Hybrid Power-Law Models of Network Traffic

Pat Devlin[1], Jeremy Kepner[2], Ashley Luo[2], Erin Meger[3]

[1]Yale, [2]MIT, [3]Université du Québec à Montréal

Abstract—The availability of large scale streaming network data has reinforced the ubiquity of power-law distributions in observations and enabled precision measurements of the distribution parameters. The increased accuracy of these measurements allows new underlying generative network models to be explored. The preferential attachment model is a natural starting point for these models. This work adds additional model components to account for observed phenomena in the distributions. In this model, preferential attachment is supplemented to provide a more accurate theoretical model of network traffic. Specifically, a probabilistic complex network model is proposed using preferential attachment as well as additional parameters to describe the newly observed prevalence of leaves and unattached nodes. Example distributions from this model are generated by considering random sampling of the networks created by the model in such a way that replicates the current data collection methods.

Index Terms—networks, Zipf-Mandlebrot, preferential attachment, modelling

I. INTRODUCTION

Recent events have underscored the increasing importance of the Internet to our civilization, necessitating a scientific understanding of this virtual universe [1], [2]. The pandemic induced a drastic increase in Internet usage due to activities such as remote education and work, streaming, social media, online shopping, and video games [3]–[6]. However, with this rise of online activity comes a potential surge of problematic internet usage that is made more urgent by the rising influence of adversarial Internet robots (botnets) on society [7]–[10]. Thus, for scientific, economic, and security reasons observing, analyzing, and modeling the Internet is essential.

The two largest efforts to capture, curate, and share Internet packet traffic data for scientific analysis are the Widely Integrated Distributed Environment (WIDE) project [11] and the Center for Applied Internet Data Analysis (CAIDA) [12]. These data have supported a variety of research projects resulting in hundreds of peer-reviewed publications [13], ranging from characterizing the global state of Internet traffic, to specific studies of the prevalence of peer-to-peer file-sharing, to testing prototype software designed to stop the spread of Internet worms. More recently, novel analysis of

trillions of network observations enabled by high performance sparse matrix mathematics and interactive supercomputing has reinforced the ubiquity of power-law distributions in these observations and enabled precision measurements of the distribution parameters [14]–[17]. The increased accuracy of these measurements allows new underlying generative network models to be explored.

Prior modeling studies primarilyly relied on collecting data through web crawls that naturally produced connected power-law networks [18]–[21]. The foundational preferential attachment (PA) random network model emerged naturally from these observations [22]–[26]. PA models rely on the connectivity of a large component to generate subsequent iterations of the model [27]–[30]. These large components, and other network cores have been studied extensively [16], [22], [31], [32]. WIDE, CAIDA, and other Internet observatories provide different vantage points that can see complete streams of traffic (Figure 1) that reveal significant rare leaves of connected components as well as entirely unattached links (Figure 2) that create deviations from standard traditional power-law models (Figure 3).

At a qualitative level, it is suspected that many of these leaves and unattached links are formed by bot traffic. These connections often behave in unpredictable manners, and tend to form other links only with similar (bot-like) connections. Looking beyond the noise from these bots, the predictable connections still grow and behave in a manner that satisfies the preferential attachment model. Crucially, the internet traffic observed in pipeline is distinct from the actual long-term traffic network of the Internet. Rather, the observed traffic is a random subnetwork of the actual network of "who in general feels like talking with whom". Even though over large time scales, big chunks of Internet traffic might look like preferential attachment, the observed behaviors from the data collection methods will be a random subnetwork of the underlying model.

Our approach to understanding this subnetwork will be with a new model that considers the time iterative selection of random subnetworks of a larger data subset. This model amplifies the randomness of the unattached links and improves our understanding of the internet beyond a preferential attachment setting. This paper will cover the introduction and explanation of the new five-parameter PALU (PA + Leaves + Unattached links) model, a basic analysis of the model including the degrees distribution and analysis of the variability of the model, and conclude with the results and opportunities for future work.

II. Network Observations

The stochastic network structure of Internet traffic is a core property of great interest to Internet stakeholders and network scientists. Of particular interest is the probability distribution $p(d)$ where d is the degree (or count) of one of several network quantities depicted in Figure 1: source packets, source fan-out, packets over a unique source-destination pair (or *link*), destination fan-in, and destination packets [17]. Amongst the earliest and most widely cited results of virtual Internet topology analysis has been the observation of the power-law relationship

$$p(d) \propto 1/d^\alpha$$

with a model exponent $1 < \alpha < 3$ for large values of d [19], [20], [33]. In our work network topology refers to the network theoretic virtual topology of sources and destinations and not the underlying physical topology of the Internet. These early observations demonstrated the importance of a few supernodes in the Internet (see Figure 2) [34]. Measurements of power-laws in Internet data stimulated investigations into a wide range of network phenomena in many domains and lay the foundation for the field of network science [25].

