Model Analysis for the BA and SBM as Internet AS connectivity Models

Benny Rubin, Jude Rizzo

May 16th 2023

1 Introduction

The internet has changed the way people engage with each other, access information, and almost every way that people interact with the world around them. There is a strong argument to be made that the internet is one of the most important networks that can be found in modern society. Despite the vast amount of literature that exists on mapping the topology of the internet, there is hardly any agreement on what it looks like or what models work best to describe it. Yet, understanding the internet and its routing behavior is critical for evaluating the behavior and performance of routing protocols, developing new designs for resource provisioning, and securing the internet as a critical infrastructure [6, 8]. Mapping out the internet turns out to be an incredibly challenging feat [8]. It is no wonder that there is no consensus on what the internet topology looks like and what models can be used to best analyze it. Understanding why requires a little more information on the underlying infrastructure that makes up the Internet.

At its simplest, the internet contains a large number of routers that forward traffic between end hosts. The routers use distributed algorithms to discover routes between systems to forward data. In reality, sets of routers are partitioned into groups owned by a single entity called an Autonomous System (AS). Within an AS, internal routing is done using those distributed protocols (such as OSPF), however a different protocol is needed to route data between Autonomous Systems. This is the role of the Border Gateway Protocol (BGP). To view the internet as a graph, Autonomous Systems can be abstracted away as nodes, and edges are the BGP peering connections between neighboring ASes [4]. The use of a dynamic routing protocol to connect the parts of the internet makes for an incredibly complex and every-changing topology.

1.1 Prior Work

[4] uses servers that collect BGP data from ASes to map out peering relationships between Autonomous Systems. This method of viewing BGP updates from ISPs and ASes is the most popular approach we've seen for mapping the

internet topology. The largest project of this type is Oregon Route Views, which collected billions of BGP announcements that can be used to re-construct network topology [9].

With so much BGP data being recorded every day (over a billion route updates), it is hard to keep track of the data and what is useful. This is compounded by the fact that a lot of it is redundant. MVP allows users to get less, and more useful data out of BGP route updates [1]. The authors accomplish this by grouping events into categories by similarity (how they change the topography) and picking data from dissimilar categories. These two approaches come from a rich set of literature that on collecting and analyzing BGP data to understand internet connectivity.

Other projects attempt to define mathematical models that accurately represent the internet [10]. We talk extensively about these projects later in the paper. A brief summary is that they focus on the node degree distribution in the internet, claiming that it follows a power law distribution: highly connected nodes are more likely to get new connections and nodes are more likely to connect with other nodes that have high node degree.

One of the key flaws in most work done to map out the internet is that they use BGP data to build their graph [8]. BGP was not intended or designed to measure the internet. In fact, one of the major design goals of BGP is to hide information. It allows ASes to express routing policies without revealing internal data such as customer and provider information. Notably, BGP data lacks internet-wide state and routing information [8]. Traceroute and ping tools, also widely used to map the internet, were similarly not designed for such as purpose. Unfortunately, there are not many better ways to map the internet, so the key insight is to be aware of where the data is coming from and what flaws might exist in inferences made from said data.

1.2 Our Work

In this paper, we argue that solely focusing on node degree distribution is not the right approach for modeling the internet AS level connectivity. There is a large amount of literature that supports that the internet follows a powerlaw distribution [10], yet this implies barely any information about the structure of the network or how it operates. There are many networks that function completely differently that have the same node degree distribution [6]. As a counterpoint, it is possible to develop random graphs with any given degree distribution that have no meaning whatsoever.

We argue that clustering characteristics are just as important to understanding a network's topology and behavior. Analyzing the topology of the internet using clustering and expansion on top of node degree distribution gives a much better view.

We analyze data gathered on real-world Autonomous System connectivity, gathering metrics such as clustering coefficients, expansion, etc. We also analyze multiple proposed graph models for the internet and apply the same metrics, comparing them to real-world data.

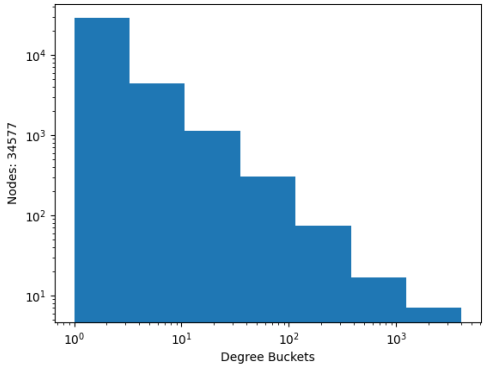

Figure 1: Internet Usage Histogram

2 Network Data

This section discusses the networkx graphs and data sets used in the analysis, as well as how they fit into the models being considered. Various aspects of the network structure, such as degree distribution, clustering coefficients, and expansion, are examined to gain a better understanding of the properties of the graph, using data from [11]

2.1 Degree Distribution

Upon analyzing the given graph from the latest BGP routing data, it is observed that the model follows a power-law distribution. The inverse power-law regression line accurately represents the trend of the data when plotted. A log-likelihood ratio value of 201, which is greater than 0, indicates that the inverse power-law distribution is more likely than the exponential distribution.

