Power Law on Network Component Sizes

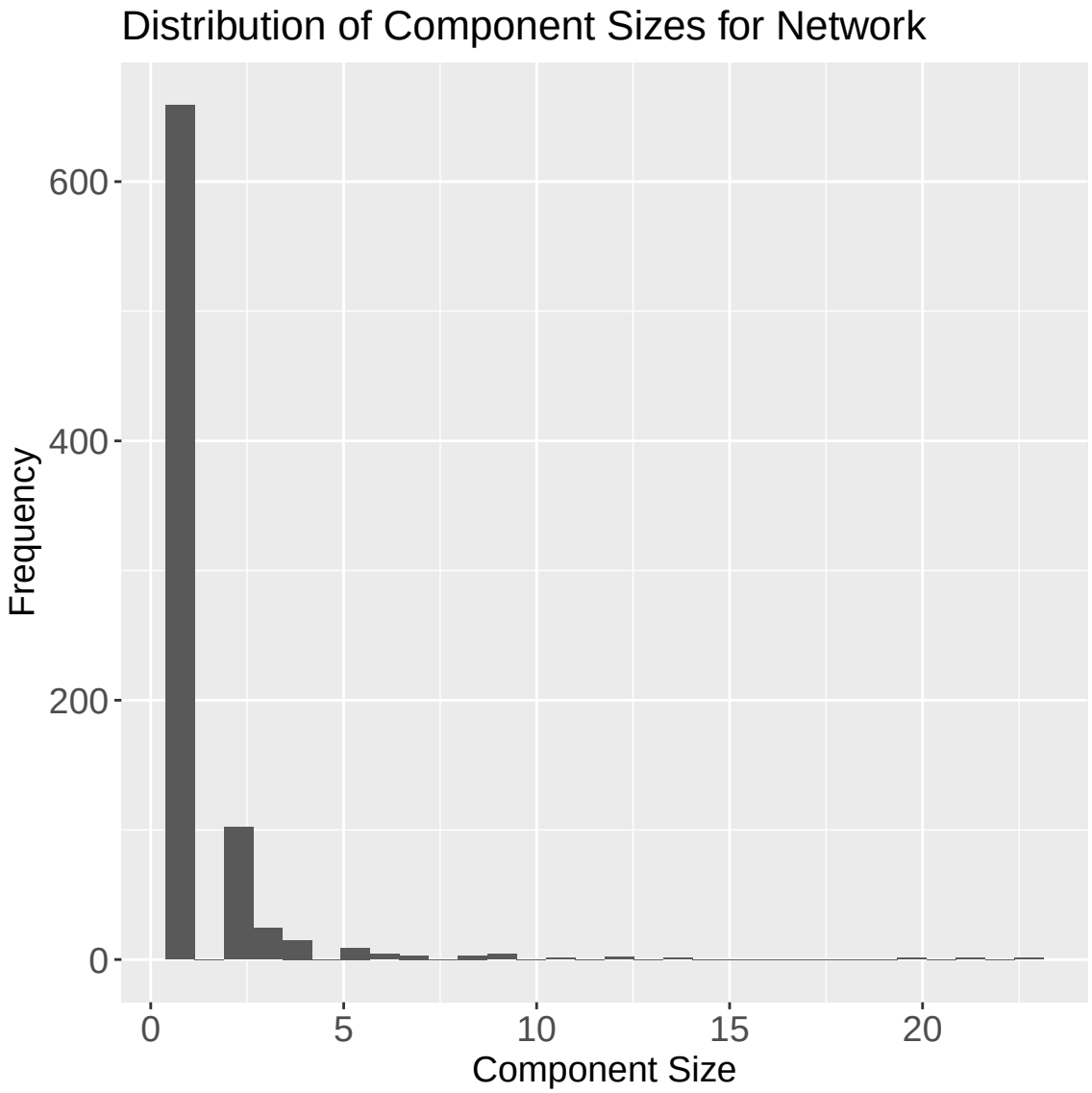
Gabrielle Lemire's project

Philip Turk

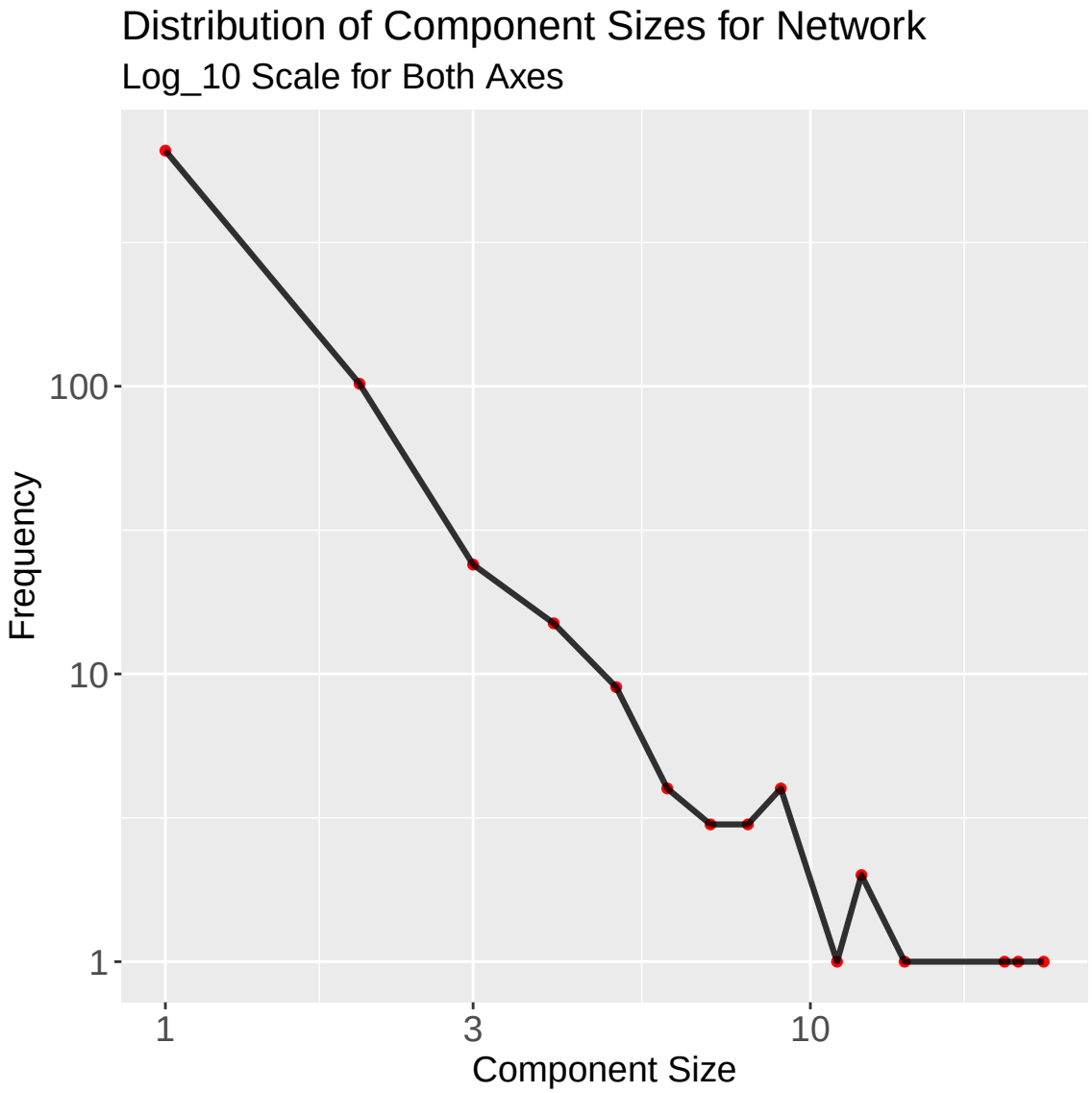
2021-08-17

Part 1

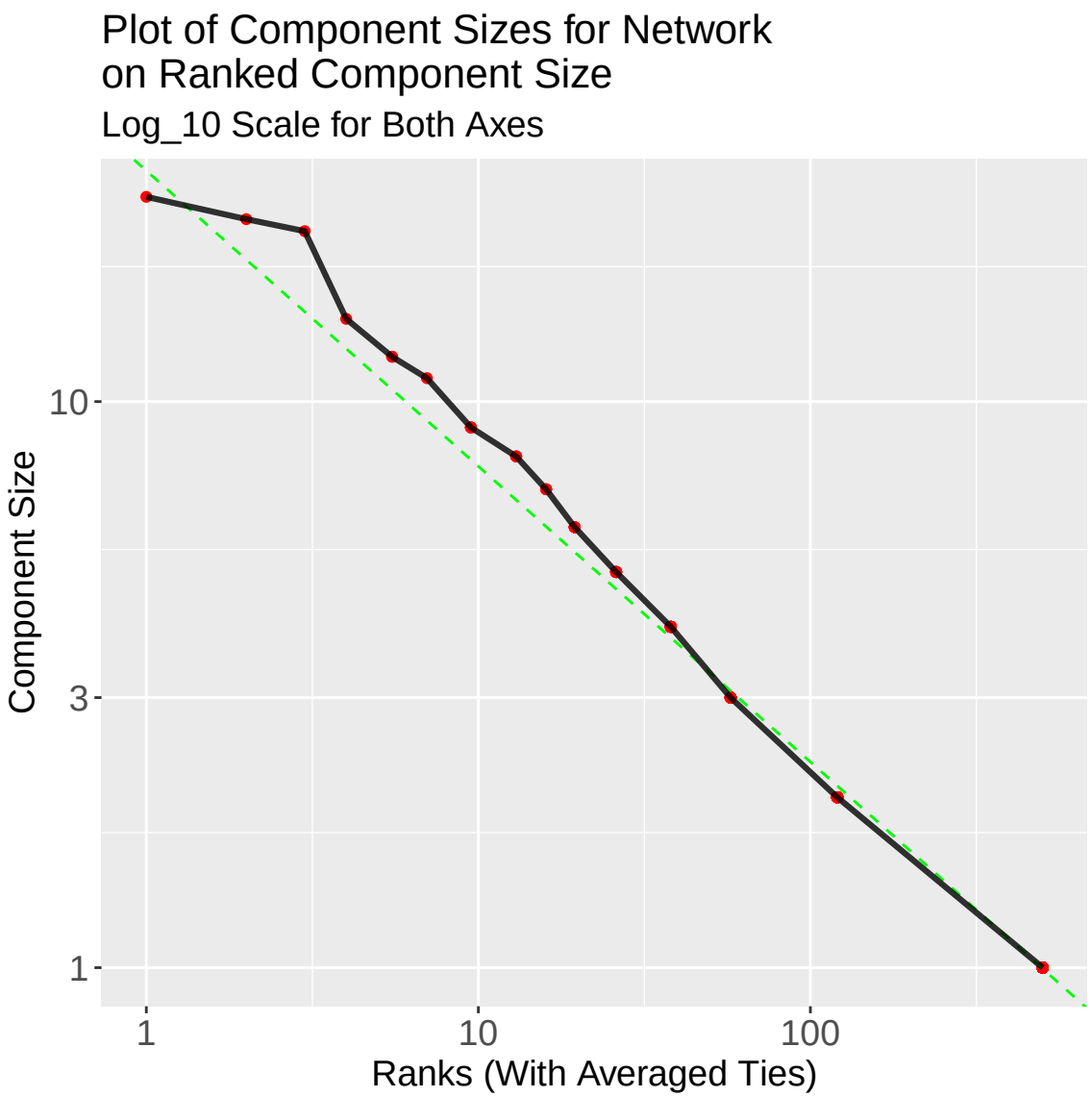
Following is a plot of the distribution of component sizes for the network (n = 1,258 nodes). I should add that sometimes in this sort of context and with more data (in text analytics, anyway), we would use relative component sizes (here, divided by 1,258). Also, it is not required to explicitly account for instances where the component size was never observed. For example, it is not required here to explicitly add in a 0 frequency if a component of size 22 was not observed.



Following is a plot of the distribution of component sizes for the network on the log base 10 scale for both the component sizes and the frequencies. We see that there is at least an approximate negative linear association.



Following is a so-called "*rank-frequency plot*" of the log base 10 component sizes for the network versus the log base 10 rank of the component size (smallest rank is the **largest** component size). Tied ranks were averaged for simplicity of exploration. There is an obvious negative linear association. If we fit the model $\log_{10}(\text{size}_i) = \beta_0 - \frac{1}{\alpha}\log_{10}(\text{rank}_i - 0.5) + \varepsilon_i$, for $i = 1, \ldots, n$ components (Gabaix and Ibragimov 2011), then we obtain an estimated α of 1.91 (in a moment, we will talk about why we have this cryptic parameterization). I've added a fitted regression line for a reference.



The power law probability mass function for (discrete) component size S is:

$$P[S = s] = \frac{s^{-\alpha}}{\zeta(\alpha, x_{\min})}$$

where:

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$$

is the Hurwitz zeta function evaluated numerically in R. These are given in Gillespie (2015).

