Small Subgraphs in Preferential Attachment Networks

Andrei Raigorodskii

Lomonosov Moscow State University,
Moscow Institute of Physics and Technology,
Yandex Division of Theoretical and Applied Research,
Moscow, Russia

The Fifth International Conference on Network Analysis, Nizhniy Novgorod, 18.05.2015–20.05.2015

The main objects

Real-world web-graph

$G = (V, E)$, where V —

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).
Sometimes multiple edges are identified. Sometimes multiple edges and even loops are allowed.

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).
Sometimes multiple edges are identified. Sometimes multiple edges and even loops are allowed.

Why do we need a model?

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).
Sometimes multiple edges are identified. Sometimes multiple edges and even loops are allowed.

Why do we need a model?
Many reasons!

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).
Sometimes multiple edges are identified. Sometimes multiple edges and even loops are allowed.

Why do we need a model?
Many reasons!

- Adjust algorithms;

The main objects

Real-world web-graph

$G = (V, E)$, where V —

- set of web-pages,
- set of web-sites,
- set of web-hosts,

and E — the set of all hyperlinks between the vertices (nodes).
Sometimes multiple edges are identified. Sometimes multiple edges and even loops are allowed.

Why do we need a model?
Many reasons!

- Adjust algorithms;
- Find unexpected structures (news, spam, etc.) using classifiers learnt on some features coming from models.

How to construct a model?

How to construct a model?

Statistics

First, find some statistical properties of web-graphs that would describe most accurately the real-world structures.

How to construct a model?

Statistics

First, find some statistical properties of web-graphs that would describe most accurately the real-world structures.

Probability Theory

Then, take a random element G which takes values in a set of graphs on n vertices and has such a distribution that w.h.p. (with high probability, i.e., with probability approaching 1 as $n \to \infty$) G has the same properties as the ones mentioned above.

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.
- Every two vertices in the giant component are connected by a path of short length (5–6, 15–20 depending on what we mean by web-graph): $\operatorname{diam} G \approx 6$ (the rule of 6 handshakes).

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.
- Every two vertices in the giant component are connected by a path of short length (5–6, 15–20 depending on what we mean by web-graph): $\operatorname{diam} G \approx 6$ (the rule of 6 handshakes).
- Web-graphs are robust when random vertices are destroyed (a giant component survives).

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.
- Every two vertices in the giant component are connected by a path of short length (5–6, 15–20 depending on what we mean by web-graph): $\operatorname{diam} G \approx 6$ (the rule of 6 handshakes).
- Web-graphs are robust when random vertices are destroyed (a giant component survives).
- Web-graphs are vulnerable to attacks onto hubs (many small components appear after a threshold is surpassed).

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.
- Every two vertices in the giant component are connected by a path of short length (5–6, 15–20 depending on what we mean by web-graph): $\operatorname{diam} G \approx 6$ (the rule of 6 handshakes).
- Web-graphs are robust when random vertices are destroyed (a giant component survives).
- Web-graphs are vulnerable to attacks onto hubs (many small components appear after a threshold is surpassed).
- The degree distribution is close to a power-law:

$$\frac{|\{v \in V : \deg v = d\}|}{n} \sim \frac{const}{d^{\gamma}},$$

where $\gamma \in (2, 3)$ depends on what we mean by web-graph.

Some properties

Barabási–Albert, Watts–Strogatz, Newman, and many others in 90s–00s.

- Web-graphs are *sparse*, i.e., their numbers of edges (links) are proportional to their numbers of vertices.
- Web-graphs have a unique "giant" connected component.
- Every two vertices in the giant component are connected by a path of short length (5–6, 15–20 depending on what we mean by web-graph): $\operatorname{diam} G \approx 6$ (the rule of 6 handshakes).
- Web-graphs are robust when random vertices are destroyed (a giant component survives).
- Web-graphs are vulnerable to attacks onto hubs (many small components appear after a threshold is surpassed).
- The degree distribution is close to a power-law:

$$\frac{|\{v \in V : \deg v = d\}|}{n} \sim \frac{const}{d^\gamma},$$

  where $\gamma \in (2, 3)$ depends on what we mean by web-graph.
- **High clustering**.

