Modeling Multi-Dimensional Datasets via a Fast Scale-Free Network Model

Shaojie Min  
College of Computer Science  
Chongqing University  
alexmin@cqu.edu.cn

Ji Liu  
College of Computer Science  
Chongqing University  
liujiboy@cqu.edu.cn

Abstract

Compared with network datasets, multi-dimensional data are much more common nowadays. If we can model multi-dimensional datasets into networks with accurate network properties, while, in the meantime, preserving the original dataset features, we can not only explore the dataset dynamic but also acquire abundant synthetic network data. This paper proposed a fast scale-free network model for large-scale multi-dimensional data not limited to the network domain. The proposed network model is dynamic and able to generate scale-free graphs within linear time regardless of the scale or field of the modeled dataset. We further argued that in a dynamic network where edge-generation probability represents influence, as the network evolves, that influence also decays. We demonstrated how this influence decay phenomenon is reflected in our model and provided a case study using the Global Terrorism Database.

Introduction

Real-life networks are often complex and hard to analyze. However, thanks to the discovery of common properties and similar patterns of networks in different fields [1] [2] [3], generative network models became not only possible but also demanding.

The network structure is vital to aspects such as information diffusion [4], community cohesion [5], knowledge pooling [6], and the world interaction phenomenon. If we can guarantee that simulated graphs are, in certain aspects, similar to real-life networks, we will be able to model and analyze the dynamics of real-world networks. For instance, Barabasi and Albert’s work in 1999, revealed the internal causality of scale-free network connections. The emergence of the Configuration Model made it possible for the generation of a synthetic graph with any given edge distribution.
Since the 1960s, plenty of complex network models have been proposed with real-life like properties including degree distribution, clustering coefficient, and average path length. But the problem of integrating those network models with real-life data remains.

Although different fields of data take a variety of focus, most of them come in the form of multi-dimensional data with individuals and their attributes. This is nowhere near the node-edge data form in a network. A network describes the relationship among its participants whereas a multi-dimensional dataset describes the properties of its included samples. As a consequence, a network aiming at modeling real-life data should find its way to establish connections using the individuals’ properties, while, in the meantime, preserving the common features of complex networks. In addition, the model should also be dynamic, cost-efficient, and sustainable to change of the input dataset because data in different fields are ever-changing, large-scale, and takes different types of properties nowadays. Some real-life data has been systematically investigated via network models [7] [8] [9] [10] [11], however, those works have been confined to a single dataset or a limited field. Jebran et al. brought up a more generative modeling method – OSN [12], in which edge generation of a graph is under the control of the dataset samples’ similarities. Despite the good generality and flexibility of OSN, some problem still remains, such as the time efficiency of generation and the incompatibility between node behavior and network dynamic, in which the node behavior we will describe in particular as influence decay.

In this paper, we proposed FSM, a Fast Generic Scale-free Network for Modeling Multi-dimensional Data. The model transforms real-life data into nodes with order, influence, and geographic attributes and the model dynamics are under the control of both node attributes and a neighbor resampling function. Prescribed model has a $O(n)$ time complexity regardless of the scale or field of input data and the network generated can both, from a global perspective, follows a scale-free degree distribution and, from a local perspective, mimic the influence decay of individuals. This is essential for modeling individuals in a network system because we argue that with the edge-generation probability representing influence, the possibility of forming connections, that influence ought to fade away as the system evolves. We further explore the usage of our model with the Global Terrorism Database (GTD) [13], and demonstrate how the influence decay is reflected in our network with Global Terrorism Database (GTD) by focusing on the 9-11 terrorist attack.

**Related Works**

Network models focusing solely on the properties and topology or with hybrid inputs have both drawn extensive research over recent decades. However, few genetic methodologies for modeling multi-dimensional data into a network have been proposed.

**Property-Based-Solely Network Models**

Properties and topology of a network, such as centralities, clusters, path length, and degree distribution, have always played a vital part in network generation. The *ER random network*
proposed by Erdős and Rényi [14] was able to simulate similar average path lengths and giant components corresponding to real-life networks while preserving the randomness of a complex network. Compare with the undirect ER random graph, *Binary Relation Model* [15], also known as the *pl model*, expend the network to a direct graph with a stochastic function on the assumption of statistically dyadic independence [16]. High clustering coefficient is also an important property of a network as the high transitivity among relationships. *Small-World Network*, a model with a high clustering coefficient, was put forward by Watts and Strogatz focusing on the small-world phenomenon [3]. Those early explorations have successfully modeled some properties of real-life networks, however, leaving the degree distribution untended.

