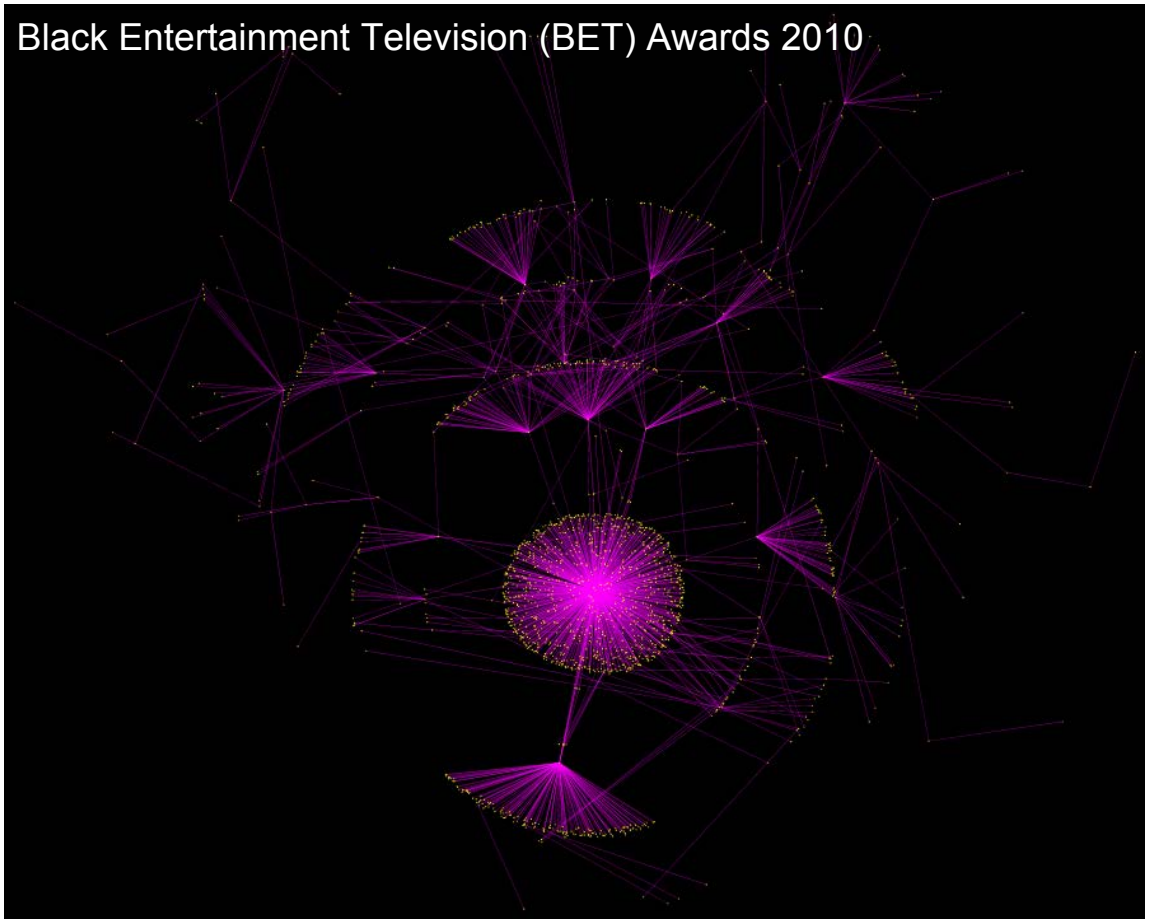
Superstar Model: ReTweets, Lady Gaga and Surgery on a Branching Process

J. Michael Steele

Simons Conference on Random Graph Processes
Austin, Texas 2016

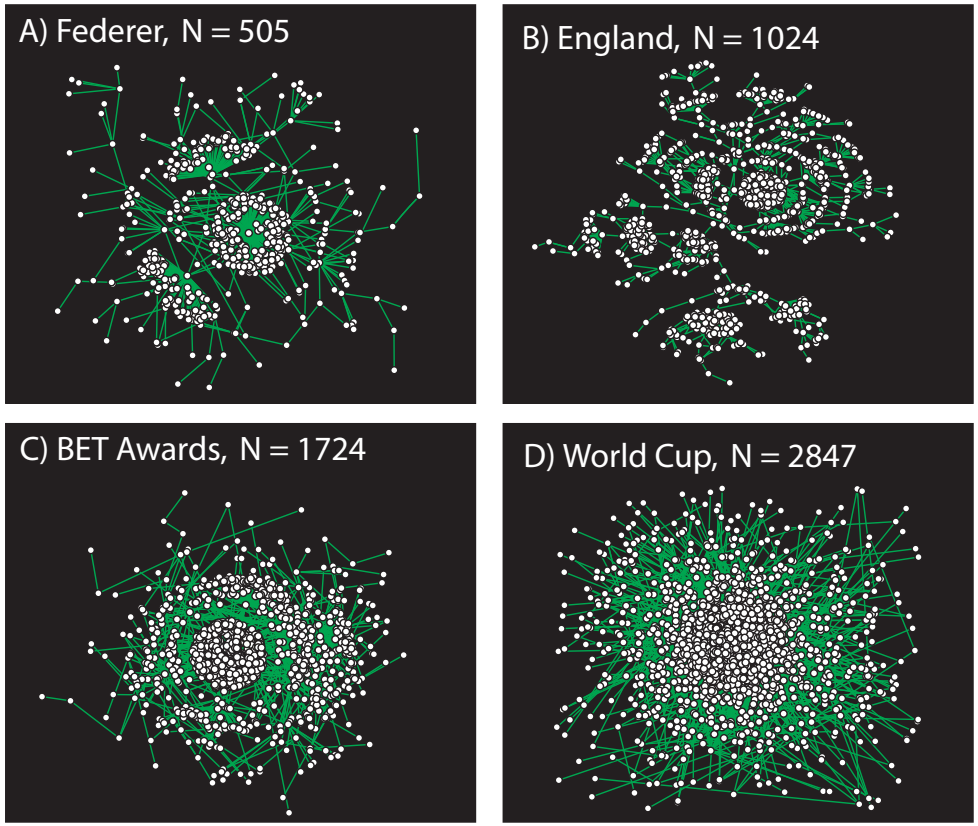
Passage from the Retweet Graph to the Superstar Model

