# Analysis 4 Instructions

# Degree Distributions and Preferential Attachment Models

### Final version due in Laulima Saturday March 5th

### 75 points

In these problems we analyze natural networks through plotting degree distributions, fitting distributions to the power law, and modeling with preferential attachment models. The objective is to identify real world processes and constraints that may have generated the networks. We use prior techniques of visualization and metrics to assist in the interpretation. The assignment builds on previous demonstrations and uses utility functions we have defined.

#### Template

You will use an .Rmd template I provide to construct your response. This template also has further guidance on response format, but does not have the detailed discussion below, so you should follow these instructions while working with the template.

#### Utilities

The template loads utility functions that you should include in your assignment folder. You may also write utility functions for plotting with legends, and making metrics tables, but are not required to do so. If you do so, it is your choice whether to place them in the Utility folder or in-line in the .Rmd. I would put the definition in the .Rmd if I wanted the reader to be able to easily inspect it, and in the Utility folder otherwise.

#### A Caution Concerning Sampling Random Models

You will be sampling from random models in this assignment. Ideally we would take many samples and average results before making claims about what the models predict. We aren’t doing this because students have varying programming backgrounds, and it would take a long time to run and compute metrics on (say) 100 samples of each model.

If you just take one sample it is likely typical but it could be atypical. To avoid drawing unwarranted conclusions based on an atypical sample, I suggest that you run the model a few times to get a feel for what it typically produces. Then, when writing up your results do not hard-code the metrics in your text: use R variables to embed the actual values. This is because the number may change when Knit re-runs all the code to generate the html. So, for example, rather than saying “the average degree is 3.78”, assign the actual value to a variable in a code chunk:

```{r} avgdeg <- mean(degree(g)) … ```

and embed in your prose a reference to the variable like: “the average degree is ```r avgdeg```”.

### Part A. Analyzing and Interpreting the Structure of a Real World Network (30 pts)

In this section, we use our analytic toolkit to characterize a natural network. This section revisits methods we learned in all three of the previous analysis assignments (visualization, metrics, and random models), but also adds interpretation of degree distribution plots and power law fits.

### Western States Power Grid

The file Networks/power\_grid.gml contains an undirected unweighted representation of the topology of the Western States Power Grid of the United States, compiled by Duncan Watts and Steven Strogatz. We can presume that the nodes are major generators and distribution centers and the edges are the physical power lines connecting them, often over long distances at great expense. The network was taken from the web site of Prof. Watts at Columbia University (http://cdg.columbia.edu/cdg/datasets, but this site is no longer available). This citation is given:

D. J. Watts and S. H. Strogatz, "Collective dynamics of `small-world' networks", Nature 393, 440-442 (1998).

You will first characterize the network structure with the tools available to us, and then discuss your interpretation in terms of domain processes or constraints:

### 1. Visualize the network (5 pts)

**(a)** Do a visualization in Gephi. Show degree with node size. Use ForceAtlas 2 with LinLog mode and set Gravity to make it sufficiently spread out that one can see the structure but not so spread out that it makes poor use of the space available. Make PDF and include it in your analysis folder. (3)

**(b)** What is/are salient feature(s) of this visualization that makes sense in terms of it being a power grid? (2)

### 2. Construct configuration model and compare on metrics (8 pts)

From here on all work is in R/igraph.

**(a)** Load the network in R/igraph, and const a configuration model using sample\_degseq with method “vl”. This will enable you to see which values are expected at random given the degree sequence.

**(b)** Compute standard metrics on the natural network and on the configuration model: average degree, log(|V|) for comparison, global transitivity, degree correlation (assortativity\_degree), and mean distance. Display the results in summary format, such as by using a data frame or tibble (recommended). Also display a table of component sizes for the natural network.

### 3. Plot degree distribution (4 pts)

Plot the degree distribution (it is the same for the configuration model) using both of these:

* Linear binning with log-log axes (2)
* Cumulative with log-log axes (2)

Use appropriate x axis adjustment. Include in your linear binning plot a legend with basic metrics pertaining to degree distribution: kmin and kmax. (Other metrics are optional.)