Clustering

Clustering

Let $G = (V, E)$, $v \in V$. Let N_v be the set of neighbours of v in G. Let $n_v = |N_v|$. If $n_v \geqslant 2$, then

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

Clustering

Let $G = (V, E)$, $v \in V$. Let N_v be the set of neighbours of v in G. Let $n_v = |N_v|$. If $n_v \geqslant 2$, then

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

Global clustering coefficient

The global clustering coefficient of G is

$$T(G) = \frac{\sum\limits_{v \in V} C_{n_v}^2 C_v}{\sum\limits_{v \in V} C_{n_v}^2}.$$

Clustering

Let $G = (V, E)$, $v \in V$. Let N_v be the set of neighbours of v in G. Let $n_v = |N_v|$. If $n_v \geqslant 2$, then

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

Global clustering coefficient

The global clustering coefficient of G is

$$T(G) = \frac{\sum\limits_{v \in V} C_{n_v}^2 C_v}{\sum\limits_{v \in V} C_{n_v}^2}.$$

Let $\sharp(H, G)$ be the number of copies of a graph H in a graph G. Then

$$T(G) = \frac{3\sharp(K_3, G)}{\sharp(P_2, G)},$$

where K_3 is a triangle and P_2 is a 2-path.

Clustering

Clustering

Average local clustering coefficient

The average local clustering coefficient of G is

$$C(G) = \frac{1}{|V|} \sum_{v \in V} C_v,$$

where, again,

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

Clustering

Average local clustering coefficient

The average local clustering coefficient of G is

$$C(G) = \frac{1}{|V|} \sum_{v \in V} C_v,$$

where, again,

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

The quantities $T(G)$ and $C(G)$ are quite different.

Clustering

Average local clustering coefficient

The average local clustering coefficient of G is

$$C(G) = \frac{1}{|V|} \sum_{v \in V} C_v,$$

where, again,

$$C_v = \frac{|\{\{x, y\} \in E : \; x, y \in N_v\}|}{C_{n_v}^2}.$$

The quantities $T(G)$ and $C(G)$ are quite different.

Let G be $K_{2,n-2}$ plus one edge between the vertices in the part of size 2. Then $C(G) \sim 1$, but $T(G) = \Theta\left(\frac{1}{n}\right)$.

Clustering

Average local clustering coefficient

The average local clustering coefficient of G is

$$C(G) = \frac{1}{|V|} \sum_{v \in V} C_v,$$

where, again,

$$C_v = \frac{|\{\{x, y\} \in E : \ x, y \in N_v\}|}{C_{n_v}^2}.$$

The quantities $T(G)$ and $C(G)$ are quite different.

Let G be $K_{2,n-2}$ plus one edge between the vertices in the part of size 2. Then $C(G) \sim 1$, but $T(G) = \Theta\left(\frac{1}{n}\right)$.

Very important! However, many inaccuracies in the literature.

Clustering: experiments and models

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$.

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations.

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations. This is wrong, provided we do not take into account multiple edges and loops (which is widely done).

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations. This is wrong, provided we do not take into account multiple edges and loops (which is widely done).

Theorem (Ostroumova, Samosvat)

If in a sequence $\{G_n\}$ of graphs, the degrees of the vertices follow a power law with exponent $\gamma \in (2, 3)$, then $T(G_n) \to 0$ as $n \to \infty$.

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations. This is wrong, provided we do not take into account multiple edges and loops (which is widely done).

Theorem (Ostroumova, Samosvat)

If in a sequence $\{G_n\}$ of graphs, the degrees of the vertices follow a power law with exponent $\gamma \in (2,3)$, then $T(G_n) \to 0$ as $n \to \infty$.

Newman might be right, provided we do take into account multiple edges and loops.

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations. This is wrong, provided we do not take into account multiple edges and loops (which is widely done).

Theorem (Ostroumova, Samosvat)

If in a sequence $\{G_n\}$ of graphs, the degrees of the vertices follow a power law with exponent $\gamma \in (2, 3)$, then $T(G_n) \to 0$ as $n \to \infty$.

Newman might be right, provided we do take into account multiple edges and loops.

Theorem (Ostroumova)

There exist sequences $\{G_n\}$ of multigraphs with loops, whose degrees of the vertices follow a power law with exponent $\gamma \in (2, 3)$ and, nevertheless, $T(G_n) \geqslant \text{const}$ as $n \to \infty$.

