Technische Universität München

Department of Mathematics

Master's Thesis

Finding influential Persons in the Spread of Vaccination Willingness
-
A Proof of Concept for a Centrality Measure

Raphael Schönball

Supervisor: Prof. Jürgen Pfeffer, Prof. Mathias Drton

Advisor: Dr. Guy Harling

Submission Date: 16th of September, 2022

I assure the single-handed composition of this master’s thesis only supported by declared resources.

München, 16th of September, 2022

Schonball

Abstract

If you want to spread an entity in a network like pro-vaccination sentiment in a village, it is vital to know the persons who will be central to the spread. This thesis is a proof of concept for a diffusion centrality measure which explains the spread of an entity in a network based on common and standardly used variables. The proof of concept is performed on the problem of spreading pro-vaccination sentiment in 75 Indian villages in a rural, low-income setting and based on this, a vaccination centrality measure is developed.

We will achieve this in six stages. First, 75 networks are generated on the basis of real-world networks from villages in Malegaon and Karnataka, India by matching similar nodes from both networks and by simulating missing values. Second, each edge of these networks is given an influence weight which stands for the capability of one adjacent node to exert influence on the other adjacent node. The weight is the scalar product of a parameter vector β with an attribute vector comprising characteristics of the edge, e.g. the degree difference or the age sum of the adjacent nodes. Third, a diffusion algorithm (PageRank / Laplacian Heat Diffusion) is run on these weighted networks to simulate the spread of pro-vaccination sentiment, to reach a final equilibrium state and to draw a value for each node from it. Fourth, this value of pro-vaccination sentiment is compared with the respective 'real-world' vaccination status from the generated network from step 1. This difference (sum of differences) over all nodes is minimized by adjusting the parameter vector β. The optimal parameters in β signify how relatively important the corresponding attribute is for the diffusion. Fifth, the optimal β vectors of all 75 networks are summarized. Sixth, the mean values of β are used to generate a centrality measure which can be applied in socio-centric contexts but also if only a small number of ego-centric networks are available like in practical field-operations.

We find that a high degree difference or a high socio-economic difference with one's neighbours or being a community leader are indicators for being central in the diffusion of pro-vaccination sentiment. The derived vaccination diffusion centrality is tested and the results are in the range of standard centrality measures.

Zusammenfassung

Wenn man etwas in einem Netzwerk verbreiten möchte, wie z. B. die Impfbereitschaft in einem Dorf, ist es von entscheidender Bedeutung, die Personen zu kennen, die für die Ausbreitung zentral sind. Diese Masterarbeit ist ein Proof of Concept für ein Diffusionszentralitätsmaß, das sich darauf konzentriert, die Verbreitung einer Einheit in einem Netzwerk auf Grundlage von allgemein und

standardmäßig verwendeten Variablen zu erklären. Zum Beweis des Konzepts wird das Maß auf die Ausbreitung von Impfbefürwortung in 75 indischen Dörfern in einem ländlichen, einkommensschwachen Umfeld angewendet und daraus wird ein Zentralitätsmaß spezifisch für Impfbefürwortung entwickelt.

Wir erreichen dies in sechs Schritten. Zunächst werden 75 Netzwerke auf der Grundlage realer Netzwerke aus Dörfern in Malegaon und Karnataka, Indien, generiert, indem ähnliche Knoten aus beiden Netzwerken zusammengeführt und fehlende Werte simuliert werden. Zweitens wird jeder Kante dieser Netze ein Einflussgewicht zugewiesen, das für die Fähigkeit eines benachbarten Knotens steht, Einfluss auf den anderen benachbarten Knoten auszuüben. Das Einflussgewicht ist das Skalarprodukt eines Parametervektors β mit einem Attributvektor, der Merkmale einer Kante enthält, z. B. die Graddifferenz oder die Alterssumme der benachbarten Knoten. Drittens wird ein Diffusionsalgorithmus (PageRank / Laplacian Heat Diffusion) auf diese gewichteten Netzwerke angewendet, um die Ausbreitung von Impfbefürwortung zu simulieren, einen endgültigen Gleichgewichtszustand zu erreichen und daraus einen Wert für jeden Knoten zu ziehen. Viertens wird dieser Wert der Impfbefürwortung mit dem entsprechenden "realen" Impfstatus aus dem in Schritt 1 generierten Netzwerk verglichen. Diese Differenz (Summe der Differenzen) über alle Knoten wird minimiert, indem der Parametervektor β angepasst wird. Die optimalen Parameter in β geben an, wie relativ wichtig das entsprechende Attribut für die Diffusion ist. Fünftens werden die optimalen β Vektoren aller 75 Netzwerke zusammengefasst. Sechstens wird aus den Mittelwerten von β ein Zentralitätsmaß gebildet, das in soziozentrischen Kontexten, aber auch bei einer geringen Anzahl von egozentrischen Netzwerken wie im praktischen Feldeinsatz angewendet werden kann.

Wir stellen fest, dass eine hohe Graddifferenz oder eine hohe sozio-ökonomische Differenz zu den Nachbarn oder eine Führungsposition in der Gemeinde Indikatoren für eine zentrale Rolle bei der Verbreitung von Impfbefürwortung sind. Die abgeleitete Impfdiffusionszentralität wird getestet und die Ergebnisse liegen im Bereich von allgemeinen Zentralitätsmaßen.

Table of Contents

Introduction

Imagine this: You are a field worker in a medical information campaign about vaccines. You arrive in a village in a remote area where nobody has heard about the vaccine. You want to target those who will most likely be central and therefore effective in circulating the news. How do you find them?

This thesis answers this question by deriving a centrality measure concerning the centrality of a node in the diffusion of pro-vaccination sentiment, called vaccination diffusion centrality. For this, diffusions of pro-vaccination sentiment are simulated in 75 networks which are based on real-world networks. The diffusion of sentiment from one node to its neighbour is modelled with various internodal attributes, like the degree difference or the age sum, and attribute parameters. The relative importance of the attributes in the diffusion process is derived by optimizing the attribute parameters in order to create a diffusion whose final equilibrium state is closest to the real-world data about each node's vaccination sentiment, deduced from its real-world vaccination status. The optimal parameters together with the attribute values of a node with all its neighbours are summed up to reach a value which expresses how central a node is in the diffusion of pro-vaccination sentiment. It is the final centrality measure. It is demonstrated how the parameters can also be used to calculate an approximate centrality measure in a few ego-networks instead of a full socio-centric network.

Established centrality measures like betweenness centrality [1], degree centrality [61], pageRank centrality [15], leading-eigenvector centrality [61], harmonic [50], or closeness centrality [61] are applicable on socio-centric networks, and so is the measure which will be introduced in this thesis, vaccination diffusion centrality. However, in practical field operations and in many research settings, socio-centric network data cost a lot of time and money and are often not fully achievable due to nonrespondents. An easier approach is interviewing various (randomly chosen or selected) persons from a village and asking them about their contacts and the links between their contacts. These ego-centric networks are part of common study settings and a lot of research has been done to facilitate the collection process, as described in [35] or in [34]. This thesis fits into common designs and methods by adding one version of our centrality which can be calculated for a node even without knowing more about its wider environment and the full network. Approximated vaccination diffusion centrality values for egos of ego-centric networks can be calculated without knowing more about the network. The only condition is that a few ego-networks from the analysed community are available as samples for the whole network and that the model parameters have been trained in similar circumstances in a socio-centric network.

Standard centrality measures are often only used to analyse the basic network topology. General information about how well a node is connected / how central the node is in a graph are returned. Our presented approach is context specific: It is obvious that diffusion mechanisms in networks change depending on which diffusion is regarded. E.g. a person who is central in the spread of legal information is not necessarily central in spreading information about IT. The proposed vaccination diffusion centrality is based on a simulation of pro-vaccination sentiment in networks and can hence be classified as a context-specific centrality measure.

Valente's paper on network intervention gives an overview of four tactics which can be used to cause behaviour change in a network [69], e.g. how to reach most people with a product. Looking at the example of marketing strategies, knowledge about the underlying diffusion mechanisms can boost a marketing campaign. The Big Seed Marketing approach for instance described by Watts, Peretti and Frumin [72] uses "tell a friend"-strategies to expand classical marketing campaigns which have a lot of viewers with viral propagation features by adding "tell a friend"-features. Many additional viewers can be reached without increasing the campaign's budget. Currently employed strategies using influencers in social media for promoting products also hope for viral propagation over networks. The analysis of diffusions like these has been subject to research for a long time. The presented vaccination diffusion centrality is focused on pro-vaccination sentiment. Later in the thesis, it is described how this proof-of-work can be expanded to other diffusion processes of other entities in different network settings.

Vaccination hesitancy has been studied intensively in the last years. Various reasons for vaccination hesitancy or uptake have been identified concerning different vaccines [65] [75]. E.g. during the Covid19 pandemic, Latkin et al. found that "family and friends discouraging vaccination […], not knowing whom to believe about vaccine safety […], and concerns that shortcuts were taken with vaccine development […] were all independent predictors of lower vaccine uptake." [45]

Various approaches to counter vaccination hesitancy have been presented like e.g. "the use of social mobilization, mass media, communication tool-based training for health-care workers, non-financial incentives and reminder/recall-based interventions" [38] as summarised by Larson et al.

This thesis is organised as follows. In chapter Background, similar research is discussed and potential variables for our model are found in the Literature section. In the Methods chapter, the proof of concept for finding importance parameters concerning the spread of pro-vaccination sentiment in a network and based on the parameters, for defining a vaccination diffusion centrality measure is explained. The Result Section shows and tests the distributions of the importance parameters for all 75 villages and compares the vaccination diffusion centrality with other standard centrality measures.

The Discussion Chapter summarises the results and gives an outlook on the generalisation of the proof of concept for any diffusion process.

Background

Before finding potential variables for the later model in the literature review subsection, an overview of related research is presented. The generalized form of the diffusion centrality presented in this work is compared with all other approaches.

Banerjee et al. [10] analysed the spread of microfinance products among inhabitants of villages in Karnataka in rural India. The 75 analysed villages are also part of the data basis of this thesis. The spread of microfinance information and uptake was simulated on the basis of given general demographic information about households, villages and some persons such that it finally corresponded to the actual microfinance information or uptake state of the households over the time of the introduction of these products. The paper introduces the communication centrality for each node which is equal to the fraction of households who would take a microfinance product if the analysed node was initially informed (a seed node) and a diffusion centrality as simplified and easier to calculate proxy for the communication centrality. Although the idea of optimizing a simulation result for fitting with a real-world outcome is shared between this paper and our thesis, this paper is based on another diffusion model (close to the Susceptible-Infected-Recovered Model) and estimates the centrality measure for each node with reference to the reached households in the simulation whereas our approach measures centrality as being central to the diffusion simulation which is independent of the starting node. The independence of the starting node has the advantage that the centrality values can be computed for all nodes at once.

Based on the works by Banerjee et al. [10], Leng et al. [46] define in their 2020 paper contextual centrality which is based on the diffusion centrality defined by Banerjee et al. [10]. To the diffusion centrality, it additionally incorporates a mixture of different network properties (including other measures like the Eigenvector and Catz Centrality) and relevant node characteristics. Our approach differs from this paper like it differs from the original Banerjee paper [10].

Kang et al. [39] [40] introduce a diffusion centrality which measures how central a specific node is in the diffusion of a property (the entity which spreads). Their general concept is useable in most diffusion processes in any social network. It can be applied to many diffusion properties and to most diffusion processes (both tipping, cascade, and homophilic models).

The reach of an average spread of property p is measured by a value called fixed point. The centrality measure is defined by the fixed point of the diffusion if the regarded node v did have the diffusion property minus the fixed point of the diffusion if the regarded node v did not have the diffusion property. So, this measure captures how much the fact that v had p boosted the average spread of p over the network in comparison to if v did not have p. [40]

In this thesis, we present a different approach for a diffusion centrality by explaining how (relatively) important personal attributes were for the diffusion (whether a variable helped the spread or rather stopped it) and then using the importance of these attributes and the attributes themselves to calculate how central a node was in the diffusion. This allows us to approximate the centrality measure even if the full socio-centric network data are not available. Additionally, we focus on the example of the diffusion of pro-vaccination sentiment in this thesis.

Another similar approach to the one which is presented in this thesis is topic-aware diffusions (or topic-aware social influence propagation models). Barbieri et al. [12] alter the influence probabilities of the Independent Cascade Model [63] and the Linear Threshold Model [20] such that they are topic, interest- and authoritativeness - dependent: The probability of each edge to diffuse awareness from the sender node to the receiver node is dependent on the following parameters:

- the ability of the beginning node of the edge of influencing the respective end of the edge concerning a topic (authoritativeness)
- how central the topic is in the information which is circulated (Each information is a mixture of different topics)

This is multiplied by the interest of the end of the edge in a topic. The mechanisms of the Linear Threshold model are the basis of this developed model. In a follow-up paper [9], the concept is applied to viral diffusion in social networks, involving a budget k and with a focus on scalability for large networks.

The information sources of this model and of the model presented in our work are different: Barbieri et al. assume information about how one person can influence another concerning a topic and how interested a person is in various topics. Our edge weights are based on personal attributes like age, degree, socio-economic status etc. which can be assumed to be more often part of standard questionnaires. Additionally, the diffusion models assume that the spread of an entity stops if in one step, no additional person is convinced. Our diffusion algorithm assumes that the information reaches every node, are passed on by every node and the question whether one person is convinced or not is finally decided according to the final equilibrium spread. This makes it independent of the starting node and thus more robust. Moreover, our model is optimized to simulate and explain a real-world spread whereas the work of Barbieri et al. focuses on how to reach most nodes.

In 2009, Tang et al. [67] started investigating the field of topic-related propagation models. Their model incorporated an interest in a topic for each user of a social network and for each piece of information propagating through the network a distribution over topics (to which topics does an information belong?). Similar to this work, the diffusion algorithm and centrality measure PageRank [15], which was introduced by Brin and Page in 1998, is used with altered edges depending on how likely the receiving node thinks that it can be influenced by the sending node and on how likely the sending node thinks that it will influence the receiving node depending on a topic. Again, the underlying information basis is different to our model. The overall approach is also different: The goal is to find central nodes of a topic-related diffusion based on nodes' interest in a topic using pageRank, whereas our model tries to explain which variables were how important for the diffusion and deriving a centrality measure from that.

The field of influence maximization is related to the topic of centrality measures. It answers the question of how to choose the optimal set of initial nodes such that most nodes are reached by a diffusion caused by these initial "seed nodes". Many papers have analysed the problem and proposed optimal solutions, like Leskovec et al. in 2007 [47], Wang [19] in 2010 or Goyal et al. [32] in 2011 and More et al. In 2013 [2]. They all consider the topology of networks, but do not take edge or node attributes and their effects on the diffusion into account.

The use of influence maximization techniques in global health settings is described in a 2012 paper by Nicholas Christakis et al. They conducted a randomized intervention study in villages in Honduras analysing which targeting methods for finding influential village members for spreading health related behaviour change would produce the highest behaviour adaption rate among the population: randomly choosing individuals; choosing individuals with most ties; or choosing nominated friends of individuals. Their findings were that nominated persons of individuals had the biggest impact on the behaviour of the population, whereas taking individuals according to their in-degree was worse than a random selection. [43]

In comparison to the field of spread maximization techniques, out approach is more explanatory and retrospective: this work focuses on how a diffusion has occurred, how it can be explained and what can be deduced for future diffusions. Spread maximization techniques rather focus on how to choose the best set of nodes to maximize the spread over the network.

All in all, as previous research focused on topics like spread maximization, diffusion centralities, and topic-related diffusions, not much research has been done to explain a diffusion on the basis of practical, easy to collect and standardly used variables using simulation techniques and how to derive an easy-to-use centrality measure from it. Vaccination hesitancy is an intensively studied topic, but no centrality measure about finding central nodes in the diffusion of vaccination willingness for information campaign targeting has been developed.

This thesis is a proof of concept for a diffusion centrality measure which focuses on explaining the spread of an entity in a network based on common and standard variables. Which personal, tie and network characteristics helped the spread along an edge and were the driving forces behind a diffusion? It can be applied to socio-centric networks and to ego-centric networks even if the rest of the network is unknown, as long as the measure's parameters could be trained on one socio-centric graph. This thesis is a proof-of-concept of the approach and as such, it explains the spread of pro-vaccination sentiment in 75 villages and derives a vaccination centrality measure from it.

Literature Review

In the Methods Section, a model about the spread of pro-vaccination sentiment will be designed. The choice of variables which can explain how pro-vaccination sentiment flows through a network is crucial for the model. Therefore, the hypotheses about why a variable is included in the model will be based on the current state of scientific findings in the respective field.

The literature review findings are presented below. Not all findings or potential variables could be included in the model due to the design choice of avoiding a large and complex model and due to the fact that not all potential variables were part of the generated dataset. It is stated below when a variable was included in the model.

Potential variables for explaining the diffusion of pro-vaccination sentiment in a network were collected from two perspectives. First of all, predictors for vaccination hesitancy were analysed. As a second perspective, indicators for influence on peers were investigated.

Vaccination Hesitancy Predictors

The problem of vaccination hesitancy has been in a special focus of scientific work recently and many strategies have been derived to effectively deal with the issue as summarised by Larson et al. [38]. The model, which will be designed in the Methods section, focuses on the diffusion of pro-vaccination sentiment. Indicators for vaccination willingness / hesitancy can be interesting for designing the model and are therefore collected in the following.

Vaccine hesitancy, defined as "delay in acceptance or refusal of vaccination despite availability of vaccination services" by MacDonald et al. [49], is a widely researched phenomenon with a recent focus on COVID-19 vaccination hesitancy.

Many predictors for vaccination hesitancy have been discussed in the last years. In 2021, Allington et al. [5] conducted a general analysis with various models of different predictors. These predictors are explained in the following and merged with additional predictors and findings of other studies (individually cited below) and especially with the findings of a paper by Hudson and Montelpare [36]. Their literature review concerning predictors for vaccination hesitancy using a broad search strategy analysed 57 studies or reports in English from 2006 until 2021 which included relevant words like 'vaccine hesitancy', 'vaccine confidence' or 'COVID-19 vaccination uptake' among others.

The following predictors for vaccination hesitancy were found.

Personal Attributes:

- Age
  Supported by four studies in various countries, elder age groups were less hesitant to vaccinations than younger age groups. The increased vaccination hesitancy of elder age groups is also attributed to an increased social media use by younger people [27][51] and worries of young parents or during pregnancy [52].

- Sex
  A positive correlation was identified between being female and being hesitant about vaccinations [4][5]. Dror et al. [26] also found being female to be a positive predictive factor for vaccination hesitancy. Callaghan et al. [16] found as well in their Covid-related study in 2020 that women were more resistant concerning vaccines.

- Parental Status
  Parents with young children were most likely to have an aversion to side effects of a vaccination. Families with more children tended to refuse vaccinations more often. [22] Having a child is a negative predictor [26].

Socio-Economic Status Attributes:

- Socioeconomic Status

  Socioeconomic status is found to be predictive both for and against vaccination hesitancy. Depending on circumstances, it has mixed effects on vaccination hesitancy. In the US, more wealthy persons were generally more vaccination hesitant, but one study found that low-income families had a less trustful attitude towards vaccinations [76]. In West-Africa, however, social status of families relates to the vaccination status of its children, causing wealthier families to being able to afford vaccines and hence having more vaccinated children [64]. In a field study in India concerning vaccination hesitancy, households with more than USD 143 monthly income were 0.7 as likely to be vaccination hesitant in comparison to richer households [71].

- Education

  The effect of education on vaccination hesitancy depends on the situation and is not a clear predictor. Whereas worries about the safety about vaccinations decrease with a higher educational level in Canada [17] and a higher education status leads to more age-appropriate vaccination uptake in Greece [22], in the US, parents who refuse vaccinations for their children tend to live in neighbourhoods of higher educational level [74]. Robertson et al. find that vaccination hesitancy is higher in groups with lower education levels in the UK as part of a study which focused on hesitancy in various population subgroups [60]. Using the WHO vaccination hesitancy scale, the education of mothers from low- and middle-income countries did not relate to vaccination hesitancy [70]. Additionally, Wagner et al. found that in an Indian field study, those with a high school diploma were 0.10 times likely of being vaccination hesitant in comparison to those with less education [71].

- Household Income

  A slight negative correlation between household income and vaccination hesitancy was identified. [4]

- Working from home

  In a US survey by King et al., persons working outside their homes (e.g. in the construction and protective service) had twice the vaccination hesitancy than those working from home [44].

- Sector of Occupation

A study from 2020 by Dror et al. [26] finds that an occupation in the health sector is a non-significant predictor for vaccination willingness. Over all sectors, Dror et al. also discovered a higher vaccination willingness among persons who lost their jobs due to the pandemic in comparison to others.

Personal Beliefs Attributes:

- Mistrust in Authorities
  Mistrust in authorities is a predictor for hesitancy. Mistrust in medical staff leads to consulting online sources. Mistrust in the government leads to doubts about the effectiveness and the safety of vaccines. The study argues that this issue can be addressed by knowledge sharing through peers. [25]

- Moral Foundation
  Amin et al. [7] focused on the issue of underlying moral values of vaccination hesitancy. In two correlational studies, they found that individual foundations in harm ("ability to feel the pain of others; kindness, gentleness, nurturance" [23]) and fairness ("reciprocal altruism; justice, rights, autonomy" [23]) as described in Haidt's Moral Foundations Theory [33] are not significantly linked to vaccination hesitancy whereas liberty (resentment toward domination by others [23]) and purity ("striving to live in an elevated, less carnal, more noble way" [23]) are linked to more hesitancy.

- Risk Aversion
  Risk Aversion leads to more vaccination willingness, but only if a disease is regarded as prevalent and / or dangerous [55]. The survey conducted among health care workers and members of the general population in Israel however finds that the self-perception of high risk for severe COVID-19 infection is a significant predictor for vaccination willingness [26]. Generally, risk averse parents tend to being vaccination hesitant which can be linked to their preference to being passive rather than taking a risk through active behaviour (also known as the omission bias) [18].

Cultural Attributes:

- Media Consumption

Vaccination hesitancy was found to be negatively correlated with the consumption of print and broadcast media whereas a slightly positive correlation of social media reliance and vaccination hesitancy could be found [4].

- Religion

  A connection between vaccination hesitancy and religion was described in the work by Volet et al. [42] and supported by a study in Venezuela by Andrade [8]. Protestants, Catholics, Jewish, Muslims, Christians, Amish, Hinduist and Sikhist had all religious reasons for not getting vaccinated. Non-halal ingredients, Ramadan fasting, faith in divine healing and that you should not go against the will of God, "the use of aborted foetal cells for vaccines' production" are among the reasons of vaccination hesitancy within these religious groups. At the same time, three studies from Ghana, Uganda and Zimbabwe also showed contrary effect, that the vaccination rate within religious groups were actually higher than the population average [42].