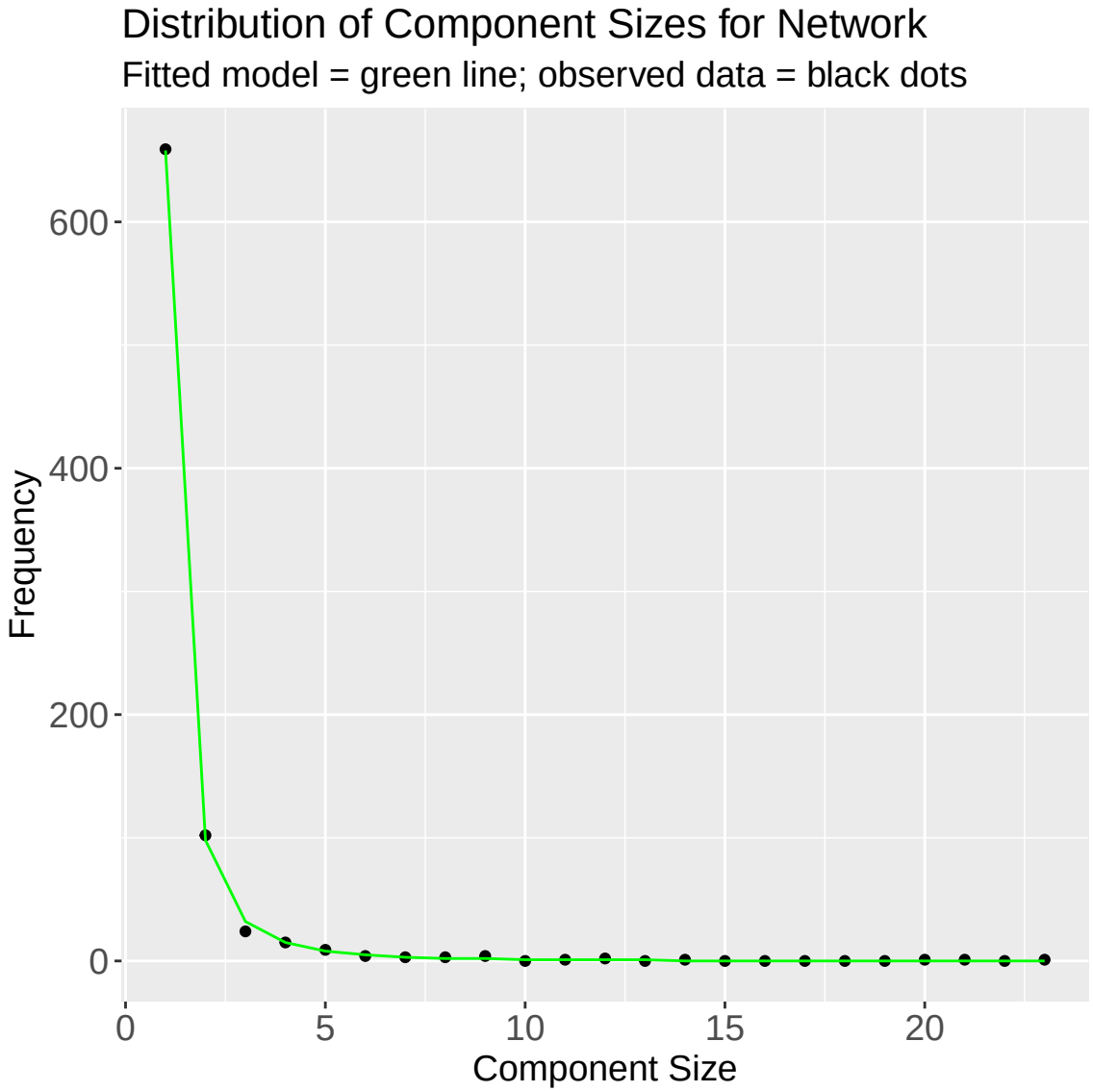
For our network, using the `poweRlaw` package in R, $x_{\min}$ is estimated to be 1, and α is estimated to be ~2.75. This is a bit distant from the value of 1.91 using OLS above, but this later technique should be thought of as a blunt force method.

We now use bootstrapping to get a confidence interval for α. Based on initial tuning, I used 3,000 bootstrap samples. The bootstrap distribution looked a touch right-skewed. A 95% bootstrap percentile confidence interval is [2.62, 2.91].

We now use the bootstrap test for power law goodness-of-fit described in Gillespie (2015) (his so-called

Algorithm 2). After initial tuning, I ran 3,000 bootstrap samples and received a p-value of 0.7238. Hence, there is insufficient evidence to suggest the data are *not* from a power law distribution.

The plot below shows the frequency distribution of the component sizes (black dots), with implied zeros added where necessary, and our fitted power law distribution overlaid (green line).



Odds-and-Ends

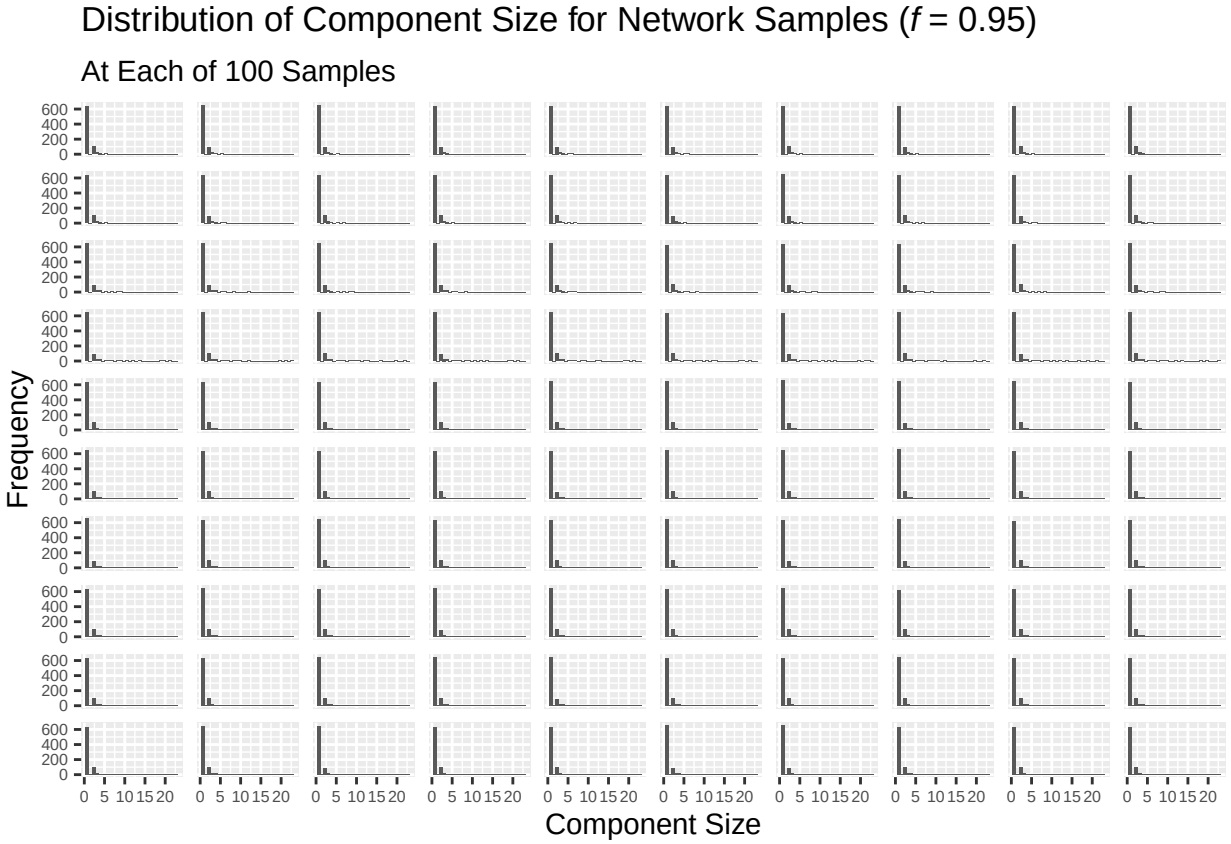
- Given how nicely the power law distribution fit the data, and the nature of the problem, I was disinclined to explore other heavy-tailed distributions for now (e.g., exponential).
- I think the use of the power law distribution for this problem is reasonable. For example, it is well-known that city size follows a power law distribution. A moment's reflection shows our situation to be analogous. That is, the nodes are the "citizens," and the aggregation into components (albeit through genetic linkage) represent "cities." Previous research shows the population of US cities follows a power law distribution with $\widehat{\alpha} = 2.30$, an estimate that is somewhat close to ours.
- Per the third point in New Modeling Attempt, I do not feel there is anything wrong with pooling the 100 simulations. In fact, in text analytics, this is routinely done. In our case, each sampling proportion setting represents the "corpus," and a simulation represents a "document." One could look at the simulation-specific empirical log-log distributions with a simulation envelope to get a sense of

network-to-network variation. You could do worse than to use the median frequency (or rounded mean frequency) over simulations at each observed component size to get a sense of marginalized behavior for networks sampled at the given sampling fraction.

- Ravi brought up the point about how the sampling is done. In my mind, I view the method as follows. Take a one-shot SRSWOR of xx% nodes from the "population" network. Construct a "sample" network based on genetic distance and obtain the component sizes. (And I assume we hope the sample network is representative of the population network with respect to component sizes, taking into account reduced frequencies, variance being affected from large sampling fractions taken from a finite population, etc.) Ravi thinks of it differently, and uses a sequential approach. I will leave it to him to explain his thought process.

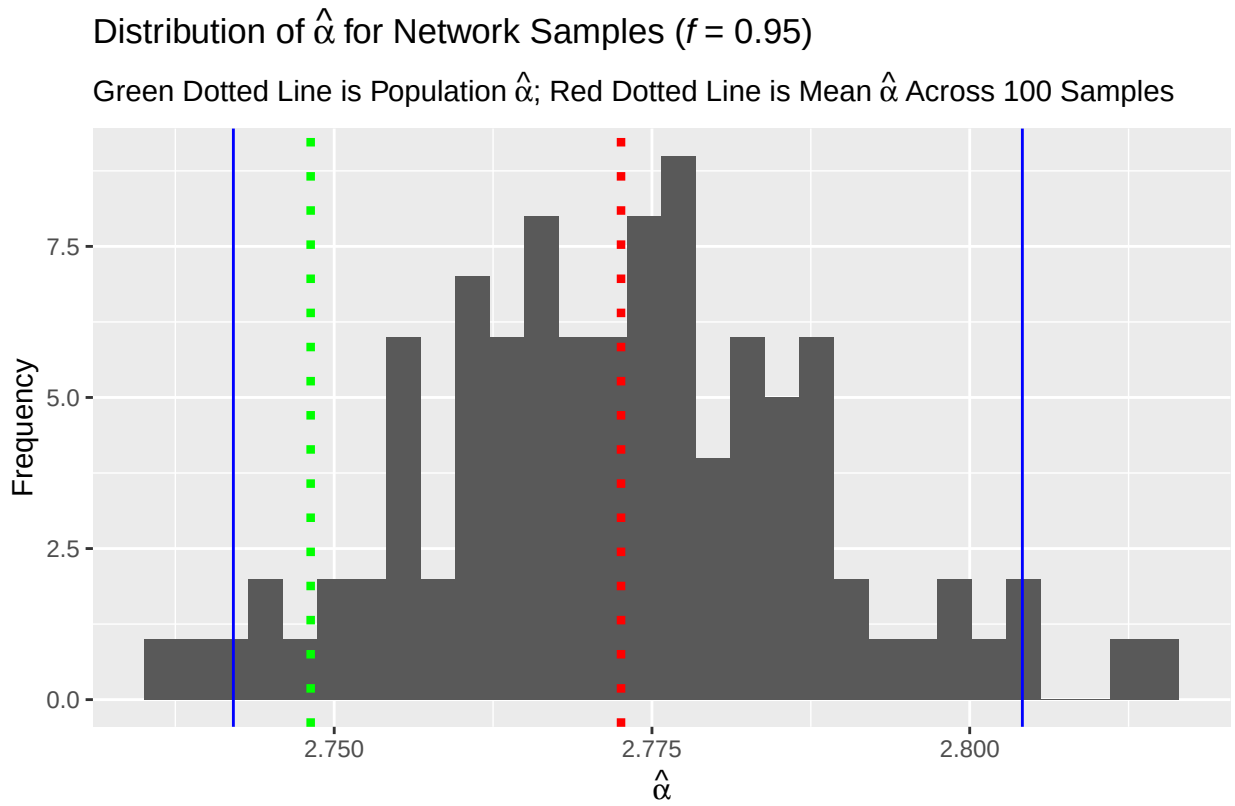
Part 2: Examining 100 Samples from Network at Sampling Fraction 0.95 (*updated*)

Let the sampling fraction f equal the sample size n divided by the number of nodes, or population size, N. I use Gabby's 100 samples where $f = 0.95$. This corresponds to $n = 1,195$ nodes. The plot below shows frequency histograms of the component sizes for all 100 samples. Obviously, it is hard to see much, but there are small, meaningful differences from sample-to-sample.

