The Clustering Coefficient of a Scale-Free Random Graph

Nicole Eggemann Steven Noble

Department of Mathematical Sciences
Brunel University

1/5/2008 / Newton Institute

Outline

Outline

Real World Networks

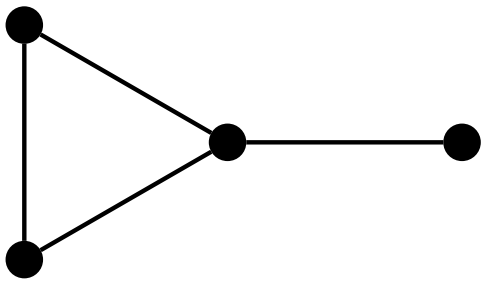
Some examples

Huge physics literature studying empirical properties of real world networks.

- World wide web. Directed graph - vertices are webpages; edges are hyperlinks.
- Internet. Undirected graph - vertices are computers / routers; edges are physical links.
- Maths collaboration network. Undirected graph - vertices are authors; edges indicate joint paper.
- Social contacts.
- Electricity power grid, phone calls, movie actor collaboration network, linguistics (synonyms), sheep movements. . . .

Key Properties of Real World Networks

- Degree sequence.
- Distribution of distances between pairs of vertices.
- Diameter.
- Connectedness and sizes of connected components.
- Clustering coefficient - 3$\times$ quotient of the number of triangles and the number of pairs of adjacent edges.



$$C(G) = 3/5.$$

The Internet

Empirical Properties of the Internet

- Govindan, Tangmunarunkit (2000): a subgraph of the internet with 150,000 vertices and 200,000 edges has $\Pr(\text{degree } d) \propto d^{-2.4}$.
- Govindan, Tangmunarunkit (2000): average path length $\simeq 11$.
- Yook et al. (2001), Pastor-Satorras et al. (2001): clustering coefficient between 0.18 and 0.3.

Summary of Empirical Results

Network	Vertices	Degree	γ	Path	Cluster
WWW	2×10^8	7.5	2.1	16	0.1078
Internet, router	150000	2.66	2.4	11	0.18
Movie actors	212250	28.78	2.3	4.54	0.79
Maths, coauthors	70975	3.9	2.5	9.5	0.59
Phone calls	53×10^6	3.16	2.1		
Synonyms	22311	13.48	2.8	4.5	0.7

(Adamic (1999), Aiello et al (2000), Barabási, Albert (1999), Broder (2000), Watts, Strogatz (1998).)

Outline

$G_{n,p}$
Definition and key properties

$G_{n,p}$ is a graph with vertices $\{1, \ldots, n\}$ such that each edge is present independently with probability p.

Theorem (Erdös+Rényi (1959–61))

- $p = c/n, 0 < c < 1$*: a.a.s. every connected component of has order* $O(\log n)$.
- $p = c/n, c > 1$*: a.a.s. G has a component with* $(\alpha(c) + o(1))n$ *vertices and other components have size* $O(\log n)$.
- $p = (\log n - \log\log n)/n$*: a.a.s. G is disconnected.*
- $p = (\log n + \log\log n)/n$*: a.a.s. G is connected.*

$G_{n,p}$
More properties

Let X_k be the number of vertices of degree k in $G_{n,c/n}$.

Theorem (Bollobás (1982))

For any $\epsilon > 0$, a.a.s.,

$$(1-\epsilon)\frac{c^k e^{-c}}{k!} \leq \frac{X_k}{n} \leq (1+\epsilon)\frac{c^k e^{-c}}{k!}.$$

Let $C(G)$ denote the clustering coefficient of G.

Folklore

$$\mathbf{E}\left[C(G_{n,p})\right] \to p, \qquad n\mathbf{E}\left[C(G_{n,c/n})\right] \to c.$$

Random regular graphs $G_{n,r}$

Theorem (Bollobás, de la Vega (1982))

Let $r \geq 3, \epsilon > 0$. Then a.a.s.

$$(1-\epsilon)\frac{\log n}{\log(r-1)} \leq \operatorname{diam}(G_{n,r}) \leq (1+\epsilon)\frac{\log n}{\log(r-1)}.$$

Bollobás–Riordan model

Definition $m = 1$

- Precise version of Barabási–Albert preferential attachment model. (Originally due to Yule (1925).)
- G_1^1 consists of a single vertex v_1 with a loop.
- G_1^t is formed from G_1^{t-1} by adding vertex v_t and an edge directed from v_t to v_s where s is chosen so that

$$\Pr(s = i) = \begin{cases} \frac{d_{t-1}(v_i)}{2t-1} & 1 \leq i \leq t-1, \\ \frac{1}{2t-1} & i = t. \end{cases}$$

- Gives a graph in which each vertex has out-degree 1.