### 4. Fitting to Power Law (5 pts)

**(a)** Fit the distribution to the power law using fit\_power\_law, taking the defaults at first but adjusting if needed to include adequate data. Display a table of the results $alpha, $xmin, $logLik, $KS.stat, and $KS.p (converting to a tibble is an easy way to do this). (3)

**(b)** Does this network display a power law distribution? Interpret the above results (not just p) to conclude how well the distribution fits the power law, and for what portion (domain) of the distribution. Then re-run with xmin chosen manually. (2)

### 5. Discussion (8 pts)

**Discuss the real world network structure using the above results**, taking everything you did (visualization and metrics) into account. In answering the following questions, your discussion should make the connection between the visualization and metric results and the domain being modeled (what one would want to do when constructing a power grid):

**(a)** What connectivity is predicted by the G(n,m) random model (compare mean degree to log(|V|), and why is the connectivity of the power grid so different? What non-random design consideration would make this so? (2)

**(b)** What mean distance is predicted by the G(n,m) random model? Why is this value so different for the power grid? Does the degree distribution explain the difference? What else might explain it? (2)

**(c)** What would explain having the transitivity and small degree saturation shown by your analysis, as compared to the random models? (Hint: consider reliability issues in a real world power grid.) (2)

**(d)** How does the degree distribution for higher degrees compare to predictions for scale free networks? (See Barabasi formula 4.18 and consider typical gamma of 2.5 to 3.5.) How might you explain this result in terms of domain considerations? (2)

### Part B. Modeling the Degree Distribution of a Real World Network (45 pts)

In this section, we try to model a natural network with variations of the Preferential Attachment model. These models are primarily concerned with matching natural degree distributions: they do not have much to say about other measures, so, unlike in the previous question, we will focus on modeling degree distribution, including minimum and maximum values, plotted features, and fit to power law. You will do these things:

* Discuss what domain processes might be present, and which plausibly have preferential attachment dynamics.
* Plot both in-degree and out-degree distributions for the network, and discuss whether the plots provide evidence for the domain processes identified above.
* Fit models to try to get the same distribution using sample\_pa\_age, modifying model parameters that correspond to some of the above domain processes.
* Once fit, interpret what worked and what didn't work to draw conclusions about the domain processes hypothesized above. (What did not work is just as important as what did work.)

### High-Energy Physics Citation Network

We will work with the High-Energy Physics Theory citation network (cit-HepTh.gml). The documentation says:

Arxiv HEP-TH (high energy physics theory) citation graph is from the e-print arXiv and covers all the citations within a dataset of 27,770 papers with 352,807 edges. **If a paper i cites paper j, the graph contains a directed edge from i to j**. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this.

The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP-TH section.

The data was originally released as a part of 2003 KDD Cup.

Initially I had difficulty modeling this network with the preferential attachment code that I wrote and demonstrated in class. I soon realized that it was because my code used total degree, while this is a directed graph and in the natural network *two distinct processes produce the in-degree and out-degree distributions*. **In-degree** indicates how many other papers chose to cite a given paper over time, while **out-degree** indicates how many papers a given paper cites at publication time. It is difficult to come up with a clean process model for two processes added together, so we need to analyze and model in-degree and out-degree separately.

### 6. Discuss domain processes that may be present (14 pts)

We will think first, before we run any code. Before we start looking for processes that might affect degree distribution, we should consider whether it even makes sense to expect these processes to be present. So, first we consider how each of 6 processes *might* be manifest in the domain of paper citations if they were present (without yet knowing whether they actually are). Since there are different processes behind in-degree and out-degree, we discuss each separately:

**(a)** For **in-degree (being cited)**, what **domain** **processes**, if any, might be understood as examples of the following (also identify those that do not apply): (6)

* Growth?
* Preference of attachment to higher in-degree?
* Nonlinearity in this preference of attachment?
* Fitness?
* Initial appeal?
* Citation preferences sensitive to aging?

**(b)** For **out-degree (citing others)**, what **domain** **processes**, if any, might be understood as examples of the following (also identify those that do not apply): (6)

* Growth?
* Preference of attachment to higher out-degree?
* Nonlinearity in this preference of attachment?
* Fitness?
* Initial appeal?
* Citation preferences sensitive to aging?

**(c)** Given the above, how plausible is it that a preferential attachment model will model in-degree and why? How plausible is it that a preferential attachment model will model out-degree and why? (2)

### 7. Plot HEP degree distributions with metrics and discuss - (12 pts)

Now we look at the actual distribution, and think further before we attempt to model.

**(a)** Plot both in-degree and out-degree, using linear binning and cumulative with log-log axes for both.

* Include in your linear binning plot a legend with basic metrics pertaining to degree distribution: kmin, kmean, and kmax.
* Include in your cumulative plot a legend with results of fit\_power\_law including xmin, apha, KS.p and KS.stat.

Other metrics are optional. (4)

**(b)** Which of the processes you identified in question 6(a) as being present *might* **explain features you see in the in-degree plot**? Refer explicitly to the features of your plots from (a) in your response, indicating how they are evidence that the process is present or absent. You may also discuss other features of the plot that strike you as requiring an explanation. (4)

**(c)** Which of the processes you identified in question 6(b) as being present *might* **explain features you see in the out-degree plot**? Refer explicitly to the features of your plots from (a) in your response, indicating how they are evidence that the process is present or absent. You may also discuss other features of the plot that strike you as requiring an explanation. (4)

*(This is a theoretical discussion, but it suggests what we might try in the empirical modeling work below. You will get credit for any conclusions that are grounded in features of the plot and relate them to hypothesized processes: you need not say exactly what I said.)*

