (u_{ij}) = (u_k) \\ u_k &= \|X_i - X_j\|_2^2. \end{aligned} \tag{4}$$

As a result, the dissimilarity score U has the form of vector of size $K \times 1$ and its entries are defined as u_k.

3.2 The finite-mixture model

When the dependent variable is obtained, we need to use a suitable probabilistic model to represent whether the specific connection transmits treatment or not. One possibility is the regime-switching model (Everitt and Hand (1981)). They show that the model defines the probability distribution of the observations in the population. It is also called the finite-mixture model, as it is a mixture of different regimes. The decision rule to which subset the observation belongs is based on unobservable/latent variable. Regime-switching models allow the dynamic behaviour of the dissimilarity score to depend on the regime which occurs at any given point in the network. In our model, there will be a latent variable S_k, which is the current state of the edge k. It can either take the value 0 or 1. They are defined as the states of the contact intensity. State 1 ($S_k = 1$) denotes strong contact and State 0 is the state of weak contact. The concept of contact can be understood as friendships, acquaintances or business interactions or any type of relationships between two people. One can relate the similarity between individuals to the

status of the contact they share. McPherson et al. (2001) demonstrate that the similarity can lead to strong friendship. He argues that sociologists demonstrate that students form friendship mostly with peers, who have similar characteristics. Inversely, Jussim and Osgood (1989) claim that interpersonal relationships can influence political preferences, school attainment or individuals' behaviours, thus making two individual more similar to each other. Moreover, Davies and Kandel (1981), Heider (2013), Newcomb (1961) show that acquaintance, friendship or business contacts enhance the similarity between individuals over time. Hence, there is a positive correlation between similarity and contact intensity. The mutual interrelation effect does not violate the assumption of the regime-switching model, therefore we will focus only on the impact of the contact state on the dissimilarity score. The state of contact intensity is able to explain the variability of the dependent variable. Hence, we can assume that the dissimilarity score u_k follows a distribution that depends on the latent variable S_k. In other words, when the individuals exhibit an intense contact and their tie is assigned to the state of strong contact, then we know that the agents are similar to each other (measured by u_k). If their tie exhibits weak contact then their dissimilarity score is higher (the individuals are less similar to each other).

The finite-mixture model with the regimes of friendship states can be represented as follows:

$$u_k = \beta_{S_k} + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2_{S_k}). \tag{5}$$

There are two regime-specific parameters β_{S_k} and $\sigma^2_{S_k}$, which need to be estimated. The dissimilarity score is positive; therefore the use of a normal distribution could be questionable. However to avoid having negative values of the modelled variable we impose all negative values of the variable to be equal 0. In addition, one can also use the truncated normal distribution to deal with the negative values of the modelled dissimilarity score.

If one is interested in the information propagation, it is unintuitive to condition the dissimilarity score on the latent variable describing the rumour acceptance. To avoid ambiguity, the latent variable S_k denotes the contact intensity as mentioned previously. To link the rumour propagation with contact intensity, it is assumed

that the probability of accepting the rumour within the strong contact equals 1 (and 0 when the rumour goes through the edge with a weak contact). In other words, the more similar the agents, the higher the probability of accepting the rumour (as the probability of a stronger contact is higher). Therefore, the probability of a strong contact is imposed to be equal to the acceptance probability of the rumour. More details about the correlation between friendship and rumour spread can be found in the subsection 3.4.

Specification of equation (5) allows us to use two different types of regime-switching mechanisms, either the finite-mixture model or the Markov-switching model. The main drawback of the standard Markov Switching is the memory property - by model specification the probability of accepting the rumour (indirectly obtained by the probability of strong contact) depends on the past. This is not the case - if the individual does not accept the event, he will not spread it further. One may show it by the use of the transition matrix. For simplicity, let us define P_{mn} as the probability of spreading information between two neighbouring individuals m and n, who take value 1 (spread information further) or 0 (refuse to pass it on). Thus we are able to show that, with imposed the no-memory property, it converges to the standard finite-mixture model,

$$\begin{pmatrix} P_{00} & 1-P_{11} \\ 1-P_{00} & P_{11} \end{pmatrix} = \begin{pmatrix} 1 & 1-P_{11} \\ 0 & P_{11} \end{pmatrix}. \tag{6}$$

As shown above, when the individual is in state 0 (contact is weak), then he will not change the contact to strong. Hence the probability P_{01} equals 0, implying P_{00} to be always 1. Hence, the model can be represented as the finite-mixture model with two regressions dependent on latent variable S_k,

$$u_k = \beta_{S_k} + \varepsilon_k, \quad S_k = \{0, 1\}. \tag{7}$$

As mentioned in equation (5), the distribution of residuals follow the standard normal distribution, but with the regime-varying variance $\sigma^2_{S_k}$. Let us denote the fixed probability of strong contact of the edge k as follows:

$$P[S_k = 1] = p. \tag{8}$$

Accordingly the probability of having a weak contact is equal to $1 - p$. It is assumed that $0 \leq p \leq 1$. In the notation set, the small letter denotes the realized values (e.g. s_k), while the capital letter means the unknown true values (e.g. S_k). Accordingly, vectors are denoted without subscript, e.g. $u = \left[u_k\right]$ and $s = \left[s_k\right]$.

The density of u_k is the mixture of two normal distributions (whose probability density function is denoted as $\phi(.)$) and $\theta = \left[\beta_1 \quad \sigma_1^2 \quad \beta_0 \quad \sigma_0^2\right]^T$ is the vector of parameters. Vector θ_{S_k} corresponds respectively to the parameters from the specific regime. The probability density function of the dissimilarity score is

$$f(u_k;\theta) = \sum_{s_k} P[S_k = s_k] f(u_k|s_k;\theta_{S_k}) = p\phi(u_k;\theta_1) + (1-p)\phi(u_k;\theta_0). \tag{9}$$

The next step requires that the parameters of the model to be obtained. One can do it by maximizing likelihood. The likelihood function is given by

$$f(u;\theta) = \prod_{k=1}^{K} f(u_k;\theta) = \prod_{k=1}^{K} \big(p\phi(u_k;\beta_1) + (1-p)\phi(u_k;\beta_0)\big). \tag{10}$$

3.3 The EM algorithm

The likelihood function presented in the equation (10) can be optimized by the Maximum Likelihood estimator (ML). However the likelihood function may have several local maxima and can be ill-behaved. As a remedy, Neal and Hinton (1998) propose an estimation technique called expectation-maximization (EM) algorithm. The EM is an iterative parameter estimation routine which provides a maximum of the log likelihood function by the use of a two step procedure: expectation and maximization.

Firstly, let us define the complete data likelihood function $f(u, s;\theta)$ based on the joint density of u_k and s_k. Supposedly, we are able to observe the realizations of the latent process S_k. It takes the values either of 0 or 1, which corresponds with the draws from the Bernoulli distribution. Hence we may find the density of the combination (u_k, s_k) (which is independent from equation (10)) as

$$f(u_k, s_k;\theta) = f(u_k|s_k;\theta)P[S_k = s_k], \tag{11}$$

which is equal to

$$f(u_k, s_k; \theta) = (p\phi(u_k; \theta_{S_k}))^{\mathbb{I}[S_k=1]}((1-p)\phi(u_k; \theta_{S_k}))^{\mathbb{I}[S_k=0]}. \tag{12}$$

Hence, the complete data likelihood function is the product of joint densities,

$$f(u, s; \theta) = \prod_k (p\phi(u_k; \theta_{S_k}))^{\mathbb{I}[S_k=1]}((1-p)\phi(u_k; \theta_{S_k})^{\mathbb{I}[S_k=0]}). \tag{13}$$

Obtaining this function allows us to find the parameter estimates and the probability of the strong contact. Next, we perform an E-step: we take the expectation of the log complete data likelihood function with respect to $s|u$ and M-step: we maximize the expected value. The procedure can be written as

$$\max_{\theta} \ \mathbf{E}_{s|u}[\ln \ f(u, S, \theta)]. \tag{14}$$

To implement the E-step, Muthen and Shedden (1999) suggest firstly to calculate the conditional probability dependant on variable u_K denoted as π_k,

$$\hat{\pi_k} = P[s_k = 1|u_k; \hat{\theta}] = \frac{\hat{p}\phi(u_k; \hat{\theta}_1)}{\hat{p}\phi(u_k; , \hat{\theta}_1) + (1-\hat{p})\phi(u_k; \hat{\theta}_0)}. \tag{15}$$

The expected log-likelihood is therefore given as

$$\mathbf{E}_{s|u}[\ln f(u, s, \theta)] = \sum_k (1-\hat{\pi}_k)\ln(1-p) + \hat{\pi}_k \ln p + \\ + (1-\hat{\pi}_k)\phi(u_k; \theta_0) + \hat{\pi}_k \ln \phi(u_k; \theta_1). \tag{16}$$

The form of the expected log-likelihood allows us to split it into three parts separately. The probability parameter $\hat{p}$ is obtained by maximizing the first part and is calculated as follows

$$\hat{p} = \frac{\sum_k \hat{\pi}_k}{K}, \tag{17}$$

where K denotes the total number of edges. By maximizing the second and third part we may obtain parameters $\hat{\beta}_1$ and $\hat{\beta}_0$ for regime $S_k = 1$ and $S_k = 0$ respec-

tively,

$$\hat{\beta}_1 = \frac{\sum_k \hat{\pi}_k u_k}{\sum_k \hat{\pi}_k} \quad \text{and} \quad \hat{\beta}_0 = \frac{\sum_k (1-\hat{\pi}_k) u_k}{\sum_k (1-\hat{\pi}_k)}. \tag{18}$$

The variance of error term $\sigma^2_{S_k}$ can also be estimated and be updated consequently,

$$\hat{\sigma}_1^2 = \frac{\sum_k \hat{\pi}_k (u_k - \hat{\beta}_1)^2}{\sum_k \hat{\pi}_k} \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{\sum_k \hat{\pi}_k (u_k - \hat{\beta}_0)^2}{\sum_k \hat{\pi}_k}. \tag{19}$$

Given the updated formulas for the parameters, the EM converges to a local minimum in the log-likelihood function and the optimized parameters are obtained.