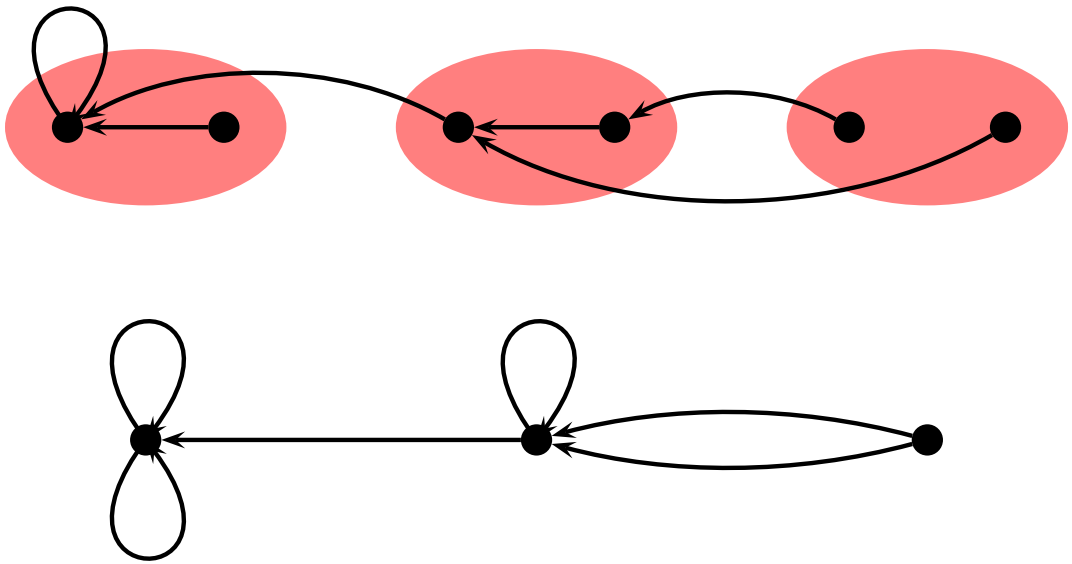
Bollobás-Riordan Model

Definition $m > 1$

- Run (G_1^t) on vertices $v_1', v_2', \ldots$.
- Form G_m^t from G_1^{mt} by merging $v_1', \ldots, v_m'$ to form v_1 etc.
- Gives a graph in which each vertex has out-degree m.
- Graph may have parallel edges.

Bollobás-Riordan Model

Example



Bollobás-Riordan Model

Diameter

Theorem (Bollobás, Riordan (2004))

Let $m \geq 2$ and $\epsilon > 0$. Then a.a.s G_m^n is connected and such that

$$(1-\epsilon)\frac{\log n}{\log\log n} \leq \operatorname{diam}(G_m^n) \leq (1+\epsilon)\frac{\log n}{\log\log n}.$$

Bollobás-Riordan Model

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády (2001))

Let $m \geq 1$ *and* $\epsilon > 0$. *Let*

$$\alpha_{m,d} = \frac{2m(m+1)}{(d+m)(d+m+1)(d+m+2)}.$$

Then if $0 \leq d \leq n^{1/15}$ *a.a.s.*

$$(1-\epsilon)\alpha_{m,d} \leq \frac{\#\textit{vertices degree } d}{n} \leq (1+\epsilon)\alpha_{m,d}.$$

Buckley-Osthus Model

Let a be a non-negative integer. Define $H^t_{a,1}$ exactly as G^t_1 except

$$\Pr(s = i) \propto \begin{cases} d_{t-1}(v_i) + a & 1 \leq i \leq t-1, \\ 1 + a & i = t. \end{cases}$$

Theorem (Buckley, Osthus (2004))

For all $d \leq n^{1/(100(a+2))}$, a.a.s, the proportion of vertices with degree d is $\theta(d^{-(3+a)})$.

Móri Model

Definition

- Two parameters m (out-degree of all vertices except first) and $\beta > 0$.
- G_1^2 consists of two vertices joined by a single edge.
- G_1^{t+1} is formed from G_1^t by adding vertex v_{t+1} and an edge directed from v_{t+1} to v_s where s is chosen so that for $1 \leq i \leq t$

$$\Pr(s = i) = \frac{d_t(v_i) + \beta}{(2 + \beta)t - 2}.$$

- Gives a connected loopless graph in which each vertex except the first has out-degree 1.
- Merge vertices as before to form $G_{m,\beta}^n$. Loops and parallel edges may now be formed.

Móri Model

Alternative formulation

$$\Pr(s = i) = \frac{d_t(v_i) + \beta}{(2+\beta)t - 2}$$
$$= A_{\beta,t}\frac{d_t(v_i)}{2t-2} + B_{\beta,t}\frac{1}{t}.$$

So either

- a random endpoint of an edge is chosen and copied (preferential attachment) or
- a random vertex is chosen and copied (uniform attachment).

Móri Model

Key properties

Let Δ_n denote the maximum degree of $G^n_{1,\beta}$.

Theorem (Mori 2005)

For every k, the sequence $\mathbf{E}\left[\left(\frac{\Delta_n}{n^{1/(2+\beta)}}\right)^k\right]$ *is bounded.*

Outline

Definition

Local clustering coefficient (Watts and Strogatz (1998))

The local clustering coefficient is given by

$$C_v(G) = \frac{\text{number of triangles including } v}{\binom{d(v)}{2}}.$$

(Global) clustering coefficient

$$C'(G) = \frac{1}{n}\sum_v C_v(G).$$

$$C(G) = \frac{\sum_v \binom{d(v)}{2} C_v(G)}{\sum_v \binom{d(v)}{2}} = \frac{3 \times \text{ total number of triangles}}{\sum_v \binom{d(v)}{2}}.$$

Main Result

Theorem (Eggemann,N)

For $\beta > 0$ we have

$$\mathbf{E}\left[C(G^n_{m,\beta})\right] \sim A_{m,\beta}\frac{\log n}{n},$$

where $A_{m,\beta}$ is a known constant.