Classification of Internet phenomena is often based on data obtained from crawling the network from a number of starting points [21]. These webcrawls naturally sample the supernodes of the network [34] and their resulting $p(d)$ are accurately fit at large values of d by single-parameter power-law models. Characterizing a network by a single power-law exponent provides one view of Internet phenomena, but more accurate and complex models are required to understand the diverse topologies seen in streaming samples of the Internet.

At a given time t, N_V consecutive valid packets are aggregated from the traffic into a sparse matrix $\mathbf{A}_t$, where $\mathbf{A}_t(i,j)$ is the number of valid packets between the source i and destination j [35]. The sum of all the entries in $\mathbf{A}_t$ is equal to N_V

$$\sum_{i,j} \mathbf{A}_t(i,j) = N_V$$

All the network quantities depicted in Figure 1 can be readily computed from $\mathbf{A}_t$ using the formulas listed in Table I. An essential step for increasing the accuracy of the statistical measures of Internet traffic is using windows with the same number of valid packets N_V. Using packet windows with the same number of valid packets produces aggregates that are consistent over a wide range of windows from $N_V = 100,000$ to $N_V = 100,000,000$. While weights are important to study, in this initial work we are studying the unweighted model. The common weights to study subsequently could be the number of packets or number of bytes sent over a link.

A. Logarithmic Pooling

A network quantity d computed from $\mathbf{A}_t$ produces a corresponding histogram denoted by $n_t(d)$, with corresponding probability
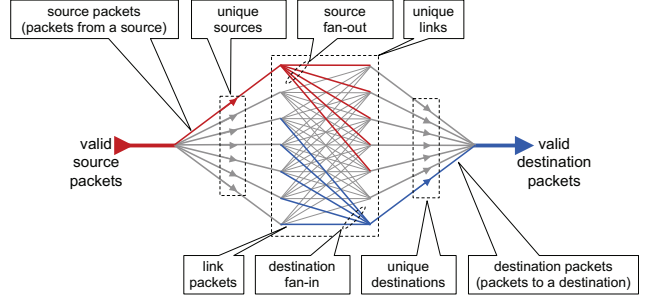
$$p_t(d) = n_t(d) / \sum_d n_t(d)$$

Fig. 1. **Streaming network traffic quantities.** Internet traffic streams of N_V valid packets are divided into a variety of quantities for analysis: source packets, source fan-out, unique source-destination pair packets (or links), destination fan-in, and destination packets [17]
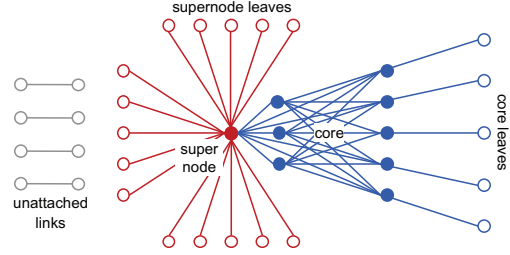

Fig. 2. **Traffic network topologies.** Internet traffic forms networks consisting of a variety of topologies: unattached links, supernode leaves connected to a supernode, densely connected core(s) with corresponding core leaves [17]

and cumulative probability

$$P_t(d) = \sum_{i=1,d} p_t(d)$$

Due to the relatively large values of d observed due to a single supernode, the measured probability at large d often exhibits large fluctuations. However, the cumulative probability lacks sufficient detail to see variations around specific values of d, so it is typical to pool the differential cumulative probability with logarithmic bins in d

$$D_t(d_i) = P_t(d_i) - P_t(d_{i-1})$$

where $d_i = 2^i$ [23]. All computed probability distributions use the same binary logarithmic pooling (binning) to allow for

TABLE I
Aggregate Network Properties

Formulas for computing aggregates from a sparse network image $\mathbf{A}_t$ at time t in both summation and matrix notation. $\mathbf{1}$ is a column vector of all 1's, $^\mathsf{T}$ is the transpose operation, and $|\ |_0$ is the zero-norm that sets each nonzero value of its argument to 1 [36].

Aggregate Property	Summation Notation	Matrix Notation
Valid packets N_V	$\sum_i \sum_j \mathbf{A}_t(i,j)$	$\mathbf{1}^\mathsf{T} \mathbf{A}_t \mathbf{1}$
Unique links	$\sum_i \sum_j \|\mathbf{A}_t(i,j)\|_0$	$\mathbf{1}^\mathsf{T} \|\mathbf{A}_t\|_0 \mathbf{1}$
Unique sources	$\sum_i \|\sum_j \mathbf{A}_t(i,j)\|_0$	$\mathbf{1}^\mathsf{T} \|\mathbf{A}_t \mathbf{1}\|_0$
Unique destinations	$\sum_j \|\sum_i \mathbf{A}_t(i,j)\|_0$	$\|\mathbf{1}^\mathsf{T} \mathbf{A}_t\|_0 \mathbf{1}$

281

consistent statistical comparison across data sets [23], [25]. The corresponding mean and standard deviation of $D_t(d_i)$ over many different consecutive values of t for a given data set are denoted $D(d_i)$ and $\sigma(d_i)$.