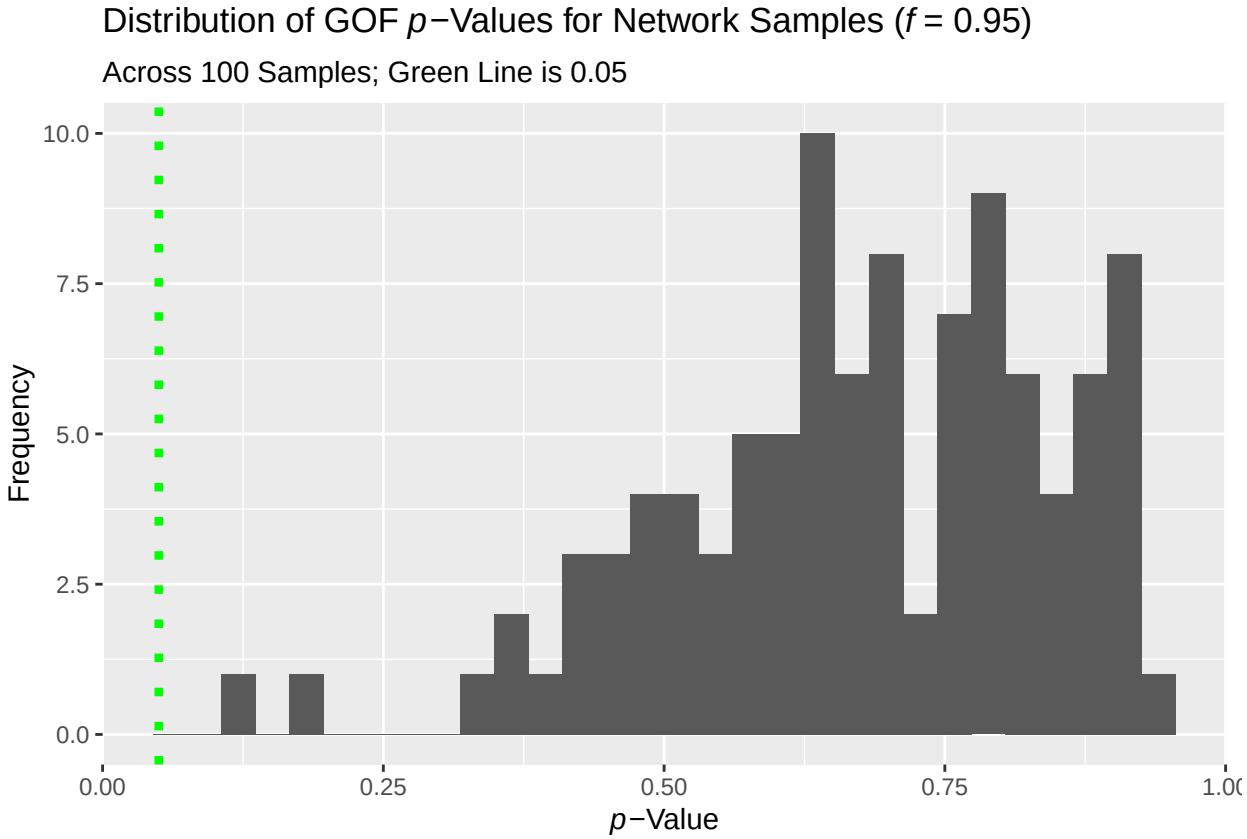


I then took each of the 100 samples and fit the power law model. In each case, I a.) obtained $\widehat{\alpha}$, and b.) computed the bootstrap test for power law goodness-of-fit exactly as described above. Total elapsed time, running 4 cores in parallel on my Mac, took a little over 2 hours; so, this is not a trivial undertaking.

The plot below shows the frequency distribution for the 100 sample $\widehat{\alpha}$'s. In all 100 cases, x_{min} was estimated to be 1. The red line represents the mean across 100 samples, while the green line represents the "population" $\widehat{\alpha} = 2.75$ from above. Blue solid lines correspond to the 0.025 and 0.975 quantiles. The difference between the red and green line will have little practical effect for the work we do here.



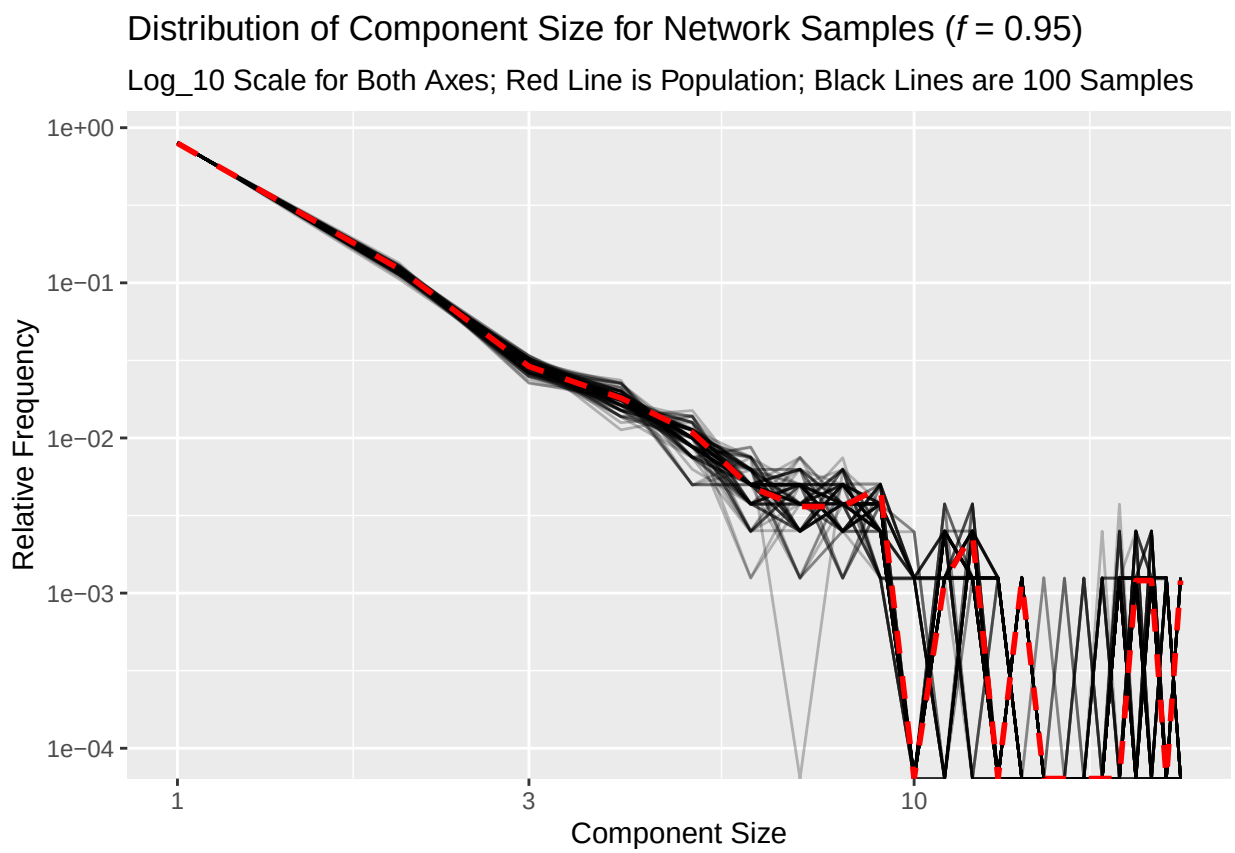
Next, we examine the frequency distribution for the p-values from the bootstrap test for power law goodness-of-fit conducted on each of the 100 samples. At least in this case, all the p-values are above 0.05 (the green line), but there is considerable variation among tests.



Care must be taken when working with samples drawn from networks because the frequencies of component sizes will necessarily decrease as f decreases. Hence, for the following two graphs, it is best to use *relative* frequencies.

The following is a plot showing the relative frequency distributions for component size ($\log_{10}$-$\log_{10}$ scale) for all 100 samples *and* the population. Because spaghetti plots with interpolation can be misleading, I use the data set explicitly containing zero frequencies for any component sizes that were not observed within any sample. When plotted on a log scale, this required me to trim the y-axis (Relative Frequency) to an arbitrarily small number (< 0.0001) as a quick workaround.

For the most part, the samples are emblematic of the population (red line). However, the samples start to show small differences out in the right tail. This is similar to a well-known problem in extreme value theory. Long, right tails are just hard to nail down.



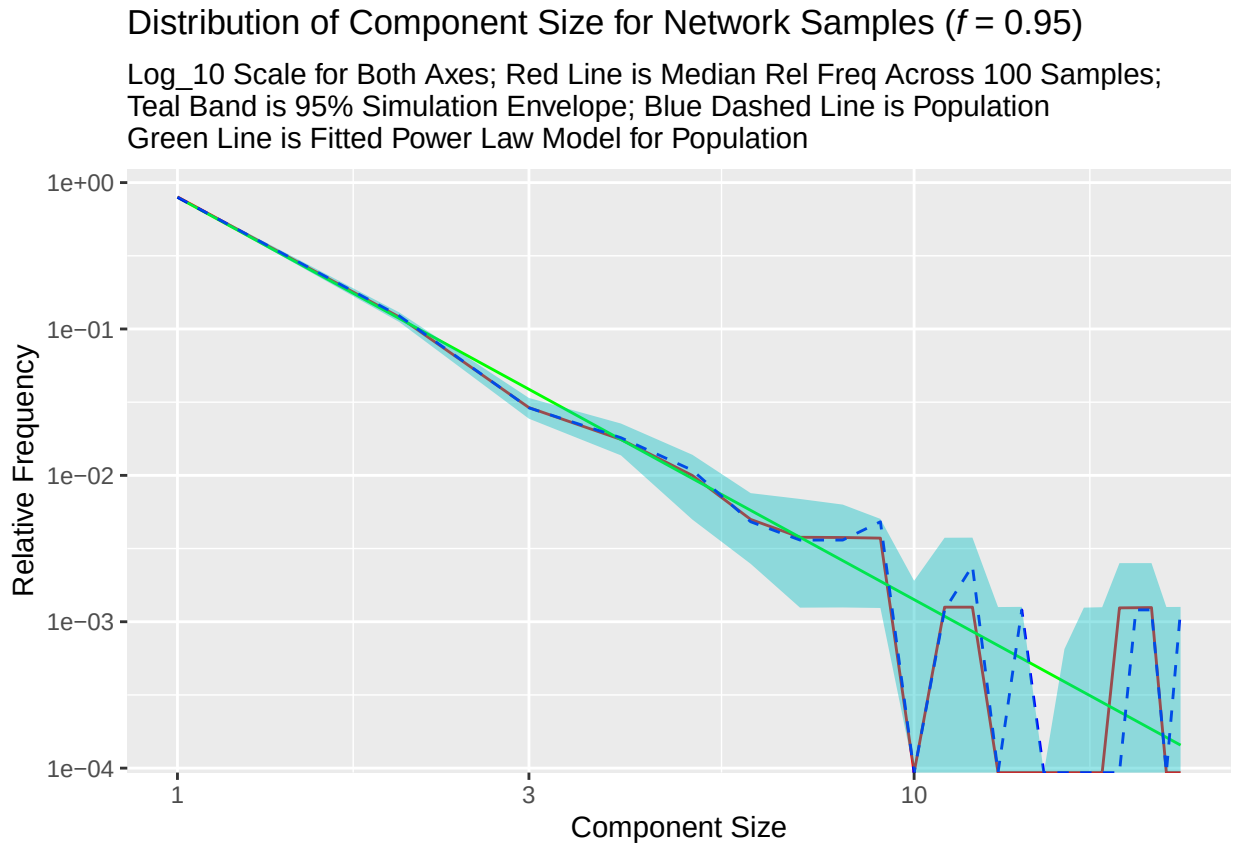
The table below shows the count of the 100 samples where at least one of the given component sizes (`size`) was observed. It also shows if at least one component of the given size was ever observed in the population (`In pop?`). For example, there were 2 samples where at least one component of size 15 was observed, yet no components of size 15 were observed in the population.