- Joint work with Shankar Bhamidi (UNC) and Tauhid Zaman (MIT) — genuine members of the Twitter generation!
- Retweet graph: Given a topic and a time frame — form all the (undirected) *retweet arcs* and look at the giant component of the graph you get.



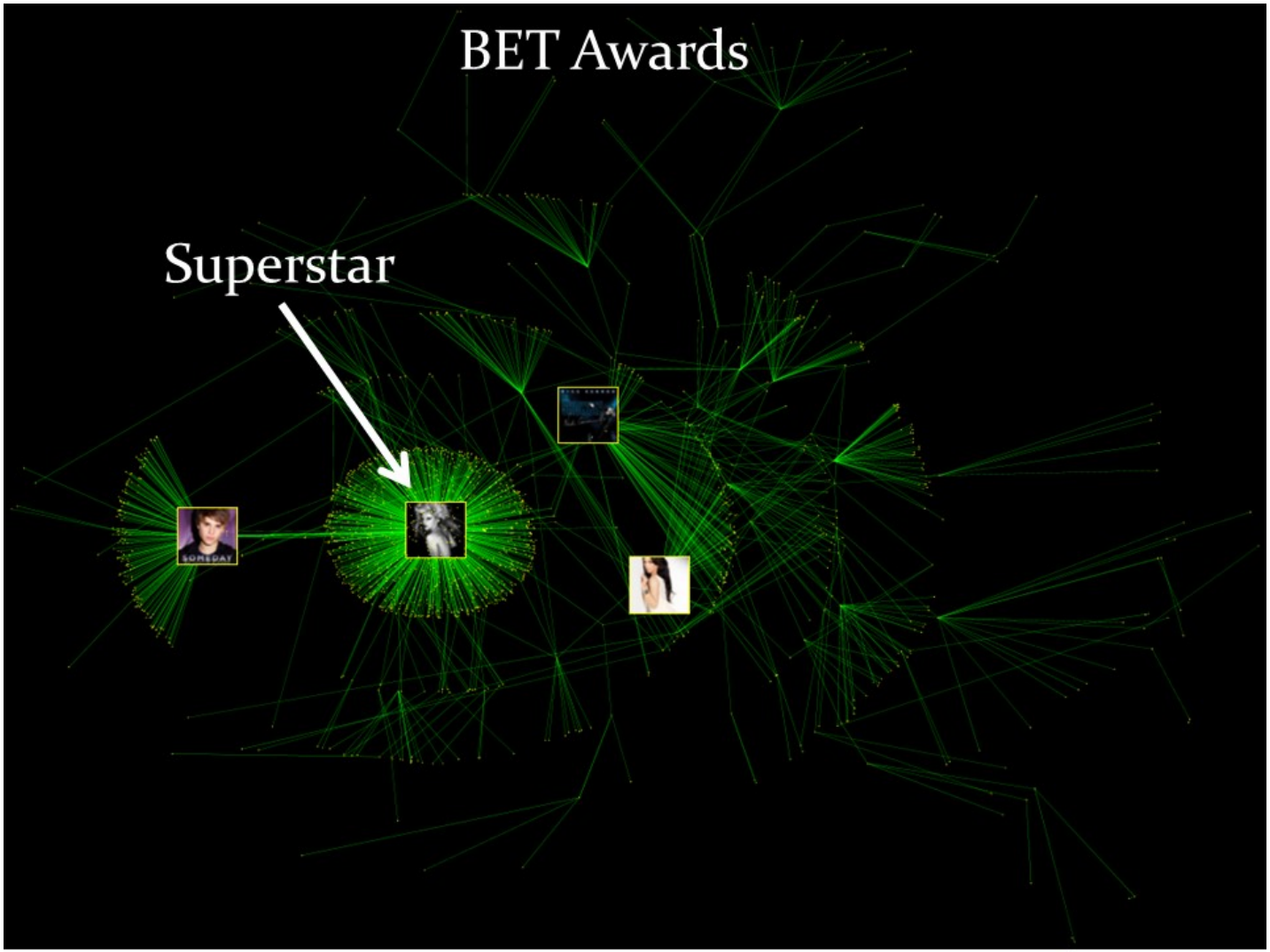
Reading the Message from Some Empirical Retweet Graphs