### 8. Fitting a Preferential Attachment Model (13 pts)

By modeling, we test our hypotheses.

**Find a model using sample\_pa\_age that fits the HEP in-degree** distribution and metrics as closely as possible, by varying appropriate parameters (read the documentation). Try nonlinear preference of attachment, aging, and initial appeal, *guided by your own reasoning in the previous questions*. Try them separately to see which might be plausible alone, and then try combinations to see which might be working together and to get your best match. Run the same plots and metrics on the model as you did on the original data in #7 to compare them.

Your first priority should be to fit the shape of the HEP and model in-degree distributions and k-max as closely as possible, followed by fitting x-min and gamma for fit\_power\_law (which will not fit the entire distribution, but should be similar for the portions it fits in both the original data and your model). If you can't match all these exactly, explore the tradeoff and then choose what to prioritize, explaining why. This will help you decide which processes are empirically plausible, to be discussed in question 9.

*(I have a solution, but you might find a different solution: You will get credit for a good match that is well justified regardless of whether it is the same as mine.)*

*Why we do not emphasize the other metrics:* k-min will never match HEP's value of 1, because every node we add has at least m edges. We don't have much control over transitivity, which will be too low, nor over mean geodesics and assortativity, which will be similar across models having similar degree distributins. (You may also consider them if you wish. Geodesics are slow to compute: I suggest leaving this out until you are close to your final model.) Over the next few weeks we will examine processes behind transitivity and assortativity that aren't modeled by sample\_pl or sample\_pl\_age.

**(a)** Provide (in R Markdown code blocks) the call to sample\_pl\_age that produces the final model you settled on, displaying the parameters you set, and also degree distribution plots and metrics. As you did for the natural network, include in your linear binning plot a legend with basic metrics pertaining to degree distribution: kmin, kmean, and kmax. Include in your cumulative plot a legend with results of fit\_power\_law including xmin, alpha, KS.p and KS.stat. (10)

**(b)** Then provide a brief summary of what you tried. What did you try with the sample\_pa\_age parameters to test each of the processes from question 2? Which worked and which didn't? (This discussion will be purely in terms of changing model parameters to match the distribution: leave discussion of the domain for the next question.) (1)

### 9. Interpretation (8 pts)

Having attempted to model, we think again about what the models told us.

**Discuss what your modeling work in #8 says about processes in the real-world domain of HEP citations** that you discussed in #6. **Identify relevant results**: refer to the shape of the curve on the plots and the parameters you had to set to get this shape (low degree saturation and/or high degree cutoff if present) with similar x-min, gamma and k-max. **Say what these results tell you about citation practices in the domain**. By "in the domain" I mean you should **talk about what people do, not what your network or analysis does.** Bring it back to the domain your client cares about: that is your job as a data scientist.

**(a)** What does your modeling suggest (identify relevant results) about the **linearity or non-linearity of attachment** in this domain, and why does this make sense in terms of what people do?

**(b)** What does your modeling say (identify relevant results) about whether there is **initial attractiveness** in this domain, and why does it make sense in terms of what people do?

**(c)** What does your model say (identify relevant results) about citation preferences sensitive to **aging** in this domain, and why does it make sense in terms of what people do?

**(d)** What is your conclusion concerning which of the above are operating **in combination**? Are there **other results** you obtained (or could not obtain) in your modeling process that lead to conclusions not covered above?

#### Issues from 2019

* Main point most students missed: The point of modeling is to see what in your model matches the domain, and hence potentially indicates an explanation of domain processes; and to identify what does not match, indicating unexplained processes. Most students' "Interpretation" was just a re-hash of what they concluded from the originally given network plots (or worse, stories they had just made up, not grounded in anything empirical), without taking into account what they had to do to model this network. Then why bother to model?
* Most points taken off in previous assignments were for not taking it back to the domain.
* A model may or may not match the domain: learn from failure as well as success and don't force a bad model.
* Consider why we cite a paper: not just because it is cited a lot by others: it has to be relevant. Also, authors like to cite their own papers.
* Many confuse preference with age, discussing aging as if it is simply the fact that links acquire nodes over time. Aging is a *change in the rate* at which nodes acquire links over time, with older papers either getting more or less than younger papers with the *same* degree.
* Use mode="in" and mode="out" everywhere it is needed. If you plot both together you are not doing this assignment.
* Don’t throw away your own work! Too often students treat questions as isolated exercises and don't see how my problem sets are designed to build up understanding. The answer to each question should build on the prior question, such as (for example) deciding how to try to model HEP in igraph.