Generative Models of Conversation Trees as a Preferential Attachment Stochastic Process

Tanner Amundsen
The Johns Hopkins University Applied Physics Laboratory
Laurel, Maryland
tamunds1@jhu.edu

***Abstract*—Online discussion forums are increasingly popular ways for humans to communicate. The ability to model and generate this discourse as conversation trees provides insight into what makes certain posts and topic popular and how these online discussion threads grow. We model online discussion threads from Reddit as conversation trees and fit the data to two discrete-time, preferential attachment stochastic process models. We find that the degree distribution of this data is well-suited to preferential attachment modeling and, empirically, these two models generate synthetic conversation trees that accurately summarize the growth dynamics of the original data set. We propose a new model that builds upon existing research by adding a parameter for a post's "provocativeness". However, we do not find this new model to improve accuracy compared to existing models.**

***Index Terms*—preferential attachment, branching process, conversation trees, cascades, social media**

I. Introduction

Online discussion forums have gained immense popularity in recent years. Users gather online to communicate around shared interests, comment on news, and collaborate on projects. These forums are often organized in "threads" which are ordered collections of dialogue, connected together by "reply trees". These threads have become a topic of interest for social scientists and computer scientists alike as social media sites grow in popularity.

One particular research area concerns modeling the generative process of these conversation trees. An obvious mathematical model for these tree-like data structures is the branching process (or Galton Watson process). A "root" user begins the conversation with some initial post. Subsequent posts are either replies to the original root post or replies to other replies. Modeling how these trees grow requires some knowledge of the distribution of the degree of each node. That is, how many replies will a given post have based on characteristics about that node.

The ability to model the growth of comment threads offers several benefits. First, understanding what makes a post popular can inform the marketing strategy of people and business that use discussion forums for advertising. Models can indicate which characteristics have positive or negative correlations with generating more comment. Second, social media sites that architect these forums are often interested in highlighting specific posts and threads that will grab users' attention on some hand-selected "feed". Generative conversation tree models can help social media sites predict when a tree will continue to grow or when it will die out. A third application of this research is the ability to compare the parameters of conversation trees among different communities.

II. Background

A. Galton Watson Models

The simplest Galton-Watson processes assume a constant and common probability distribution $p(x)$ for the number of children each node will generate [1]. The key requirement of this (simplest) model is that this probability distribution does not change from individual form individual, nor from generation from generation. Using the notation of generating functions, this model can be summarized:

$$g_n(t) = g_{n-k}(g_k(t)) \quad \text{for } k = 1, 2, ..., n-1$$

However, under this assumption, parameterizing the branching process is simple: identify the expected value of offspring for each node and call it m. Using m, one can determine the probability of extinction of the branching process, the size of each generation, and many other useful metrics. This model proves to be insufficient for modeling online conversation threads as branching processes where several factors like popularity, recency, and originality play in role in attracting replies. Not only is the degree distribution, then, different from node to node, but it also changes over time.

B. Preferential Attachment

Beyond Galton-Watson model, other types of stochastic processes have been identified as modeling real life social networks quite well. In particular, it has been observed that many social graphs exhibit degree distributions that follow an approximate power law [2]. These networks, sometimes called "scale-free" networks exhibit small diameters, low clustering, and heavy tailed degree distributions [3]. The phenomenon that leads to this kind of network is termed preferential attachment. The generative model associated with this stochastic process is called the Barabási-Albert model.

The simplest form of the Barabási-Albert model, sometimes called the "rich-get-richer" model, relates the probability that an incoming node will attach to node i as

$$p_i = \frac{d_i}{\sum_{j=1}^{n} d)_j}$$

That is, highly connected nodes are more likely to get attached to than less connected nodes. This model was created for graphs that don't necessarily preclude cycles. Conversation trees, on the other hand, must be directed acyclic graphs (DAGs) so generative models for these kinds of graphs tend to enforce that restriction. Node degree is only one way to parameterize attachment preference and the next section discusses other ways to do that.