Table 1: Count of 100 Samples With One or More Components of Given Size

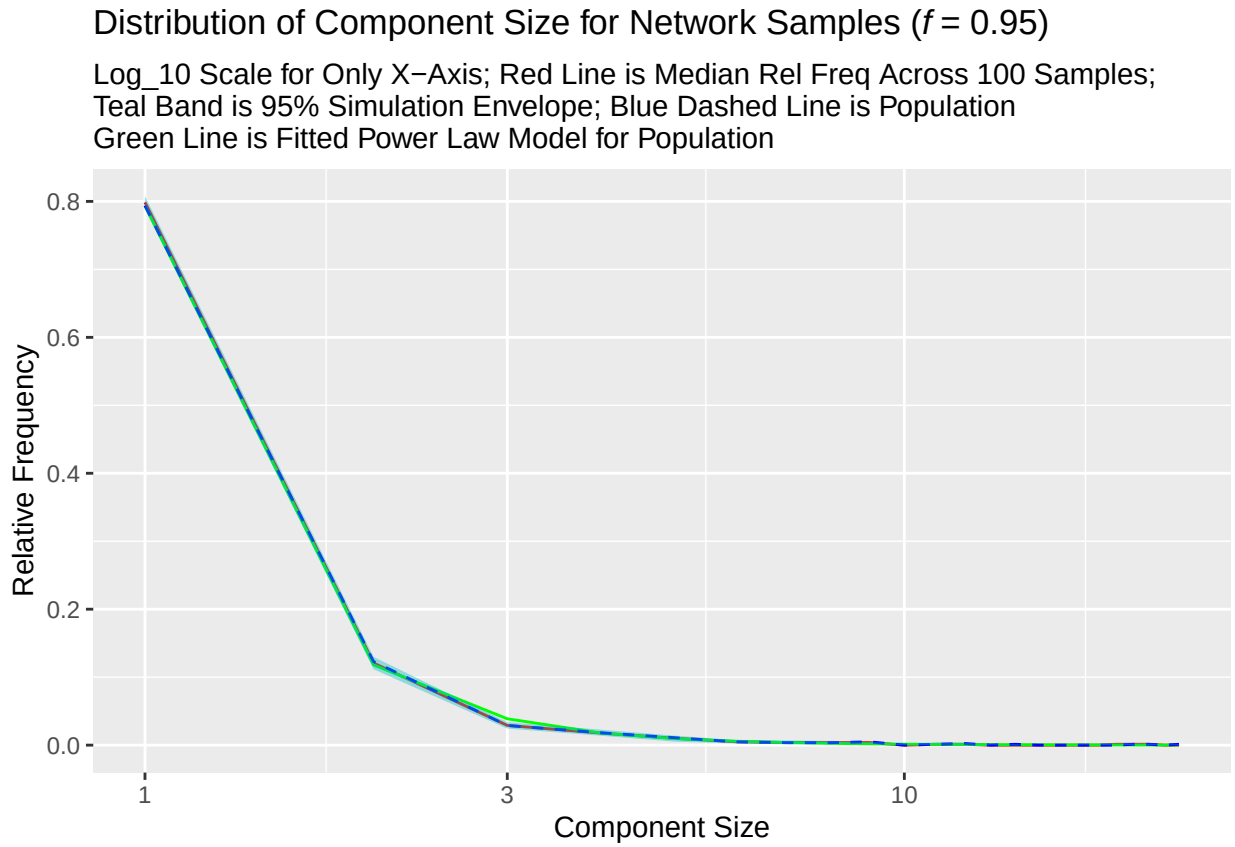
size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
n	100	100	100	100	100	100	99	100	100	34	80	88	35	41	2	3	6	13	56	60	63	32	30
In pop?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	No	Yes	Yes	No	Yes

The following is a "do-over" of the previous plot. At each component size from 1-to-23, I took the 0.025 and 0.975 quantiles of the 100 relative frequencies to form a "simulation envelope" (teal band). The patchy constriction of the teal band can be understood using the table above. For example, only 2 samples were observed containing at least one component of size 15. Specifically, Samples #80 and #95 each returned one component of size 15 (with relative frequencies of 1/798 and 1/793, respectively) with the other 98 samples returning implied relative frequencies equal to zero. More samples would be needed to form a sensible 95% simulation envelope.

Continuing, the blue, dashed line corresponds to the population. The red line corresponds to the median relative frequency; notice that it tracks nicely with the blue, dashed population line until we get further out in the right tail. The green line represents our fitted power law model from Part One. For the most part, both lines fluctuate around what is predicted by the power law. Note the log scale makes any apparent discrepancies look much worse than they are; indeed, the bootstrap test for power law goodness-of-fit for the population (in Part One) gave a p-value of 0.7238.



What would the plot look like if we were to plot it in "real-time"; that is, remove the log scale for relative frequency? Looking below gives us a sense of a very different story!



The approach given here would be more useful if we had more samples (1,000, say). We would then extend it to look at a sequence of sampling fractions f.

In closing, and as Ravi brought up last week, this report is not meant to suggest that a power law model will uniformly be "best" across all types of network compositions.

Part 3: Comparing Distributions for Population Network

In order to take a deeper dive into an appropriate model for component size for the population network, we compare the following candidates: power law (as defined above), discrete log-normal, Yule-Simon, discrete exponential, and Poisson. One has to be a bit careful here. The reason is because these distributions have been modified, or "normalized," so that they adhere to the general form of a power-law distribution (see Clauset, Shalizi, and Newman (2009), for example).

To compare fitted models, the thought occurred to me to use Akaike's Information Criterion (AIC), as opposed to all possible pairs of likelihood ratio tests (Vuong's test). We use Akaike's Information Criterion defined as -2*log-likelihood(mle) + 2*npar where npar represents the number of parameters in the fitted model. (There is a technical quibble about how to treat estimation of x_{min}. Specifically, should we acknowledge this in the penalty term of AIC? For the sake of parsimony, I do not do so here.)

I fit the aforementioned five models to the observed component sizes from the population network, rightfully setting $x_{min} = 1$ for each, and computed the AIC value. These are displayed in the table below:

As a general rule, the smaller the AIC value, the better the fit *and* models that are within 2 AIC units of each other are considered to provide equivalent fit (although 2 is now thought to be too low these days and some lobby for 4 or 5 as a cutoff). In our case, both the power law and discrete log-normal models clearly provide the best relative fit to the data.

Model	AIC
Power Law	1373.49
Disc logN	1375.32
Yule-Simon	1389.79
Disc Exp	1615.22
Poisson	2017.57

Using Vuong's test, we use the power law and discrete log-normal models to test:

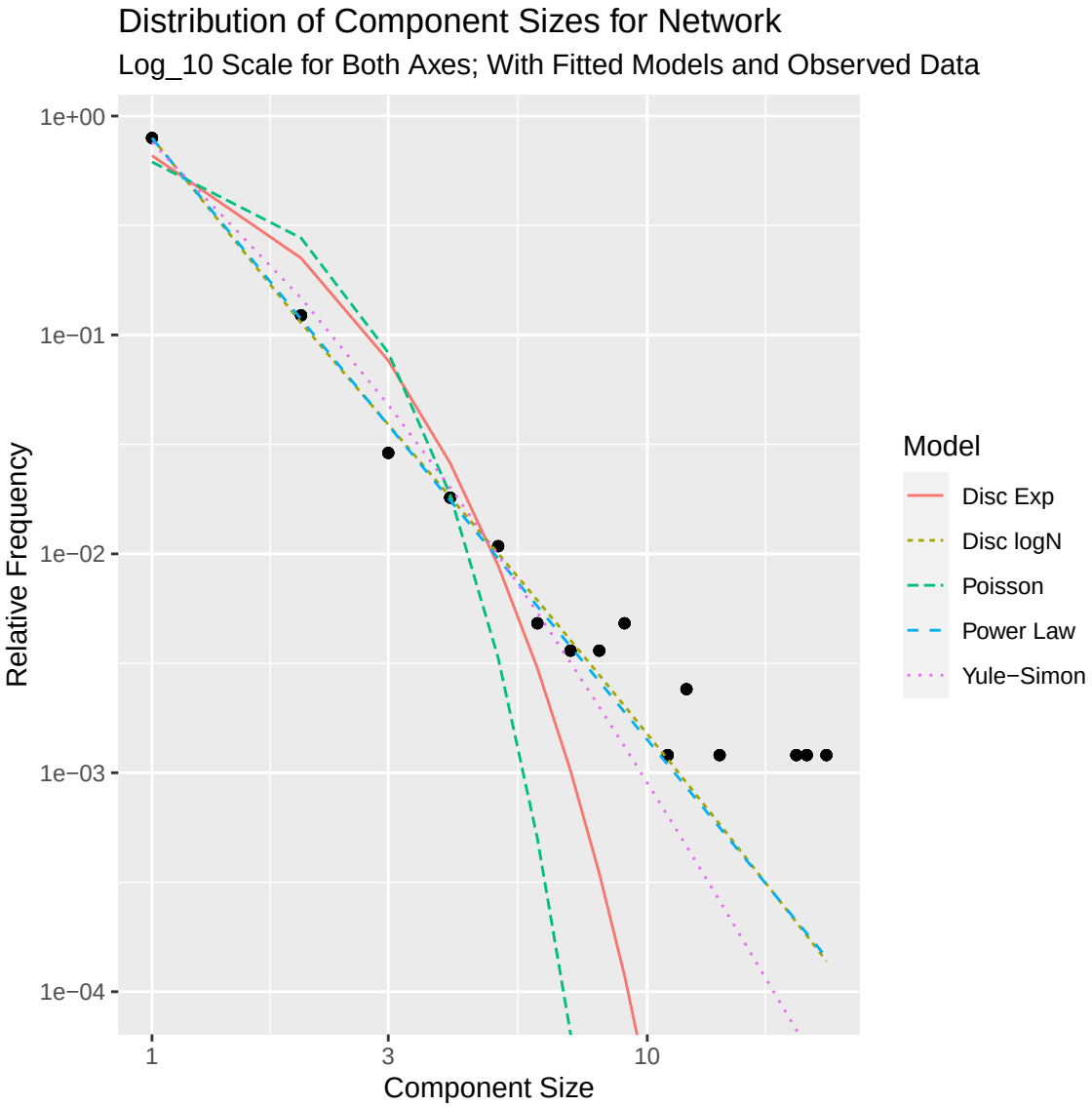
$$H_0 : \text{both models are equally far from the "true" model}$$

versus:

$$H_a : \text{one of the models is closer to the "true" model}$$

Unfortunately, and as Gabby and I discussed, this test cannot make any decision about whether the closer distribution is the "true" distribution. Said another way, if this test is significant, it does not mean that one model is a "good" fit, only that it is better than the other model. In any case, the p-value of 0.87 suggests that both models perform equally well, a hardly surprising result given the AIC values.

All of what we have said here is affirmed by the plot of the five model fits below. It appears sensible to continue to work with the (simple?) power-law distribution for this particular population network. At its worst, out in the right tail, the difference between the observed versus fitted relative frequency is roughly 0.001. To me, this is practically meaningless for our purposes.

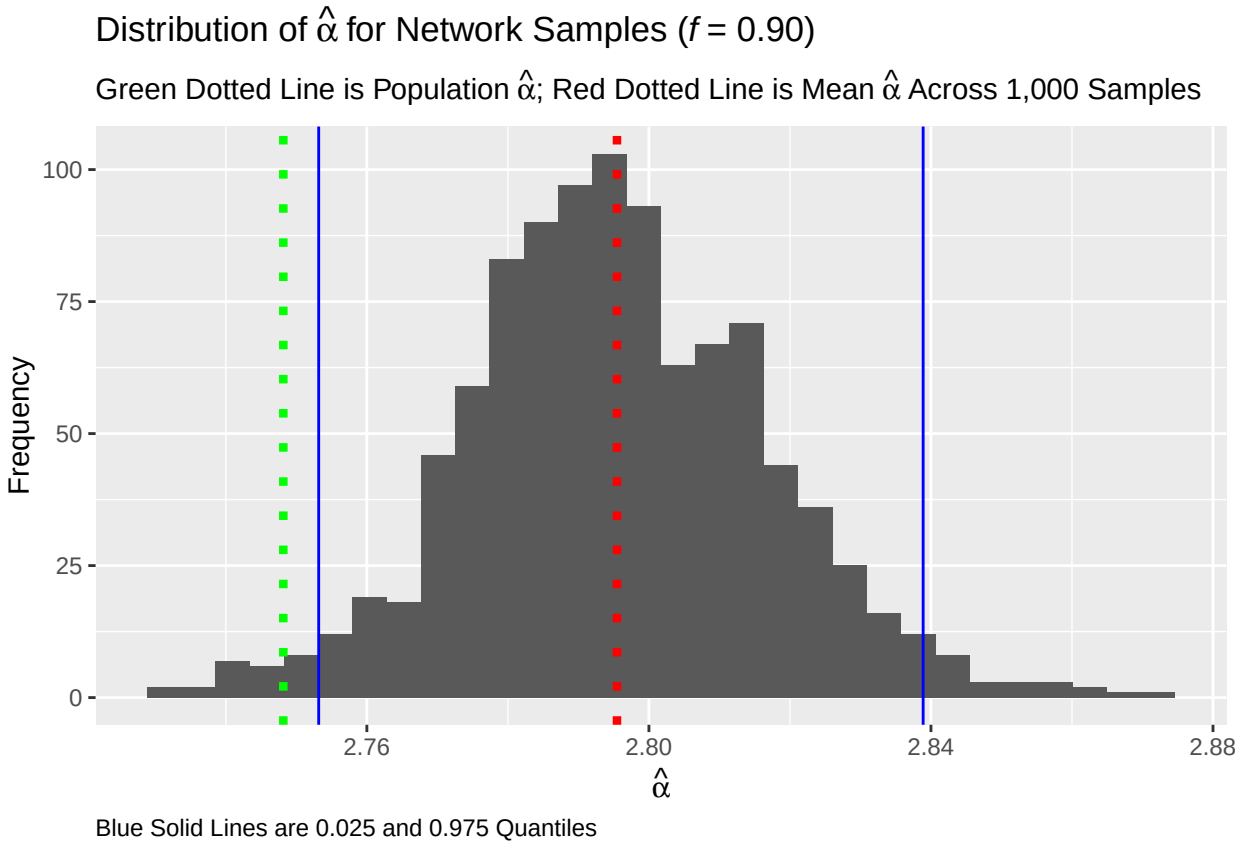


Part 4: Examining 1,000 Samples from Network at Sampling Fractions 0.90, 0.50, and 0.20

We take 1,000 samples from the population network with sampling fractions $f = 0.90$ ($n = 1,132$), 0.50 ($n = 629$), and 0.20 ($n = 252$) and repeat a few approaches described previously. In addition, we add a few new approaches. It should be noted that all of this can best be accomplished through high-performance computing; doing this on a desktop will take many hours.

