Understanding the Impact of Network Structure on Propagation Dynamics based on Mobile Big Data

Yuanfang Chen†§, Lei Shu§, Noel Crespi†, Gyu Myoung Lee‡, Mohsen Guizani∗

†Institut Mines-Télécom, Télécom SudParis, France
§Guangdong University of Petrochemical Technology, China
‡Liverpool John Moores University, Liverpool, UK
∗Qatar University, Qatar
Email: yuanfang.chen.2009@ieee.org, lei.shu@ieee.org, noel.crespi@mines-telecom.fr, G.M.Lee@ljmu.ac.uk, mguizani@ieee.org

***Abstract*—Understanding the propagation dynamics of information/an epidemic on complex networks is very important for discovering and controlling a terrorist attack, and even for predicting a disease outbreak. As an effective method, with analyzing the structure of a propagation network, a large number of previous studies have analyzed the propagation dynamics. Most of these studies are based on a special network structure to make such analysis. However, a propagation network has dynamically changed structure during the propagation. How to track, recognize and model such dynamic change is a big challenge. Along with the popularity of smart devices and the rapid development of the Internet of Things (IoT), massive mobile data is automatically collected. In this article, as a typical use case, we investigate the impact of network structure on epidemic propagation dynamics by analyzing the massive mobile data collected from smart devices carried by the volunteers of Ebola outbreak areas. From this investigation, we obtain two observations. Based on these observations and the analytical ability of Apache Spark on streaming data and graphs, we propose a simple model to track and recognize the dynamic structure of a network. Moreover, we introduce and discuss open issues and future work for developing this proposed recognition model.**

***Keywords*—Network structure, propagation dynamics, mobile big data, Internet of Things.**

I. Introduction

Information/epidemic propagation dynamics [1], [2], [3], [4] has been extensively studied by network-enabled science, e.g., graph theory, network theory, and probability theory. When information/epidemic propagation is modelled over networks, it is usual to assume that the propagation has the same probability over each link. Even if different links have respective propagation probabilities, such modelling is not enough to reflect the real propagation pattern in the physical world. As the important feature of networks, network structure needs to be considered [5], because the patterns of propagation are different in different network structures. Most of studies use a special network structure to make such modelling, and even analysis, for example, the propagation network of an epidemic is best described as having exponential degree distribution. However, a propagation network has dynamically changed structure along with the propagation of information/an epidemic.

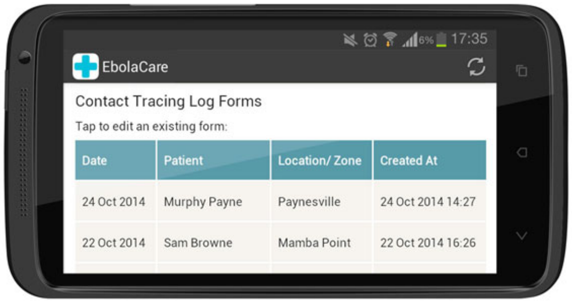
On the basis of the above description, firstly, we need to figure out: whether the structure of complex networks impacts the propagation dynamics of information/an epidemic [6]. It is an open issue. Despite a lack of direct experimental evidence to support such "structure-propagation" hypothesis, a number of theoretical studies have shown that the topological structure of complex networks (mostly scale-free and small-world topologies) leads to markedly different propagation dynamics compared with the predicted by standard propagation models. For example, in the literature [7], Michael Small *et al.* examine the global spatio-temporal distribution of avian influenza cases in both wild and domestic birds, and they find that the cases and the links between these cases during an outbreak form a scale-free network. It means that such an avian influenza outbreak will continue to propagate even with a very small propagation rate. In contrast, by standard mathematical models of disease propagation [8], the propagation of this avian influenza has been controlled and even has stopped. This may cause to miss the best time of vaccination, and thus may increase the probability of another outbreak.

Then, based on the above explanation, understanding the impact of network structure on propagation dynamics is very important, and recognizing the dynamic structure of a network is a gap in the previous studies of propagation dynamics.

This article reviews the advance of propagation dynamics, and then as a typical use case, we investigate the impact of network structure on epidemic propagation dynamics, by analyzing the massive mobile data collected from the GPS-enabled wireless devices carried by volunteers (Fig. 1 illustrates an example). On this basis, we propose a model to recognize the dynamic structure of a network. Finally, open issues and future work are provided and discussed for developing this model.

