Comparison between Configuration Model with i.i.d. Binomial(n, p) Degree Distribution and the Corresponding Erdős-Rényi Binomial Random Graph

Saptashwa Baisya
BS2226

A Joint work with Somarddha Das

13th May, 2025

Contents

This report entails the simulation studies carried out by us in order to study the (possible) differences between Configuration Model with i.i.d. Binomial(n,p) Degree Distribution and the Corresponding Erdős-Rényi Binomial Random Graph. All the guesses are simulation based and require theoretical backing either in the form of a proof or a disproof.
We recall the following definitions:

1 Definitions

- $G(n,p)$: Consider a simple graph with vertex set $V = [n]$ and edge set E. Label the $\binom{n}{2}$ pair of vertices $p_1, \cdots, p_{\binom{n}{2}}$. Let $e_1, \cdots, e_{\binom{n}{2}} \sim Ber(p)$. Set $E = \{p_i | e_i = 1\}$. $G = (V, E)$ is said to be an Erdős-Rényi Binomial Random Graph and it is said $G \sim G_{n,p}$ or $G \sim ER_{n,p}$
- $Conf(n,p)$: Given a sequence of n non-negative integers $d_1, \cdots, d_n$ (with an even sum), one can construct a (possibly multi-) graph (which is random) on $V = [n]$ by the method of joining half-edges such that vertex i has degree d_i $\forall i$. Let $d_1, \cdots, d_n \sim Bin(n-1, p)$ and if $\sum_i d_i$ is odd, we increase d_1 by 1. (To guarantee that the sum of degrees is even). Let the random graph generated by the above procedure be G, then we say that $G \sim Conf_{n,p}$

2 Works

2.1 Probability of Simplicity:

Before getting to the comparisons, we would like to study the $Conf(n,p)$ model. In case of the usual configuration model, we have the following result:
Theorem 2.1.1: Let $D_1, \cdots, D_n, \cdots$ be iid copies of D, where $\mathbb{E}(D^2) < \infty$ and $\mathbb{P}(D \geq 1) = 1$. G_n denote the configuration model obtained by using $D_1, \cdots, D_n$ (adjusted, if required) if S_n, M_n denote the number of self loops and multiple edges in G_n then

$$(S_n, M_n) \xrightarrow{d} (S, M)$$

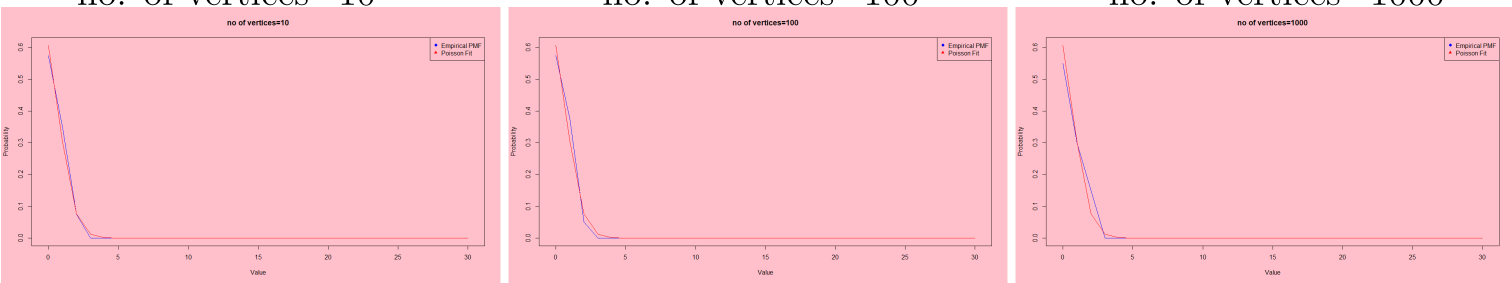
where $\nu = \dfrac{\mathbb{E}(D(D-1))}{\mathbb{E}(D)}$, $S \sim Poi\left(\dfrac{\nu}{2}\right)$, $M \sim Poi\left(\dfrac{\nu^2}{4}\right)$ *and S, M are independent.*
Corollary: $\lim_{n \to \infty} \mathbb{P}(G_n \text{ is simple}) = e^{-\frac{\nu}{2} - \frac{\nu^2}{4}}$
We would like to see if such a result holds incase $G_n \sim Conf(n, \dfrac{\lambda}{n})$. If $G \sim Conf(n, \dfrac{\lambda}{n})$: Then for large enough n, the degree distribution will be roughly $Poi(\lambda)$. Thus it seems reasonable to guess that the above theorem *might* hold true in this case, with $D \sim Poi(\lambda)$.

Note: In that case $\nu = \dfrac{\mathbb{E}(D(D-1))}{\mathbb{E}(D)} = \lambda$.