If you have two nodes that were unconnected before with respect to genetic similarity, then that relationship would not change just because those two nodes were selected into a sample. Furthermore, it becomes increasingly unlikely to sample entire larger components from the population as the sampling fraction decreases. Therefore, we would anticipate in advance that we will increasingly trim the right tail of the distribution of component sizes.

For the $f = 0.90$ case, x_{min} was estimated to be 1 for all samples. We see that on average, a sample returns a slightly higher estimate of the shape parameter α (2.80) than the one estimated for the population (2.75). But what is the practical effect of this? The answer is "hardly anything." The greatest disparity between the population and a typical sample with respect to the probability mass function is when the component size is 1. Specifically, the probability the component size is 1 for the population is 0.7932, while it is 0.8012 for the typical sample.



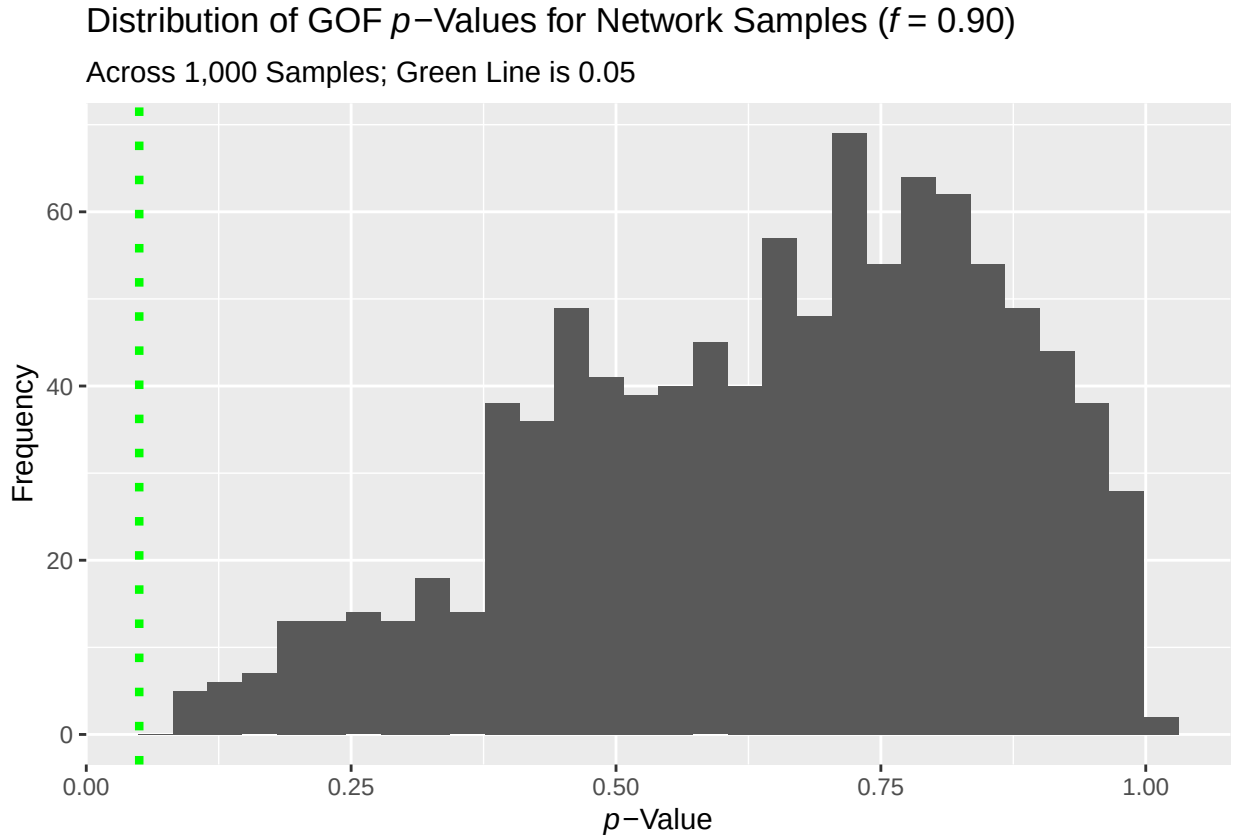
The mean of the power law distribution is:

$$\frac{\alpha - 1}{\alpha - 2}$$

Hence, by the invariance property of MLEs, we can estimate the mean for each sample and then take the grand mean to get a sense of what the average-sized component is. Our answer here is 1.56. In reality, if we directly compute the grand mean for our 1,000 samples, we get 1.48, and the actual mean for the population network is 1.52.

By taking the variance of the $\hat{\alpha}$ for our 1,000 samples (0.00044) and adding the square of the approximate "bias" $\left(\bar{\hat{\alpha}} - \hat{\alpha}_{pop}\right)$ (0.00224), we can obtain an MSE-type of statistic equal to 0.0027.

Next, we examine the frequency distribution for the p-values from the bootstrap test for power law goodness-of-fit conducted on each of the 1,000 samples. All the p-values are above 0.05 (the green line), but there is considerable variation among tests.



The table below shows the count of the 1,000 samples where at least one of the given component sizes (size) was observed. Once again, it also shows if at least one component of the given size was ever observed in the population (In pop?). It is interesting to see some of the differences at the sample level versus the population. For example, more often than not, we will observe a component of size 10 or 19 in our sample, but do not observe components of these sizes in the population. In only 89 out of 1,000 samples do we observe the largest component realized in the population (size of 23).

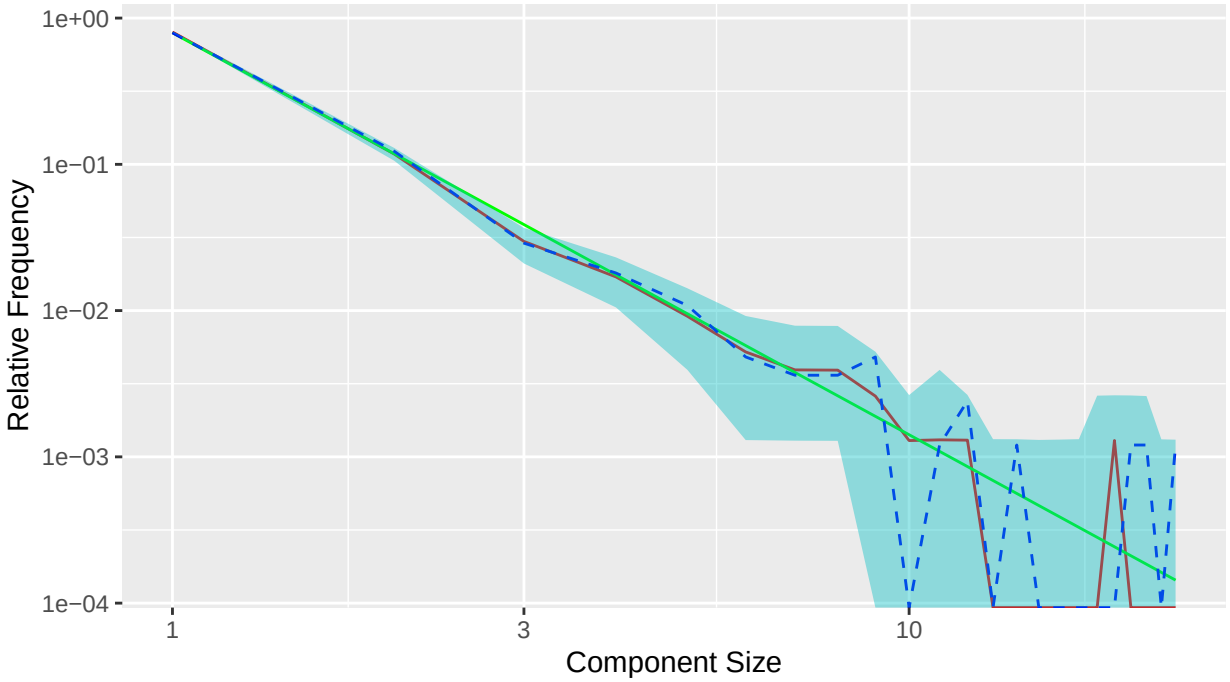
Table 2: Count of 1,000 Samples With One or More Components of Given Size

size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
n	1000	1000	1000	1000	1000	992	980	979	895	549	770	631	332	257	73	127	247	448	544	459	332	230	89
In pop?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	No	Yes	Yes	No	Yes

The plot below shows the distribution of component sizes with our simulation envelope as detailed in Part 2. We see the estimated power law, the observed relative frequency from the population, and the median relative frequency across the 1,000 samples are congruent. The sum of the squared deviations between the later two is only 0.00012. No surprise here; our sample fraction of 0.90 is quite large. Any differences we see between the population and the samples are minor and exaggerated by the $\log_{10}$ scale. For example, looking at the "difference" for the component size of 10, the median relative frequency is only 0.00129. Perhaps we should remove the the $\log_{10}$ scale for the y-axis to show this in another light.

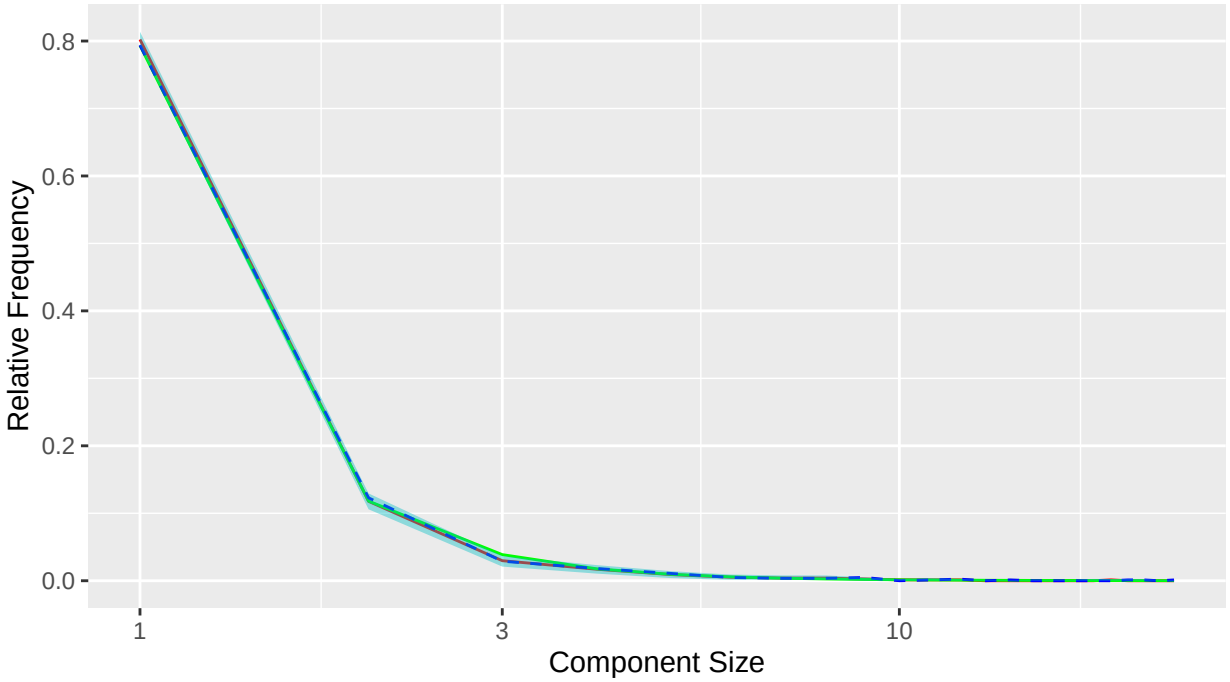
Distribution of Component Size for Network Samples (f = 0.90)

Log_10 Scale for Both Axes; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population