- Retweet graphs were constructed for 13 different public events [1]
  - Sports, breaking news stories, and entertainment events
  - Time range for each topic was between 4-6 hours
- Empirically the graphs are very tree-like (almost no cycles)
- Empirically the graphs each have one giant component — this is what we model
- The graphs are taken as undirected — and the the degrees tell the whole story



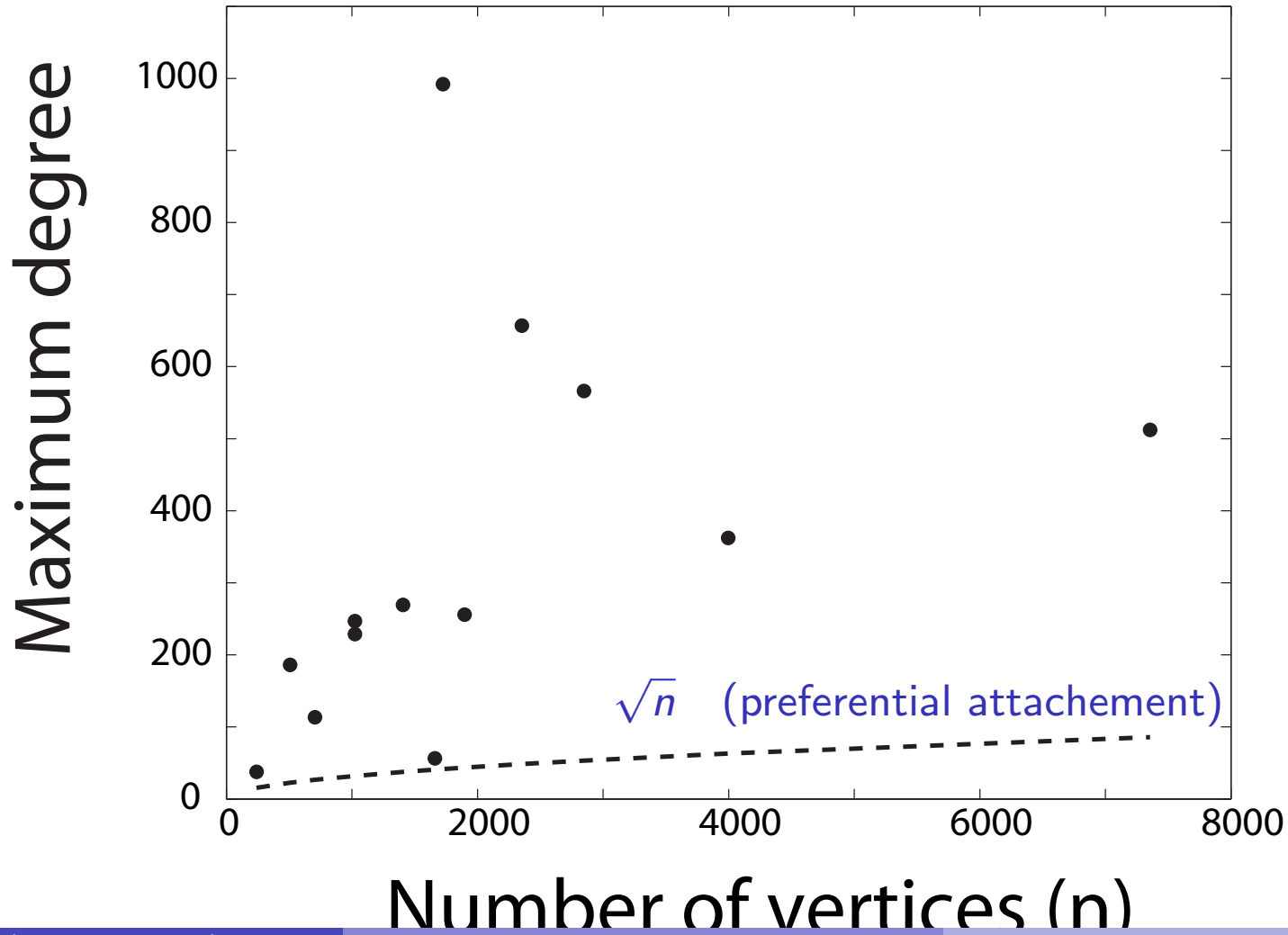
[1] Data courtesy of Microsoft Research, Cambridge, MA

BET 2010 Data — with Labels

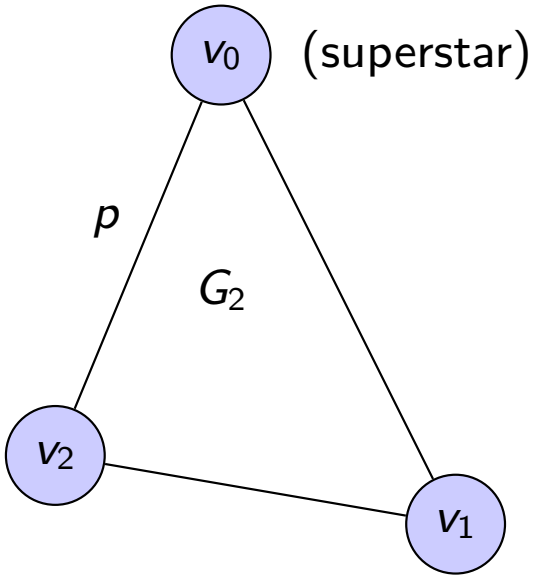


What Goes Wrong with Plain Vanilla Preferential Attachment?