In summary, the scientific contributions of this article are listed as follows:

- The impact of network structure on epidemic propagation dynamics is investigated, by analyzing the massive mobile data collected from the wireless devices carried by volunteers.
- By the investigation, we obtain two observations which are the motivation to design the recognition model of dynamic structure.



EbolaCare

Contact Tracing Log Forms

Tap to edit an existing form:

Date	Patient	Location/ Zone	Created At
24 Oct 2014	Murphy Payne	Paynesville	24 Oct 2014 14:27
22 Oct 2014	Sam Browne	Mamba Point	22 Oct 2014 16:26

Fig. 1: Contact Tracing module of Ebola Care [9]. It can track each individual who contacts with a sick Ebola patient. The collected data by this application is shared with the World Health Organization (WHO). WHO is using information from hundreds of aid organizations to make big strategic decisions.

- A streaming data and graph based model is designed to recognize the dynamic structure of a network. A propagation network can be formulated as a dynamic graph with processing massive streaming data. The streaming data processing is an important research issue in Big Data analytics as well.

This article is structured as follows. The advance of propagation dynamics is introduced in Section II. As a typical use case, in Section III, we investigate the impact of network structure on epidemic propagation dynamics based on the propagation network of the Ebola outbreak in 2014. On the basis of this investigation, Section IV proposes a recognition model of network structure. This model is designed to recognize the dynamic structure of a propagation network. For developing this model to practical applications, in Section V, we present the open issues and future work. This article is concluded in Section VI.

II. Advance of Propagation Dynamics

It is important to understand the propagation processes arising over the networks with different structures. For example, in knowledge mining, how a behaviour on a specially structured network to impact the nodes of the network, is worth understanding. Such understanding is helpful to model the behaviour as well.

In recent years, there is an increased effort to study propagation dynamics based on a variety of complex networks. Recent achievements can be divided into two categories based on different types of networks:

- Propagation dynamics on social networks [10]. On such networks, information is the main research target. Exponential and power-law models that reflect network structure have been widely used to model the dynamics of information propagation.
- Propagation dynamics on contact networks [11]. A contact network describes the real relationships among individuals/ecosystems in the physical world. Based on the real relationships from the physical world, the propagation dynamics on contact networks is different from the propagation dynamics on social networks. With the development of IoT (Internet of Things) and the help of various sensors and wireless devices, some researchers have paid their attention to this propagation dynamics, and have obtained some achievements in: (i) the propagation of infectious diseases, and (ii) the propagation of contaminants. Analyzing and studying the dynamics of propagation among individuals/ecosystems can help us understand and control the dynamic behaviours on these real networks.

As an important aspect of propagation dynamics, theoretical studies on the "structure-propagation" hypothesis are classified into two classes: information-related and epidemic-related propagation dynamics on respective complex networks.

Information-related propagation dynamics. As important recent achievements in information-related propagation dynamics [1], Jure Leskovec *et al.* have obtained three interesting observations, along with tracking information propagation among media sites and blogs: (i) The information pathways for general recurrent topics are more stable across time than for on-going news events. It means that the former has a more stable network structure; (ii) clusters of news media sites and blogs often emerge and vanish in a matter of days for on-going news events. From this observation, we can acquire that hub nodes (clusters) exist in an information propagation network. As a key element to reflect a network structure, clusters are dynamically varying over time, and different information propagation networks have different clusters; (iii) major events, for example, large-scale civil unrest such as Libyan civil wars and the Syrian uprising, increase the number of information pathways among blogs, and also increase the network centrality of blogs and social media sites.

Epidemic-related propagation dynamics. As a recent achievement in epidemic-related propagation dynamics [12], Louis Kim *et al.* propose a parameter estimation method by learning network characteristics and disease dynamics. This method is applied to the data collected during the 2009 H1N1 epidemic. On this basis, they find the outbreak network is best fitted into a scale-free network. This finding implies that random vaccination alone will not efficiently stop the propagation of influenza, and instead vaccination should be based on understanding the propagation dynamics of the epidemic with exploiting the special structure of an outbreak network.