3.4 The SI model

Once we obtain all estimates in our models $\hat{p}$ and $\hat{\theta}$, we are able to apply them to the contagion model. $\hat{\beta}_1$ and $\hat{\beta}_0$ indicate the value of the intercept for the two regimes, respectively. The values of the homophily specify in which state the specific edge (and node) is. Then, when we obtain $\hat{p}$ we may subcategorize all links into specific regimes and perform a diffusion propagation. Moreover, observing individual-specific $\hat{\pi}_k$ allows us to decide which regime was relevant for each of the connections.

The obtained parameters do not correspond directly to the diffusion process. The aim of the EM algorithm is to achieve parameters for contacts' regimes rather than for rumour propagation. The only correspondence between the contact status and the propagation of the rumour occurs by imposing the transmission rates to be equal to the probability of the strong contact. One should note that the diffusion process does not distinguish between a strong and weak contact, but only uses the EM probabilities parameters to set the transmission rate of the rumour spread. To research on the diffusion over the network, let us assume that there are two stages: S (susceptible) and I (infected). All agents are at the beginning susceptible - they do not accept the event. Then we initialize a diffuser - an infected agent who has an information and spreads it further. As the propagation advances, it is assumed that, after some time T, all individuals will receive the rumour.

In the general framework of the SI model, there is given a transmission rate λ which denotes the probability of passing the disease further. As we have obtained

$\hat{p}$ and per-individual $\hat{\pi}_k$, we impose $\hat{\lambda} = \hat{p}$ and we create an individual-specific transmission rate $\hat{\lambda}_k = \hat{\pi}_k$. Newman (2018) defines $S(t)$ as the number of individuals who are susceptible at time t and $C(t)$ as the number of the infected. Let us also define s and c as the average fraction of susceptible and infected individuals, so $c + s = 1$. Thus, we may obtain the rate of the change of c and s for a fixed transmission rate by differentiating, which can be written down as follows:

$$\begin{aligned} \frac{dc}{dt} &= \hat{\lambda} sc \\ \frac{ds}{dt} &= -\hat{\lambda} sc \end{aligned} \tag{20}$$

The notation for individual-specific $\hat{\lambda}_k$ differs adequately only on the added subscript k. The equation for the change of c can be simplified as:

$$\frac{dc}{dt} = \hat{\lambda} c(1 - c) \tag{21}$$

This equation, which is widely used in population growth models, is called the logistic growth equation (Newman (2018)). The use of $\hat{\lambda}$ and $\hat{\lambda}_k$ allows us to compare the propagation over the network under the two specifications, either for a fixed probability in every edge or for an individual-specific probability where each link is treated separately. Hence, the methodology presented leaves room for further analysis.

The techniques presented in this chapter allow us to propagate diffusion and analyze the outcomes. The contagion model SI has been widely used in the modelling of the diffusion process. Its alternations have a larger number of states and offer broader application. The transmission rate $\hat{\lambda}$ is usually set by the researcher. However, the finite-mixture model enables modelling the transmission probabilities based on the individuals characteristics, thus being more applicable to real networks. Moreover, such an approach extends the current literature and can be a valid contribution to the existing methodology about diffusion propagation over the network.

4 Data

The starting point is to properly define network formation by the use of different techniques. The next step is to construct diffusion model, which is performed on previously created networks. There will be two data sets used in the research on diffusion process. One describes the school students' friendship network and the second reflects the corporate law partnership structure.

The first dataset used for an inference consists of 100 students from a single school in the AddHealth data set. The data set contains of characteristics of students and their friendship links. We use four students characteristics: sex (0 for male, 1 for female), race (white = 1, black = 2, hispanic = 3, asian = 4, mixed/other = 5), grade (ranging from 7 to 12) and age. There is also a match-specific characteristic in which for each friend named, the student was asked to check off whether he/she participated in any of five activities with the friend. The activities are as follows:

- you went to (his/her) house in the last seven days.
- you met (him/her) after school to hang out or go somewhere in the last seven days.
- you spent time with (him/her) last weekend.
- you talked with (him/her) about a problem in the last seven days.
- you talked with (him/her) on the telephone in the last seven days.

All activities have been summed to create a valued network. This variable ranges from 1 - reporting the student as friend but without any common activities to 6, which means to perform all 5 activities with the friend. Moreover, all friendship links are undirected and static over time. The summary statistics is given in Table 1:

The second data set comes from a research study of corporate law partnership network in the New England, US carried out by Lazega et al. (2001). It reflects the connections among 71 attorneys. There are 7 characteristics included: status (0 = partner; 1 = associate), sex (0= male, 1 female), office location (1 = Boston,

Table 1: Summary Statistics of Student Characteristics (N=100)

Characteristic	Mean	Standard deviation	Median	Min	Max
Sex	0.50	0.50	0.50	0	1
Race	1.6	1.41	1	1	5
Grade	9.56	1.52	10	7	12
Age	17.19	1.24	17.25	14.09	20.21
No. of activities in common	2.42	1.34	2	1	6

2 = Hartford, 3 = Providence), working years, age, work practice (0 = litigation, 1 = corporate) and finished law school (1 = Harvard, Yale, 2 = University of Connecticut School and 3 = others). Moreover, there is also variable describing the intensity of the connection between individuals. There are three types of connection : strong-coworker cooperation, advice partnership or friendship. The connections have been summed and the variable created denotes the number of connections, the two individuals have in common, varying from 1 to 3 if agents share a link. The links in network are undirected and static over time. Their summarized statistics are shown in Table 2:

Table 2: Summary Statistics of Attorneys (N=71)

Characteristic	Mean	Standard deviation	Median	Min	Max
Status	0.49	0.50	0	0	1
Sex	0.25	0.44	0	0	1
Office location	1.38	0.59	1	1	3
Working years	10.56	9.61	7	1	32
Age	41.85	10.25	39	26	67
Work practice	0.42	0.50	0	0	1
Law school	2.18	0.76	2	1	3
Intensity of connection	1.75	0.73	2	1	3

Thus we have two different networks. One describing the students' relationships and the other reflecting the law partnerships. The school network has an average degree node (the average number of adjacent edges to nodes) of 5.66, while the attorney network has an average degree node of 28.39. A transitivity (also called clustering coefficient) measures the probability that the adjacent vertices of a vertex are connected. The first data set has the clustering coefficient of 0.35 and the second of 0.55. There are no isolated nodes in the second data set

and only 4 nodes in the school network. For the purposes of the treatment propagation, these individuals will be omitted. The networks of students and attorneys have been visualised in the Appendix A in Figure 1 and Figure 2, respectively. The two obtained networks much differ in terms of their randomization. Students network is less randomized and is characterized by lower degree node than the attorney network. Moreover, there are few agents, who have connection only with one person. The law partnership network exhibits higher degree node and higher transitivity measure. The attorneys share much more links with each other, hence it is assumed that the diffusion will propagate faster under this structure.

5 Estimation of parameters

The estimated parameters of the finite-mixture model are presented in this section. In Section 5.1 we present the results of the EM-algorithm estimation for student network. Section 5.2 shows the obtained parameters for corporate law partnership network.

5.1 Student network

The finite-mixture model for the student network is exactly the copy of the equation (7). We regress the dissimilarity score on the regime-specific constant parameter and a noise. The dissimilarity score is defined as in equation (4) with the adjustment for a match-specific characteristic (denoted later as m_k), which has been explained in Data section. The increasing value of a connection characteristic suggests a stronger contact, while the growing value of dissimilarity score suggests a weak relationship. To compromise the direction of these variables, we will use the inverse of link-specific variable $\frac{1}{m_k}$. To include match-specific variable one may either make a sum or a product of the dissimilarity score and the variable $\frac{1}{m_k}$. The choice, which approach fits better to the given data is decided by comparing the modelled dissimilarity score to the real dissimilarity score. Arbitrarily, the sum of variables is finally chosen. The results of estimation for the product of variables is presented in the Appendix B. The final dissimilarity score (including m_k) is given as follows.

$$u_k = \sqrt{\frac{\sum_l \left(x_{il} - x_{jl}\right)^2}{L}} + \frac{1}{m_k} \tag{22}$$

In the case of sharing many activities, the dissimilarity score should be changed the fewest, while lack of common actions should boost the dissimilarity score. The second part of equation (22) works inversely to the match-specific characteristic. For instance, if the individuals perform no activities together then to the initial dissimilarity score we add $\frac{1}{1} = 1$. If they share all mentioned activities (e.g. 5 activities), then their dissimilarity score increases only by $\frac{1}{5}$. Hence, the dissimilarity score barely increases and the chance of having strong relationship is consequently

higher.

The EM-algorithm starts with a proper choice of initializing values. It can avoid trapping into local minima as well as it speeds up the convergence process. The starting values are chosen based on K-means clustering (Murphy (2012)). The vector of dissimilarity score u is split into two subsets. Then the means, variances of subsets and fractions of values assigned to specific cluster are picked as the initializers of $\hat{\beta}_{S_k}$, $\hat{\sigma}^2_{S_k}$ and $\hat{p}$ respectively. The decision, which state stands for the strong friendship depends on $\hat{\beta}_{S_k}$ parameter. The low $\hat{\beta}_{S_k}$ suggests that the value of dissimilarity score is lower, what means that individuals are closer to each other in terms of the characteristics. Hence, the low $\hat{\beta}_{S_K}$ indicates strong relationship, while the high signalizes the weak contact. The results of the EM-algorithm are presented in Table 3.