B. Modified Zipf-Mandelbrot Model

Measurements of $D(d_i)$ can reveal many properties of network traffic, such as the fraction of nodes with only one connection $D(d = 1)$ and the size of the supernode

$$d_{\max} = \operatorname{argmax}(D(d) > 0) \qquad (1)$$

Effective classification of a network with a low parameter model allows these and many other properties to be summarized and computed efficiently. In the standard Zipf-Mandelbrot model typically used in linguistic contexts, d is a ranking with $d = 1$ corresponding to the most popular value [37]–[39]. To accurately classify the network data using the full range of d, the Zipf-Mandelbrot model is modified so that d is a measured network quantity instead of a rank index

$$p(d; \alpha, \delta) \propto 1/(d + \delta)^\alpha$$

The inclusion of a second model offset parameter δ allows the model to accurately fit small values of d, in particular $d = 1$, which has the highest observed probability in these streaming data. The model exponent α has a larger impact on the model at large values of d while the model offset δ has a larger impact on the model at small values of d and in particular at $d = 1$.

The unnormalized modified Zipf-Mandelbrot model is denoted

$$\rho(d; \alpha, \delta) = \frac{1}{(d + \delta)^\alpha}$$

with corresponding gradient

$$\partial_\delta \rho(d; \alpha, \delta) = \frac{-\alpha}{(d + \delta)^{\alpha+1}} = -\alpha \rho(d; \alpha + 1, \delta)$$

The normalized model probability is given by

$$p(d; \alpha, \delta) = \frac{\rho(d; \alpha, \delta)}{\sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta)}$$

where d_{max} is the largest value of the network quantity d. The cumulative model probability is the sum

$$P(d_i; \alpha, \delta) = \sum_{d=1}^{d_i} p(d; \alpha, \delta)$$

The corresponding differential cumulative model probability is

$$D(d_i; \alpha, \delta) = P(d_i; \alpha, \delta) - P(d_{i-1}; \alpha, \delta)$$

where $d_i = 2^i$.

Minimizing the differences between the observed differential cumulative distributions allows accurate Zipf-Mandelbrot parameters to be selected for a given set of observations. Figure 3 provides a representative sample of the hundreds of fits from [17], illustrating the effectiveness of Zipf-Mandelbrot model in describing these observations.
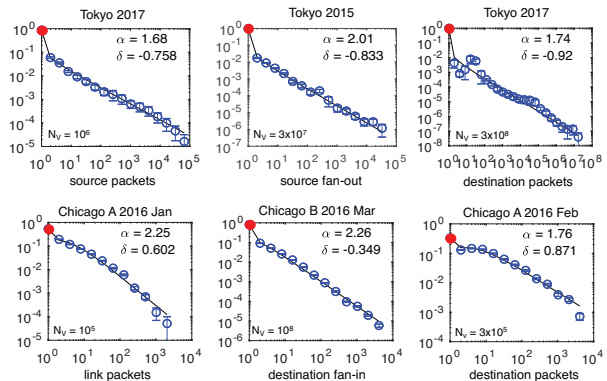

Fig. 3. **Measured Distributions and Model Fits.** A selection of measured differential cumulative probabilities spanning different locations, dates, and packet windows. Blue circles are measured data with ± 1-σ error bars. Red dot highlights the significance of leaves and unattached links. Black lines are the best-fit modified Zipf-Mandelbrot models with parameters α and δ and is an excellent fit in all the examples except the upper right.

III. PA + Leaves + Unattached (PALU) Model

The effectiveness of the empirical Zipf-Mandelbrot model drives the need to explore new underlying generative network models that can produce these more complex probability distributions. The new model extends the PA model with explicit terms accounting for the leaves attached to the PA core as well as unattached links.

The model can be conceptualized in two parts: the *underlying network*, and the *observed network*. The underlying network can be considered as the "true" information of the traffic connections that occur and how frequent these connections are. With our current methods, the full image of this network cannot be detected, and our consideration of this underlying network is theoretical. There are three main pieces that make up this network: the *core* which is constructed by preferential attachment; a set of degree 1 nodes called *leaves* that are adjacent to nodes in the core; and *unattached nodes* that are not connected to the core, and have very low connection within the set itself.