Degree distributions of large-scale networks have been observed to follow the power law distribution [2]. Nevertheless, the previously mentioned *ER random graph* with a fixed edge forming probability possesses a binomial degree distribution, and for the *Small-World Network*, most nodes have similar degrees. Network models with power-law degree distribution are referred to as scale-free networks. The very first scale-free network model was proposed by Barabasi and Albert in 1999 [2]. The edge generation probability of the proposed *Preferential Attachment Model (BA Model)* is under the control of node degree, resulting in a phenomenon of “rich get richer” which ultimately leads to a power law degree distribution. The *fitness-Weighted Preferential Attachment Model*, proposed by Juan and Jorge [17], extended the *BA Model* by considering the edge number of new nodes as a random variable rather than a constant.

The above-mentioned models’ dynamics are solely under the restraint of network properties. For example, edge generation of *Small-World Network* is based on the topology among nodes, and all forms of the *Preferential Attachment Model* depend heavily on the degree of nodes. The properties of networks are essential, however, as those networks evolve, the generation-dependent properties also alter, increasing the complexity of network generation. Compared with depending solely on the properties of network, several network models with hybrid inputs were put forward.

**Network Generation with Hybrid Inputs**

We refer to network models, whose evolution is not only dependent on their own network properties, as networks with hybrid inputs. A simple example of a hybrid-input network is the *Configuration Model* [18]. It takes a sample sequence as input and outputs a graph with a degree distribution corresponding to that given sequence. *DSNG-M* uses provided community partitioning and an initial graph to generate a dynamic graph [19]. A more complex work, *NetGAN* [20], introduced sequential characteristics of the graph itself into the process of edge generation.

Fragkiskos and Maksim argue in *OSN* that similarity also plays an important role in network evolution [20]. The similarities among nodes are calculated in *OSN* via samples’ demography like age, gender, profession, and location information in a real-life dataset [12], then the network evolution is based on those similarities alongside with preferential attachment.

**Data Modelling via Complex Networks**
Wen et al argued that the ultimate goal of the data-driven network model is to create a synthetic network as the *Digital Twin* of reality [21]. The generated network and original data should have interrelations and be able to reflect some properties of one another. Data from a variety of research fields have been investigated using network models [9, 22-25]. The above-mentioned OSN was a practice of modeling social networks by converting users with attributes into nodes with similarities. In the field of Mobile instant messaging (MIM) systems, Ebrahim and Behrouz proposed *social communication network (SCN)* according to the special structural properties of communications in MIM [26].

**The Model**

![Diagram](Image)

**Nodes Generation**

As a first step, we generate \(N\) nodes with attributes, denoted as \(V = (T, C, I)\), according to the \(N\) samples in a multi-dimensional dataset. \(T, C, I\) represent nodes’ order, geographic and influence attributes respectfully.

As for the order attribute, \(T\) refers to a **non-strict partial order relation** “\(\leq\)” among nodes that is reflexive, antisymmetric, and transitive. That is to say, any \(t_i, t_j, t_k \in T\) satisfies:

i. Reflexivity: \(t_i \leq t_i\), i.e. every element is related to itself.

ii. Anti-symmetry: if \(t_i \leq t_j\) and \(t_j \leq t_i\) then \(t_i = t_j\), i.e. no two distinct elements precede each other.

iii. Transitivity: if \(t_i \leq t_j\) and \(t_j \leq t_k\) then \(t_i \leq t_k\).

The order attribute \(T\) defines the “order of arrival” of nodes to the graph. According to the order attribute sequence \(t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_{n-1} \leq t_n\), we acquire \(n\) nodes with arrival sequence \(v_1 \leq v_2 \leq v_3 \leq \cdots \leq v_{n-1} \leq v_n\). While the most common \(T\) is the timestamp, any other suitable numerical features in a multi-dimensional dataset can be utilized as \(T\).
For the influence attribute $I$ of nodes, due to the general applicability of the Pareto Principle, given a multi-dimensional dataset regardless of its field, on most occasions, one sample can be reduced to a single value (i.e. influence rate, $InfRt \in I$) that belongs to a population following a loose power-law distribution (see figure 2) using a linear function:

$$InfRt_t = Scale(aX_t + \beta)$$

where $X_t$ are the features of sample $i$ and $Scale(\cdot)$ is a standardization function.