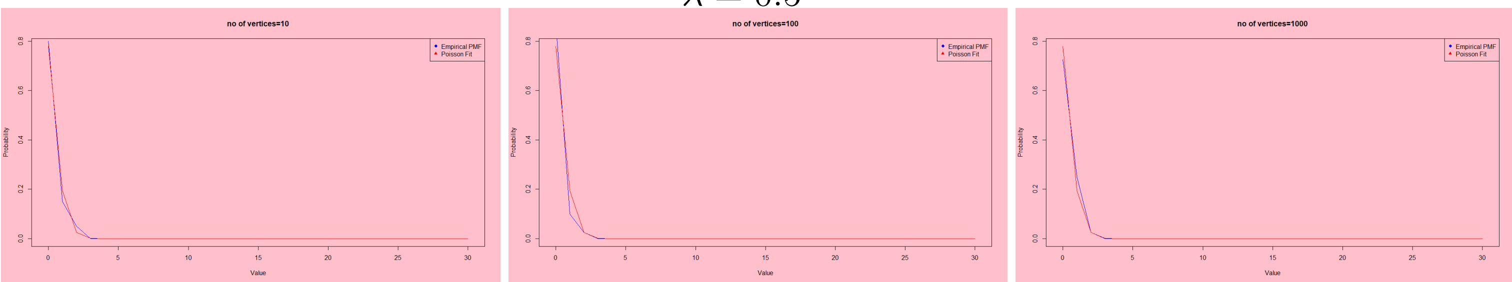
For $G_n \sim Conf(n, \dfrac{\lambda}{n})$, we define S_n, M_n similarly.

- We check if $S_n \xrightarrow{d} Poi\left(\dfrac{\lambda}{2}\right)$

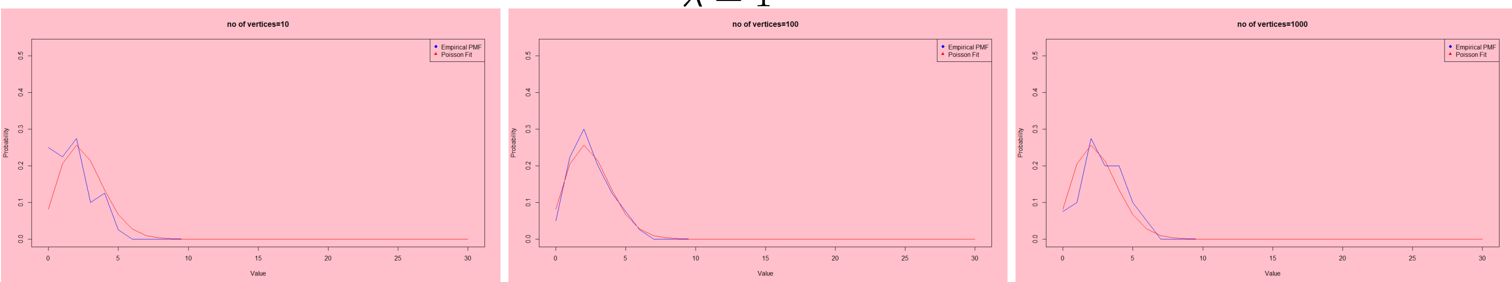
no. of vertices=10 no. of vertices=100 no. of vertices=1000



$\lambda = 0.5$



$\lambda = 1$



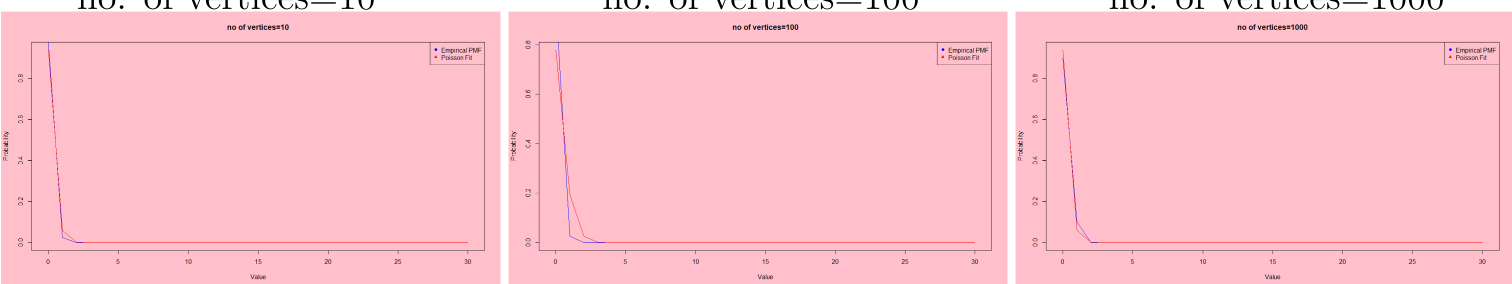
$\lambda = 5$

In each plot, the red curve is the probability mass function for $Poi(\dfrac{\lambda}{2})$, while the blue curve is the emperical probability mass function.