Looking through a window of a certain size with respect to the number of observed connections (packets), a random subnetwork of the underlying network is witnessed, which we call the *observed network*. In this particular set of data, we observe all possible data that passes from one gateway to another until a given number of connections are observed, as described in Section II. For further details on the collection of the data see [17]. Conceptually, we can consider selecting this subnetwork by randomly deleting edges and nodes of the underlying network.

In reality these edge connections are directed since internet communication is directional, however for the sake of the model we will consider this undirected. Using a directed model has a small impact on overall the degree distribution analysis [20]. We will leave this discussion for later in the paper.

A webcrawl is more likely to detect nodes in the core, but

less likely to detect the leaves and the unattached nodes. The measurements taken by MAWI, CAIDA, and others through trunkline observations of trillions of connections reveal these leaves and unattached nodes included in this new model [17].

A. Defining the Model

The PALU model requires four parameters: the clustering of the unattached nodes, the proportions of nodes in each section of the underlying network, the preferential attachment of the core in the underlying network, and the size of the sampled network from the underlying network.

1) Let $\lambda \in [0, 20]$ define the average degree of the unattached nodes in the underlying network
2) The parameters C, L, U represent the proportions of nodes in each of the core, the leaves, and the unattached nodes in the underlying network, conforming to the relationship

$$C + L + U(1 + \lambda - e^{-\lambda}) = 1$$

3) To describe the preferential attachment of the core, we require the parameter $\alpha \in [1.5, 3]$ as the exponent of power-law decay of the degree distribution.
4) The window size is described by the parameter $p \in [0, 1]$ as the proportion of the underlying network that is being observed. As the window size increases, p will get closer to 1. Specifically, this parameter is the probability that an edge in the underlying network will appear (be selected) in the observed network.

The model is completely determined by five parameters, and the relationship in (1) determines the sixth ($d_{\max}$). Importantly, for a given network, the parameters λ, C, L, U, and α should be the same regardless of the window size. As the window size increases, the only parameter that will change is p, as it is more likely to see more edges.

IV. PRELIMINARY ANALYSIS

In the following, we reference the Riemann zeta function, which is a well-known function given by

$$\zeta(\alpha) = \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}}$$

In this paper, α was determined to be $1.5 \leq \alpha \leq 3$ experimentally from the log-log plot of degree distribution in [17]. This function is supported in MATLAB with the built-in function zeta(x), and was determined here to be $1.202 \leq \zeta(\alpha) \leq 2.612$, where this will assist in the probabilistic analysis.

Given a window size p, the fraction of nodes in the underlying network that we expect to see in our observed network is

$$V = \frac{Cp^{\alpha-1}}{(\alpha - 1)\zeta(\alpha)} + Lp + U(1 + \lambda p - e^{-\lambda p})$$

Our model predicts the following about the observed network, where d is taken to be some integer greater than 1

$$\frac{\text{\# core nodes}}{\text{total \# nodes}} \approx \left(\frac{Cp^{\alpha-1}}{(\alpha - 1)\zeta(\alpha)}\right)\frac{1}{V}$$

$$\frac{\text{\# leaves nodes}}{\text{total \# nodes}} \approx Lp/V$$

$$\frac{\text{\# unattached nodes}}{\text{total \# nodes}} \approx \frac{U(1 + \lambda p - e^{-\lambda p})}{V}$$

$$\frac{\text{\# unattached links}}{\text{total \# nodes}} \approx \frac{U \cdot \lambda p \cdot e^{-\lambda p}}{V}$$

$$\frac{\text{\# degree 1 nodes}}{\text{total \# nodes}} \approx \frac{\frac{Cp^{\alpha}}{\zeta(\alpha)} + Lp + U \cdot \lambda p \cdot (1 + e^{-\lambda p})}{V}$$

$$\frac{\text{\# degree } d \text{ nodes}}{\text{total \# nodes}} \approx \frac{\frac{Cp^{\alpha}}{\zeta(\alpha)}d^{-\alpha} + Ue^{-\lambda p}\frac{(\lambda p)^d}{d!}}{V}$$

$$\approx \frac{\frac{Cp^{\alpha}}{\zeta(\alpha)}d^{-\alpha} + Ue^{-\lambda p}\left(\frac{e\lambda p}{d}\right)^d}{V}$$

$$\approx \frac{\frac{Cp^{\alpha}}{\zeta(\alpha)}d^{-\alpha}}{V}$$

The last two approximations are very good when $\log(d) > 1$, in which case we have

$$\log\left(\frac{\text{\# nodes of degree } d}{\text{total \# nodes}}\right) \approx -\alpha \log(d) + \beta$$

where β is a constant independent of d. This provides an effective estimation for α via linear regression in a log-log plot.