Table 3: Parameter estimators and log-likelihood value for EM-algorithm of the student network

EM	$\hat{p}$	$1-\hat{p}$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}^2_1$	$\hat{\sigma}^2_0$	$\log L$	iterations
	0.63	0.37	2.19	4.67	0.73	1.14	-489.45	89

The finite-mixture model applied to student network shows clearly that the probability of assignment to the strong contact is relatively high. On average, there is higher probability to be assigned to the strong relationship than to weak. Moreover not only has the state 1 lower coefficient but also the lower variance. It means, the strong contact is more concentrated around its constant than the weak contact, whose error term exhibits the larger variance. As the relative values between constant parameters are large, it means both subsets do not interfere with each other. The small number of iterations indicates the initialization scheme was properly chosen. To visualize the fit to data, one may use the density plot of original u together with density plot of estimated $f(u;\theta) = \hat{p}\phi(u_k;\hat{\beta}_1) + (1-\hat{p})\phi(u_k;\hat{\beta}_0)$.

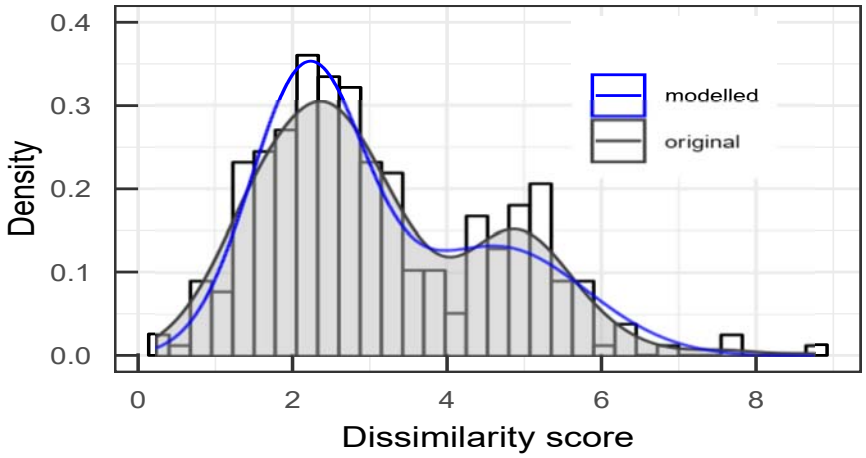


Figure 2: The density plot of original and modelled variables in the student network

The graph presents the histogram of original data values with the original and the modelled density functions (presented in the form of lines). The plot indicates the modelled density fits the data relatively well. The density line of modelled variable follows the original density function, deviating only around estimated values of $\hat{\beta}_{S_k}$'s. Moreover, there is drop around the dissimilarity score of 4, which is probably a threshold - the point where individual-specific probability $\hat{\pi}_k$ is around 0.5.

The EM-algorithm presents also individual-specific probability vector $\hat{\pi}$, which varies per each observation. The distribution of the observation-specific probability of being in the state of strong relationship is presented in the Figure 3.

The majority of individual-specific probabilities is close to 1. There is also many links exhibiting probabilities close to 0 of having the strong contact. It poses a threat that some of the connections manifest the weak contact, hence they are untransmittable. When performing process of rumour propagation, one may expect that overall probability $\hat{p}$ should offer the faster spread of the event over the network. The use of $\hat{\pi}$ may probably block the percolation of rumour by some untransmittable links, hence the diffusion may not reach every individual. The graph of student network with shown $\hat{\pi}$ probabilities for each link can be found in Appendix E.

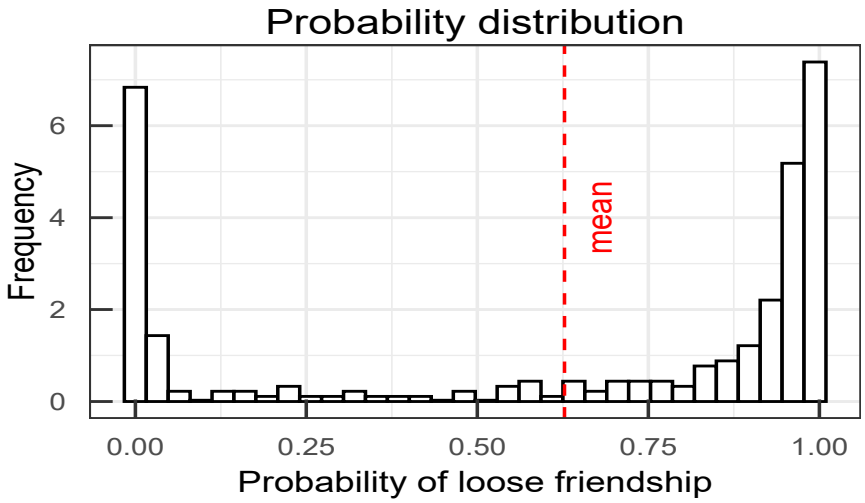


Figure 3: Histogram for distribution of $\hat{\pi}$ in the student network

5.2 Corporate law partnership network

The estimation of parameters proceeds in the similar way to the student network. We use the equation (7) on which the finite-mixture model will be applied. The dissimilarity score u_k will be regressed on the regime-specific constant and the noise (whose variance is also regime-specific). Similarly to the student network, the dissimilarity of the connection is enriched by the match-specific characteristic (m_k). In the attorney network it denotes the intensity of connection. Hence, analogically to the student network, the new dissimilarity score is defined as the sum of the euclidean distance in individuals' characteristics and the ratio $\frac{1}{m_k}$ (exactly as the equation (22)). The estimation of parameters for a product of match-specific characteristic and the dissimilarity score can be found in the Appendix B. The impact of the match-individual characteristic on the regime choice is assumed not to be so influential as in the student network. The volatility and the range of the variable (from 1 to 3) is smaller than in the other data set, hence the regime-switching mechanism is affected mostly by the characteristics' distance between agents.

The initialization for the Expectation-Maximization algorithm proceeds the same way as the student network. The initial values are chosen based on the K-means clustering. Then the EM-algorithm is performed accordingly. The results

are shown in Table 4.

Table 4: Parameter estimators and log-likelihood value for EM-algorithm of the attorney network

EM	$\hat{p}$	$1-\hat{p}$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}_1^2$	$\hat{\sigma}_0^2$	$\log L$	iterations
	0.41	0.59	8.04	21.53	3.64	9.12	-3106.24	181

The algorithm converges in 181 iterations, what indicates that the optimal parameters lay far from the initial values. The probability of the strong contact is lower compared to the student network. In the attorney network, the variation within the network is much bigger than in the other data set. One may see it in the values of the $\hat{\beta}_1$ parameter – it oscillates around 8 – while $\hat{\beta}_0$ is around 21.5. Similarly the situation is repeated in the variance. The standard deviation of the error term is larger in the weak contact state, while the noise in the state 1 exhibits the low variance. Moreover, the volatility is higher in the attorney than in the student network. One of the reason may be the high number of connections. As the attorney network exhibits more link (measured by average degree node), the more various the edge list is. The attorney network represents the corporal partnership in which lawyers make connections for business purposes, rather than based on their characteristics. Hence, some of the agents share a link, although they are substantially distant from each other. The fit of the model to the observed density is shown in the Figure 4.

The density plot shows the fit of the finite-mixture model to the original model is not exact. The original density plot is negatively skewed and does not exhibit any threshold point, which would be explained by the switch of the regimes. Moreover, original density plot is not characterized by any bell-shaved curve. The most of the dissimilarity score is concentrated around the value of 6. Then the variable decreases linearly in frequency. The finite-mixture model indicates that the utility score value of 16 can be a threshold, in which the edges are assigned to state 0. However, the original density plot is almost linear in its shape and based on the original data we are not able to observe the regime-switching threshold. It can be an indication that the assumption about the normal distribution of the error may not hold. One can try to apply other distributions of the error term to explain its variability better.

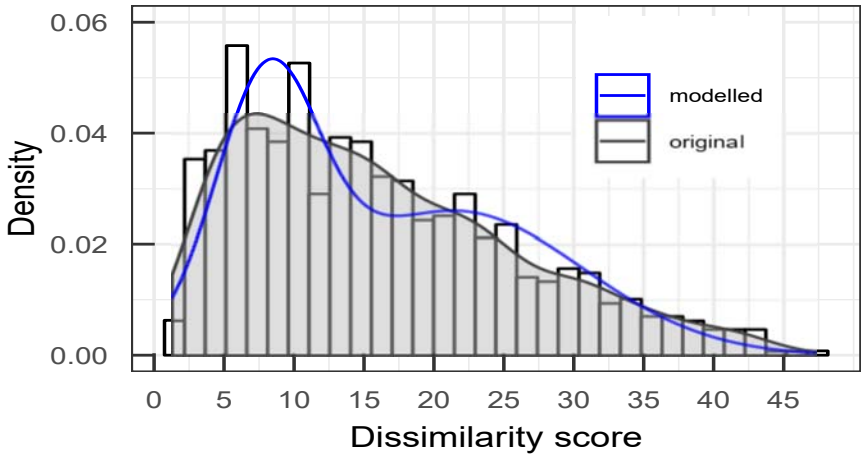


Figure 4: The density plot of original and modelled variables in the attorney network

The distribution of the individual-specific probability vector $\hat{\pi}$ of the strong contact is presented in the Figure 5.