However, network structure is time-varying along with information/epidemic propagation on the network. It is necessary to recognize the dynamic structure of such a network.

III. Impact of Network Structure

As a typical use case, the impact of network structure on epidemic propagation dynamics is investigated based on the propagation network of the Ebola outbreak in 2014.

By using the wireless communication devices carried by the volunteers of epidemic areas, new cases (with corresponding locations and the relationships between these cases) are reported to a control center. Then, these reported cases with corresponding locations construct a propagation network (an example is illustrated in Fig. 2). During an epidemic, the network is time-varying along with the propagation of an infectious disease. The propagation network can be modelled as a dynamic graph G_t. The weight $w_{\{i,j\}}$ is the transmission probability ($p_{\{i,j\}}$) of a disease from vertex i to vertex j (on the corresponding edge $e_{\{i,j\}}$).

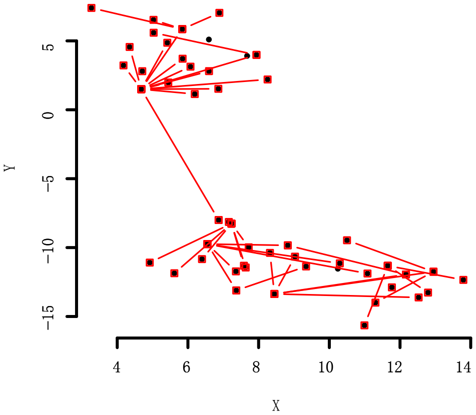


Fig. 2: An example of a propagation network during an epidemic. This example displays 50 cases and their relationships (contact). X and Y are only used to denote the relative locations of cases (no units). These cases come from three typical countries and seven regions of the Ebola outbreak in 2014. Three countries are: Guinea, Nigeria and Liberia. Seven regions are: Gueckedou, Macenta, Kissidougou, Conakry, Monrovia, Lagos and Port Harcourt. The black nodes of this network are cases (suspected and confirmed), and if there is an edge between two nodes, it means that there is contact between the corresponding individuals of the two cases.

Outbreak Data. The outbreak data of the Ebola in West Africa from March 2014 is used as real surveillance data to analyze the impact of network structure on propagation dynamics.

As the latest disease outbreak, until February 15, 2015, Ebola has killed 9380 people, and the total cases have reached 23253. Researchers generally believe that from a two-year-old boy of Guinea to his mother, sister and grandmother (a propagation network), Ebola rapidly spreads in West Africa from March 2014.

The reported Ebola cases with time series and location information are collected by the World Health Organization (WHO), as well as the ministries of health of epidemic countries. In this study, we select part of the data from three typical outbreak countries, Guinea, Nigeria and Liberia. Guinea is the source of this outbreak, and has relatively high quantity of confirmed cases (2727, as of February 15, 2015), and Nigeria is far away from the source of the outbreak, and has relatively low quantity of confirmed cases (19, as of February 15, 2015), and Liberia is close to the source of the outbreak, and has high quantity of confirmed cases (3149, as of February 15, 2015). In addition, the seven regions of these three countries are: Gueckedou of Guinea, Macenta of Guinea, Kissidougou of Guinea, Conakry of Guinea, Monrovia of Liberia, Lagos of Nigeria, and Port Harcourt of Nigeria. Moreover, these variables are included in the outbreak data:

1) Case ID. A unique number indicates a case.
2) Source ID. Source identification indicates the source of an infection for a case.
3) Date. It is the date that a case is reported.
4) Location. It indicates the coordinates (longitude and latitude) of a reported case.

Investigation and Results.

Investigation. In this article, we analyze the degree distribution[1] of the propagation network that is constructed by the collected Ebola outbreak data. Such analyses are completed by:
(i) Conducting the Maximum-Likelihood Fitting (MLF) to fit the calculated degree distribution of the propagation network into exponential, normal, poisson and power-law distributions.

Definition 1 (Maximum-Likelihood Fitting)**.** *The Maximum-Likelihood Fitting (MLF) for a set of data points* $\theta = \{\theta_1, ..., \theta_m\}$ *is the method that maximizes* $lik(\theta)$*, where* $lik(\theta)$ *is the probability of observing the given data as a function of* θ*.*

For example, the data points $\theta = \{\theta_1, ..., \theta_m\}$ are the degrees of nodes, and $lik(\theta) \Rightarrow lik(exp(\theta))$ is the probability that the data points conform to the exponential distribution.