In Figure one, we plot the degree distribution. Note that on the y-axis we have a logarithmic scale for the number of nodes, and on the x-axis we have a logarithmic scale. Their linear relationship suggests the scale-free property - with the MLE of the scale free exponent being equal to the slope of the above line, which is roughly equal to -1.2.

2.2 Clustering Coefficients

The global clustering coefficient employed in this study is defined as the ratio of the number of triangles in the graph to the total possible number of triangles, as defined above:

$$T(G) = \frac{3\delta(G))}{\tau(G)}$$

To compute the global clustering coefficient, in our graph, we calculate the triangles and triplets directly using the networkx library. Given the relatively small graph size of 6,000 nodes, this method is computationally feasible. The resulting global clustering coefficient is approximately 0.71.

Year	Month	Nodes	Edges	Clustering
2011	01	18292	37414	0.460
2011	12	19246	38486	0.405
2012	01	19263	38783	0.402
2012	12	20077	40753	0.421
2013	01	20037	40655	0.431
2013	12	22706	44462	0.412
2014	01	22633	44253	0.400
2014	12	23990	47709	0.457
2015	01	23749	48788	0.551
2015	12	25215	51268	0.505
2016	01	25008	50434	0.505
2016	12	25283	50236	0.480
2017	01	25133	49715	0.456
2017	12	24934	51009	0.532
2018	01	27770	57298	0.532
2018	12	31144	61709	0.503
2019	02	31484	62582	0.510
2019	12	32261	66266	0.548
2020	01	34574	72439	0.578
2020	03	33947	71177	0.581
2020	05	34577	74290	0.609

Table 1: Table of clustering coefficients over time

In this table, we see that the global clustering coefficient actually grows with time, and does not seem to be approaching 0 - an important observation we will take note of when analyzing the theoretical result of later models.

Another metric we considered was edge expansion, defined as:

$$\Phi(G) = \min_{\substack{S \subseteq V \\ S \neq \emptyset}} \frac{|\delta(S)|}{|S|}$$

where $\delta(S)$ is the set of edges with one endpoint in S and the other in $V \setminus S$. We note here that expansion is NP-hard to calculate exactly, and approximating it using Cheeger's inequality on the entirety of the graph is still too computationally expensive to calculate. However, we note that when approximated by sampling, we find values $< 1e^{-2}$, suggesting a near 0 expansion for the whole graph.

3 BA Model and Analysis

3.1 BA model and assumptions

The BA model was designed to model social networks, the world wide web, and biological networks that exhibit the scale-free property, taking into account the following assumptions:

1. Growth: The network grows over time by adding new nodes.
2. Preferential attachment: New nodes are more likely to attach to existing nodes that have a high degree (i.e., many connections) in the network.
3. Scale-free network: The resulting network has a power-law distribution of node degrees, meaning that there are a few highly connected nodes (hubs) and many poorly connected nodes.

   The model incorporates these assumptions through the following algorithm

   (a) Initialize the graph G_0 nodes.
   (b) Add a new node to the network.
   (c) Connect the new node to m existing nodes, where the probability of connecting to a node i is proportional to its degree k_i in the network, such that $P(i) = k_i / \sum_j k_j$.
   (d) Repeat steps 2-3 until the network has n nodes.

3.2 Degree distribution

An important tool that we will use later in this paper is the expected degree of a node per iteration.

We can derive analytically the expected degree distribution of a BA model given any initial parameters, as well as a variance on the random variable α, the exponential in the BA distribution. For a given node, we have (from Network Science by Barabasi) that the degree distribution of the graph (which will be represented by $P(k) \sim k^{-3}$, where k is the degree of a node.

As a brief aside, if we wanted to find the expected degree of a given node in a graph, instead of summing up all of the above over given t, we can instead divide twice the number of edges by the number of nodes,

$$\frac{2(mt + |E_0|)}{t + |V_0|}$$

, where E_0 is the set of all the initial edges and $|V_0|$ is the set of the initial vertices.

Below, we analyze the clustering behavior of the BA model

3.3 Global Clustering Coefficient for the BA model

We have used analytical techniques to estimate the expected value of the global clustering coefficient for the BA model.

We use the global clustering coefficient as defined in [5] as

$$T(G) = \frac{3\delta(G))}{\tau(G)}$$

Where

$$\tau(G) : \text{total number of triples after t steps} \tag{1}$$
$$\delta(G) : \text{total number of triangles after t steps} \tag{2}$$

We use the technique outlined in Network Science [2], where we derive a temporal differential equation based on the unit change in the metric - in this case the global clustering coefficient.