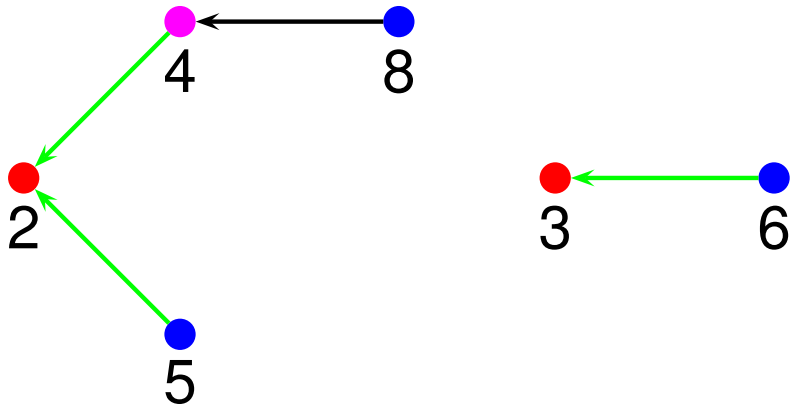
Finding the Clustering Coefficient

- Find the probability that Móri tree ($m = 1$ case) contains a given subgraph.
- Use this to find the expected number of triangles and the expectation of $\sum_v \binom{d(v)}{2}$.
- Show that $\sum_v \binom{d(v)}{2}$ is tightly concentrated around its expectation so that the expected clustering coefficient may be computed as three times the quotient of the two expectations.

Feasible Forests

A feasible forest S is a labelled directed forest in which each vertex has at most one out-going edge and each directed edge (v_i, v_j) has $i > j$.

- $c_S(i) = \#$ edges in $E(S)$ from $\{v_i, \ldots, v_n\}$ to $\{v_1, \ldots, v_{i-1}\}$.
- $V^- = \{v_i \in V(S) : \text{ indegree } > 0\}$.
- $V^+ = \{v_i \in V(S) : \text{ outdegree } > 0\}$.



Probability of Finding a Subgraph

Let S be a feasible forest. Then

Lemma (Bollobás, Riordan (2004))

$$\Pr(S \subset G^n_{1,0}) = \prod_{i \in V^-} d^{in}_S(i)! \prod_{i \in V^+} \frac{1}{2i-4} \prod_{i > 2 \notin V^+} \left(1 + \frac{c_S(i)}{2i-4}\right).$$

Lemma (Eggemann, N (2010))

$$\Pr(S \subset G^n_{1,\beta}) = \frac{\beta}{\beta + d^{in}_S(1)} \prod_{i \in V^-} \frac{\Gamma(1+\beta+d^{in}_S(i))}{\Gamma(1+\beta)}$$
$$\cdot \prod_{i \in V^+} \frac{1}{(2+\beta)(i-1)-2} \prod_{i \notin V^+} \left(1 + \frac{c_S(i)}{(2+\beta)(i-1)-2}\right).$$

A More Useful Expression

Lemma (Eggemann+N)

For $\beta > 0$ and S a feasible forest.

$$\Pr(S \subset G^n_{1,\beta})$$

$$= \frac{\beta}{d^{in}_S(1)+\beta} \prod_{i:v_i \in V^-} \frac{\Gamma(1+d^{in}_S(i)+\beta)}{\Gamma(1+\beta)}$$

$$\cdot \prod_{(v_i,v_j)\in E(S):i>j} \frac{1}{(2+\beta)(i^{1+\beta}j)^{1/(2+\beta)}} \cdot \exp\left(O\left(\sum_{j=2}^{k} \frac{c_S(s_j)^2}{j-1}\right)\right).$$

Calculating the Expected Number of Triangles

Proposition (Bollobás+Riordan)

For $\beta = 0$, the expected number of triangles in $G^n_{m,\beta}$ is

$$\frac{m(m-1)(m+1)}{48}(\log n)^3 + O((\log n)^2).$$

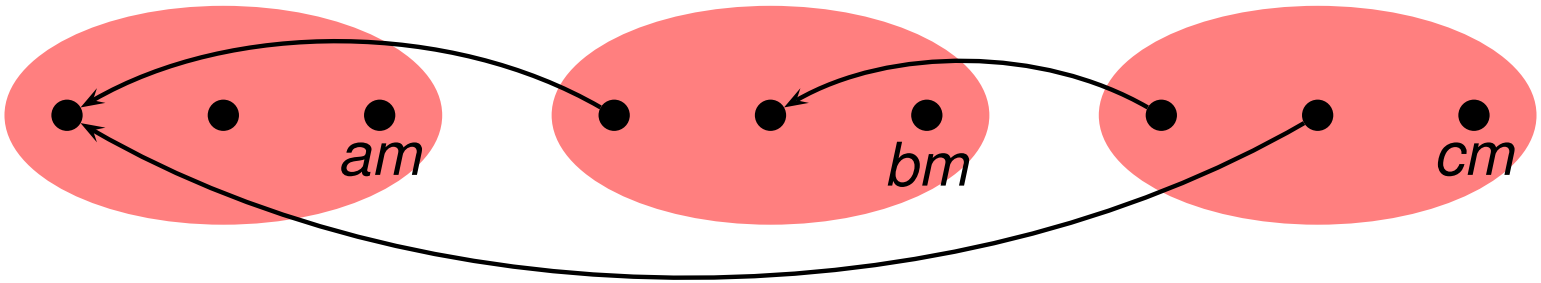
Proposition (Eggemann+N)