Clustering: experiments and models

Experimentally, $C(WWW)$ seems to be constant. However, no real data for $T(WWW)$. Newman asserts that $T(WWW)$ is constant as well without any explanations. This is wrong, provided we do not take into account multiple edges and loops (which is widely done).

Theorem (Ostroumova, Samosvat)

If in a sequence $\{G_n\}$ of graphs, the degrees of the vertices follow a power law with exponent $\gamma \in (2,3)$, then $T(G_n) \to 0$ as $n \to \infty$.

Newman might be right, provided we do take into account multiple edges and loops.

Theorem (Ostroumova)

There exist sequences $\{G_n\}$ of multigraphs with loops, whose degrees of the vertices follow a power law with exponent $\gamma \in (2,3)$ and, nevertheless, $T(G_n) \geqslant \text{const}$ as $n \to \infty$.

However, what is $T(G)$, if G has multiple edges and loops? Many different definitions, and Newman does not say a word about this subtlety!

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Case $m = 1$

G_1^1 — graph with one vertex v_1 and one loop.

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Case $m = 1$

G_1^1 — graph with one vertex v_1 and one loop.
Given G_1^{n-1} we can make G_1^n by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \dots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{G_1^{n-1}}(v_s)}{2n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{1}{2n-1} & s = n \end{cases}$$

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Case $m = 1$

G_1^1 — graph with one vertex v_1 and one loop.
Given G_1^{n-1} we can make G_1^n by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \ldots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{G_1^{n-1}}(v_s)}{2n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{1}{2n-1} & s = n \end{cases}$$

Preferential attachment suggested by Barabási and Albert in 1999.

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Case $m = 1$

G_1^1 — graph with one vertex v_1 and one loop.
Given G_1^{n-1} we can make G_1^n by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \ldots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{G_1^{n-1}}(v_s)}{2n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{1}{2n-1} & s = n \end{cases}$$

Preferential attachment suggested by Barabási and Albert in 1999.

Case $m > 1$

Given G_1^{mn} we can make G_m^n by gluing $\{v_1, \ldots, v_m\}$ into v_1' , $\{v_{m+1}, \ldots, v_{2m}\}$ into v_2', and so on.

Bollobás–Riordan model

Construct a random graph G_m^n with n vertices and mn edges, $m \in \mathbb{N}$.
Let $d_G(v)$ be the degree of a vertex v in a graph G.

Case $m = 1$

G_1^1 — graph with one vertex v_1 and one loop.
Given G_1^{n-1} we can make G_1^n by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \ldots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{G_1^{n-1}}(v_s)}{2n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{1}{2n-1} & s = n \end{cases}$$

Preferential attachment suggested by Barabási and Albert in 1999.

Case $m > 1$

Given G_1^{mn} we can make G_m^n by gluing $\{v_1, \ldots, v_m\}$ into v_1' , $\{v_{m+1}, \ldots, v_{2m}\}$ into v_2', and so on.

The random graph G_m^n is certainly sparse. What's about other properties?

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \ \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \ \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.
- Not too great, since the exponent in the power-law is a bit different from the experimental ones ($\gamma \in (2,3)$).

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.
- Not too great, since the exponent in the power-law is a bit different from the experimental ones ($\gamma \in (2, 3)$).
- Bad, since $d \leqslant n^{1/15}$, which is non-realistic.

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.
- Not too great, since the exponent in the power-law is a bit different from the experimental ones ($\gamma \in (2,3)$).
- Bad, since $d \leqslant n^{1/15}$, which is non-realistic.
- The last problem completely removed by Evgeniy Grechnikov: analog of B–R–S–T-theorem with an arbitrary d.

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.
- Not too great, since the exponent in the power-law is a bit different from the experimental ones ($\gamma \in (2,3)$).
- Bad, since $d \leqslant n^{1/15}$, which is non-realistic.
- The last problem completely removed by Evgeniy Grechnikov: analog of B–R–S–T-theorem with an arbitrary d.
- Tune the model somehow to get other exponents in the power-law?

Degree distribution

Theorem (Bollobás, Riordan, Spencer, Tusnády)

If $d \leqslant n^{1/15}$, then w.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(m)}{d^3}.$$

- Great, since we get a power-law.
- Not too great, since the exponent in the power-law is a bit different from the experimental ones ($\gamma \in (2, 3)$).
- Bad, since $d \leqslant n^{1/15}$, which is non-realistic.
- The last problem completely removed by Evgeniy Grechnikov: analog of B–R–S–T-theorem with an arbitrary d.
- Tune the model somehow to get other exponents in the power-law?
- Let's discuss clustering before.