A. Logarithmic Pooling

In order to consider a selection of nodes over a logarithmically binned degree interval, simply find the cumulative sum of the corresponding estimate for all the values of d within the interval. For example with $i > 3$, the number of nodes of degree between 2^i and 2^{i+1}, we have the following summation:

$$\frac{1}{V}\sum_{d=2^i}^{2^{i+1}}\frac{Cp^{\alpha}}{\zeta(\alpha)}d^{-\alpha} = \frac{Cp^{\alpha}}{\zeta(\alpha)V}\sum_{d=2^i}^{2^{i+1}}d^{-\alpha}$$

$$\approx \frac{Cp^{\alpha}}{\zeta(\alpha)V}\int_{2^i}^{2^{i+1}}x^{-\alpha}dx$$

$$= \frac{Cp^{\alpha}}{\zeta(\alpha)V}\left(\frac{1 - 2^{1-\alpha}}{\alpha - 1}\right)\cdot\left(2^i\right)^{1-\alpha}.$$

Taking logs, where γ is some constant that does not depend on i, we get this is approximately

$$(1 - \alpha)\log(2^i) + \gamma.$$

283

Importantly, taking intervals of degrees for large i, a $\log$ plot will have the slope of the regression line as $1 - \alpha$, and not $-\alpha$ as it would be in the non-interval case. When i is small (e.g., $1 \leq i \leq 3$), the approximation for the number of nodes between degree a and degree b with $2 \leq a, b \leq 2^3$ becomes

$$\sum_{d=a}^{b} \frac{\# \text{ nodes of degree } d}{\text{total } \# \text{ nodes}}$$

and the right-hand-side can be computed by summing up all the terms, carefully approximating with the $d!$ term and the Lp term when $d = 1$. This differs from the large-i estimate in that for large i, we can discard small terms on the right-hand-side and safely estimate the sum by an integral. In both Fig 3 and Fig 4 our parameters are given for the underlying probability distribution, but we are plotting the differential cumulative distribution which will result in power law exponent one unit higher [40].

B. Simplified Degree Distributions

The ratios for distinct degrees d can be simplified using the following parameters

$$\text{if } d = 1, \quad \frac{\# \text{ degree 1 nodes}}{\text{total } \# \text{ nodes}} \quad \approx \quad c + l + u \qquad (2)$$

$$\text{if } d \geq 2, \quad \frac{\# \text{ degree } d \text{ nodes}}{\text{total } \# \text{ nodes}} \quad \approx \quad c \cdot d^{-\alpha} + u \left(\frac{\Lambda}{d}\right)^d \quad (3)$$

$$\text{if } d \geq 10, \quad \frac{\# \text{ degree } d \text{ nodes}}{\text{total } \# \text{ nodes}} \quad \approx \quad c \cdot d^{-\alpha} \qquad (4)$$

where c, l, u, and Λ are constants that do not depend on d

$$
\begin{aligned}
c &= C p^\alpha / \zeta(\alpha) V \\
l &= L p / V \\
u &= U \exp(-\lambda p) / V \\
\Lambda &= e \lambda p
\end{aligned}
$$

For c, l, and u, each is proportional to the number of nodes in the core, unattached, and leaves respectively. These three constants do depend on the parameter p, related to the window size. The parameter Λ depends on p as well and is related to the clustering. The constant α is the same as the above section. All of the parameters above should be positive.

To fit these parameters, all of which are in terms of the discrete parameters, we can consider the following

(a) One first fits (4) to the long-term behavior of the degree distribution. A log-log plot will have a linear behavior whose slope is equal to $-\alpha$ and constant term equal to $\log(c)$, as we will see in the discussion in the following section. This will result in fitting c and α.
(b) Subsequently (3) will fit small values of d ($\log(d) < 10$) estimating u and Λ.
(c) It is then possible to solve for l exactly by using (2).

In (b) it is possible to fit the parameters by subtracting the $cd^{-\alpha}$ term from both sides of (3) and then summing up both sides will give a value of roughly $u \cdot (e^\Lambda - 1 - \Lambda)$, which would be a more robust estimate than the point-wise estimates of (3).

After having computed c and α from (4), computing

$$\sum_{d=2}^{\infty} d \left[\frac{\# \text{ degree } d \text{ nodes}}{\text{total } \# \text{ nodes}} - cd^{-\alpha} \right]$$

should be equal to roughly $(e^\Lambda - 1)\Lambda u$. Thus, resulting in

$$\frac{\displaystyle\sum_{d=2}^{\infty} d \left[\frac{\# \text{ degree } d \text{ nodes}}{\text{total } \# \text{ nodes}} - cd^{-\alpha} \right]}{\displaystyle\sum_{d=2}^{\infty} \left[\frac{\# \text{ degree } d \text{ nodes}}{\text{total } \# \text{ nodes}} - cd^{-\alpha} \right]} \approx \Lambda + \frac{\Lambda^2}{e^\Lambda - \Lambda - 1}$$

Computing the two summations on the left-hand-side from the data can be used to to approximate Λ by numerically solving the above.

