Web-graph Models and Applications

Andrei Raigorodskii

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

The Yandex School of Data Analysis International Conference, 27 September – 2 October 2013

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?

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

How to construct a model?

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

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).
- Many triangles — high clusternig.

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).
- Many triangles — high clusternig.
- 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.

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 $\{v_1, \ldots, v_n\}$ with probability

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

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

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

Preferential attachment!

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!

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!

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?

Diameter and giant components

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Great, since for real values of n, we get $\frac{\ln n}{\ln \ln n} \in [5, 15]$.

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Great, since for real values of n, we get $\frac{\ln n}{\ln \ln n} \in [5, 15]$.

Theorem (Bollobás, Riordan)

If $p \in (0, 1)$ and we make a random subgraph $G_{m,p}^n$ of the graph G_m^n by deleting its vertices independently each with probability p, then w.h.p. $G_{m,p}^n$ contains a connected component of size $\asymp n$.

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Great, since for real values of n, we get $\frac{\ln n}{\ln \ln n} \in [5, 15]$.

Theorem (Bollobás, Riordan)

If $p \in (0,1)$ and we make a random subgraph $G_{m,p}^n$ of the graph G_m^n by deleting its vertices independently each with probability p, then w.h.p. $G_{m,p}^n$ contains a connected component of size $\asymp n$.

Great, since we have the robustness property.

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Great, since for real values of n, we get $\frac{\ln n}{\ln \ln n} \in [5, 15]$.

Theorem (Bollobás, Riordan)

If $p \in (0, 1)$ and we make a random subgraph $G_{m,p}^n$ of the graph G_m^n by deleting its vertices independently each with probability p, then w.h.p. $G_{m,p}^n$ contains a connected component of size $\asymp n$.

Great, since we have the robustness property.

Theorem (Bollobás, Riordan)

If $c \in (0, 1)$ and we make a random subgraph $G_{m,c}^n$ of the graph G_m^n by deleting its $[cn]$ first vertices, then for $c \leqslant (m-1)/(m+1)$, w.h.p. $G_{m,c}^n$ contains a connected component of size $\asymp n$, and for $c > (m-1)/(m+1)$, w.h.p. all the connected components of $G_{m,c}^n$ are of size $o(n)$.

Diameter and giant components

Theorem (Bollobás, Riordan)

If $m \geqslant 2$, then w.h.p. $\operatorname{diam} G_m^n \sim \frac{\ln n}{\ln \ln n}$.

Great, since for real values of n, we get $\frac{\ln n}{\ln \ln n} \in [5, 15]$.

Theorem (Bollobás, Riordan)

If $p \in (0,1)$ and we make a random subgraph $G_{m,p}^n$ of the graph G_m^n by deleting its vertices independently each with probability p, then w.h.p. $G_{m,p}^n$ contains a connected component of size $\asymp n$.

Great, since we have the robustness property.

Theorem (Bollobás, Riordan)

If $c \in (0,1)$ and we make a random subgraph $G_{m,c}^n$ of the graph G_m^n by deleting its $[cn]$ first vertices, then for $c \leqslant (m-1)/(m+1)$, w.h.p. $G_{m,c}^n$ contains a connected component of size $\asymp n$, and for $c > (m-1)/(m+1)$, w.h.p. all the connected components of $G_{m,c}^n$ are of size $o(n)$.

Great, since we have the vulnerability to attacks on the hubs.

Degree distribution

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 recently 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 recently 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?

Clustering

Clustering

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

Clustering

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

Clustering coefficient

The global clustering coefficient of G is

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

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

Clustering

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

Clustering coefficient

The global clustering coefficient of G is

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

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

Roughly speaking, $T(G)$ is the probability that two neighbours of a vertex of G are themselves joined by an edge.

Clustering

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

Clustering coefficient

The global clustering coefficient of G is

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

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

Roughly speaking, $T(G)$ is the probability that two neighbours of a vertex of G are themselves joined by an edge.

