Regression Chain Model for Predicting Epidemic Variables

Kirti Jain[1](✉), Vasudha Bhatnagar[1], and Sharanjit Kaur[2]

[1] Department of Computer Science, University of Delhi, Delhi, India
{kjain1,vbhatnagar}@cs.du.ac.in
[2] Acharya Narendra Dev College, University of Delhi, Delhi, India
sharanjitkaur@andc.du.ac.in

Abstract. Real-time detection and forecasting of disease dynamics is critical for healthcare authorities during epidemics. In this paper, we report a systematic investigation into the possibility of predicting three epidemic variables, viz., *peak day*, *peak infections*, and *span* of the epidemic using the Regression Chain Model.

We construct a dataset, *EpiNet*, using 35K synthetic networks of varied sizes and belonging to three network families. The dataset consists of five network features and three target variables obtained by simulating the SEIR epidemic model on the networks. We train Regression Chain Model (RCM) using four popular machine learning algorithms to predict the target variables. The model generally performs fairly well for *peak day* and *peak infections*, but the performance degrades for the *span* variable. Our preliminary investigation motivates further inquiry into the use of RCMs to replace computationally expensive epidemic simulations on larger networks.

Keywords: Contact network · Topological properties · Epidemic variables · Machine learning · Regression chain model

1 Introduction

The COVID-19 pandemic provided an unprecedented boost to research in epidemic modeling. When an epidemic spreads to a large population, early and real-time estimations of the disease infectivity are of critical importance for healthcare policy planners and administrators for managing and controlling the spread of disease. However, it is often impractical or impossible to continuously monitor the entire population to estimate the size, span, and severity of the epidemic.

1.1 Background and Motivation

Compartmental mathematical models like susceptible-infected-susceptible (SIS), susceptible-infected-recovered (SIR), susceptible-exposed-infected-recovered (SEIR), etc., have served as indispensable tools for estimating the epidemic

R. Thomson et al. (Eds.): SBP-BRiMS 2023, LNCS 14161, pp. 285–294, 2023.
https://doi.org/10.1007/978-3-031-43129-6_28

dynamics for almost a century [5]. The simplifying assumption of uniform interactions within the population (homogeneous mixing) is a well-understood caveat of these models [1]. This assumption not only affects the quality of estimates but also overlooks the complexity of the dynamics being modeled. This limitation has promoted research related to network-based simulations for understanding disease dynamics [5]. Network-based simulations of the epidemic models deliver comparatively more realistic approximates of epidemic dynamics due to the incorporation of connectivity patterns in the population. However, the cost of network-based simulations escalates steeply with the network size. This is the prime motivation to find alternatives to expensive simulations on large networks.

During network simulation of epidemics, the notable role played by the structure and the topological properties of contact networks in the spread of contagion has been established in several studies [6,9,10,12,13]. It is reasonable to infer that topological properties of networks carry the potential for predicting the epidemic variables, viz., *peak day*, *peak cases*, and *span*. Peak cases are the maximum number of infected cases on a given day. The day when the cases are maximum is the peak day and the span denotes the time period between the first and last infected cases.

Rodrigues et al. used machine learning models to identify and rank the topological properties of the network that are crucial to estimate the outbreak size of the epidemic [11]. The major limitation of this work is the use of features of a small subset of nodes, which can be misleading due to stochasticity and non-linearity in the simulation of epidemic spread. Bucur et al. used centrality measures as features to predict outbreak sizes in networks limited to ten nodes [3]. They empirically demonstrate that it is possible to accurately predict the outbreak using network measures in isomorphic networks. However, the network size used in this study is unrealistically small, and the method does not scale-up for application in the real world. Pérez-Ortiz et al. employ thirty network properties to predict the average percentage of infected individuals using linear regression [10]. Since distance-based network properties are computationally expensive, this limits the practical applicability of this approach to large networks.