C. Parameterizing Post Attractiveness

Extensive research exists in the area of random graphs, Markov chains, and branching processes where the degree distribution is some function of each nodes "attractiveness". Generative models for discussion threads build upon this theory to develop models that parameterize the attractiveness of individual posts within discussion threads.

In [4], the author surveys seven state of the art generative models for modeling online discussion threads as stochastic processes with preferential attachment. Several of these generative models rely on continuous time data of when each comment was posted. Others modeled the process as a discrete time series. Within the realm of discrete time models, this paper first focuses on fitting the model proposed of [5] to Reddit Comment threads. This model provides a distribution over all nodes in the thread at time $t-1$ of the probability that the incoming node at time t will attach to those nodes. The model factors in

1) Popularity - the degree of all nodes at time t (parameterized by α)
2) Novelty - the age of all nodes at time t (parameterize by τ)
3) Root bias - accounts for the tendency of new posts to reply to the original post by default (parameterized by β)

Formally, the probability of the next parent of incoming node X_t follows the probability

$$P(X_t = k|\alpha,\tau,\beta) = \frac{\alpha deg_k + \beta\delta_{0,k} + \tau^{age_k}}{\sum_{t=2}^{k} \alpha deg_t + \beta\delta_{0,t} + \tau^{age_t}} \tag{1}$$

Where the $\delta_{0,k}$ is the Kronecker delta function indicating if node k is the root node or not.

D. Inclusion of Message Type Bias

This paper is interested in seeing how well the model in [5] fits to our data. We are also interested in augmenting this model to include a 4th parameter that captures how provocative a post is. We hypothesize that posts that are clearly seeking a response are more likely to garner comments than more neutral posts.

In particular, we are interested biasing the degree distribution to more than just the root node (the β parameter in [5]). In particular, we hypothesize that message *type* is a possible indicator of a message's overall attractiveness for subsequent commenters. By "type" we are referring to whether a node is a question, an answer, a reaction, an elaboration, a complaint, etc. The dataset chosen for this research contains

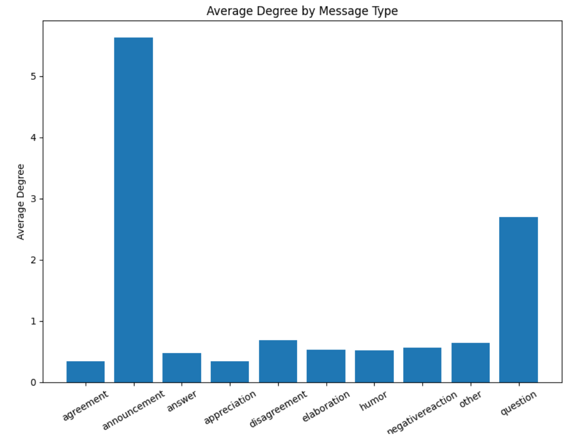

Fig. 1. Node degree by post type. Message type was human-labeled using one of the following labels: Announcement, Elaboration, Humor, Appreciation, Question, Answer, Agreement, Negative Reaction, Disagreement, Other. Note the clear outliers in average degree among question type and announcement type posts. Source: [6]

human-labeled characterizations of each comment. In our preliminary analysis (see Figure 1) we found that comments that are formulated as questions or announcements are more likely to garner a response. We propose an additional term be added to the generative model in (1) to paramaterize a node's "provocativeness":

$$P(X_t = k|\alpha,\tau,\beta,\rho) = \frac{\alpha deg_k + \beta\delta_{0,k} + \tau^{age_k} + \rho\delta_{?,k}}{\sum_{t=2}^{k} \alpha deg_t + \beta\delta_{0,t} + \tau^{age_t} + \rho\delta_{?,t}} \tag{2}$$

Where $\delta_{?,k}$ is the Kronecker delta function indicating whether a comment was labeled as provocative.