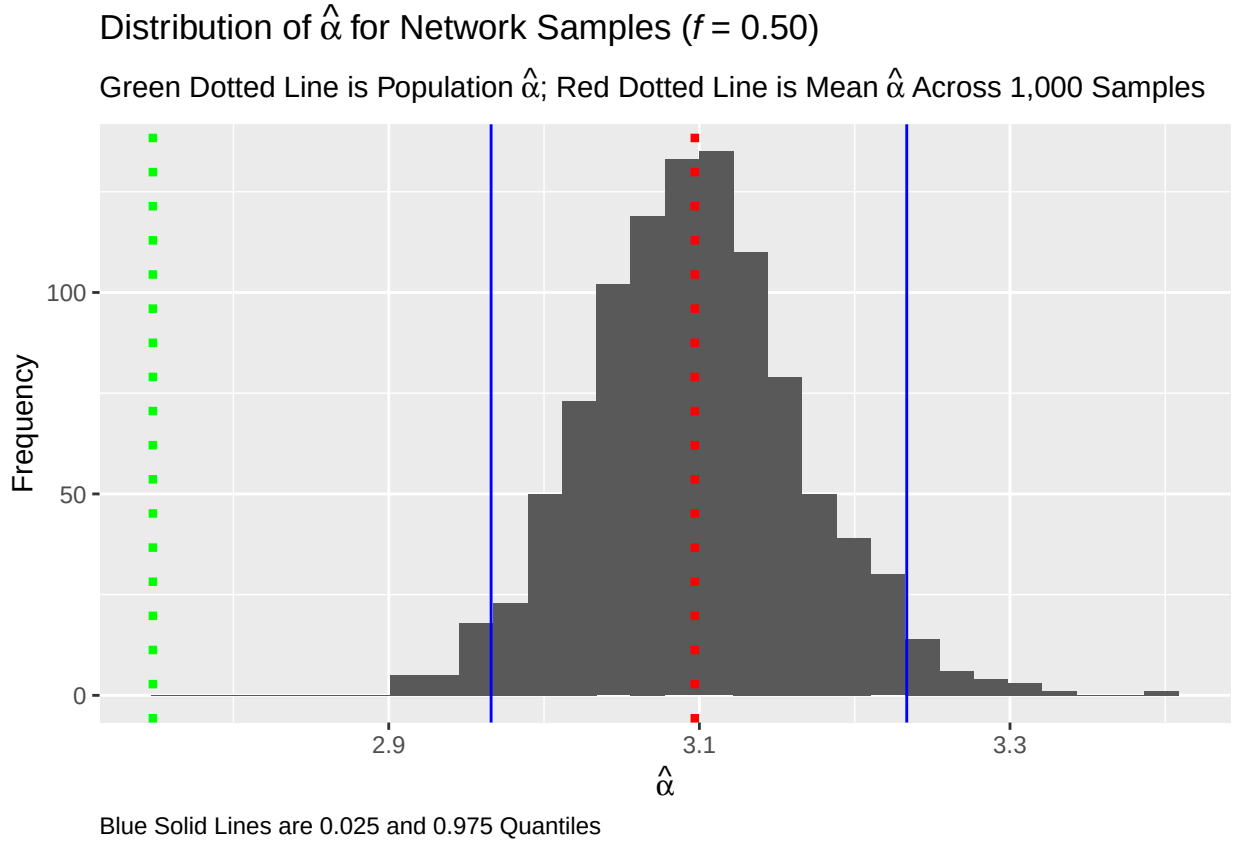


Distribution of Component Size for Network Samples (f = 0.90)

Log_10 Scale for Only X−Axis; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population



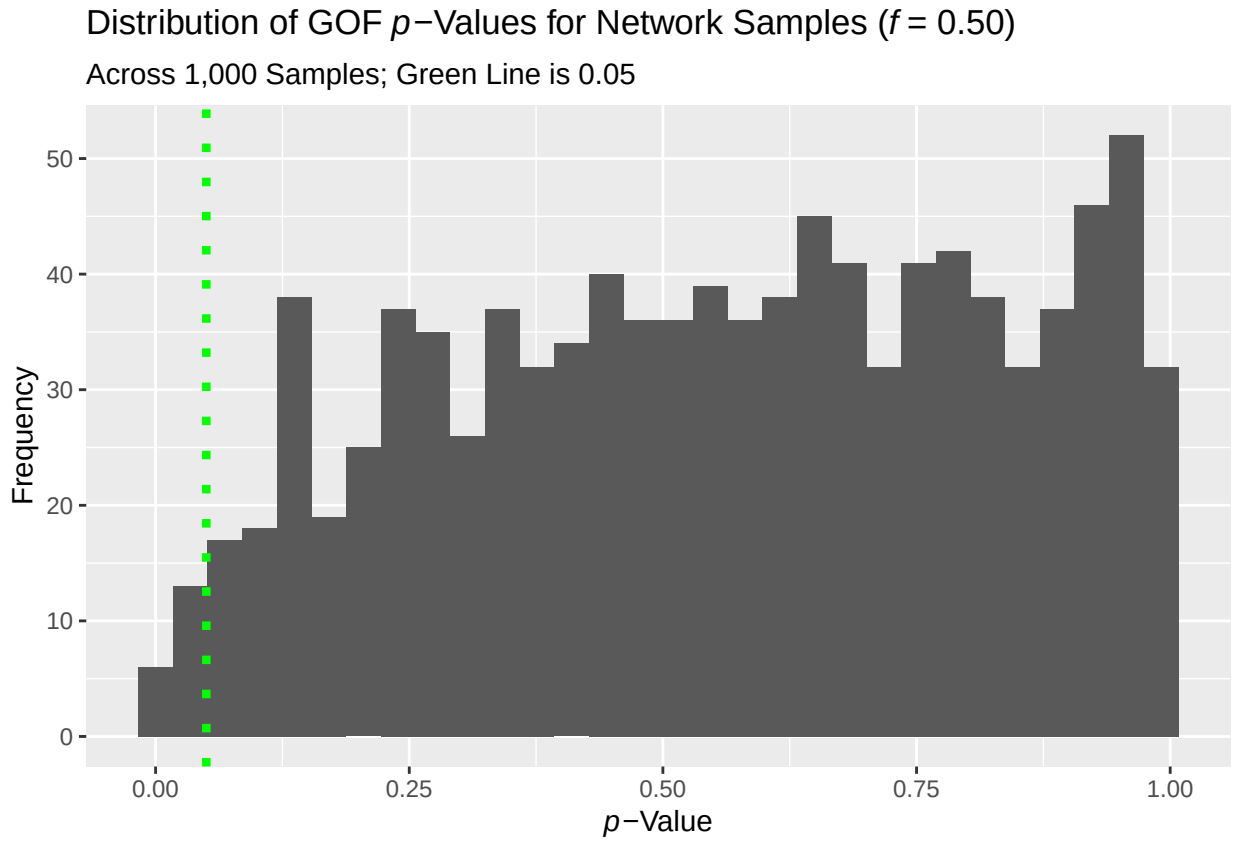
For the $f = 0.50$ case, x_{min} was estimated to be 1 for all samples. We see that on average, a sample is starting to return a more meaningfully higher (biased) estimate of the shape parameter α (3.10) than the one estimated for the population (2.75). The greatest disparity between the population and a typical sample with respect to the probability mass function is when the component size is 1. Specifically, the probability the component size is 1 for the population is 0.7932, while it is 0.8447 for the typical sample. Certainly, this is more pronounced difference than the $f = 0.90$ case.



If we estimate the mean component size for each sample using $\hat{\alpha}$ and then take the grand mean, our answer here is 1.48. In reality, if we directly compute the grand mean for our 1,000 samples, we get 1.31, and the actual mean for the population network is 1.52.

By taking the variance of the $\hat{\alpha}$ for our 1,000 samples (0.00472) and adding the square of the approximate "bias" $\left(\overline{\hat{\alpha}} - \hat{\alpha}_{pop}\right)$ (0.12176), we can obtain an MSE-type of statistic equal to 0.1265.

Next, we examine the frequency distribution for the p-values from the bootstrap test for power law goodness-of-fit conducted on each of the 1,000 samples. We now observe 19 p-values at or below 0.05 (the green line).



The table below shows the count of the 1,000 samples where at least one of the given component sizes (size) was observed. Once again, it also shows if at least one component of the given size was ever observed in the population (In pop?). Compared to the previous $f = 0.90$ case, we are clearly starting to chip away at the right tail.

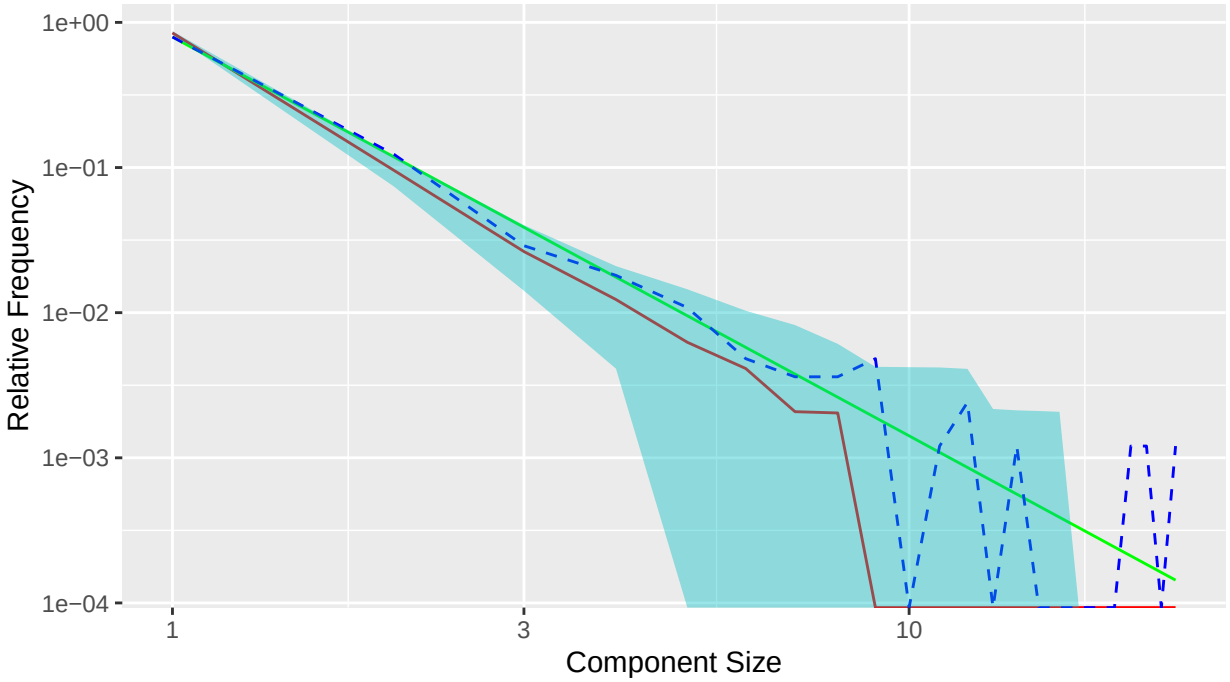
Table 3: Count of 1,000 Samples With One or More Components of Given Size

size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
n	1000	1000	1000	999	969	852	717	542	459	403	336	319	231	158	85	39	17	7	2	0	0	0	0
In pop?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	No	Yes	Yes	No	Yes

The plot below shows the distribution of component sizes with our simulation envelope as detailed in Part 2. We now see some minor differences appearing between the observed relative frequency from the population and the median relative frequency across the 1,000 samples. The sum of the squared deviations between the later two is now 0.0036. We remove the the $\log_{10}$ scale for the y-axis to show this in another light.

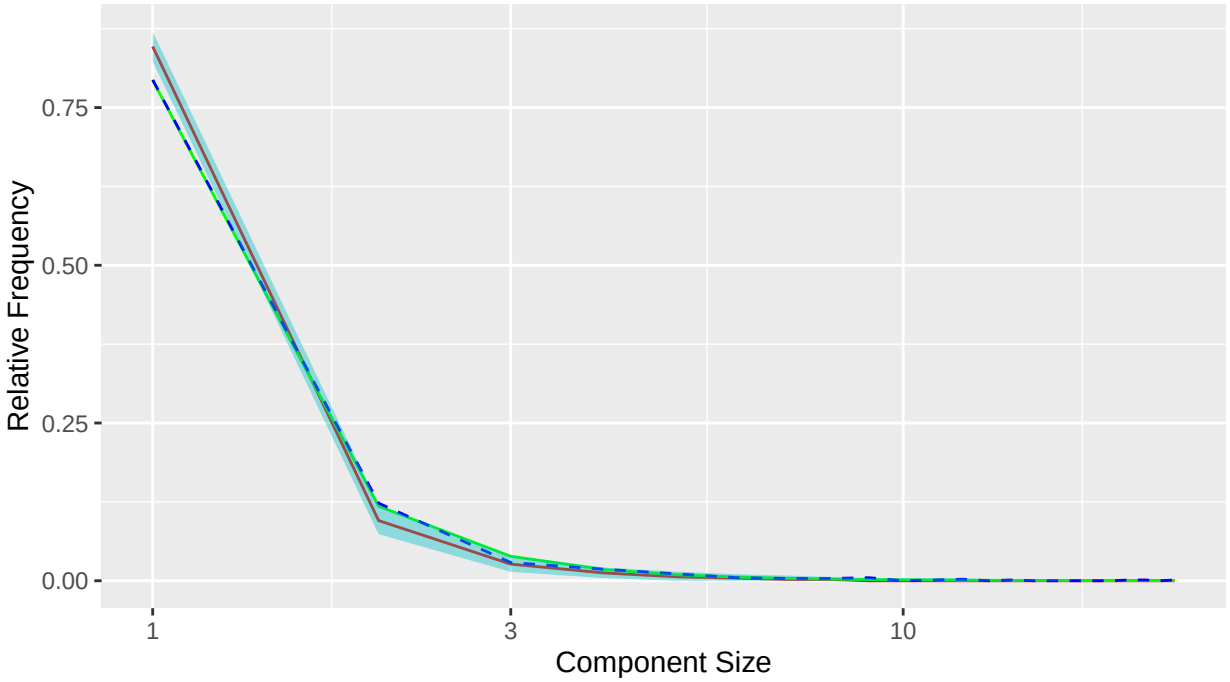
Distribution of Component Size for Network Samples ($f = 0.50$)

Log_10 Scale for Both Axes; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population



Distribution of Component Size for Network Samples ($f = 0.50$)