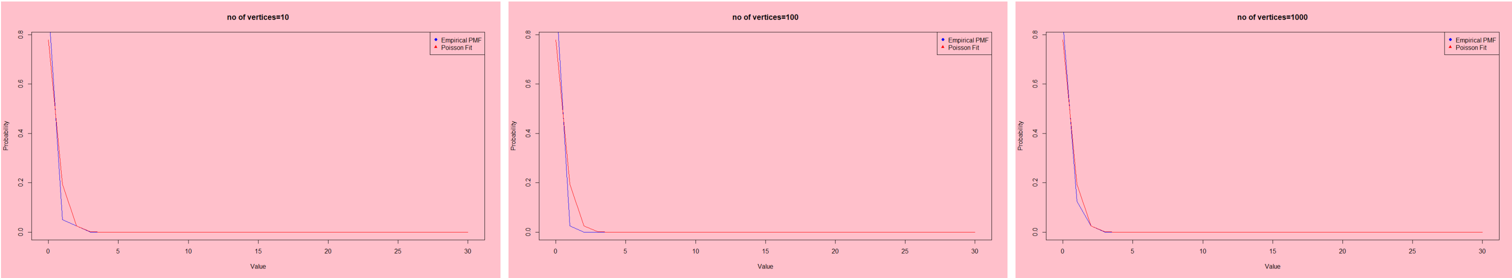
The Poisson fit, appears to be a good one, especially for large values of n (no. of vertices).

- We check if $M_n \xrightarrow{d} Poi\left(\dfrac{\lambda^2}{4}\right)$

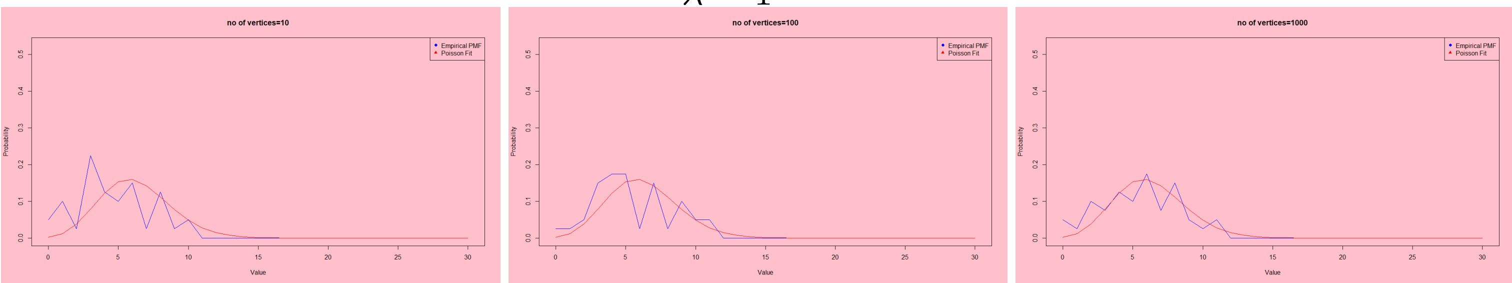
no. of vertices=10 no. of vertices=100 no. of vertices=1000



$\lambda = 0.5$



$$\lambda = 1$$



$$\lambda = 5$$

In each plot, the red curve is the probability mass function for $Poi\left(\frac{\lambda^2}{4}\right)$, while the blue curve is the emperical probability mass function.

For small vaues of λ, the fit becomes good for small values of n as the number of edges are very small. However for larger values of λ, the Poisson fit becomes better as n increases.