(ii) Calculating and comparing the estimated standard deviations and the estimated variance-covariance matrices of these fittings.

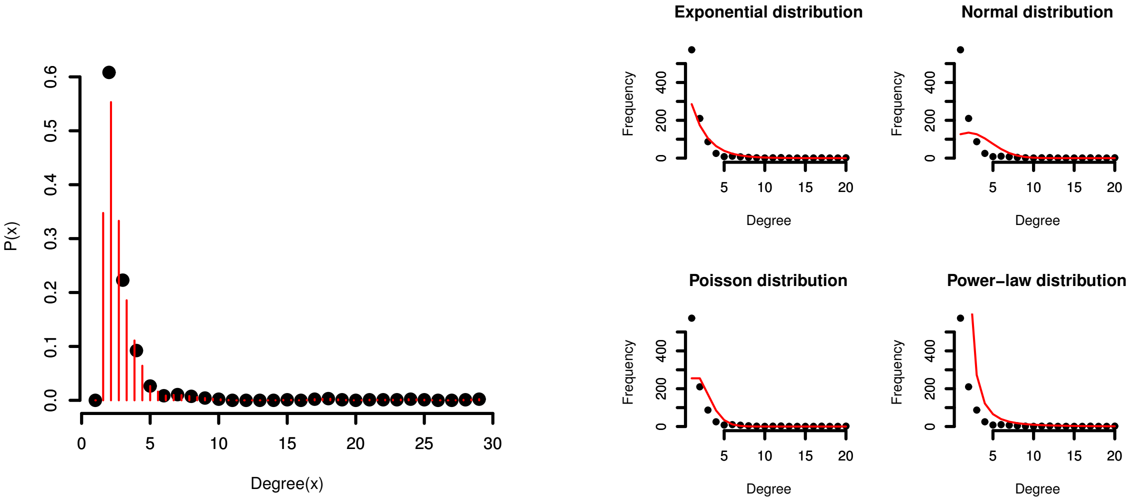
Figure 3a illustrates the degree distribution of the propagation network constructed by Ebola outbreak data. And then the maximum-likelihood fitting is conducted to fit the calculated degree distribution into exponential, normal, poisson and power-law distributions. Finally, the estimated standard deviations and the estimated variance-covariance matrices of these fittings are measured to quantify: "**how many differences between two different distributions**". The results of these fittings are illustrated in Fig. 3b.

By the maximum-likelihood fitting, we can fit the calculated degree distribution into exponential, normal, poisson, and power-law distributions, and the fitted parameter values for these distributions are listed as follows:
(i) rate parameter $\lambda = 0.50159915$ for the exponential distribution.
(ii) $\mu = 1.99362380$ and $\sigma = 2.77914691$ for the normal distribution.
(iii) $\lambda = 1.9936238$ for the poisson distribution.
(iv) $x_{min} = 2$ and $\alpha = 2.803973$ for the power-law distribution.

Table I provides the estimated standard deviations and the estimated variance-covariance matrices for the parameter values of these fittings.

Comparing the estimated standard deviations and estimated variance-covariance matrices listed in Tab. I, the minimum standard deviation is 0.01635166. This minimum standard deviation is corresponding to the exponential distribution with the rate parameter $\lambda = 0.50159915$. This result indicates that the

[1] Degree distribution is the basic and most important structure knowledge of a network. It is the probability distribution of degrees over the propagation network.



(a) Degree distribution of the propagation network constructed by Ebola outbreak data. There are 942 nodes and 938 edges in this network. The black spots are the probability distribution of nodes' degrees.

(b) The maximum-likelihood fitting of degree distributions. The degree distribution of the constructed propagation network is fitted into exponential, normal, poisson and power-law distributions by the maximum-likelihood fitting. The black spots indicate the calculated probability distribution of nodes' degrees, and the red lines are the corresponding fittings for exponential, normal, poisson and power-law distributions.

