About the local clustering coefficient in preferential attachment graphs

Alexander Krot, Liudmila Ostroumova (Prokhorenkova)

Moscow Institute of Physics and Technology
Moscow State University
Yandex

May, 2015

Plan

1. Models based on the preferential attachment
2. Theoretical analysis of the general case and the local clustering coefficient
3. Conclusion

Degree distribution

Real-world networks often have the power law degree distribution:

$$\frac{\#\{v : \deg(v) = d\}}{n} \approx \frac{c}{d^\gamma},$$

where $2 < \gamma < 3$.

Clustering coefficient

Global clustering coefficient of a graph G:

$$C_1(n) = \frac{3\#(\text{triangles in } G)}{\#(\text{pairs of adjacent edges in } G)}.$$

Clustering coefficient

Global clustering coefficient of a graph G:

$$C_1(n) = \frac{3\#(\text{triangles in } G)}{\#(\text{pairs of adjacent edges in } G)}.$$

Average local clustering coefficient

- T^i is the number of edges between the neighbors of a vertex i
- P_2^i is the number of pairs of neighbors
- $C(i) = \frac{T^i}{P_2^i}$ is the local clustering coefficient for a vertex i
- $C_2(n) = \frac{1}{n}\sum_{i=1}^{n} C(i)$ – average local clustering coefficient

Preferential attachment

Idea of preferential attachment [Barabási, Albert]:

- Start with a small graph
- At every step we add new vertex with m edges
- The probability that a new vertex will be connected to a vertex i is proportional to the degree of i

PA-class of models

- Start from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges

PA-class of models

- Start from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges
- We make G_m^{n+1} from G_m^n by adding a new vertex $n+1$ with m edges

PA-class of models

- Start from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges
- We make G_m^{n+1} from G_m^n by adding a new vertex $n+1$ with m edges
- **PA-condition**: the probability that the degree of a vertex i increases by one equals

$$A\frac{deg(i)}{n} + B\frac{1}{n} + O\left(\frac{(deg(i))^2}{n^2}\right)$$

PA-class of models

- Start from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges
- We make G_m^{n+1} from G_m^n by adding a new vertex $n+1$ with m edges
- **PA-condition**: the probability that the degree of a vertex i increases by one equals

$$A\frac{deg(i)}{n} + B\frac{1}{n} + O\left(\frac{(deg(i))^2}{n^2}\right)$$

- The probability of adding a multiple edge is $O\left(\frac{(deg(i))^2}{n^2}\right)$

PA-class of models

- Start from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges
- We make G_m^{n+1} from G_m^n by adding a new vertex $n+1$ with m edges
- **PA-condition**: the probability that the degree of a vertex i increases by one equals

$$A\frac{deg(i)}{n} + B\frac{1}{n} + O\left(\frac{(deg(i))^2}{n^2}\right)$$

- The probability of adding a multiple edge is $O\left(\frac{(deg(i))^2}{n^2}\right)$
- $2mA + B = m$, $0 \le A \le 1$

T-subclass

Triangles property:

The probability that the degree of two vertices i and j increases by one equals

$$e_{ij} \frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right)$$

Here e_{ij} is the number of edges between vertices i and j in G_m^n and D is a positive constant.

Bollobás–Riordan, Buckley–Osthus, Móri, etc.

- Fix some positive number a – "initial attractiveness". (Bollobás–Riordan model: $a = 1$).
- Start with a graph with one vertex and m loops.
- At n-th step add one vertex with m edges.
- We add m edges one by one. The probability to add an edge $n \to i$ at each step is proportional to $indeg(i) + am$.

Bollobás–Riordan, Buckley–Osthus, Móri, etc.

- Fix some positive number a – "initial attractiveness". (Bollobás–Riordan model: $a = 1$).
- Start with a graph with one vertex and m loops.
- At n-th step add one vertex with m edges.
- We add m edges one by one. The probability to add an edge $n \to i$ at each step is proportional to $indeg(i) + am$.

- **Outdegree:** m
- **Triangles property:** $D = 0$
- **PA-condition:** $A = \frac{1}{1+a}$
- **Degree distribution:** Power law with $\gamma = 2 + a$
- **Global clustering:** $\frac{(\log n)^2}{n}$ $(a = 1)$, $\frac{\log n}{n}$ $(a > 1)$

Degree distribution

Let $N_n(d)$ be the number of vertices with degree d in G_m^n.

