# Introduction and Comparison On Random Graph Generation Algorithms

郑清源

1. School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

**Abstract**

In network science and data analysis, randomly generated graphs are widely applied to evaluate the performance of existing algorithms and verify developed models. This article explains and compares a few typical random graph generation algorithms, discuss their different mathematical and/or practical properties, and then lists their advantages and disadvantages in practical usage.

**Keywords** Random graph, Stochastic process

## 1 Introduction

In an era where quantity of information explodes and most of them are interconnected and correlative, data analysis on sophisticated graphs and complex networks is increasingly valued by researchers and business owners. A lot of data mining algorithms and models are developed to extract and exploit valuable and potentially profitable information from existing graphs. To verify them, tons of data has to be collected, well organized, and correctly labelled in order to construct a graph ready for analysis. However, such process involves with a large amount of manual work that is very painful and costly. In order to enable agile verification and iteration of new models, generation of heterogeneous random graphs has become very important. Fortunately, many effective algorithms are already proposed and broadly adopted.

This article would list a couple of typical algorithms of graph generation for random undirected graphs, discuss the most fundamental ideas under the hood, and then compare them on different and practical perspectives. Afterward, some real-world graphs are provided in comparison to samples of our random graphs.

## 2 Random Graphs

In mathematical points of view, a random graph is described by either probability distribution of vertices and edges or a stochastic process that generate it. Practically, two types of construction process are very common:

1. Create all nodes at once, and connect them upon a specific probability distribution
2. Add nodes by ones or batches, and wire(rewire) them on growth

Given some parameters determined, such as number of vertices and edges, a sample of the random graph could be created and put into use.

Before digging into concrete generation algorithms, some important indices for realistic graph evaluations are listed below:

1. **degree distribution**: probability distribution of degree of nodes in the graph
2. **average path length**: average smallest number of edges between any two nodes in the graph
3. **clustering coefficient**: an index that describe how nodes form communities

## 3 Graph Generation Algorithms

### 3.1 Erdős–Rényi model

Considering the first type of construction process, the Erdős–Rényi model further specify that the presence of edge between any two vertices is equally and independently probable, i.e. observes a uniform distribution. It may either be parameterized as $G(n, M)$ or $G(n, p)$, where $n$ is the number of vertices, $m$ is the number of edges and $p$ is the probability of a connection to form.

In $G(n,M)$ notation, the total number of edges is deterministic, i.e. $M$. Thus the scale of the sample is known beforehand and spawn probability of edges may vary. However, it is obvious that the ML estimation of connection rate satisfy the formula:

$$p' = \frac{M}{C_n^2}$$

Meanwhile, in $G(n,p)$ abstraction, the edge count $M'$ is a random variable with a binomial probability distribution