- Caste

  Wagner et al. found that those with scheduled castes or scheduled tribes were 3.48 times more likely of being vaccination hesitant in comparison to an 'other caste' group. Those with a 'backward or unknown caste' were 0.89 times as likely of being vaccination hesitant [70]. The scheduled tribes and scheduled castes belong to the group of untouchables in the Indian caste system and are not part of the regular four caste groups called varnas. The 'other and backward caste' group is lower than the highest three varnas and higher as the Scheduled Tribes and Scheduled Castes and comprise about half of the Indian population [77].

- Language

  Geipel et al. found out in a study in Hong Kong that vaccination hesitancy could be decreased when information was received in English in comparison to the respondent's native Chinese [29]. Aktürk et al. found in a study in Munich that vaccination willingness campaigns were much more effective when being conducted in the mother tongue of the recipient [3].

- Rurality

  Generally, persons from rural areas had less confidence in vaccinations than persons from urban areas [4]. Uptake is influenced by accessibility in rural areas [22].

Indicators for Social Influence on Peers

In this section, various types of variables about social influence were identified in the literature, e.g. “the ways in which individuals change their behaviour to meet the demands of a social environment.” [41]
Herbert Kelman describes three forms of social influence: Compliance (Agreeing to others but keeping oneself's differing opinion to oneself), Identification (Agreeing to someone due to his/her social acceptance or respect, e.g. a celebrity) and Internalization (Making someone else’s opinion your own opinion). [41]

Social influence indicators are fundamental to the diffusion of behaviour patterns in a social network. Concerning the spread of pro-vaccination sentiment, a node with a high social influence on a neighbouring node is likely to spread much pro-vaccination sentiment to its peer. Therefore, the focus of this section was to look for types of indicators for social influence to include some specific indicators in the model which will be designed later.

Ambler et al. published a study in 2021 [6] about social influence on risk taking in a rural setting in Malawi. 1028 farmers in rural Malawi were given a special amount of cash to invest it. The money was provided given the information that a test person had invested as well. The test person was either “a randomly selected individual (peer), the elected chair of the farmer club (formal leader), or a professional extension agent assigned to work with the club (external leader)” [6]. Behaviour decisions by peers were found to be most influential on individual behaviour, followed by decisions taken by elected leaders. The decision by external leaders had the least influence on the farmers’ decisions. Furthermore, vaccination hesitancy in a group can often be linked to an influential leader in the community, as described by Goldstein et al. [31]. By building the trust of community leaders, effective information campaigns for polio vaccination uptake could be launched.

Risselada et al. [59] find in their study about opinion leadership in social networks that degree centrality is a significant indicator for opinion leadership. Opinion leaders are defined as consumers that exert a disproportionate influence on those around them [44], hence the findings by Risselada can be expanded to having an effect on social influence.

Research by Ni et al. [53] shows that choosing nodes according to their Betweenness-Centrality has a positive effect on the total spread of this diffusion over the whole network. Banerjee et al. find as well that nodes with a high eigenvector centrality have a high influence on their neighbours [10].

Chung et al. deal with the topic of tie vs. network attributes for explaining social influence. They find that both tie and network attributes have an effect on the success of information campaigns. In specific, they point out that tie level factors like embeddedness have more influence on the effectiveness of social-oriented crowdsource-campaigns whereas network factors like centrality have a bigger effect on technology-oriented campaigns. [21]

Susarla describes how measures like a high average neighbour degree and a high clustering coefficient can be important to measure the cohesion of a cluster. This paper investigates the success of Youtube videos after their upload. The cohesion is found to be an indicator for a network-wide diffusion process [66].

Ross [62] finds in a study about the effectiveness of a psychologist's advice on mothers with low socioeconomic status that educational and socioeconomic differences have a positive effect on social influence. The expert's advice was more influential than a peer's advice. [62]

So, this section summarised existing literature both about social influence and about vaccination hesitancy. Both topics are relevant for the analysis of pro-vaccination sentiment. They interact with each other. A community leader has an influence on the community's vaccination hesitancy [31] and a high socio-economic difference increases the social influence of the higher end of a node when at the same time depending on the circumstances, a high socio-economic status can lead to a low level of vaccination hesitancy [62][64]. Due to the fact that the goal of this work is the relative importance of parameters, independence of the variables was not desirable. When looking at absolute importance parameters which could be transferred to other variable constellations, the independence would be crucial.

Hence, to simulate the diffusion of pro-vaccination sentiment in a network, both topics affect the diffusion. Variables from both topics are thus valuable for explaining the diffusion and therefore, variables from both topics are included in the model.

Methods

Summary

The goal of this thesis is, as a proof of concept, to explain how relatively important various personal, tie and network attributes are for the diffusion of pro-vaccination sentiment in a social network in a rural, low-income setting, and to derive a centrality measure from it. We will achieve this in six stages. First, 75 networks are generated on the basis of real-world networks from villages in Malegaon and Karnataka, India by matching similar nodes from both networks and by simulating missing values. Second, each edge of these networks is given an influence weight which stands for the capability of one adjacent node to exert influence on the other adjacent node. The weight is the scalar product of a parameter vector β with an attribute vector comprising characteristics of the edge, e.g. the degree difference or the age sum of the adjacent nodes. Third, a diffusion algorithm (PageRank / Laplacian Heat Diffusion) is run on these weighted networks in order to simulate the spread of pro-vaccination sentiment, to reach a final equilibrium state and to draw a value for each node from it. Fourth, this value of pro-vaccination sentiment is compared with the respective 'real-world' vaccination status from the generated network from step 1. This difference (sum of differences) over all nodes is minimized by adjusting the parameter vector β. The optimal parameters in β signify how relatively important the corresponding attribute is for the diffusion. Fifth, the optimal β vectors of all 75 networks are summarized by showing their distributions. Sixth, the mean values of β are used to generate a centrality measure which can be applied in socio-centric contexts but also if only a small number of ego-centric networks are available like in practical field-operations. A sample application is performed; central nodes are detected in a network and the findings are compared with standard centrality measures.

A systematic overview of the Methods as a flow-chart diagram is provided in the Appendix.

Details

Network Generation

Since we did not have a single empirical network dataset available containing both substantial personal information about respondents and at the same time vaccination uptake information, we combined two network datasets from India to generate one which we could use for analysis. Random network models, like the Erdos-Rényi random network model, the Barabási-Albert network model or the Watts-Strogatz network model, could not be used in this thesis. Generally, they do not focus on personal (or nodal) attributes, their interdependences, and their distributions over the whole network. The Erdos-Rényi random network model, which was used e.g. in [37] or [68], assumes that tie formation is independent of previously existing ties and does not produce scale-free network features. The also popular scale-free Barabási-Albert network model [11] which was used e.g. in [68] produces scale-free structures, but does not consider the interdependences of personal attributes, their effects on cluster formation and on distributions of single attributes over the whole network. The Watts and Strogatz "small-world" network model [73] which was e.g. used in [1], does not generally produce a scale-free degree distribution and is therefore not likely to produce realistic scale-free attribute distributions over the whole network as well.

So, this thesis needed graphs which were closest to real-world graphs in order to get realistic homophily and clustering values because our approach relies on simulating diffusions which are the closest possible to real-world diffusions; close to real-world homophily values with respect to e.g. education are central in explaining the process of passing on information, i.e. generally who is talking to whom and to which extent are educational differences central in the diffusion of pro-vaccination sentiment; close to real-world clusters are also important for the diffusion due to the fact that clusters change the way how information spread over the network and how often/intense individuals are or are not confronted with an information.

As random networks would not have created the desirable results, to create many personal attributes and realistic interdependences between these variables both for one individual but also with respect to their effects on tie formation and realistic distributions of attributes over the whole network, two real-world networks were matched.

For the personal attribute information, we used a dataset comprising 75 sociocentric networks covering 75 villages in Karnataka, India that was originally used to study the diffusion of microfinance products [10].

The households in each village provide the nodes for one network. Personal data from some individuals from the households were also collected. To have most information about each node, we added individual-level data of the leader of the household. When the data of the household leader were missing, the spouse of the head of household was selected. Each household is thus represented by one person.

The nodal data comprise information such as general household descriptions (number of rooms, roof type) and personal information about the household leader like age, sub-caste, education, language, native home, and occupation. Two households have a tie if any household member of at least one of the households indicated a relationship based on advice-giving or -seeking with someone from another household. Only individuals from about half of all households were interviewed; these households were randomly selected out of all households in all networks, stratified by religion and geographic sub-region. The ties are undirected.

There were households where nobody was personally interviewed and only the general information about the household were available. It was not the focus of this proof-of-concept to incorporate missing data in the model, so to reach a realistic database, the missing information had to be simulated based on the available general household data. As it has been proven to be a good method for simulating missing data [67][57], random forest models were trained and tested for each variable which had missing values on the basis of all general household information (religion, roof type, number of rooms, number of beds, electricity available, whether the property is rented or owned, whether this household is the home of a community leader, degree), tie information (number of triads, local clustering coefficient (also known as local transitivity)) and network information (Betweenness centrality). Due to the large number of variables which had to be simulated for about half of the 14904 nodes in all networks, the results could only be tuned using autotuning methods and all available feature variables were used for all predictions. Depending on whether the variable type is continuous or categorical, a random forest regression or classification was run. The number of trees was set to 500 and the model was trained on 80 % of the dataset and tested on 20% of the dataset[1]. The simulated variables were age, gender, education, mother tongue, English abilities, village native, whether one works outside or not, whether monetary savings are available, and the caste. Depending on the original format of the variable, the output had to be adjusted (numbers rounded to integer etc.). To improve the performance of the models, autotuning was applied to the number of variables randomly sampled as candidates at each split.

[1] For replication purposes, use the implementations of the R randomForest and caret packages with these additional parameters:
mtry=[optimizedParameter], na.action=na.omits

Autotuning

To improve the performance of the random forest prediction for filling up empty data entries, parameter tuning was performed. Due to the large number of different variables which were simulated and various predictors in multiple villages, individual approaches were not feasible and autotuning was used. The parameter "mtry" describes the number of variables randomly sampled as candidates at each split. Making use of a function of the R-package "randomForest" [14], the out of bag error estimate was minimized and the optimal value for "mtry" was returned. Fig. 2 shows the out of bag error for various values and the respective minimum. This method was used to decrease the error rates of the various random forest models.

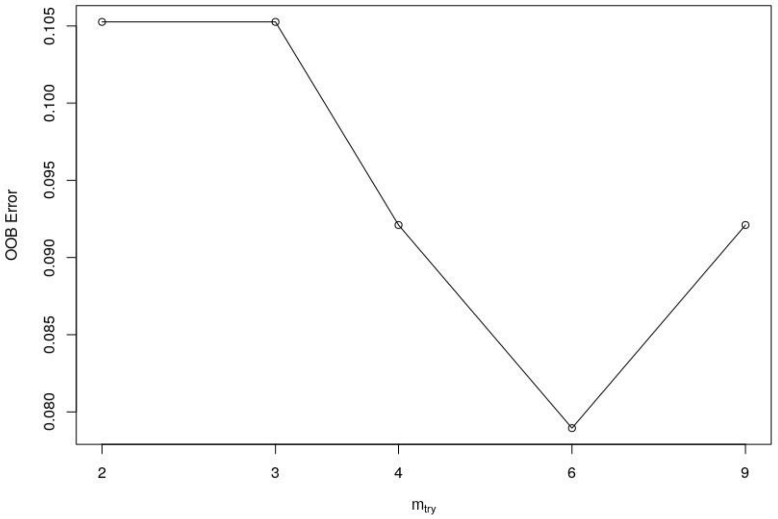


Figure 1: Example Parameter tuning for "mtry" and finding the minimum error at m=6

The dataset from Karnataka did not include vaccination related information. A second dataset from Malegaon, India by JP Onnela et al. was chosen. This paper investigates polio vaccination hesitancy in social networks. Interviews with family heads from 2462 households in 25 neighbourhoods were conducted. Each household is a node in the network. Ties between two households were formed when one nominated the other based on advice-seeking. Each household could only nominate up to 4 other households. As well as including household attributes like whether the household has a TV, a cooking cylinder, a toilet, how many people live in the household, the dataset furthermore includes a variable for parents' acceptance or refusal of polio vaccines for eligible children in the household: accepting; reluctant; refusing; and not applicable (no vaccine eligible household member). [56]

In the process of exploring the dataset, a web application visualising the network and its multiple parameters was implemented. It can be accessed online[2].

To create a single dataset, we added the attributes of the most similar node from Malegaon to each node from Karnataka, matching with replacement. The matching was based on a nearest neighbour approach, taking the absolute difference in each variable between the respective data points (also known as Manhattan distance or L1-norm).

Definition 1. The optimal node index $\widehat{j}_i$ from Malegaon for the i. node from Karnataka is

$$\widehat{j}_i = argmin_{j=1,\dots,M} \textstyle\sum_{z=1}^{Z} \|x_{iz} - y_{jz}\|_1$$

with x_{iz} all nodes from Karnataka and their shared attributes z, $i = 1, \dots, N$,
y_{jz} all nodes from Malegaon and their shared attributes z, $j = 1, \dots, M$,
$z = 1, \dots, Z$ all shared attributes of the nodes,
N the number of nodes from Karnataka, M the number of nodes from Malegaon,
Z the number of shared attributes,
and $\widehat{j}_i$ the optimal node from Malegaon for the i. node from Karnataka.

Due to the scarcity of shared variables, all existing shared variables were selected as matching variables. The vaccination status as a central analysis variable was not used in both the simulation of data and the matching. For obtaining comparability between the variables from the different data sets, they needed some adjustments.

Matching Variables

- Degree

  Nodes of similar degree were supposed to be matched. However, different questioning and other backgrounds of the surveys led to non-comparable outcomes.

  To create comparability, all degree values are converted in relative position normalization, hence they were ranked in increasing order and the rank was replaced by a percentage (current rank divided by the maximum rank).

[2] https://schoenball.shinyapps.io/Malegaon; due to performance reasons, please use firefox or google chrome.

Definition 2. Relative Position Normalization

The relative position normalization of a vector $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$ is

$$\tilde{x} = \frac{1}{N} \begin{bmatrix} |\{x_l | x_l < x_1\}| \\ |\{x_l | x_l < x_2\}| \\ \vdots \\ |\{x_l | x_l < x_N\}| \end{bmatrix}, \text{ with } l \; = 1, ..., N \; .$$

E.g. the lowest degree was replaced by a value close to 0%, the highest value was replaced with 100%. This facilitated the matching of two nodes according to where in the degree distribution they appear in comparison to all other nodes of the respective network.

- Closed Triads

  Nodes with a similar number of closed triads were to be matched. However, the number of triads differed a lot, also because of the fact that in Malegaon, each person should only name up to four alters. So, each data point was converted using the relative position normalization.

- Betweenness-Centrality

  The Betweenness-Centrality [28] of each node was used as a variable for the matching as well. Again, the values of both networks could not be compared. Hence, in analogy to the degree distribution, each data point was converted using the relative position normalization.

- Education

  To match nodes in terms of the educational background, the education systems had to be compared.

  The education system in Karnataka is based on standards, junior colleges and universities. In this analysis, the first until the fifth standard were classified as basic school education. They were put, together with no education, in category 1. Standard 6 until 10, which is equivalent to having passed the S.S.L.C. exam, was considered as middle or high school education and were hence in category 2. After the S.S.L.C. exam, pre-colleges are offered which were considered along with an uncompleted university degree, a finished university degree or above as category 3. Other diplomas were considered as category 4.[1] These categories were matched with equivalent standards from Malegaon. There, no education and primary school were summarised in category 1, middle and high school education in category 2, university students and graduates and above were in category 3 and religious and professional trainings and other educational backgrounds were summarised in category 4.

In the distance cost function, a penalty of 1 was added in case the two compared data points from Karnataka and Malegaon were from different education categories. If they were from the same education category, no penalty was added.

- Toilets and Latrines

  The Malegaon data set differentiates between having a toilet and having no toilet. The Karnataka data set differentiates between having no toilet, a common toilet available or owning a toilet. The last two options were summarised as having a toilet.

  The penalisation of differences between two compared to data points were performed in analogy to the education status.

- Number of Rooms

  Both data sets comprise the number of rooms in a house. The number of rooms were subtracted from each other and in analogy to the degree normalization, each data point was converted using the relative position normalization.

- Socio economic status

  The socio-economic status was estimated based on additional information in both data sets. In the Malegaon data set, the availability of a TV, a telephone and a cooking cylinder were combined with the (in analogy to the degree normalization) normed ratio of rooms per persons. Each of these variables had the same weight for the socio-economic status of persons in Malegaon.

  The socio-economic status in Karnataka was estimated by the availability of electricity in a household, by the existence of monetary savings and by the fact whether the corresponding house is rented or owned.

  The resulting socio-economic variables from both data sets were normed such that a comparison of different socio-economic levels was possible (so, each variable from Karnataka was normed to one and each made up a third of the whole socio-economic variable for Karnataka. This was similarly performed for the Malegaon data set).

To measure the success of the matching procedure, the distributions of the best fit differences per matching variable are shown. Additionally, various ego variables of one of the created networks were compared with variables of the original networks from Malegaon and Karnataka by generating histograms of central variables in order to make a visual comparison possible. Moreover, it was calculated which rows (household) from Malegaon were how often the best fit for a row (household) in the Karnataka data set. This was performed for all 14904 households in 75 villages.

Diffusion, Optimization and Centrality Measure

Influence Weights

We then created an adjacency matrix for our combined network comprising weighted directed edges to represent the capability of nodes to influence one-another concerning pro-vaccination sentiment. A weight is a scalar product of a parameter vector, called importance parameters, with various personal, link and network attributes, chosen based on a review of relevant literature and available variables.

Definition 3. Influence Weights α

The influence weight of the edge between node i and node j is

$$\begin{aligned} \alpha_{ij}(\beta) \ = \quad & \beta_1 \ * \ A_1 && A_1 \text{ being e.g. difference in degree} \\ + & \beta_2 \ * \ A_2 && A_2 \text{ being e.g. sum of age} \\ + & \beta_3 \ * \ \ldots && \\ & \vdots && \\ + & \beta_K \ * \ A_K && A_K \text{ being e.g. difference in socio-economic status} \end{aligned}$$

with $\beta_k \geq 0$ being the importance parameter vector entries,
initialized by $\beta \ = \ (1, ..., 1)^T$,
$A_1, ..., \ A_k$ being the attributes,
and $k = 1, ..., K, \ K$ is the number of attributes.

Definition 4. Influence Matrix A

All $\alpha_{ij}(\beta)$ make up the influence Matrix

$$A(\beta) \ = \begin{bmatrix} 0 & \cdots & \alpha_{1j}(\beta) & \cdots & \cdots \\ \vdots & 0 & \vdots & \vdots & \vdots \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \alpha_{i1}(\beta) & \cdots & \alpha_{ij}(\beta) & 0 & \vdots \\ \vdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

Personal Attributes

Variable	Type	Considered Relation	References
Age	Integer between 18 and 80	Numerical Sum	[52], [51], [27]
Community Leader	Yes / No	Categorical: 1 – community leader 0 - otherwise	[6], [31]
Socioeconomic Difference	Value between 0 and 1 equally based on • number of rooms per person ratio • electricity availability • latrine ownership • house ownership • tv ownership • phone ownership • cooking cylinder ownership	Numerical Difference	[62], [76], [64], [70]
Degree	Integer	Numerical Difference	[59], [44]
Tie / Edge Attributes			
Variable	Type	Considered Relation	References
Closed triads	Integer	Increase of passing on pro-vaccination sentiment with more closed triads at the beginning of an edge	[66], [21]
Average Neighbour Degree	Numerical Value	Increase of passing on pro-vaccination sentiment with a larger average neighbour degree at the beginning of an edge	[66], [21]
Transitivity / clustering coefficient	Numerical Value	Increase of passing on pro-vaccination sentiment with a larger clustering coefficient at the beginning of an edge	[66], [21]

Network Attributes			
Variable	Type	Considered Relation	References
Betweenness-Centrality	Numerical Value	The higher the Betweenness Centrality, the more effectively pro-vaccination sentiment is passed on.	[53], [10]

In the following chapter, the importance parameters (the parameters of the linear combinations of attributes) will be optimised such that the influence weights are specifically optimised for the influence on pro-vaccination sentiment.

The choice of the variables is based on the literature review from the previous section. In the following, the specific reasons why a variable was included are explained and specifics about its integration in the model are defined:

a. Age

   Generally looking at vaccination willingness and age, the literature shows that elder people tend to be less vaccination hesitant than younger. Therefore, hypothesis was made that the flow of pro-vaccination sentiment is higher the older both ends of a tie are. The sum of both ends of a tie was taken as a measure for how pro-vaccination sentiment is spreading along the tie between the two nodes.

b. Community leader

   The works of Ambler et al. [6] show that being a community leader has a positive effect on influence and the work of Goldstein et al. [31] reports that the influence of a community leader is also observable with respect to vaccination willingness or hesitancy. If a community leader has received pro-vaccination sentiment, then in the model, the leader passes it on very effectively (the contribution to the total edge weight by the attribute community leader is 1 if the beginning of the edge is a community leader and 0 if not).

c. Socio-economic difference

   According to the literature, a positive connection between a socio-economic difference and social influence is existent. The numerical difference between two nodes contributes to the total edge weight.

d. Difference of degree

   Based on the literature review, degree centrality [54] was found to be significant for opinion leadership and hence social influence. If one node has a higher degree as a neighbour node, it is more likely to have more social influence than its neighbour. Therefore, a difference in degree was included in the analysis.

e. Number of Triads, Clustering Coefficient and Average Neighbour Degree

   As pointed out by the literature, having a greater tie-based centrality helps to initiate a greater diffusion process. Therefore, it is assumed that the higher values in closed triads, in the clustering coefficient and in average neighbour degree, the more effective is the process of passing sentiment on.

f. Difference in Betweenness centrality

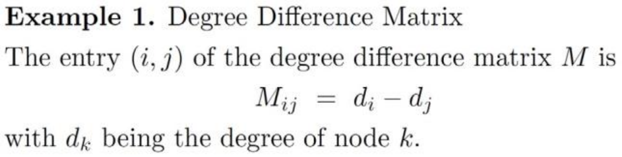
   The Betweenness-Centrality is as well investigated as it has been shown that choosing nodes as seeds by their Betweenness-Centrality score has a positive effect on the reach of the diffusion. A positive effect of eigenvector centrality on influence has been shown as well. In the model, the higher the value in Betweenness-Centrality, the more effective is sentiment passed on to another node.