Clustering: the B–R model

Clustering: the B–R model

Theorem (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Clustering: the B–R model

Theorem (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

- To calculate the value $\mathbf{E}(T(G_m^n))$, one needs to know the number of triangles and the number of 2-paths.

Clustering: the B–R model

Theorem (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

- To calculate the value $\mathbf{E}(T(G_m^n))$, one needs to know the number of triangles and the number of 2-paths.
- Recall that $\sharp(H, G)$ is the number of copies of a graph H in a graph G.

Clustering: the B–R model

Theorem (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

- To calculate the value $\mathbf{E}(T(G_m^n))$, one needs to know the number of triangles and the number of 2-paths.
- Recall that $\sharp(H, G)$ is the number of copies of a graph H in a graph G.
- A general and nice result was proved by Ryabchenko and Samosvat.

Clustering: the B–R model

Theorem (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

- To calculate the value $\mathbf{E}(T(G_m^n))$, one needs to know the number of triangles and the number of 2-paths.
- Recall that $\sharp(H, G)$ is the number of copies of a graph H in a graph G.
- A general and nice result was proved by Ryabchenko and Samosvat.

Theorem (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

Clustering: comments on the theorems for the B–R model

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

- Theorem 2 agrees with Theorem 1: $\mathbf{E}(\sharp(K_3, G_m^n)) \asymp \ln^3 n$, $\mathbf{E}(\sharp(P_2, G_m^n)) \asymp n \ln n$.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

- Theorem 2 agrees with Theorem 1: $\mathbf{E}(\sharp(K_3, G_m^n)) \asymp \ln^3 n$, $\mathbf{E}(\sharp(P_2, G_m^n)) \asymp n \ln n$.
- By Theorem 2 the number of K_4 etc. is asymptotically constant: bad.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

- Theorem 2 agrees with Theorem 1: $\mathbf{E}(\sharp(K_3, G_m^n)) \asymp \ln^3 n$, $\mathbf{E}(\sharp(P_2, G_m^n)) \asymp n \ln n$.
- By Theorem 2 the number of K_4 etc. is asymptotically constant: bad.
- Unfortunately, $C(G_m^n)$ also approaches 0: even worse.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

- Theorem 2 agrees with Theorem 1: $\mathbf{E}(\sharp(K_3, G_m^n)) \asymp \ln^3 n$, $\mathbf{E}(\sharp(P_2, G_m^n)) \asymp n \ln n$.
- By Theorem 2 the number of K_4 etc. is asymptotically constant: bad.
- Unfortunately, $C(G_m^n)$ also approaches 0: even worse.
- And, once again, $\gamma = 3$, not $\gamma \in (2, 3)$.

Clustering: comments on the theorems for the B–R model

Theorem 1 (Bollobás, Riordan)

The expected value of $T(G_m^n)$ tends to 0 as $n \to \infty$: $\mathbf{E}(T(G_m^n)) \asymp \frac{\ln^2 n}{n}$.

Theorem 2 (Ryabchenko, Samosvat)

For any H, $\mathbf{E}(\sharp(H, G_m^n)) \asymp n^{\sharp(d_i=0)} \cdot (\sqrt{n})^{\sharp(d_i=1)} \cdot (\ln n)^{\sharp(d_i=2)}$, where $\sharp(d_i = k)$ is the number of vertices of degree k in H.

- Theorem 2 agrees with Theorem 1: $\mathbf{E}(\sharp(K_3, G_m^n)) \asymp \ln^3 n$, $\mathbf{E}(\sharp(P_2, G_m^n)) \asymp n \ln n$.
- By Theorem 2 the number of K_4 etc. is asymptotically constant: bad.
- Unfortunately, $C(G_m^n)$ also approaches 0: even worse.
- And, once again, $\gamma = 3$, not $\gamma \in (2, 3)$.
- So let's tune the model and try to calculate again the number of **small subgraphs**!

Buckley–Osthus model

Buckley–Osthus model

Which problems we had in the model of Bollobás–Riordan? Non-realistic exponent in the power-law, non-realistic clustering. Can solve the first problem! The following model is very close to the first one, but it has one important new parameter $a > 0$ called *initial attractiveness* of a vertex.