- One finds Max degree in empirically observed retweet graphs have the order of the graph size, i.e. MaxDeg $\sim pn$
- Preferential attachment would predict sub-linear max degree



The Superstar Model — It's Completely Determined by p



- Attach to superstar with probability p
- Else with probability $1-p$ attach to one of the non-superstar vertices.
- Non-SS Attachment Rule: probability proportional to its degree (preferential attachment rule)

The only model parameter is p: The super star parameter

This is a very simple model: But (1) it has empirical benefits and (2) it is tractable — though not particularly easy.

The Degree of the Superstar Under the Superstar Model

Remark (Built-In Easy Fact)

Let $\deg(v_0, G_n)$ *be the degree of the superstar in* G_n*. We then have that*

$$\frac{\deg(v_0, G_n)}{n} \to p \qquad \textit{with probability 1 as } n \to \infty$$

- Empirically the Superstar degree is $\Theta(n)$ and the Superstar Model "Bakes this into the Cake"
- But that is ALL that is baked in...
- The value of p *predicts other features of the graph*
- The Superstar Model is *TESTABLE*.

The Most Starry of the Non-Superstars

Theorem

Let $\deg_{\max}(G_n)$ *be the maximal non-superstar degree in* G_n*, i.e.*

$$\deg_{\max}(G_n) = \max_{1 \le i \le n} \deg(v_i, G_n)..$$

If we set

$$\gamma = \frac{1-p}{2-p}.$$

then here is a non-degenerate, strictly positive, random variable Δ^* *such that*

$$n^{-\gamma}\deg_{\max}(G_n)) \to \Delta^* \qquad \textit{with probability 1 as } n \to \infty$$

- Maximal non-superstar degree is little-oh of the degree of the Superstar
- The Super Star Model makes an explicit **prediction** for the growth rate of maximum degree of a non-superstar.

Realized Degree Distribution in the Superstar Model

Theorem

Let $F(k, G_n)$ be the realized degree distribution of G_n under the Superstar model,

$$F(k, G_n) = n^{-1} \, |\{1 \le j \le n : \deg(v_j, G_n) = k\}|$$

and introduce the superstar model probability mass function

$$f_{SSM}(k, p) = \frac{2-p}{1-p}(k-1)! \prod_{i=1}^{k} \left(i + \frac{2-p}{1-p} \right)^{-1}.$$

We then have

$$F(k, G_n) \to f_{SSM}(k, p) \qquad \text{with probability 1 as } n \to \infty$$

- KEY POINT: The degree distribution scales like $k^{-\beta}$, where $\beta = 3 + p/(1-p)$
- This contrasts with the preferential attachment model which scales like k^{-3}

Superstar Model vs Preferential Attachment

Model	Superstar Model	Preferential Attachment
Superstar Degree	$\sim pn$	*NA*
Maximal non-superstar degree exponent	$\frac{1-p}{2-p}$	$\frac{1}{2}$
Degree distribution power-law exponent	$3 + \frac{p}{1-p}$	3

Superstar Model Predictions

- Use actual data $\widehat{G}_n$ to fit the superstar degree and predict the degree distribution
- Consider the observed degree distribution for each empirical retweet graph:

$$F(k, \widehat{G}_n) = n^{-1} \, |\{1 \le j \le n : \, \deg(v_j, G_n) = k\}|$$

- Consider the theoretical asymptotic degree distribution under the Superstar Model

$$f_{SSM}(k, p) = \frac{2-p}{1-p}(k-1)! \prod_{i=1}^{k} \left(i + \frac{2-p}{1-p} \right)^{-1}.$$