For $\beta > 0$, the expected number of triangles in $G^n_{m,\beta}$ is

$$\left(m(m-1)\frac{(1+\beta)^2}{\beta^2} + (m-1)^2\frac{(1+\beta)^3}{\beta^2(2+\beta)} \right) \log n + O(1).$$

How Triangles Look in the Tree

Type 1



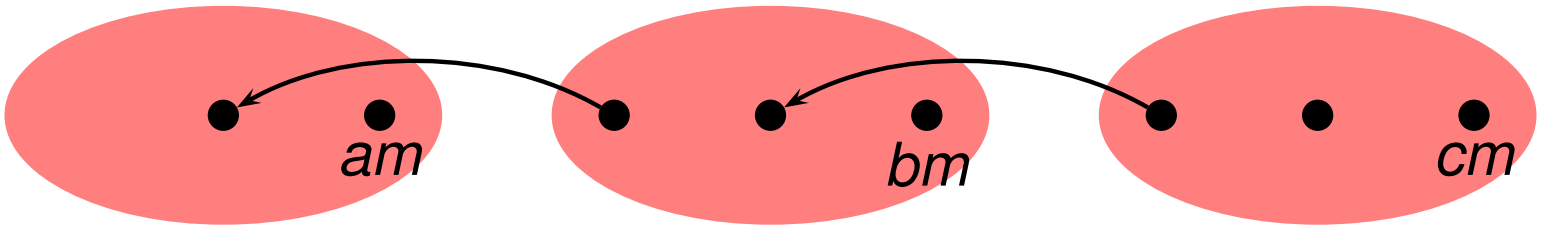
$$\Pr(S \subseteq G^{mn}_{1,\beta})$$

$$= \frac{\Gamma(3+\beta)\Gamma(2+\beta)}{(\Gamma(1+\beta))^2} \frac{1}{(2+\beta)^3} \frac{1}{m^3} \left(\frac{1}{a^2 b (bc^2)^{1+\beta}} \right)^{1/(2+\beta)} (1 + O(1/a)$$

Number of such triangles is $m^4(m-1)$.

How Triangles Look in the Tree

Type 2



$$\Pr(S \subseteq G_{1,\beta}^{mn})$$

$$= \frac{(\Gamma(2+\beta))^3}{(\Gamma(1+\beta))^3} \frac{1}{(2+\beta)^3} \frac{1}{m^3} \left(\frac{1}{a^2 b (bc^2)^{1+\beta}} \right)^{1/(2+\beta)} (1 + O(1/a)).$$

Number of such triangles is $m^4(m-1)^2$.

Total Number of Pairs of Adjacent Edges

Proposition (Bollobás+Riordan)

For $\beta = 0$, the expected number of pairs of adjacent edges in $G^n_{m,\beta}$ is

$$(1 + o(1))\frac{m(m+1)}{2} n \log n.$$

Proposition (Eggemann+N)

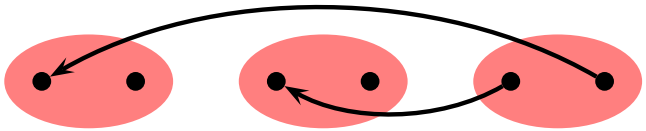
For $\beta > 0$, the expected number of pairs of adjacent edges in $G^n_{m,\beta}$ is

$$\left(\frac{2+5\beta}{2\beta}m^2 + \frac{2-\beta}{2\beta}m\right) n + O(n^{2/(2+\beta)})$$

Non-Degenerate Pairs of Adjacent Edges

$$\left(m^{\frac{2+\beta}{\beta}} + m(m-1)^{\frac{1+\beta}{\beta}}\right) n + O(n^{2/(2+\beta)})$$

$$m^2 n + O(n^{2/(2+\beta)})$$



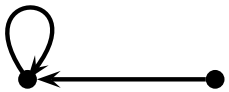
$$\binom{m}{2} n + O(n^{1/(2+\beta)})$$

Degenerate Pairs of Adjacent Edges

$O(\log n)$

$O(1)$

$O(\log n)$



$O(n^{1/(2+\beta)})$

Main Concentration Result

Theorem

For any $\epsilon > 0$, the number $\sum_v \binom{d_n(v)}{2}$ of pairs of adjacent edges in $G^n_{m,\beta}$ is concentrated about its expected value within $O(n^{(4+\beta)/(4+2\beta)+\epsilon})$.
More precisely, for any $\epsilon > 0$,

$$\Pr\left(\left|\sum_v \binom{d_n(v)}{2} - \mathbf{E}\left[\sum_v \binom{d_n(v)}{2}\right]\right| \geq n^{\frac{4+\beta}{4+2\beta}+\epsilon}\right) \longrightarrow 0$$

as $n \longrightarrow \infty$.

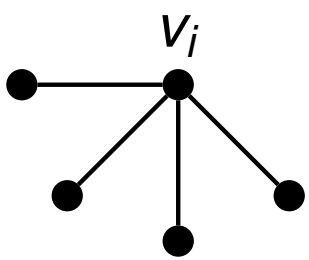
Martingale Result

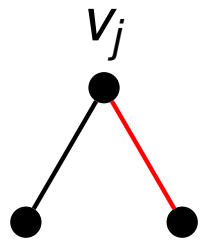
- Let $g(x; x_1, \ldots, x_{s-1}) = \mathbf{E}[F(X_1, \ldots, X_n) | x_1, \ldots, x_{s-1}, x]$.
- Let $\text{ran}(x_1, \ldots, x_{s-1}) = \sup_{x,y}\{|g(x) - g(y)|\}$.
- Let $R^2(x_1, \ldots, x_n) = \sum_{k=2}^{n} (\text{ran}(x_2, \ldots, x_{k-1}))^2$.
- Let $r^2 = \sup_{\mathbf{x} \in \Omega \setminus \mathcal{B}} R^2(\mathbf{x})$.