The advantage to this methodology is that it presumably reduces the estimate to one with substantially less variance. Estimating the right-hand side analytically, would provide roughly accurate estimates for large Λ, but for $\Lambda \approx 0$ the estimates would become roughly $2 + \Lambda/3$, by expanding its Taylor series. After having an estimate of Λ, c, and α, (3) can be used to estimate u by a linear regression.

A benefit to the efficacy of the model is that an arbitrary choice of u and Λ would not lead to an inaccurate computation. In choosing these variables arbitrarily, the network would maintain the structural topology for $d = 1$ since the information provided in choosing l will mostly determine the structure. The model would be even more accurate for large values of d by (3), and for the very few values of d in between, the chosen values u and Λ could be modified until the desired results were acquired.

V. DERIVATION OF THE ESTIMATES

The estimates from the simplistic model follow immediately from the corresponding estimates in the refined model. For the derivation, the aim is to count nodes rather than the ratios determined with the variables C, L, and U. Let N be the number of nodes in the underlying network. Rigorously, we let C_N be the number of nodes in the core and L_N be the number of leaf nodes in the underlying network. We will carefully define U_N later. We begin the model of the observed network by beginning with the underlying network. We obtain our observed subnetwork by retaining each edge independently with probability p, creating an Erdős–Rényi random subnetwork of the underlying network [29].

The core nodes of the underlying network are assumed to be generated according to a preferential attachment model, independent of the other parts of the underlying network. The number of core nodes of the underlying network having degree d follows a power-law distribution of the form $d^{-\alpha}/\zeta(\alpha)$. Thus, the number of degree $d > 0$ nodes in the observed network is well-approximated by

$$pd^{-\alpha}/\zeta(\alpha)$$

Any degree d node in the underlying network will have degree distributed in the observed network according to the binomial distribution

$$Bin(d, p) \approx dp \pm \sqrt{d(1-p)p} \approx dp$$

To estimate the number of core nodes with degree greater than 0 in our observed network, we sum these estimates, which we in turn approximate as a Riemann integral, at the cost of a negligible error term. The number of leaf nodes in the observed network has a binomial distribution with mean pL_N and variance $(1-p)pL_N$, and the random variable is approximated by its mean [29].

Returning to the rigorous derivation of the unattached nodes. In this part of the model we generate U_N-many stars, each of which has a random number of non-central nodes, where the number of non-central nodes is given by independent identically distributed Poisson random variables with mean λ. The Poisson random variable represents the modeling of the U_N central nodes and a large number of potential leaves, each of which chooses to attach to a central node with some sufficiently low probability. With this distribution, the total number of nodes in the unattached portion of the underlying network has a distribution precisely given by

$$U_N + \sum_{j=1}^{U_N} Po(\lambda).$$

Since the sum of independent Poisson random variables is again Poisson, this simplifies to

$$U_N + Po(U_N\lambda).$$

However, of these central nodes, there will be $Bin(U_N, e^{-\lambda})$ of them which are isolated nodes. As these cannot be seen by examining traffic between nodes, we remove these nodes from this model. The model demonstrating a non-zero number of isolated nodes gives significant evidence to the existence of these nodes in the true picture of the internet, regardless that they cannot be observed in our current methods of data collection.

Estimating the degree sequence within our observed network amounts to the following. If we first sample $Y \sim Po(\lambda)$ and then we take the sum of Y independent $\{0, 1\}$-Bernoulli random variables of mean p, we seek to find the resulting distribution, that is to analyze $Bin(Po(\lambda), p)$. This is well-known to be simply $Po(\lambda p)$—as seen from an elementary computation—and additionally motivated our choice of Poisson random variables.

All to say, the unattached version of our observed subnetwork is distributed as U_N stars each of which have $Po(\lambda p)$ non-central leaves, from which the results follow immediately by independence and estimating the terms by their means, which is valid large values of U_N.