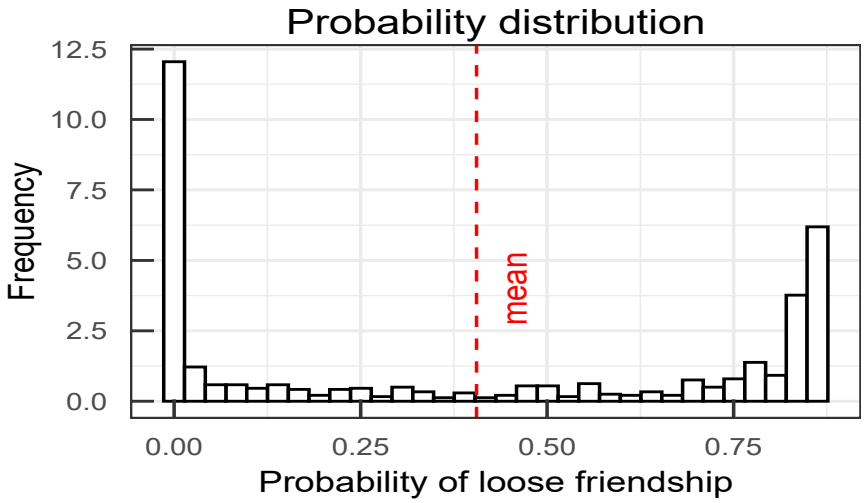


Figure 5: Histogram for distribution of $\hat{\pi}$ in the attorney network

The conclusion is contrary to the student network. Most of the individual-specific probabilities are close to 0. Some of the edges can be the dead end of the rumour percolation, if one chooses the individual-specific probabilities as the transmission rates. The general probability $\hat{p}$ might offer faster spread of the event than the

individual-specific. As the network exhibits more links than the student network it is possible that even with many zero-probabilities the event will be able to propagate through neighbouring connections. The graph of the law corporal partnership network with individual-specific probability vector is presented in the Appendix E.

6 Application of the diffusion model

This section applies the contagion model SI with parameters estimated by the EM algorithm to the empirical data. The application of the disease model follows the notation defined in the Section Methodology. One model takes into account a general probability of spread, while another is determined by a individual-specific probability, hence varying per edge. Section 6.1 analyses how the diffusion (with the same design) spreads in the simulated network with the fixed probability of the spread. We are interested to perform the same type of diffusion on the artificial networks to see how the design of networks and densities of networks may affect the spread itself. Section 6.2 analyzes the students friendship network, while Section 6.3 describes the law partnership network. The design of the simulation is followed by the notation defined in the Section 3.4. On each data set we perform the SI model, but with the empirical data we use two different configurations of the same SI model depending on the choice of the probability. This model considers two possible state of information reception. It is either susceptible (which in our setting means the individual vulnerable to accept the rumour) or infected (which denotes the person who accepted the information and is willing to pass it further). While standard SI model uses a fixed λ as a diffusion rate, the second configuration of the diffusion model in the empirical data set is enriched with a diffusion rate λ_k varying per link k. For artificially created network, we only use a fixed diffusion rate (λ without subscript) as simulating characteristics of the individuals on such a network may be unrealistic and hence obstructs the impact of sole density on the diffusion spread.

The simulation itself is designed as follows. We pick one individual, whose status is manually changed to I (infected). Other individuals in the network have the status S (susceptible). Then the diffusion process starts. The infected individual communicates with his susceptible neighbours and tries to pass the information further. The neighbours either accept it and further communicate with their yet uninfected (susceptible) neighbours or refuse to accept it from the infected individual. However, if the neighbour did not accept the rumour from the infected individual, he still may accept it from the other neighbouring infected individuals in the latter stage. The decision, whether the information is passed depends

on the probability of the event also defined as a diffusion rate. As mentioned in the Section Methodology, the transmission rate $\hat{\lambda}$ takes the same value as $\hat{p}$ (so $\hat{\lambda} = \hat{p}$) and the individual-specific probability of accepting the rumour $\hat{\pi}_k$ equals the edge-specific transmission rate $\hat{\lambda}_k$. As the probabilities of strong friendship and transmission rate take the same value, the words are being used interchangeably. The decision whether the information got passed depends on the uniform distribution. If the draw from uniform distribution for specific link is smaller than $\hat{\lambda}$ or $\hat{\lambda}_k$ then the receiver accepts the information, otherwise not. Draws from the uniform distribution are randomized, so every replication differs from each other.

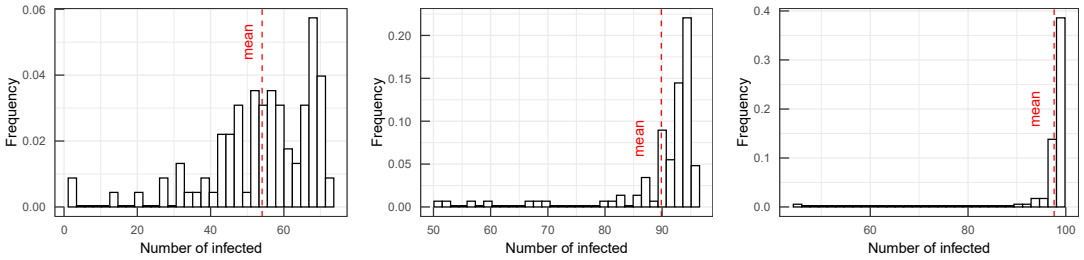
Once we have picked the initial individual, then we start the diffusion. Each infected agent tries to pass the information simultaneously to every susceptible neighbour and if fails, the diffuser does not try to pass the information again. It means that the rumour might not arrive to all of the individuals in the network.

Such a scenario reflects only one possible replication. To check how the simulation looks when different initial diffusers are chosen we will perform the simulation per each individual as a initializer, which gives us N (total number of nodes) possible replications. Moreover, to ensure that the results are reproducible, the simulation of the diffusion model will be performed 100 times per each replication and the results will be averaged out accordingly. The total number of simulation is equal to $N \times 100$. Moreover the words time and iterations are used interchangeably and accordingly one time unit/one iteration means one step in the diffusion process. For instance, if the diffusion process has in total 5 steps until it dies off, then it means we have 5 iterations/time units.

In the diffusion simulation, we will see how the infection time and range differs per each different network. We will present the distribution of the average infection range and time, together with the cumulative density functions of the diffusion averaged out over the all replications. Within the artificial networks, the analysis will be considered only for fixed transmission rate, while in the empirical networks we will analyse outcomes under different types of transmission rate. Moreover, we will present the average maximal range of the infection per different replication. Finally, we will observe, how many initial diffusers guarantee the rumour to be delivered everywhere.

6.1 Diffusion in simulated networks

In this chapter, we create three different networks depending on the number of links connecting the individuals. Every network created has in total 100 nodes (individuals). The only differing feature is the number of links. First network has 180 links, second 300 and the third one has 450 connections. Such a configuration may ensure us, whether the density of the network has an important impact on the diffusion process, while simultaneously the number of nodes stays the same. The larger number of links may speed up and broaden the whole diffusion process, however it is important to verify whether the differences in the diffusion time and the range are noticeable. Moreover it is also important how the process proceeds in each stage, what might be visualised by CDF function. For this purpose we use Erdős–Rényi model (see Figure 1) to build our artificial networks. It guarantees us the network is constructed completely at random, hence we may investigate the sole impact of density on the diffusion range. The artificial networks are attached to the Appendix C. In the simulated diffusion process in these networks we set the transmission rate $\hat{\lambda}$ equal to 0.5. The distributions of the average diffusion range under three different configurations is presented in the Figure 6.

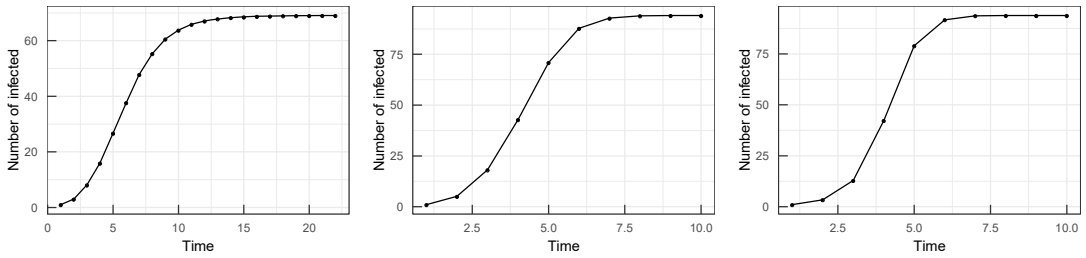


(a) with 180 links (b) with 300 links (c) with 450 links

Figure 6: Distributions of the average diffusion range in three different Erdős–Rényi networks with 100 nodes

The diffusion in the network with the smallest number of links does not arrive to many of the individuals (it arrives averagely to around 55 individuals but never to all of them). When the density is increased (the number of links is doubled to 300 connections) the diffusion is able to arrive to the most of the individuals (the mean of the diffusion range is around 90). The more links we add to the

network, the average diffusion range increases in each setting. Moreover, under the first configuration, some of the initial nodes do not pass the information even to the closest neighbours, leaving the range equal 0. Within the network with 300 links, the information always arrives to at least 50 individuals. In the most dense network, the diffusion with various initial agents arrives almost always to all of the individuals, reaching ca. 95 agents as the shortest propagation range. The distribution of the average diffusion time is available in the Appendix D. To investigate how the percolation proceeds within different networks, let us present the cumulative density function (CDF) for each simulated network as in Figure 7.



(a) with 180 links (b) with 300 links (c) with 450 links

Figure 7: CDF in three different Erdős–Rényi networks with 100 nodes

One may notice that propagation in the least dense network requires not only more time but also it is slower as it requires around 9 steps until it arrives to half of the individuals. Additionally after 5 steps, there are less than 30 individuals infected in this network. If one analyzes the more dense networks (with 300 and 450 links), then he may notice that the rumour reaches half of the total nodes after 5 steps (4 less compared to the first network). After around 7 time unit, the propagation reaches around 90 individuals and then it slowly dies out. One may see that the initial 3 steps of the propagation in the most dense network are smaller compared to the diffusion in the network with 300 links, however later on it slightly outperforms the latter one (the convergence starts after 6 steps in the most dense network). The CDF in the last network may indicate that the diffusion does not arrive to all of the individuals, as cumulative density is averaged over all initializations and the ones, which do not arrive to all agents decrease the final cumulative range of the infection.

This subsection shows that the manipulation with the density of the network does have an impact on the diffusion range. Not only varies the diffusion range depending on the density, but also the time of the propagation shortens the more dense the network is. Moreover when the networks are dense (in our example, we treat the networks with 300 and 450 links as dense), the difference between them is not so visible as compared to the network with only 180 links. The empirical networks also exhibit different density. The student network has 283 edges, while the attorney network around 860 links. As the size is more than 3 times bigger, one may expect two divergent diffusion processes. This also means that one should not compare real student and attorney networks as they have different density and the impact of the sole similarity between individuals on the diffusion process might be considerably covered by the influence of the density structure of the network.

