Clustering properties of Spatial Preferential Attachment model

Lenar Iskhakov[1] Bogumił Kamiński[2] Maksim Mironov[1]
Paweł Prałat[4] Liudmila Prokhorenkova[1,3]

[1]Moscow Institute of Physics and Technology, Moscow, Russia

[2]Warsaw School of Economics, Warsaw, Poland

[3]Yandex, Moscow, Russia

[4]Ryerson University, Toronto, Canada

Introduction

We consider a graph, where nodes are people, web-pages, articles etc. and edges between them are friendship, links, mentioning, citation etc.

Some typical observed properties of real-world graphs:

- low diameter: of order $O(\log n)$
- clustered structure: locally dense, globally sparse
- power-law degree distribution: number of nodes of degree d is proportional to $d^{-\lambda}$
- etc.

SPA model

- SPA model combines **geometry** and **preferential attachment**, it produces an oriented graph
- Final graph G_n is constructed step by step
- At the beginning we have empty graph $G_0 = (V_0 = \varnothing, E_0 = \varnothing)$
- Then at time t we add a new node v_t, which is chosen uniformly at random from unit hyper-cube H in $\mathbb{R}^m$

SPA model

- Every node at time t has its own "sphere of influence", its volume is defined by the number of in-coming edges:

$$S(v,t) = \frac{A_1 deg^-(v,t) + A_2}{t}$$

- To avoid boundary effects we use a torus metric derived from any of the L_p norms. This means that for any two points x, y in H:

$$d(x,y) = \min\{||x - y + u||_p : u \in \{-1,0,1\}^m\}$$

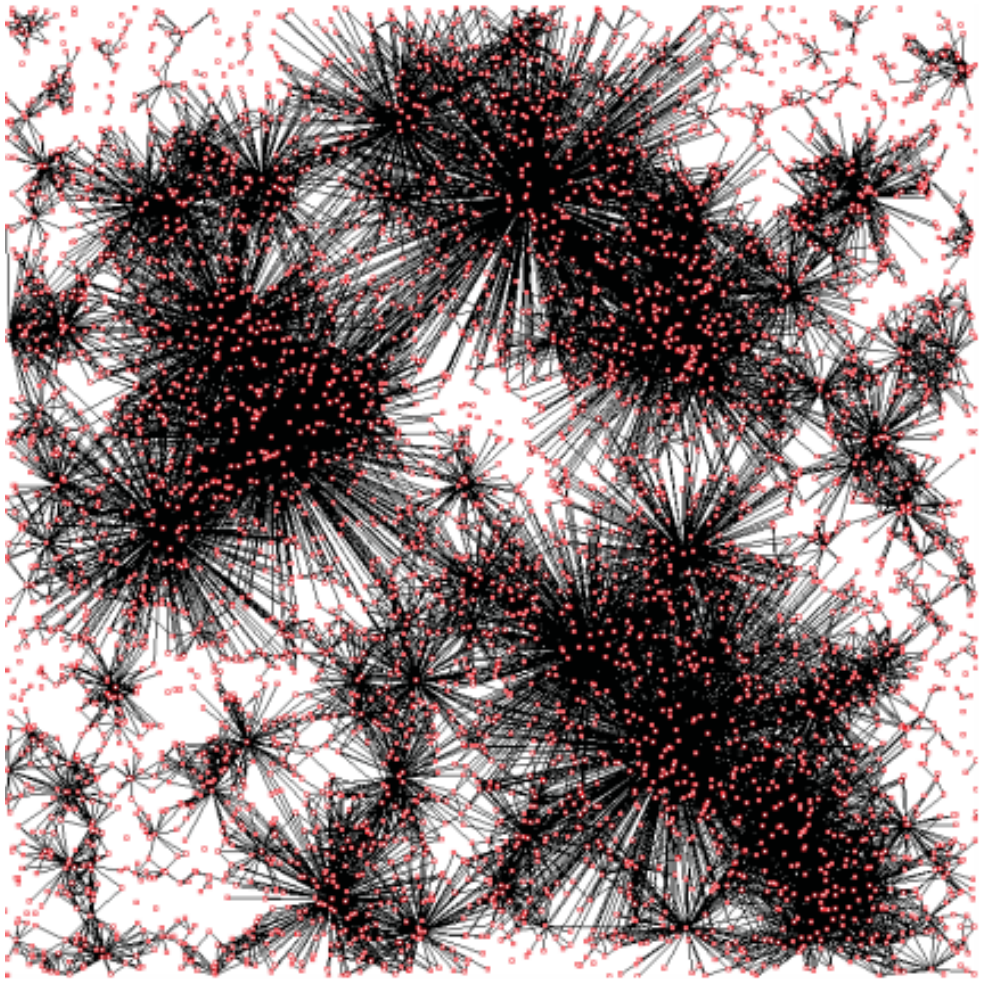
- At time t for all vertices $u \in V_{t-1}$, such that $v_t \in S(u,t)$ new directed edge (v_t, u) is established independently with probability p