Fig. 3: Degree distribution and the maximum-likelihood fitting for the propagation network of the Ebola outbreak

TABLE I: Estimated standard deviations and estimated variance-covariance matrices

Distribution	Standard deviation	Variance-covariance matrix
Exponential	λ (rate parameter): 0.01635166	$\begin{array}{c\|c} & \lambda \text{ (rate parameter)} \\ \hline \lambda \text{ (rate parameter)} & 2.673769e-04 \end{array}$
Normal	μ (mean): 0.09059760, σ (standard deviation (sd)): 0.06406218	$\begin{array}{c\|c\|c} & \mu & \sigma \\ \hline \mu & 0.008207925 & 0.000000000 \\ \hline \sigma & 0.000000000 & 0.004103963 \end{array}$
Poisson	λ (lambda): 0.0460285	$\begin{array}{c\|c} & \lambda \text{ (lambda)} \\ \hline \lambda \text{ (lambda)} & 0.002118623 \end{array}$
Power-law	$x_{min} + \alpha$: 0.03831463	NULL

degree distribution of the propagation network is approximate to the exponential distribution with $\lambda = 0.50159915$.

However, based on the description of the network that is studied in this article, the propagation network is time-varying along with the propagation of an infectious disease. As an example, the analytical results of the subnetwork constructed by 96 time periods of August 26th, 2014, are illustrated in Fig. 4 and Tab. II.

By the maximum-likelihood fitting for the subnetwork, the results of the parameter estimation for different distributions are listed as follows: (i) the rate parameter $\lambda = 0.74796748$ for the exponential distribution, (ii) $\mu = 1.33695652$ and $\sigma = 1.00841216$ for the normal distribution, (iii) $\lambda = 1.33695652$ for the poisson distribution, and (iv) $x_{min} = 1$ and $\alpha = 3.041947$ for the power-law distribution.

Table II provides the estimated standard deviations and the estimated variance-covariance matrices of the above fittings.

By observing the fittings and the estimated results, the degree distribution of the subnetwork is approximate to the power-law distribution with $x_{min} = 1$ and $\alpha = 3.041947$.

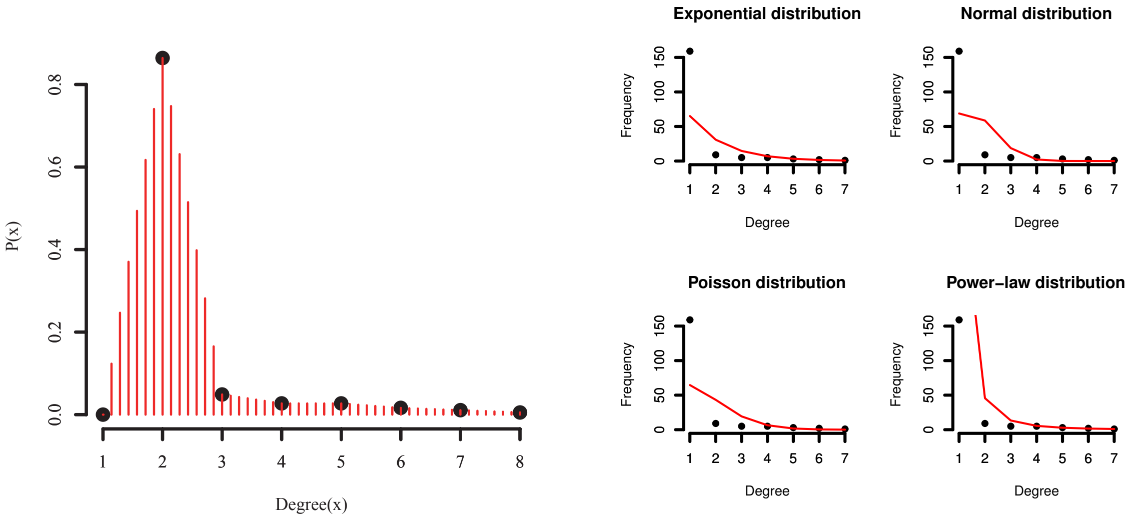
Results. Based on the above detailed analyses on the structure of the networks, we obtain these observations:

- In Fig. 3a and Fig. 4a, the probabilities (P(x)) of 1 degree are both zero. It means that the suspected or confirmed cases at least contact with two other cases. It is an important feature of network structure, but the maximum-likelihood fitting to the four kinds of degree distributions fails to capture it.
- Network structure is time-varying during an epidemic.