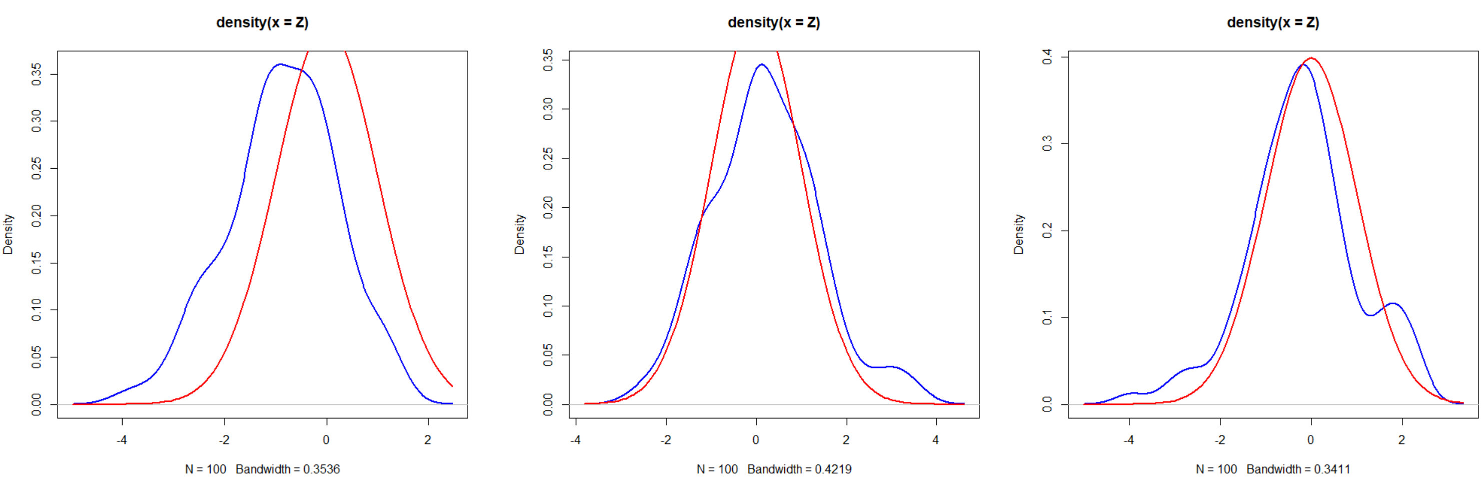
- We now finally check if asymptotically S_n, M_n are independent:

  We take help of Kendall's Tau (τ).

  Let τ_m denote the kendalls tau calculated using the points $(X_1, Y_1), \cdots, (X_m, Y_m)$ (where all the points are iid). Then under the assumption that X_1, Y_1 are independent

  $$\sqrt{m}\tau_m \xrightarrow{d} N\left(0, \frac{4}{9}\right)$$

  For various values of n, we plot the empirical density of a properly scaled τ (calculated using 40 pair of values (value of m in above expression) of S_n, M_n) and overlay the pdf of $N(0,1)$ over it, to see if independence might hold.



No. of vertices=10 No. of vertices=100 No. of vertices=1000

Here, the red curve is the pdf of $N(0,1)$, whereas the blue curve is the estimated pmf of $\frac{3}{2}\sqrt{n}\tau_n$.

We observe that the 2 curves are close, especially for larger number of vertices, which indicates that independence might actually hold.

Using these observations, we state the following conjecture analogous to Theorem 2.1.1:

Conjecture 2.1.1: Let $G_n \sim Conf(n, \frac{\lambda}{n})$. S_n, M_n denote the number of self-loops and multiple edges in G_n. Then

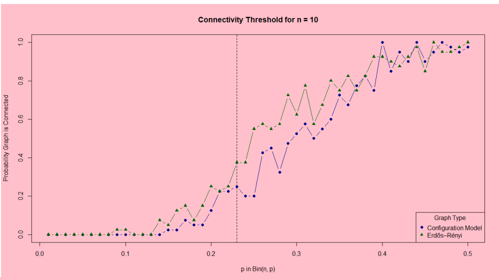
$$(S_n, M_n) \xrightarrow{d} (S, M)$$

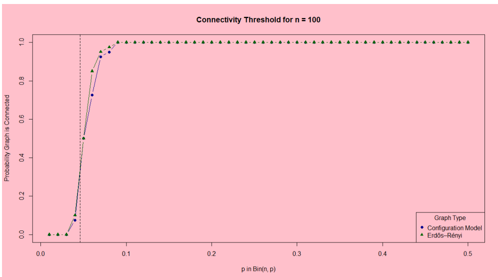
where $S \sim Poi\left(\frac{\lambda}{2}\right)$, $M \sim Poi\left(\frac{\lambda^2}{4}\right)$ and S, M are independent.

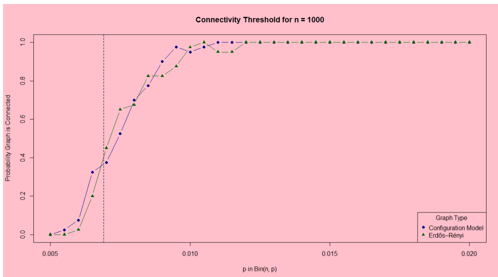
2.2 Threshold of Some Known Properties

We now return to the comparison of $G_{n,p}$ and $Conf(n,p)$ Now we look at thresholds of some known properties.