$$P(M') = p^M(1-p)^{C_n^2-M}, 0 < M < C_n^2 + 1$$

and an expected value

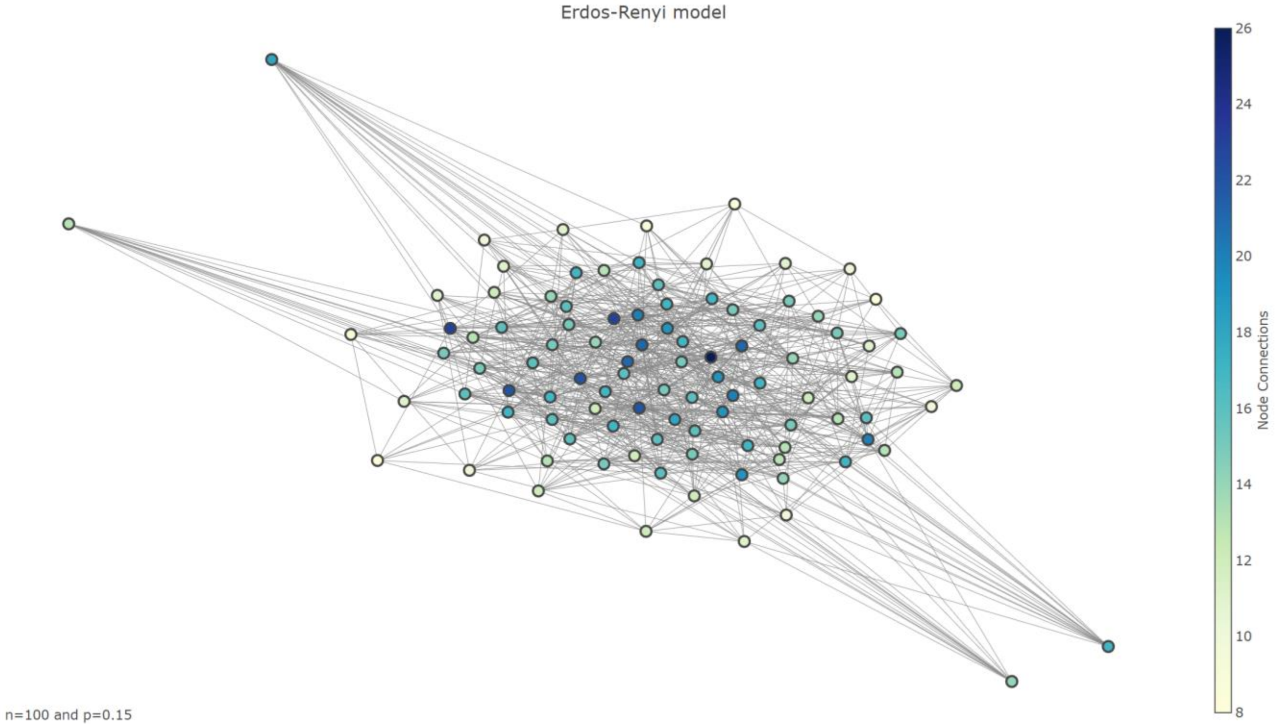
$$E(M') = p \times C_n^2$$

In this article, the latter model is preferred to be referred as ER model. By definition, ER model has a binomial degree distribution

$$P(k) = C_{n-1}^k p^k (1-p)^{n-1-k}$$

where k denotes the degree of a random vertex, and a fairly small clustering coefficient. The graph it samples is usually very dense and such topology is rare in reality. Also, because of no random nature of the graph, it is not guarantee one totally inter-connected graph is sampled instead of a forest. However, it is provable that there is usually a major component in branches of forest, if any.

A sample of ER model is given below, where $n = 100$ and $p = 0.15$, and vertices are colored by its degree.



3.2 Barabási–Albert model

In real-world, many networks have the property of being scale-free; that is, their degree distribution follows a power law, instead of being binomial. Suppose P(k) denotes the proportion nodes that have k outgoing edges. The power law could be described as $P(k) \sim k^{-\gamma}$, where $\gamma$ typically falls in the range $2 < \gamma < 3$.

The construction process of scale-free networks usually (but not necessarily) possess the following schemes:

1. Construct the whole graph by adding new nodes
2. Apply preferential attachment: vertices with more connections are more popular

Similar to the ER model, Barabási–Albert model requires a linear preferential attachment probability. For an existing node $n_i$, a new node would have a chance of $p_i = \frac{k_i}{\sum_j k_j}$ to be attached or connected, noting $k_i$ is degree of the particular vertex. Usually, the number of edges to be attached for each new node is denoted as a constant m. Therefore, the model may be denoted as $G(n,m)$.

For a model starts with a single isolated vertex, the edge count is also a deterministic constant $(n-1) \times m$. Different from previously mentioned ER model, graph samples of BA model usually forms a certain number of central nodes and possess a very recognizable topology structure. And more importantly, they are scale-free, where degree distribution would be around