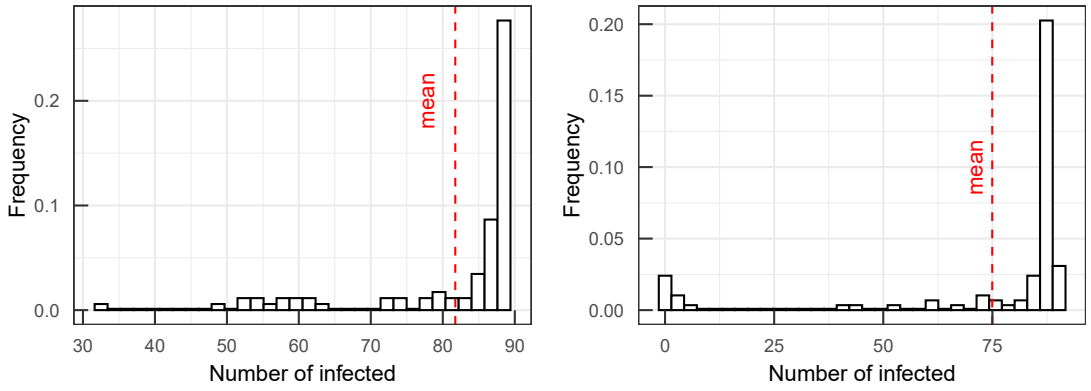
6.2 Diffusion in student network

The structure of the student network is less randomized (coincidental) than the law partnership network (See Appendix A). One can observe two major groups (cliques) concentrating most of the nodes. It may indicate, the diffusion propagates fast there, due to vast number of edges. However a few outlying vertices surrounding the concentrations may slow down the rumour percolation. There are four individuals unconnected to the main system, which will be omitted in the diffusion process.

With fixed probability, we know that all of the passages are transmittable, so the diffusion is independent from the similarity between specific individuals. When including similarity between individuals, the individual-specific probability may block some paths if the calculated probability equals 0. In such situation, the diffusion needs to go around the untransmittable edge, hence prolonging the infection time or rapidly stopping the propagation, if none of the neighbours accepts the rumour.

As we have in total 96 connected individuals in the system, there are 9600 simulations performed. The fixed probability follows the results from Table 8 and equals 0.63. The individual-specific probability varies per link and its distribution is shown in the Figure 3.

The distribution of the average maximal range of the diffusion, which varies per initialization have been presented in the form of the histograms in the Figure 8.



(a) with fixed $\hat{\lambda}$ (b) with individual-specific $\hat{\lambda}_k$

Figure 8: Distributions of the average diffusion range in the student network

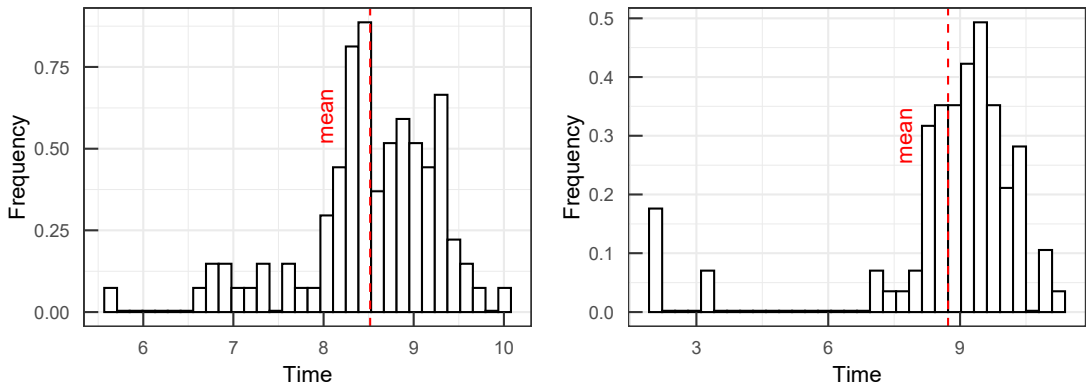
Table 5: Summary statistics for diffusions in the student network

	Min.	1st Qu	Median	Mean	3rd Qu.	Max.
with $\hat{\lambda}$	33.25	81.42	87.67	81.74	88.44	89.25
with $\hat{\lambda}_k$	0.00	83.42	86.66	74.97	87.72	90.28

The average infection range of the diffusion with fixed $\hat{\lambda}$ varies between 33 and 89, what means that on average the information does not arrive to all the individuals. The average number of infected individuals equals 82 with fixed propagation rate. 75% of all diffusions lies in the range between 81 and 89 in such scenario. The broadest diffusion occurs, when starting in the clique and the smallest, if it starts from the isolated agent. As the transmission rate stays the same for each link, these results are predictable. We are ensured, there is constant probability in each link, therefore the propagation is not impacted by the specific character of the connection. Hence, the starting point of the diffusion can be main bolstering factor. If initializer starts with a vast number of neighbours, the propagation proceeds far compared to outlying individual, who needs few carriers until the rumour arrives to the cliques.

The distribution of the individual-specific transmission rate is presented in the sub-figure (b) of Figure 8. Individual-specific transmission rate denotes here the probability of accepting the rumour, equal to probability of strong contact similarly like with fixed tranmission rate, but with link-varying values. The furthest propagation occurs, when the initializer is located in the clique (the infection range oscillates around 86-90). The smallest percolation occurs, when the diffusion starts from the isolated individuals. Few outlying individuals (mostly isolated but also the ones with too few links) do not spread the rumour in any replications at all or no further than to the nearest neighbours due to the links with the assigned probability of 0.

It is also possible to present the average diffusion time per each type of the probability. The longer diffusion time means simultaneously that the infection spreads further. The time is measured by the number of steps, the diffusion makes, until it dies out. Based on the Figure 9 and its scale on x-axis, one may assume that the propagation time in the network with fixed $\hat{\lambda}$ is very similar with individual-varying $\hat{\lambda}_k$.



(a) with fixed $\hat{\lambda}$

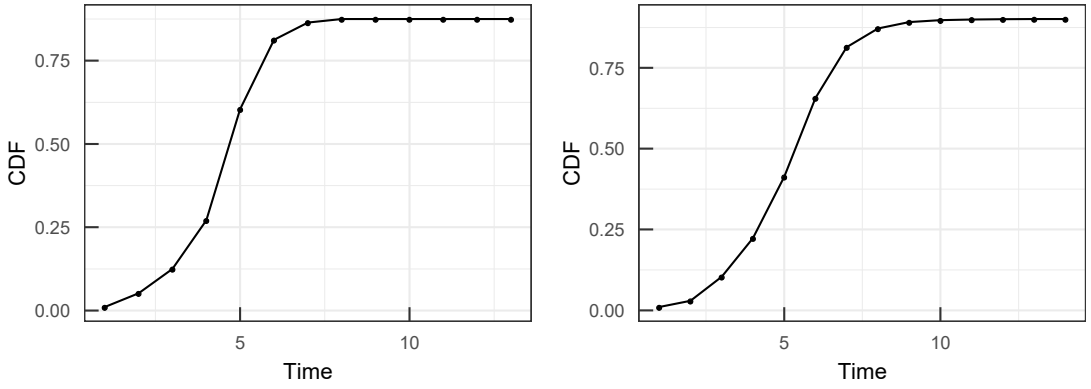
(b) with individual-specific $\hat{\lambda}_k$

Figure 9: Distributions of the diffusion time in the student network

The diffusion time reflects the range of the diffusion itself. The most of replications' times with fixed probability are gathered around the value 8.5, the same can be observed in the diffusion with individual-specific $\hat{\lambda}_k$. The minimum spreading times are equal to 5 and 2 for fixed and varying probabilities respectively. The diffusion time is strictly correlated with diffusion range. If the diffusion stops after

few steps (3-6), one may assume that the diffusion range is respectively slower.

One can visualize the diffusion process by the use of cumulative density functions on the proportion of the infected individuals to see how the rumour propagates. The CDF is a function of the diffusion averaged out by all of the replications. The cumulative density function does not reach the value of 1, as in both scenarios (with varying and fixed rate) the infection does not arrive to all of the agents. The cumulative density functions (CDF) for fixed and varying transmission rates are displayed in the Figure 10.



(a) with fixed $\hat{\lambda}$ (b) with individual-specific $\hat{\lambda}_k$

Figure 10: Cumulative Density functions in the student network

One may say, the CDF functions in both scenarios resemble the logistic function. The diffusion slowly increases at the beginning, then the propagation accelerates and after reaching around 85 % of individuals it slowly dies out. In other words, propagation (in both scenarios) starts in the vertex. Later on, via edges, the diffusion arrives to the concentration of the nodes. There, on each time unit, the information is accepted by many neighbours and propagates fast. When, the treatment arrives to the outlying individuals, the diffusion dies out. In the diffusion with constant probability, scarcity of connection to the outlying individuals may be the reason, why on average, the diffusion does not arrive to all agents in the network. In the propagation process with varying probability, the infection proceeds similarly in the initial stage. Although there are some untranslatable links through which the diffusion cannot continue, there are also edges with high acceptance probability, hence increasing chance of the information to be passed

further. In the latter stage, the only uninfected individuals are outliers. As they are relatively distant in similarity to the neighbours, they received low acceptance probability. As they also are characterized by small number of neighbours (1-2), the information does not arrive to them.