Technical Implementation

Due to the costly calculations involved in this analysis, code efficiency was emphasized throughout the design of the technical implementation.

The influence weights for each edge of the network were generated by summing up relations from various ego, tie or network attributes multiplied by the respective importance parameter. The relations were differences, sums or simple values of one of the adjacent nodes. Due to code efficiency, a matrix for each attribute relation (e.g. degree difference or age sum) was generated for the whole network. The matrix is an adjacency matrix concerning the attribute relation. The three types of relations are briefly discussed:

The difference matrices are introduced with the example of the degree difference matrix:



Example 1. Degree Difference Matrix
The entry (i, j) of the degree difference matrix M is

$$M_{ij} = d_i - d_j$$

with d_k being the degree of node k.

The matrix which involves a sum was the age matrix. The ages of both ends of the network are summed up.

The matrices comprising only information about the beginning end of an edge, hence the node which passes on pro-vaccination sentiment to another node, are simply adjacency matrices which encode the value (e.g. being a community leader) of a variable at the beginning end of an edge.

To reach comparability with other variables in the model, to ensure transferability of the results to other models, measures and networks, the matrices were normalized. To avoid being overly influenced by outliers, the median was chosen for the normalization.

The difference matrices are skew-symmetric:

$$A^T = -A$$

Therefore, the median of all matrix entries is by definition 0 and therefore not adequate for normalization. So, negative values of the difference matrices (degree and socio-economic status) are interpreted as implying a weak influence on the other node, positive values as a strong influence. So, the negative values were shifted to the positive domain by adding the minimum value to all others. Then, the normalization was performed by dividing by the median.

Definition 5. Matrix Normalization for difference matrices

The entries of Matrix M are normalized and shifted to $\tilde{M}$:

$$\widetilde{M_{ij}} = \frac{M_{ij} + |\min_{i,j\ =\ 1,\ldots,N}(\{M_{ij}\})|}{\text{median}(\{M_{ij} + |\min_{i,j\ =\ 1,\ldots,N}(\{M_{ij}\})\ |\ i,j\ =\ 1,\ldots,N\})}\ .$$

The other matrix types could be simply normalized by dividing by the median.

Definition 6. Matrix Normalization for other matrices

The entries of Matrix M are normalized to $\tilde{M}$:

$$\widetilde{M_{ij}} = \frac{M_{ij}}{\text{median}(\{M_{ij}|\ i,j\ =\ 1,\ldots,N\})}\ .$$

Diffusion Process Simulation with PageRank

We simulate a diffusion process across the network based on the nodes and edges in our combined network, initially without reference to the imputed vaccine willingness values from step 1. We use a mixture between a cascade and a homophilic model, as summarised by Kang [39]. The PageRank model is a probabilistic diffusion algorithm that does not assume that the overall sum of pro-vaccination sentiment increases during the diffusion. The resulting distribution of nodal values shows the intensity of some property, in this case pro-vaccination sentiment, at each node in the network, with nodes with high values close to 1 being generally vaccination-accepting and those with low values close to 0 being vaccination refusing.

PageRank is initially computed[3] on the network with the influence weights, based on arbitrarily chosen initial importance parameters for the linear combinations (e.g., 1 in every entry) [15]. Then, one unit of pro-vaccination sentiment is given to a random household (hereafter, "agent") of the initially neutral network. This is congruent with the agent being given pro-vaccination sentiment by an exterior information source like a health information campaign, doctor, or distant friend. This is encoded by having a vector with one entry for each node comprising the amount of pro-vaccination sentiment which each node has currently; hence, the initial distribution is given by a vector with 0 in all entries except for the entry of the randomly selected household which has a value of 1.

In the next iteration step, this agent meets one of its neighbours and passes on the sentiment. The sentiment influences this neighbour with a probability of the directed influence weight of the edge connecting both. Due to the fact that it is not known which neighbour the agent will meet, a probabilistic approach is chosen: It is assumed that the agent will meet all of its neighbours with the same probability (uniform distribution). So, if an agent has 3 neighbours, the likelihood that the agent passes on its sentiment to one of these neighbours is one third multiplied by the edge influence weight.

Now, the pro-vaccination distribution is given by a vector with non-zero entries for each neighbour of the initially chosen household. This is calculated in practice by multiplying the influence weight adjacency matrix with the initial pro-vaccination sentiment distribution vector (with only one entry of 1 for the initially chosen household).

Each new iteration step is reached by multiplying the adjacency matrix with the previous distribution vector. After subsequent iterations, the diffusion results in a likelihood for each node on how likely it is that one node has been influenced by pro-vaccination sentiment. The final distribution or equilibrium state is reached when the sum of the changes in likelihoods of all nodes is below a predefined threshold.

The PageRank model makes a number of assumptions relevant to this work. First, it assumes that those who are more often influenced by positive pro-vaccination sentiment are more likely to accept vaccinations.

[3] For replication purposes, use the R-igraph implemenation using the standard parameters: https://igraph.org/r/doc/page_rank.html

Second, PageRank assumes that at each iteration step a small value of pro-vaccination sentiment is added to each node. This acknowledges the fact that there are non-network ways to be exposed to pro-vaccination sentiment, including newspapers, relatives from other regions or countries, internet sources and social media sentiment.

Third, it assumes that after one agent has passed on the sentiment to one of his neighbours, it loses the sentiment. So, pro-vaccination sentiment is passed on to the neighbours in probability and does not stay with the agent. Due to the probabilistic characteristic of the diffusion and the bonus for each node in each iteration (see above), each agent will be active at every step of the iteration.

In order to let sentiment diffuse over the network, the power iteration method is applied. The iteration process is repeated until no significant changes in the distribution of sentiment vector are found (threshold 10^6) from one iteration to the next one. This equilibrium distribution of pro-vaccination sentiment in the network is a unique fixed point of the iteration process, and also approximately the eigenvector of the leading eigenvalue of 1 of the adjacency matrix. The uniqueness of the equilibrium state ensures that the selection of the random person in the beginning does not alter the results of the diffusion process and supports the fact that fundamental network and behaviour diffusion properties are investigated, independent of time or starting point of the diffusion.

Alternative Diffusion Processes

Alternative diffusion algorithms like Markov Random Walk with restart or Nearest Neighbours diffusion were considered. Only one, the Laplacian Heat Diffusion, is presented here with its basic functionalities and a brief discussion about its applicability in our model.; the others can be found in the Appendix.

Laplacian Heat Diffusion Algorithm

This algorithm simulates a diffusion over a network in analogy to the diffusion of heat in some metal body. The Laplacian matrix L is created by adding the negative adjacency matrix A to a matrix D with the nodes' degrees on the diagonal:

$$L = D - A$$

Intuitively, the eigenvalues and eigenvectors of matrix L correspond to wavelets of heat (or in our case pro-vaccination sentiment) which diffuse through the network over time. Adding up all the relevant wavelets returns the heat at a given time.

$$H_t = H_0 * exp(-Lt)$$

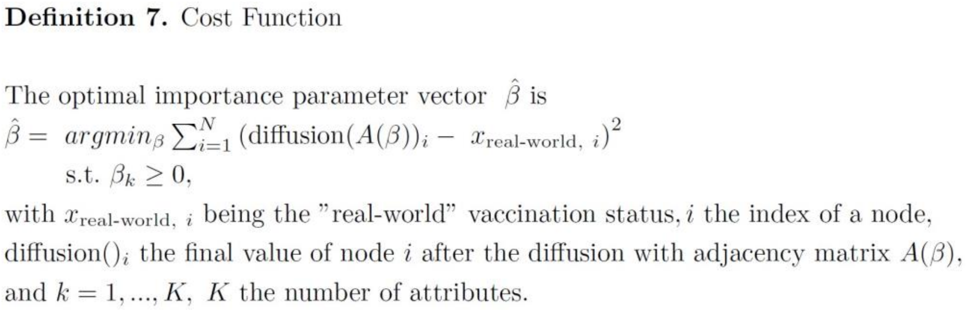
H_t being the heat distribution at time t, and exp the matrix exponential [48]. The spread is simulated until the change of the heat (or pro-vaccination sentiment) from one time step to another is below a defined threshold and an equilibrium state is reached.

This algorithm is, like PageRank, not dependent on the starting point of the diffusion. By simulating heat waves through the network, it finds basic network features. The diffusion spreads to all neighbours and not only one, so more social nodes (with more contacts) can diffuse more pro-vaccination sentiment than less social nodes. Like PageRank, the amount of pro-vaccination sentiment or heat does not increase during the diffusion but if one node influences a neighbour, the initial node loses pro-vaccination sentiment/heat after having passed it on. PageRank includes a pro-vaccination argument bonus for each node in each iteration (sentiment can come from other sources rather than social contacts) which the Laplacian Heat Diffusion Algorithm does not.

Parameter Optimisation

To optimize the importance parameters for the diffusion process, we compared the outcome values of the diffusion processes with the 'real-world' vaccination opinion values in our matched network. The goal is to minimize the squared difference between the calculated diffusion values and the 'real-world' data (the categories vaccination refusing, reluctant, and accepting). Therefore, the 'real-world' data had to be converted to numerical values on a scale between 0 and 1 due to the fact of the diffusion values being on this scale. For 'vaccination accepting' nodes, this value was set to 1 as the optimization should return the highest diffusion values for these nodes. For 'vaccination refusing' nodes, the value of 0 was chosen to ensure that those nodes had the smallest possible diffusion value. For identifying a value for 'vaccination reluctant' nodes, the distribution of the three categories in the 'real-world' network were examined. The percentage of vaccination refusing and half of the percentage of the vaccination reluctant nodes were added up in order to define a mean PageRank value as an aim for a vaccination reluctant node after the optimization.

The cost function was defined as the difference between the 'real-world' values which we just defined and the diffusion values. In the optimization, the importance parameters of the influence weights were optimized. These parameters describe the relative importance of personal attributes for being central in the diffusion process of pro-vaccination sentiment.



Definition 7. Cost Function

The optimal importance parameter vector $\hat{\beta}$ is

$$\hat{\beta} = \ argmin_{\beta} \sum_{i=1}^{N} (\text{diffusion}(A(\beta))_i - \ x_{\text{real-world, } i})^2$$
$$\text{s.t. } \beta_k \geq 0,$$

with $x_{\text{real-world, } i}$ being the "real-world" vaccination status, i the index of a node, diffusion$()_i$ the final value of node i after the diffusion with adjacency matrix $A(\beta)$, and $k = 1, ..., K$, K the number of attributes.

The practical implementation was performed with BOBYQA by Powell [58][4]. Due to the fact that no single variable which is part of the influence weights should be given more weight in the beginning of the optimization in comparison to others, the same initial importance parameter value (e.g. 3) was chosen for all variables. The independence of this choice on the outcome of the optimization is discussed.

The variables which were multiplied with the importance parameters and summed up to generate the influence weights were normalized as well (by division with the median and shifting negative entries in the positive domain) to reach comparability. This is also central for ensuring transferability of the resulting importance parameters to other networks, models, and measures with the same variable constellation.

Summary Statistics

The process which is described above was repeated for all 75 networks from Karnataka and summary statistics were generated. Boxplots of the distributions of the importance parameters are shown on one scale such that direct visual comparisons are possible.

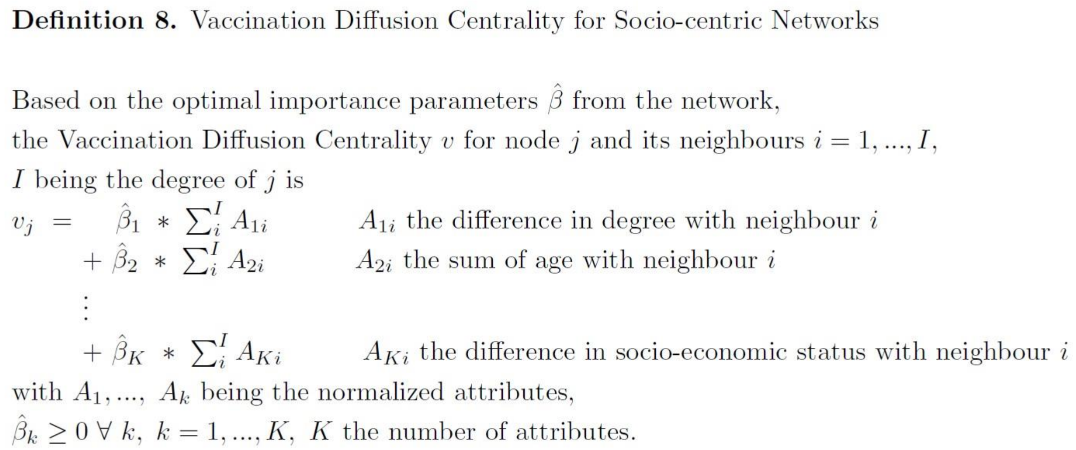
[4] For replication purposes, use the R implementation: https://www.rdocumentation.org/packages/nloptr/versions/2.0.3/topics/bobyqa

Vaccination Diffusion Centrality

To show the applicability of these findings, a centrality measure both for a socio-centric network and for single ego-centric networks was generated. It reveals how central both the edges which are adjacent to one node were and how central the node itself was in the diffusion of pro-vaccination sentiment.

Vaccination Diffusion Centrality for Socio-Centric Networks

The centrality measure for a socio-centric network was developed by making use of the network-specific importance parameters. The importance parameters were calculated and the centrality measure was developed by taking into account how important various variables are.



Definition 8. Vaccination Diffusion Centrality for Socio-centric Networks

Based on the optimal importance parameters $\hat{\beta}$ from the network,
the Vaccination Diffusion Centrality v for node j and its neighbours $i = 1, ..., I$,
I being the degree of j is

$$\begin{aligned} v_j \;=\; & \hat{\beta}_1 * \textstyle\sum_i^I A_{1i} && A_{1i} \text{ the difference in degree with neighbour } i \\ & + \hat{\beta}_2 * \textstyle\sum_i^I A_{2i} && A_{2i} \text{ the sum of age with neighbour } i \\ & \vdots \\ & + \hat{\beta}_K * \textstyle\sum_i^I A_{Ki} && A_{Ki} \text{ the difference in socio-economic status with neighbour } i \end{aligned}$$

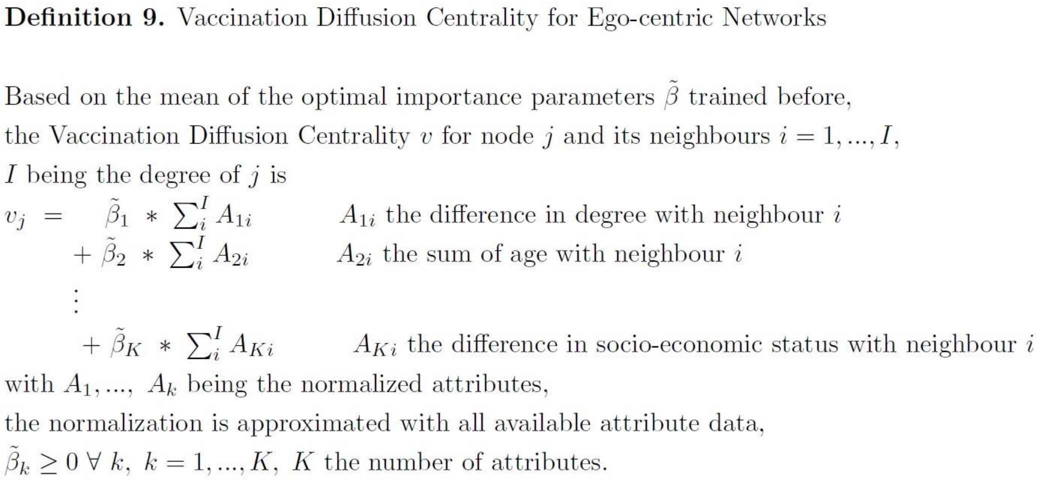
with $A_1, ..., A_k$ being the normalized attributes,
$\hat{\beta}_k \geq 0 \; \forall \; k, \; k = 1, ..., K, \; K$ the number of attributes.

For instance, the age difference importance parameter was multiplied with the age difference variable of one node with its neighbours (taking the sum of all age differences with all its neighbours; the variables were normalized by shifting entries in the positive domain and dividing by the median). This was performed for all variables in the model. Their sum resulted in a centrality measure (in the following vaccination diffusion centrality) concerning the diffusion of pro-vaccination sentiment. It was demanded before that $\beta_i \geq 0, i = 1, \dots, K \geq 0$, such that $\widehat{\beta_{\imath}} \geq 0, i = 1, \dots, K$ as well. This ensures that the centrality measure is always greater or equal to 0 and so actually a mathematical measure. This limitation was dropped later due to the fact that all diffusion vaccination centralities were greater or equal to 0 also without the limitation of the parameter space. The reasons for this and possible cases when this would not be the case need to be analysed in future research. As can be seen from analogy with the degree centrality measure, the other conditions for a centrality measure, null empty set and countable additivity, are fulfilled as well. (The introduced centrality measure only differs from the degree centrality measure in that way that the edge weights in our measure, implying the centrality of the edge in the diffusion, are summed up whereas the degree centrality measure simply counts the edges.)

The outcomes were compared with other centrality measures (Betweenness centrality, Node centrality, PageRank centrality, Eigenvector centrality, Harmonic centrality). The network was plotted and the nodes were coloured according to their vaccination diffusion centrality with a focus on central persons.

Vaccination Diffusion Centrality for Ego-Centric Network Data

Socio-centric network data are costly and most often not existent in on-the-ground health information interventions. Ego-centric networks are more often available as a data set. However, due to the fact that we did not have the full socio-centric network data, we could not calculate the network-specific importance parameters. Therefore, we relied on the analysis earlier in this work: The network-specific importance parameters were replaced with the average value (mean) of the importance parameters from all 75 villages excluding the one network which we want to use for testing this approach. The importance parameters needed to be recalculated for this case with all variables except for the Betweenness centrality which needed to be excluded because it cannot be calculated on the basis of some parts of the network.



Definition 9. Vaccination Diffusion Centrality for Ego-centric Networks

Based on the mean of the optimal importance parameters $\tilde{\beta}$ trained before,
the Vaccination Diffusion Centrality v for node j and its neighbours $i = 1, ..., I$,
I being the degree of j is

$$\begin{aligned} v_j \;=\; & \tilde{\beta}_1 \, * \, \textstyle\sum_i^I A_{1i} && A_{1i} \text{ the difference in degree with neighbour } i \\ & + \tilde{\beta}_2 \, * \, \textstyle\sum_i^I A_{2i} && A_{2i} \text{ the sum of age with neighbour } i \\ & \vdots \\ & + \tilde{\beta}_K \, * \, \textstyle\sum_i^I A_{Ki} && A_{Ki} \text{ the difference in socio-economic status with neighbour } i \end{aligned}$$

with $A_1, ..., \; A_k$ being the normalized attributes,
the normalization is approximated with all available attribute data,
$\tilde{\beta}_k \geq 0 \; \forall \; k, \; k = 1, ..., K, \; K$ the number of attributes.

One other approximation had to be included: The variables (degree difference, age sum etc.) had to be normalized by dividing by the median. Due to the lack of full network data, we needed to approximate median of the respective variables by taking the median of all available nodes (the medians of the variables of all egos of the ego-centric networks, potentially also including personal information from alters). As in the case of the socio-centric measure, when the variables included negative values, all entries in this variable were shifted in the positive domain by adding the negative minimum value to all entries.

With these two approximations, a locally applicable vaccination diffusion centrality could be generated. Due to the fact that this local version of the diffusion centrality has the advantage that only a few ego-networks need to be collected in order for it to work, a test with a low number of ego networks ($N = 15$) which were chosen by a random function was performed.

Results

This section is subdivided into two subsections: In the first section, we look at side results about how the networks were generated and simulated. In the second section, we look at normalization and optimization results and finally at the resulting importance parameters. Afterwards, the two derived centrality measures for both socio-centric and ego-centric network data are tested and finally compared with established centrality measures.

Network Generation

The first dataset contained 75 villages in Karnataka, India with an average of 198.7 households per village. The average number of contacts is 2.1 per household.

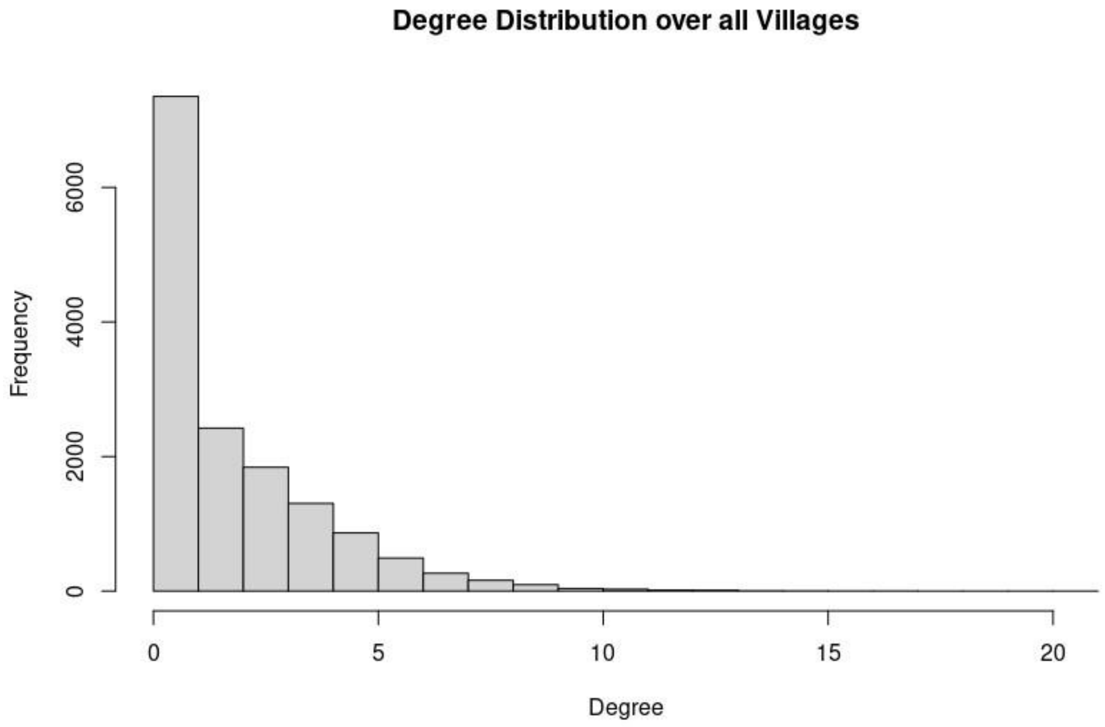


Figure 2: Degree Distribution over all 75 villages

On average, each node's contacts have 4.3 contacts. The average time which an individual has already lived in the village is 25.8 years. This distribution seems to follow the power law. The Erdos-Rényi e.g. model would not have produced such a distribution. Scale-free degree distributions are of course central to the Barabási-Albert model [11], but the interdependences of these distributions with other personal, tie and network attributes is difficult to model with random network generation approaches. The gender and age distributions of the respondents over all households in all networks were displayed here.