SPA model

There are at least three features that distinguish the SPA model from previous models:

- A new node can choose its links purely based on local information.
- The out-degree is not a constant nor chosen according to a pre-determined distribution, but arises naturally from the model.
- The varying size of the influence regions allows long edges between nodes that are spaced far apart. (This implies a certain small world property.)

SPA graph example



Example of simulation of SPA model with $t = 5000, p = 1.0$

SPA model: degree distribution

Important property of SPA is power law with $\lambda = 1 + \frac{1}{p}$.

Theorem (Aiello, Bonato, Cooper, Janssen, Prałat)

A.a.s.

$$N(0,t) = (1+o(1))\frac{t}{1+p},$$

and for all d satisfying $\omega \le d \le \left(\frac{t}{\log^8 t}\right)^{\frac{p}{4p+2}}$,

$$N(d,t) = t(1+o(1))cd^{-1-\frac{1}{p}}.$$

for some constant c.

Here $N(d,t)$ is a number of vertices of in-degree d at time t.

SPA model: degree distribution

Some results also about out-degrees:

Theorem (Aiello, Bonato, Cooper, Janssen, Prałat)

A.a.s.

$$\max_{0 \le i \le t} deg^+(v_i, t) \ge (1 + o(1))p\frac{\log t}{\log \log t}.$$

However, a.a.s. all nodes have out-degree $O(\log^2 t)$.

Theorem (Aiello, Bonato, Cooper, Janssen, Prałat)

A.a.s. $deg^+(v_t, t) = O(\log^2 t)$.

Clustering coefficient

We want to analyze SPA model in terms of clustering. One way to measure it is using average local clustering coefficient:

$$c^{av}(G) = \frac{1}{|V(G)|} \sum_{v \in V(G)} c^{loc}(v),$$

where

$$c^{loc}(v) = \frac{\text{number of edges between neighbours of v}}{\binom{\deg(v)}{2}}$$

Here is a result about c^{av} for undirected SPA graph:

Theorem (Jacob, Mörters)

There exists a strictly positive value c such that $c^{av}(G_n) \to c$ in probability with $n \to \infty$.

Coefficient $c^{av}_{\varepsilon,d}$

Lets consider set of vertices with in-degree ε-close to d:

$$V_{\varepsilon,d}(G) = \left\{u \in V(G) \mid deg^-(v) \in [(1-\varepsilon)d, (1+\varepsilon)d]\right\}.$$

Then the average local clustering coefficient among all vertices with in-degree ε-close to d in graph G is as follows:

$$c^{av}_{\varepsilon,d}(G) = \frac{1}{|V_{\varepsilon,d}(G)|} \sum_{v \in V_{\varepsilon,d}(G)} c^{loc}(v).$$

According to Ravasz and Barabaśi there are several examples of big real-world networks like WWW, actor network and others, where $c^{av}_d (:= c^{av}_{0,d})$ can be approximated by d^{-1}.