In our approach, we find expected values for the numerator and denominator separately, and argue that their quotient should be asymptotically close to the expected value of the true expected global clustering coefficient. The argument is summarized below and the full derivation is in the appendix. Using a differential equation approach, we model the expected change in the global clustering coefficient, $T(t) = \frac{3\delta(t)}{\tau(t)}$, as new nodes and edges are added to the network.

3.3.1 Evaluating the expected number of triples

We derive the time-dependent function for the expected number of triples, $E[\tau(t)]$. In the limit as $t \to \infty$, we find that the expected number of triples grows linearly: $E[\tau(t)] \approx (\binom{m}{2} + 2m^2)t$.

3.3.2 Evaluating the expected number of triangles

Similarly, we derive the time-dependent function for the expected number of triangles, $E[\delta(t)]$. In the limit as $t \to \infty$ with small initial values, we find that the expected number of triangles grows logarithmically: $E[\delta(t)] \approx \binom{m}{2}(A + B)\ln|t|$, where A and B are constants dependent on the initial graph.

3.3.3 Approximating the expected global clustering coefficient

Dividing the expected number of triangles by the expected number of triples, we find that the growth of the global clustering coefficient is asymptotic of the order $\frac{\ln(t)}{t}$, which approaches 0 as $t \to \infty$.

A full derivation for the above is included in the appendix, as well as an exact formula for the expectations of $E(\tau(t))$ and $E(\delta(t))$

4 SBM and analysis

Unlike the BA model, the Stochastic Block Model does not use growth - instead, it starts with a fixed number of nodes, and generates edges probablistically, and independently, depending on a preassigned group assignment of each node.
.

More formally, we define a $k \times k$ matrix $P = [p_{ij}]$ where p_{ij} is the probability of an edge between a vertex in block C_i and a vertex in block V_j. The generation process of the edges is then as follows:

1. For each pair of vertices i and j such that $i \in C_a$ and $j \in C_b$:
2. Generate an edge between i and j with probability p_{ab}, independently of all other edges.

We will be examining a version of the Stochastic Block model called the strongly assortative model. In this case, we strongly preference the probability that an arbitrary edge forms within any group is higher than the probability that an arbitrary edge forms between nodes among different groups. Quanlitatively, this will result in:

$$P_{ii} >> P_{ij} \forall i, j$$

4.1 Degree Distribution

We will prove that the degree distribution of the SBM is not scale free, and more precisely, that it follows a Poisson distribution. Firstly,

Given 2 independent Poisson random variables, X and Y, where

$$X \sim \pi(a) \text{ and } Y \sim \pi(b)$$

, where $\pi(\lambda)$ indicates a Poisson distribution

$$X + Y \sim \pi(a + b)$$

The degree distribution of a node in a given cluster of a strongly assortative Stochastic Block Model (SBM) follows a Poisson distribution.

Consider a strongly assortative SBM with n nodes partitioned into k clusters $C_1, C_2, ..., C_k$. Let n_i denote the number of nodes in cluster C_i and p_{ij} be the probability of an edge between a node in cluster C_i and a node in cluster C_j.

Consider a node v in cluster i. The degree of this node can be written as the sum of indicator random variables representing potential edges to other nodes. We can partition these edges into groups based on the clusters that the other nodes belong to.

For each cluster C_j, define X_{ij} as the sum of indicator random variables corresponding to potential edges between v and each node in cluster C_j. Each X_{ij} is a sum of n_j independent Bernoulli random variables with success probability

p_{ij}, and therefore X_{ij} follows a Poisson distribution with parameter $\lambda_{ij} = n_j p_{ij}$ by the Law of Rare Events.

The degree of the node v is then $D = \sum_{j=1}^{k} X_{ij}$, the sum of independent Poisson random variables. Therefore, by the above statement about Poisson distributions, the degree distribution of a node in a given cluster of a SBM follows a Poisson distribution with parameter $\lambda_i = \sum_{j=1}^{k} \lambda_{ij}$.

We then can say the following:

The degree distribution of the entire Stochastic Block Model (SBM) follows a Poisson distribution. Furthermore, if λ_i is the parameter for the Poisson distribution of cluster i and n_i is the number of nodes in cluster i, then the parameter Λ for the Poisson distribution of the entire SBM is given by $\Lambda = \sum_{i=1}^{k} \frac{n_i}{n} \lambda_i$, where $n = \sum_{i=1}^{k} n_i$ is the total number of nodes.

In an SBM, the degree distribution of the entire model is the weighted average of the degree distributions of each of the clusters, with the weights being the proportions of nodes in each cluster.

By Theorem 1, we know that the degree distribution of a node in a given cluster follows a Poisson distribution. Hence, the average of these Poisson distributions will also follow a Poisson distribution, due to the properties of the Poisson distribution and the law of total probability.