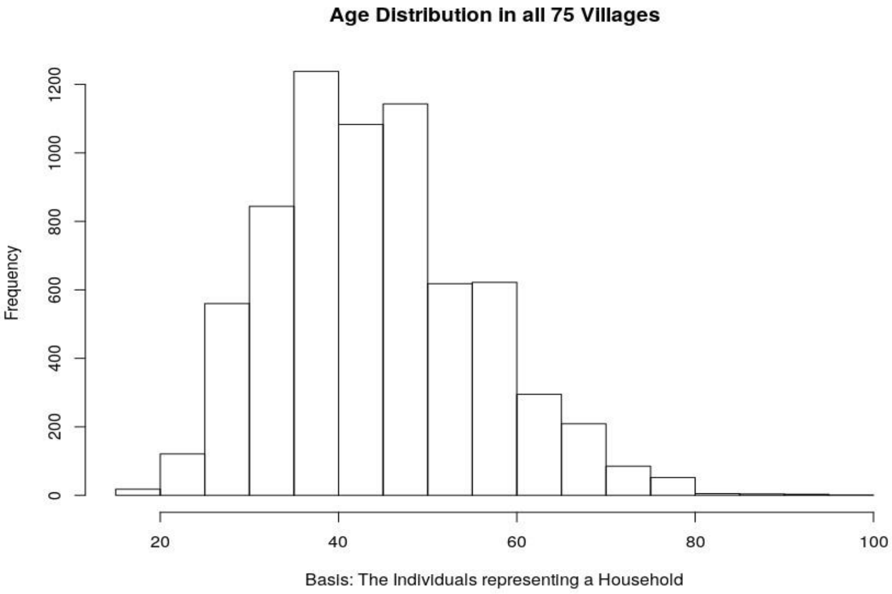


Figure 3: The Distribution of Age of the chosen individuals in all 75 villages

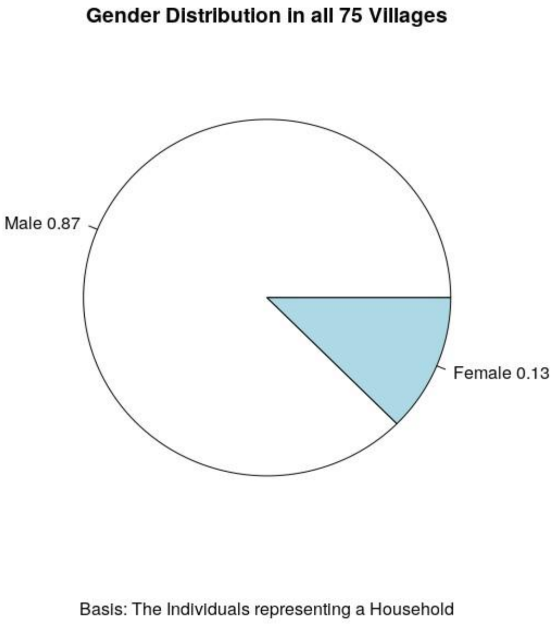


Figure 4: The Gender Distribution of the chosen individuals in all 75 villages. Missing gender information were simulated in the next step.

In the next step, the occupations and basic needs like a bank account and electricity were looked at.

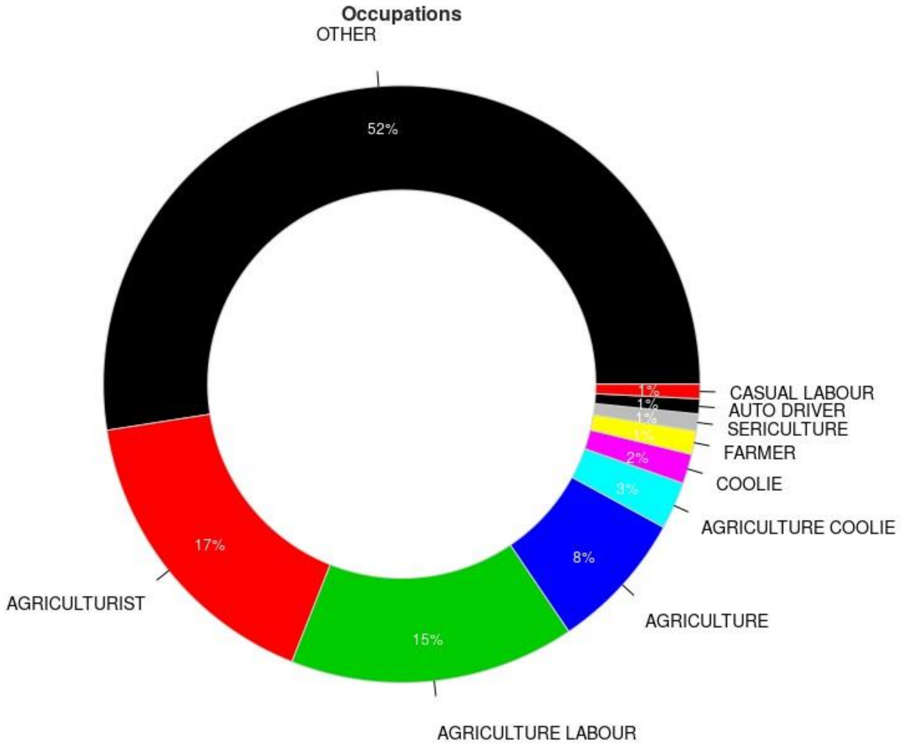


Figure 5: All Occupations from the analysed Individuals in 19404 Villages

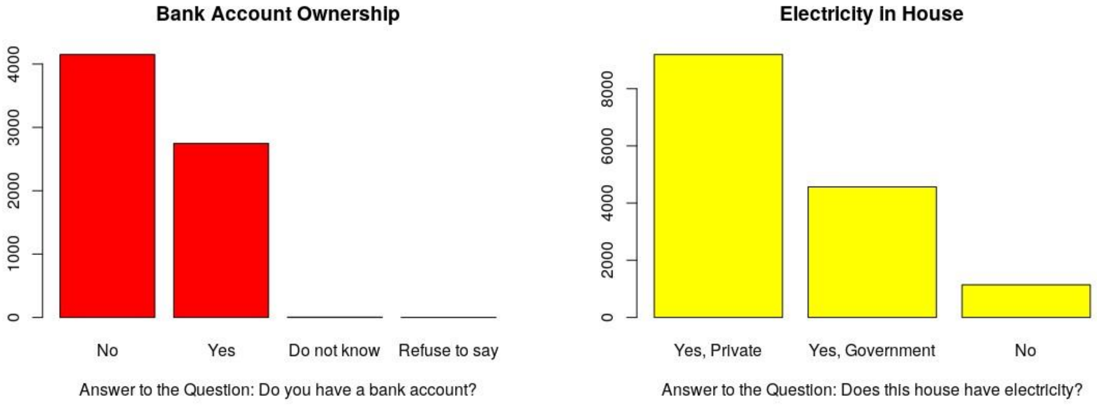


Figure 6: Bank Account Ownership and Electricity in the respondents' Houses.

Most houses have two rooms and 73.4 % of the houses have neither a common or an own latrine.

Most respondents have Kannada as a mother tongue and do not speak English:

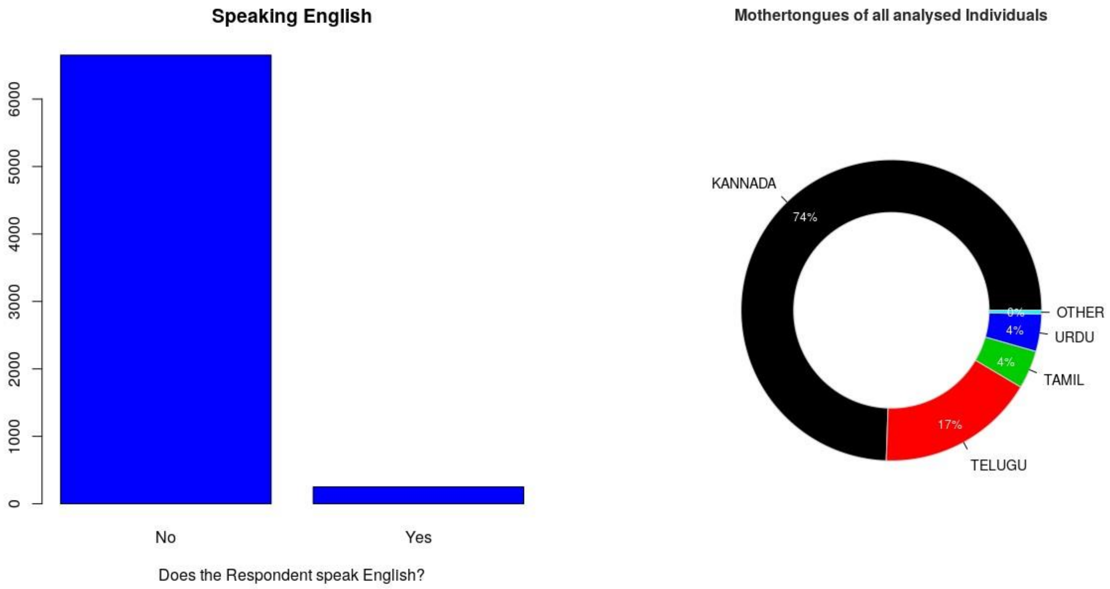


Figure 7: English Ability of Respondent and Mother Tongue

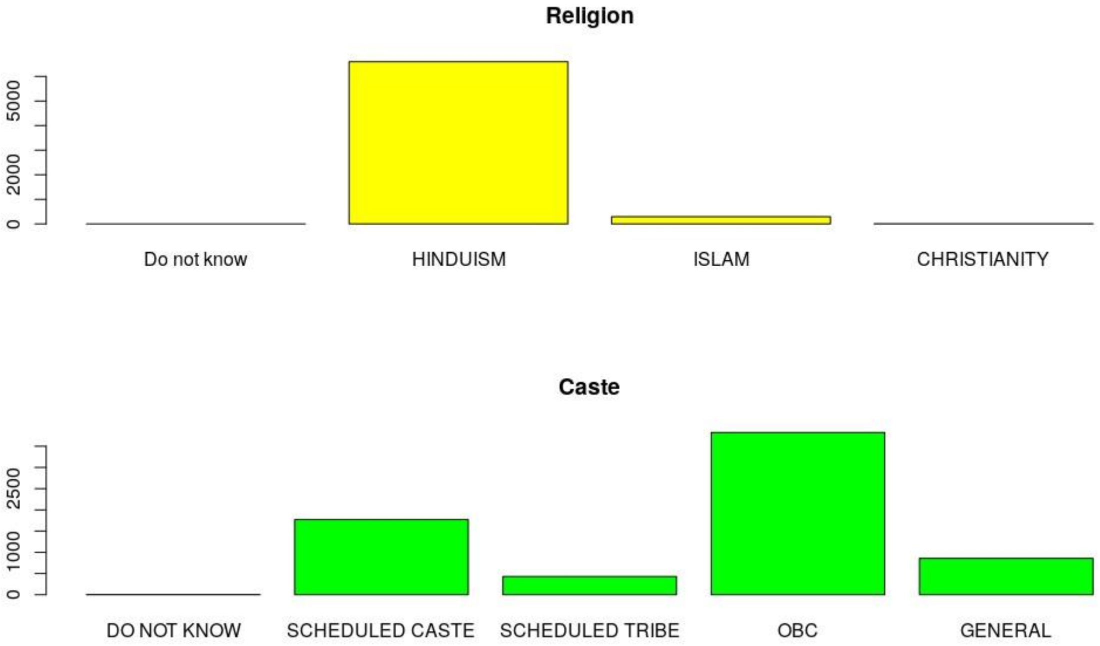


Figure 8: Religion and Caste of all 14904 included Individuals. OBC refers to Other Backward Classes, meaning that this group has been classified as socially and educationally backward in 1991.

The information about the households in Karnataka were expanded by adding the information of one household representative, if there were any information available. Over all 14904 households in all 75 networks, in 95.6 % of the cases, the household head was chosen. The spouse of the household head was chosen in 3.6 % of the cases. Other members were chosen in 0.8 % of the cases.

If there were no individuals interviews in one household, the missing data were simulated based on the household information. The random forest prediction both for classification and for regression models produced the following performance measures over all 75 networks.

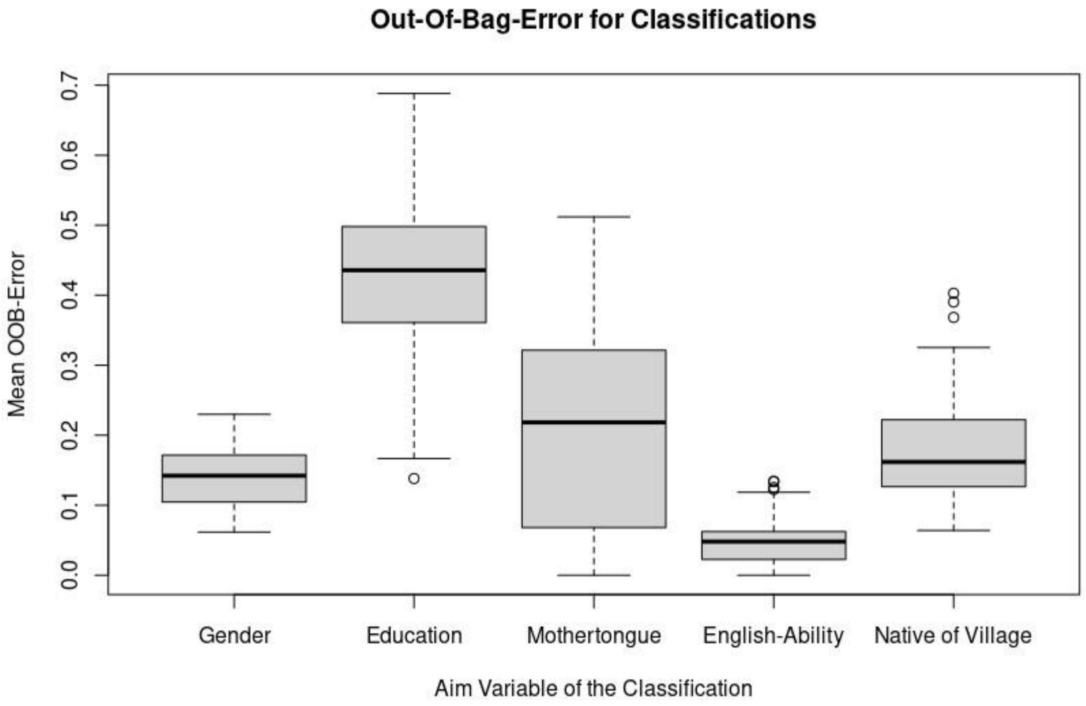


Figure 9: The Mean OOB-Error when testing the model for a certain aim variable in one network is calculated. The distribution of these values over al 75 networks is shown in this plot.

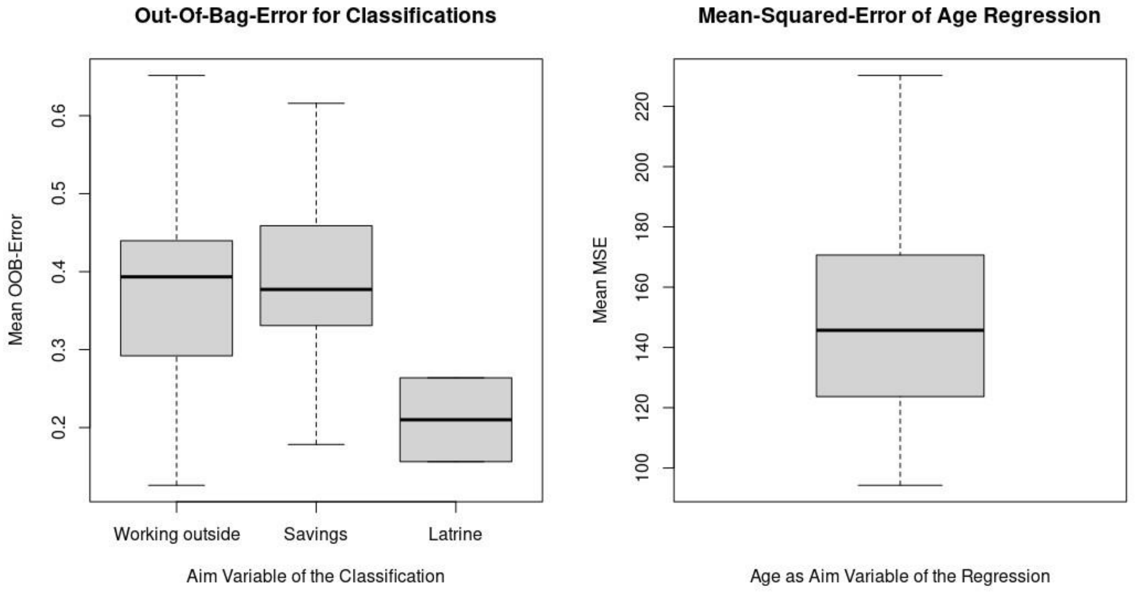


Figure 10: For the explanation of the left plot see the figure above. The right plot shows the distributions of the mean MSE over all 75 networks when creating a regression model for the variable “Age”.

The training performance of the model was comparatively better for the ability to speak English and for the estimation of the gender which can also be attributed to the fact that most respondents were men and there were only few who could speak English, meaning that the variances of these attributes were not very large. It was comparatively worse for education, working outside and savings which is due to a comparatively larger variance.

In the following, the success of the matching is investigated. As the underlying network structure from the chosen Karnataka network is the base of the generated network and hence the Karnataka ego characteristics are the same in the matched network, only the ego variables from Malegaon and from the matched network are shown. Each Karnataka data point draws the best fit from Malegaon (with replacement). Hence, the distributions of the resulting variables from the matched Network are subsamples from the Malegaon distribution.

Over all networks, 334 different nodes from the 706 nodes Malegaon were chosen as the minimal distance match. This is due to distribution differences, for instance in one matching variable, the number of rooms: Karnataka households had most of the times 2 and 3 rooms, Malegaon households 1. So, nodes with only 1 room in Malegaon were comparatively less frequently chosen than with 2 or 3 rooms, even when the variable was normalized using the relative position normalization defined above. Comparatively, the variance in the number of rooms in Karnataka is higher than in Malegaon.

So, nodes from Malegaon with more rooms were more likely to have a minimal distance to households in Karnataka than a node with less rooms.

On average, one of the nodes from Malegaon was matched 44.6 times with a node from Karnataka over all networks. The node from Malegaon which was matched most often times was node 39 with 2 043 out of the total 14 904 matches. This node's vaccination status is that there were no vaccines eligible. 4 persons live in one room in this household and no latrine is available. The household has a TV and a phone, but no cooking cylinder. The education status is either no education or a primary school degree.

To measure the overall success of the matching, the differences of all variables between all 14 904 households in the dataset from Karnataka with their best fit households in the Malegaon data set was analysed.

The difference in the number of rooms was on average the highest, followed by the degree difference. The smallest average distance was measured in the 'triads' variable. Categorical differences in education level and the existence of a latrine in the household occurred rarely.
The distributions are visualized in corresponding boxplots:

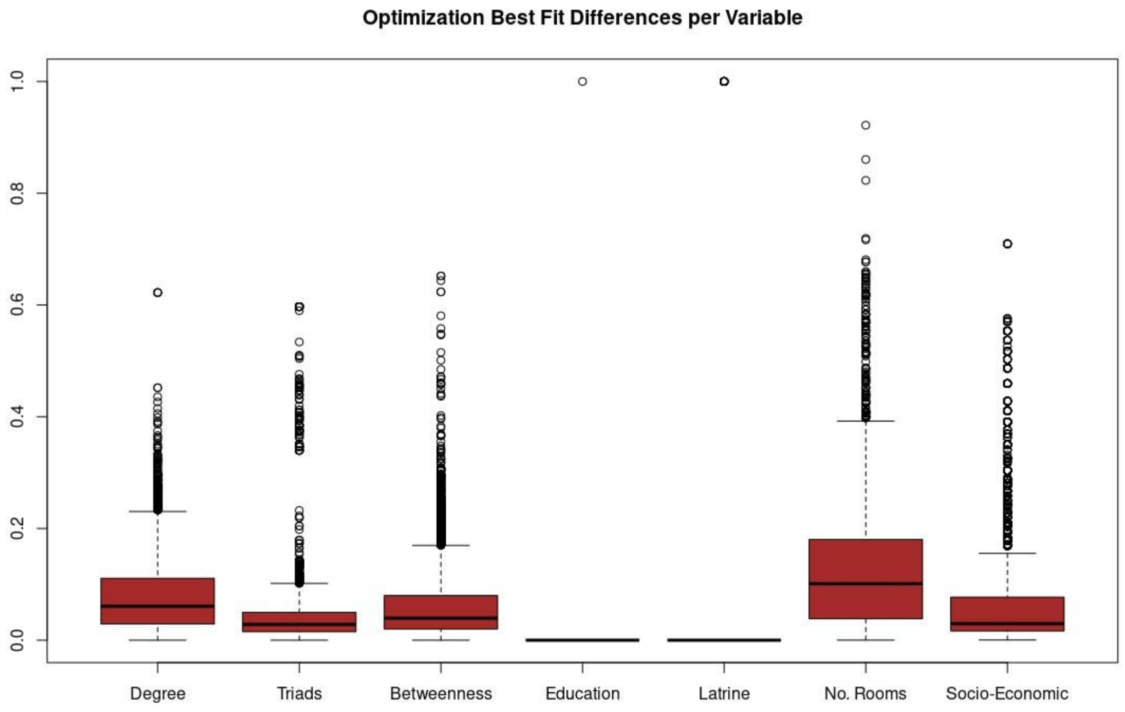


Figure 11: Differences of all variables between all 14904 households in the dataset from Karnataka with their best fit households in the Malegaon data set. All variables are normalized according to their rank position except for the Education and Latrine Variable

In the following, it will be shown which vaccination status distribution and other distributions result from this choice for one of the final matched networks which was selected randomly.