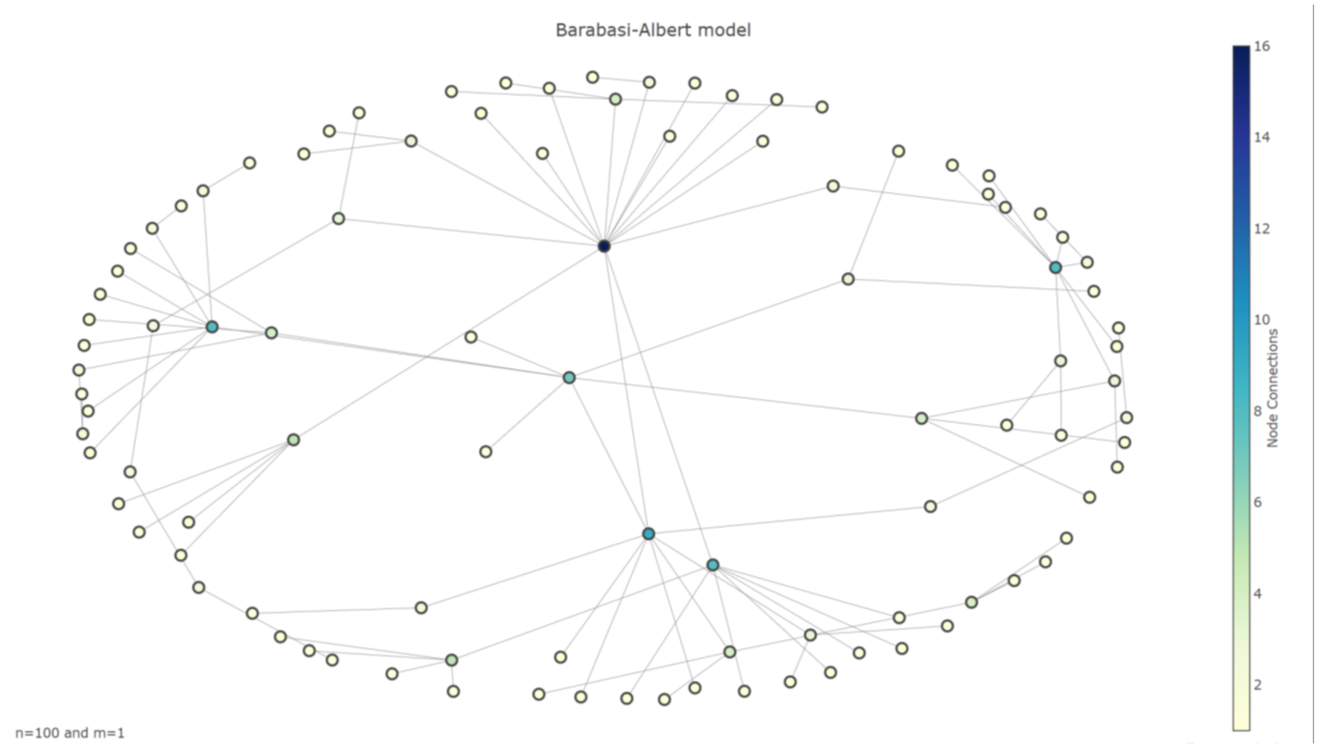
$$P(k) \sim k^{-3}$$

and average path length would be around

$$l \sim \frac{\ln(N)}{\ln\big(ln(N)\big)}$$

To be noted, preferential attachment is not a requirement. And different random distribution for new attachment might work better in some circumstances.

The following image shows a sample of BA model, where $n = 100$ and $m = 1$. It is apparent that the network BA model generates is very centralized, and most end-points rely on a few inter-connected hubs to communicate with each other, which, in some sense, is very realistic.