Therefore, the degree distribution of the entire SBM follows a Poisson distribution with parameter $\Lambda = \sum_{i=1}^{k} \frac{n_i}{n} \lambda_i$.

4.2 Triples and the Global Clustering Coefficient

Unlike the BA model, we will show that the clustering coefficient of the SBM does not approach 0 with larger graphs, and instead can be bounded below.

Using our result from above, let the degree distribution of our SBM graph G be the Poisson distribution with Poisson parameter Λ. Consider that for a random node v, the number of triples with v as a center is

$$\binom{deg(v)}{2} = \frac{deg(v)(deg(v)-1)}{2}$$

Therefore we can say

$$E[\tau(V)] = \frac{1}{2}E[deg(v)^2] - \frac{1}{2}E[deg(v)]$$

Now, we can utilize that

$$Var(deg(v)) = E[deg(v)^2] - E[deg(v)]^2$$

Since deg(v) follows a Poisson distribution with parameter Λ

$$\Lambda = E[deg(v)^2] - \Lambda^2$$

$$\Lambda + \Lambda^2 = \Lambda(\Lambda - 1) = E[deg(v)^2]$$

Therefore, we have

$$E[\tau(v)] = \frac{1}{2}(\Lambda^2 + \Lambda) + \frac{1}{2}\Lambda$$

.

Now, we make the following observations. Let

$$P = min_{\forall i}(P_{ii}|P_i i > 0)$$

and let

$$Q = min_{\forall i,j}(P_{ij}|P_i j > 0)$$

.

Firstly, note that given any triple exists, we can bound the probability that is is a triangle below by the constant Q, which means for larger graphs, the global clustering efficient $T \geq Q$.

However, this is a very weak bound, and if we utilize the assumption that $P_{ij} << P_{kk}$ for all i, j, and k, we realize that the most common type of triples will be those among large groups, and therefore, we can qualitatively note that P will serve as a better approximator for the global clustering coefficient.

4.3 Expansions

Consider that for strongly assortative SBMs, we can consider sets S to be the individual clusters. (We know if there are more than 1 clusters, we have that atleast one of the clusters are less than half of the size of the total graph). Recall that expansion is defined as

$$h(G) = \min_{S \subseteq V, |S| \leq |V|/2} \frac{|E(S, \bar{S})|}{\min(|E(S)|, |E(\bar{S})|)}$$

In this model, since $R << P$, a majority of edges for any given node are within its own cluster. When we cut through a set S that contains an entire cluster, we mainly cut through the relatively fewer inter-cluster edges, which occur with probability R. Thus, the number of cut edges, $|E(S, \bar{S})|$, is proportional to R times the number of nodes in the in the cluster.

On the other hand, the number of edges within the set, $|E(S)|$, is proportional to P times the number of nodes in the cluster. Therefore, we can bound the expansion from above by a value of the order to R/P.

5 Custom Model

To accommodate the assumptions of both of these models, we attempt to balance the assumptions that lead to the rise of the beneficial properties within each model. Let's list the assumptions necessary for the behavior the data agrees within our model:

1. BA model
   (a) Interactive Growth
   (b) Preferential Attachment
2. Strong Assortative Stochastic Block Model
   (a) Preexisting clustering structure
   (b) Edges within a cluster have higher probability of forming compared to edges outside of a cluster

We implement the following algorithm: Start with an initial graph, $G = G(E_0, V_0)$. We will add a new node each iteration (until a desired number of added nodes, T, is reached). Our model will take in 3 parameters, $0 \leq \delta_1 << \delta_0 \leq 1$ and ϵ - which will be explained later. Each time we add a new node, the following will be done:

1. First, assign the node v to cluster C_i with probability
$$\frac{|C_i|}{\sum_{j=0}^{k} |C_j| + \epsilon}$$
2. Once the node has been assigned to a cluster, C_i, it connects with another node in its cluster, v_0 with probability $\delta_0 deg(v_0)/(deg(C_i))$
3. The probability that it connects to another node, v_1 in a different cluster is probability $\delta_1 \frac{outdeg(v_1)}{|E|}$, where —E— is the total number of edges in the graph, and $outdeg(v_1)$ is the proportion of edges that leave the cluster.

5.1 Group size distribution

First, let's note the growth of groups over time. Firstly, consider a group C_i. We will show that each of these groups grows linearly Let $|C_i|$ be the size of a group. We use differentials to model the continuous expectation of the size of a group as new nodes are added.

$$\frac{\partial |C_i|}{\partial t} = \frac{|C_i|}{t + |V_0| + \epsilon}$$
$$ln(|C_i|) = ln|t + |V_0| + \epsilon| + C$$
$$C_i = C(t + |V_0| + \epsilon)$$

We see from solving this differential equation that each group grows linearly, in proportion to t.