Degree distribution

Let $N_n(d)$ be the number of vertices with degree d in G_m^n.

Expectation

For every $d \geq m$ we have

$$\mathrm{E}N_n(d) = c(m,d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right),$$

where

$$c(m,d) = \frac{\Gamma\left(d+\frac{B}{A}\right)\Gamma\left(m+\frac{B+1}{A}\right)}{A\,\Gamma\left(d+\frac{B+A+1}{A}\right)\Gamma\left(m+\frac{B}{A}\right)} \sim \frac{\Gamma\left(m+\frac{B+1}{A}\right)d^{-1-\frac{1}{A}}}{A\,\Gamma\left(m+\frac{B}{A}\right)}$$

and $\Gamma(x)$ is the gamma function.

Degree distribution

Concentration

For every $d = d(n)$ we have

$$\mathsf{P}\left(|N_n(d) - \mathrm{E}N_n(d)| \geq d\sqrt{n}\log n\right) = O\left(n^{-\log n}\right).$$

Therefore, for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : |N_n(d) - \mathrm{E}N_n(d)| \geq \varphi(n)\,\mathrm{E}N_n(d)\right) = 0\,.$$

Global clustering

Let $P_2(n)$ be the number of all path of length 2 in G_m^n.

$P_2(n)$

(1) If $2A < 1$, then **whp** $P_2(n) \sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) \frac{n}{1-2A}$.

(2) If $2A = 1$, then **whp** $P_2(n) \propto n \log(n)$.

(3) If $2A > 1$, then **whp** $P_2(n) \propto n^{2A}$.

Global clustering

Let $P_2(n)$ be the number of all path of length 2 in G_m^n.

$P_2(n)$

(1) If $2A < 1$, then **whp** $P_2(n) \sim \left(2m(A+B) + \frac{m(m-1)}{2}\right) \frac{n}{1-2A}$.

(2) If $2A = 1$, then **whp** $P_2(n) \propto n \log(n)$.

(3) If $2A > 1$, then **whp** $P_2(n) \propto n^{2A}$.

Triangles

Whp the number of triangles $T(n) \sim D\, n$.

Global clustering

Global clustering

(1) If $2A < 1$ then **whp** $C_1(n) \sim \frac{3(1-2A)D}{\left(2m(A+B)+\frac{m(m-1)}{2}\right)}$.

(2) If $2A = 1$ then **whp** $C_1(n) \propto (\log n)^{-1}$.

(2) If $2A > 1$ then **whp** $C_1(n) \propto n^{1-2A}$.

Average local clustering

Average local clustering

Whp

$$C_2(n) \geq \frac{1}{n} \sum_{i:\deg(i)=m} C(i) \geq \frac{2cD}{m(m+1)}\,.$$

Local clustering: definition

Let $T_n(d)$ be the number of triangles on the vertices of degree d in G_m^n.

Let $N_n(d)$ be the number of vertices of degree d in G_m^n.

Local clustering coefficient over vertices of degree d**:**

$$C(d) = \frac{T_n(d)}{N_n(d)\frac{d(d-1)}{2}}$$

Local clustering: number of triangles

Let $T_n(d)$ be the number of triangles on the vertices of degree d in G_m^n. Then

Expectation

If $2A < 1$ then

$$\mathrm{E}T_n(d) = K(d)\left(n + \theta\left(C \cdot d^{2+\frac{1}{A}}\right)\right),$$

If $2A = 1$ then

$$\mathrm{E}T_n(d) = K(d)\left(n + \theta\left(C \cdot d^{2+\frac{1}{A}} \cdot \log(n)\right)\right),$$

If $2A > 1$ then

$$\mathrm{E}T_n(d) = K(d)\left(n + \theta\left(C \cdot d^{2+\frac{1}{A}} \cdot n^{2A-1}\right)\right),$$

where

$$K(d) = \left[\frac{\sum_{i=1}^{d} \frac{D(i-1)}{m[A(i-1)+B]}}{\sum_{i=1}^{m} \frac{(i-1)}{m[A(i-1)+B]}}\right] \cdot c(m,d) d \to \infty \frac{D}{A \cdot \sum_{i=1}^{m} \frac{(i-1)}{[A(i-1)+B]}} \cdot \frac{\Gamma\left(m + \frac{B+1}{A}\right)}{A\Gamma\left(m + \frac{B}{A}\right)} \cdot d^{-\frac{1}{A}}$$

and $\theta(X)$ is some function such that $|\theta(X)| < X$.