1.2 Research Contributions

A critical analysis of recent related works reveals the following research gaps: i) use of computationally expensive topological properties to predict outbreak size, ii) prediction of only one epidemic variable, iii) effectiveness reported on small networks, iv) unavailability of data to reproduce results. Our empirical study fills these gaps and contributes in the following manner. We

i. use five inexpensive topological features of the networks that distinctively influence the pathogen spread (Sect. 2.1).

ii. address the prediction of three epidemic variables, viz., *peak day*, *peak cases*, and *span* using the Regression Chain Model (Sect. 2.2–2.3). We empirically validate our conjecture and demonstrate the possibility of accurately predicting epidemic variables using Regression Chain Model in a restricted environment. Our results encourage further study in this direction (Sect. 3).

iii. curate a dataset, *EpiNet*, with five topological features and three target variables obtained by simulating the SEIR[1] epidemic model on synthetic networks belonging to three diverse families. The dataset will permit the reproduction of results and promote further investigation in this direction by the research community (Sect. 2.4).

Organization: Section 2 describes the methodology used in this study. Experimental settings along with the results are presented in Sect. 3. Conclusion and future work are given in Sect. 4.

2 Methodology

In this section, we describe the methodology for network construction, topological features, the epidemic spread model, and the regression chain model, along with the estimators and the metrics used to evaluate the performance. We also give details of the construction of the dataset.

2.1 Network Models and Topological Properties

We use three network models belonging to diverse families, viz., Erdos-Renyi model, Watts-Strogatz model, and Stochastic Block model to construct random, small-world, and community-based networks, respectively. These networks are extensively used for modeling social structure in epidemiology [1,8]. For each constructed network, we select those features of contact networks that impact disease dynamics most strongly. We compute the following topological properties and use them as features to train the Regression Chain Model (RCM).

i. **Average Degree:** Average degree $\bar{k}$ is the global property of the network that determines the speed of transmission of disease in the network [12]. It is computed as $\bar{k} = \frac{1}{N}\sum_{i}^{N} k_i = \frac{2M}{N}$, where k_i, M, and N denote the degree of node i, total edges, and the number of nodes, respectively. *It is established that individuals with a higher average degree have more chances of contracting/transmitting the disease* [6,12,13].

ii. **Normalized Network Density:** Density d is defined as the ratio of the number of edges M over the maximum possible number of edges in the network of N nodes, and is computed as $d = \frac{2M}{N(N-1)}$. Following [14], we compute the normalized network density as $\bar{d} = 1 + \frac{log\, d}{log\, N/2}$ so that the density is comparable across networks of all sizes. *It is established that networks with higher density, favor rapid transmission of the pathogen, and lead to higher number of infected individuals* [6,12,13].

iii. **Degree Variance:** This metric characterizes degree heterogeneity within a network [15]. For a graph of size N, degree variance v is computed as $v = \frac{1}{N}\sum_{i}^{N}(k_i - \bar{k})^2$. *Moreno et al. show that networks with higher heterogeneity in degree cause stronger outbreak incidence* [7]. *In such networks,*

[1] Note that the choice of epidemic model and parameters are disease-specific.

the infection spreads more rapidly in comparison to networks with lower variance in node degree.

iv. **Average Clustering Coefficient:** The clustering coefficient captures the degree to which its neighbors are linked to each other. For a node i with degree k_i, its clustering coefficient is defined as $c_i = \frac{2m_i}{k_i(k_i-1)}$, where m_i represents the number of links between the k_i neighbors of node i. Note that $c_i \in [0,1]$ is a local property of the node. The average clustering coefficient of a graph ($\bar{c}$) is the global property and is computed as $\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i$. *It is established that the clustering coefficient is an important topological characteristic that prevails in human social networks and affects pathogen transmission* [6,12,13,16].

v. **Average Shortest Path Length:** It is defined as the shortest distance averaged over all pairs of nodes in a network. Since the average shortest path length of large networks is computationally expensive, we approximate it as $\bar{p} = \exp\left(\frac{1}{\theta_1\bar{d}+\theta_0}\right)$ using normalized density ($\bar{d}$) as given by [14]. The parameters θ_1 and θ_0 are derived from the regression of $\bar{d}$ and inverse of $log\, p$, where p is the shortest path length of the sampled small networks from each network type. *It is shown that networks with short average path length exhibit fast disease spread* [6,12,13].

2.2 Epidemic Spread on Networks

Different compartmental models in epidemiology have been successfully used to model the spread of numerous contagious diseases, including COVID-19, Ebola, Chikungunya, Measles, etc. [1]. We use the susceptible-exposed-infected-recovered (SEIR) model that addresses the exposed period commonly found in most transmissible diseases. The model assumes immunity to re-infection. The population is divided into compartments, and at each time step of the dynamics, an individual can be in one of four possible states: susceptible (S), exposed (E), infected (I), or recovered (R). An infected person exposes susceptible neighbors with probability β. The exposed individuals transit to the infected state with probability α. Infected persons eventually recover with probability γ. We assume a constant population with a uniform birth and death rate for simplicity.