Let's also note that because of the parameter ϵ in the denominator, we have that a node v is not added to any group with probability $\epsilon/t + |V_0|$. We see here that the expected number of groups can be written as

$$k = k_0 + H(t + |V_0|) - H_(|V_0|)$$

This grows like the harmonic series - another factor expressed in our model that reflects the assumption that New AS clusters can occur but tend to become less likely as more and larger clusters are already present in the system.

5.2 Degree Distrubtion

We will show that this model wields a close relative scale-free property.

First, let's write the appropriate differential equation corresponding to the growth rate of the in-cluster node k in group C_I. We will assume the graph has already gone through many iterations, and we can therefore remove the $+ |E_0|$ from our calculations.

$$\frac{\partial k_i}{\partial t} = \frac{|C_i|}{|E| + \epsilon}\delta_0 \frac{k_i}{\sum k_j \in C_i}$$

Since each cluster grows linearly, and each time a new node is added to the cluster, it add an expected value of δ_0 edges, increasing the total degree of the cluster by $2\delta_0$ on each time a new node is added to that cluster.

Now consider that then this differential becomes

$$\frac{\partial k_i}{\partial t} = \frac{|C_i|}{|E| + \epsilon}\delta_0 \frac{k_i}{2\delta_0 \frac{|C_i|}{|E|+\epsilon} t}$$

$$\frac{\partial k_i}{\partial t} = \frac{k}{2t}$$

We have that this differential is exactly that of the BA model where $m = 1$ - as is expected, since in the special case of isolated groups, we simply have a single BA model. We thereby observe that the degree distribution that is proportinal to k^{-3}

Now, lets measure the change in the out degree of a node.

$$\frac{\partial k_o}{\partial t} = (1 - \frac{|C_i|}{|E| + \epsilon})\delta_1 \frac{k_o}{2\delta_1(1 - \frac{|C_i|}{|E|+\epsilon})t}$$

$$\frac{\partial k_o}{\partial t} = \frac{k}{2t}$$

Now, we see a similar construction to above, where the out degree of the node, and can again make some cancellations. Again, per iteration, the expected number of out nodes is δ_1 per com

We can note that the total degree of a node, which we can write as $k_i + k_o$, will be donimated by the k_i term for larger clusters (which will form the majority of our model), so we can approximate our models degree distribution but.

5.3 Clustering behavior, Expansion, and conclusion.

In this model, our clustering coefficient remains low because of a "ports and hubs" structure - where ports are nodes with a high out-degree and nodes are . Firstly, we will make the assumption that since δ_1 is small, the number of triples and triangles outside

We know that the global clustering coefficient grows like $\delta_0 \frac{ln(t)}{t}$, by the same reasoning used in the appendix we used to derive the clustering coefficient of the BA model - just multiplied by the δ_0 constant (since we are neglecting the small count of cross-cluster triangles). We could counteract this by setting δ_0 to be an adaptive model, but doing so would complicate the differential equation in the degree distribution section to a non-homogenous differential equation (and the order could depend on the function chosen to represent δ_0).

However, we can conclude that the expansion of this graph is very close to 0, because again of the assumption that $\delta_1 << \delta_0$, as we can use the same reasoning we did in the SBM - as the expected number of edges out of a cluster grows linearly with time, and the expected number of edges in a cluster grows linearly with time, we have an expansion bounded above by a constant of the order δ_0/δ_1

5.4 Conclusion

While this model doesn't have the exact properties of either the SBM or the BA model, it serves as a healthy medium to maintain a clustered structure while still maintaining the scale-free law. As a philosophical note, this model leads to the generation of "ports" and "hubs", where a port within is a node with high out degree of a cluster, and a hub is a node with high in degree in a cluster. A high δ_0 can be used manually to raise the global clustering coefficient.

Going forward, we realize this model is not perfect. Like the Holmes-Kim model, while our expected global clustering still approaches in 0 in time, our hyperparameters offer tweaks to skew this constant upward. An adaptive δ_0 would lead to a non-zero - but would also lead to much more complex behavior regarding degree distributions as it would no longer be a first-order homogeneous differential equation, as delta itself would change in time - and solving such is beyond the scope of this paper - but we are excited to explore it in the future.

A Full derivation of BA model global clustering metrics

A.1 Degree distribution

A.1.1 Analyzing expected degree for a given node

An important tool that we will use later in this paper is the expected degree of a node per iteration.

We can derive analytically the expected degree distribution of a BA model given any initial parameters, as well as a variance on the random variable α, the exponential in the BA distribution. For a given node, we have (from Network Science by Barabasi) that the degree distribution of the graph (which will be represented by $P(k) \sim k^{-3}$, where k is the degree of a node.