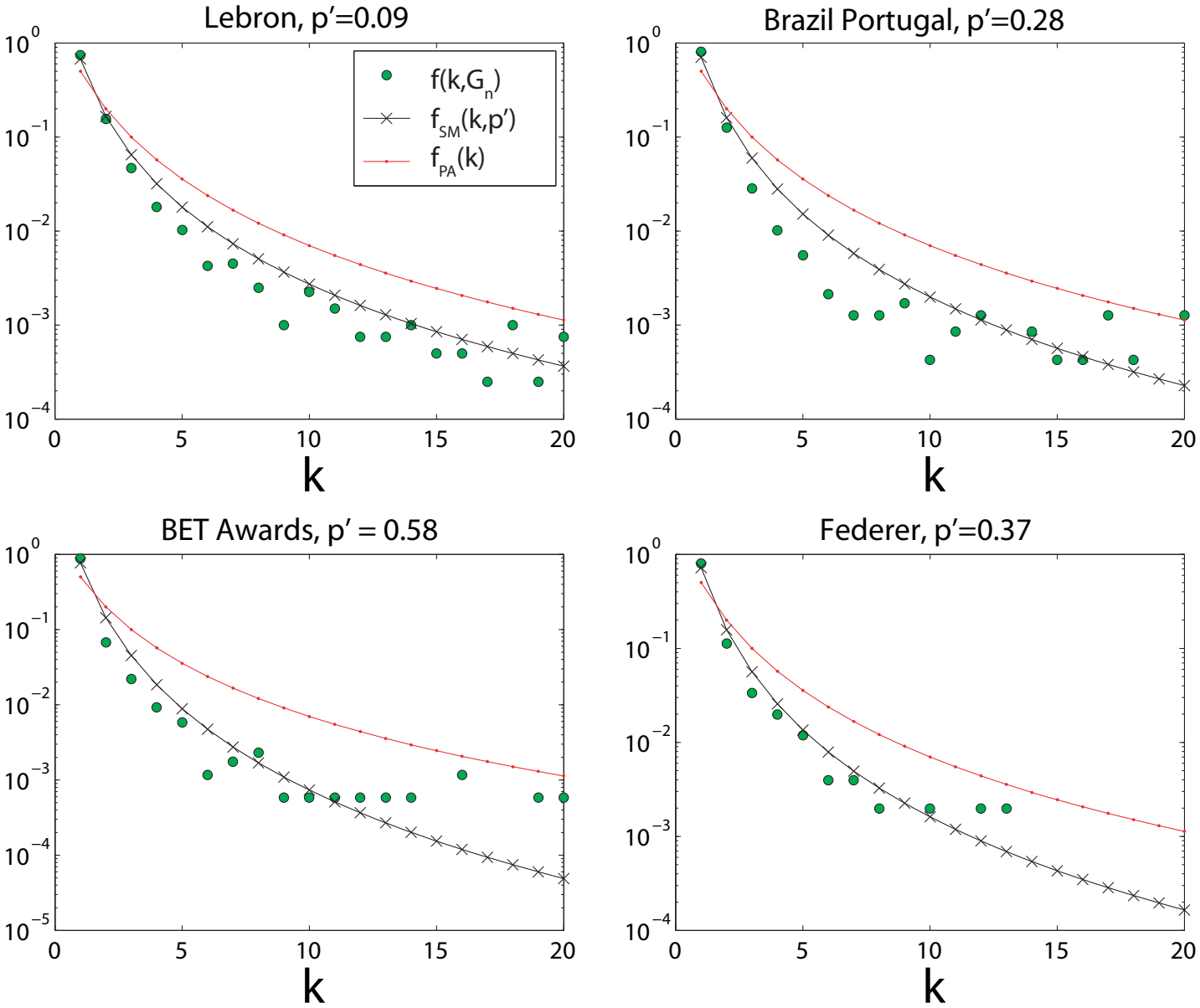
- Bottom Line: We get a pretty impressive fit "observed vs predicted"

$$F(k, \widehat{G}_n) \approx f_{SM}(k, \hat{p}) \qquad \text{where} \qquad \hat{p} = \frac{\text{observed superstar degree}}{n}$$

- Basis for Tests: Preferential Attachment always predicts...