Buckley–Osthus model

Which problems we had in the model of Bollobás–Riordan? Non-realistic exponent in the power-law, non-realistic clustering. Can solve the first problem! The following model is very close to the first one, but it has one important new parameter $a > 0$ called *initial attractiveness* of a vertex.

Case $m = 1$

$H_{a,1}^1$ — graph with one vertex v_1 and one loop.

Buckley–Osthus model

Which problems we had in the model of Bollobás–Riordan? Non-realistic exponent in the power-law, non-realistic clustering. Can solve the first problem! The following model is very close to the first one, but it has one important new parameter $a > 0$ called *initial attractiveness* of a vertex.

Case $m = 1$

$H_{a,1}^1$ — graph with one vertex v_1 and one loop.
Given $H_{a,1}^{n-1}$ we can make $H_{a,1}^n$ by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \dots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{H_{a,1}^{n-1}}(v_s)+a-1}{(a+1)n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{a}{(a+1)n-1} & s = n \end{cases}$$

Buckley–Osthus model

Which problems we had in the model of Bollobás–Riordan? Non-realistic exponent in the power-law, non-realistic clustering. Can solve the first problem! The following model is very close to the first one, but it has one important new parameter $a > 0$ called *initial attractiveness* of a vertex.

Case $m = 1$

$H_{a,1}^1$ — graph with one vertex v_1 and one loop.
Given $H_{a,1}^{n-1}$ we can make $H_{a,1}^n$ by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \ldots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{H_{a,1}^{n-1}}(v_s)+a-1}{(a+1)n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{a}{(a+1)n-1} & s = n \end{cases}$$

For $a = 1$, we get the model of Bollobás–Riordan.

Buckley–Osthus model

Which problems we had in the model of Bollobás–Riordan? Non-realistic exponent in the power-law, non-realistic clustering. Can solve the first problem! The following model is very close to the first one, but it has one important new parameter $a > 0$ called *initial attractiveness* of a vertex.

Case $m = 1$

$H^1_{a,1}$ — graph with one vertex v_1 and one loop.
Given $H^{n-1}_{a,1}$ we can make $H^n_{a,1}$ by adding vertex v_n and an edge from it to a vertex v_i, picked from $\{v_1, \ldots, v_n\}$ with probability

$$\mathbf{P}(i = s) = \begin{cases} \frac{d_{H^{n-1}_{a,1}}(v_s)+a-1}{(a+1)n-1} & 1 \leqslant s \leqslant n-1 \\ \frac{a}{(a+1)n-1} & s = n \end{cases}$$

For $a = 1$, we get the model of Bollobás–Riordan.

Case $m > 1$

Given $H^{mn}_{a,1}$ we can make $H^n_{a,m}$ by gluing $\{v_1, \ldots, v_m\}$ into v'_1 , $\{v_{m+1}, \ldots, v_{2m}\}$ into v'_2, and so on.

Buckley–Osthus model: degree distribution

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H_{a,m}^n : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H^n_{a,m} : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

- Great, since now we can tune the model to get the expected exponent.

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H^n_{a,m} : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

- Great, since now we can tune the model to get the expected exponent.
- Bad, since $d \leqslant n^{1/(100(a+1))}$.

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H^n_{a,m} : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

- Great, since now we can tune the model to get the expected exponent.
- Bad, since $d \leqslant n^{1/(100(a+1))}$.
- Completely removed by Grechnikov.

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H^n_{a,m} : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

- Great, since now we can tune the model to get the expected exponent.
- Bad, since $d \leqslant n^{1/(100(a+1))}$.
- Completely removed by Grechnikov.

Assertion (Grechnikov, Zhukovskii, Vinogradov, Ostroumova, Pritykin, Gusev, Raigorodskii)

If the reality agrees with a Buckley–Osthus model, then most likely $a \approx 0.27$.

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

If $d \leqslant n^{1/(100(a+1))}$, then w.h.p.

$$\frac{|\{v \in H_{a,m}^n : \deg v = d\}|}{n} \sim \frac{const(a,m)}{d^{a+2}}.$$

- Great, since now we can tune the model to get the expected exponent.
- Bad, since $d \leqslant n^{1/(100(a+1))}$.
- Completely removed by Grechnikov.