![Figure 2. the InfRt distribution of several multi-dimensional datasets. (a. Global Terrorism Database [13]; b. UK Road Safety Dataset [27]; c. COVID-19 Vaccination Progress Dataset [28]; d. IP Network Traffic Flows Dataset [29]; e. Popular YouTube Channel Data [30])](image)

For other cases where reduced values cannot be regarded as following a power-law population, we consider them as random samples. We also managed to convert this type of dataset to a scale-free network.

**Graph Dynamics**

The next step is modeling the graph dynamics via the $N$ nodes generated above. Through the order attribute $T$ of nodes, we are able to form an arrival sequence of nodes. For each $t \in T$, let $v_t$ describe the currently arrived node and $G_t = (V_t, E_t)$ describe the current graph. The potential neighbors of $v_t$, denoted as $PNbr_{v_t}$, are sampled through a neighbor searching algorithm, denoted as $Y(\cdot)$. After the potential neighbors are determined, edges are generated with the control of the nodes’ influence attribute.

**Neighbors Resampling**

The potential neighbor number $k_t$ of for node $v_t$ is determined by $v_t$’s influence attribute $InfRt_{v_t}$ regardless of its geographic attribute $C_{v_t}$:

$$k_t = \min(n_{t-1}, \eta \ast InfRt_{v_t}^\theta)$$

where $n_{t-1} = |V_{t-1}|$ and $\eta, \theta \in R^+$ represent the constant controlling the global edge scale and the scale variants among nodes respectfully.

As we have discussed previously, most multi-dimensional datasets can be transformed into a list of nodes possessing scale-free influence attributes, $I$. For the cases where $I$ follow a power-law distribution, $\theta$ is set to 1, as for other cases in which the variable’s population is considered random, $\theta$ is generally between 3-9.
Once $k_t$ is settled, neighbor searching $Y(\cdot)$ is performed according to the geographic and order attributes among node $v_t$ and nodes in $V_{t-1}$. We calculate the distance of two nodes using Minkowski Distance as follows:

$$d(i,j) = \sqrt{m(t_i - t_j)^p + \frac{n}{D} \sum_{\lambda=1}^{D} (c_{i,\lambda} - c_{j,\lambda})^p}$$

where $t_i, c_i$ represent the order attribute and the geographic attribute of node $i$ and $j$ respectively, $m, n$ represents their weights, and $D$ is the geographic dimension.

Noted that the order and geographic attributes are both considered when calculating the distance. This is because we think the impact that order and geographic properties have on graph dynamics should be all taken into consideration.

Moreover, although we utilize the Minkowski Distance as a demonstration, the distance function will not affect the network property. As a result, any means of distance can be employed here. Lastly, the $PNbr_{v_t}$ is acquired as followed:

$$PNbr_{v_t} = Y_{k_t}(v_t, V_{t-1})$$

Edges Generation

As for edge generation for $G_t$, let $E_{v.t,j}$ represents the edge between node $v_t$ and $j \in V_{t-1}$, and $E_{v.t,j} = 1$ indicates the existence of an undirected edge. The node $v_t$ connects to $j$ with probability:

$$P(E_{v.t,j} = 1) = \Gamma(\infRt_{v_t}, \infRt_j), j \in PNbr_{v_t}$$

where $\Gamma(\cdot)$ is a generalized function to reduce influence rates to a single value and, in this paper, we apply multiplication as $\Gamma(\cdot)$.

Evaluation and Simulation

Network Property Evaluation

Scale-Free Distribution

We tested our model’s degree distribution for two scenarios mentioned earlier — $\infRt$ is either scale-free or random. The few nodes with large degree (tail of the power-law
distribution) in a scale-free graph are referred to as Gnode (giant node).

For the scale-free InfRt scenario, similarity among BA Model, Configuration Model, and our model lies in the tail of degree distribution (i.e. scattered points forming several horizontal lines), particularly, node numbers with large edge counts are mostly 1. This is due to the fact that Gnodes are rare and, moreover, Gnodes always contain a large number of edges. It is not common for two Gnodes to have the same amount of edges.