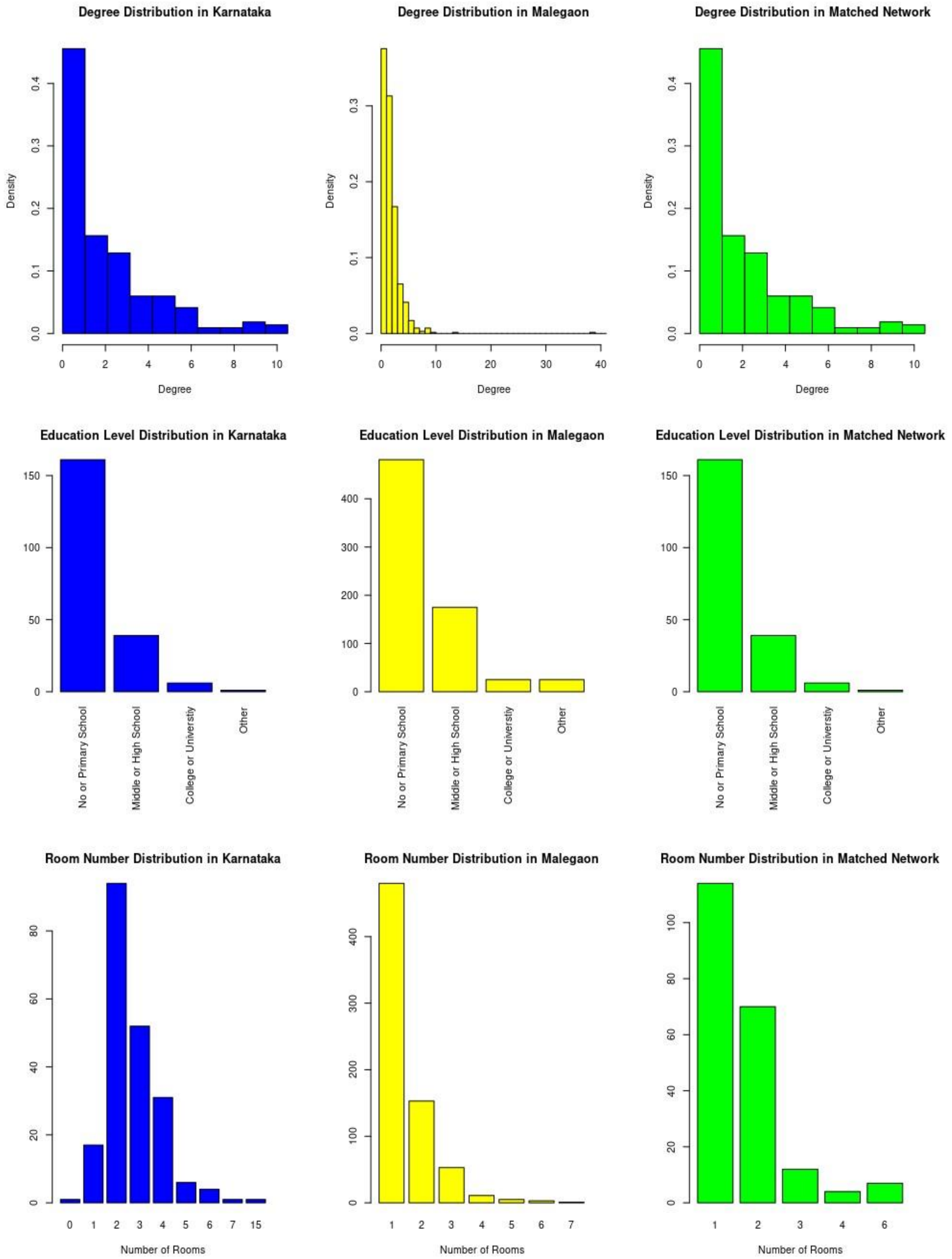


Figure 12: The blue plots are distributions from Karnataka, the yellow from Malegaon and the green from the Matched Network. The Matched Network shows similar attributes as real-world networks.

The degree distributions from the matched and the Malegaon network differ. The difference between the average degree of these two distributions can be attributed to the fact that in Malegaon, each questioned individual could name up to 4 contacts. This limit of number of mentions was not given in Karnataka.

The education levels of the Matched Network are similar to the Karnataka levels. The number of rooms distribution of the Matched Network is a mixture from Malegaon and Karnataka.

Concerning the vaccination status, a decrease of the “No vaccine eligibles” and “refusing” households was found in the matched network in comparison with the original network from Malegaon. The number of “reluctant” nodes increased comparatively.

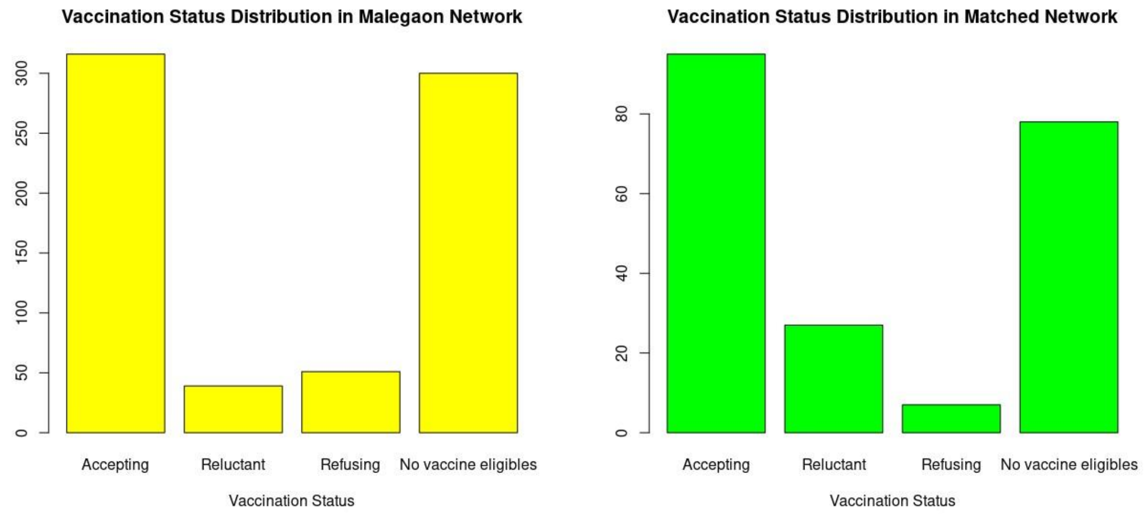


Figure 13: The Vaccination Status Distributions form Malegaon and from the Matched Network.

Based on the described matching procedure, an example resulting graph is shown below with the node colour referring to the vaccination status.

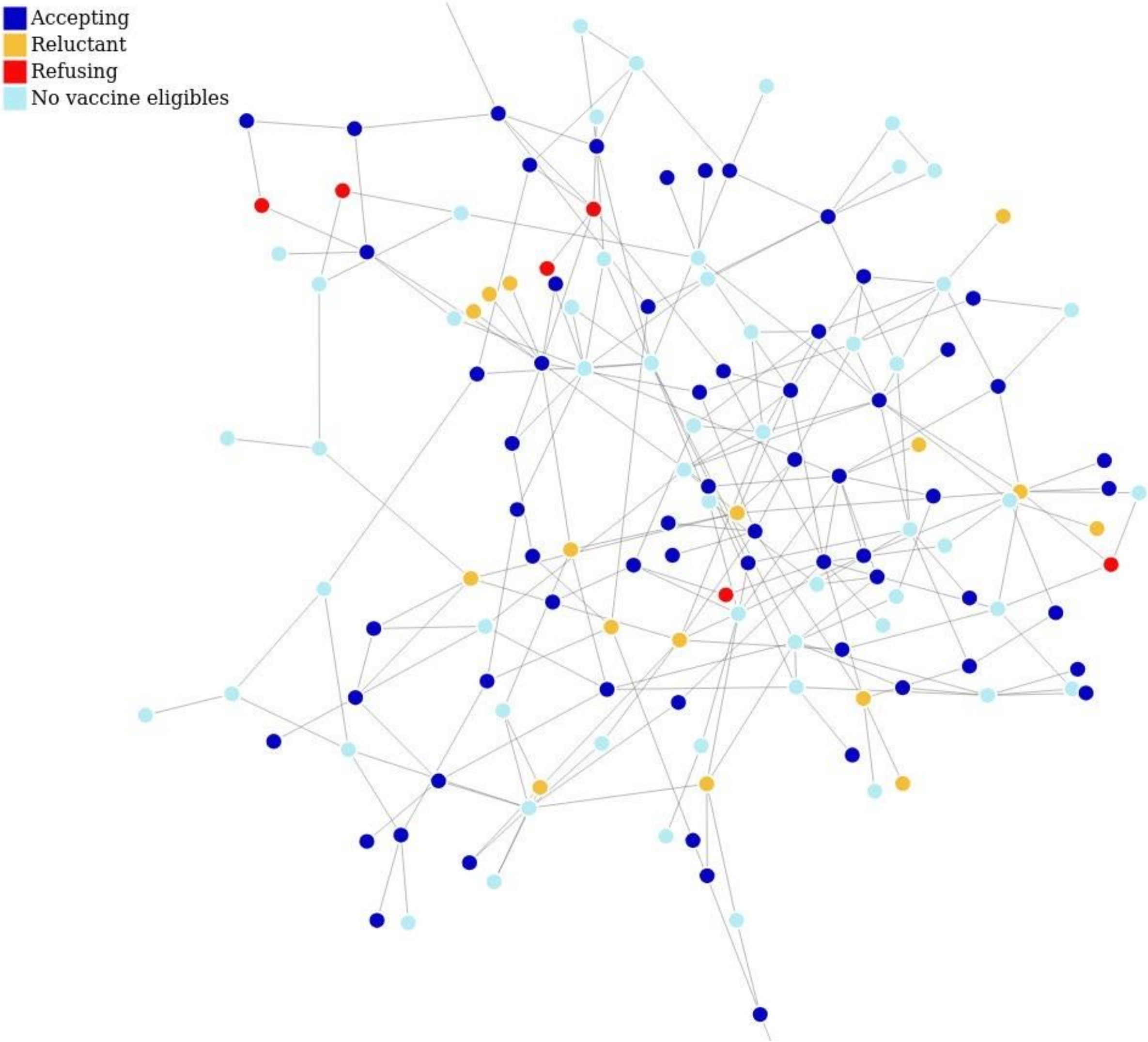


Figure 14: A visualisation of one resulting village network

Diffusion, Optimization and Centrality Measure

This section describes results of the influence weight matrices generation. It is shown how they were normalized to reach transferability and comparability and that their sum forms the adjacency matrix for the network with influence edge weights. The Optimization subsection investigates the stability of the optimization results. The final importance parameters are shown in Summary Statistics. Afterwards, the resulting centrality measures are described.

Influence Weights

The normalizations of the attribute matrices before including them in the diffusion model result in the following distributions of all matrix entries which show that the normalization leads to attribute values which are comparable. The data are taken from a sample network.

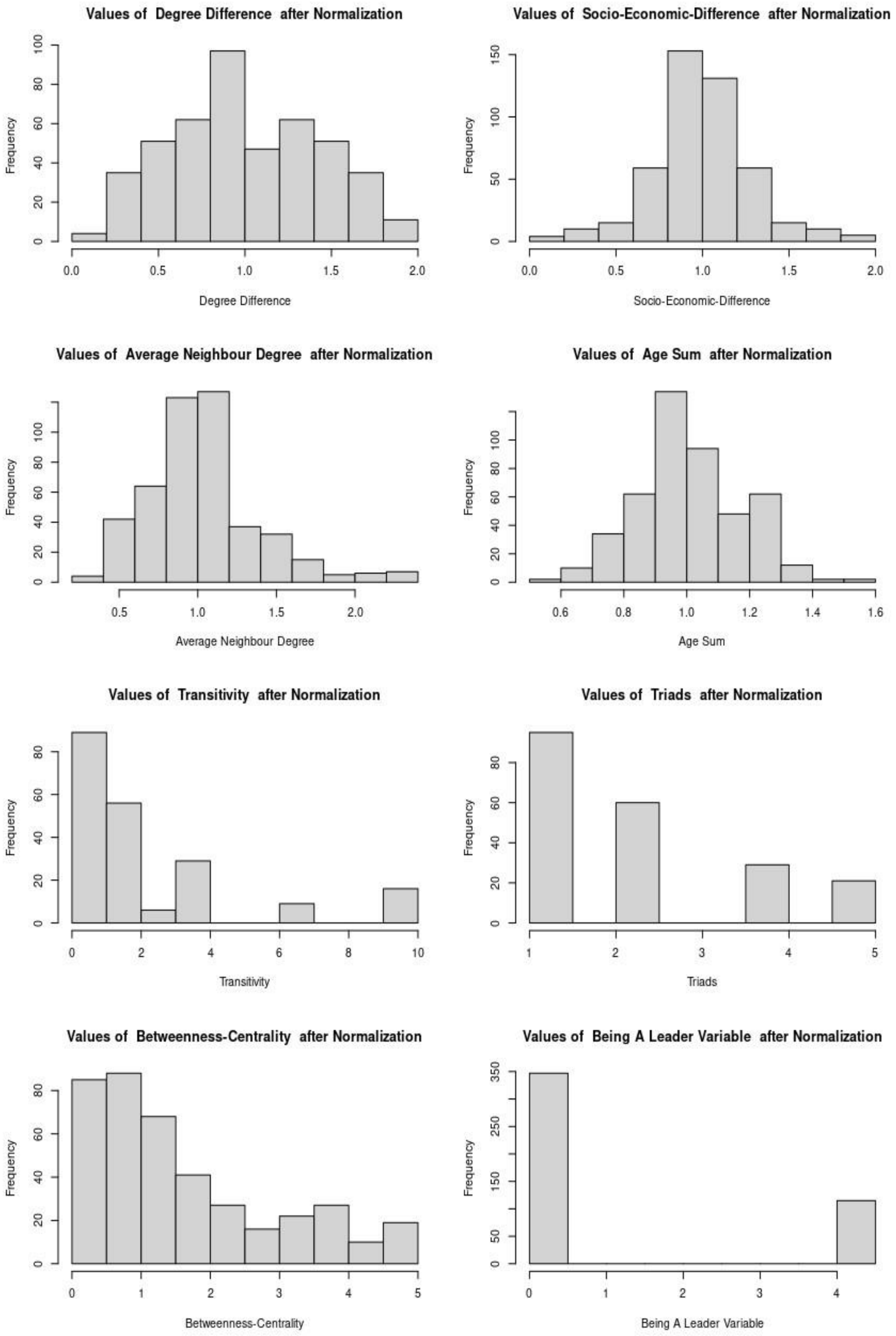


Figure 15: Distributions of attributes after normalization

So, the variables are comparable to each other and their respective importance parameters are directly comparable as well.

All attribute matrices were multiplied with their respective importance parameters and added up to the adjacency matrix which was used in the diffusion.

Optimization

The outcome of the optimization changed with a different initial value of the importance parameters. This was explored further. As described earlier, the same initial importance parameter value was chosen for all variables to avoid biased initial conditions.

The following hypothesis was tested on a randomly chosen subset of all villages (N=20): Multiplying the initial values of the importance parameters with a factor changes the scale, but not the relation between the final importance parameters.

Due to the normalization of the columns of the adjacency matrix which is performed by the diffusion algorithm pageRank, the columns' entries can be multiplied by any positive factor (of the real numbers). The normalization by pageRank would revert a prior multiplication of columns with a factor. Hence, also the importance parameters could be multiplied by this arbitrary factor before being multiplied with their attribute matrix. PageRank would revert this as well. Nevertheless, multiplying the initial importance parameters by a factor changes the baseline of the comparison as each parameters starts not at 1, but e.g. at 3 or 10.

So, the factor does not change the relations between the final importance parameters, which contain Information about the relative importance of their respective variables.
The tests of the hypothesis had the following results:

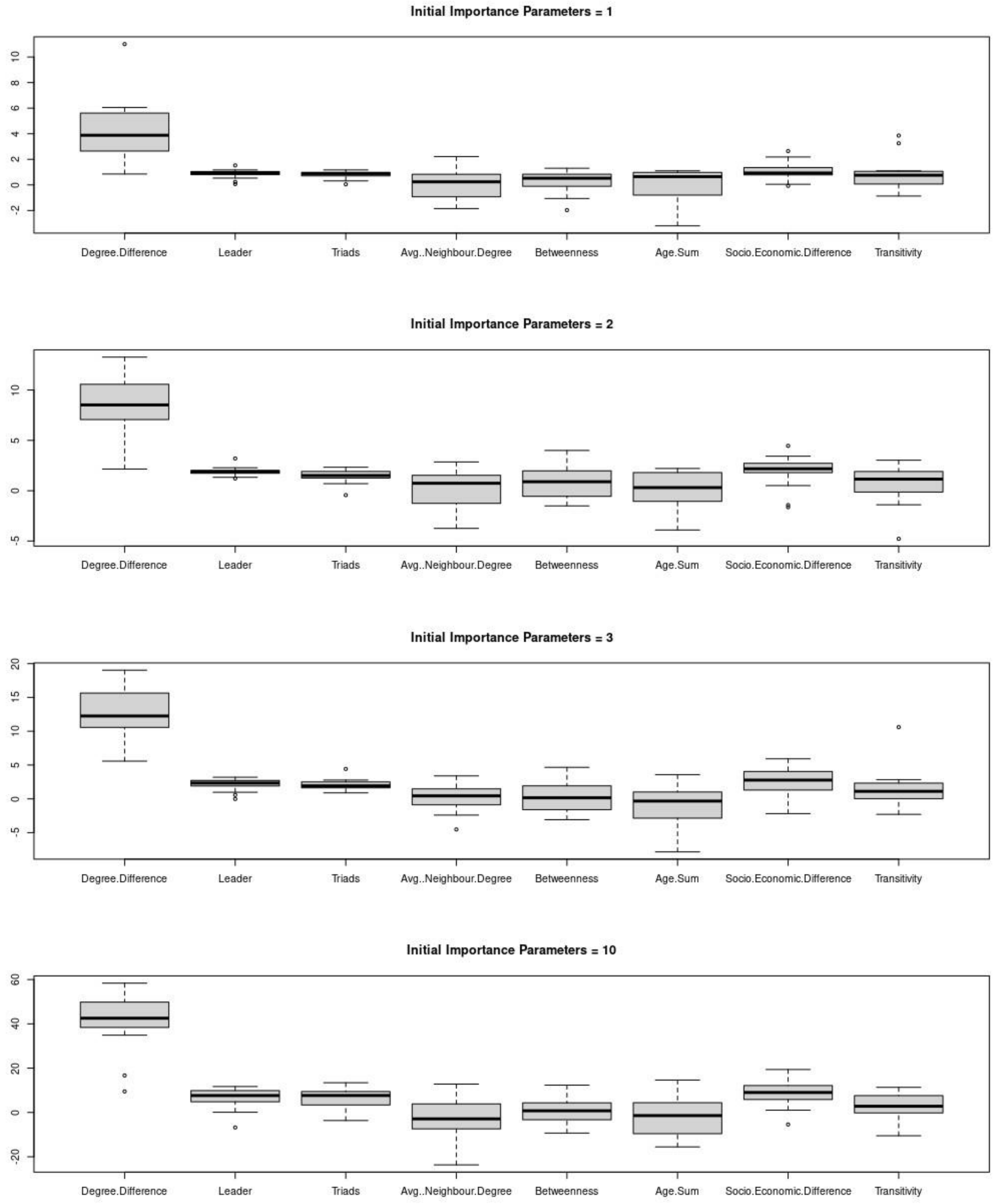


Figure 16: Optimization Outcomes due to various initial importance parameters

Despite some variability which is caused by varying optimization outcomes, general trends can be detected which can be interpreted as supportive of the hypothesis above. For the later analyses, the importance parameters were initialized with the value 3.

Further investigations of the topic like exploring the relations of the final importance parameters under the condition of another optimization algorithm can be subject to future research.

Summary Statistics

The resulting importance parameter describe the relative importance of the explanatory variables in the diffusion process of pro-vaccination sentiment. The importance parameters found with the generated networks in this work are depicted below by showing the distributions of these parameters over all 75 networks.

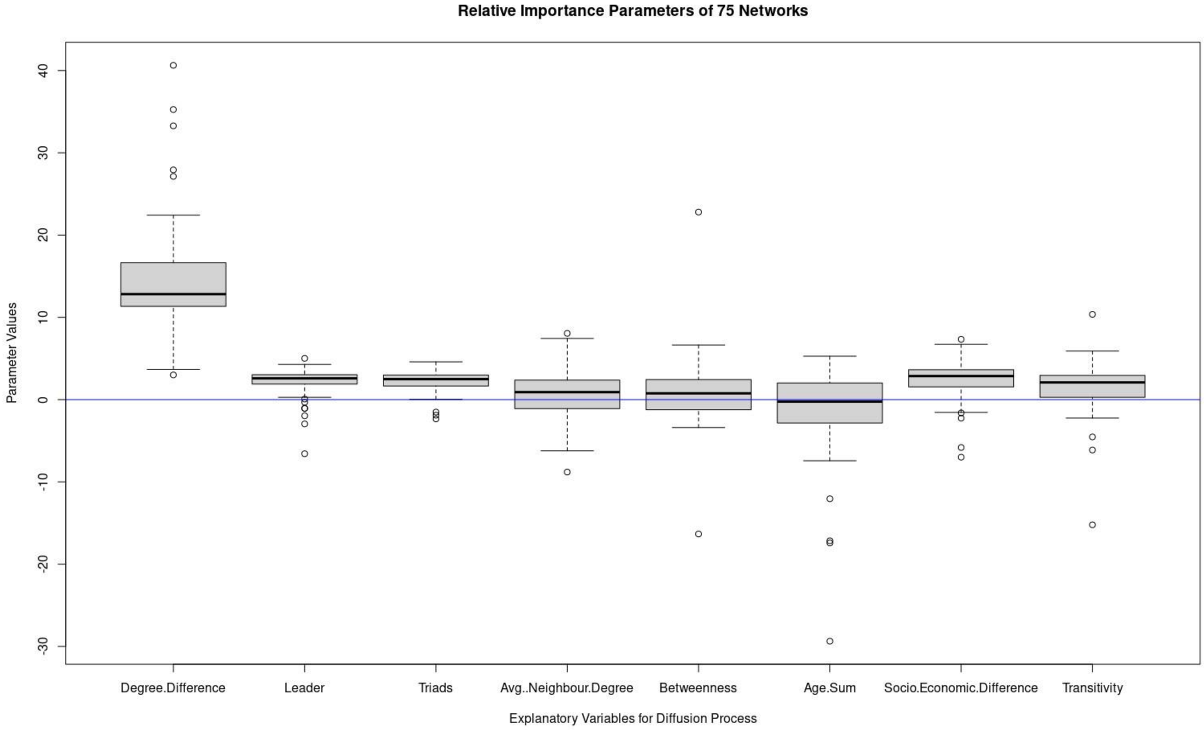


Figure 17: The distributions of relative Importance Parameters over 75 analysed networks

A high degree difference importance parameter suggests that the degree difference was a central variable in explaining the flow of pro-vaccination sentiment through edges. Nodes with a high average of degree differences with their neighbours are adjacent to many rather central edges concerning the flow of pro-vaccination sentiment. Due to the fact that the flow through adjacent edges also goes through the node itself, they are as well central in the flow of pro-vaccination sentiment.