- **Connectivity Threshold:**





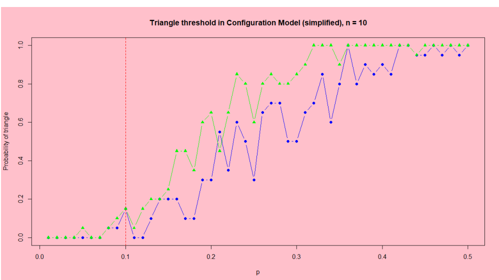


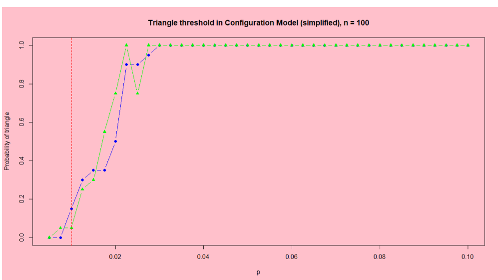
No. of vertices=10 No. of vertices=100 No. of vertices=1000

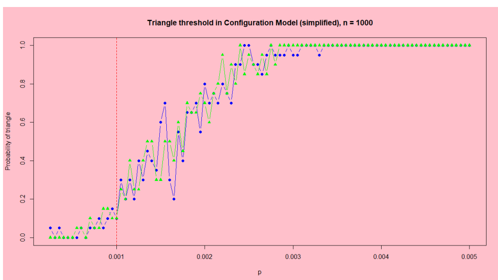
We look at the plots of probability of connectivity of both $G_{n,p}$ (green) and $Conf(n,p)$ (blue) as a function of p after fixing n. The 2 curves appear to be very close. The vertical line $x = \frac{log(n)}{n}$ (connectivity threshold for $G_{n,p}$) has been added to aid the visualization. This leads us to the following conjecture:

Conjecture 2.2.1: $\hat{p}_n = \frac{log(n)}{n}$ is the connectivity threshold for $Conf(n,p)$

- **Triangle Containment Threshold**





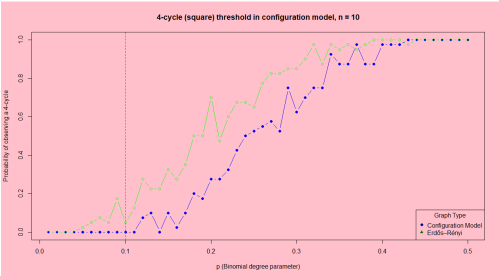


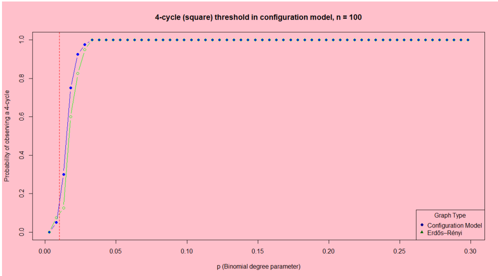
No. of vertices=10 No. of vertices=100 No. of vertices=1000

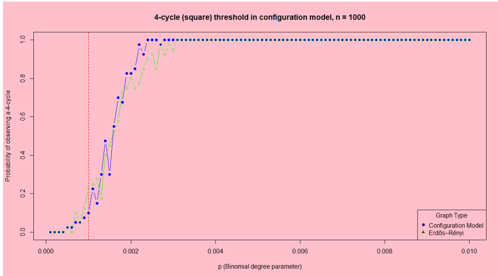
We repeat the same exercise here, and we add the vertical line $x = \frac{1}{n}$ (Triangle Containment Threshold for $G_{n,p}$). This leads us to the following conjecture:

Conjecture 2.2.2: $\hat{p}_n = \frac{1}{n}$ *is the triangle containment threshold for* $Conf(n,p)$

- **Square Containment Threshold**





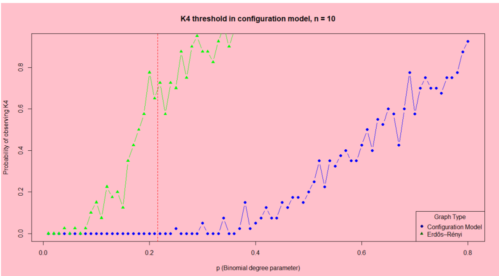


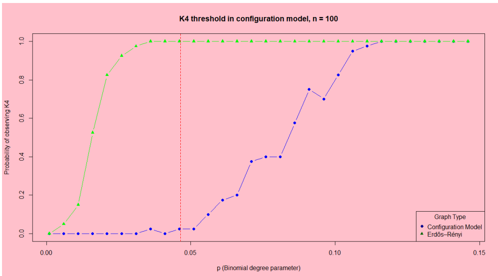
No. of vertices=10 No. of vertices=100 No. of vertices=1000

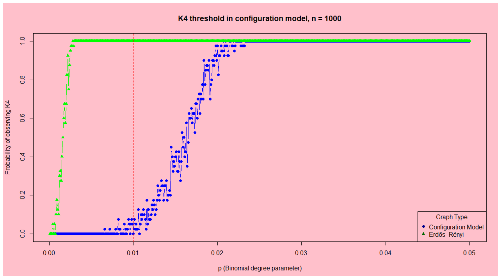
We repeat the same exercise here, and we add the vertical line $x = \frac{1}{n}$ (Square Containment Threshold for $G_{n,p}$). This leads us to the following conjecture:

Conjecture 2.2.3: $\hat{p}_n = \frac{1}{n}$ *is the square containment threshold for* $Conf(n,p)$

- **K_4 Containment Threshold**







No. of vertices=10 No. of vertices=100 No. of vertices=1000

We repeat the same exercise here, and we add the vertical line $x = \frac{1}{n^{\frac{2}{3}}}$ (K_4 Containment Threshold for $G_{n,p}$). This leads us to the following conjecture:

Conjecture 2.2.4: $\hat{p}_n = \frac{1}{n^{\frac{2}{3}}}$ *is the* K_4 *containment threshold for* $Conf(n,p)$

2.3 Size of the largest connected component

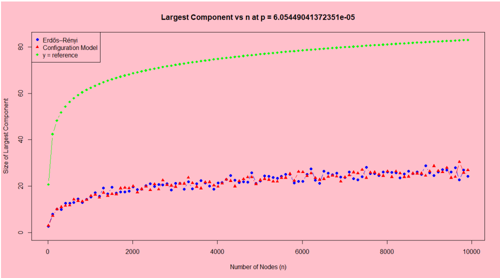
For $Conf(n, \frac{\lambda}{n})$, we define (analogously to $G_{n,\frac{\lambda}{n}}$) the sub-critical regime to be $\lambda < 1$ and the super-critical regime to be $\lambda > 1$. Now for each of the regimes, we fix a value of λ and we plot the size of the largest connected component as a function of n (the total number of vertices) for both $G_{n,p}$ and $Conf(n,p)$

- **The Sub-Critical Regime:** $\lambda < 1$

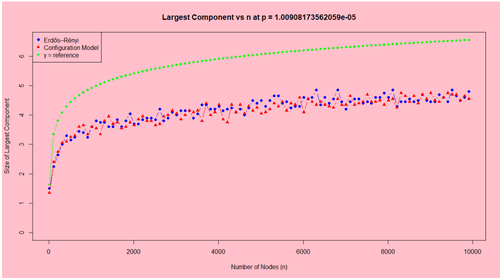
Theorem 2.3.1: If $G_n \sim G_{n,\frac{\lambda}{n}}$*, and* $|C_{max}|$ *denote the size of the largest connected component of* G_n*, then*

$$\frac{|C_{max}|}{log(n)} \xrightarrow{P} \frac{1}{I_\lambda} \text{ , where } I_\lambda = \lambda - 1 - log(\lambda)$$

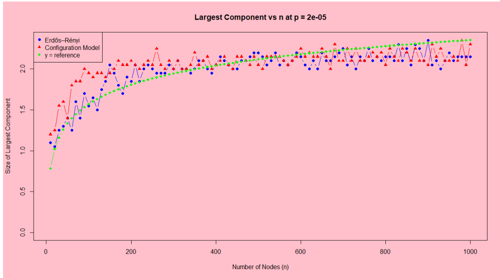
We wish to see if a similar result holds for $Conf\left(n, \frac{\lambda}{n}\right)$.



$\lambda = 0.6$



$\lambda = 0.1$



$\lambda = 0.02$

The red curve indicates the size of the largest connected component in the $Conf(n, \frac{\lambda}{n})$, whereas the blue curve indicates the size of the largest connected component in the $G_{n,\frac{\lambda}{n}}$. The green curve is the reference size $\frac{log(n)}{I_\lambda}$. It appears that the 2 curves near to each other however may are very far from the reference curve. However we expect them to come close, except for the fact that it will happen after a very large n, this is because the rate of increase of $log(x)$ decreases as x increases.

We also observe that as λ moves farther from the critical value of 1, the green curve gets closer to the red and blue curves, i.e. the convergence occurs for relatively small values of n.

Conjecture 2.3.1: If $G_n \sim conf(n, \frac{\lambda}{n})$, and $|C_{max}|$ denote the size of the largest connected component of G_n, then

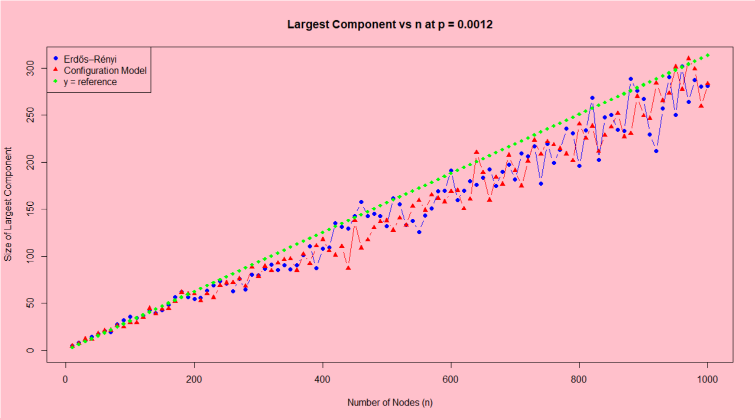
$$\frac{|C_{max}|}{log(n)} \xrightarrow{P} \frac{1}{I_\lambda} \text{ , where } I_\lambda = \lambda - 1 - log(\lambda)$$