VI. ZIPF-MANDELBROT CONNECTION

Finally, a one-parameter (unnormalized) approximation of Equation (3) can be used to compare with the empirical Zipf-Mandelbrot distributions. Using the approximation

$$(\Lambda/d)^d \approx r^{(1-d)}$$

where r is a positive number to be fit to the data. This approximation effectively changes the underlying distribution from Poisson to an equally valid Geometric distribution, resulting in the distribution

$$cd^{-\alpha} + ur^{(1-d)}$$

which can be rescaled as

$$d^{-\alpha} + (u/c)r^{(1-d)}$$

The above expressions can be aligned with the Zipf-Mandelbrot parameters by setting

$$u/c = (1+\delta)^{-\alpha} - 1$$

resulting in

$$\text{PALU}(d) \propto d^{-\alpha} + r^{(1-d)}((1+\delta)^{-\alpha} - 1) \qquad (5)$$

In Figure 4, the $\text{PALU}(d)$ degree distribution is shown with a selection of curve families that demonstrate the the PALU model can be made to fit a Zipf-Mandlebrot distribution for d. For any given power law exponent α and offset parameter δ, the Zipf-Mandelbrot distribution can be well-approximated by Equation (5) by varying r. In general, the model $\text{PALU}(d)$ tends towards Zipf-Mandlebrot. In addition, the $\text{PALU}(d)$ model has the potential to explain some observations that deviate from the Zipf-Mandlebrot distribution (see Figure 3 upper right).

Connecting back to the original values of u and c, we have

$$u/c = (U/C)\exp(-\lambda p)\zeta(\alpha)/p^\alpha$$

and combining with the same u/c expression in $\text{PALU}(d)$ gives

$$(1+\delta)^{-\alpha} = (U/C)\exp(-\lambda p)\zeta(\alpha)p^{-\alpha} + 1$$

Thus, the model described in Equation 5 captures the PA aspect of the model mostly in the first term and the rest of the model is captured via r and δ through the second term.

VII. CONCLUSIONS

In this paper, we presented a modified preferential attachment model to better represent new observations of network topology observed from streaming data. From the figures, we found the PALU model to fit the Zipf-Mandelbrot distribution very well given the right parameters.

With the PALU model as a stepping stone, we can continue to examine the space of the network and more effective models, including comparing the data from [17] to the model itself, extended the analysis done in Fig 4. Some opportunities for future work that we are interested in include further investigation of the PALU model such as its various applications in other big data environments and deeper study into
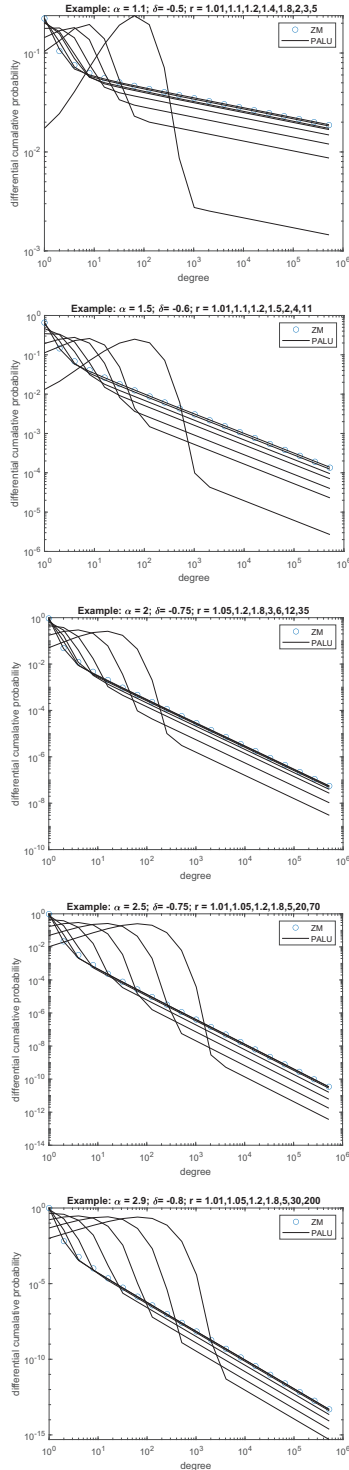
Fig. 4. **PALU model curve families.** Examples of degree distributions that are possible with the PALU model using varying parameters, and how they relate to their base Zipf-Mandelbrot (ZM) differential cumulative distributions. From the top to the bottom we vary the power law exponent α between 2 and 3. In each, δ is the model offset for each α, and in each figure, we vary r to create the family of curves, as described in section V with Equation (5). Further work will involve comparing these families with the model.

the degree distribution and clustering coefficients. The PALU model research can also extend to the case of weighted edges where potential weights could be the number of packets or number of bytes sent along a link. Exploration of other models are also possible, such as combining preferential attachment with the Erdos-Renyi model and determining if there is a better fitting model than the Zipf-Mandelbrot distribution. It would be worthwhile to explore the existence and importance of isolated nodes, and analyzing aspects of the network by studying the behaviors of a random sampling or a preferential attachment graph. Finally, extrapolating the results of the PALU model to observe and define the large clusters of small disconnected components may also be of interest.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, p. 1200970, 2011.

[2] B. Li, J. Springer, G. Bebis, and M. H. Gunes, "A survey of network flow applications," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 567–581, 2013.