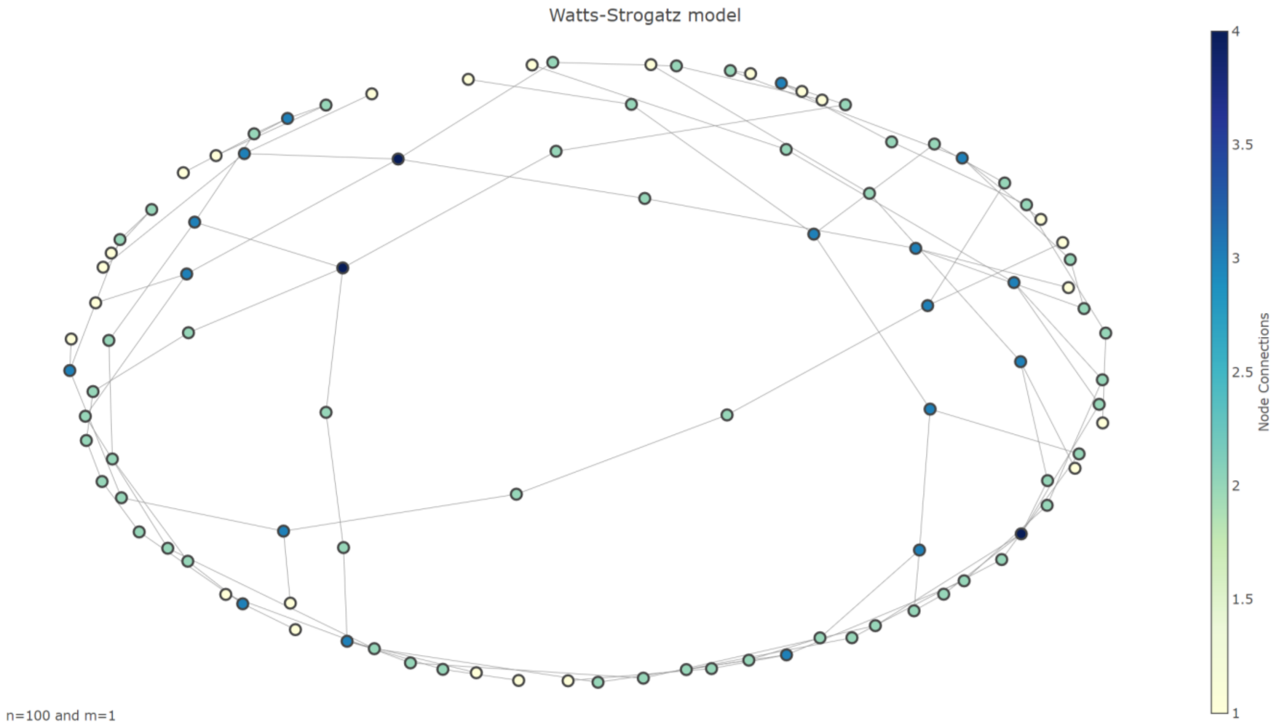


3.3 Watts–Strogatz model

On top of being scale-free, many real-world networks are also observed as following the analogy of small-world phenomenon, where length of path between any two nodes are relatively short in comparison to the scale of whole graph. Mathematically, a small-world network has a small average node distance and a significantly large clustering coefficient.

And the most trivial method for generating small-world graph is Watts-Strogatz model. This is also very simple an algorithm. It starts with a regular ring lattice graph with $N$ nodes connected to its $K$ neighbors. For each lattice edge from $n_i$ to $n_j$ where $i < j$, there is a chance of $\beta$ that a rewire process occur and a random new node $n_k$ is selected in place of $n_j$ in uniform distribution. In addition, the algorithm should always avoid some $n_k$ that may lead to self loop or duplicate edges. After all edges are visited, a small-world graph is done.

Note properties such as degree distribution and average path length is rather complicated and rely heavily on the parameter \beta. Here, the formula would be skipped, and a simulation of the model is beneath:



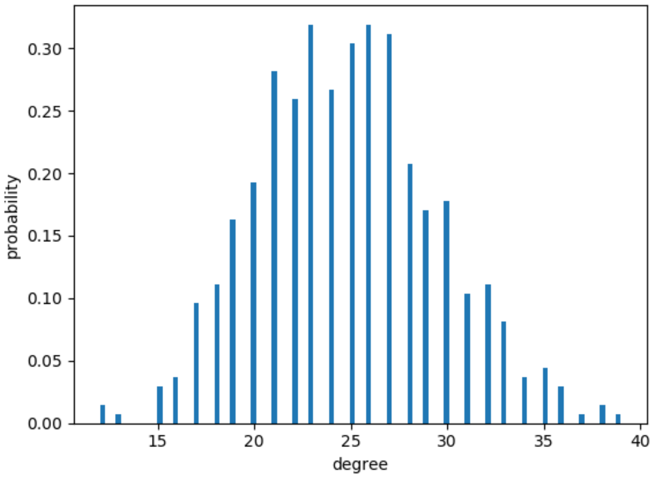
3.4 Ensemble model

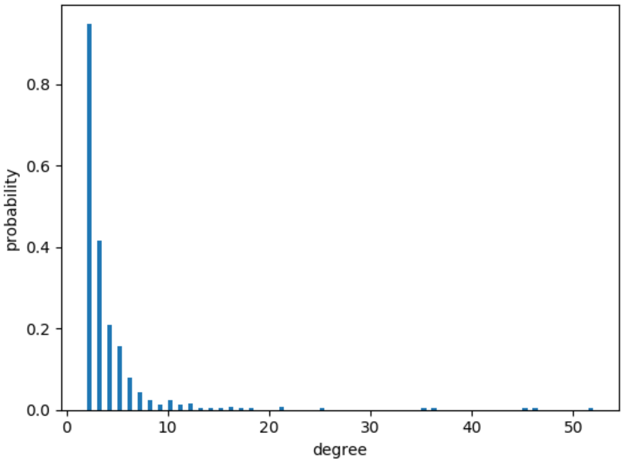
According to the discussion above, although being carefully designed, both BA model and WS model cannot perfectly abstract our real-world needs. However, as graph construction process is incremental, we may first construct several sub-graphs as an autonomous system and then merge them into a single large network.