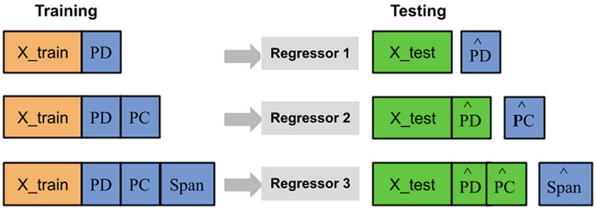


Fig. 1. RCM with the order as peak day (PD), peak cases (PC), and span (Span). Regressor 1 predicts PD, Regressor 2 predicts PC using predicted $\hat{PD}$ and Regressor 3 predicts Span using predicted $\hat{PD}$ and $\hat{PC}$.

2.3 Regression Chain Model

A Regression Chain Model (RCM) is an ensemble built over a chain of regressors to capture dependencies between the output variables (interested reader may refer to [2] for an extensive account on RCM). We set the order for the regression chain as *peak day* (PD), *peak cases* (PC), and *span* (Span). The order was empirically found to deliver the best quality predictions among all possible orders of the three target variables. Figure 1 shows the framework for the Regression Chain Modeling. We use four representative base algorithms (estimators), viz., Decision Tree, Random Forest, Kernel Ridge, and XG Boost. To quantify the assessment of prediction quality, we use two performance metrics. The first metric is the predicted coefficient of determination, $\langle R^2 \rangle$, and the second is the root mean squared error, $\langle RMSE \rangle$, both averaged over the three target variables. The higher value of $\langle R^2 \rangle$ indicates a better fit of the model with higher predictive performance, while the lower value of $\langle RMSE \rangle$ implies higher predictive accuracy.

2.4 Dataset Construction

In the absence of any real dataset for predicting epidemic variables, we curate a rich dataset called *EpiNet*, consisting of five network properties (features) and three epidemic variables (targets to be predicted) for networks with varying sizes (N) and average degrees ($\bar{k}$). We generate 15K small networks ($N \in$ [20K–60K]), 10K medium networks ($N \in$ [60K–150K]), and 10K large networks ($N \in$ [150K–300K]), with equal number of instances belonging to three network families[2] - Random, Small-world, and Community-based networks. We set average degrees for all generated networks in the range [6, 40], as observed in real-life social networks from the SNAP[3] library.

We compute five topological properties mentioned in Sect. 2.1 for each constructed network. Subsequently, we simulate epidemic spread using the SEIR model for specific parameters and note three dependent epidemic variables, viz., peak day (in year), the fraction of infections on the peak day, and the span of the epidemic (in year). To mitigate the effect of stochasticity, we average epidemic variables over *ten* simulation runs of the SEIR spreading process on each network. Hence, we get a feature vector of five network properties and three target variables for each network. Based on the network sizes, we split *EpiNet* into three partitions, corresponding to small networks (D-SN), medium networks (D-MN), and large networks (D-LN). The dataset can be used for further investigation by the research community and is available on GitHub[4].

3 Experiments and Results

This section presents the details of the experiments carried out to examine the feasibility of using RCM as a substitute for expensive simulation of epidemic

[2] We omit scale-free networks as they are inappropriate to study epidemic spread [4].

[3] https://snap.stanford.edu/data/#socnets.

[4] https://github.com/kirtiJain25/EpiNet.

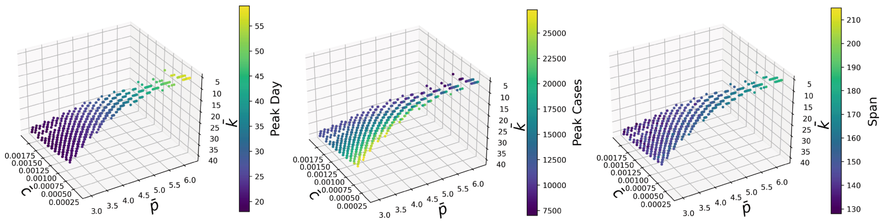
spread on networks for a given set of epidemic parameters. Following Pérez-Ortiz et al., we use $\beta = 0.155, \alpha = 1/5.2$ and $\gamma = 1/12.39$, for SEIR epidemic simulation and obtain the target variables in the training set. In line with our objective, we formulate the following research questions.