Nevertheless, the difference lies in the head of the distribution — The gradient of the curve head is flatter than the tail in our model, whereas for BA Model and Configuration Model, the gradient seems to remain the same.

We found that for real-life networks, the degree distribution is more similar to our model.

Figure 4 shows the degree distribution of several real-life networks from different fields and our model’s corresponding simulations. We can see that, regardless of the field or node scale
of the real-life network, the gradient of the distribution head is always flatter than the tail. This property of networks is correctly captured by our model while others cannot.

**Intrinsic Values versus Network Properties**

For modeling real-life datasets as complex networks, the edge-generation probability represents influence, the possibility of forming connections of nodes. While a system evolves, the influences of individuals within also decay, because it is obvious that any individual’s influence cannot last forever in a system. As a consequence, when modeling multi-dimensional data via network models, any node’s probability of forming edges should also decay alongside the evolution of the network. This is what we call the information decay phenomenon.

Now, we compare the simulation results of *Gnodes* in both *BA Model* and our model to illustrate information decay. The *Gnodes* edge-forming dynamics of the *BA Model* are shown in figure 5.a.

![Figure 5. Random sampled Gnodes’ edge-forming probability dynamics with the new arriving node (t in x-axis represents the node v_t arrival), under the simulation of 1000 nodes in total and 100 nodes for BA initial graph (figure 5.a for the BA Model; figure 5.b for our model).](image)

**Chart 1. Statistics of Gnodes Edge-Forming Probabilities in BA Model**

| node  | v_0      | v_16     | v_18     |
|-------|----------|----------|----------|
| max   | 0.04368  | 0.04598  | 0.04031  |
| 75%   | 0.02282  | 0.02253  | 0.01970  |
| 50%   | 0.01728  | 0.01961  | 0.01398  |
| 25%   | 0.01543  | 0.01565  | 0.01201  |
| min   | 0.01365  | 0.01360  | 0.01049  |
| std   | 0.006899 | 0.00713  | 0.006601 |
As for our model, the probabilities dissolve at around 200 nodes after $G_{nodes}$' own arrivals, satisfying the phenomenon that “an individual’s influence decays as the whole system evolves”.

Moreover, although the edge-forming probabilities decay in our model, the property determining the edge formation for any node $v$, $InfRt_v$, remains as a constant. What has changed is that the result of neighbor resampling of the new arrival is more and more unlikely to include $v$ as a result of introducing the order attribute into distance calculation.

This is exactly what we tend to accomplish when using network modeling multi-dimensional data-- the notable decayed influences of nodes and the intrinsic property of nodes, i.e. $InfRt$, unchanged. Because previous models, in which forming edges is based on the properties of the network, have an inevitable preference that a node’s property also alters while the network evolves. Nevertheless, in many real-life situations, as the surroundings changes, individuals within that environment tend to remain in the status quo.

To conclude, compared to the node’s degree property which grows alongside the network evolution, our model’s methodology is more suitable for modeling real-life datasets.

**Large Scale with High Generation Proficiency**

The time complexity of our model is solely dependent upon the process of finding $PNbr$. This is the result of two factors: 1. each new node only has to consider the possibility of having edges with a limited number of existing nodes; 2. the determinant of forming an edge is $InfRt$, a constant. As a consequence, a constant time of finding $PNbr$ will lead to a model time complexity of $O(n)$, in which $n$ represents the network node number.

In this section, we demonstrate the time complexity of our model by employing Locality-Sensitive Hashing (LSH) [35], an approximate nearest neighbor (ANN) algorithm, denoted as $\Upsilon(\cdot)$. We simulated a network with 5000 nodes and documented the time cost of each node adding to the graph. Figure 6 shows the result.

![Figure 6](image_url)

Figure 6. The time complexity of our model with LSH and its fitting curve, original scatters were regularly sampled into 1/50 of its original size for legibility.

The fitting curve in Figure 6 is $0.0152x^{1.0014}$ with covariance matrix:
The time complexity of our model with LSH as \( Y(\cdot) \) is \( O(n) \) because of LSH’s \( O(1) \) query time. This makes it possible for modeling any large-scale multi-dimensional datasets as the time cost will be tolerable regardless of the dataset scale. This scale-invariant time complexity is particularly crucial for any network model and can be more clearly demonstrated with the proportional time cost relative to a small-scale graph cost.