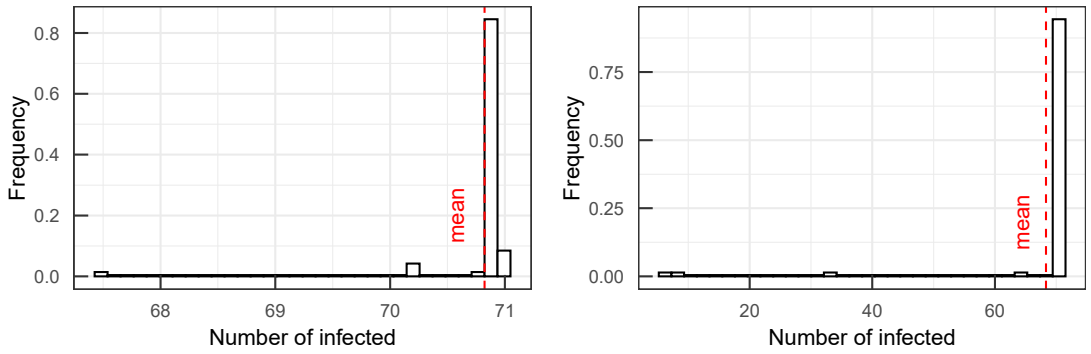
The diffusion process in the student network shows the rumour does not arrive to all the individuals. The propagation with fixed $\hat{\lambda}$ offers on average sligtly broader rumour spread compared to individual-varying diffusion rate, however the differences are minimal (See Table 6). Moreover, the choice of the initial diffuser plays an important role on the rumour spread. The Figure 8 shows that average percolation range with fixed probability of accepting the rumour percolation range varies from 30 to 90 and applying the varying transmission rate decreases the possible range minimum even more. The wrongly chosen initial student can also immediately block the spread with individual-varying probability, so the information is not even delivered to the nearest neighbours. In the real-life student networks, one may assume the diffusion does not spread constantly, but rather is affected by the type of the connection between students, therefore the individual-specific diffusion rate is better applicable to the real-life networks.

6.3 Diffusion in the Corporate law partnership network

The attorney network as given in the Appendix A consists out of 71 lawyers, exhibiting a lot of mutual connections within the network. The number of links - 859 is more than three times larger than in the student network. It implicates the diffusion should be spread within shorter amount of time, to the larger range of recipients. Moreover, there are not any outliers observable, what may indicate that the diffusion may be not constrained by the sparsity. Moreover, the character of the network differs from the student friendship network. The attorney network is relatively dense, exhibiting higher average degree node compared to the school network.

The transmission rate in the diffusion process takes the values of the probability parameters calculated in the Section 4 ($\hat{\lambda}_k = \hat{\pi}_k$ and $\hat{\lambda} = \hat{p}$). The parameter $\hat{\lambda}$ equals 0.41 and it is smaller compared to the student network. There are 71 individuals and the simulation is performed 100 times per each starting individual.

Hence we run 7100 simulations.



(a) with fixed $\hat{\lambda}$ (b) with individual-specific $\hat{\lambda}_k$

Figure 11: Distributions of the average diffusion range in the attorney network

Table 6: Summary statistics for diffusions in the attorney network

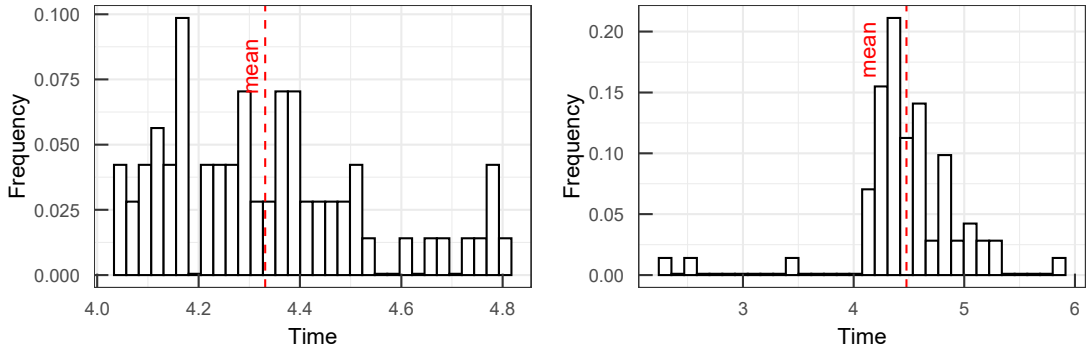
	Min.	1st Qu	Median	Mean	3rd Qu.	Max.
with $\hat{\lambda}$	67.45	70.88	70.90	70.82	70.92	70.96
with $\hat{\lambda}_k$	6.69	70.69	70.76	68.32	70.81	70.94

The average infection range (presented in the Figure 11) varies less compared to the student network. With fixed $\hat{\lambda}$, the average infected number of agents is around 70, with the minimum and maximum infection reach of 67 and 71 respectively. Moreover, most of the diffusion simulations arrives to almost all of the individuals, hence ensuring that the rumour is acknowledged by everyone. Only under few replications, the infection range is smaller than 70 individuals.

In the diffusion simulation with individual-varying probability of accepting the rumour, the propagation exhibits similar characteristics to the diffusion with the fixed probability. Most of the replications let the diffusion arrive to around 70 individuals. Other infection ranges are rarely observable. However, one may also observe, under few initializations, that the range of the diffusion is smaller than 60 individuals. For few individuals, the percolation stops immediately after the arrival to the first neighbours. Scenarios with different types of probability are very similar to each other, however under the agent-specific probability one may

observe replications, where the number of infected individuals is relatively small due to the fact, that some of links untransmittable or transmittable with very small probability. However, as the average degree node is relatively large, the diffusion can still proceed via the transmittable connections.

As the diffusion arrives to the majority of the individuals, it can be useful to demonstrate the average time of the propagation under different initializations. The more blocked links between individuals, the longer the time of the diffusion (if the infection does not die out earlier). Hence, with the similar diffusion range, the average time can help us asses, how various initializations differ. The histograms of average infection times in the attorney network are presented in Figure 12.



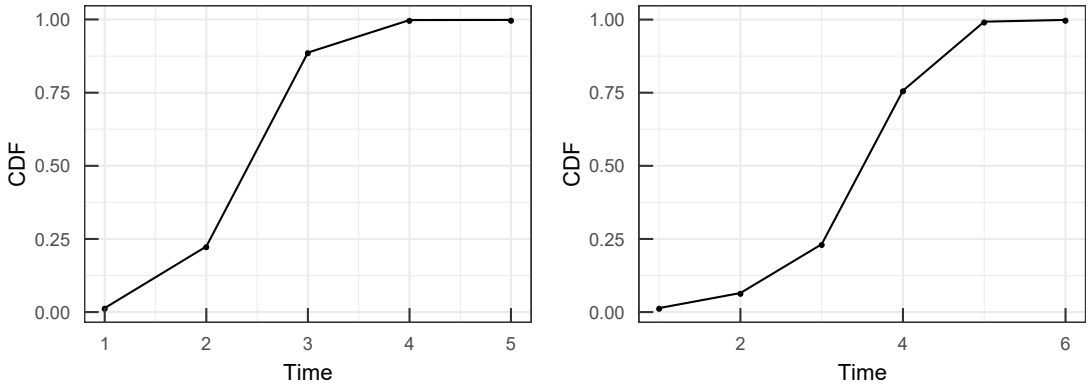
(a) with fixed $\hat{\lambda}$ (b) with individual-specific $\hat{\lambda}_k$

Figure 12: Distributions of the diffusion time in the attorney network

The infection times with fixed probability or individual-varying are very similar to each other. The average times of diffusion are 4.33 and 4.48 for fixed and varying scenarios respectively. The infection proceeds with the similar average velocity and arrives to the same number of recipients under these two different scenarios. The minimal values of the diffusion time under the agent-varying probability come from these simulations, in which the infection rapidly died out, arriving only to a few individuals.

To research on the diffusion process among attorneys, one may use the CDF function of the cumulative proportion of the infected over all individuals under the furthest initialization.

The Cumulative Density Functions in the attorney network recreates the S-



(a) with fixed $\hat{\lambda}$ (b) with individual-specific $\hat{\lambda}_k$

Figure 13: Cumulative Density functions in the attorney network

shaped logistic function, similarly to the student network. With fixed probability and the initializer himself passes the information to around 25 % individuals on average. Next, the rumour arrives to ca. 85% agents and then arrives to the corners of the network, where its speed slowly runs out. On average after 5 steps, the infection is received by all the individuals in the attorney network.

The percolation in the simulation with link-varying probability proceeds slightly slower. In the first step, the information is passed only to around 10 % of individuals. Next, it reaches ca. 25 % of agents. In the latter stages it reaches 75 % and the propagation slows down, when arriving to the most outlain individuals. On average under the furthest diffusion, the whole process proceeds in 5 steps when the acceptance probability is fixed and in 6 stages, if one uses link-specific probability. In the second scenario, the highest increase occurs in the 3 stage, while with fixed rate in the 2 step. It may indicate that few links are inadmissible due to low individual probability, so the percolation with $\hat{\lambda}_k$ requires one more step to arrive via different links to specific agents.

Within the network of the law partnership, the use of fixed probability is less informative, as the propagation proceeds always similar, no matter which initial diffuser is chosen. With varying probability, the rumour in most of the cases arrives to all or almost all individuals (69-71 individuals). However, there are few diffusers, who will not spread the rumour far if they are chosen as initializers. Moreover, the cumulative proportion of the individual infected exhibits S-curve in

both scenarios, but with different velocity and the number of stages. Hence, the attorney network does not allow to fully recognize the impact of the individuals' characteristics on the diffusion process.

7 Conclusions

The goal of this paper has been to construct a model of information spread that would depend on the similarities between individuals. In order to accomplish the task, the suitable data sets and methods have been chosen. Firstly we have created artificial networks and spread the rumour in such networks. To demonstrate the model under empirical networks, firstly the scarcer school network, and then secondly the more dense attorney network have been chosen. Once the network was known, we have used characteristics of individuals to obtain the similarity score between them. Later, this score was used to model the probability of strong contact between individuals using the finite-mixture model in which we split the dissimilarity score into two subsets denoting the states of the strong or weak contacts. Afterwards we applied the EM-algorithm to the finite-mixture model to obtain two types of probabilities, one averaged over all the links between the individuals and one with varying values dependent on the links between individuals. Finally, the paper examines how the propagation proceeds under the SI contagion model with the parameters obtained in previous steps.