Theorem (Azuma, Hoeffding, McDiarmid)

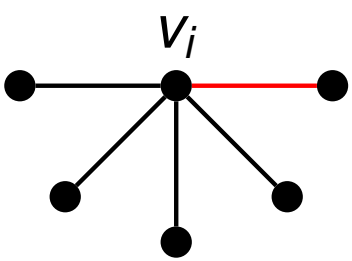
$$\Pr(|F(\mathbf{X}) - \mathbf{E}[F(\mathbf{X})]| \geq t) \leq 2(e^{-2t^2/r^2} + \Pr(\mathbf{X} \in \mathcal{B})).$$

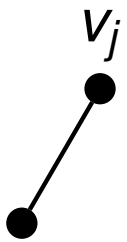
Basic Idea



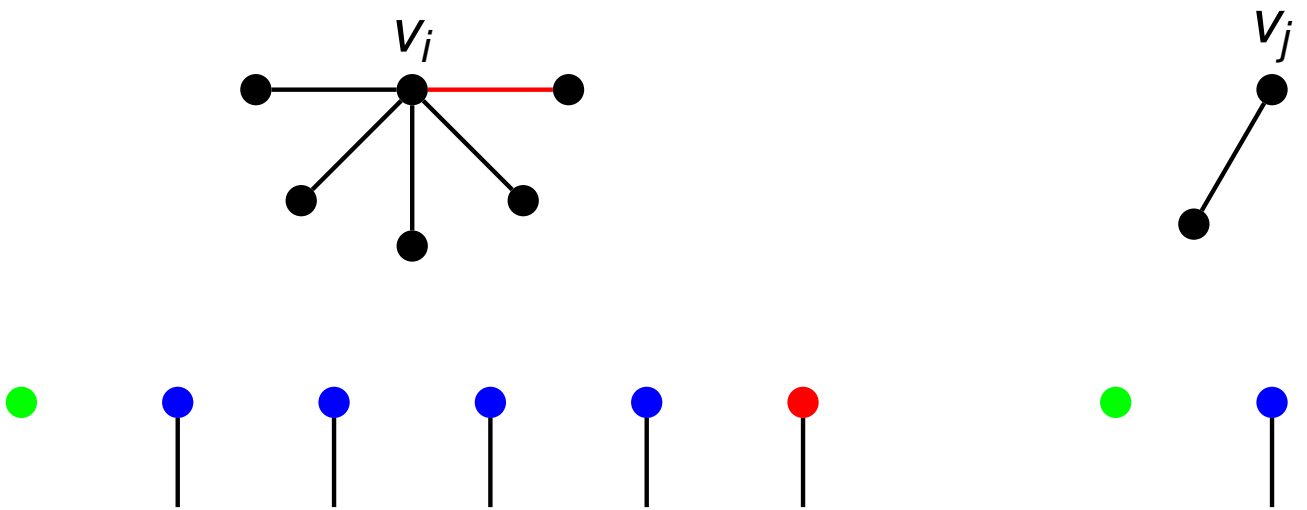


Basic Idea

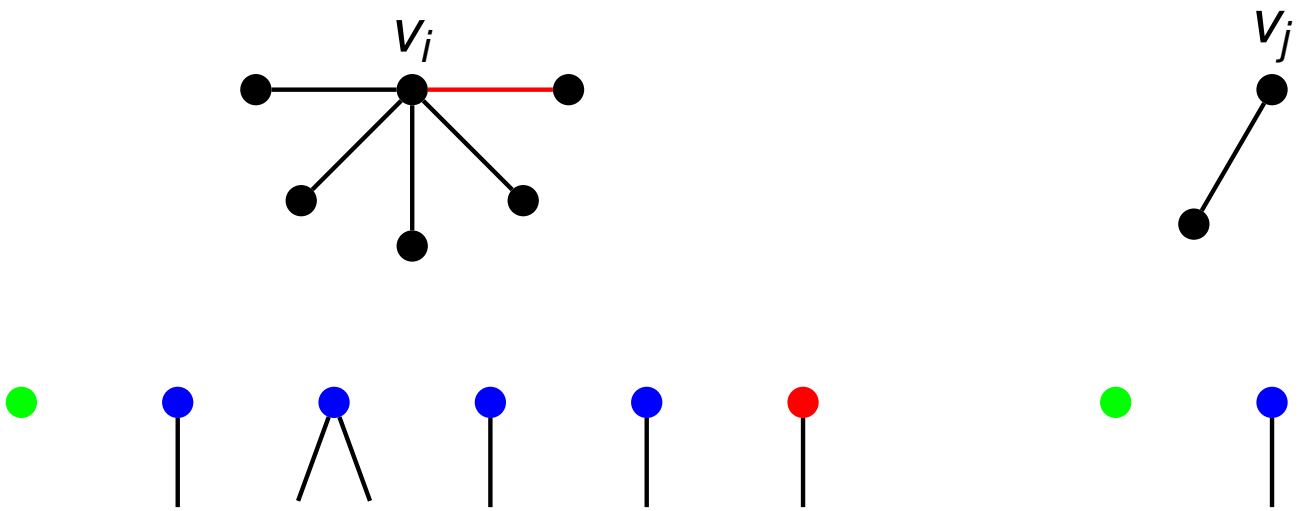




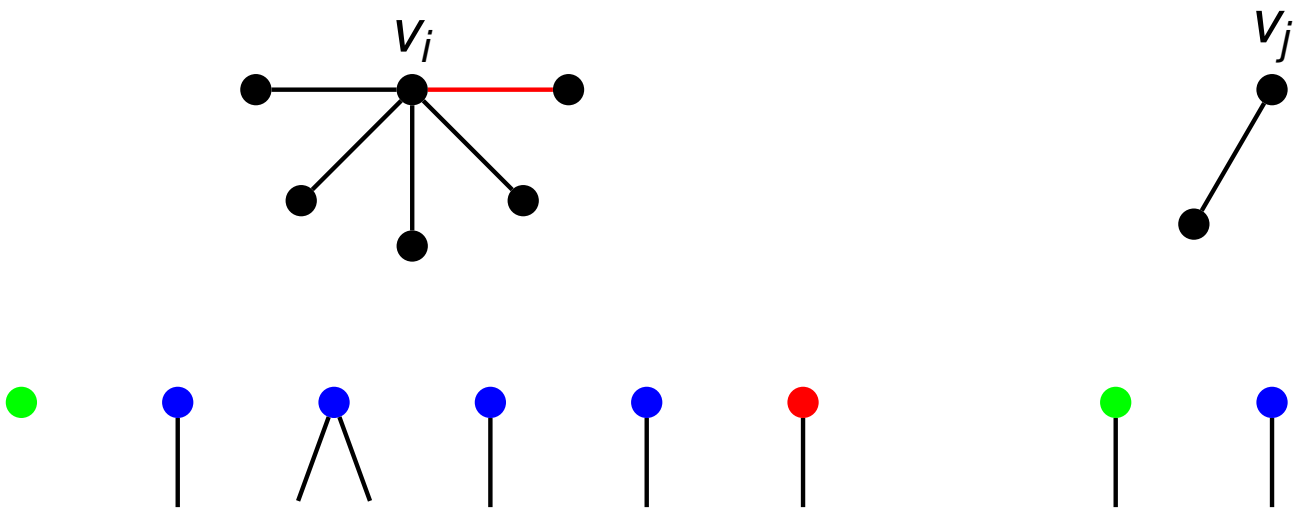
Basic Idea