Throughout the rest of the paper, we refer to (1) as Model 1 and (2) as Model 2.

III. Dataset

A. Data Collection

The dataset used for this model comes from HuggingFace "Coarse Discourse" [6]. This corpus contains data scraped from public comment threads on Reddit. The total dataset contains roughly 9,000 threads comprising over 100,000 comments and was manually annotated via paid crowdsourcing. These conversation threads themselves were chosen randomly from all over the site. Reddit is organized into "subreddits" which are forums organized around a specific topic. Each thread in the dataset starts with an initial post and subsequent posts follow it with a field indicating who each comment is replying to.

The dataset most obviously applies to discourse classification for ML models because each comment is hand-labeled with a "type" field indicating if the comment was a question, answer, elaboration, appreciation, agreement, negative reaction etc. However, we are interested in modeling this data as a discrete time, preferential attachment stochastic process using the two models described earlier.

B. Tree Formation

Given that every comment has a field indicating which other comment they are replying to, and because the comments were scraped in time order, we are able to create a discrete time tree representation of each thread. One succinct way of representing discrete time trees is the parent vector representation used in [5] where a vector π with length equal to the size of the thread minus 1. Each node is identified by the discrete integer time step at which it appeared. The value at π_i corresponds to the parent of the comment posted at $t = i$. This representation is illustrated in Figure 2.

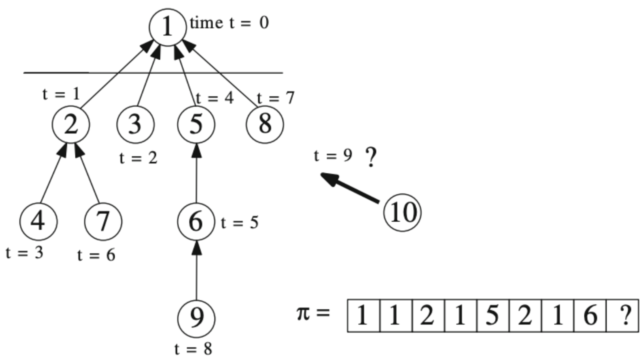

Fig. 2. A small example of how a tree can be represented using parent vector representation π. Source: [5]

C. Thread Characteristics

Figures 3 and 4 show the distribution of thread size and depth in our database, respectively. We observe that both size and depth have heavy-tailed distributions. Next in figure 5 we see the distribution of thread size versus average depth.

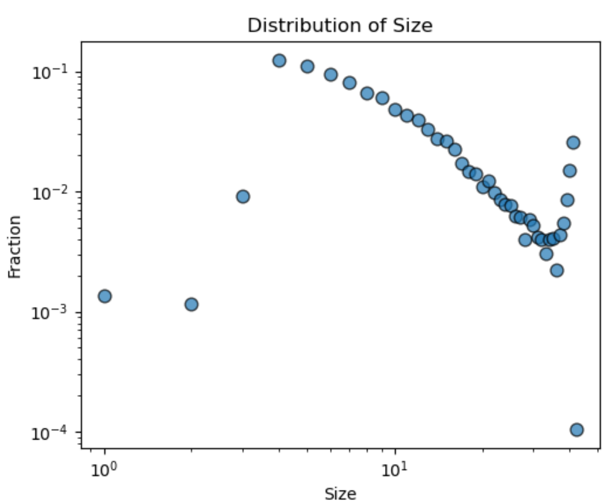

Fig. 3. Distribution of thread size (total number of comments) for Reddit Coarse Discourse. Source: [6]

Figure 6 plots the distribution of node degree which, in this case, is the number of direct replies to a node plus one