There are some other definitions of clustering coefficients.

Clustering

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

Clustering coefficient

The global clustering coefficient of G is

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

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

Roughly speaking, $T(G)$ is the probability that two neighbours of a vertex of G are themselves joined by an edge.

There are some other definitions of clustering coefficients.

Anyway, experimentally, clustering coefficients are constant.

Clustering

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

Clustering coefficient

The global clustering coefficient of G is

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

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

Roughly speaking, $T(G)$ is the probability that two neighbours of a vertex of G are themselves joined by an edge.

There are some other definitions of clustering coefficients.

Anyway, experimentally, clustering coefficients are constant.

Theorem (Bollobás, Riordan)

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

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 $m = 1$

$H^1_{a,1}$ — 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 $\{v_1, \ldots, v_n\}$ with probability

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

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.

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.

Case $m > 1$

Given $H_{a,1}^{mn}$ we can make $H_{a,m}^n$ 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.

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

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

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 $d \leqslant n^{1/(100(a+1))}$.

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 $d \leqslant n^{1/(100(a+1))}$.

Recently completely removed by Grechnikov.

Buckley–Osthus model: degree distribution

Theorem (Buckley, Osthus)

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

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

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

Bad, since $d \leqslant n^{1/(100(a+1))}$.

Recently completely removed by Grechnikov.

However, still problems with clustering!

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 $d \leqslant n^{1/(100(a+1))}$.

Recently completely removed by Grechnikov.

However, still problems with clustering!

Many more further great features of the model instead!

Buckley–Osthus model: second degrees of vertices

Buckley–Osthus model: second degrees of vertices

Let

$$d_2(t) = |\{\{i,j\}: \ i \neq t, j \neq t, \{i,t\} \in E(H_{a,1}^n), \{i,j\} \in E(H_{a,1}^n)\}|.$$

Buckley–Osthus model: second degrees of vertices

Let

$$d_2(t) = |\{\{i,j\}: \ i \neq t, j \neq t, \{i,t\} \in E(H_{a,1}^n), \{i,j\} \in E(H_{a,1}^n)\}|.$$

So we calculate the number of edges of $H_{a,1}^n$ that are joined with a neighbour of a given vertex t.

Buckley–Osthus model: second degrees of vertices

Let

$$d_2(t) = |\{\{i,j\} : \ i \neq t, j \neq t, \{i,t\} \in E(H_{a,1}^n), \{i,j\} \in E(H_{a,1}^n)\}|.$$

So we calculate the number of edges of $H_{a,1}^n$ that are joined with a neighbour of a given vertex t.

Theorem (Ostroumova, Grechnikov, Kupavskiy, Tetali)

W.h.p.

$$\frac{|\{i = 1, \ldots, n : \ d_2(i) = d\}|}{n} \sim \frac{const(a)}{d^{a+1}}.$$

Buckley–Osthus model: second degrees of vertices

Let

$$d_2(t) = |\{\{i,j\}: \ i \neq t, j \neq t, \{i,t\} \in E(H_{a,1}^n), \{i,j\} \in E(H_{a,1}^n)\}|.$$

So we calculate the number of edges of $H_{a,1}^n$ that are joined with a neighbour of a given vertex t.

Theorem (Ostroumova, Grechnikov, Kupavskiy, Tetali)

W.h.p.

$$\frac{|\{i = 1, \ldots, n: \ d_2(i) = d\}|}{n} \sim \frac{const(a)}{d^{a+1}}.$$

Fits quite well to the real data.

Buckley–Osthus model: the number of edges between vertices of given degrees

Buckley–Osthus model: the number of edges between vertices of given degrees

Let $X_n(d_1, d_2)$ be the total number of edges between vertices of given degrees.

Buckley–Osthus model: the number of edges between vertices of given degrees

Let $X_n(d_1, d_2)$ be the total number of edges between vertices of given degrees.