[3] W. Ali, "Online and remote learning in higher education institutes: A necessity in light of covid-19 pandemic.," *Higher Education Studies*, vol. 10, no. 3, pp. 16–25, 2020.

[4] M. A. Almaiah, A. Al-Khasawneh, and A. Althunibat, "Exploring the critical challenges and factors influencing the e-learning system usage during covid-19 pandemic," *Education and Information Technologies*, vol. 25, pp. 5261–5280, 2020.

[5] M. Candela, V. Luconi, and A. Vecchio, "Impact of the covid-19 pandemic on the internet latency: A large-scale study," *Computer Networks*, vol. 182, p. 107495, 2020.

[6] N. Pandey, A. Pal, *et al.*, "Impact of digital surge during covid-19 pandemic: A viewpoint on research and practice," *International Journal of Information Management*, vol. 55, p. 102171, 2020.

[7] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[8] "https://go.neosit.com/bot-report-2019."

[9] J. Healey and R. K. Knake, *Zero botnets: Building a global effort to clean up the internet*. Council on Foreign Relations, 2018.

[10] W. Ahmad, "Why botnets persist: Designing effective technical and policy interventions," 2019.

[11] K. Cho, K. Mitsuya, and A. Kato, "Traffic data repository at the wide project," in *Proceedings of USENIX 2000 Annual Technical Conference: FREENIX Track*, pp. 263–270, 2000.

[12] K. Claffy, "Internet tomography," *Nature, Web Matter*, 1999.

[13] "http://www.caida.org/data/publications/."

[14] V. Gadepally, J. Kepner, L. Milechin, W. Arcand, D. Bestor, B. Bergeron, C. Byun, M. Hubbell, M. Houle, M. Jones, *et al.*, "Hyperscaling internet graph analysis with d4m on the mit supercloud," *IEEE High Performance Extreme Computing Conference (HPEC)*, 2018.

[15] A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, *et al.*, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," *IEEE High Performance Extreme Computing Conference (HPEC)*, 2018.

[16] J. Kepner and H. Jananthan, *Mathematics of big data: Spreadsheets, databases, matrices, and graphs*. MIT Press, 2018.

[17] J. Kepner, K. Cho, and K. Claffy, "New phenomena in large-scale internet traffic," *arXiv preprint cs.NI/1904.04396*, 2019.

[18] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking (ToN)*, vol. 2, no. 1, pp. 1–15, 1994.

[19] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, p. 130, 1999.

[20] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[21] C. Olston, M. Najork, *et al.*, "Web crawling," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 3, pp. 175–246, 2010.

[22] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical review E*, vol. 64, no. 2, p. 025102, 2001.

[23] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[24] A.-L. Barabási, "Scale-free networks: a decade and beyond," *science*, vol. 325, no. 5939, pp. 412–413, 2009.

[25] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[26] P. Sheridan and T. Onodera, "A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions," *Scientific reports*, vol. 8, no. 1, p. 2811, 2018.

[27] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[28] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pp. 171–180, 2000.

[29] F. Chung and L. Lu, *Complex graphs and networks*. No. 107, American Mathematical Soc., 2006.

[30] A. Bonato, H. Chuangpishit, S. English, B. Kay, and E. Meger, "The iterated local model for social networks," *Discrete Applied Mathematics*, 2020.

[31] R. Yang and P. Bogdan, "Controlling the multifractal generating measures of complex networks," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[32] Y. Xue and P. Bogdan, "Reliable multi-fractal characterization of weighted complex networks: algorithms and implications," *Scientific reports*, vol. 7, no. 1, pp. 1–22, 2017.

[33] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, ACM, 2005.

[34] J. Cao, Y. Jin, A. Chen, T. Bu, and Z.-L. Zhang, "Identifying high cardinality internet hosts," in *INFOCOM 2009, IEEE*, pp. 810–818, IEEE, 2009.

[35] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.

[36] J. Karvanen and A. Cichocki, "Measuring sparseness of noisy signals," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 125–130, 2003.

[37] B. Mandelbrot, "An informational theory of the statistical structure of language," *Communication theory*, vol. 84, pp. 486–502, 1953.

[38] M. A. Montemurro, "Beyond the zipf–mandelbrot law in quantitative linguistics," *Physica A: Statistical Mechanics and its Applications*, vol. 300, no. 3-4, pp. 567–578, 2001.

[39] O. Saleh and M. Hefeeda, "Modeling and caching of peer-to-peer traffic," in *Network Protocols, 2006. ICNP'06. Proceedings of the 2006 14th IEEE International Conference on*, pp. 249–258, IEEE, 2006.

[40] E. P. White, B. J. Enquist, and J. L. Green, "On estimating the exponent of power-law frequency distributions," *Ecology*, vol. 89, no. 4, pp. 905–912, 2008.