Assertion (Grechnikov, Zhukovskii, Vinogradov, Ostroumova, Pritykin, Gusev, Raigorodskii)

If the reality agrees with a Buckley–Osthus model, then most likely $a \approx 0.27$.

What's about clustering and, more generally, small subgraphs?

Buckley–Osthus model: clustering

Buckley–Osthus model: clustering

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(K_3, H^n_{a,m})) \asymp \ln n$ as $n \to \infty$.

Buckley–Osthus model: clustering

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(K_3, H^n_{a,m})) \asymp \ln n$ as $n \to \infty$.

It's remarkable that for $a = 1$ (i.e., for the B–R model) we had $\ln^2 n$ instead of $\ln n$.

Buckley–Osthus model: clustering

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(K_3, H^n_{a,m})) \asymp \ln n$ as $n \to \infty$.

It's remarkable that for $a = 1$ (i.e., for the B–R model) we had $\ln^2 n$ instead of $\ln n$.

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(P_2, H^n_{a,m})) \asymp n$ as $n \to \infty$.

Buckley–Osthus model: clustering

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(K_3, H^n_{a,m})) \asymp \ln n$ as $n \to \infty$.

It's remarkable that for $a = 1$ (i.e., for the B–R model) we had $\ln^2 n$ instead of $\ln n$.

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(\sharp(P_2, H^n_{a,m})) \asymp n$ as $n \to \infty$.

Theorem (Eggemann, Noble)

If $a > 1$, then $\mathbf{E}(T(H^n_{a,m})) \asymp \frac{\ln n}{n}$ as $n \to \infty$.

Buckley–Osthus model: small subgraphs

Buckley–Osthus model: small subgraphs

Very recently Tilga has proved far-reaching generalizations and refinements of the theorems by Bollobás–Riordan, Ryabchenko–Samosvat, and Eggemann–Noble.

Buckley–Osthus model: small subgraphs

Very recently Tilga has proved far-reaching generalizations and refinements of the theorems by Bollobás–Riordan, Ryabchenko–Samosvat, and Eggemann–Noble.

Theorem (Tilga)

For any $a > 0$ and any fixed graph F, the order of magnitude of $\mathbf{E}(\sharp(F, H^n_{a,m}))$ is found.

Buckley–Osthus model: small subgraphs

Very recently Tilga has proved far-reaching generalizations and refinements of the theorems by Bollobás–Riordan, Ryabchenko–Samosvat, and Eggemann–Noble.

Theorem (Tilga)

For any $a > 0$ and any fixed graph F, the order of magnitude of $\mathbf{E}(\sharp(F, H^n_{a,m}))$ is found.

The exact statement is quite cumbersome involving many parameters and cases. So we just give several most important and short enough corollaries.

Buckley–Osthus model: paths and cliques

Buckley–Osthus model: paths and cliques

Theorem (Tilga)

Let $m \geqslant 2$ and $a < 1$, $\lambda = \frac{1}{a+1}$. Let P_l be a path of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(P_l, H^n_{a,m})\right) = \begin{cases} n^{(2\lambda-1)k+1} \cdot \Theta(m^l) & \text{for } \; l = 2k, \\ n^{(2\lambda-1)k+1} \cdot \ln n \cdot \Theta(m^l) & \text{for } \; l = 2k+1. \end{cases}$$

Buckley–Osthus model: paths and cliques

Theorem (Tilga)

Let $m \geqslant 2$ and $a < 1$, $\lambda = \frac{1}{a+1}$. Let P_l be a path of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(P_l, H_{a,m}^n)\right) = \begin{cases} n^{(2\lambda-1)k+1} \cdot \Theta(m^l) & \text{for} \quad l = 2k, \\ n^{(2\lambda-1)k+1} \cdot \ln n \cdot \Theta(m^l) & \text{for} \quad l = 2k+1. \end{cases}$$

Theorem (Tilga)

Let K_k be a clique of size k, where $4 \leqslant k \leqslant m+1$. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(K_k, H_{a,m}^n)\right) = \begin{cases} n^{1+(\lambda-1)(k-1)} \cdot \Theta(m^{C_k^2}) & \text{for} \quad a < \frac{1}{k-2}, \\ \ln n \cdot \Theta(m^{C_k^2}) & \text{for} \quad a = \frac{1}{k-2}, \\ \Theta(m^{C_k^2}) & \text{for} \quad a > \frac{1}{k-2}. \end{cases}$$