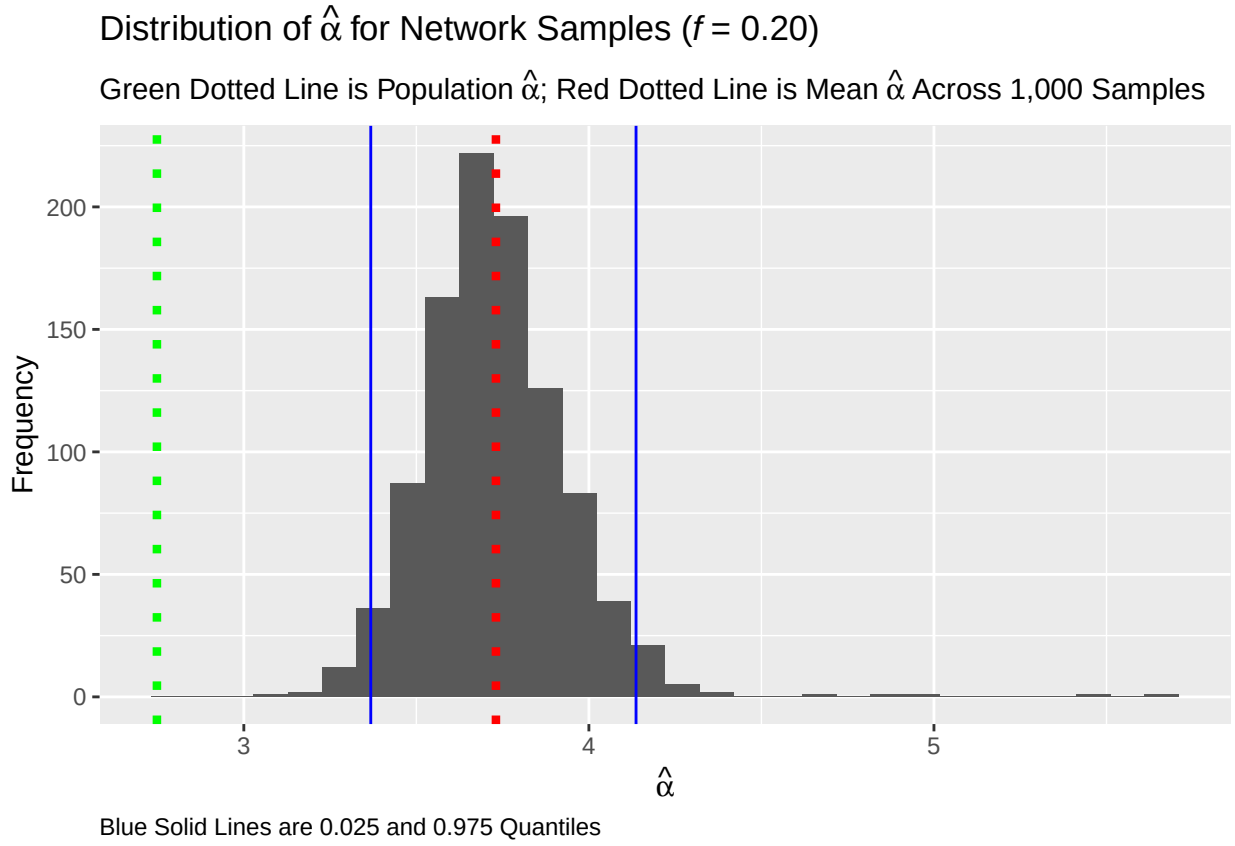
Log_10 Scale for Both Axes; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population



For the $f = 0.20$ case, x_{min} was estimated to be 1 for 996 samples, and 2 for 4 samples. In hindsight, I should have fixed x_{min} to be 1 for all the samples, but we ignore this minor annoyance for now.

We see that on average, a sample now appears to return a substantially (biased) estimate of the shape parameter α (3.73) compared to what we observe estimating α for the population (2.75). The greatest disparity between the population and a typical sample with respect to the probability mass function is when the component size is 1. Specifically, the probability the component size is 1 for the population is 0.7932, while it has now climbed to 0.9062 for the typical sample. To me, this is a difference that could not be ignored.

It is worth noting the appearance of relatively large values of $\hat{\alpha}$.



If we estimate the mean component size for each sample using $\hat{\alpha}$ and then take the grand mean, our answer here is 1.37. In reality, if we directly compute the grand mean for our 1,000 samples, we get 1.14, and the actual mean for the population network is 1.52.

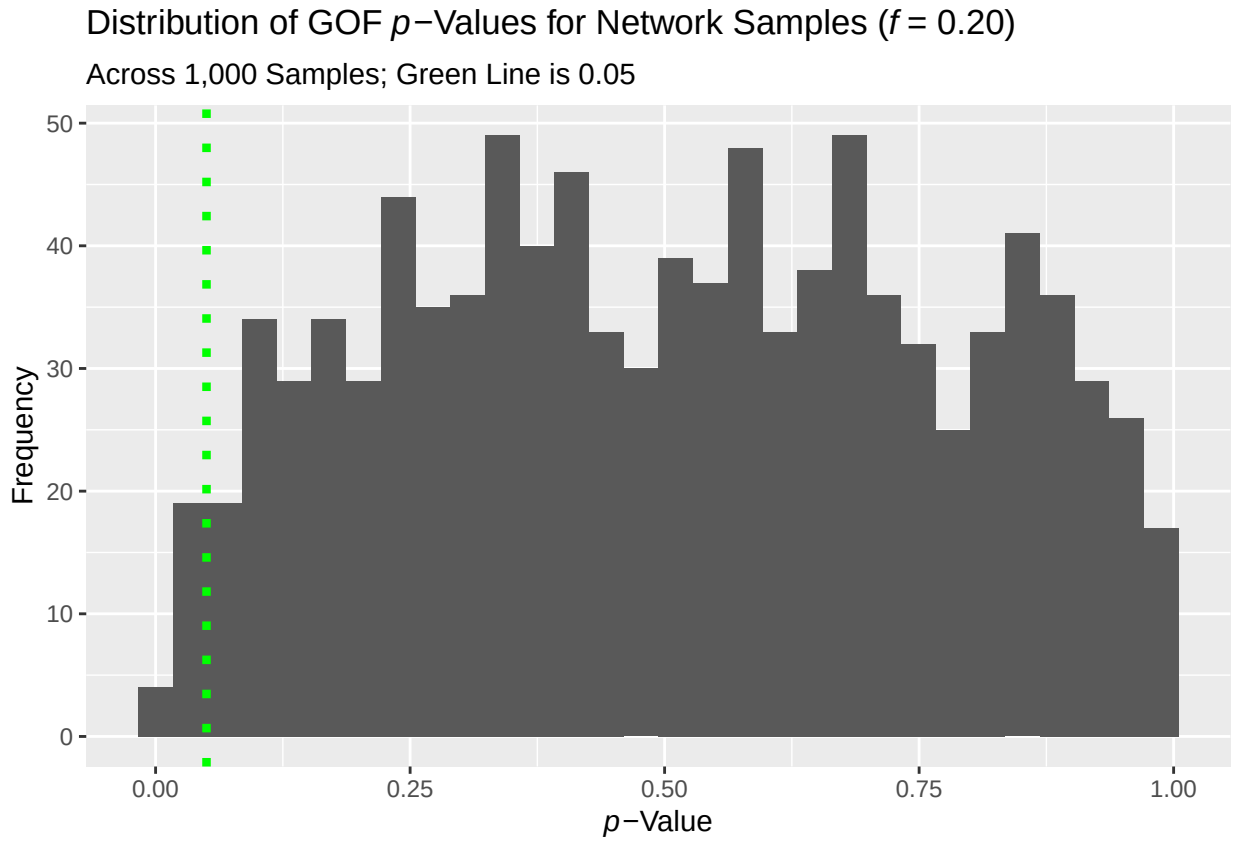
Conclusion: The sampling fraction had little appreciable effect on estimating mean component size.

By taking the variance of the $\hat{\alpha}$ for our 1,000 samples (0.04683) and adding the square of the approximate "bias" $\left(\bar{\hat{\alpha}} - \hat{\alpha}_{pop}\right)$ (0.96513), we can obtain an MSE-type of statistic equal to 1.0120.

Conclusion: What is driving the increase in our 'MSE?' Moving from sampling fractions $f = 0.90$ to 0.50 to 0.20, the sample variance of $\hat{\alpha}$ went from 0.00044 to 0.00472 to 0.04683. However, more dramatically, the squared bias went from 0.00224 to 0.12176 to 0.96513.

Next, we examine the frequency distribution for the p-values from the bootstrap test for power law goodness-of-fit conducted on each of the 1,000 samples. We now observe 23 p-values at or below 0.05 (the green line).

Conclusion: I thought the GOF tests would suffer much more than they did. The 23 small p-values do NOT associate with large estimates of α.



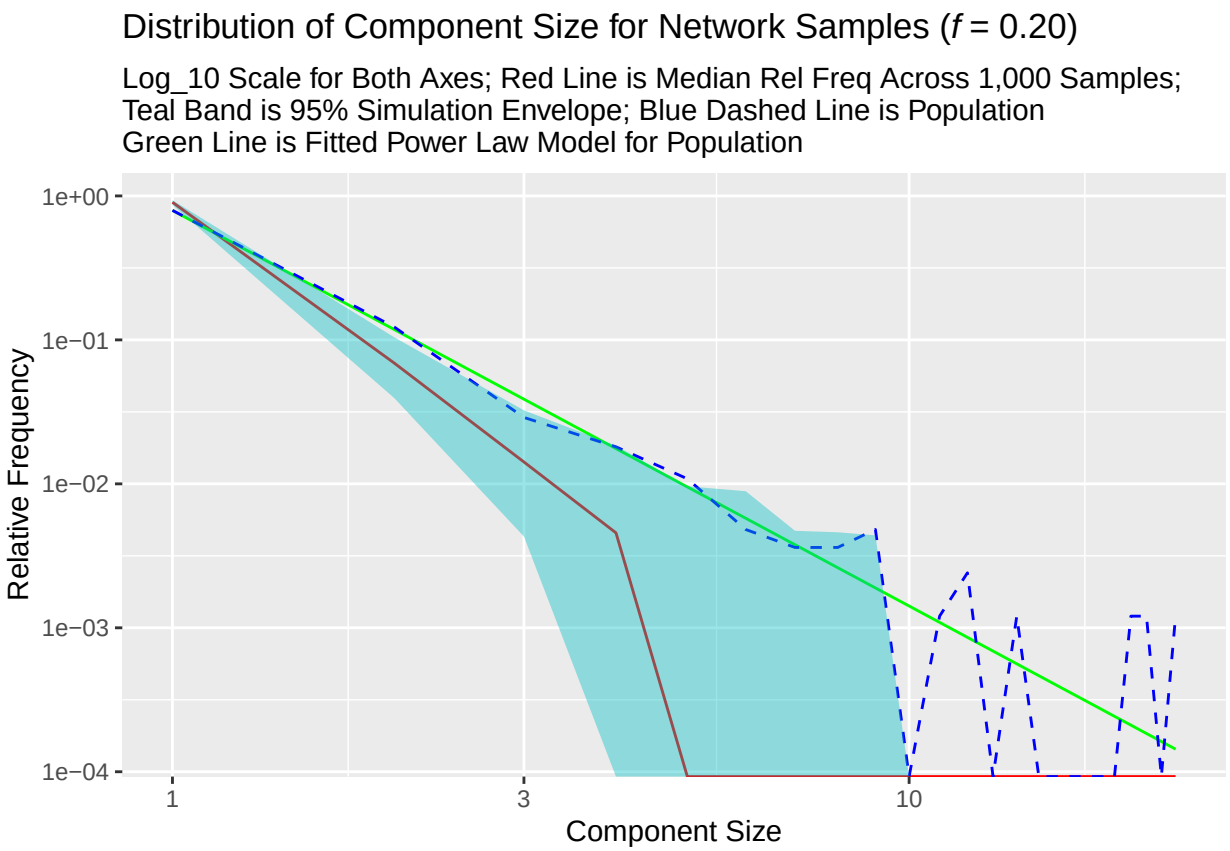
The table below shows the count of the 1,000 samples where at least one of the given component sizes (size) was observed. Once again, it also shows if at least one component of the given size was ever observed in the population (In pop?). The right tail is now very much undermined. Only one sample had the largest observed component size of 11.