Theorem (Grechnikov)

W.h.p.

$$\frac{X_n(d_1, d_2)}{n} \sim c(a, m) \left(\frac{(d_1 + d_2)^{1-a}}{d_1^2 d_2^2} \right).$$

Buckley–Osthus model: the number of edges between vertices of given degrees

Let $X_n(d_1, d_2)$ be the total number of edges between vertices of given degrees.

Theorem (Grechnikov)

W.h.p.

$$\frac{X_n(d_1, d_2)}{n} \sim c(a, m) \left(\frac{(d_1 + d_2)^{1-a}}{d_1^2 d_2^2} \right).$$

How this is important, we will see soon.

Buckley–Osthus model: “power and glory”

Theorem (Grechnikov)

Let $d_1 \geqslant m$ and $d_2 \geqslant m$. Let $X = X_n(d_1, d_2)$. There exists a function $c_X(d_1, d_2)$ such that

$$\mathbf{E}X_n(d_1, d_2) = c_X(d_1, d_2)n + O_{a,m}(1)$$

and

$$c_X(d_1, d_2) = \frac{\Gamma(d_1 - m + ma)\Gamma(d_2 - m + ma)}{\Gamma(d_1 - m + ma + 2)\Gamma(d_2 - m + ma + 2)} \times$$
$$\times \frac{\Gamma(d_1 + d_2 - 2m + 2ma + 3)}{\Gamma(d_1 + d_2 - 2m + 2ma + a + 2)} ma(a+1) \frac{\Gamma(ma + a + 1)}{\Gamma(ma)} \times$$
$$\times \left(1 + \theta(d_1, d_2) \frac{(d_1 - m + ma + 1)(d_2 - m + ma + 1)}{(d_1 + d_2 - 2m + 2ma + 1)(d_1 + d_2 - 2m + 2ma + 2)}\right),$$

where

$$-4 + \frac{2}{1 + ma} \leqslant \theta(d_1, d_2) \leqslant a \frac{\Gamma(ma + 1)\Gamma(2ma + a + 3)}{\Gamma(2ma + 2)\Gamma(ma + a + 2)}.$$

Bollobás–Riordan model: "power and glory"

Theorem (Grechnikov)

If $d_1 < k$, $d_2 < k$ or $d_1 = d_2 = k$, then $X = 0$. If $d_1 \geqslant k, d_2 \geqslant k$ and $d_1 + d_2 \geqslant 2k + 1$, then the expected value of X is

$$\mathbf{E}X = \frac{k(k+1)}{d_1(d_1+1)d_2(d_2+1)} \left(1 - \frac{C_{2k+2}^{k+1} C_{d_1+d_2-2k}^{d_1-k}}{C_{d_1+d_2+2}^{d_1+1}}\right)(2kt+1) -$$

$$- \sum_{n=1}^{k} \frac{C_{d_1+d_2-2n}^{d_1-n}}{d_1 d_2 C_{d_1+d_2}^{d_1}} \left(\frac{(2n)!}{n!(n+1)!} \frac{k+1}{2k} + [n=k] \frac{(2k)!}{2(k-1)!^2}\right) -$$

$$- [d_1 = k] \frac{(k-1)(k+1)}{2kd_2(d_2+1)} - [d_2 = k] \frac{(k-1)(k+1)}{2kd_1(d_1+1)} + O_{k,d_1,d_2}\left(\frac{1}{t}\right).$$

Buckley–Osthus model: one more evidence of its quality

Buckley–Osthus model: one more evidence of its quality

Assume that the web-graph is governed by the Buckley–Osthus model. What is the most likely parameter a?

Buckley–Osthus model: one more evidence of its quality

Assume that the web-graph is governed by the Buckley–Osthus model. What is the most likely parameter a?

We may try to find an optimal a by comparing the reality with the fact that the number of vertices of degree d is close to d^{-2-a} (Grechnikov).