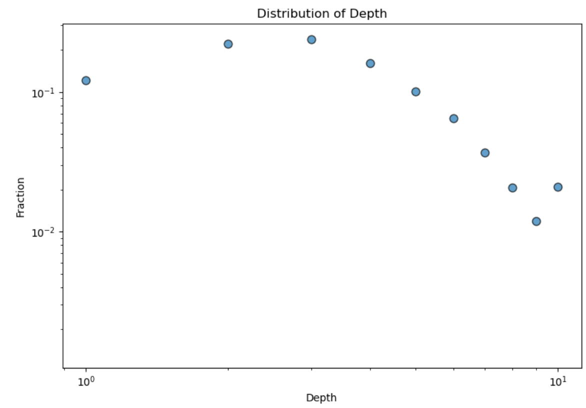

Fig. 4. Distribution of thread depth (longest distance from root to leaf) for Reddit Coarse Discourse. Source: [6]

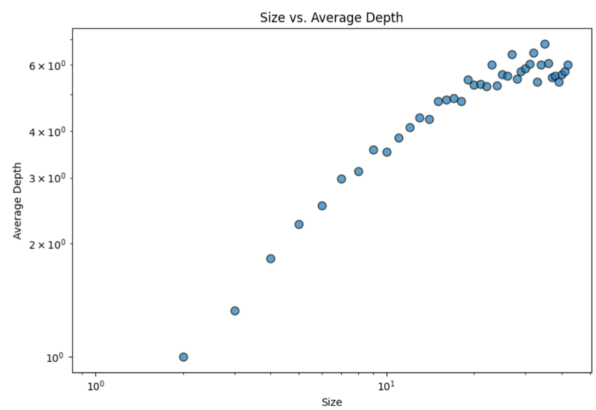

Fig. 5. Size versus average depth (longest distance from root to leaf) for Reddit Coarse Discourse. Source: [6]

since every node except the root has an additional attachment to its parent. We observe that the distribution of degree follows an approximate power law i.e. $p(d) \propto d^{-\alpha}$ for some α. This behavior is a good indicator that some type of preferential attachment generative process is at work in network creation. Some variation of the Barbási-Albert model will fit better to our data than other social network models like the Erdős–Rényi graph model or the Galton-Watson process model [7] [2].

The summary statistics of these tree characteristics are tabulated below.

Parameter	Mean	Standard Deviation
Thread Size	12.23	9.65
Thread Depth	3.56	2.03
Node Degree	0.92	2.04

We note here that the discussion thread sizes are smaller than many of the average discussion thread sizes analyzed in [4] and other existing literature. We acknowledge this as a limitation in Section VI.

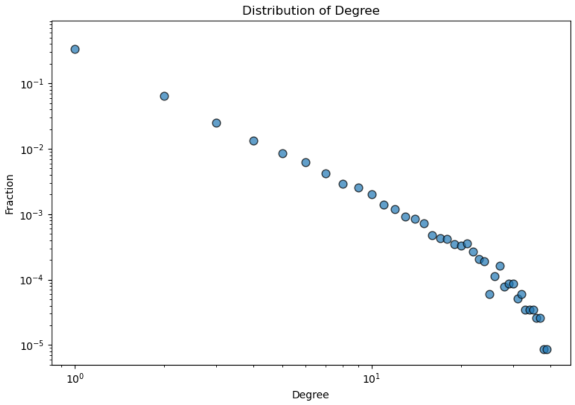

Fig. 6. Degree distribution all comments in Reddit Coarse Discourse. Note that the degree distribution follows an approximates power law sequence. A result that is expected for time series random trees with preferential attachment models dependent on node degree/popularity. Source: [6]

IV. METHODS

A. Parameter Estimation

The work of fitting a generative model to a preferential attachment stochastic process can be done analytically (if certain assumptions are made) or performing some kind of gradient-descent based optimization algorithm on the log likelihood function of your model.