As a brief aside, if we wanted to find the expected degree of a given node in a graph, instead of summing up all of the above over given t, we can instead divide twice the number of edges by the number of nodes,

$$\frac{2(mt + |E_0|)}{t + |V_0|}$$

, where E_0 is the set of all the initial edges and $|V_0|$ is the set of the initial vertices.

A.2 Expected Clustering Coeffficients and Distributions

We can also analytically derive the expectations for the global clustering coefficient and the distribution and expected values of the local clustering coefficient for the BA model. We will start with the global clustering coefficient - and will derive an expression for this using a differential equation to model the expected change.

A.2.1 Global Clustering Coefficient for the BA model

Variables and Definitions Before we start, let's define some variables:

t : number of iterations (time steps)

t_0 : initial number of nodes in the graph

N : total number of nodes in the graph after t iterations ($N = t + t_0$)

m : number of edges added to the graph with each new node

p_i : probability of connecting to node i, which is proportional to its degree $k_i (p_i = \frac{k_i}{\sum_{j=1}^{N} k_j})$

$\tau(t)$: total number of triples after t steps

$\delta(t)$: total number of triangles after t steps

E_i : the set of edges in the graph after i iterations

V_i : the set of nodes in the graph after i iterations

Recall that the global clustering coefficient

$$T(G) = \frac{3\delta(G))}{\tau(G)}$$

Using our notation above, we rewrite this as

$$T(t) = \frac{3\delta(t))}{\tau(t)}$$

Now, let $\delta(t)$ be the expected number of triangles in the graph after t iterations. We're interested in finding the differential equation that describes the change in $\delta(t)$ as new nodes and edges are added. To model the change in this fraction find time dependent functions for our numerator and denominator. In both of our cases, we will follow a similar approach to section 5.4 of Network Science for deriving the expected degree distribution of the BA model.

Evaluating the expected number of triples First, lets evaluate the $E(\tau(t))$. Let $\Delta\tau(t)$ be the number of new triples added on the t'th node. We can write that

$$\tau(t) = \tau(t-1) + \Delta\tau(t)$$

Likewise, using expectation to go from discreet random variable to a continuous variable,

$$E[\tau(t)] = E[\tau(t-1)] + E[\Delta\tau(t)]$$

Now, we can take derivatives with respect to t on both sides to create a model on how our number of triplets can change over time. However, when we take derivatives, realize that $E[\tau(t-1)]$ is a constant given the iteration, so we get

$$\frac{\partial E[\tau(t)]}{\partial t} = \frac{\partial E[\Delta\tau(t)]}{\partial t}$$

So we need only analyze the expectation of $\Delta\tau(t)$ - the number of triangles that are created with the new node - which intuitively makes sense as these define the number of triples added per time, or more closely define a rate. We can split the new triples added into.

Before we go on, we should define the center of a triple

$$\tau = \{\{v_1, v_2, v_3\} \in V, \{(v_1, v_2)(v_3, v_2)\} \in E\}$$

where V is the set of all nodes and E is the set of all edges in a graph G - to be v_2, the node that connects to the other 2 nodes in the triple.

Now, consider that there are two types of triples, those that have v_t as their center, which we will say are captured by the function $\Delta\tau_{v_t}(t)$ and those that do not, which will be captured by the function $\Delta\tau_{\not v_t}(t)$

$$E[\Delta\tau(t)] = E[\Delta\tau_{v_t}(t)] + E[\Delta\tau_{\not v_t}(t)]$$

Therefore we can rewrite our differential as

$$\frac{\partial E[\Delta\tau(t)]}{\partial t} = \frac{\partial E[\Delta\tau_{v_t}(t)]}{\partial t} + \frac{\partial E[\Delta\tau_{\not v_t}(t)}{\partial t}]$$

Now, let the new node being added be called v_t. . The first type is simple to calculate, there are m new edges added each iteration, and if any 2 form a triple, we have $\binom{m}{2}$ new triples. So:

$$\frac{\partial E[\Delta\tau_{v_t}(t)]}{\partial t} = \binom{m}{2}$$

As for $E[\Delta\tau_{\not v_t}(t)]$, consider that every time an edge is added from v_t to any other node, let's call it $v_{t'}$, then deg $(v_{t'})$ triples are formed, as $v_{t'}$ is the center for a triple with one edge connecting to v_t. Consider combining this idea with linearity (we connect n new edges each time):

$$\frac{\partial E[\Delta\tau_{\not v_t}(t)]}{\partial t} = E[\sum_{1}^{m} deg(v \in V_{t-1})] = mE[deg(v \in V_{t-1})]$$

However, we've already calculated

$$E[deg(v \in V_{t-1})]$$

to be

$$\frac{2(mt + |E_0|)}{t + |V_0|}$$

So therefore we have that

$$\frac{\partial E[\Delta\tau_{\not v_t}(t)]}{\partial t} = 2m\frac{(mt + |E_0|)}{t + |V_0|}$$