A socio-economic difference is also important for explaining the flow of pro-vaccination sentiment as well as the fact that the flow emitting node is a community leader. The other network attributes (average neighbour degree, betweenness centrality, transitivity, triads) were also important for explaining the diffusion. The age sum importance parameter ranks as the lowest parameter. It can be concluded that the age sum is not as central to the flow of pro-vaccination sentiment as other variables and not an adequate variable to explain how the diffusion has spread. The age sum is relatively less important for the diffusion process.

Vaccination Diffusion Centrality

Vaccination Diffusion Centrality for Socio-Centric Networks

The optimized importance parameter values in a network can be used for designing a centrality measure concerning the diffusion of pro-vaccination sentiment (in the following vaccination diffusion centrality).

The vaccination diffusion centrality was calculated for one sample network of the previously generated 75 networks. The optimized importance parameters for the network were multiplied with the respective explanatory variables (age sum, degree difference...) for each node. Summing up these values returns the targeted centrality measures for each node. The measure is a value for each node which indicates the centrality of the node’s adjacent edges in the diffusion of pro-vaccination sentiment and therefore the centrality of the node in the diffusion.

This was calculated for all nodes. The centrality measure could be efficiently computed by summing up the rows of the final adjacency matrix based on the final importance parameters after the optimization.

The distribution of the vaccination diffusion centrality measure follows the power law which is similar to comparable centrality measures [2] [30], as shown below.

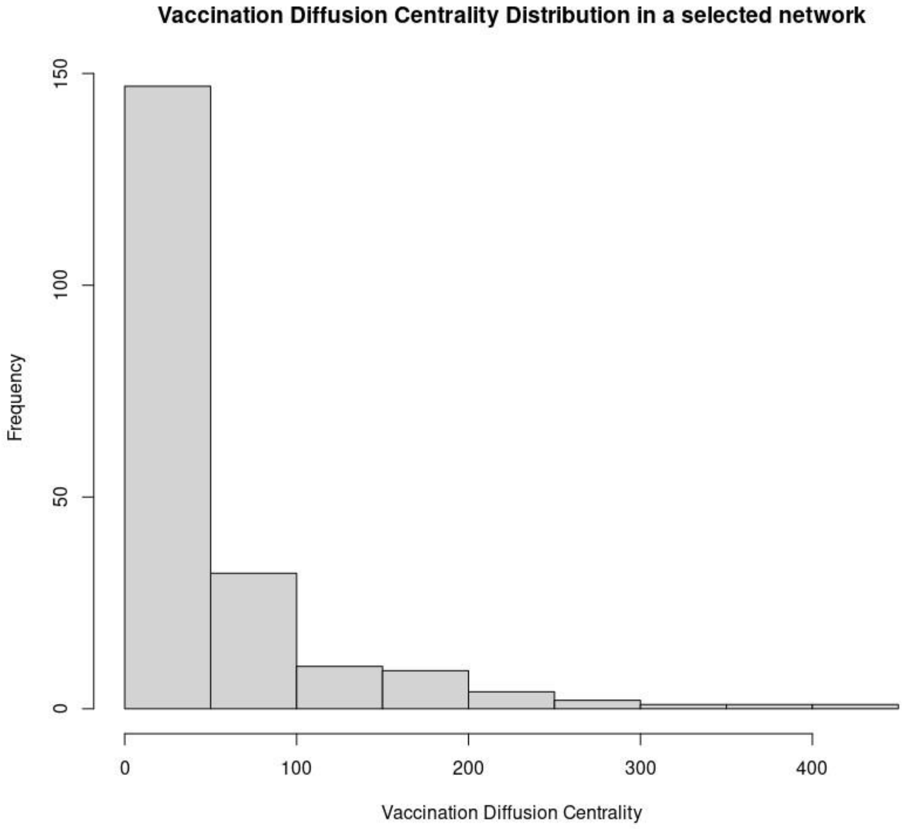


Figure 18: The Distribution of Vaccination Diffusion Centrality over all nodes of a network

The distribution of vaccination diffusion centrality was visualized in a complete network plot:

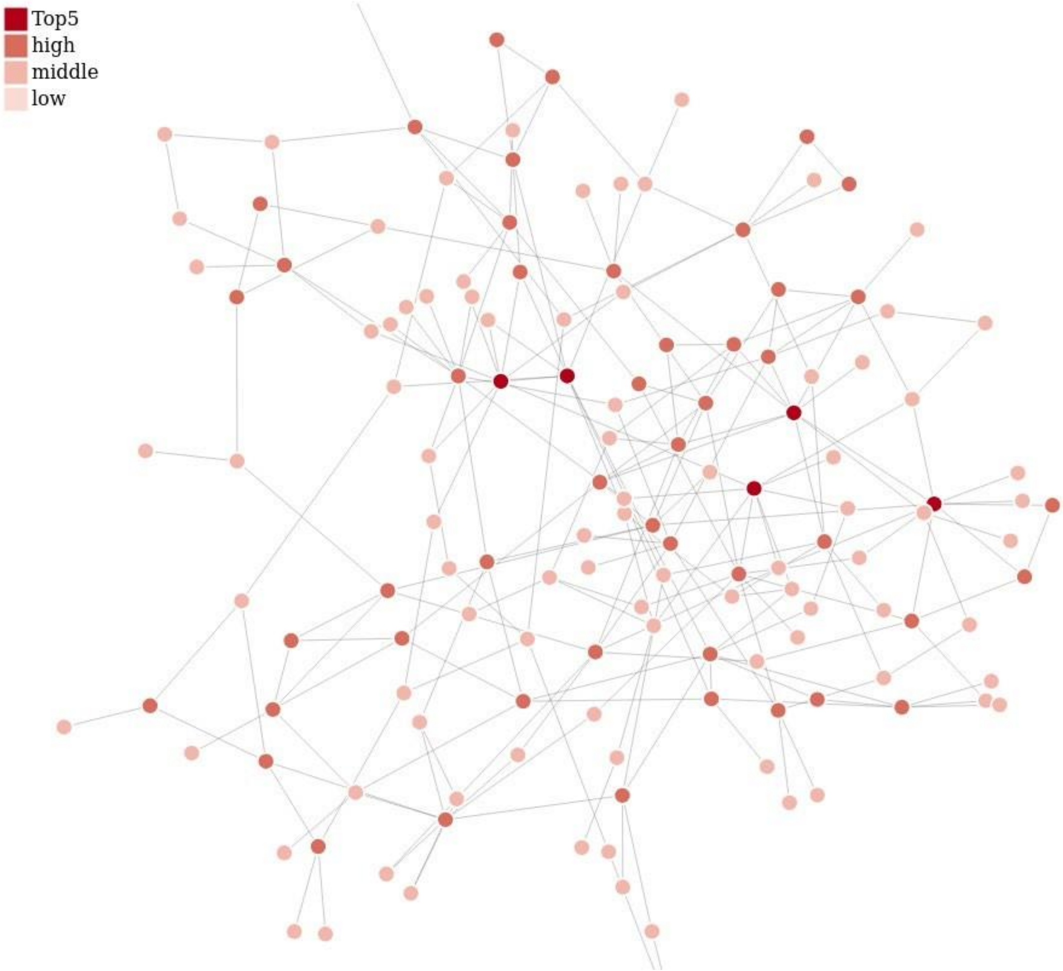


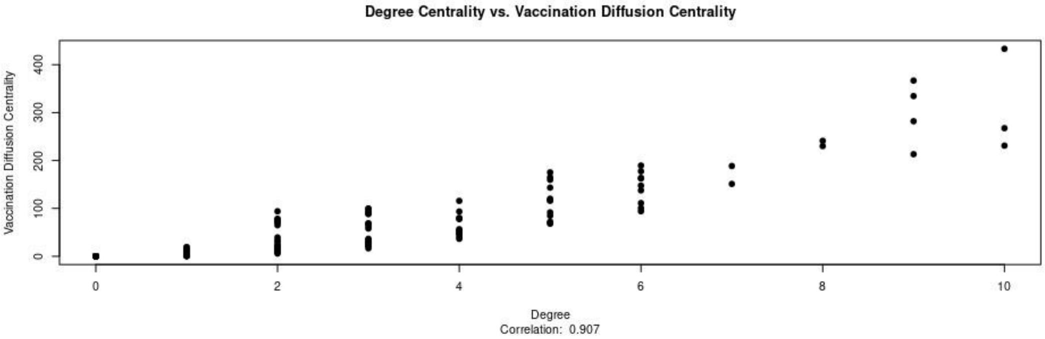
Figure 19: The Top 5 nodes are shown in dark red, the nodes in the "high" group belong to the upper quartile (75%-100%), the middle group to the 2nd and 3rd quartile and the "low" nodes to the 0% - 25% quartile.

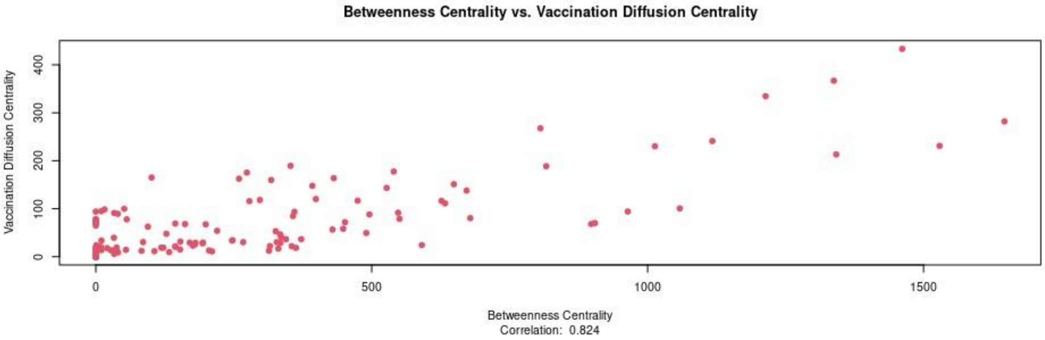
The values were compared with standard centrality measures (betweenness-centrality, degree centrality, pageRank centrality, harmonic centrality, eigenvector centrality).

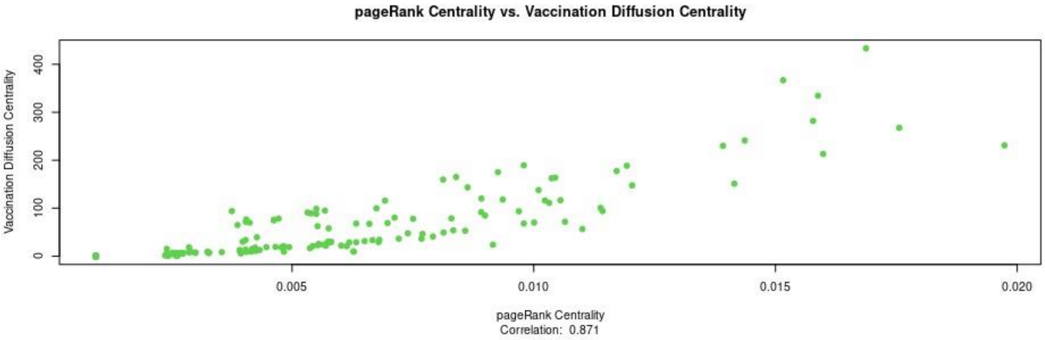
The harmonic centrality by Marchiori and Latora [50] is similar to the closeness centrality, for a node, it is the mean inverse distance to all other nodes with the difference to the closeness centrality that disconnected nodes receive a distance value of 0.

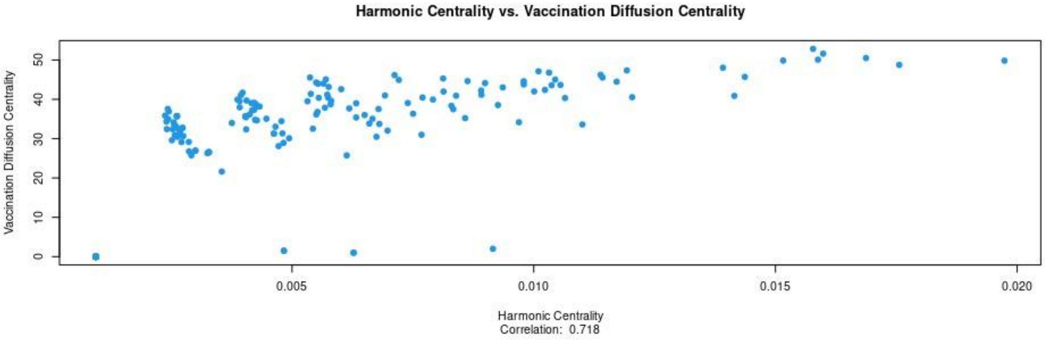
The eigenvalue centrality of Bonacich [13] takes the first eigenvector of the adjacency matrix as the distribution of centrality over all nodes. As the first eigenvector can be interpreted as the largest extent of how far the matrix function can change vectors, this vector comprises many key properties of the adjacency matrix.

It can be deduced from the plots below that the vaccination diffusion centrality results are in the range of established centrality measures.


Degree Centrality vs. Vaccination Diffusion Centrality
Vaccination Diffusion Centrality
0
100
200
300
400
0
2
4
6
8
10
Degree
Correlation: 0.907


Betweenness Centrality vs. Vaccination Diffusion Centrality
Vaccination Diffusion Centrality
0
100
200
300
400
0
500
1000
1500
Betweenness Centrality
Correlation: 0.824


pageRank Centrality vs. Vaccination Diffusion Centrality
Vaccination Diffusion Centrality
0
100
200
300
400
0.005
0.010
0.015
0.020
pageRank Centrality
Correlation: 0.871


Harmonic Centrality vs. Vaccination Diffusion Centrality
Vaccination Diffusion Centrality
0
10
20
30
40
50
0.005
0.010
0.015
0.020
Harmonic Centrality
Correlation: 0.718

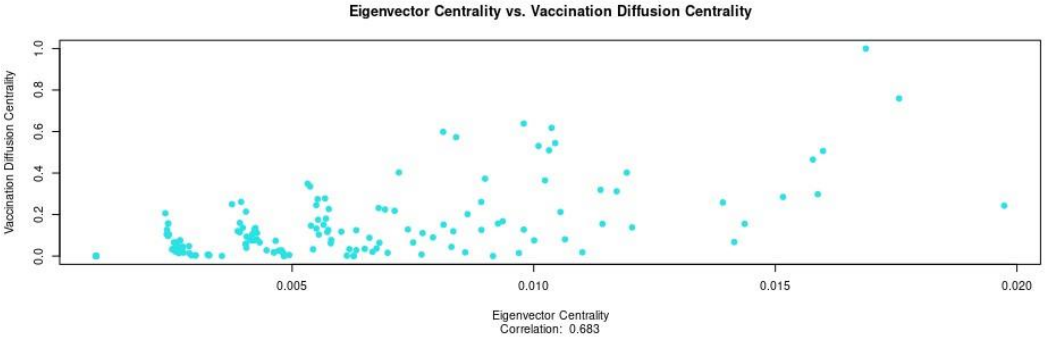

Eigenvector Centrality vs. Vaccination Diffusion Centrality
Vaccination Diffusion Centrality
0.0
0.2
0.4
0.6
0.8
1.0
0.005
0.010
0.015
0.020
Eigenvector Centrality
Correlation: 0.683

Figure 20: The Vaccination Diffusion Centrality versus established centrality measures

This result is not surprising due to the fact that the betweenness centrality is a summand of the vaccination diffusion centrality and on average makes up 11.6% of the vaccination diffusion centrality. The degree centrality is indirectly included as part of the degree difference.

Vaccination Diffusion Centrality for Ego-Centric Network Data

Full socio-centric network data are cost-intensive and oftentimes not available in standard circumstances of health information interventions. So, a local vaccination diffusion centrality measure was developed for the case that only a few ego-centric networks are available. It was calculated with the means of the importance parameters from 74 villages (one village was excluded which was used for later testing) and with approximated medians and minima of the explanatory variables for the normalization.

To test the local vaccination diffusion centrality, the model needed to be altered due to the existence of the betweenness centrality as an explanatory variable which needs a full socio-centric network and cannot be calculated on the basis of a few ego networks. Hence, the variable was excluded and the importance parameters were recalculated. The resulting parameters from 74 villages are shown below.

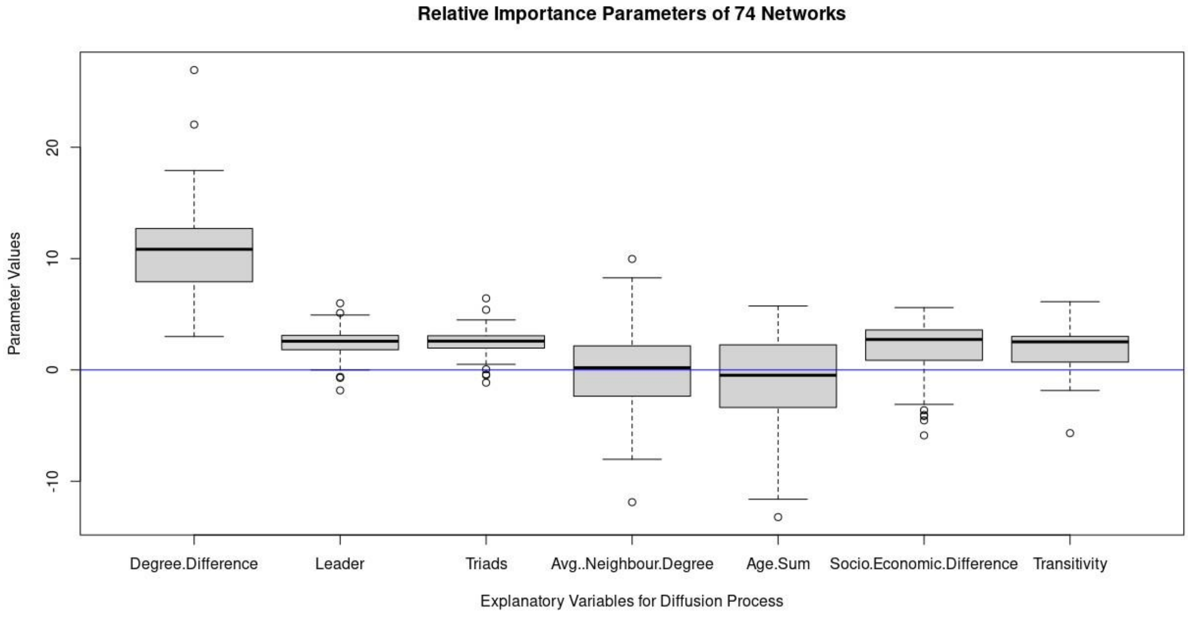


Figure 21: Relative Importance Parameters of 74 networks without Betweenness centrality variable

On the basis of the mean importance parameters, the local centrality measure with ego and tie variables and their respective importance parameters and approximated medians for normalization was derived as described earlier (multiplying the importance parameters with the variables and finally summing everything up).

As an example, 15 nodes from the unseen 75th network were extracted. The ego networks were taken as a basis for calculating their vaccination diffusion centrality. The choice of the number of ego networks for this example is arbitrary and future research has to clarify the sensitivity of the measure when changing the number of nodes potentially with the goal of finding a recommended minimum number of nodes to ensure a certain level of quality. The ego networks' centrality values are depicted below.

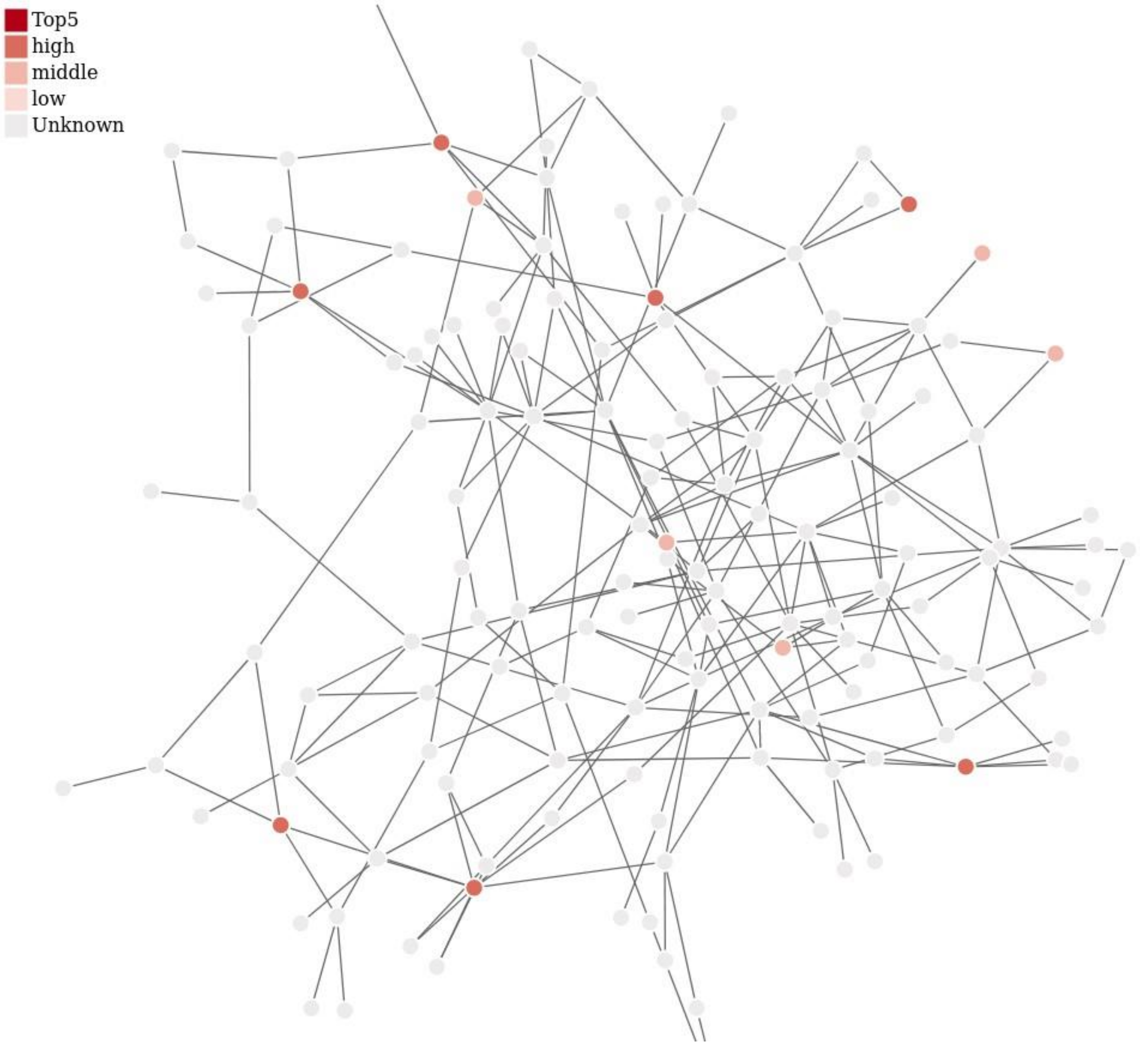


Figure 22: The centrality measure can also be calculated (as an approximation) when only ego network data are available.