Idea of the proof

- The following recurrent formula holds:

$$\mathrm{E}T_{n+1}(d) = \mathrm{E}T_n(d) - \mathrm{E}T_n(d) \cdot \left[\frac{Ad+B}{n} + O\left(\frac{d^2}{n^2}\right) \right] +$$

$$+\mathrm{E}T_n(d-1) \cdot \left[\frac{A(d-1)+B}{n} + O\left(\frac{(d-1)^2}{n^2}\right) \right] +$$

$$+ \sum_{\substack{j: j \text{ is a neighbor} \\ \text{of a vertex of degree } d}} (d-1) N_n(d-1) \left[\frac{D}{mn} + O\left(\frac{(d-1)\cdot d_j}{n^2}\right) \right]$$

Idea of the proof

- The following recurrent formula holds:

$$\mathrm{E}T_{n+1}(d) = \mathrm{E}T_n(d) - \mathrm{E}T_n(d) \cdot \left[\frac{Ad+B}{n} + O\left(\frac{d^2}{n^2}\right)\right] +$$

$$+\mathrm{E}T_n(d-1) \cdot \left[\frac{A(d-1)+B}{n} + O\left(\frac{(d-1)^2}{n^2}\right)\right] +$$

$$+ \sum_{\substack{j: j \text{ is a neighbor} \\ \text{of a vertex of degree } d}} (d-1)N_n(d-1)\left[\frac{D}{mn} + O\left(\frac{(d-1)\cdot d_j}{n^2}\right)\right]$$

- Next we prove that $\mathrm{E}T_n(d) = K(d)\,(n + \Theta(...))$, where $K(d) = K(d-1)\frac{Ad-A+B}{Ad+B+1} + c(m, d-1)\frac{D(d-1)}{m(Ad+B+1)}$

Number of triangles

Let G_m^n belong to T-subclass of PA-class. Then

Concentration

If $2A < 1$ then

$$\mathsf{P}\left(|T_n(d) - \mathrm{E}T_n(d)| \geq d^2\sqrt{n}\log n\right) = O\left(n^{-\log n}\right),$$

Therefore, for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : |T_n(d) - \mathrm{E}T_n(d)| \geq \varphi(n)\,\mathrm{E}T_n(d)\right) = 0\,.$$

Number of triangles

Concentration

If $2A = 1$ then

$$\mathsf{P}\left(|T_n(d) - \mathrm{E}T_n(d)| \geq d^2\sqrt{n}\,\log^2 n\right) = O\left(n^{-\log n}\right),$$

Therefore, for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : |T_n(d) - \mathrm{E}T_n(d)| \geq \varphi(n)\,\mathrm{E}T_n(d)\right) = 0\,.$$

Number of triangles

Concentration

If $1 < 2A < \frac{3}{2}$ then

$$\mathsf{P}\left(|T_n(d) - \mathrm{E}T_n(d)| \geq d^2\, n^{2A-\frac{1}{2}}\, \log n\right) = O\left(n^{-\log n}\right),$$

Therefore, for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A(3-4A)-\delta}{4A+2}} : |T_n(d) - \mathrm{E}T_n(d)| \geq \varphi(n)\, \mathrm{E}T_n(d)\right) = 0\,.$$

Idea of the proof

Azuma, Hoeffding

Let $(X_i)_{i=0}^n$ be a martingale such that $|X_i - X_{i-1}| \leq c_i$ for any $1 \leq i \leq n$. Then

$$\mathsf{P}\left(|X_n - X_0| \geq x\right) \leq 2e^{-\frac{x^2}{2\sum_{i=1}^n c_i^2}}$$

for any $x > 0$.

Idea of the proof

Azuma, Hoeffding

Let $(X_i)_{i=0}^n$ be a martingale such that $|X_i - X_{i-1}| \le c_i$ for any $1 \le i \le n$. Then

$$\mathrm{P}\left(|X_n - X_0| \ge x\right) \le 2e^{-\frac{x^2}{2\sum_{i=1}^n c_i^2}}$$

for any $x > 0$.