- **The Super-Critical Regime:** $\lambda > 1$

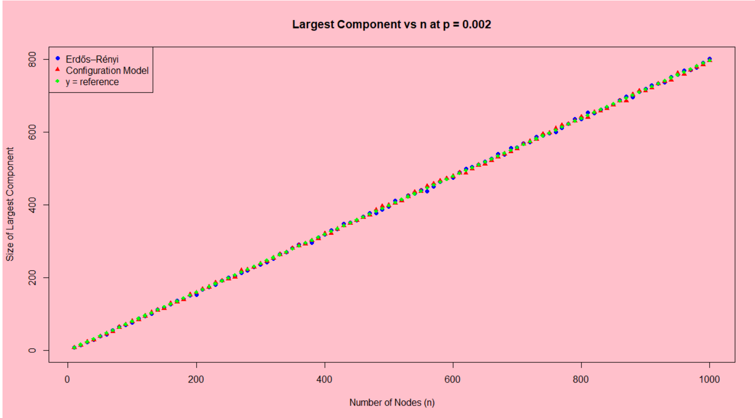
Theorem 2.3.2: If $G_n \sim G_{n,\frac{\lambda}{n}}$, and $|C_{max}|$ denote the size of the largest connected component of G_n, then

$$\frac{|C_{max}|}{n} \xrightarrow{P} \zeta_\lambda \text{ , where } \zeta_\lambda \text{ is the non zero-root of } 1 - x - e^{\lambda x}$$

We wish to see if a similar result holds for $Conf\left(n, \frac{\lambda}{n}\right)$.



$\lambda = 1.2$



$\lambda = 2$

The red curve indicates the size of the largest connected component in the $Conf(n, \dfrac{\lambda}{n})$, whereas the blue curve indicates the size of the largest connected component in the $G_{n,\frac{\lambda}{n}}$. The green curve is the reference size $n\zeta_\lambda$.

The 3 curves are close, but they come closer as λ moves farther from the critical value of $\lambda = 1$.

Conjecture 2.3.2: If $G_n \sim conf(n, \frac{\lambda}{n})$, and $|C_{max}|$ denote the size of the largest connected component of G_n, then

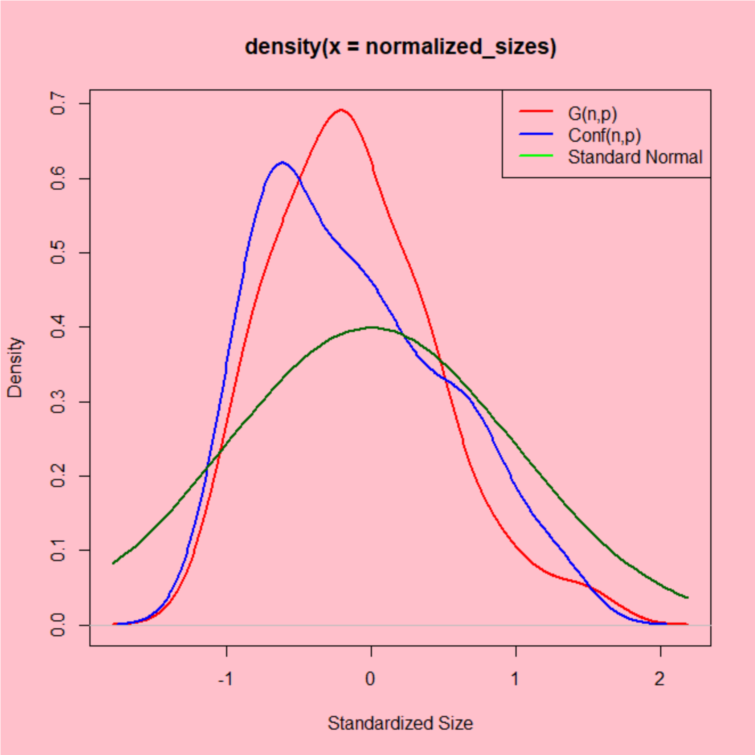
$$\frac{|C_{max}|}{n} \xrightarrow{P} \zeta_\lambda \text{ , where } \zeta_\lambda \text{ is the non zero-root of } 1 - x - e^{\lambda x}$$