Wrapping it all up, we therefore have

$$\frac{\partial E[\tau(t)]}{\partial t} = \binom{m}{2} + 2m\frac{(mt + |E_0|)}{t + |V_0|}$$

Now, we can integrate to yield an exact solution:

$$E[\tau(t)] = \binom{m}{2}t + 2m \cdot ((|E_0| - |V_0|m)\ln(t + |V_0|) + mt) + \tau(G_0)$$

However in the limit as $t \to \infty$, we get our differential approahces

$$\frac{\partial E[\tau(t)]}{\partial t} = \binom{m}{2} + 2m\frac{(mt)}{t} = \binom{m}{2} + 2m^2$$

And get an asymptotically linear growth function

$$E[\tau(t)] = (\binom{m}{2} + 2m^2)t$$

Evaluating the expected number of triangles In the numerator, we have that

When a new node (n) is added at time t, it will form a triangle if it connects to two nodes that are already connected. Let's define the expected number of triangles formed by the new node (n) as $\Delta\delta$T(t). By the same reasoning as above we have

$$\frac{E[\partial\delta(t)]}{\partial t} = \frac{E[\partial\Delta\delta(t)]}{\partial t}$$

Now consider again that a new triangle is formed when a new node connects to 2 edges that are already connected. Since the new node v_i connects to m new nodes, it can form $\binom{m}{2}$ seperate pairs of nodes.

Now, X_{ij} be the independent variable representing 1 if nodes i and j are connected by an edge and 0 otherwise. We aim to find $E[X_{i,j}|i,j<t]$, since all the previous nodes we're added before t. However, this is simple. We have a total of

$$m(t-1)+|E_0|$$

distinct edges in a graph, each connecting a distinct pair of nodes, and

$$\binom{t-1+|V_0|}{2}$$

distinct pairs of nodes in the graph. Therefore, we have that

$$E[X_{ij}] = \frac{m(t-1)+|E_0|}{\binom{t-1+|V_0|}{2}}$$

Therefore, by linearity, we have that per time increment, we can represent the expected change in triangles as

$$\frac{\partial E[\Delta\delta(t)]}{\partial t} = \binom{m}{2} E[X_{ij}] = \binom{m}{2} \frac{m(t-1)+|E_0|}{\binom{t-1+|V_0|}{2}}$$

Integrating the above expression using partial fraction decomposition, we get

$$E[\delta(t)] = \int \binom{m}{2} \frac{m(t-1)+|E_0|}{\binom{t-1+|V_0|}{2}} dt = \binom{m}{2} \left[A\ln|t+|V_0|-2| + B\ln|t+|V_0|-3|\right] + \delta(t_0)$$

Where

$$A = \frac{\binom{m}{2}(|V_0|-1)(m(|V_0|-2)+|E_0|)}{(2|V_0|-5)(|V_0|-2)}$$

$$B = \frac{\binom{m}{2}(|V_0|-1)(m(|V_0|-2)+|E_0|)}{(2|V_0|-5)(|V_0|-1)}$$

If we wanted to simplify the above expressions at $t \to \infty$, small $|V_0|$, and small $|E_0|$:, we can simplify to get below:

For small $|V_0|$, the terms inside the absolute value of the logarithms will be dominated by t. So, we can approximate the logarithms as follows:

$$\ln|t + |V_0| - 2| \approx \ln|t|$$
$$\ln|t + |V_0| - 3| \approx \ln|t|$$

Now the expression becomes:

$$\binom{m}{2}[A \ln|t| + B \ln|t|] + \delta(t_0)$$

Combine the terms inside the parentheses:

$$\binom{m}{2}[(A + B) \ln|t|] + \delta(t_0)$$

In the limit $t \to \infty$, the Dirac delta function term, $\delta(t_0)$, will become irrelevant, and we are left with:

$$\binom{m}{2}[(A + B) \ln|t|]$$

The sum of A and B can be computed as follows:

$$A + B = \frac{\binom{m}{2}(|V_0| - 1)(m(|V_0| - 2) + |E_0|)}{(2|V_0| - 5)(|V_0| - 2)} + \frac{\binom{m}{2}(|V_0| - 1)(m(|V_0| - 2) + |E_0|)}{(2|V_0| - 5)(|V_0| - 1)}$$

Since both terms have a common factor of

$$\binom{m}{2}(|V_0| - 1)(m(|V_0| - 2) + |E_0|)$$

, we can rewrite the sum as:

So, the simplified expression in the limit $t \to \infty$, with small $|V_0|$ and small $|E_0|$, is:

$$\binom{m}{2}[(A + B) \ln|t|]$$

A note on approximating the expected global clustering coefficient Dividing our expected values above, we get the horrifying expression:

$$3\frac{\binom{m}{2}[A \ln|t + |V_0| - 2| + B \ln|t + |V_0| - 3|] + \delta(t_0)}{\binom{m}{2}t + 2m \cdot ((|E_0| - |V_0|m) \ln(t + |V_0|) + mt) + \tau(G_0)}$$

In the limit as $t \to \infty$, this becomes

$$\frac{3\binom{m}{2}[(A + B) \ln|t|]}{(\binom{m}{2} + 2m^2)t}$$

Which implies that the growth of the global clustering coefficient is asymptotically of the order $\frac{ln(t)}{t}$ which approaches 0 as we get to infinity.