$$f_{PA}(k) = \frac{4}{k(k+1)(k+2)}$$

Degree distribution

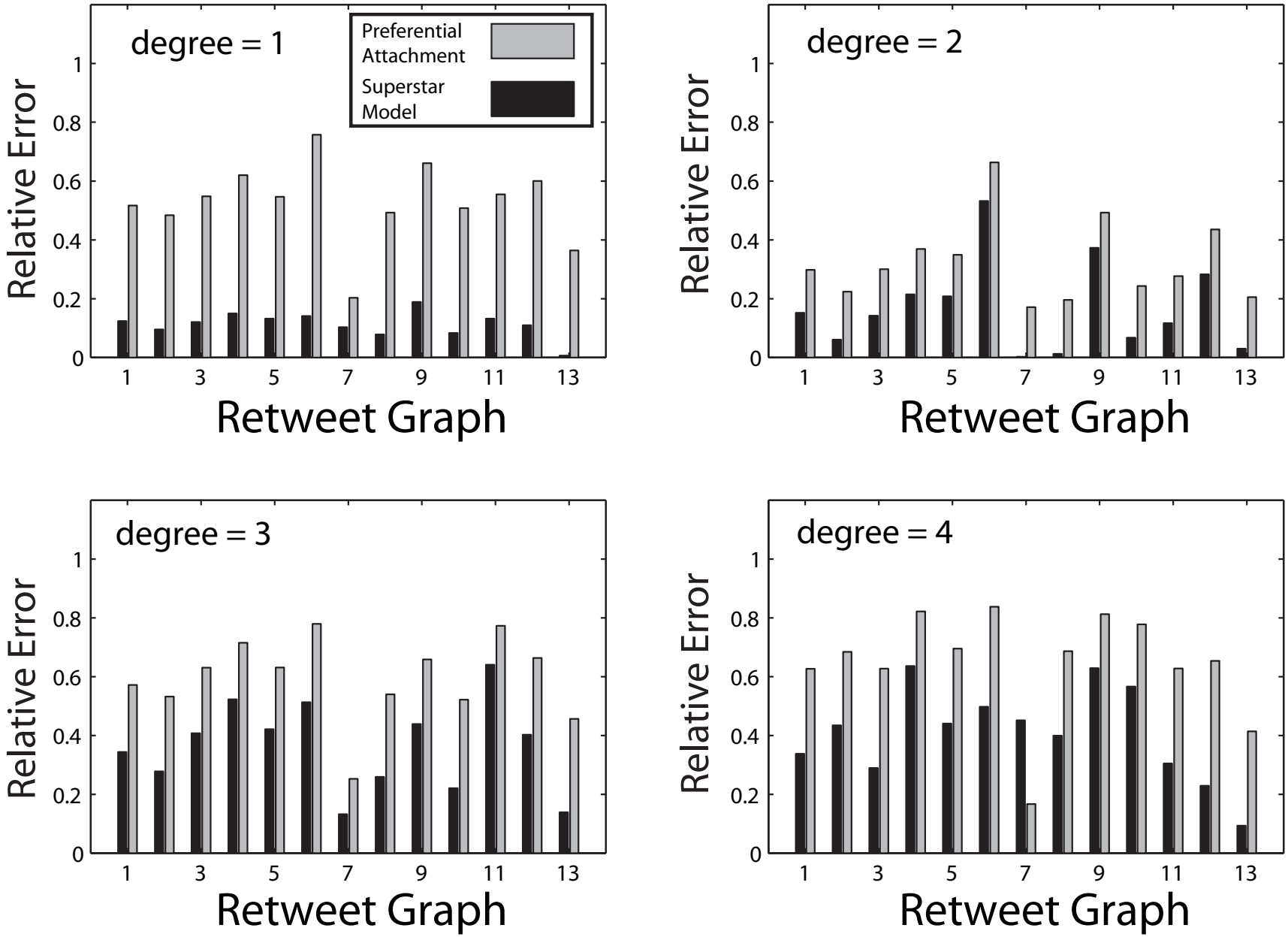


Degree distribution Comparison

Compare relative error of the Superstar Model and Preferential Attachment for different degrees k

Model	Superstar Model	Preferential Attachment
Relative Error	$\frac{\lvert f(k, G_n) - f_{SM}(k, p')\rvert}{f_{SM}(k, p')}$	$\frac{\lvert f(k, G_n) - f_{SM}(k, p')\rvert}{f_{SM}(k, p')}$

Degree Distribution Comparison



The Superstar Model and the Realized Degree Distribution: Bottom Line

- The Superstar Model implies a mathematical link between the superstar degree and the degree distribution of the non-superstars.
- When we look at Twitter data for actual events, we see (1) a superstar and (2) a degree distribution of non-superstars that is more compatible with the superstar model than with the preferential attachment model.
- The first property was "baked" into our model, but the second was not. It's an honest discovery.
- Next: How Can one Analyze the Superstar Model?

Basic Link: Branching Processes

- Proto-Idea: Branching processes have a natural role almost anytime one considers a stochastically evolving tree.

- More Concrete Observation: If the birth rates depend on the number of children, the arithmetic of the Poisson process relates lovingly to the arithmetic of preferential attachment — this is sweet.

- Creating the Superstar: Yule processes don't come with a superstar. Still, it is not terribly hard to move to multi-type branching processes. In a world with multiple types, you have the possibility of doing some surgery that let you build a super star.

- Realistic Expectations: The paper is a reasonably dense 35 pages. Some of the branching process theory is drawn from the dark well of experts; it's not off-the-shelf stuff. Still, if you want the deeper parts of the theory (e.g. the distribution of the maximum degree of the non-superstars) then you have to pay the piper.

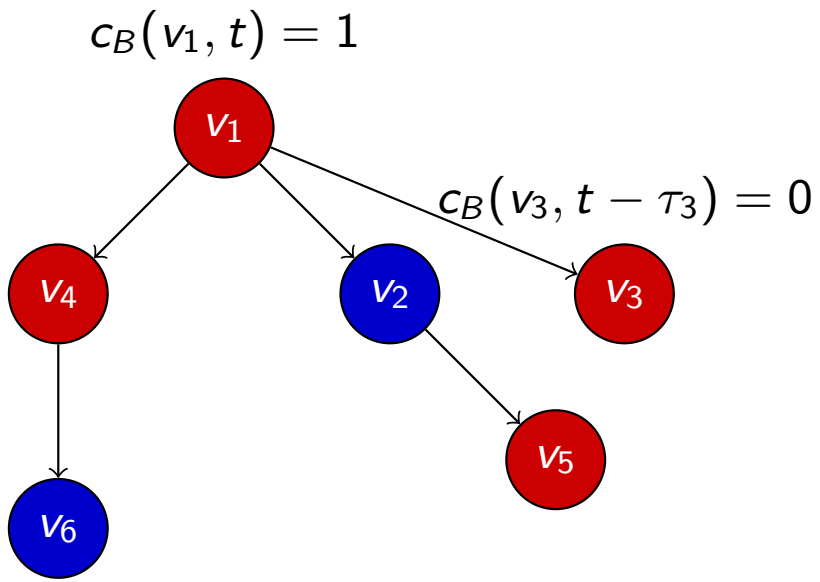
- News You Can Use? One can see the benefits of using multi-type branching processes. One can see that the connection between the Yule process and preferential attachment is natural. This is enough to get you rolling in a variety of applied probability models (social net works are a good start — but they are not the only game.)