The local vaccination diffusion centrality was also found to lie within the range of standard centrality measures.

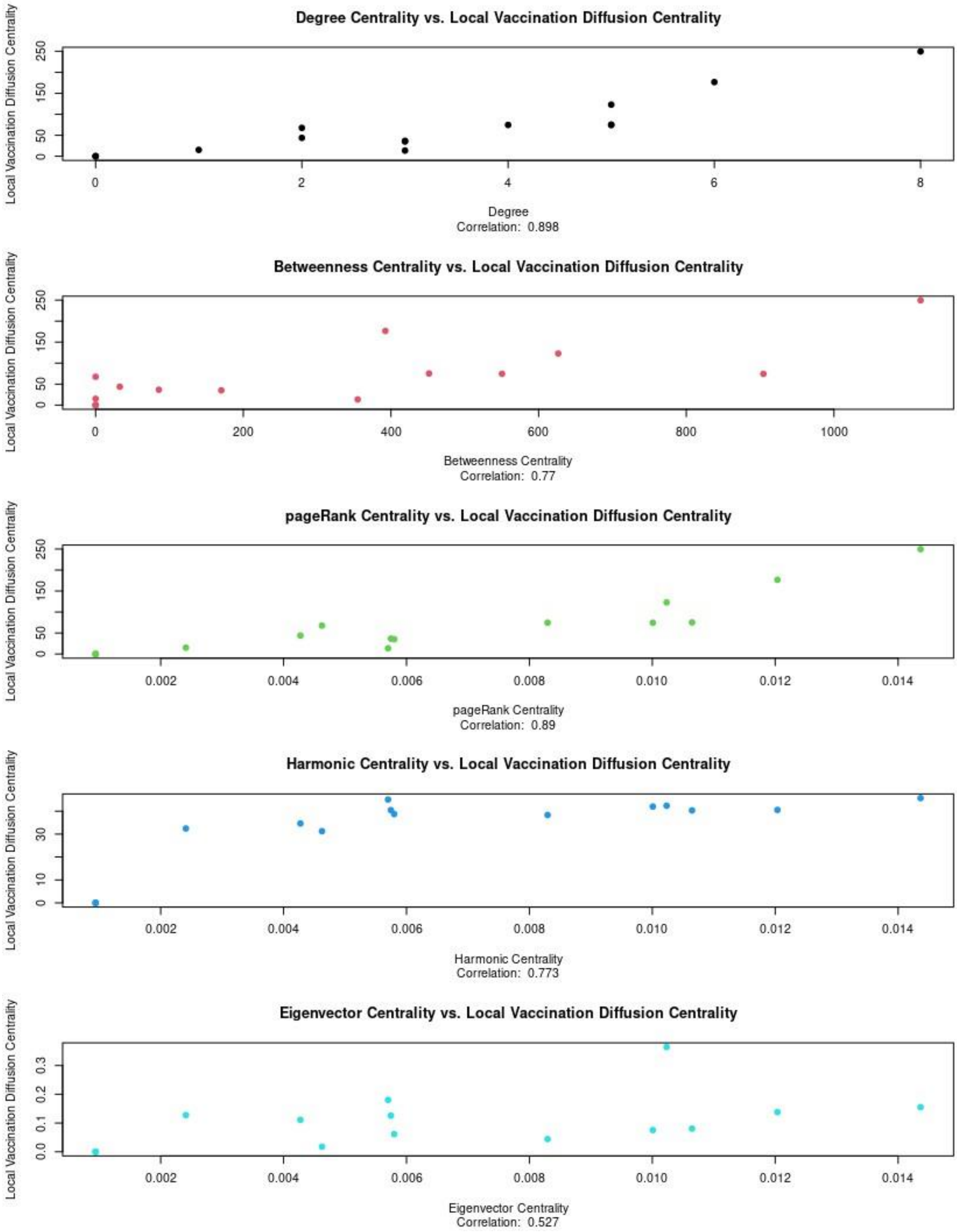


Figure 23: The local Vaccination Diffusion Centrality versus established centrality measures

Discussion

While there has been a lot of research about diffusion centralities, spread maximization, and topic-related diffusions, there has not been much research on explaining a diffusion on the basis of practical, easy to collect and standardly used variables using simulation techniques and how to derive an easy-to-use centrality measure from it. While there has been much research about vaccination hesitancy, a centrality measure about finding central nodes in the diffusion of vaccination willingness for information campaign targeting has not been developed.

This thesis is a proof of concept for an easy-to-use diffusion centrality measure which focuses on explaining the spread of an entity in a network. The proof of concept has been performed on the problem of the spread of pro-vaccination sentiment in 75 villages and a vaccination centrality measure has been developed.

In this process, we learned that the degree difference of two adjacent nodes is more suitable to explain the spread of pro-vaccination sentiment than all other variables analysed. The difference in socioeconomic status, whether the information sending node is a community leader or not, or the average neighbour degree of the originating edge were also found to be central in explaining the spread, however, to a much lesser extent than the degree difference. The sum of the age of a node and the age of its neighbour was least suitable to explain the spread of pro-vaccination sentiment along this respective edge.

These findings, encoded in the importance parameters, can support the process of designing questionnaires concerning diffusion processes with limited length when one needs to decide which variables are to be included and which are not.

The derived socio-centric vaccination centrality measure returned results which were similar to standard centrality measures. Their correlation can lead to the conclusion that the centrality measure lies in the range of standard measures.

The same was found for the locally applicable version of the centrality measure. Relying on average importance parameters for assessing the centrality of an ego in an ego-centric network led to similar results as for the socio-centric version.

The chosen approach explains the diffusion of a specific topic and the simulation does not only rely on basic topological network information, but on ego, tie and network attributes. This enables analysing the relative importance between all different types of attributes and finally allows the simulation of more realistic diffusion models.

The matching of networks from Karnataka and Malegaon produced networks with distributions which are in the range of real-world network distributions. The simulated data could create a complete and realistic dataset. The relative position normalization helped to take the comparative local status into account before matching. Various measures could be taken in the future in order to avoid that only a fraction of the nodes from one of the datasets are actually matched, like introducing a random function when choosing from multiple best fit nodes.

As this proof of concept has proven to function on the analysed networks, an outlook on the general approach to derive a centrality measure concerning the diffusion of an item/topic in a network can be sketched:

Applicability and Transferability to other Diffusion Phenomenon

The previously presented centrality can be generalized and applied to other diffusion processes. This general approach has not been done by previous approaches like in [10][12][39][40][46]. So, instructions for a general diffusion centrality are described:

To generate a diffusion centrality concerning a specific topic, the following information must be available:

- Aim variable:
  Which diffusion phenomenon do you want to explain and find out which nodes were central in its spread?
  For instance, smoking uptake, the adoption of new smart phones, or uptake of a new sport activity can be aim variables.
- Explanatory variables:
  Which variables could explain how the aim variable has spread over the network? E.g. being in the same peer group as a smoker could explain smoking uptake; a difference in technology expertise between individuals could explain the adoption of new smartphone; or an advice by a sporting friend could explain the uptake of a new sport.
- One full socio-centric network of the respective community to train the importance parameters.

The resulting importance parameters imply the relative importance of the considered explanatory variables. As it has been done as a proof of concept in this thesis, the importance parameters of the explanatory variables can be calculated by simulating a diffusion process and optimising the parameters such that the real-world distribution of the aim variable is approximately the same as the one simulated in the diffusion process.

The explanatory variables must be defined. For each of these, at least three types of relations are possible: Differences, sums or simply the value of one node adjacent to an edge. Each variable must be divided by the median to normalize it and to reach comparability and transferability of the resulting importance parameters. For the difference matrices, all values have to be shifted into the positive domain by adding the absolute value of the minimum to all values before dividing by the median. The variables are multiplied with the importance parameters and finally summed up.

The result is a centrality measure which implies the centrality of a node in the diffusion of the aim variable. (E.g. how important was the node for the spread of smoking in its peer group?).

Various additional questions arise with the formulation of the generalized diffusion centrality: For the diffusion of which entities and for which networks and diffusion mechanisms is this approach feasible / not feasible? How does the centrality measure scale with a large number of nodes? Which coding (and programming language) could be used to ensure efficiency? These questions will go along with the development of the more general centrality.

As already pointed out in the various sections, this proof of concept's applicability is limited. The generated data basis only allows to judge the feasibility of the approach. Some data points had to be simulated based on other data points. The networks were undirected, a directed network could come closer to real world networks. The nodes represented households, an analysis based on individuals could give more insights and could model the real-world more closely. The diffusion algorithm was additionally based on multiple assumptions e.g. that every node gets external pro-vaccination sentiment from media sources or from relatives from outside the village in every iteration or that the diffusion model does not take into account whether a node is more socially active than other nodes.

After having discussed the general applicability of the concept, the question is how the vaccination-specific approach can be used in practice: Can this centrality measure be used in practical operations with the goal of choosing persons from communities as initial seeds to reach a maximum spread over a network? Not directly. The question of influence maximization has been a topic in research for a long time. The proposed vaccination diffusion centrality correlates with other centrality measures which have been proven of being an indicator for reaching a maximum spread over a network, e.g. in the papers of Dihyat [24], Ni [53], and Susarla [66]. The conclusion that a person with a high vaccination diffusion centrality is a better target for reaching a maximum spread in vaccination information campaigns is however not valid. The presented approach must first be tested on real-world data (not on simulated data which are the basis of this work) and second, it must be tested whether a high vaccination diffusion centrality actually goes along with a maximum spread over the network as in the case of other centrality measures.

Coming back to the beginning of this thesis and to you being a field worker of a vaccination information campaign. How do you find the nodes who are most likely to be central in circulating your argument for vaccines?

The answer is that there are already various general solutions in research for the most effective strategy. Specifically for vaccination campaigns, this thesis proposes another one: Collect some ego networks of individuals in the village and find those with the highest vaccination diffusion centrality; hence find those with much more contacts than their friends, those who have a higher socio-economic status, or those who are community leaders.

References

[1]

I. Achitouv, ‘Propagation of epidemics in a polarized society: impact of clustering among unvaccinated individuals’. arXiv, Jun. 01, 2022. Accessed: Aug. 25, 2022. [Online]. Available: http://arxiv.org/abs/2206.00357

[2]

M. Akbarzadeh, S. Memarmontazerin, and S. Soleimani, ‘Where to look for power Laws in urban road networks?’, Appl Netw Sci, vol. 3, no. 1, Art. no. 1, Dec. 2018, doi: 10.1007/s41109-018-0060-9.

[3]

Z. Aktürk, K. Linde, A. Hapfelmeier, R. Kunisch, and A. Schneider, ‘COVID-19 vaccine hesitancy in people with migratory backgrounds: a cross-sectional study among Turkish- and German-speaking citizens in Munich’, BMC Infectious Diseases, vol. 21, no. 1, p. 1214, Dec. 2021, doi: 10.1186/s12879021-06940-9.

[4]

D. Allington, S. McAndrew, V. Moxham-Hall, and B. Duffy, ‘Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the COVID-19 pandemic’, Psychological Medicine, pp. 1–12, Apr. 2021, doi: 10.1017/S0033291721001434.

[5]

D. Allington, S. McAndrew, V. L. Moxham-Hall, and B. Duffy, ‘Media usage predicts intention to be vaccinated against SARS-CoV-2 in the US and the UK’, Vaccine, vol. 39, no. 18, pp. 2595–2603, Apr. 2021, doi: 10.1016/j.vaccine.2021.02.054.

[6]

K. Ambler, S. Godlonton, and M. P. Recalde, ‘Follow the leader? A field experiment on social influence’, Journal of Economic Behavior & Organization, vol. 188, pp. 1280–1297, Aug. 2021, doi: 10.1016/j.jebo.2021.05.022.

[7]

A. B. Amin et al., ‘Association of moral values with vaccine hesitancy’, Nat Hum Behav, vol. 1, no. 12, pp. 873–880, Dec. 2017, doi: 10.1038/s41562-017-0256-5.

[8]

G. Andrade, ‘Predictive demographic factors of Covid-19 vaccine hesitancy in Venezuela: A crosssectional study’, Vacunas, vol. 23, pp. S22–S25, May 2022, doi: 10.1016/j.vacun.2021.07.009.

[9]

C. Aslay, N. Barbieri, F. Bonchi, and R. Baeza-Yates, ‘Online Topic-aware Influence Maximization Queries’. OpenProceedings.org, Athens, 2014. doi: 10.5441/002/EDBT.2014.28.

[10]

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, ‘The Diffusion of Microfinance’, Science, vol. 341, no. 6144, Jul. 2013, doi: 10.1126/science.1236498.

[11]

A.-L. Barabási and R. Albert, ‘Emergence of Scaling in Random Networks’, Science, vol. 286, no. 5439, pp. 509–512, Oct. 1999, doi: 10.1126/science.286.5439.509.

[12]

N. Barbieri, F. Bonchi, and G. Manco, ‘Topic-aware social influence propagation models’, Knowl Inf Syst, vol. 37, no. 3, pp. 555–584, Dec. 2013, doi: 10.1007/s10115-013-0646-6.

[13]

P. Bonacich, ‘Power and Centrality: A Family of Measures’, American Journal of Sociology, vol. 92, no. 5, pp. 1170–1182, Mar. 1987, doi: 10.1086/228631.

[14]

L. Breiman and A. Cutler, ‘Random Forests’. Accessed: Sep. 06, 2022. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/

[15]

S. Brin and L. Page, ‘The anatomy of a large-scale hypertextual Web search engine’, Computer Networks and ISDN Systems, vol. 30, no. 1, pp. 107–117, Apr. 1998, doi: 10.1016/S01697552(98)00110-X.

[16]

S. Callaghan et al., ‘Correlates and disparities of intention to vaccinate against COVID-19’, Soc Sci Med, vol. 272, p. 113638, Mar. 2021, doi: 10.1016/j.socscimed.2020.113638.

[17]

R. M. Carpiano, A. N. Polonijo, N. Gilbert, L. Cantin, and E. Dubé, ‘Socioeconomic status differences in parental immunization attitudes and child immunization in Canada: Findings from the 2013 Childhood National Immunization Coverage Survey (CNICS)’, Preventive Medicine, vol. 123, pp. 278–287, Jun. 2019, doi: 10.1016/j.ypmed.2019.03.033.

[18]

C. J. Charpentier, J. Aylward, J. P. Roiser, and O. J. Robinson, ‘Enhanced Risk Aversion, But Not Loss Aversion, in Unmedicated Pathological Anxiety’, Biol Psychiatry, vol. 81, no. 12, pp. 1014–1022, Jun. 2017, doi: 10.1016/j.biopsych.2016.12.010.

[19]

W. Chen, C. Wang, and Y. Wang, ‘Scalable influence maximization for prevalent viral marketing in large-scale social networks’, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10, Washington, DC, USA, 2010, pp. 1029–1038. doi: 10.1145/1835804.1835934.

[20]

W. Chen, Y. Yuan, and L. Zhang, ‘Scalable Influence Maximization in Social Networks under the Linear Threshold Model’, in 2010 IEEE International Conference on Data Mining, Dec. 2010, pp. 88–97. doi: 10.1109/ICDM.2010.118.

[21]

Y. Chung, Y. Li, and J. Jia, ‘Exploring embeddedness, centrality, and social influence on backer behavior: the role of backer networks in crowdfunding’, J. of the Acad. Mark. Sci., vol. 49, no. 5, pp. 925–946, Sep. 2021, doi: 10.1007/s11747-021-00779-x.

[22]

K. Danis, T. Georgakopoulou, T. Stavrou, D. Laggas, and T. Panagiotopoulos, ‘Socioeconomic factors play a more important role in childhood vaccination coverage than parental perceptions: a crosssectional study in Greece’, Vaccine, vol. 28, no. 7, pp. 1861–1869, Feb. 2010, doi:

10.1016/j.vaccine.2009.11.078.

[23]

David Dobolyi, ‘Moral Foundations Theory | moralfoundations.org’, Sep. 13, 2022. https://moralfoundations.org/ (accessed Sep. 13, 2022).

[24]

M. M. H. Dihyat, K. Malik, M. A. Khan, and B. Imran, ‘Detecting Ideal Instagram Influencer Using Social Network Analysis’. 2021. Accessed: Sep. 15, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Detecting-Ideal-Instagram-Influencer-Using-Social-Dihyat-Malik/58e19a63a60d6e4cd93672d3b3db10ae3a370b75

[25]

K. M. Douglas et al., ‘Understanding Conspiracy Theories’, Political Psychology, vol. 40, no. S1, pp. 3–35, 2019, doi: 10.1111/pops.12568.

[26]

A. A. Dror et al., ‘Vaccine hesitancy: the next challenge in the fight against COVID-19’, Eur J Epidemiol, vol. 35, no. 8, pp. 775–779, Aug. 2020, doi: 10.1007/s10654-020-00671-y.

[27]

K. J. Fietkiewicz, E. Lins, K. S. Baran, and W. G. Stock, ‘Inter-Generational Comparison of Social Media Use: Investigating the Online Behavior of Different Generational Cohorts’, in 2016 49th Hawaii International Conference on System Sciences (HICSS), Jan. 2016, pp. 3829–3838. doi: 10.1109/HICSS.2016.477.

[28]

K. C. Freeman, ‘A Set of Measures of Centrality Based on Betweenness’, Sociometry, vol. 40, no. 1, pp. 35–41, 1977, doi: 10.2307/3033543.

[29]

J. Geipel, L. H. Grant, and B. Keysar, ‘Use of a language intervention to reduce vaccine hesitancy’, Sci Rep, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s41598-021-04249-w.

[30]

K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, ‘Classification of scale-free networks’, Proceedings of the National Academy of Sciences, vol. 99, no. 20, pp. 12583–12588, Oct. 2002, doi: 10.1073/pnas.202301299.

[31]

S. Goldstein, N. E. MacDonald, and S. Guirguis, ‘Health communication and vaccine hesitancy’, Vaccine, vol. 33, no. 34, pp. 4212–4214, Aug. 2015, doi: 10.1016/j.vaccine.2015.04.042.

[32]

A. Goyal, W. Lu, and L. V. S. Lakshmanan, ‘SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model’, in 2011 IEEE 11th International Conference on Data Mining, Dec. 2011, pp. 211–220. doi: 10.1109/ICDM.2011.132.

[33]

J. Graham et al., ‘Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism’, in Advances in Experimental Social Psychology, vol. 47, P. Devine and A. Plant, Eds. Academic Press, 2013, p. 68. doi: 10.1016/B978-0-12-407236-7.00002-4.

[34]

G. Harling and A. C. Tsai, ‘Using social networks to understand and overcome implementation barriers in the global HIV response’, J Acquir Immune Defic Syndr, vol. 82, no. Suppl 3, pp. S244–S252, Dec. 2019, doi: 10.1097/QAI.0000000000002203.

[35]

B. Hollstein, T. Töpfer, and J. Pfeffer, ‘Collecting egocentric network data with visual tools: A comparative study’, Network Science, vol. 8, pp. 223–250, Feb. 2020, doi: 10.1017/nws.2020.4.

[36]

A. Hudson and W. J. Montelpare, ‘Predictors of Vaccine Hesitancy: Implications for COVID-19 Public Health Messaging’, International Journal of Environmental Research and Public Health, vol. 18, no. 15, Art. no. 15, Jan. 2021, doi: 10.3390/ijerph18158054.

[37]

G. Jain, A. B. Sreenivas, S. Gupta, and A. A. Tiwari, in Causes and Symptoms of Socio-Cultural Polarization: Polarization Around the Vaccine Development for COVID-19, Springer, 2022, pp. 51–72.

Accessed: Aug. 25, 2022. [Online]. Available: https://econpapers.repec.org/bookchap/sprsprchp/978-981-16-5268-4_5f3.htm

[38]

C. Jarrett, R. Wilson, M. O'Leary, E. Eckersberger, and H. J. Larson, 'Strategies for addressing vaccine hesitancy – A systematic review', Vaccine, vol. 33, no. 34, pp. 4180–4190, Aug. 2015, doi: 10.1016/j.vaccine.2015.04.040.

[39]

C. Kang, S. Kraus, C. Molinaro, F. Spezzano, and V. S. Subrahmanian, 'Diffusion centrality: A paradigm to maximize spread in social networks', Artificial Intelligence, vol. 239, pp. 70–96, Oct. 2016, doi: 10.1016/j.artint.2016.06.008.

[40]

C. Kang, C. Molinaro, S. Kraus, Y. Shavitt, and V. S. Subrahmanian, 'Diffusion Centrality in Social Networks', in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug. 2012, pp. 558–564. doi: 10.1109/ASONAM.2012.95.

[41]

H. C. Kelman, 'Compliance, identification, and internalization three processes of attitude change', Journal of Conflict Resolution, vol. 2, no. 1, pp. 51–60, 1958.

[42]

A. Kibongani Volet, C. Scavone, D. Catalán-Matamoros, and A. Capuano, 'Vaccine Hesitancy Among Religious Groups: Reasons Underlying This Phenomenon and Communication Strategies to Rebuild Trust', Frontiers in Public Health, vol. 10, 2022, Accessed: Sep. 04, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.824560

[43]

D. A. Kim et al., 'Social network targeting to maximise population behaviour change: a cluster randomised controlled trial', The Lancet, vol. 386, no. 9989, pp. 145–153, Jul. 2015, doi: 10.1016/S0140-6736(15)60095-2.

[44]

C. W. King and J. O. Summers, ‘Overlap of Opinion Leadership across Consumer Product Categories’, Journal of Marketing Research, vol. 7, no. 1, pp. 43–50, 1970, doi: 10.2307/3149505.