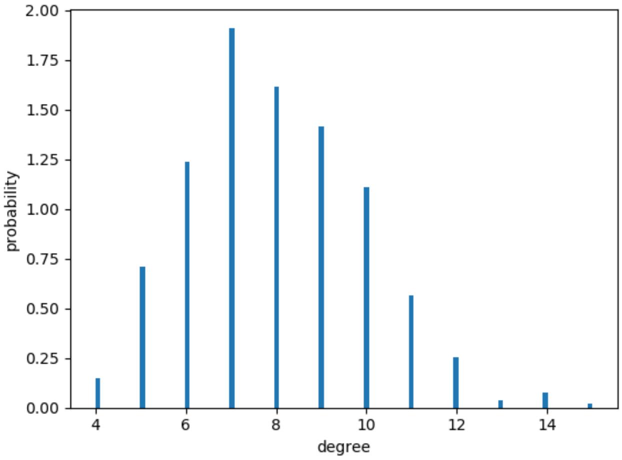
## 4 Monte Carlo Simulation On Statistical Properties

To be noted, the simulation graph below is mostly smoothened by averaging results from multiple epochs. Also, the simulation results are always listed as, respectively, ER, BA and WS model.

3.1 Degree Distribution

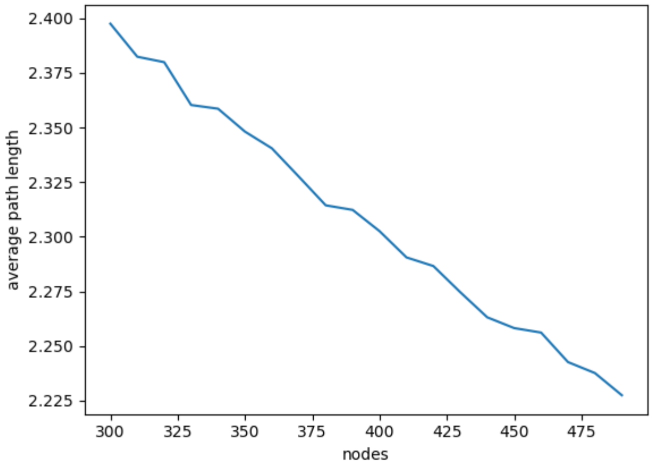


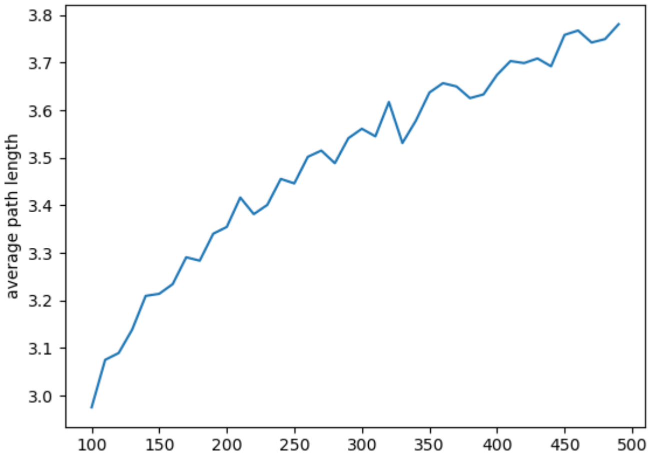


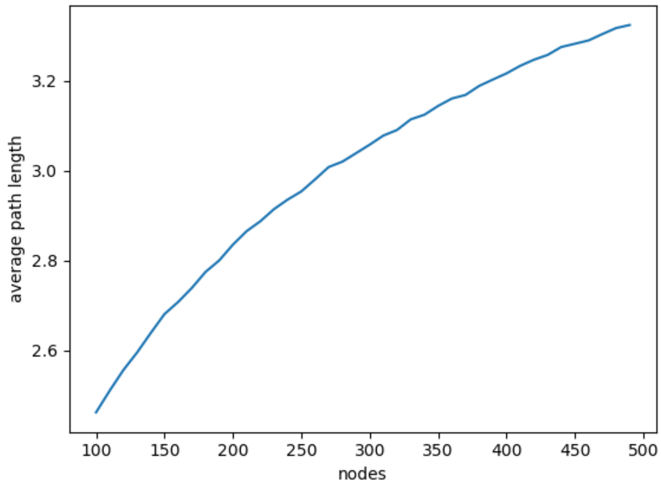


A Poisson distribution for the ER model and power-law distribution for the BA model can be easily observed. The degree distribution of WS model presents a Poisson distribution as well, but as it is less densely connected as the ER model, it is less obvious.