The paper shows that one can apply the finite-mixture model under the normality assumption of the error term to the network analysis, which may contribute to the current literature on this subject. The EM-algorithm might also be a valuable addition to the current methodology of the diffusion processes in the networks since the way the probabilities are modelled takes into account the characteristics of individuals. Moreover, the simulation of diffusion process has shown that the simulation differs when using fixed or varying probabilities. On average, the diffusion range is similar in both scenarios, but in favour of the fixed transmission rate. Nevertheless, it has been shown that the choice of a suitable starting diffuser is crucial. Intuitively, when the percolation starts in the clique, it is delivered to a larger number of agents. When it starts from a separated individual, there is a higher chance of an immediate die-off. Applying an individual-varying probability is more realistic than the fixed probability, and it may reflect the single character of the connections between individuals. It has also been shown that the research on the diffusion process is strictly related to the network structure. The more dense the network, the broader the range of the diffusion, and also the effect of

the network structure might be a crucial factor in the propagation of the rumour.

The extent of this paper could be extended by investigating the impact of the past observations on the current reception of the information. It might be useful to concentrate more on the impact of the network structure on the diffusion process. Future studies might also focus on the impact of the number of neighbours on the information acceptance (e.g. by taking into account how many neighbours are willing to pass the information onto the individual). One can try to alter the dissimilarity score by using different metrics. Moreover, relaxing the time constraint might be a valuable addition, which not only allows connections to vary but also it allows the probabilities to be time dependent. Such a research might bring the topic closer to real-world networks.

References

Acemoglu, D., Johnson, S., Robinson, J. A., and Yared, P. (2008). Income and democracy. *American Economic Review*, 98(3):808–42.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Bailey, N. T. et al. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Campbell, A. (2013). Word-of-mouth communication and percolation in social networks. *American Economic Review*, 103(6):2466–98.

Davies, M. and Kandel, D. B. (1981). Parental and peer influences on adolescents' educational plans: Some further evidence. *American journal of Sociology*, 87(2):363–387.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Everett, B. (1984). *An Introduction to Latent Variable Models*. Springer.

Everitt, B. S. and Hand, D. J. (1981). Finite mixture distribution. *Journal of the American Statistical Association*, 77.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

Jussim, L. and Osgood, D. W. (1989). Influence and similarity among friends: An integrative model applied to incarcerated adolescents. *Social Psychology Quarterly*, pages 98–112.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, 115:700–721.

Lamberson, P. (2010). Social learning in social networks. *The BE Journal of Theoretical Economics*, 10(1).

Lamberson, P. (2016). Diffusion in networks. *The Oxford handbook of the economics of networks. Oxford University Press, Oxford*, pages 479–503.

Lazega, E. et al. (2001). *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press. Harcourt Brace & Company.

McKendrick, A. G. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:1–34.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pages 355–368.

Newcomb, T. M. (1961). The acquaintance process as a prototype of human interaction.

Newman, M. (2018). *Networks*. Oxford university press.

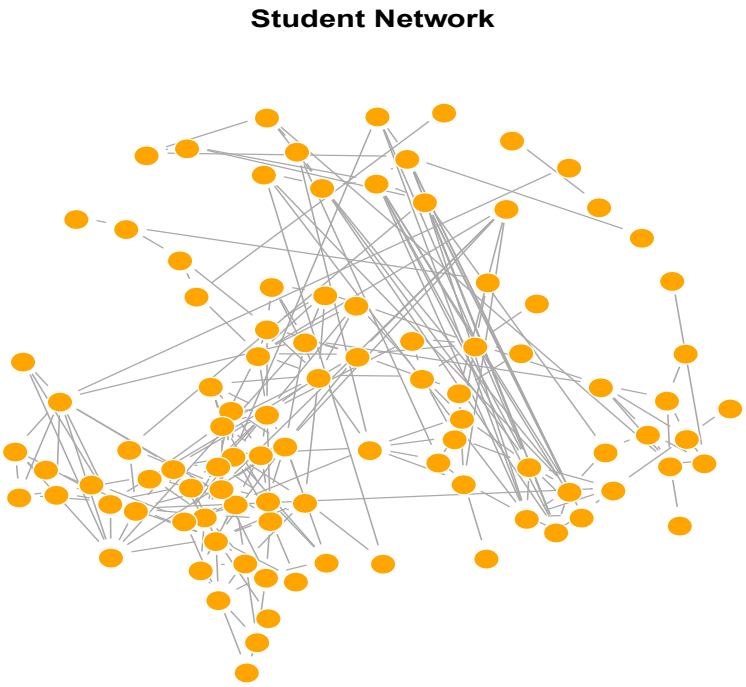
Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Danish Institute for Educational Research.

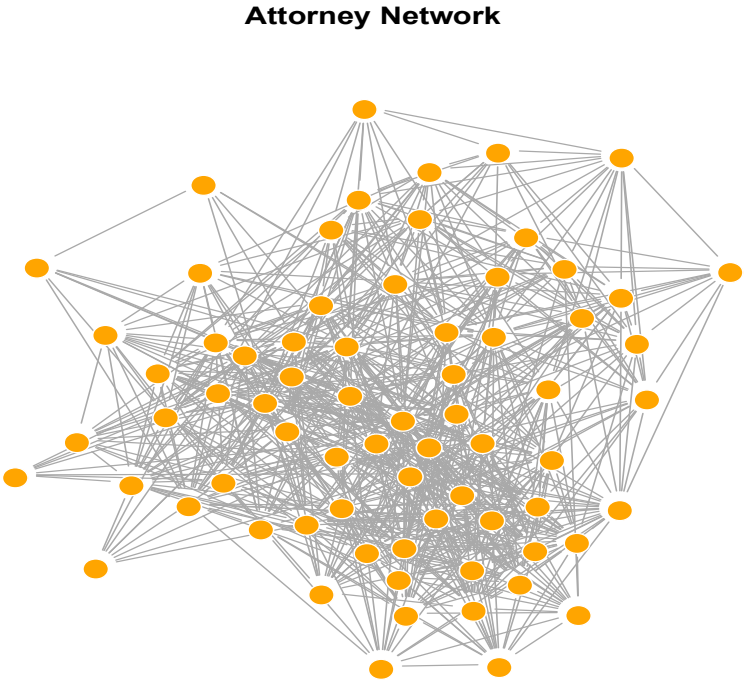
Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440.

Appendix

A Empirical Networks





B Estimation of parameters by EM-algorithm with different configuration

The estimation of parameters with a product of variables

$$u_k = \sqrt{\frac{\sum_l \left(x_{il} - x_{jl}\right)^2}{L}} * \frac{1}{m_k}$$

Table 7: Parameter estimators and log-likelihood value for EM-algorithm of the student network with the product of variables

EM	$\hat{p}$	$1-\hat{p}$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}_1^2$	$\hat{\sigma}_0^2$	$\log L$	iterations
	0.57	0.43	0.76	2.51	0.40	1.33	-398.94	75

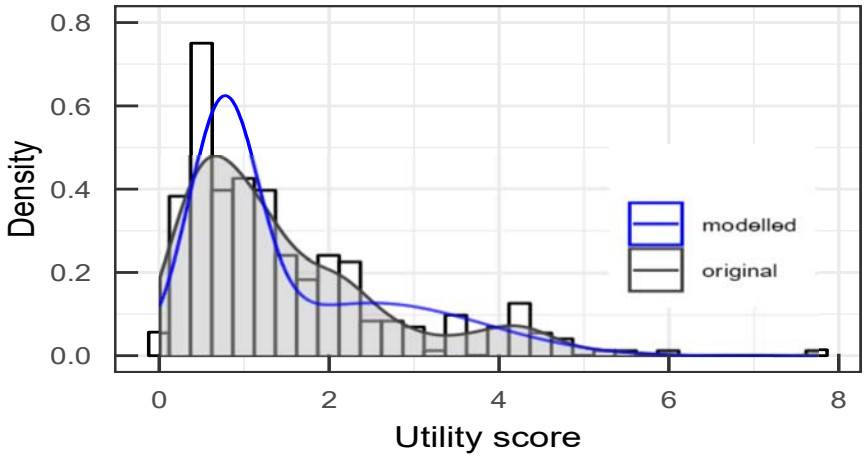


Figure 14: The density plot of original and modelled variables in the student network of the product of variables

Table 8: Parameter estimators and log-likelihood value for EM-algorithm of the attorney network with the product of variables

EM	$\hat{p}$	$1-\hat{p}$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}_1^2$	$\hat{\sigma}_0^2$	$\log L$	iterations
	0.62	0.38	5.78	18.48	3.26	9.23	-2892.99	87