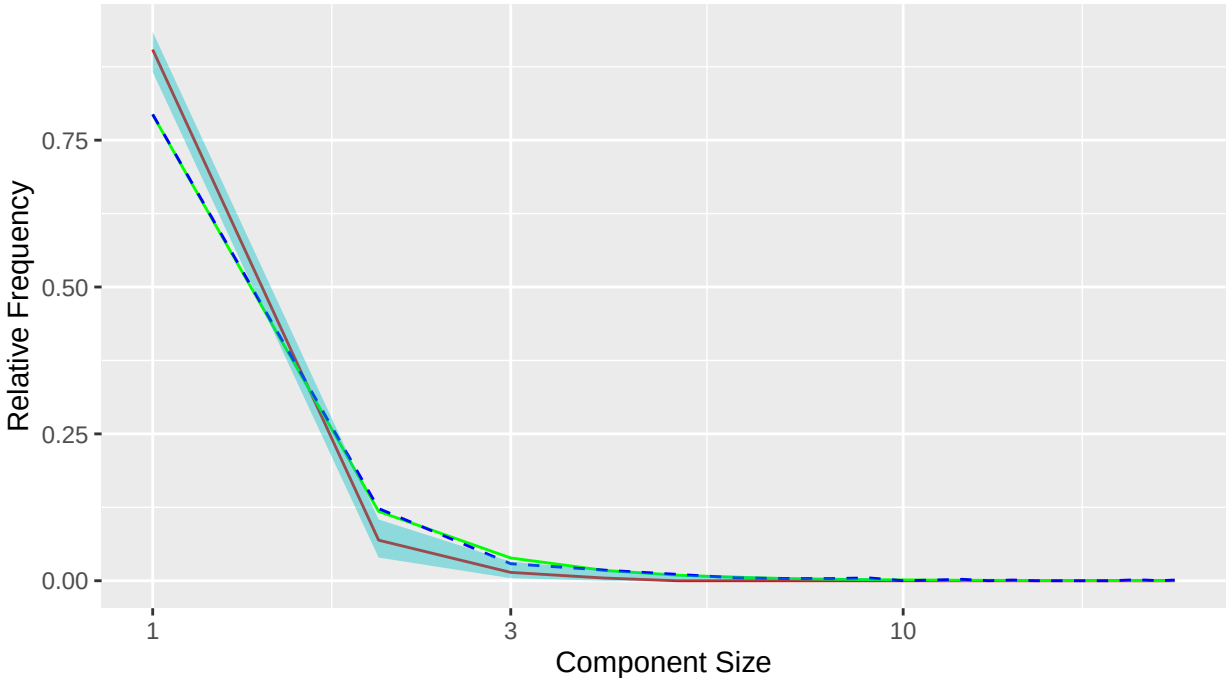
Conclusion: We lose the right tail with decreasing sampling fraction f

Table 4: Count of 1,000 Samples With One or More Components of Given Size

size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
n	1000	1000	976	759	451	258	134	58	26	7	1	0	0	0	0	0	0	0	0	0	0	0	0
In pop?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	No	Yes	Yes	No	Yes

The plot below shows the distribution of component sizes with our simulation envelope as detailed in Part 2. We now see a substantial difference appearing between the observed relative frequency from the population and the median relative frequency across the 1,000 samples. The sum of the squared deviations between the later two has now increased to 0.0156. We remove the the $\log_{10}$ scale for the y-axis to show this in another light.

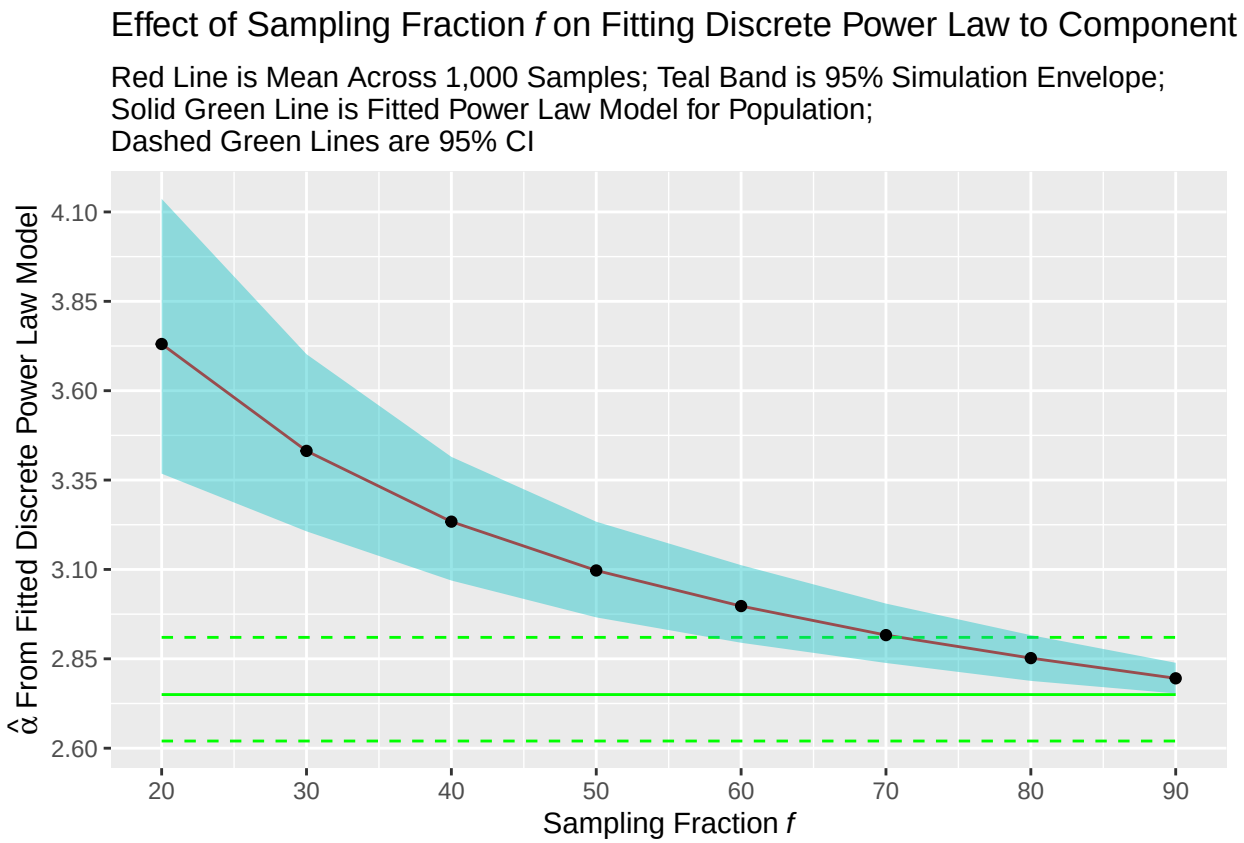

Distribution of Component Size for Network Samples (f = 0.20)
Log_10 Scale for Both Axes; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population
Relative Frequency
1e+00
1e−01
1e−02
1e−03
1e−04
1
3
10
Component Size


Distribution of Component Size for Network Samples (f = 0.20)
Log_10 Scale for Both Axes; Red Line is Median Rel Freq Across 1,000 Samples;
Teal Band is 95% Simulation Envelope; Blue Dashed Line is Population
Green Line is Fitted Power Law Model for Population
Relative Frequency
0.75
0.50
0.25
0.00
1
3
10
Component Size

Conclusion: Taking a sample of 20% of the nodes from the population, looking at genetic similarity, and the resulting observed component sizes in the sample, would not be an adequate representation of the unobserved component sizes in the population.

Part 5: Effect of Sampling Fraction f on Fitting Discrete Power Law Model to Component Size

In this section, we use the same approaches as described previously. For a sequence of sampling fractions $f \in \{0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90\}$, we repeat the process of fitting the discrete power law model on component size for 1,000 samples at each f. Next, with respect to $\widehat{\alpha}$, we extract the mean and the 0.025 and 0.975 quantiles (see red line and teal band in plot below). This is compared to $\widehat{\alpha}$ and its 95% confidence interval obtained from fitting the power law model to the entire population network (see green lines in plot below). This whole process is quite computationally expensive.



Some interesting results can be clearly observed. For example, we can see that below a sampling fraction $f = 0.70$, the estimate of the mean of the sampling distribution for $\widehat{\alpha}$ (black dot) falls outside the 95% confidence interval for α (dashed green lines) with respect to the population network. This would suggest a sample size "cutoff" for which bias in estimation of α will become an issue. Notice that samples corresponding to $f = 0.55$ or less, say, will yield $\widehat{\alpha}$ that will not be representative of the population network; the smaller the sampling fraction, the more the teal band moves away from the green lines. Finally, and unsurprisingly, it is clear that the smaller the sampling fraction, the more sampling variability there is in estimating α (blue band increasingly fans out).

Assuming a power model for component size is reasonable, the next big question would be, "*If we know "small" sample sizes are not representative of the population network, can we determine a bias correction factor for estimation of α?*" In studying the plot and thinking about this problem, it would feel to me that the answer has to be "yes." Perhaps this is where Nicole's paper comes into play; I am also thinking about other approaches. Stay tuned :)

Part 6: Bias and Exponential Decay

We can use nonlinear least squares to fit an exponential decay model to our estimates of α ($n = 8,000$) for our sequence of sampling fractions f. A plausible 3-parameter exponential decay model for $\widehat{\alpha}$ would be:

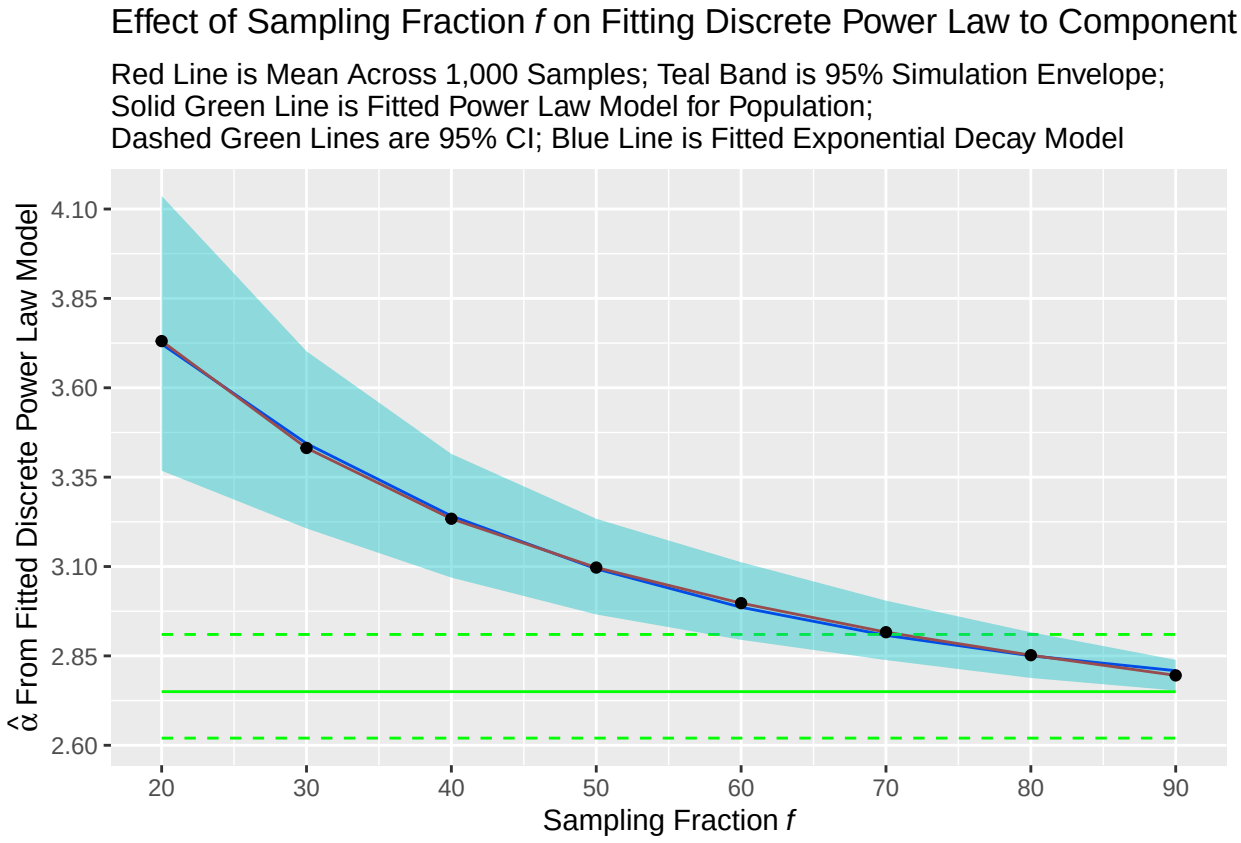
$$\widehat{\alpha} = (c + (d - c)\exp(-rf)) + \varepsilon$$

where c is the lower/right asymptote as $f \to 100$, d is the expected value of Y at $f = 0$, r is the "steepness," or rate of decay, f is the sampling fraction (expressed as a percentage), and ε is a normal error term with mean 0 and constant variance σ^2.

If $Bias[\widehat{\alpha}] = E[\widehat{\alpha}] - \alpha$, and we set $c = \alpha$, then it is easy to see that $Bias[\widehat{\alpha}] = (d - \alpha)\exp(-rf))$.

I should also add that we could fit this model using generalized least squares. In that case, the errors are allowed to be correlated and/or have unequal variances. However, at this point, that level of detail is not neccesary.

Upon fitting this model, we observe $\widehat{c} = 2.70$, $\widehat{d} = 4.63$, and $\widehat{r} = 0.0316$. In the plot below, we take the previous plot and add a blue line corresponding to the fitted exponential decay model. The fit remarkably tracks the interpolated mean (red line). Based on my experience with finite population sampling, I personally do not believe this is coincidence nor that we have blundered upon a model that conforms to what we see. Said another way, I truly do believe that bias for $\widehat{\alpha}$ is a nonlinear function of the sampling fraction f.



References

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703.

Gabaix, Xavier, and Rustam Ibragimov. 2011. "Rank- 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents." *Journal of Business & Economic Statistics* 29 (1): 24–39.

Gillespie, Colin S. 2015. "Fitting Heavy Tailed Distributions: The **poweRlaw** Package." *Journal of Statistical Software* 64 (2). https://doi.org/10.18637/jss.v064.i02.