3.2 Average Path Length

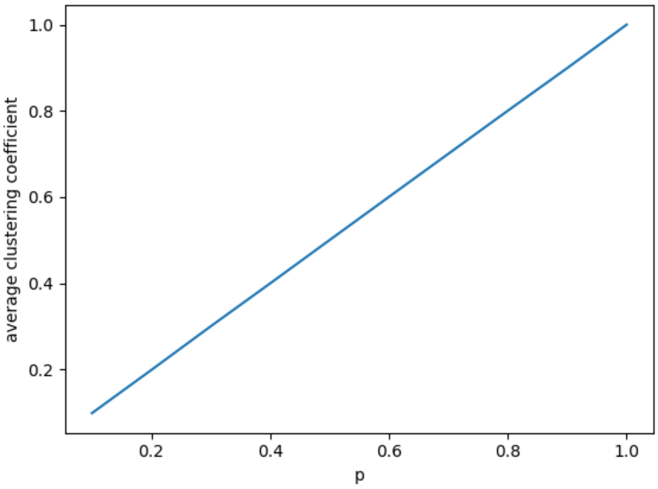


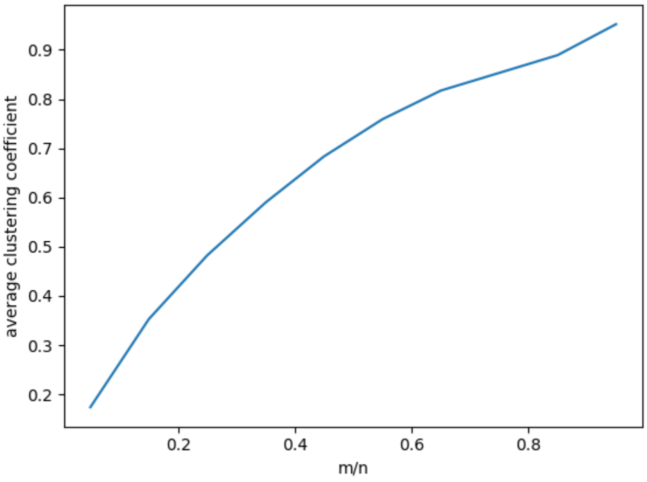


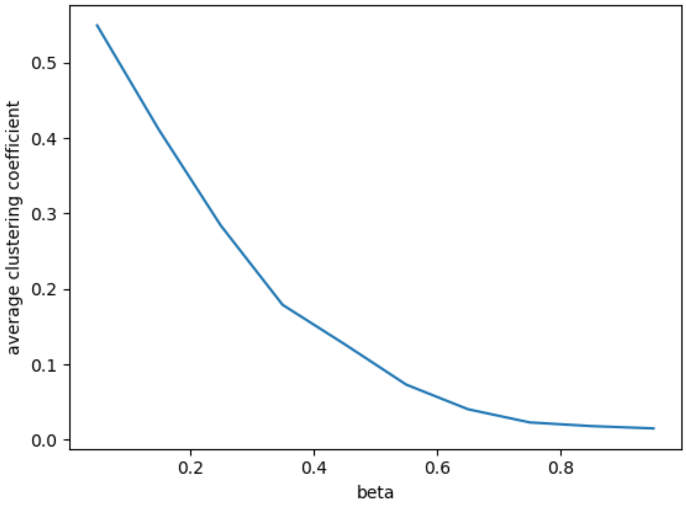


To be noted, for ER model, as p persists as a constant, the number of edges grows faster than the scale of the graph(Note $E(M') = p \times C_n^2 \sim p \times n^2$), so the average path length actually drops as the count of nodes increases. Apparently, these three models have an average path length that grows in logarithm as the graph scales, which is expected and realistic in real-world.