- $X_i(d) = \mathrm{E}(T_n(d) \mid G_m^i)$, $i = 0, \ldots, n$.
- Note that $X_0(d) = \mathrm{E}T_n(d)$ and $X_n(d) = T_n(d)$.
- $X_n(d)$ is a martingale.
- For any $i = 0, \ldots, n-1$: $|X_{i+1}(d) - X_i(d)| \le Md^2$ (for the case $2A < 1$), where $M > 0$ is some constant.

Idea of the proof

Fix $0 \le i \le n-1$ and some graph G_m^i.

$$\left|\mathrm{E}\left(T_n(d) \mid G_m^{i+1}\right) - \mathrm{E}\left(T_n(d) \mid G_m^i\right)\right| \le$$
$$\le \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\}.$$

Idea of the proof

Fix $0 \leq i \leq n-1$ and some graph G_m^i.

$$\left|\mathrm{E}\left(T_n(d) \mid G_m^{i+1}\right) - \mathrm{E}\left(T_n(d) \mid G_m^i\right)\right| \leq$$
$$\leq \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\}.$$

$\hat{G}_m^{i+1} = \arg\max \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1})$, $\bar{G}_m^{i+1} = \arg\min \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1})$.

Idea of the proof

Fix $0 \leq i \leq n-1$ and some graph G_m^i.

$$\left|\mathrm{E}\left(T_n(d) \mid G_m^{i+1}\right) - \mathrm{E}\left(T_n(d) \mid G_m^i\right)\right| \leq$$
$$\leq \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\}.$$

$$\hat{G}_m^{i+1} = \arg\max \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1}), \ \bar{G}_m^{i+1} = \arg\min \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1}).$$

For $i+1 \leq t \leq n$ put $\delta_t^i(d) = \mathrm{E}(T_t(d) \mid \hat{G}_m^{i+1}) - \mathrm{E}(T_t(d) \mid \bar{G}_m^{i+1})$.

Idea of the proof

Fix $0 \le i \le n-1$ and some graph G_m^i.

$$\left| \mathrm{E}\left(T_n(d) \mid G_m^{i+1}\right) - \mathrm{E}\left(T_n(d) \mid G_m^{i}\right) \right| \le$$
$$\le \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{ \mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{ \mathrm{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\}.$$

$\hat{G}_m^{i+1} = \arg\max \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1})$, $\bar{G}_m^{i+1} = \arg\min \mathrm{E}(T_n(d) \mid \tilde{G}_m^{i+1})$.

For $i+1 \le t \le n$ put $\delta_t^i(d) = \mathrm{E}(T_t(d) \mid \hat{G}_m^{i+1}) - \mathrm{E}(T_t(d) \mid \bar{G}_m^{i+1})$.

$$\delta_{t+1}^i(d) \le \delta_t^i(d) \left(1 - \frac{Ad+B}{t} + O\left(\frac{d^2}{t^2}\right)\right) +$$
$$+\delta_t^i(d-1)\left(\frac{A(d-1)+B}{t} + O\left(\frac{d^2}{t^2}\right)\right) + O\left(\frac{d^2}{t}\right).$$

Local clustering

Let G_m^n belong to T-subclass of PA-class. Then

Local clustering coefficient

If $2A \leq 1$ then for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : \left| C(d) - \frac{2D}{A \cdot \sum_{i=1}^{m} \frac{(i-1)}{[A(i-1)+B]}} \cdot \frac{1}{d} \right| \geq \varphi(n)\, \frac{1}{d} \right) = 0$$

If $1 < 2A < \frac{3}{2}$ then for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A(3-4A)-\delta}{4A+2}} : \left| C(d) - \frac{2D}{A \cdot \sum_{i=1}^{m} \frac{(i-1)}{[A(i-1)+B]}} \cdot \frac{1}{d} \right| \geq \varphi(n)\, \frac{1}{d} \right) = 0$$

Generalized preferential attachment models:

- Power law degree distribution with any exponent $\gamma > 2$
- Constant global clustering coefficient **only** for $\gamma > 3$
- Constant average local clustering coefficient
- Local clustering $C(d) \sim \frac{C}{d}$ for $\gamma > \frac{7}{3}$

Thank You!
Questions?