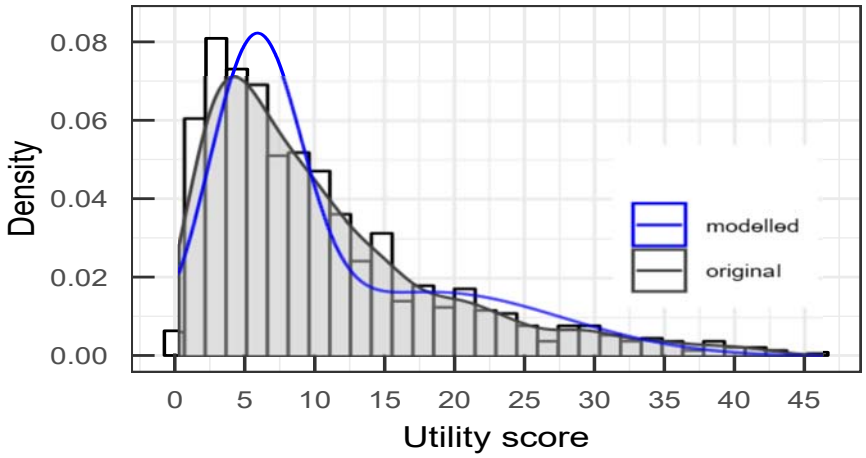
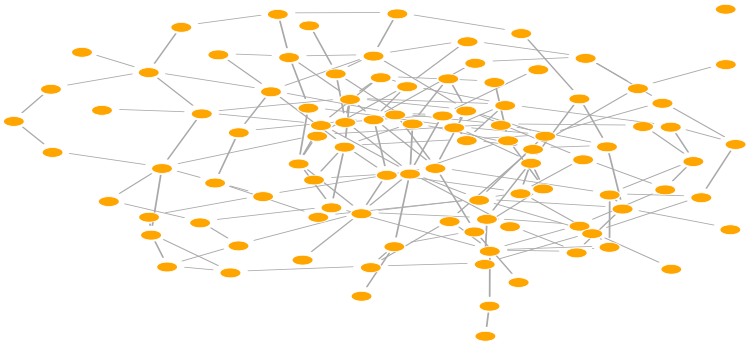
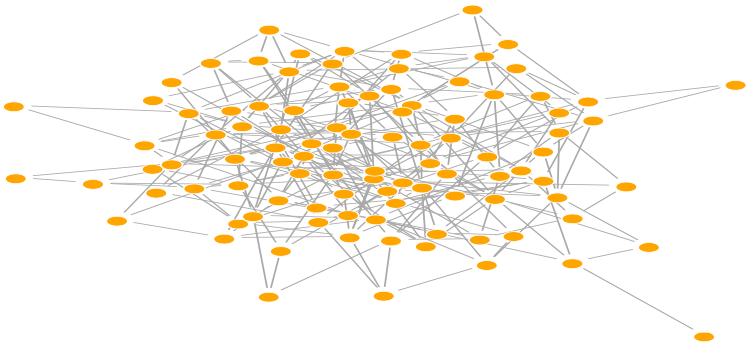


Figure 15: The density plot of original and modelled variables in the attorney network of the product of variables

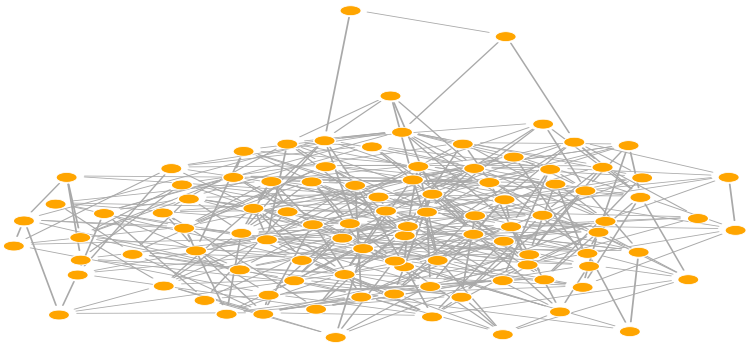
C Simulated Networks



Erdős–Rényi network with 100 nodes and 180 links

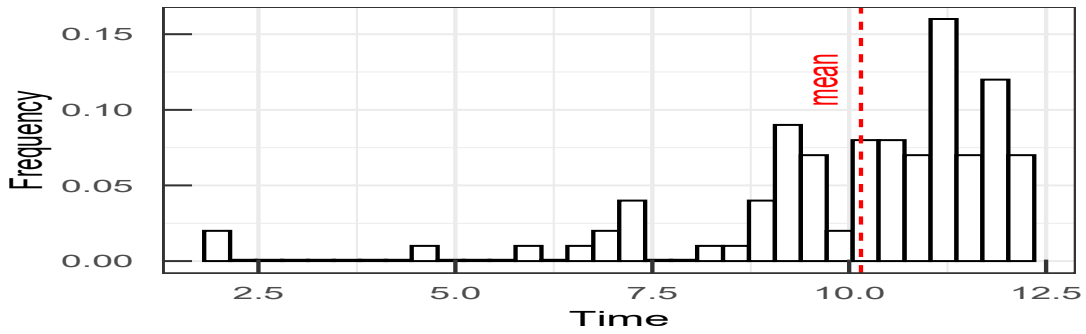


Erdős–Rényi network with 100 nodes and 300 links

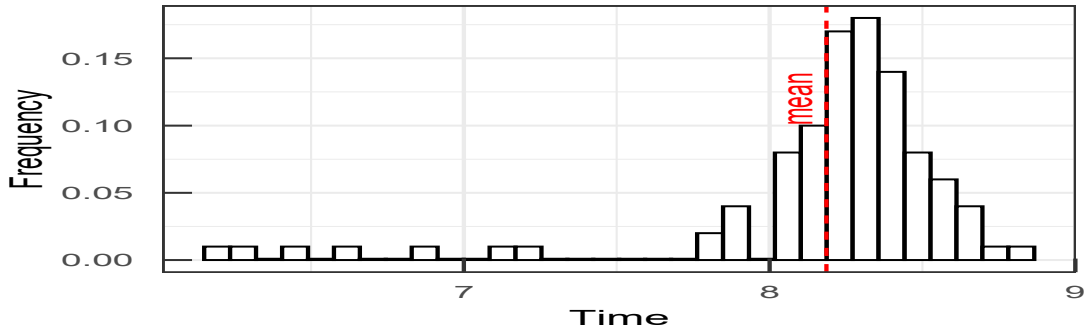


Erdős–Rényi network with 100 nodes and 450 links

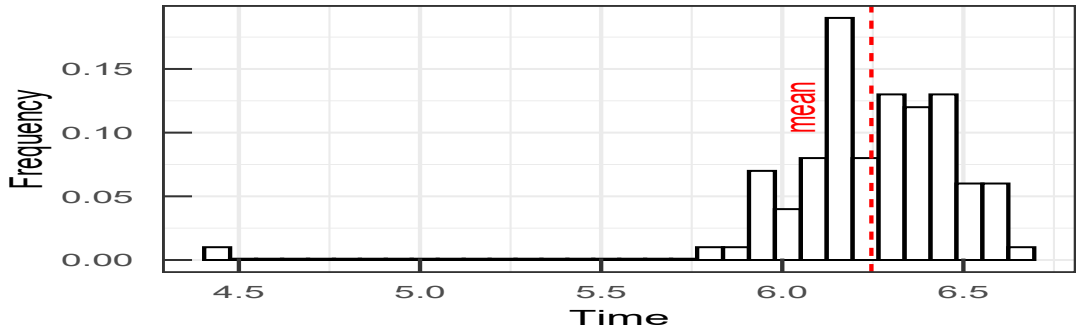
D Diffusion times in the simulated networks



Erdős–Rényi network with 100 nodes and 180 links

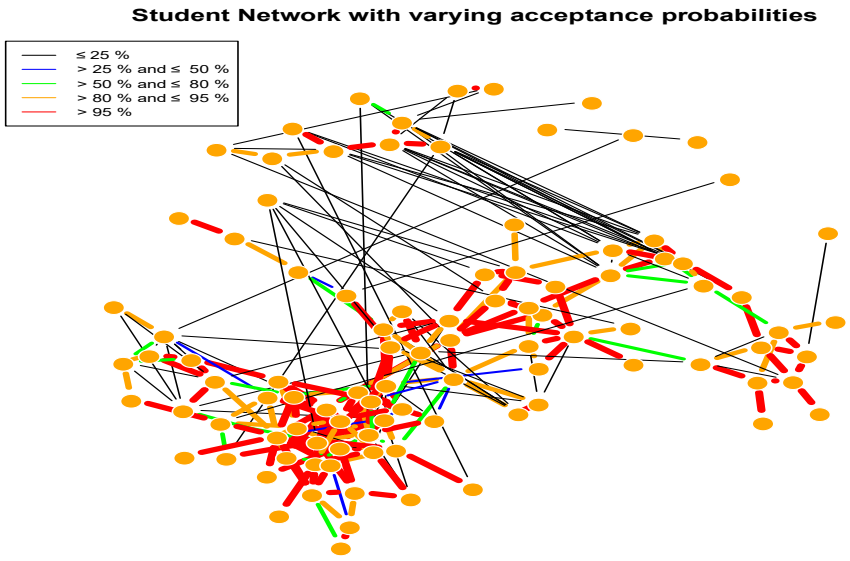


Erdős–Rényi network with 100 nodes and 300 links

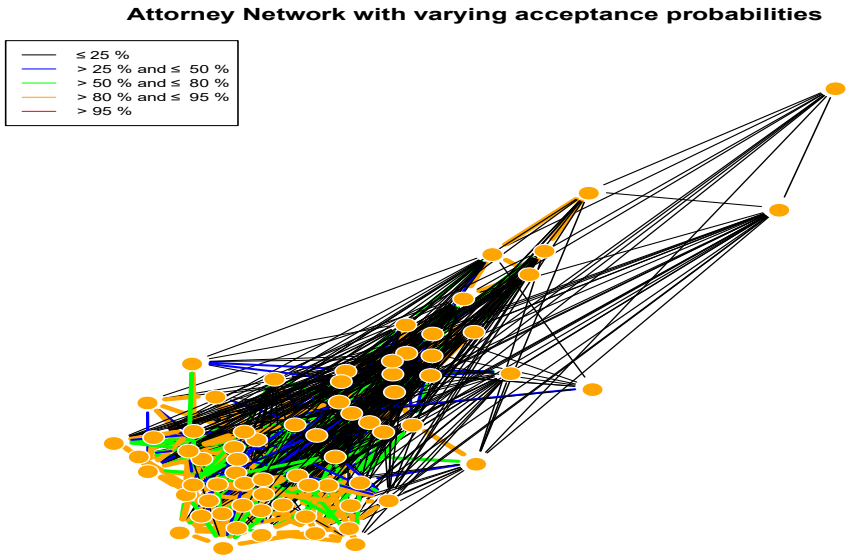


Erdős–Rényi network with 100 nodes and 450 links

E The graphs with individual-specific probabilities



Student network



Attorney network