i. *How do the network properties influence epidemic variables for the specific epidemic model and epidemic parameters?* (Sect. 3.1)
ii. *Is it possible to predict three epidemic variables with reasonable accuracy using Regression Chain Model?* (Sect. 3.2)
iii. *How sensitive is the performance of RCM to network size?* (Sect. 3.3)

We create the networks using the Igraph library and simulate the SEIR epidemic model in Python (64bits, v3.7.2) on an Intel(R) Core(TM) i7 CPU @1.80 GHz with 16 GB RAM. We train RCM using the Scikit-learn library.

3.1 Influence of Network Properties on Epidemic Variables

We study the relationship between three selected network topological characteristics, viz., average degree, average clustering coefficient, and average shortest path length with three epidemic variables through scatter plots (Fig. 2).



(a) (b) (c)

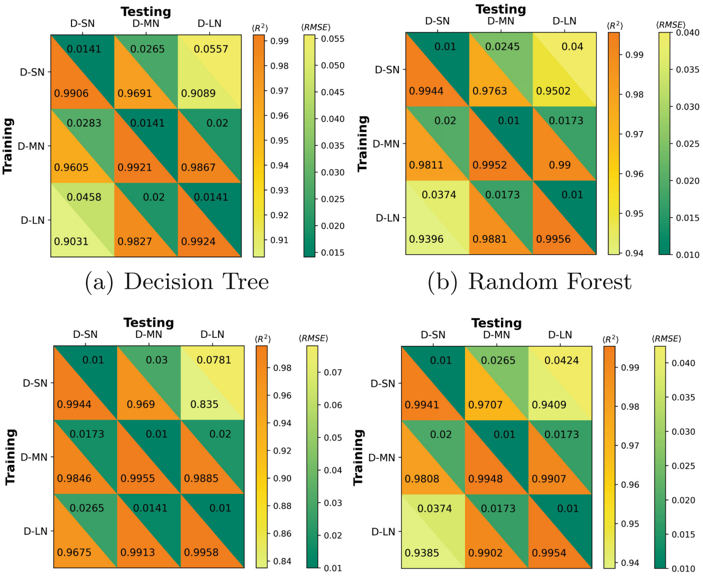
Fig. 2. Effect of three selected network properties, viz., average degree ($\bar{k}$), average clustering coefficient ($\bar{c}$), and average shortest path length ($\bar{p}$) on three epidemic variables, viz., peak day, peak cases, and epidemic span in Figs (a)–(c) respectively.

It is clear from figures (a)–(c) that topological properties have a profound impact on the three epidemic variables. Networks with low average degrees, high average path lengths, and small clustering coefficients have delayed *peak days* with reduced *peak infections* and larger *spans*. In addition, low path length and high average degree fuel the epidemic spread leading to early *peak day* and higher *peak infections*. We also observe that *peak day* has a negative relationship with *peak cases*. Early peak day results in a higher number of cases, and vice versa. On the other hand *peak day* and *span* are positively correlated. The earlier the peak day, the shorter the epidemic duration.

Thus, it is sufficient to conclude that network properties influence epidemic dynamics and are potent to be used as features to predict three epidemic variables.

3.2 Model Validation

We use ten-fold cross-validation method to assess the competence of regression chain models for all three training sets, followed by testing the model on unseen data. Figure 3 shows the heatmap of $\langle R^2 \rangle$ scores and $\langle RMSE \rangle$ values of the RCM using four base algorithms, (a) Decision Tree, (b) Random Forest, (c) Kernel Ridge, and (d) XG Boost. The lower and upper triangles of each cell show $\langle R^2 \rangle$ scores and $\langle RMSE \rangle$ values respectively.



(a) Decision Tree

(b) Random Forest

(c) Kernel Ridge

(d) XG Boost

Fig. 3. Predictive performance of the RCMs trained and tested on topological properties of small (D-SN), medium (D-MN), and large (D-LN) networks. The lower and upper triangles in each cell denote $\langle R^2 \rangle$ scores and $\langle RMSE \rangle$ values, respectively.