1) Analytical Approaches: The analytical approach for parameter estimation in preferential attachment stochastic processes often relies on evaluating the distribution of the node degree. In [8] the authors propose an analytical result for the degree distribution of where preference is dependent on node depth. They find that the distribution $f_d^{(n)}$ of nodes with degree d at the nth layer of the tree has an algebraic tail with a depth-dependent exponent:

$$f_d^{(n)} \sim d^{-\gamma(n)}, \qquad \gamma(n) = 2 + \frac{n-1}{\langle n-1 \rangle} \tag{3}$$

where the brackets denote an average over all nodes. Other models (sometimes called the "rich get richer" models) construct the degree distribution of nodes at time $t+1$ to be positively correlated with their degree at time t. In chapter 8 of [9], the authors discuss preferential attachment models of random graphs. They denote the state of a random graph at time t to be PA_t, the degree of node k to be $D_k(t)$ and a discrete time-ordered vertex as v_t. They establish the growth rule as:

$$P(v_{t+1} \to v_k | PA_t(1,\delta)) = \frac{D_k(t)+\delta}{t(2+\delta)} \tag{4}$$

where in the notation $PA(m,\delta)$, m is the number of new connections made at each time step and δ is a parameter of the graph. In words, this rule defines the probability that an incoming node will attach to an existing node. They also prove that for $m = 1$ and $\delta > -1$, that

$$E[D_k(t)+\delta] = (1+\delta)\frac{\Gamma(t+1)\Gamma(k-\frac{1}{2+\delta})}{\Gamma(t+\frac{1+\delta}{2+\delta})\Gamma(i)} \tag{5}$$

Which $\implies \exists\, a_m$ such that

$$E[D_k(t)] \sim a_m \left(\frac{t}{k}\right)^{1/(2+\delta)} \quad \text{for } k,t \text{ large} \tag{6}$$

Where the last results comes from Stirling's approximation applied to (5). The final result in (6) is powerful and allows us to bound the degree distribution of each node at each time step by a power law.

Now, to estimate the parameters of Model 1 proposed in [5], the authors use the above results from [9] to bound the expected value of the number comments under a given comment k at a given time t:

$$\left(\frac{t}{k}\right)^{1/2} < E[d_{k,t}] < C\left(\frac{t}{k}\right)^{1/2} \tag{7}$$

Where C is derived from the recursive definition of the model defined in Equation (1) and is upper bounded by an exponential term

$$e^{\frac{\tau}{1-\tau}}$$

where τ is the novelty parameter of the model. However, since C ranges from 1 to ∞, the bound in (7) proves to be too rough in practice for analytical estimation of the degree distribution. Given that, we instead used maximum likelihood estimation for the parameter estimation of Models 1 and 2.

2) Maximum Likelihood Approach: The more practical method for estimating the parameters of Models 1 and 2 is optimizing a maximum likelihood function using observed data. We first make a few assumptions about the data: we assume that the threads in the dataset are independent. Define $\Pi := \{\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, ..., \boldsymbol{\pi_N}\}$ to be the set of all N threads in parent vector notation $\boldsymbol{\pi}_i$ with respective size $|\boldsymbol{\pi_i}|$ and we want to estimate the parameters $\boldsymbol{\theta} := \{\alpha, \beta, \tau, \rho\}$. We define ϕ_1 and ϕ_2 as the attractiveness functions defined in numerators of the Models 1 and 2 respectively. We also define $Z^{(1)}$ and $Z^{(2)}$ to be the normalizing factors for each of those probabilities. Then the set of parameters $\hat{\boldsymbol{\theta}}$ that best estimate the data Π is defined by the likelihood function:

$$
\begin{aligned}
L(\mathbf{\Pi}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\pi_i|\theta) \\
&= \prod_{i=1}^{N} \prod_{t=2}^{|\pi_i|} p(\pi_{t,i}|\boldsymbol{\pi_{1:(t-1),i}}, \boldsymbol{\theta}) \\
&= \prod_{i=1}^{N} \prod_{t=2}^{|\pi_i|} \frac{\phi_1(\pi_{t,i})}{Z_{t,i}^{(1)}} \quad \text{For Model 1} \\
&= \prod_{i=1}^{N} \prod_{t=2}^{|\pi_i|} \frac{\phi_2(\pi_{t,i})}{Z_{t,i}^{(2)}} \quad \text{For Model 2}
\end{aligned}
$$