[45]

C. Latkin et al., ‘A longitudinal study of vaccine hesitancy attitudes and social influence as predictors of COVID-19 vaccine uptake in the US’, Human Vaccines & Immunotherapeutics, vol. 18, no. 5, p. 2043102, Nov. 2022, doi: 10.1080/21645515.2022.2043102.

[46]

Y. Leng, Y. Sella, R. Ruiz, and A. Pentland, ‘Contextual centrality: going beyond network structure’, Sci Rep, vol. 10, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41598-020-62857-4.

[47]

J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, ‘Cost-effective outbreak detection in networks’, in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, Aug. 2007, pp. 420–429. doi: 10.1145/1281192.1281239.

[48]

S. Lu et al., ‘Use of Laplacian Heat Diffusion Algorithm to Infer Novel Genes With Functions Related to Uveitis’, Frontiers in Genetics, vol. 9, 2018, Accessed: Sep. 03, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2018.00425

[49]

N. E. MacDonald and SAGE Working Group on Vaccine Hesitancy, ‘Vaccine hesitancy: Definition, scope and determinants’, Vaccine, vol. 33, no. 34, pp. 4161–4164, Aug. 2015, doi: 10.1016/j.vaccine.2015.04.036.

[50]

M. Marchiori and V. Latora, ‘Harmony in the Small-World’, Physica A: Statistical Mechanics and its Applications, vol. 285, no. 3–4, pp. 539–546, Oct. 2000, doi: 10.1016/S0378-4371(00)00311-3.

[51]

T. Mitra, S. Counts, and J. Pennebaker, ‘Understanding Anti-Vaccination Attitudes in Social Media’, Proceedings of the International AAAI Conference on Web and Social Media, vol. 10, no. 1, Art. no. 1, 2016.

[52]

F. S. Mohd Azizi, Y. Kew, and F. M. Moy, ‘Vaccine hesitancy among parents in a multi-ethnic country, Malaysia’, Vaccine, vol. 35, no. 22, pp. 2955–2961, May 2017, doi: 10.1016/j.vaccine.2017.04.010.

[53]

C. Ni, J. Yang, and D. Kong, ‘Sequential seeding strategy for social influence diffusion with improved entropy-based centrality’, Physica A: Statistical Mechanics and its Applications, vol. 545, p. 123659, May 2020, doi: 10.1016/j.physa.2019.123659.

[54]

J. Nieminen, ‘On the centrality in a graph’, Scandinavian Journal of Psychology, vol. 15, no. 1, pp. 332–336, 1974, doi: 10.1111/j.1467-9450.1974.tb00598.x.

[55]

G. Noyman-Veksler, D. Greenberg, I. Grotto, and G. Shahar, ‘Parents’ malevolent personification of mass vaccination solidifies vaccine hesitancy’, J Health Psychol, vol. 26, no. 12, pp. 2164–2172, Oct. 2021, doi: 10.1177/1359105320903475.

[56]

J.-P. Onnela et al., ‘Polio vaccine hesitancy in the networks and neighborhoods of Malegaon, India’, Social Science & Medicine, vol. 153, pp. 99–106, März 2016, doi: 10.1016/j.socscimed.2016.01.024.

[57]

A. Pantanowitz and T. Marwala, ‘Missing Data Imputation Through the Use of the Random Forest Algorithm’, in Advances in Computational Intelligence, Berlin, Heidelberg, 2009, pp. 53–62. doi: 10.1007/978-3-642-03156-4_6.

[58]

M. Powell, ‘The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives’, Technical Report, Cambridge University Department of Applied Mathematics and Theoretical Physics, Jan. 2009.

[59]

H. Risselada, P. C. Verhoef, and T. H. A. Bijmolt, ‘Indicators of opinion leadership in customer networks: self-reports and degree centrality’, Mark Lett, vol. 27, no. 3, pp. 449–460, Sep. 2016, doi: 10.1007/s11002-015-9369-7.

[60]

E. Robertson et al., ‘Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study’, Brain, Behavior, and Immunity, vol. 94, pp. 41–50, Mai 2021, doi: 10.1016/j.bbi.2021.03.008.

[61]

F. A. Rodrigues, ‘Network Centrality: An Introduction’, in A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems, E. E. N. Macau, Ed. Cham: Springer International Publishing, 2019, pp. 177–196. doi: 10.1007/978-3-319-78512-7_10.

[62]

J. A. Ross, ‘Influence of Expert and Peer upon Negro Mothers of Low Socioeconomic Status’, The Journal of Social Psychology, vol. 89, no. 1, pp. 79–84, Feb. 1973, doi: 10.1080/00224545.1973.9922570.

[63]

K. Saito, R. Nakano, and M. Kimura, ‘Prediction of Information Diffusion Probabilities for Independent Cascade Model’, in Knowledge-Based Intelligent Information and Engineering Systems, I. Lovrek, R. J. Howlett, and L. C. Jain, Eds. Berlin, Heidelberg: Springer, 2008, pp. 67–75.

[64]

D. Sia, P. Fournier, J.-F. Kobiané, and B. K. Sondo, ‘Rates of coverage and determinants of complete vaccination of children in rural areas of Burkina Faso (1998-2003)’, BMC Public Health, vol. 9, no. 1, p. 416, Nov. 2009, doi: 10.1186/1471-2458-9-416.

[65]

L. E. Smith, R. Amlôt, J. Weinman, J. Yiend, and G. J. Rubin, ‘A systematic review of factors affecting vaccine uptake in young children’, Vaccine, vol. 35, no. 45, pp. 6059–6069, Oct. 2017, doi: 10.1016/j.vaccine.2017.09.046.

[66]

A. Susarla, J.-H. Oh, and Y. Tan, ‘Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube’, Information Systems Research, vol. 23, no. 1, pp. 23–41, Mar. 2012, doi: 10.1287/isre.1100.0339.

[67]

F. Tang and H. Ishwaran, ‘Random Forest Missing Data Algorithms’, Stat Anal Data Min, vol. 10, no. 6, pp. 363–377, Dec. 2017, doi: 10.1002/sam.11348.

[68]

J. N. A. Tetteh, V. K. Nguyen, and E. A. Hernandez-Vargas, ‘Network models to evaluate vaccine strategies towards herd immunity in COVID-19’, Journal of Theoretical Biology, vol. 531, p. 110894, Dec. 2021, doi: 10.1016/j.jtbi.2021.110894.

[69]

T. W. Valente, ‘Network Interventions’, Science, vol. 337, no. 6090, pp. 49–53, 2012.

[70]

A. L. Wagner et al., ‘Comparisons of Vaccine Hesitancy across Five Low- and Middle-Income Countries’, Vaccines (Basel), vol. 7, no. 4, p. 155, Oct. 2019, doi: 10.3390/vaccines7040155.

[71]

A. L. Wagner, A. R. Shotwell, M. L. Boulton, B. F. Carlson, and J. L. Mathew, ‘Demographics of Vaccine Hesitancy in Chandigarh, India’, Frontiers in Medicine, vol. 7, 2021, Accessed: Aug. 29, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2020.585579

[72]

D. J. Watts, J. Peretti, M. Frumin, and D. Watts, ‘Viral Marketing for the Real World Duncan J. Watts, Jonah Peretti, and Michael Frumin’, Harvard Business Review, vol. 85, no. 5, Jan. 2007, [Online]. Available: https://www.microsoft.com/en-us/research/publication/viral-marketing-for-the-realworld-duncan-j-watts-jonah-peretti-and-michael-frumin/

[73]

D. J. Watts and S. H. Strogatz, ‘Collective dynamics of “small-world” networks’, Nature, vol. 393, no. 6684, Art. no. 6684, Jun. 1998, doi: 10.1038/30918.

[74]

F. Wei et al., ‘Identification and characteristics of vaccine refusers’, BMC Pediatrics, vol. 9, no. 1, p. 18, März 2009, doi: 10.1186/1471-2431-9-18.

[75]

A. Wheelock, A. Thomson, and N. Sevdalis, ‘Social and psychological factors underlying adult vaccination behavior: lessons from seasonal influenza vaccination in the US and the UK’, Expert Review of Vaccines, vol. 12, no. 8, pp. 893–901, Aug. 2013, doi: 10.1586/14760584.2013.814841.

[76]

A. C. Wu et al., ‘Postpartum Mothers’ Attitudes, Knowledge, and Trust Regarding Vaccination’, Matern Child Health J, vol. 12, no. 6, pp. 766–773, Nov. 2008, doi: 10.1007/s10995-007-0302-4.

[77]

‘Who Are the Scheduled Castes, Scheduled Tribes, and OBCs?’, Vakilsearch | Blog, May 31, 2022. https://vakilsearch.com/blog/who-are-the-scheduled-castes-scheduled-tribes-and-obcs/ (accessed Sep. 04, 2022).

Table of Figures

Appendix

Alternative Diffusion Algorithms

Additional diffusion algorithms to the previously presented in the Methods Section were analysed. Their basic functionality and applicability to our problem are described in the following.

Markov Random Walk with Restart

In this diffusion algorithm, each new step is calculated by multiplying the adjacency matrix with the distribution vector (of pro-vaccination arguments e.g.). Restart is added to the diffusion (or more generally the stochastic process) by adding a percentage of the initial distribution vector to the newly generated vector in each iteration step. So, if the adjacency matrix is W, the initial distribution vector p_0 and the current distribution vector p_t with t being a natural number (describing e.g. time after the initialization of the diffusion), then $p_{t+1} = (1 - r)Wp_t + rp_0$.

This diffusion process does not have any advantages in comparison to PageRank. Due to the fact that the initial distribution has to be normalized and the columns of W must be normalized as well, the process models that the amount of information stays the same. Additionally, adding a proportion of the initial distribution vector in each iteration step makes the process dependent on the initial vector, which would have to be chosen as well and limits the possibility of stating something about the basic network structures independent of the starting point of the diffusion. As with PageRank, this model also assumes that pro-vaccination sentiment is passed on to a neighbour and then put aside until another person reactivates the node.

Nearest Neighbours

This diffusion algorithm starts at a selected or random node in a graph with weighted edges. The diffusion algorithm selects the neighbour with the least edge weight until a pre-determined depth k is reached.

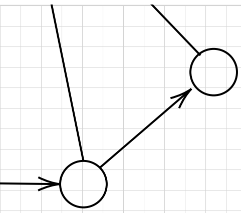
This algorithm is dependent on the initial node of the diffusion. Applied to the task of the thesis, it would assume that the diffusion only spreads along the edges with the least weight, so to the neighbour node which is most likely to be influenced. The algorithm only selects the neighbour with the maximum edge weight. It does not choose among the neighbours according to a probability approach. If two neighbours are almost equally likely of being chosen, the principle of ‘the winner takes it all’ is applied which excludes all other neighbours which would in reality also have a certain probability of being chosen. Furthermore, the application of the algorithm could only result in a classification of Accepting and Refusing, not in three classes of the used dataset: refusing, hesitant and accepting. Additionally, the envisioned optimization of the edge weights would optimize for an optimal chain of nodes which would lead to the best fit with the ‘real-world' dataset. The order of the nodes in the diffusion process would matter whereas an independent process would be envisioned. It would not consider all neighbours of a node at the same time but only one.

Overview Diagram

On the following pages, an overview of the Methods Section is provided.

Who is central in the diffusion of pro-vaccination sentiment?

For this proof-of-concept, the required network data including information about vaccination wiilingness / hesitancy are simulated based on real-world data.



We generate one network dataset by combining two real-world datasets.

Karnataka Networks

Socio-centric network data from 75 villages

Undirected ties between households (HH) based on advice-seeking and -giving

Information about HH:
number of rooms
rooftype
...

Information about HH leader:
age
sub-caste
education
language
native home
occupation
...

Missing data about HH leader are simulated based on HH information using random forest prediction.

Generated, Merged Networks

Network Structure from Karnataka

```
Merging the Networks:
for each node in Karnataka
  Choose optimal node
  from Malegaon
  with minimal distance in
     - Degree *
     - Closed Triads *
     - In-Betweenness
       Centrality *
     - Education Status ***
     - Ownership of Toilets ***
     - Number of Rooms **
     - Socio economic status**
```

* order rank normed to 1

**normed to 1 (division with max value)

*** Category equal or not equal

Malegaon Network

Ca. 2500 household heads from 25 neighbourhoods

Directed ties based on advice-seeking nominations (only up to 4 nominations per interviewee)

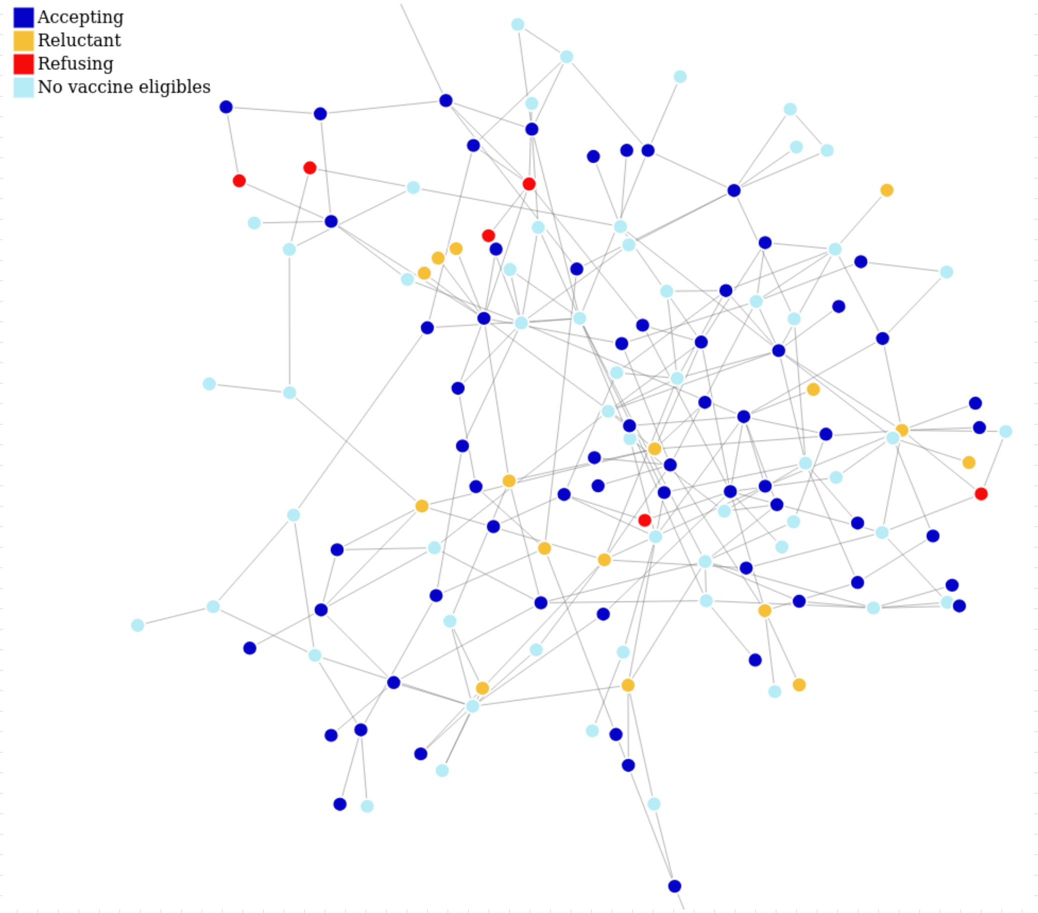
Information about HH:
- TV-, cooking cylinder-, toilet- ownership
- numbers of person in HH
- number of rooms
- opinion concerning polio vaccinations

The optimal node index $\hat{j}_i$ from Malegaon for the i. node from Karnataka is

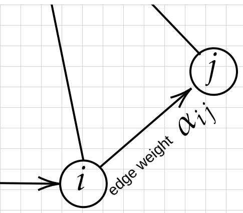
$$\hat{j}_i = argmin_{j=1,...,M} \sum_{z=1}^{Z} ||x_{iz} - y_{jz}||_1$$

with $x_{iz},\ i = 1,...,N$ all nodes from Karnataka and their shared attributes z
and $y_{jz},\ j = 1,...,M$ all nodes from Malegaon and their shared attributes z
and $z = 1,...,Z$ all shared attributes of the nodes
and $\hat{j}_i$ the optimal node from Malegaon for the i. node from Karnataka.

Generated Network with Vaccination Willingness / Hesitancy



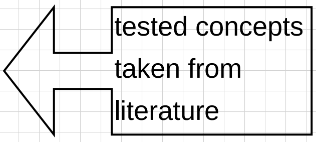
Afterwards, based on the generated network, the relative importance of attributes in the diffusion is found.



The likelier one node influences a neighbour towards pro-vaccination sentiment, the higher the edge weight α, also called influence weight. α is a linear combination of various attributes and dependent on β.

Between node i and node j , the influence weight is

$$\begin{aligned} \alpha_{ij}(\beta) \ = \ & \beta_1 * A_1 && A_1 \text{ being e.g. difference in degree} \\ & + \beta_2 * A_2 && A_2 \text{ being e.g. sum of age} \\ & + \beta_3 * ... \\ & \ \vdots \\ & + \ \beta_K * A_K && A_K \text{ being e.g. difference in socio-economic status} \end{aligned}$$



tested concepts taken from literature

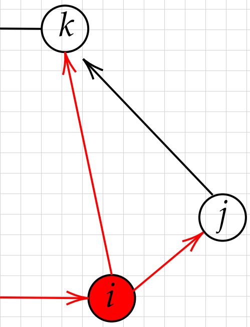
with $\beta_k \geq 0$ being the importance parameter vector entries, initialized by e.g. $\beta \ = \ (1,..., 1)^T$,

$A_1,..., \ A_k$ being the normalized attributes,

and $k = 1,..., K, \ K$ the number of attributes.

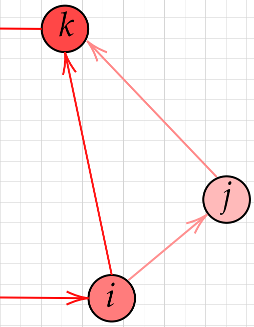
The Adjacency Matrix $A(\beta)$ is formed with all $\alpha_{ij}(\beta)$.

$$A(\beta) = \begin{bmatrix} 0 & \cdots & \alpha_{1j}(\beta) & \cdots & \cdots \\ \vdots & 0 & \vdots & \vdots & \vdots \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \alpha_{i1}(\beta) & \cdots & \alpha_{ij}(\beta) & 0 & \vdots \\ \vdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$



1 unit of pro-vaccination sentiment is inserted in the network.
A diffusion process (PageRank / Laplacian Heat Diffusion) is initiated.

Final equilibrium state of the diffusion



$$\begin{pmatrix} 0.05 \\ 0.2 \\ 0 \\ 0.16 \\ \vdots \end{pmatrix}$$

Final diffusion state at Node j

Final diffusion state at Node k

Final diffusion state at Node l

Final diffusion state at Node i

Difference with generated real-world data: Cost function definition

The optimal importance parameter vector $\widehat{\beta}$ is

$$\widehat{\beta} = argmin_{\beta} \sum_{i=1}^{N} \left(\text{diffusion}(A(\beta))_i - x_{\text{real-world, } i} \right)^2$$

s.t. $\beta_k \geq 0$,

with $x_{\text{real-world, } i}$ being the "real-world" vaccination status,
i the index of a node,
diffusion$()_i$ the final value of node i
after the diffusion with adjacency matrix $A(\beta)$,
and $k = 1, \dots, K$, K the number of attributes.

For example $\widehat{\beta} = argmin_{\beta} \left\| \begin{pmatrix} 0.05 \\ 0.2 \\ 0 \\ 0.16 \\ \vdots \end{pmatrix} - \begin{pmatrix} 0.07 \\ 0.24 \\ 0 \\ 0.24 \\ \vdots \end{pmatrix} \right\|_2^2$

Final diffusion state values dependent on β

Real-world values (Vaccination Willingness)

Optimising the Cost Function

$$\text{The optimal } \widehat{\beta} = \begin{pmatrix} 3.21 \\ 0.34 \\ \vdots \\ 1.78 \end{pmatrix}$$

The algorithm raised the value from the initial 1 to 3.21, so the degree difference is central in explaining the diffusion.

The sum of age with one's neighbours does not play a important role in the diffusion.

A difference in socio-economic status is somewhat important for the diffusion.

With the optimal $\widehat{\beta}$, the result of the diffusion is closest to the real-world data.
The different entries of $\widehat{\beta}$ indicate the relative importance of the respective attributes for the diffusion.

E.g. when looking for central nodes, having a high degree difference to and a high difference in socio-economic status with one's neighbours makes one more central in diffusion processes than having a high sum of age.

These relative importance factors can be used to assess the importance of including a variable vs. another variable in a questionnaire. They can also be used to take the variables' relative importance into account when designing a centrality measure for finding central nodes.

As a last step, a proof of concept with a derived centrality measure is performed.

1. Derive a vaccination diffusion centrality measure for a full socio-centric network

Calculate the optimal importance parameters $\widehat{\beta}$ for the specific network.

Vaccination Diffusion Centrality v for node j and its neighbours $i = 1,...,I,$
I being the degree of j :

$$v_j \;=\; \widehat{\beta}_1 * \sum_i^I A_{1i}$$

A_{1i} the difference in degree with neighbour i

$$+\,\widehat{\beta}_2 * \sum_i^I A_{2i}$$

A_{2i} the sum of age with neighbour i

$$\vdots$$

$$+\,\widehat{\beta}_K * \sum_i^I A_{Ki}$$

A_{Ki} the difference in socio-economic status with neighbour i

with $A_1,...,\; A_k$ being the normalized attributes,
$\widehat{\beta}_k \geq 0 \;\forall\, k,\; k = 1,...,K,\; K$ the number of attributes.

All attributes are normalized by dividing by the median of the attribute of all nodes.
If attribute values are negative, they are shifted to the positive domain before.

2. When only ego-networks are available, derive a locally applicable diffusion centrality measure

→ Same approach as 1. except for:

- Insert the means of $\widehat{\beta}$, so $\widetilde{\beta}$, of all analysed 75 villages into the formula for v_j
- Normalize the attributes by dividing by the medians and adding the minimum of all available attribute data in the ego networks

3. Outlook: Generalized Diffusion Centrality for any diffusion

It is described how the derived diffusion centrality can be applied to investigate any diffusion process in a network.
Based on an aim variable (e.g. smoking among youths) and explanatory variables (e.g. difference in age, difference in degree) and a full socio-centric network (a youth group), the importance parameters can be calculated and a centrality measure concerning the diffusion of smoking according to the approach above can be developed. The locally applicable version can be generated with the $\widehat{\beta}$ from the network.