As a consequence, we compared several previous works’ network generation time to our model in which we denoted the time cost of any graph with 1000 nodes as \( \{1000\} \). Included previous models are some well-known property-based-solely network models-- Small-World Network, Preferential Attachment Model, a classic network generation with hybrid inputs -- Configuration Model, and OSN which, like our model, also focuses on modeling datasets into scale-free networks.

### Chart 2. The Proportional Time Cost of Several Models, the unit is \( \{1000\} \)

|                  | Barabasi Albert Graph | Erdos Renyi Graph | Watts Strogatz Graph | OSN Model | Our Model |
|------------------|------------------------|-------------------|----------------------|-----------|-----------|
| 1000 nodes       | 1                      | 1                 | 1                    | 1         | 1         |
| 2000 nodes       | 3.74                   | 3.78              | 3.39                 | 9.17      | 3.95      |
| 5000 nodes       | 23.29                  | 26.07             | 24.14                | 130.50    | 26.13     |
| 10000 nodes      | 97.07                  | 124.55            | 465.14               | 1112.65   | 108.88    |

### Multi-Dimensional Data Modeling

In this section, we further demonstrate the procedure of FSM and, in particular, the influence decay we mentioned earlier through the 9-11 terrorist attack using the Global Terrorism Database (GTD), an open-source event-based database including information on terrorist events around the world since 1970 (currently updated through 2020) [13]. Since the 9-11 attack has been one of the most severe terrorist events with profound impact on the world, we thought modeling the terrorist events confined to 9-11 via our network model would provide more obvious disclosure about the influence decay phenomenon.

As a first step, we filtered the GTD data to a subset with events that occurred between 2001 and 2005. To represent terrorist attack events as nodes with order, geographic, and influence attributes, i.e. \( V = (T, C, I) \), we first calculate the InfRt sequence with a min-max(a−b) standardization, in which the InfRt of node \( i \) is calculated as follows:

\[
\text{InfRt}_{i, \text{MinMax}} = \frac{(\alpha X_i + \beta) - \min_{j \in N} \alpha X_j + \beta}{\max_{j \in N} \alpha X_j + \beta - \min_{j \in N} \alpha X_j + \beta} \ast (b - a) + a
\]

The linear function \( \alpha X_j + \beta \) is then implemented as follows:

\[
1 \ast nperps + 1 \ast nwound + 1.5 \ast nkills
\]
in which \textit{nperps}, \textit{nwound}, and \textit{nkills} mean the number of perpetrators, the total number of injured, and the total number of fatalities respectively.

As GTD already includes the time and location of attacks, we attained the order attribute \textbf{T} by utilizing the time information (\textit{iyear, imonth, iday} in GTD), and the geographic attribute \textbf{C} was acquired from coordinates (\textit{longitude, latitude}). On account of missing values within coordinate columns in GTD, the coordinates of the smallest known administrative area center will be employed as the incident’s missing coordinates. The smallest known administrative area information is within \textit{city, vicinity, location, and specificity} columns.

To demonstrate the influence decay phenomenon, we compared the network dynamics of our model with the \textbf{BA Model's} all the way till the end of 2005. The influence decay should be observed from a local perspective, i.e. node view, so we focused on the edge dynamic of node 9-11.

![GTD Nodes 2001.1.1-9.11(without 9-11 Attack)](image1)

![GTD Nodes 2001.1.1-9.11(with 9-11 Attack)](image2)

Figure 7. The nodes between 1.1st 2001 and 9.11th 2001 of modeling GTD. The darker the color is, the greater the InfRt of node is. (Above: without 9-11 attack; Below: with 9-11 attack)

Noted that \textbf{BA Model} requires an initial graph, we constructed a network of all 1417 recorded terrorist attacks between January 1st, 2011 and September 11th, 2001 using our network model and also use the result as the initial graph for \textbf{BA Model}. Figure 7 shows all the nodes between 1.1st, 2001 and 9.11th, 2001. We can see that due to the massive damage made by 9-11 attack, all relative influences of other nodes weaken once node 9-11 is appended.
After the construction of the initial graph, we simultaneously kept track of the edge dynamic of node 9-11 in both models. We found that in our model, the edge dynamic of node 9-11 remained stationary after around December 2001 whereas the dynamic in BA Model kept evolving throughout the simulation.