Applying the log function and changing the sign we obtain the negative log likelihood function.

$$
-\log L(\mathbf{\Pi}|\boldsymbol{\theta}) = -\sum_{i=1}^{N} \sum_{t=2}^{|\pi_i|} \log \phi_1(\pi_{t,i}) - \log Z_{t,i}^{(1)} \quad \text{For Model 1}
$$

$$
-\log L(\mathbf{\Pi}|\boldsymbol{\theta}) = -\sum_{i=1}^{N} \sum_{t=2}^{|\pi_i|} \log \phi_2(\pi_{t,i}) - \log Z_{t,i}^{(2)} \quad \text{For Model 2}
$$

which serves as the function to minimize. We used a bootstrap approach with 23 samples of size 5000 drawn with replacement from the Coarse Discourse comment trees. The algorithm we used to minimize the loss function for each bootstrap sample was the Scikit-Learn implementation of the Limited memory BFGS (LM-BFGS) [10]. Future work should use more bootstrap samples and investigate the merits of different optimization techniques.

3) Results of Negative Log-Likelihood Minimization: For Model 1, the results of the optimization yielded the following parameters:

Parameter	Mean	Standard Deviation
α	0.0505	0.1076
β	1.1778	0.4688
τ	0.7009	0.0762

For Model 2, the results of the optimization yielded the following parameters:

Parameter	Mean	Standard Deviation
α	0.1252	0.2344
β	1.0839	0.4536
τ	0.6955	0.0836
ρ	0.1907	0.2454

4) Synthetic Tree Generation: Using the parameter estimates for both of our models, we were able to create synthetic conversation trees. We are interested in creating synthetic trees from our model in order to compare the goodness of fit of both models. However, there are more practical applications of synthetic tree generation already discussion in Section I. Below is the algorithm we implemented for synthetic tree generation (1)

1) For each tree π in the training data (in parent vector form):
   a) Let $v = \text{length}(\pi)$.
   b) If we are using Model 2, let p be a boolean vector representing which nodes of tree π are a "provocative" type.
   c) Let $s = [1]$ initialize the synthetic tree (in parent vector form) as just a root node and a second node as a child of the root node.
   d) For t from 2 to v:
      i) Let π_t be π from time $[0, t)$ and p_t be p from time $[0, t]$
      ii) For each node k in π_t:
         A) Calculate the normalized attractiveness node k in t_k. Using $\phi_1(\pi_{t,k})/Z_1(t_k)$ for Model 1 and $\phi_2(\pi_{t,k}, p_{t,k})/Z_2(t_k, p_t)$ for Model 2
      iii) Let Φ_t be the normalized attractiveness of each node that exists in π_t (calculated above)
      iv) Sample over Φ_t as a discrete random variable. Let k_t^* be the node chosen
      v) Append k_t^* to s indicating that k_t^* is the parent of the node incoming at time t

By the end of this algorithm, we have roughly $9,000$ synthetic conversation trees for both Model 1 and Model 2. Doing descriptive analysis on these trees as we did with the initial training data set can help us measure the performance of both models.

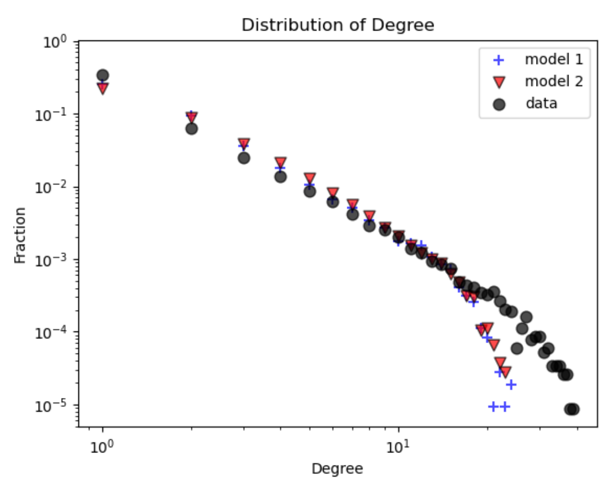

Fig. 7. Degree distribution of (1) the original Reddit dataset, (2) synthetic trees generated with Model 1, and (2) synthetic trees generated with Model 2. Both models generate synthetic trees with very similar degree distributions (approximate Power Law). Source: [6]