3.3 Average Clustering Coefficient



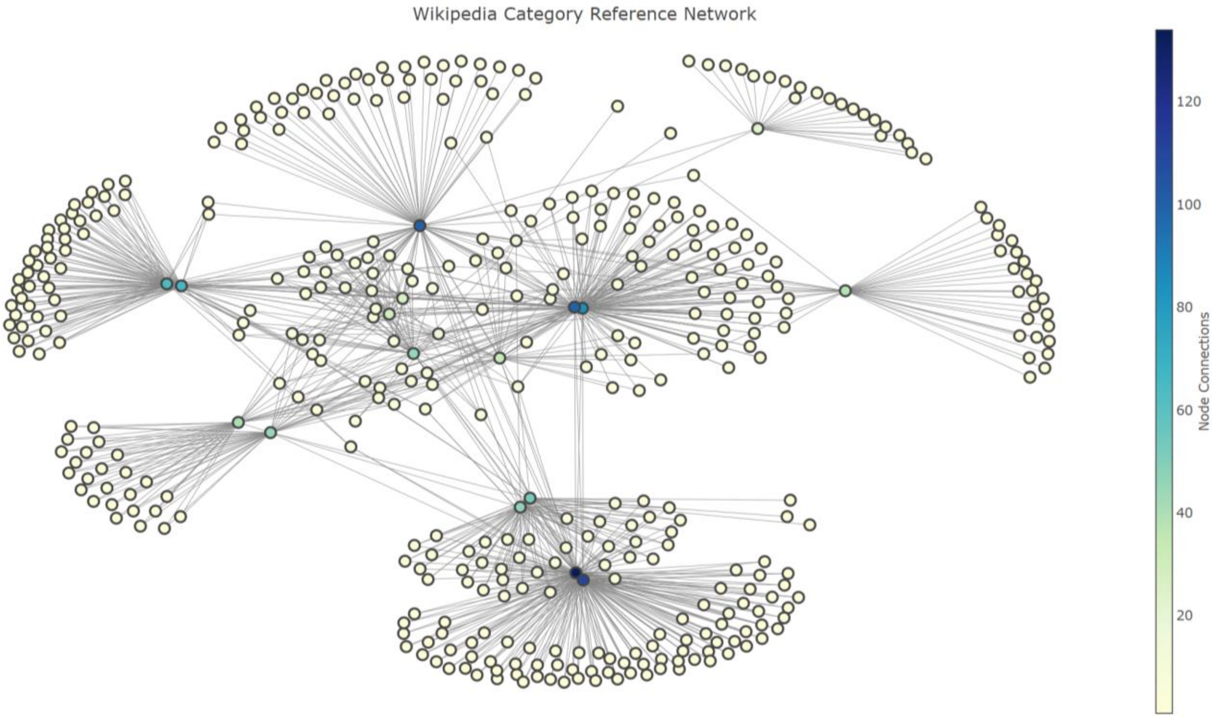




Although ER model and BA model can have a significantly large average clustering coefficient, those parameter range is usually not typical and rarely happen. For example, $p$ in ER model is usually very small, especially when the graph generated is fairly large. In such situation, the average clustering coefficient remains tiny all the time.

## 4 Real-world Network

The following image shows a category-to-category map collected from Wikipedia, in comparison to the random graph generated above.



This image shows very typical properties of real-world image which follows patterns like being scale-free and highly clustered. Some major topic covers secondary topics and each topic interconnects and references each other by, usually, those major topics.

In the network above, the average shortest path length is 2.9590 and the average clustering coefficient is 0.7006. Neither of models presented above may generate such a network, so they still need further polish and even trade-offs for practical usage.

## References

1. Erdős, P.; Rényi, A. (1959). "On Random Graphs. I". Publicationes Mathematicae
2. Albert, Réka; Barabási, Albert-László (2002). "Statistical mechanics of complex networks". Reviews of Modern Physics.
3. Watts, D. J.; Strogatz, S. H. (1998). "Collective dynamics of 'small-world' networks". Nature.