Buckley–Osthus model: one more evidence of its quality

Assume that the web-graph is governed by the Buckley–Osthus model. What is the most likely parameter a?

We may try to find an optimal a by comparing the reality with the fact that the number of vertices of degree d is close to d^{-2-a} (Grechnikov).

We may try to find an optimal a by comparing the reality with the fact that the number of edges between vertices of degree d_1 and d_2 is close to $(d_1 + d_2)^{1-a} d_1^{-2} d_2^{-2}$ (Grechnikov).

Buckley–Osthus model: one more evidence of its quality

Assume that the web-graph is governed by the Buckley–Osthus model. What is the most likely parameter a?

We may try to find an optimal a by comparing the reality with the fact that the number of vertices of degree d is close to d^{-2-a} (Grechnikov).

We may try to find an optimal a by comparing the reality with the fact that the number of edges between vertices of degree d_1 and d_2 is close to $(d_1 + d_2)^{1-a} d_1^{-2} d_2^{-2}$ (Grechnikov).

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

In both cases, the optimum is at the same $a \approx 0.27$.

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

An algorithm

- Calculate the total degrees of all the vertices of H (in the complete web-graph).

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

An algorithm

- Calculate the total degrees of all the vertices of H (in the complete web-graph).
- For each pair of vertices of H calculate the expected number of edges between them using Step 1 and Grechnikov's results.

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

An algorithm

- Calculate the total degrees of all the vertices of H (in the complete web-graph).
- For each pair of vertices of H calculate the expected number of edges between them using Step 1 and Grechnikov's results.
- Sum all the values found at Step 2.

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

An algorithm

- Calculate the total degrees of all the vertices of H (in the complete web-graph).
- For each pair of vertices of H calculate the expected number of edges between them using Step 1 and Grechnikov's results.
- Sum all the values found at Step 2.
- Compare the result of Step 3 with the real number of edges in H.

Buckley–Osthus model: an application

We have seen that the model fits quite well the reality. How could we apply it?

Assume that a subgraph H of the real web-graph has been found by an algorithm. How could we check automatically whether this graph is "expected" or it probably represents an "unnatural" structure like a spam construction or an "explosion" (say, important news)?

An algorithm

- Calculate the total degrees of all the vertices of H (in the complete web-graph).
- For each pair of vertices of H calculate the expected number of edges between them using Step 1 and Grechnikov's results.
- Sum all the values found at Step 2.
- Compare the result of Step 3 with the real number of edges in H.

The difference between the real and the expected values can be used as a feature.

A new general class of models

A new general class of models

Buckley–Osthus is not good for clustering. Can one do anything?

A new general class of models

Buckley–Osthus is not good for clustering. Can one do anything?

Many multiparametric models.

A new general class of models

Buckley–Osthus is not good for clustering. Can one do anything?

Many multiparametric models.

A break-through is due to Ryabchenko, Samosvat, Ostroumova.

A new general class of models

Buckley–Osthus is not good for clustering. Can one do anything?

Many multiparametric models.

A break-through is due to Ryabchenko, Samosvat, Ostroumova.

The PA-class

Let G_m^n $(n \geqslant n_0)$ be a graph with n vertices $\{1, \ldots, n\}$ and mn edges obtained as a result of the following random graph process. We start at the time n_0 from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges. On the $(n+1)$-th step $(n \geqslant n_0)$, we make the graph G_m^{n+1} from G_m^n by adding a new vertex $n+1$ and m edges connecting this vertex to some m vertices from the set $\{1, \ldots, n, n+1\}$. Denote by d_v^n the degree of a vertex v in G_m^n. Assume that for some constants A and B the following conditions are satisfied:

A new general class of models: continuation

The PA-class conditions

$$\mathbf{P}\left(d_v^{n+1} = d_v^n \mid G_m^n\right) = 1 - A\frac{d_v^n}{n} - B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\ , \qquad (1)$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + 1 \mid G_m^n\right) = A\frac{d_v^n}{n} + B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\ , \qquad (2)$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + j \mid G_m^n\right) = O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 2 \leqslant j \leqslant m,\ 1 \leqslant v \leqslant n\ , \qquad (3)$$

$$\mathbf{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right),\ 1 \leqslant j \leqslant m\ . \qquad (4)$$

A new general class of models: continuation

The PA-class conditions

$$\mathbf{P}\left(d_v^{n+1} = d_v^n \mid G_m^n\right) = 1 - A\frac{d_v^n}{n} - B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right), \ 1 \leqslant v \leqslant n \ , \tag{1}$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + 1 \mid G_m^n\right) = A\frac{d_v^n}{n} + B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right), \ 1 \leqslant v \leqslant n \ , \tag{2}$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + j \mid G_m^n\right) = O\left(\frac{\left(d_v^n\right)^2}{n^2}\right), \ 2 \leqslant j \leqslant m, \ 1 \leqslant v \leqslant n \ , \tag{3}$$

$$\mathbf{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right), \ 1 \leqslant j \leqslant m \ . \tag{4}$$

For $A = 1/2$, $B = 0$, we get Bollobás–Riordan.

A new general class of models: continuation

The PA-class conditions

$$\mathbf{P}\left(d_v^{n+1} = d_v^n \mid G_m^n\right) = 1 - A\frac{d_v^n}{n} - B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\,, \quad (1)$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + 1 \mid G_m^n\right) = A\frac{d_v^n}{n} + B\frac{1}{n} + O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 1 \leqslant v \leqslant n\,, \quad (2)$$

$$\mathbf{P}\left(d_v^{n+1} = d_v^n + j \mid G_m^n\right) = O\left(\frac{\left(d_v^n\right)^2}{n^2}\right),\ 2 \leqslant j \leqslant m,\ 1 \leqslant v \leqslant n\,, \quad (3)$$

$$\mathbf{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right),\ 1 \leqslant j \leqslant m\,. \quad (4)$$

For $A = 1/2$, $B = 0$, we get Bollobás–Riordan.

For $A = 1/(2+a)$, $B = ma/(2+a)$, we get Buckley–Osthus.

A new general class of models: results

A new general class of models: results

Theorem (Ostroumova, Ryabchenko, Samosvat)

W.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(A, B, m)}{d^{1+1/A}}.$$

A new general class of models: results

Theorem (Ostroumova, Ryabchenko, Samosvat)

W.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(A,B,m)}{d^{1+1/A}}.$$

Theorem (Ostroumova, Ryabchenko, Samosvat)

- If $2A < 1$ then w.h.p. $T(n) \sim c(A,B,m)$,
- If $2A = 1$ then w.h.p. $T(n) \sim \frac{c'(A,B,m)}{\ln n}$,
- If $2A > 1$ then for any $\varepsilon > 0$ w.h.p. $n^{1-2A-\varepsilon} \leqslant T(n) \leqslant n^{1-2A+\varepsilon}$.

A new general class of models: results

Theorem (Ostroumova, Ryabchenko, Samosvat)

W.h.p.

$$\frac{|\{v \in G_m^n : \deg v = d\}|}{n} \sim \frac{const(A, B, m)}{d^{1+1/A}}.$$

Theorem (Ostroumova, Ryabchenko, Samosvat)

- If $2A < 1$ then w.h.p. $T(n) \sim c(A, B, m)$,
- If $2A = 1$ then w.h.p. $T(n) \sim \frac{c'(A,B,m)}{\ln n}$,
- If $2A > 1$ then for any $\varepsilon > 0$ w.h.p. $n^{1-2A-\varepsilon} \leqslant T(n) \leqslant n^{1-2A+\varepsilon}$.

Great, since in the first case, we have a constant clustering together with power-law!