Buckley–Osthus model: paths and cliques

Theorem (Tilga)

Let $m \geqslant 2$ and $a < 1$, $\lambda = \frac{1}{a+1}$. Let P_l be a path of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(P_l, H^n_{a,m})\right) = \begin{cases} n^{(2\lambda-1)k+1} \cdot \Theta(m^l) & \text{for} \quad l = 2k, \\ n^{(2\lambda-1)k+1} \cdot \ln n \cdot \Theta(m^l) & \text{for} \quad l = 2k+1. \end{cases}$$

Theorem (Tilga)

Let K_k be a clique of size k, where $4 \leqslant k \leqslant m+1$. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(K_k, H^n_{a,m})\right) = \begin{cases} n^{1+(\lambda-1)(k-1)} \cdot \Theta(m^{C_k^2}) & \text{for} \quad a < \frac{1}{k-2}, \\ \ln n \cdot \Theta(m^{C_k^2}) & \text{for} \quad a = \frac{1}{k-2}, \\ \Theta(m^{C_k^2}) & \text{for} \quad a > \frac{1}{k-2}. \end{cases}$$

For example, if $a = \frac{1}{3}$ (close to 0.27), then the number of K_5 is about $\log n$, and the number of K_4 is about $\sqrt[4]{n}$. Much more realistic than in the B–R model!

Buckley–Osthus model: cycles and bicliques

Buckley–Osthus model: cycles and bicliques

Theorem (Tilga)

Let C_l be a cycle of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(C_l, H^n_{a,m})\right) = \begin{cases} n^{(2\lambda-1)k} \cdot \Theta(m^l) & \text{for } \; l = 2k, \\ n^{(2\lambda-1)k} \cdot \ln n \cdot \Theta(m^l) & \text{for } \; l = 2k+1. \end{cases}$$

Buckley–Osthus model: cycles and bicliques

Theorem (Tilga)

Let C_l be a cycle of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(C_l, H^n_{a,m})\right) = \begin{cases} n^{(2\lambda-1)k} \cdot \Theta(m^l) & \text{for } \; l = 2k, \\ n^{(2\lambda-1)k} \cdot \ln n \cdot \Theta(m^l) & \text{for } \; l = 2k+1. \end{cases}$$

Theorem (Tilga)

Let $K_{k,l}$ be a biclique with $2 \leqslant l \leqslant \min\{k, m\}$. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(K_{k,l}, H^n_{a,m})\right) = \begin{cases} n^{k(1+(\lambda-1)l)} \cdot \Theta(m^{kl}) & \text{for } \; a < \frac{1}{l-1}, \\ (\ln n)^k \cdot \Theta(m^{kl}) & \text{for } \; a = \frac{1}{l-1}, \\ \Theta(m^{kl}) & \text{for } \; a > \frac{1}{l-1}. \end{cases}$$

Buckley–Osthus model: cycles and bicliques

Theorem (Tilga)

Let C_l be a cycle of length l. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(C_l, H^n_{a,m})\right) = \begin{cases} n^{(2\lambda-1)k} \cdot \Theta(m^l) & \text{for} \quad l = 2k, \\ n^{(2\lambda-1)k} \cdot \ln n \cdot \Theta(m^l) & \text{for} \quad l = 2k+1. \end{cases}$$

Theorem (Tilga)

Let $K_{k,l}$ be a biclique with $2 \leqslant l \leqslant \min\{k, m\}$. Then for $n \to \infty$,

$$\mathbf{E}\left(\sharp(K_{k,l}, H^n_{a,m})\right) = \begin{cases} n^{k(1+(\lambda-1)l)} \cdot \Theta(m^{kl}) & \text{for} \quad a < \frac{1}{l-1}, \\ (\ln n)^k \cdot \Theta(m^{kl}) & \text{for} \quad a = \frac{1}{l-1}, \\ \Theta(m^{kl}) & \text{for} \quad a > \frac{1}{l-1}. \end{cases}$$

The number of bicliques shows how many communities are formed. For example, if $a = \frac{1}{3}$ (close to 0.27), then there are many $K_{k,4}$ and a lot of $K_{k,3}$, which was impossible in the B–R model (there are no vertices of degree < 3 in such graphs).