V. Discussion and Conclusion

The parameter estimate for Models 1 and 2 are very similar for the α, β, and τ terms. The clearest conclusion from these model estimates is a high degree of root bias. The estimates for β in both models indicate that posters on Reddit are very likely to comment under the original post. When comparing the β terms to the α terms, it is important to note that with small conversation trees (average of 12.23 comments per thread), a high root bias likely means that the most popular node will also be the root node. This means that both models are susceptible to a multicolinearity between the popularity term α and the root bias term β. Multicolinearity occurs when two or more parameters in your model are correlated. Correlated features can create instability in training, where the values of the two parameters oscillate and lead to divergence [11]. Future work should attempt to measure and address this multicolinearity. In [5], they create 3 alternative to their full model by taking out 1 of the 3 parameters, α, β, τ. A similar approach here may help address the multicolinearity.

τ being less than 1 in both models indicates that older nodes are less attractive than newer ones. That is, posters on Reddit appreciate novelty and are are more likely to comment under a more recent post than an older one. In Model 2, our estimate of ρ supports our initial analysis and indicates that 'question' type and 'announcement' type posts are indeed more attractive than posts of different types. We don't however find that adding this term helps the predictive or generative power of Model 2 over Model 1 as we see in the results of synthetic tree generation and as we will discuss further below.

Figure 7 shows the degree distributions of that original data, and both sets of synthetic trees generating using Model 1 and Model 2. We observe that both models generate trees with very similar degree distributions. The behavior is, again, an approximate Power Law as is expected in graphs with preferential attachment. We also observe that both models tend to lose accuracy when it comes to generating nodes with very high degrees of connectivity. That is, the training data had a least a few instances of nodes with 12 or 13 children, but the synthetic trees produced fewer node with such high degree. The degree distributions of the Model 1 and Model 2 trees are very similar. Model 1 approximates the data slightly better in the lower domain of degree, whereas Model 2 approximate the data slightly better at high degree values.

Figure 8 shows a plot of size versus average depth of the original conversation trees, and the two sets of synthetic conversation trees. This plot shows that both models generate sythetic trees with mostly similar size to depth ratios. Both models tended to generate "shorter" and "fatter" trees meaning that incoming nodes were more likely to attach to nodes at higher levels than to leaf nodes. Practically, this means that the models generated synthetic conversation threads that were slightly less linear that the Reddit conversation threads. Synthetic comments were more likely to go under the original post or to replies already under the original post. This plot also shows Model 1 as fitting the original data slightly better than Model 2.

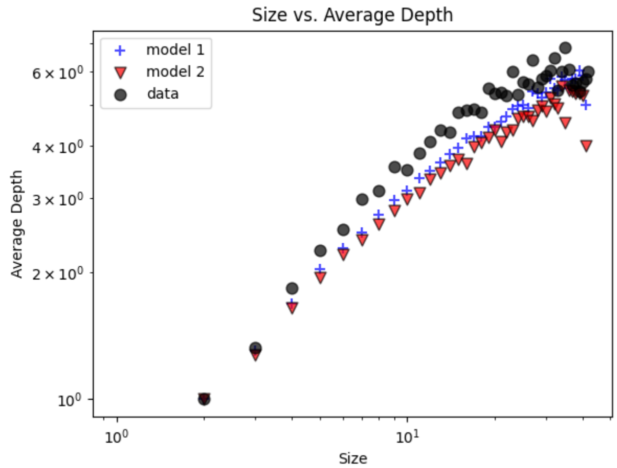

Fig. 8. Tree size versus average depth of (1) the original Reddit dataset, (2) synthetic trees generated with Model 1, and (2) synthetic trees generated with Model 2. Both models generate synthetic trees with very distributions of size versus average depth. Source: [6]