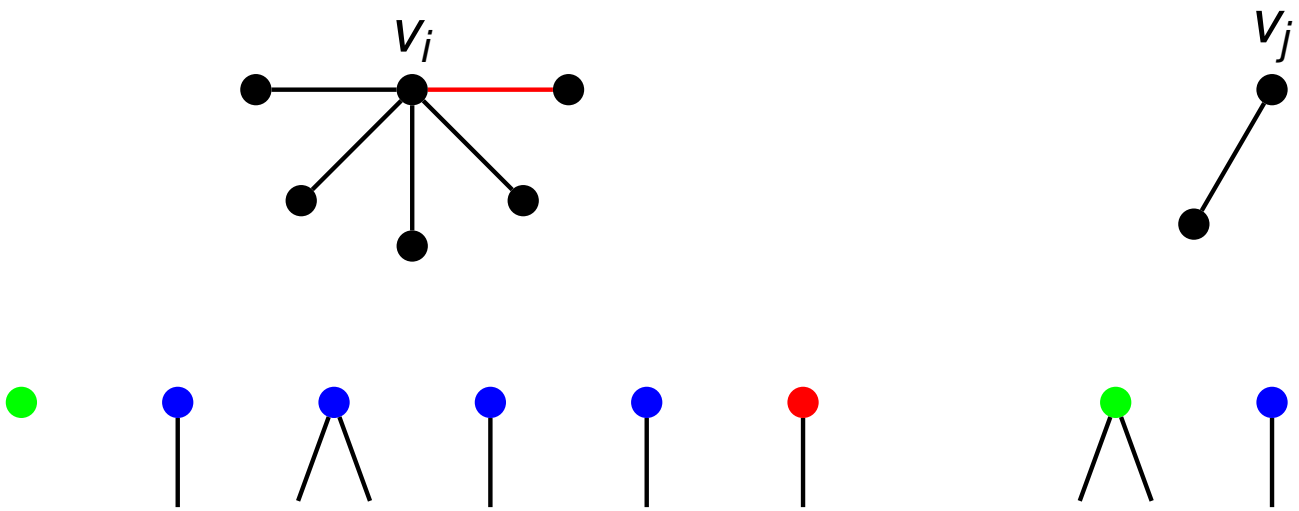
Basic Idea



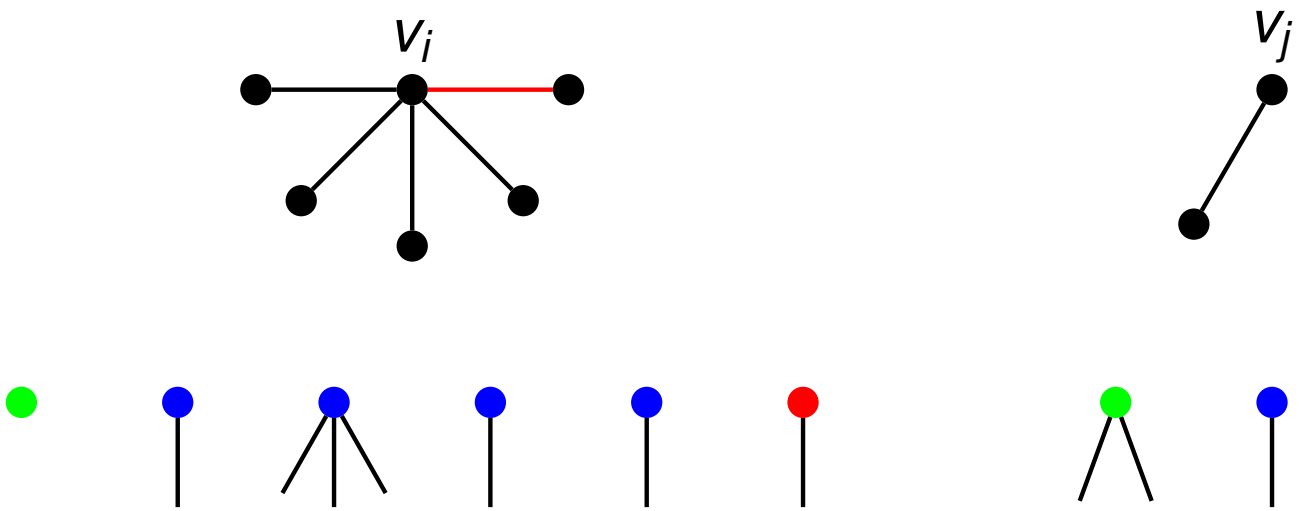
Basic Idea



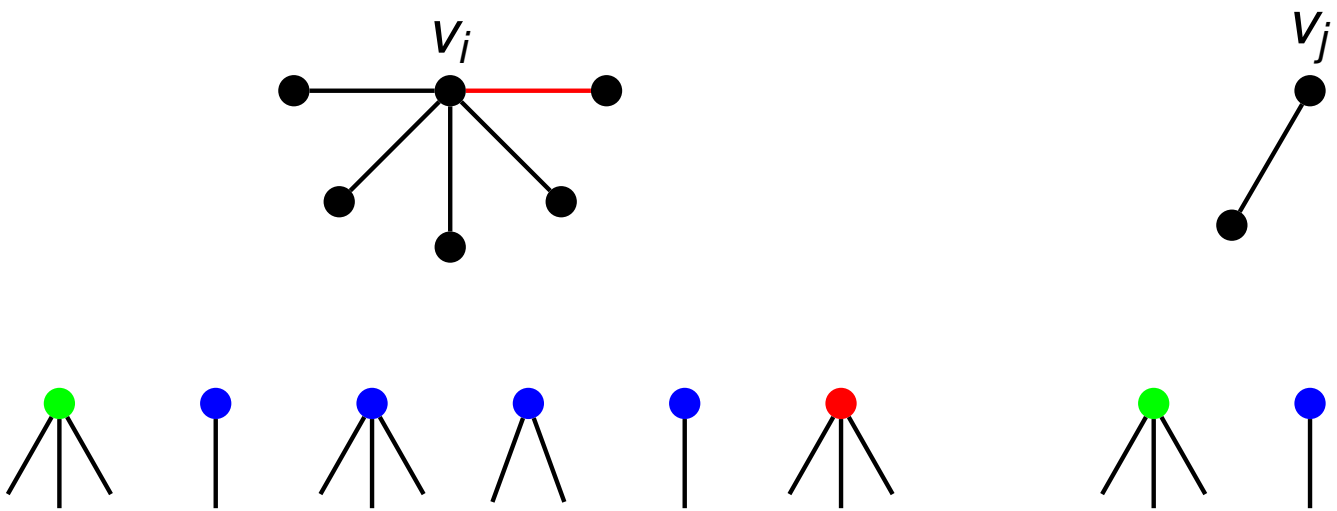
Basic Idea



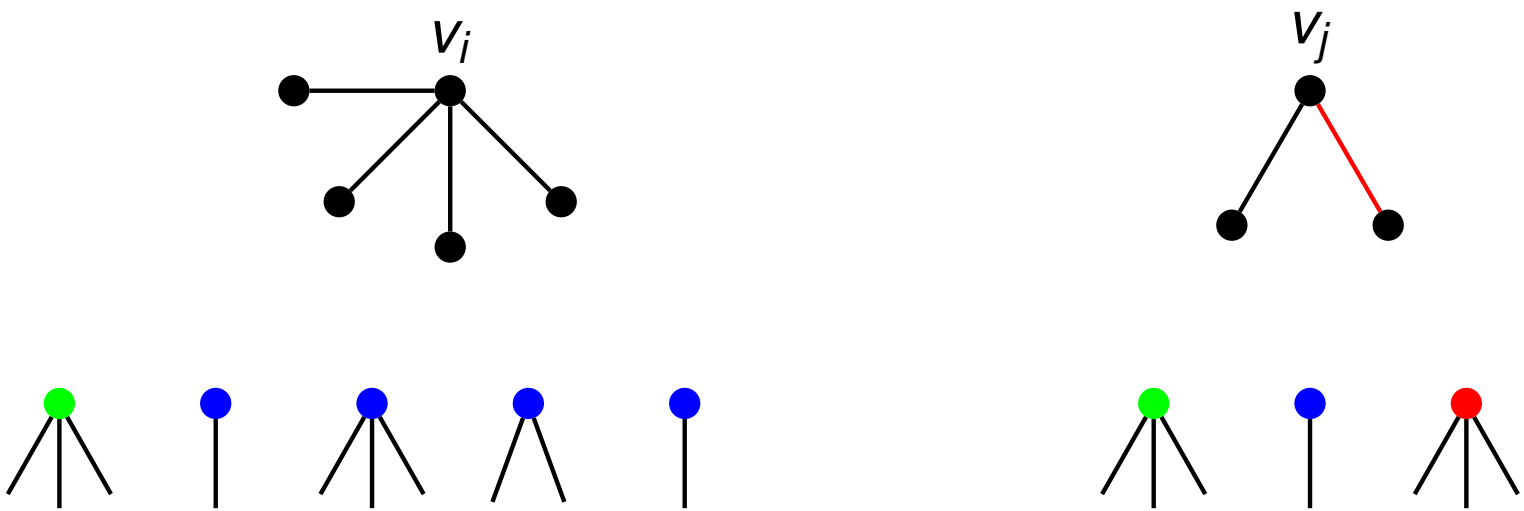
Basic Idea



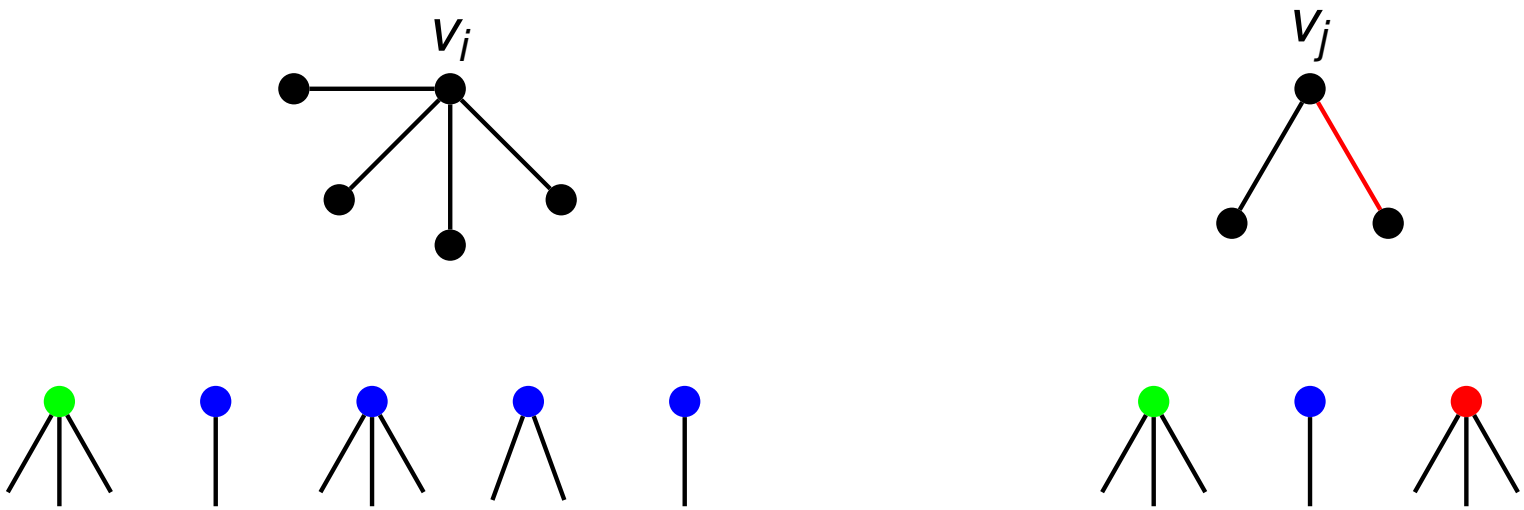
Basic Idea



Basic Idea



Basic Idea



$$|d_t(v_i) - d_t(v_j)|\mathbf{E}[|X||X'|].$$

Details of Proof

$$\text{ran} \leq \Delta_t (n/t)^{2/(2+\beta)}.$$

$$\mathcal{B} = \left\{ \mathbf{x} : \sum_{t=1}^{n} \left(\frac{\Delta_t}{t^{2/(2+\beta)}} \right)^2 > n^{\beta/(2+\beta)+\epsilon} \right\}$$

$$R^2 \leq n^{4/(2+\beta)} \sum_{t=1}^{n} \left(\frac{\Delta_t}{t^{2/(2+\beta)}} \right)^2 \leq n^{\frac{4+\beta}{2+\beta}+\epsilon}$$

$$\Pr\left(\left| \sum_v \binom{d_n(v)}{2} - \mathbf{E}\left[\sum_v \binom{d_n(v)}{2} \right] \right| \geq n^{\frac{4+\beta}{4+2\beta}+\epsilon} \right)$$
$$\leq 2\Pr(\mathcal{B}) + 2\exp(-2n^{\epsilon}).$$

Summary

- The expectation of the clustering coefficient of $G^n_{m,\beta}$ is asymptotically $A_{m,\beta} \log n/n$.
- There is a phase transition at $\beta = 0$.
- The BA model / Móri model does not display one of the key properties of webgraphs.

- Is the clustering coefficient tightly concentrated?