- **CLT for Giant Components in Super-Critical Regime:**

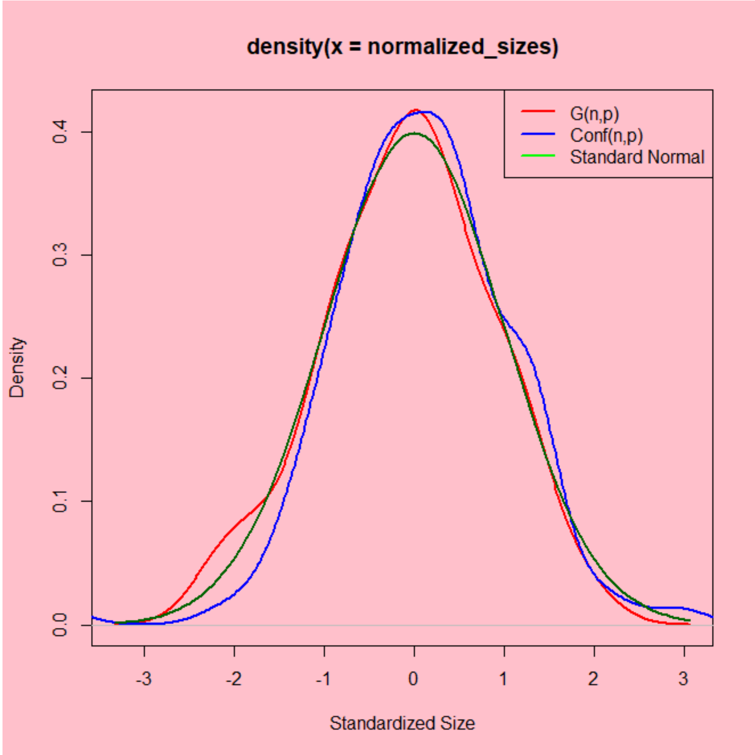
Theorem 2.3.1: Fix $\lambda > 1$, if $G_n \sim G_{n,\frac{\lambda}{n}}$, and $|C_{max}|$ denote the size of the largest connected component of G_n, then

$$\frac{|C_{max}| - n\zeta_\lambda}{\sqrt{n \cdot \sigma_\lambda^2}} \xrightarrow{d} N(0,1), \text{ where } \sigma_\lambda^2 = \frac{\zeta_\lambda(1-\zeta_\lambda)}{(1-\lambda+\lambda\zeta_\lambda)^2}$$

We wish to see if a similar result holds for $Conf\left(n, \dfrac{\lambda}{n}\right)$.



$\lambda = 1.2$



$\lambda = 2$

For $n = 1000$ and the specified value of λ, we plot the density of $\dfrac{|C_{max}| - n\zeta_\lambda}{\sqrt{n \cdot \sigma_\lambda^2}}$ for $Conf\left(n, \dfrac{\lambda}{n}\right)$ (in blue) $G_{n,\frac{\lambda}{n}}$ (in red) and overlay the $N(0,1)$ pdf on them. Here too, we observe that as λ moves away from the critical value, the 3 curves get closer, indicating that as we move away from the critical value, for smaller n, we have a better fit of the asymptotic distribution.

Conjecture 2.3.3: Fix $\lambda > 1$, if $G_n \sim Conf\left(n, \dfrac{\lambda}{n}\right)$, and $|C_{max}|$ denote the size of the largest connected component of G_n, then

$$\frac{|C_{max}| - n\zeta_\lambda}{\sqrt{n \cdot \sigma_\lambda^2}} \xrightarrow{d} N(0,1), \text{ where } \sigma_\lambda^2 = \frac{\zeta_\lambda(1-\zeta_\lambda)}{(1-\lambda+\lambda\zeta_\lambda)^2}$$

3 Discussion

In this report, we see that the configuration model with $\text{Bin}(n-1,p)$ degrees closely mimics certain $G_{n,p}$ behavior. This might be because

- Degree distribution in $G_{n,p}$ is $Bin(n-1,p)$ and for large enough n, the dependence between any 2 degrees "weakens"

However some differences crop up, especially while we look at count statistics like the total number of $\triangle$s, $\square$s in the erased model. This might be because

- While looking for thresholds, in contrast to the count statistics, we are just interested in the **indicator** of the event. Edges contributing to self-loops and multiple edges do not affect the indicator of the event, but they do affect the count statistics.

4 Concluding Remarks

- There might be a result for $Conf\left(n, \dfrac{\lambda}{n}\right)$ analogous to the Results of Simplicity in the usual Configuration Model with iid degree.
- Connectivity Threshold, $\triangle$, $\square$, K_4 Containment Threshold appear to be same for both $G_{n,p}$ and $Conf(n, p)$.
- Sub-critical and super-critical Law of large number behaviour seem to be matching for both $G_{n,p}$ and $Conf(n, p)$.
- We also see that the CLT that we have for giant components in supercritical regime in $G_{n,p}$ might also hold for $Conf(n, p)$.

5 References

- *Random Graphs and Complex Networks, Volume I.* Cambridge University Press, 2016.