The edge dynamic result of 9-11 in our model further illustrates the influence decay we mentioned earlier in which any individual in a system has a limited influence that decays as
the system evolves. What we are trying to demonstrate in this section is a network approach to model multi-dimensional data in which, from a micro-perspective, the individual acts like real-world while, from a macro-perspective, the property of the whole system is preserved.

Discussion

An Alternative Method for Neighbors Resampling

During the process of neighbors resampling, we obtained potential neighbors, i.e. $PNbr$, through a neighbor searching function $\Upsilon(\cdot)$. Although, we have demonstrated the function takes the order and geographic attributes of nodes as input, we think the potential neighbor searching doesn’t have to utilize node attributes. An alternative implementation of $\Upsilon(\cdot)$ can take the original dataset's features as input. As a consequence, we can acquire $PNbr$ directly from the original dataset through KNN, Clustering, or any other algorithms that produce classification results. This alternative produces another benefit besides more flexibility of the model -- with $PNbr$ acquired beforehand, the model will not only reduce its running time but time complexity will also be fixed to $O(n)$. Further research can test the performance of different means of neighbors resampling by comparing the digital twin between generated network and original datasets [21].

The Current Time Complexity With LSH

By employing Locality-Sensitive Hashing (LSH) as $\Upsilon(\cdot)$ we demonstrate the linear time complexity of our model. For the reason that the process of LSH and edge generation are not simultaneous, we were able to build the hash table of LSH in advance, whose time complexity, by the way, is also $O(n)$. However, one thing needs to be mentioned is that when we built the hash table for all nodes, we need to keep records of all the intermediate results in memory because the final LSH will contain all the nodes, some of which are yet to arrive for the graph at time $t$ when sampling $PNbr_{t_1}$. This is how we shift the time burden to the memory burden. This topic can be further investigated by choosing different functions as $\Upsilon(\cdot)$, or apply the alternative method for neighbors resampling we have mentioned earlier in this section.

Conclusion

In this paper, we proposed FSM, a generic multi-dimensional data modeling network with fast generation time. Synthetic networks can capture important features of the data modeled and, in the meanwhile, preserve the universal properties of graphs. Another essential identity of our model is that the behavior of nodes within reflects the influence decay while previous works cannot. Due to the flexibility and time efficiency of our model, it can be used for tracing dynamic of system, investigating network topology, generating synthetic network data, etc.
References

1. Travers, J. and S. Milgram, *An experimental study of the small world problem.* An experimental study of the small world problem. Science, 1999. 286(5439): p. 509-512.

2. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. 286(5439): p. 509-512.

3. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998. 393(6684): p. 440-442.

4. Burt, R.S., *Structural holes and good ideas.* American Journal of Sociology, 2004. 110(2): p. 349-399.

5. Beyerlein, K. and J.R. Hipp, *Social capital, too much of a good thing? American religious traditions and community crime.* Social Forces, 2005. 84(2): p. 995-1013.

6. Jiang, X. and Y. Li, *An empirical investigation of knowledge management and innovative performance: The case of alliances.* Research Policy, 2009. 38(2): p. 358-368.

7. Rahmani, F., C. Valmohammadi, and K. Fathi, *Detecting fraudulent transactions in banking cards using scale-free graphs.* Concurrency and Computation-Practice & Experience, 2022. 34(19).

8. Roeder, M., et al. *LEMMING - Example-based Mimicking of Knowledge Graphs.* in 15th IEEE International Conference on Semantic Computing (ICSC). 2021. Electr Network.

9. Chung, W., B. Rao, and L. Wang, *Interaction Models for Detecting Nodal Activities in Temporal Social Media Networks.* Acm Transactions on Management Information Systems, 2019. 10(4).

10. Gursoy, F. and B. Badur, *An Agent-Based Modeling Approach to Brain Drain.* IEEE Transactions on Computational Social Systems, 2022. 9(2): p. 356-365.

11. Inoue, M., P. Thong, and H. Shimodaira, *Joint estimation of non-parametric transitivity and preferential attachment functions in scientific co-authorship networks.* Journal of Informetrics, 2020. 14(3).

12. Khan, J. and S. Lee, *Online Social Networks (OSN) Evolution Model Based on Homophily and Preferential Attachment.* Symmetry-Basel, 2018. 10(11).