VI. Limitations and Future Work

We acknowledge that certain aspects of this study were limited in scope. Firstly, the data set included trees that are much smaller and shallower than (1) as is actually common on Reddit and (2) other data sets used in research on generative models for conversation trees [4]. We expect that both of our models' performances would decrease with larger thread sizes where thread linearity is more common. Another limitation of this study was the relatively small number of bootstrap samples used in fitting the model to that data. Given existing literature on the topic, a hyperparameter analysis would help us evaluate different bootstrap sample sizes to monitor the convergence of our model. Another limitation of our study was the bounds of our optimization when estimating the parameters using negative-log-likelihood minimization. The L-BFGS minimization algorithm requires bounds to be placed on parameters you are estimating for quicker convergence. Existing literature on the topic indicated that a bound on the α, β, and τ terms of $[0, 100]$ was reasonable for our application [5] [4]. However, future work should experiment with different bounds, especially on novel terms like ρ, to see if the same results would be obtained.

As mentioned in Section V, future work should be performed to address possible multicolinearity in the models used. Another area of interest for future work could be the inclusion of authorship into the model. Much of the existing research on generative models of conversation threads distinguish between posts made by the same author in the same thread. This study did not consider authorship at all however it can add predictive and generative power to the model. A third area of future interest would be to slightly modify the algorithm for synthetic tree generation. Instead of using existing message type vectors to inform the ρ parameter on provocativeness, one could generate message type as a Bernoulli random variable based on some empirical distribution. This would offer more diverse message type compositions in the synthetic trees and decouple tree size and type since tree size is already being taken from the existing data.

References

[1] A. Gut, *An Intermediate Course in Probability by Allan Gut.*, 2nd ed., ser. Springer Texts in Statistics. New York, NY: Springer New York, 2009.

[2] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, pp. 47–97, 2002.

[3] H. Hexmoor, *Computational network science : an algorithmic approach / Henry Hexmoor.*, 1st ed., ser. Computer Science Reviews and Trends. Waltham, Massachusetts: Morgan Kaufmann, 2015 - 2015.

[4] P. Aragón, V. Gómez, D. García, and A. Kaltenbrunner, "Generative models of online discussion threads: state of the art and research challenges," *Journal of Internet Services and Applications*, vol. 8, no. 1, p. 15, October 5 2017. [Online]. Available: https://doi.org/10.1186/s13174-017-0066-z

[5] V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner, "A likelihood-based framework for the analysis of discussion threads," *World Wide Web*, vol. 16, no. 5, pp. 645–675, November 1 2013. [Online]. Available: https://doi.org/10.1007/s11280-012-0162-8

[6] A. X. Zhang, B. Culbertson, and P. Paritosh, "Characterizing online discussion using coarse discourse sequences," in *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ser. ICWSM '17, 2017.

[7] B. Fotouhi and M. G. Rabbat, "Degree correlation in scale-free graphs," *The European physical journal. B, Condensed matter physics*, vol. 86, no. 12, 2013.

[8] E. Ben-Naim and P. L. Krapivsky, "Stratification in the preferential attachment network," *Journal of Physics A: Mathematical and Theoretical*, vol. 42, no. 47, p. 475001, nov 2009. [Online]. Available: https://dx.doi.org/10.1088/1751-8113/42/47/475001

[9] R. v. d. Hofstad, *Random Graphs and Complex Networks*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.

[10] R. BYRD, P. LU, J. NOCEDAL, and C. ZHU, "A limited memory algorithm for bound constrained optimization," *SIAM journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[11] A. S. Goldberger, *A course in econometrics / Arthur S. Goldberger.* Cambridge, Mass: Harvard University Press, 1991.