It is clear that the cross-validated performance for all regressors (diagonal cells in the heatmap) is high. The non-diagonal cells in the heatmap correspond to the performance of the model on unseen data. We observe that performance degrades marginally, i.e. low $\langle R^2 \rangle$ scores and high $\langle RMSE \rangle$, for the model trained on D-SN and tested on D-LN, and vice versa. The overall high $\langle R^2 \rangle$ scores and low $\langle RMSE \rangle$, in all cases and for all base algorithms demonstrate the competence of the RCMs for predicting epidemic variables in networks.

Figure 4 shows the R^2 and RMSE scores of the model trained on D-SN and tested on D-MN and D-LN. Each metric is computed for three epidemic variables individually. We show the cross-validated performance on D-SN (dark blue colored bars in Fig. 4). We observe high R^2 and low RMSE scores for *peak day*

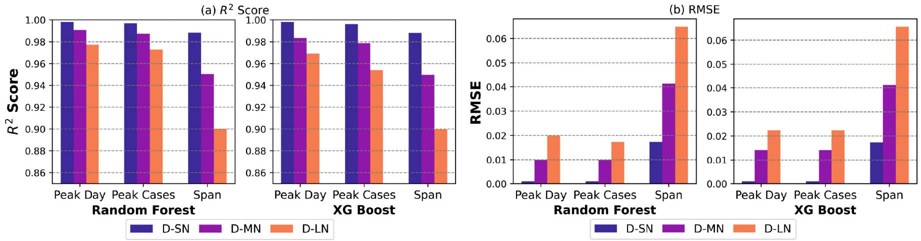


Fig. 4. Predictive performance of three epidemic variables by RCMs trained on D-SN and tested on D-SN, D-MN, and D-LN.

and *peak cases*, while the predicted R^2 score for *span* variable is comparatively low with high RMSE value for medium (D-MN) and large (D-LN) networks (Figs. 4(a) and (b)). This pulls down the overall $\langle R^2 \rangle$ score and raises the $\langle RMSE \rangle$ value when the model trained on small networks is used to predict epidemic variables for larger networks.

We conjecture that the prediction error for the *span* of the epidemic arises due to non-linearity in topological features with increasing network size. Our results motivate further study in this direction to improve predictive models so that all three epidemic variables are predicted accurately without performing costly epidemic simulations on the contact networks. *Nevertheless, it is reasonable to conclude from this experiment that RCMs are capable of predicting epidemic variables with high accuracy on similar-sized networks and may deliver slightly degraded performance on different-sized networks.*

3.3 Sensitivity Analysis

Having observed that the model performance degrades for networks of dis-similar sizes, we examine the sensitivity of the predicted variable to the size of the network. We train the model using D-SN, pool D-MN, and D-LN data sets, and test the model on the pooled test set. We group instances into eight batches (Fig. 5) with approximately equal numbers of records per batch and report batch-wise R^2 scores and RMSE values for Random Forest-based RCM.

We observe a marginal decrease in R^2 scores and a marginal increase in RMSE values for *peak days* and *peak cases* with increasing network size. However, the R^2 score degrades notably for the *span* variable for larger networks ($N > 120K$). This observation ratifies our earlier observation that the regression chain model trained using topological properties of small networks (D-SN) is capable of reliably predicting *peak day* and *peak cases* for medium and large networks, thereby saving computational expense incurred by epidemic simulations. However, the prediction of the *span* variable is not accurate. This is because the model is unable to capture the relationship between the network features and the duration of the epidemic spread.

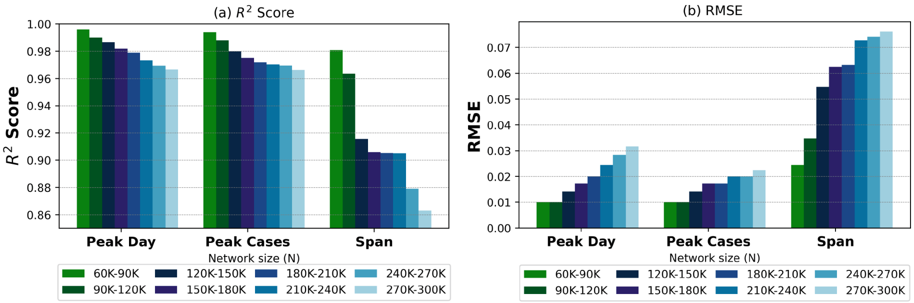


Fig. 5. Predictive accuracy for three epidemic variables for Random Forest-based regressor trained on small networks (D-SN) and tested on larger networks (D-MN and D-LN) grouped by network size.