IV. Recognition Model

How to accurately recognize the dynamic structure of a propagation network is a valuable research issue.

During an epidemic, a disease propagates along a propagation network, and such propagation makes the structure of the propagation network be dynamically changed. By recognizing the dynamic structure of a propagation network, we can acquire the propagation dynamics of a disease. Moreover, it is important to quantify and predict the propagation dynamics during an epidemic. If the quantification and prediction can be achieved for a disease outbreak, it will be helpful to allocate



(a) Degree distribution for the subnetwork of the propagation network. There are 96 time periods of August 26th, 2014 in this subnetwork. The black spots are the probability distribution of nodes' degrees.

(b) The maximum-likelihood fitting of degree distributions. The degree distribution of the subnetwork is fitted into exponential, normal, poisson and power-law distributions by the maximum-likelihood fitting. The black spots indicate the probability distribution of nodes' degrees, and the red lines are the corresponding fittings for exponential, normal, poisson and power-law distributions.

Fig. 4: Degree distribution and the maximum-likelihood fitting for the subnetwork of the propagation network

TABLE II: Estimated standard deviations and estimated variance-covariance matrices

Distribution	Standard deviation	Variance-covariance matrix		
Exponential	λ (rate parameter): 0.05514089		λ (rate parameter)	
		λ (rate parameter)	0.003040518	
Normal	μ (mean): 0.07434113, σ (standard deviation (sd)): 0.05256712		μ	σ
		μ	0.005526604	0.000000000
		σ	0.000000000	0.002763302
Poisson	λ (lambda): 0.08524123		λ (lambda)	
		λ (lambda)	0.007266068	
Power-law	$x_{min} + \alpha$: 0.02865438	NULL		

public health resources and respond to public health events, accurately and duly.

Based on the analytical ability of Apache Spark [13] on streaming data and graphs, we propose a recognition model of network structure. The work flow of this model is illustrated in Fig. 5.

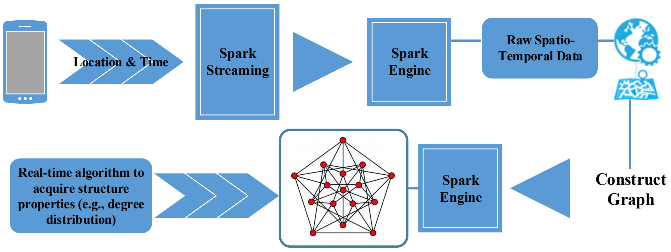


Fig. 5: Work flow of our recognition model.

In this model, there are three main parts:

1) The input stream is the cases' spatio-temporal data with GPS location information and time stamps. The GPS location of a case is associated with the physical location where the case is found and reported.
2) On the basis of the input stream, the cases and their relationships are used to construct a graph. In this graph, each vertex corresponds to a case, and the distance between two vertices can be calculated by the coordinates of corresponding two cases. Moreover, the graph is time-varying: the vertex and edge sets are changed over time, along with the propagation of a disease.
   By the processing of Spark Streaming [2], and based on the time stamps of cases, the dynamic change of the graph can be tracked.
3) Based on the graph tracking and a real-time algorithm, corresponding structure properties (e.g., degree distribution) can be calculated for this dynamic graph.

[2] Spark Streaming provides a language-integrated API to stream processing. It makes the processing of streaming data be easy as processing batch data.

Two components of Apache Spark are used to construct and process the dynamic graph, Graphx [14] and MLlib (Machine Learning Library) [15]. Such processing is conducted: (i) tracking the dynamic change of the graph, and (ii) calculating the structure properties of the graph.

1) Tracking the dynamic change of a graph. For achieving the dynamic tracking, a graph-parallel system, PowerGraph [16] is used. PowerGraph has been implemented using the GraphX interface. It consists of three phases: Gather → Apply → Scatter (GAS). During the *gather* phase, the *gather* and *sum* functions are used to collect the neighbour information of an active vertex i. In the *apply* phase, the output of the *gather* phase is consumed along with updating the information of vertex i. The *scatter* phase uses the new information of vertex i to update the information on adjacent edges. If a vertex is not activated in the *gather* phase, then this vertex is skipped in the subsequent phases.
2) Calculating the structure properties of a dynamic graph. As an important structure property, we provide an example on how to calculate the degree distribution of a graph. In GraphX, we can get the degree of each vertex by calling the degree method. This method returns a value, *VertexRDD*, which is the degree of each vertex. On this basis, the Estimator of MLlib is used to fit these degrees to produce a model. An Estimator implements a method *fit(.)*. Based on this fitting, we can understand how these degrees are distributed, for example, a logistic regression model. This logistic regression measures the cumulative logistic distribution of these degrees.