Now, we've left out two major details in the above, which WE HAVE YET TO GET TO. Consider that the numerator of this fraction, the number of triangles, is the sum of a bunch of 0-1 random variables + a constant in expectation (sum of whther or not two nodes are connnected), and the denominator of this fraction can also be written as the sum Consider that our global clustering coefficient is defined as $T(t) = \dfrac{3\delta(t)}{\tau(t)}$ Since the numerator and denominator of this fraction are not conditionally neccesarily independent given t, by simply dividing the expectations for the

Lastly, if it was only the limiting behavior, we could bound the denominator by the $\tau(t) \geq \tau_{v_t}(t)$, which reduces the deoninator to just $\frac{m}{2}t + \tau(t_0)$, and is no longer a random variable since every term we are adding exactly $\binom{m}{2}$ triples, which means our clustering coefficient still approahces 0, as the denominator can be bounded below by a linear function. Therefore, we can say

$$\lim_{t\to\infty} E(T(t)) \leq \lim_{t\to\infty} E(3\delta(t)) * \frac{1}{\binom{m}{2}t} = 0$$

References

[1] Thomas Alfroy, Thomas Holterbach, and Cristel Pelsser. "MVP: Measuring Internet Routing from the Most Valuable Points". In: *Proceedings of the 22nd ACM Internet Measurement Conference*. IMC '22. Nice, France: Association for Computing Machinery, 2022, pp. 770–771. ISBN: 9781450392594. DOI: 10.1145/3517745.3563031. URL: https://doi.org/10.1145/3517745.3563031.

[2] Albert-László Barabási. *Network Science*. Cambridge, United Kingdom: Cambridge University Press, 2016. URL: http://networksciencebook.com/.

[3] Brian Karrer and M. E. J. Newman. "Stochastic blockmodels and community structure in networks". In: *Physical Review E* 83 (1 2011), p. 016107. DOI: 10.1103/PhysRevE.83.016107. URL: https://link.aps.org/doi/10.1103/PhysRevE.83.016107.

[4] Akmal Khan et al. "AS-Level Topology Collection through Looking Glass Servers". In: *Proceedings of the 2013 Conference on Internet Measurement Conference*. IMC '13. Barcelona, Spain: Association for Computing Machinery, 2013, pp. 235–242. ISBN: 9781450319539. DOI: 10.1145/2504730.2504758. URL: https://doi.org/10.1145/2504730.2504758.

[5] FirstName LastName. "Approximating Clustering-Coefficient and Transitivity". In: (2004). URL: https://i11www.iti.kit.edu/_media/projects/spp1126/files/sw-acct-05.pdf.

[6] Lun Li et al. “A First-Principles Approach to Understanding the Internet’s Router-Level Topology”. In: *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. SIGCOMM ’04. Portland, Oregon, USA: Association for Computing Machinery, 2004, pp. 3–14. ISBN: 1581138628. DOI: 10.1145/1015467.1015470. URL: https://doi.org/10.1145/1015467.1015470.

[7] Stefano Mossa et al. “Truncation of power law behavior in ”scale-free” network models due to information filtering”. In: *Physical Review E* 71 (4 2005), p. 046141. DOI: 10.1103/PhysRevE.71.046141. URL: https://journals.aps.org/pre/pdf/10.1103/PhysRevE.71.046141.

[8] Matthew Roughan et al. “10 Lessons from 10 Years of Measuring and Modeling the Internet’s Autonomous Systems”. In: *IEEE Journal on Selected Areas in Communications* 29.9 (2011), pp. 1810–1821. DOI: 10.1109/JSAC.2011.111006.

[9] *Route Views BGP Data*. https://www.routeviews.org/routeviews/. Accessed: 2023-03.

[10] G. Siganos et al. “Power laws and the AS-level Internet topology”. In: *IEEE/ACM Transactions on Networking* 11.4 (2003), pp. 514–524. DOI: 10.1109/TNET.2003.815300.

[11] *The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 2011-2020*. https://www.caida.org/catalog/datasets/ipv4_routed_24_topology_dataset/.

[12] Shi Zhou and Raul J. Mondragon. “Towards Modelling The Internet Topology – The Interactive Growth Model”. In: Member of IEEE & IEE. Queen Mary, University of London. Mile End Road, London, E1 4NS, United Kingdom, 2003.