We conclude that basic network characteristics are insufficient for accurately predicting the span variable, and further study is required to understand the role of other topological properties on the span of the epidemic.

3.4 Discussion

Our empirical study shows promising results and indicates that inexpensive models trained on small networks can predict two epidemic variables with reasonably high accuracy. Prediction of the third variable (epidemic span) is challenging. However, certain important issues must be noted before using this approach.

i. Sensitivity of the method to population size necessitates curating training data sets for different network sizes for accurate predictions. Models trained on networks of vastly dissimilar sizes may deliver inaccurate predictions.
ii. Since the constructed data set is disease-specific, the training set needs to be curated using the appropriate epidemic spread model and its parameters.
iii. Several observable and unobserved variables influence the epidemic spread in the complex landscape of social, political, and economic realities in the real world. To account for these factors, the simulation of the epidemic spread needs to be tweaked accordingly. Our simulations in this study do not account for any external factor and hence offer the *best-case* results.

We believe that this work will help advance the study of recognizing machine learning models that proxy for expensive network-based simulation.

4 Conclusion

In this research, we examine the possibility of predicting three epidemic variables using the regression chain model (RCM). We curate a rich data set called *EpiNet*, consisting of five network properties (features) and three epidemic variables (targets captured using SEIR epidemic model) for 35K networks of varying types and sizes. The dataset is split into three partitions (small - D-SN, medium - D-MN, large - D-LN), and we train RCM using four popular regressors.

Our results establish the possibility of predicting three epidemic variables, viz. *peak day*, *peak cases*, and *span*, using a Regression Chain Model trained on the topological properties of the underlying contact networks as a substitute for costly epidemic simulations on large networks. Detailed analysis of the predicted variables reveals that prediction accuracy for the *span* variable is lower compared to that of *peak days* and *peak cases*. Further study is warranted to understand the additional topological characteristics required for its accurate prediction.

References

1. Barabási, A.L.: Network Science Book. Cambridge University Press, Cambridge (2014). http://barabasi.com/networksciencebook
2. Borchani, H., Varando, G., Bielza, C., Larranaga, P.: A survey on multi-output regression. Wiley Interdisc. Rev. Data Mining Knowl. Disc. **5**(5), 216–233 (2015)
3. Bucur, D., Holme, P.: Beyond ranking nodes: predicting epidemic outbreak sizes by network centralities. PLoS Comput. Biol. **16**(7), 1–20 (2020)
4. Du, M.: Contact tracing as a measure to combat covid-19 and other infectious diseases. Am. J. Infect. Control **50**(6), 638–644 (2022)
5. Hethcote, H.W.: The basic epidemiology models: models, expressions for R0, parameter estimation and applications, vol. 16, chap. 1, pp. 1–61. World Scientific Publishing, Singapore (2008)
6. Keeling, M.J., Eames, K.T.D.: Networks and epidemic models. J. Royal Soc. Interface **2**(4), 295–307 (2005)
7. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. Eur. Phys. J. B-Cond. Matter Complex Syst. **26**(4), 521–529 (2002)
8. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
9. Newman, M.E.: Spread of epidemic disease on networks. Phys. Rev. E **66**(1), 016128 (2002)
10. Pérez-Ortiz, et al.: Network topological determinants of pathogen spread. Sci. Rep. **12**(1), 1–13 (2022)
11. Rodrigues, F.A., et al.: A machine learning approach to predicting dynamical observables from network structure. arXiv preprint arXiv:1910.00544 (2019)
12. Shirley, M.D., Rushton, S.P.: The impacts of network topology on disease spread. Ecol. Complex. **2**(3), 287–299 (2005)
13. Small, M., Cavanagh, D.: Modelling strong control measures for epidemic propagation with networks - a COVID-19 case study. IEEE Access **8**, 109719–109731 (2020)
14. Smith, R.D.: Average path length in complex networks: patterns and predictions. arXiv preprint arXiv:0710.2947 (2007)
15. Snijders, T.A.: The degree variance: an index of graph heterogeneity. Social Netw. **3**(3), 163–174 (1981)
16. Volz, E.M., et al.: Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput. Biol. **7**, e1002042 (2011)