13. *National Consortium for the Study of Terrorism and Responses to Terrorism.* 2022; Available from: https://www.start.umd.edu/gtd/.

14. Erdos, P. and A. Renyi, *ON THE EVOLUTION OF RANDOM GRAPHS.* Bulletin of the International Statistical Institute, 1960. 38(4): p. 343-347.

15. Breiger, R.L., *An Exponential Family of Probability-Distributions for Directed-Graphs - Comment.* Journal of the American Statistical Association, 1981. 76(373): p. 51-53.

16. Pattison, P. and S. Wasserman, *Logit models and logistic regressions for social networks: II. Multivariate relations.* British Journal of Mathematical and Statistical Psychology, 1999. 52(2): p. 169-193.

17. Romero, J., J. Finke, and A. Salazar, *Fitness-Weighted Preferential Attachment with Varying Number of New Connections.* in 8th International Conference on Complex Networks and Their Applications (COMPLEX NETWORKS). 2019. Calouste Gulbenkian Fdn, Lisbon, PORTUGAL.
18. Newman, M.E.J., *The structure and function of complex networks*. Siam Review, 2003. **45**(2): p. 167-256.
19. Duan, B., et al. *Dynamic Social Networks Generator Based on Modularity: DSNG-M*. in 2nd International Conference on Data Intelligence and Security (ICDIS). 2019. Tx.
20. Bojchevski, A., et al. *NetGAN: Generating Graphs via Random Walks*. in 35th International Conference on Machine Learning (ICML). 2018. Stockholm, SWEDEN.
21. Wen, J., B. Gabrys, and K. Musial, *Toward Digital Twin Oriented Modeling of Complex Networked Systems and Their Dynamics: A Comprehensive Survey*. Ieee Access, 2022. **10**: p. 66886-66923.
22. Danovski, K. and M. Brede. *Effects of Population Structure on the Evolution of Linguistic Convention*. in 10th International Conference on Complex Networks and Their Applications (COMPLEX NETWORKS). 2021. Polytechn Univ Madrid, Madrid, SPAIN.
23. Han, X. and J. Wang, *Modelling and Analyzing the Semantic Evolution of Social Media User Behaviors during Disaster Events: A Case Study of COVID-19*. Isprs International Journal of Geo-Information, 2022. **11**(7).
24. Lai, G., et al., *Revealing Taxi Interaction Network of Urban Functional Area Units in Shenzhen, China*. Isprs International Journal of Geo-Information, 2022. **11**(7).
25. Sottile, S., O. Kahramanogullari, and M. Sensi, *How network properties and epidemic parameters influence stochastic SIR dynamics on scale-free random networks*. Journal of Simulation, 2022.
26. Sahafizadeh, E. and B.T. Ladani, *A Model for Social Communication Network in Mobile Instant Messaging Systems*. Ieee Transactions on Computational Social Systems, 2020. **7**(1): p. 68-83.
27. *Road Safety Data*. 15 October 2022; Available from: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.
28. Preda, G. *COVID-19 World Vaccination Progress*. 2022; Available from: https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress.
29. Rojas, J.S. *IP Network Traffic Flows Labeled with 75 Apps*. 2018; Available from: https://www.kaggle.com/datasets/jsrojas/ip-network-traffic-flows-labeled-with-87-apps.
30. Jha, S. *Most Subscribed YouTube Channels*. 2022; Available from: https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data.
31. Rozemberczki, B., C. Allen, and R. Sarkar. *Multi-scale Attributed Node Embedding*. 2019.
32. McAuley, J. and J. Leskovec, *Learning to Discover Social Circles in Ego Networks*, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. 2012, Curran Associates Inc.: Lake Tahoe, Nevada. p. 539–547.
33. Rozemberczki, B. and R. Sarkar, *Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models*, in Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20). 2020. p. 1325–1334.
34. Leskovec, J., J. Kleinberg, and C. Faloutsos, *Graph Evolution: Densification and Shrinking Diameters*. ACM Trans. Knowl. Discov. Data, 2007. **1**(1): p. 2–es.
35. Datar, M., et al., *Locality sensitive hashing scheme based on p-stable distributions.* 
*SoCG '04. Scg '04 Proceedings of the Twentieth Annual Symposium on Computational Geometry, 2004.* 34(2): p. 253--262.