V. Open Issues and Future Work

The above proposed recognition model is based on Spark to process and analyze streaming data and the data-based graph. For developing this model to practical applications, these open issues are worth studying as future work:

- Constructing a graph based on spatio-temporal data, with satisfying specific requirements. Once the spatio-temporal data is persistently input, Spark can process it by Spark Streaming, and further through the graph algorithms of Spark Engine, a dynamic graph is constructed. From "data" to "graph", there is an important process, "transformation". To a certain problem, such a process has specific requirements, for example, a suitable period of time needs to be decided to construct the graph that is used to denote the real-time propagation network of an epidemic.
- Real-time algorithm design to track the dynamic change of a graph, based on a modified function. The function can be designed based on different tracking requirements on structure properties.
  Spark is designed for big data analytics. It means that Spark is a fast and general engine for large-scale data processing. In this engine, the processing is based on parallelization technology, so traditional serial algorithms are not suitably used in here. Based on the real-time and parallel algorithm, the dynamic change of a graph can be captured, and the structure properties of the graph can be calculated.

VI. Conclusion

By investigating the impact of network structure on epidemic propagation dynamics, this article has obtained two observations, as the basis of the model design to recognize the dynamic structure of a network. This model is based on the ability of Spark to process and analyze streaming data and the data-based graph. This model can calculate the structure properties of a dynamic network. Based on this calculation, the structure of a network can be recognized in real time. Moreover, we have introduced and discussed the open issues of developing such a model to practical applications.

References

[1] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. SCHÖLKOPF, "Uncovering the structure and temporal dynamics of information propagation," *Network Science*, vol. 2, no. 01, pp. 26–65, 2014.

[2] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modeling propagation dynamics of social network worms," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 8, pp. 1633–1643, 2013.

[3] G. Theodorakopoulos, J.-Y. Le Boudec, and J. S. Baras, "Selfish response to epidemic propagation," *Automatic Control, IEEE Transactions on*, vol. 58, no. 2, pp. 363–376, 2013.

[4] X. Fu, M. Small, and G. Chen, *Propagation dynamics on complex networks: models, methods and stability analysis*. John Wiley & Sons, 2013.

[5] J. Guo, H. Guo, and Z. Wang, "An activation force-based affinity measure for analyzing complex networks," *Scientific reports*, vol. 1, 2011.

[6] S. Pei and H. A. Makse, "Spreading dynamics in complex networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, no. 12, p. P12002, 2013.

[7] M. Small, D. M. Walker, and C. K. Tse, "Scale-free distribution of avian influenza outbreaks," *Physical review letters*, vol. 99, no. 18, p. 188702, 2007.

[8] F. Brauer *et al.*, *Mathematical models for communicable diseases*. SIAM, 2012, vol. 84.

[9] "Ebola care app," http://www.appsagainstebola.org/#the-app.

[10] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1059–1068.

[11] G. M. Vazquez-Prokopec, D. Bisanzio, S. T. Stoddard, V. Paz-Soldan, A. C. Morrison, J. P. Elder, J. Ramirez-Paredes, E. S. Halsey, T. J. Kochel, T. W. Scott *et al.*, "Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment," *PloS one*, vol. 8, no. 4, p. e58802, 2013.

[12] L. Kim, M. Abramson, K. Drakopoulos, S. Kolitz, and A. Ozdaglar, "Estimating social network structure and propagation dynamics for an infectious disease," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2014, pp. 85–93.

[13] "Apache spark," http://spark.apache.org/.

[14] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica, "Graphx: Graph processing in a distributed dataflow framework," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 599–613.

[15] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen *et al.*, "Mllib: Machine learning in apache spark," *arXiv preprint arXiv:1505.06807*, 2015.

[16] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin, "Powergraph: Distributed graph-parallel computation on natural graphs," in *Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*, 2012, pp. 17–30.