Introduction of a Special Branching Process

- Two types of vertices: red and blue
- Each vertex gives birth to vertices according to a non-homogeneous Poisson process that has rate proportional to (1+ number of blue children)

$$c_B(v, t) = \text{number of blue children of } v \text{ at } t \text{ time units after the birth of v}$$

- At birth vertex is painted red with probability p and painted blue with probability $1 - p$

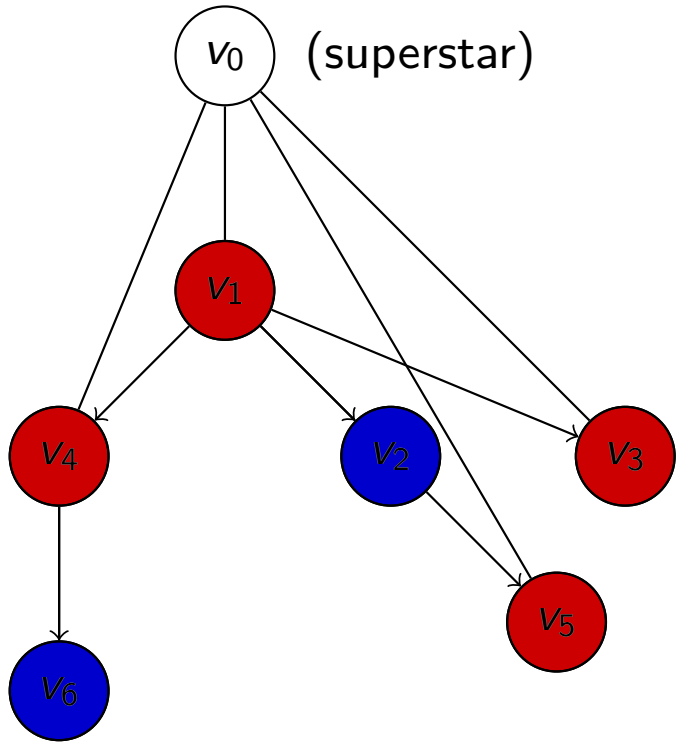


$$\begin{aligned} \mathcal{F}(t) &= \text{Branching process at time } t \\ \tau_n &= \inf\{t : |\mathcal{F}(t)| = n\} \end{aligned}$$

Surgery: From BP Model to Superstar Model

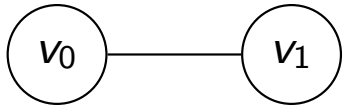
- Add an exogenous superstar vertex v_0 to the vertex set
- For each red vertex remove the edge from parent and create an undirected edge to the superstar vertex v_0
- With the surgery done, all edges are made undirected and all colors are erased

$\mathcal{F}(\tau_6)$



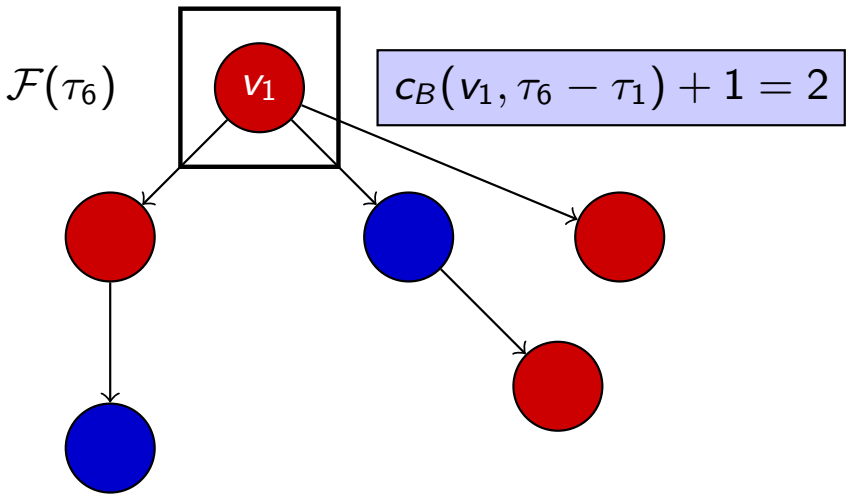
Relating the BP Construction with the Superstar Model

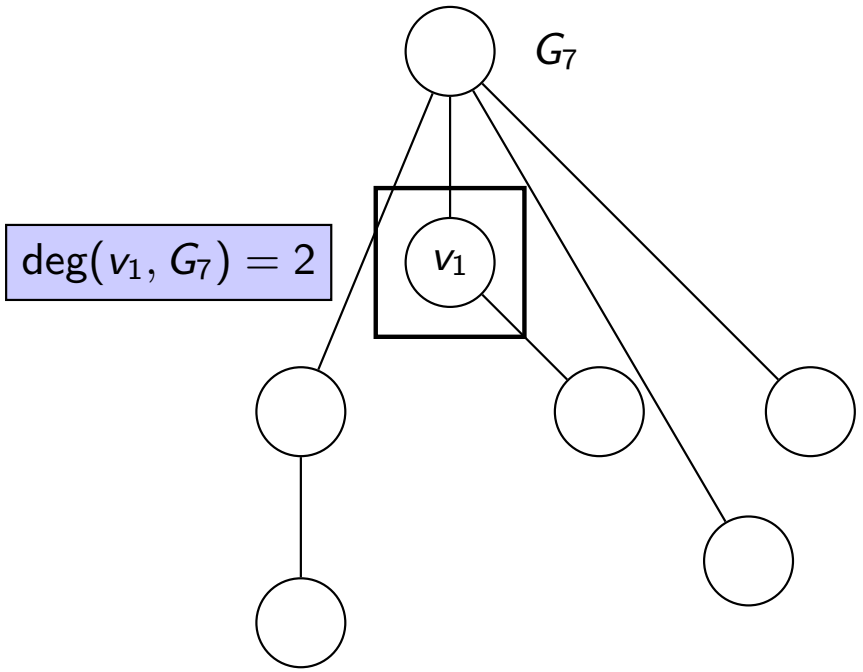
- Claim: $S(\tau_n)$ is "probabilistically the same" as G_{n+1}
- Base case: $S(\tau_1) = G_2$



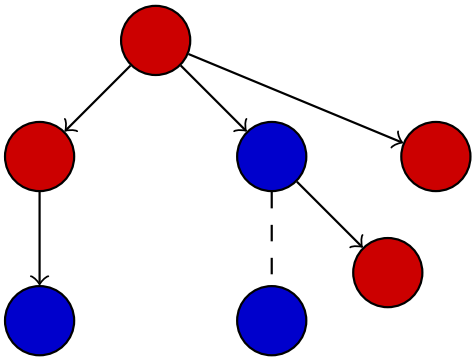
- Need to show that $S(\tau_n)$ and G_{n+1} have same probabilistic evolution
- Superstar: probability of joining superstar = probability of red vertex being born = p
- Same probability for S and G
- Non-superstars: degree of vertex = number of blue children + 1

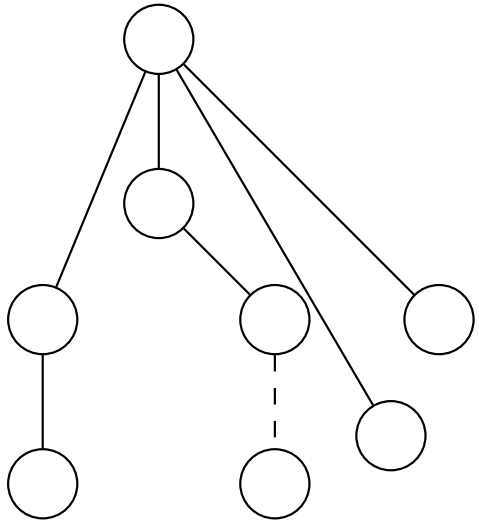
$$\deg(v_k, G_{n+1}) = c_B(v_k, \tau_n - \tau_k) + 1$$





Further Linking of the BP Model and the Superstar Model





$$\mathbb{P}\left(v_n \text{ joins } v_k | G_n\right) = \mathbb{P}\left(v_n \text{ is blue and born to } v_k | \mathcal{F}(\tau_{n-1})\right)$$

$$\mathbb{P}\left(v_n \text{ joins } v_k | G_n\right) = (1-p)\frac{\deg(v_k, G_n)}{\sum_{v_j \in G_n \setminus v_0} \deg(v_j, G_n)}$$

$$= (1-p)\frac{\deg(v_k, G_n)}{2(n-1) - \deg(v_0, G_n)}$$

$$\mathbb{P}\left(v_n \text{ is blue and born to } v_k | \mathcal{F}(\tau_{n-1})\right) = (1-p)\frac{c_B(v_k, \tau_n - \tau_k) + 1}{\sum_{v_k \in \mathcal{F}(\tau_{n-1})} c_B(v_k, \tau_n - \tau_k) + 1}$$

$$= (1-p)\frac{\deg(v_k, G_n)}{2(n-1) - \deg(v_0, G_n)}$$

Non-Superstar Degree

Theorem

There exists a strictly positive, non-degenerate, random variable W such that

$$|\mathcal{F}(t)|e^{-(2-p)t} \to W \qquad \textit{with probability 1 as } t \to \infty$$

The number of blue children is a Yule process with rate $1-p$

$$c_B(v_j,t)e^{-(1-p)t} \to T \qquad \text{where} \qquad T \sim \text{Exp}(1-p)$$

$$\begin{aligned}
\frac{\deg(v_j,G_n)}{n^{(2-p)^{-1}(1-p)}} &\approx \frac{c_B(v_j,\tau_n-\tau_j)}{|\mathcal{F}(\tau_{n-1})|^{(2-p)^{-1}(1-p)}} \\
&= \frac{c_B(v_j,\tau_n-\tau_j)e^{-(1-p)\tau_n}}{(|\mathcal{F}(\tau_{n-1})|e^{-(2-p)\tau_n})^{(2-p)^{-1}(1-p)}} \\
&\to \frac{T}{W^{(2-p)^{-1}(1-p)}} \qquad \text{with probability 1}
\end{aligned}$$

What Did I Learn?

- Value of Simple but Honest "Variation": This is one of the most reliable process in science. Too old but famous examples: Neyman Scott models and GARCH model. Nice company for the Superstar Model

- Nature of Difficulty: Things are often substantially harder than they look at first blush. He we took quite an obvious variation on the Preferential Attachment model, and we were led to quite different mathematics. Still the implications of this work do tell us something even about the PA model. One can pass from the SS model to the PA model by letting $p \to 0$.

- Using the SS Model:
  - The Superstar Model "looks like" perferential attachment with a twist — but the differences are HUGE!
  - It's easy to use since it is easy to reject. The plain vanilla SS Model is rigid. It if works it's great; if it doesn't you'll find out quickly.
  - This is the charm of a one-parameter model where the parameter is easy to estimate.
  - Still, if modeling needs demand changes, further parameters can be introduced.

Thank you!

Thanks Again to My Co-Authors on This Project

- Shankar Bhamidi
- Tauhid Zaman