Results

For SPA model we have the following result about clustering coefficient, which shows similarity with some real-world large networks:

Theorem

Let $\varepsilon, \delta \in (0, 1/2)$ *be any two constant, and*

$$\log^{C_1(\delta,p,A_1)} n \leq d := d(n) \leq n^{pA_1-\delta}.$$

Then a.a.s.

- *Clustering coefficient*

$$c^{av}_{\varepsilon,d}(G_n) = \Theta(1/d).$$

- *Moreover, for almost all vertices* v *in* $V_{\varepsilon,d}(G_n)$ *have*

$$c^{loc}_{\varepsilon,d}(v) = \Theta(1/d).$$

Proof ideas: forecasting in-degree

We need to introduce a new value T_v as

$$T_v := \min\{t : deg^-(v,t) > \omega \log n\}.$$

This is a very important variable because of the following.

Lemma (Janssen, Prałat, Wilson)

A.s.s. For every vertex v and for all, $T_v \leq t \leq n$,

$$\deg^-(v,t) = \deg^-(v,T_v)\left(\frac{t}{T_v}\right)^{pA_1}(1+o(1)).$$

This means that since node v accumulated $\omega \log n$ in-neighbours we can forecast its in-degree and, hence, sphere of influence.

Proof ideas

Let us divide $c^{loc}_{\varepsilon,d}(v)$ for $v \in V(G_n)$ in two parts:

$$c^{loc}_{\varepsilon,d}(v) = c^{loc}_{old}(v) + c^{loc}_{new}(v),$$

where

$$\begin{aligned} c^{loc}_{old}(v) &= |E_{old}(N^-(v,n))| \Big/ \binom{\deg^-(v,n)}{2}, \\ c^{loc}_{new}(v) &= |E_{new}(N^-(v,n))| \Big/ \binom{\deg^-(v,n)}{2}. \end{aligned}$$

And

$$\begin{aligned} E_{old}(N^-(v,n)) &= \{(u,w) \in E_n : u \in N^-(v,n), w \in N^-(v,T_v)\}, \\ E_{new}(N^-(v,n)) &= E(N^-(v,n)) \setminus E_{old}(N^-(v,n)) \end{aligned}$$

Proof ideas

What do we need *old* and *new* edges for:

- We can say nothing about first $\omega \log n$ in-neighbours of vertex v (before time T_v), but hopefully there are not too much of them.
- Since we know a.a.s. in-degree of node v after time T_v, we know its "sphere of influence" and can understand how new edges appear and influence on node v.

Theorem

A.s.s. for any vertex v such that $\deg^-(v,n) = d = d(n) \geq \omega \log n$ holds that

$$\begin{aligned} c^{loc}_{new}(v,n) &= \Theta(1/d), \\ c^{loc}_{old}(v,n) &= O(\omega \log n/d). \end{aligned}$$

Results

Unfortunately, if vertex v lands in a densely populated region of hyper-cube H, it might happen that $c_{old}^{loc}(v)$ is much larger than $1/d$.

Theorem

A.a.s. exist vertex v in graph G_n such that

$$c_{old}^{loc}(v) \gg 1/d.$$

That is why we can't change "almost all" to "all" in the theorem few slides ago.

Simulations: constructing graph in SPA model

- Complexity of naive algorithm is $\Theta(n^2)$
- Idea: divide all verticies into two groups *heavy* H and *light* L
- Light verticies have in-degree at most some D, their spheres are bounded. The number of light nodes depends on D and can be estimated
- We divide unit square into k squares with side $1/\sqrt{k}$, k is chosen such that every light sphere touches not more than 9 small squares
- New node can fall into any heavy sphere, but for light spheres we can check just some of them - only squares around new node
- We want k to be as large as possible, so number of comparisons is a function of D and we can find the optimum, which gives complexity of

$$\Theta(n^{2-1/(pA+1)})$$

Simulations: algorithm efficiency

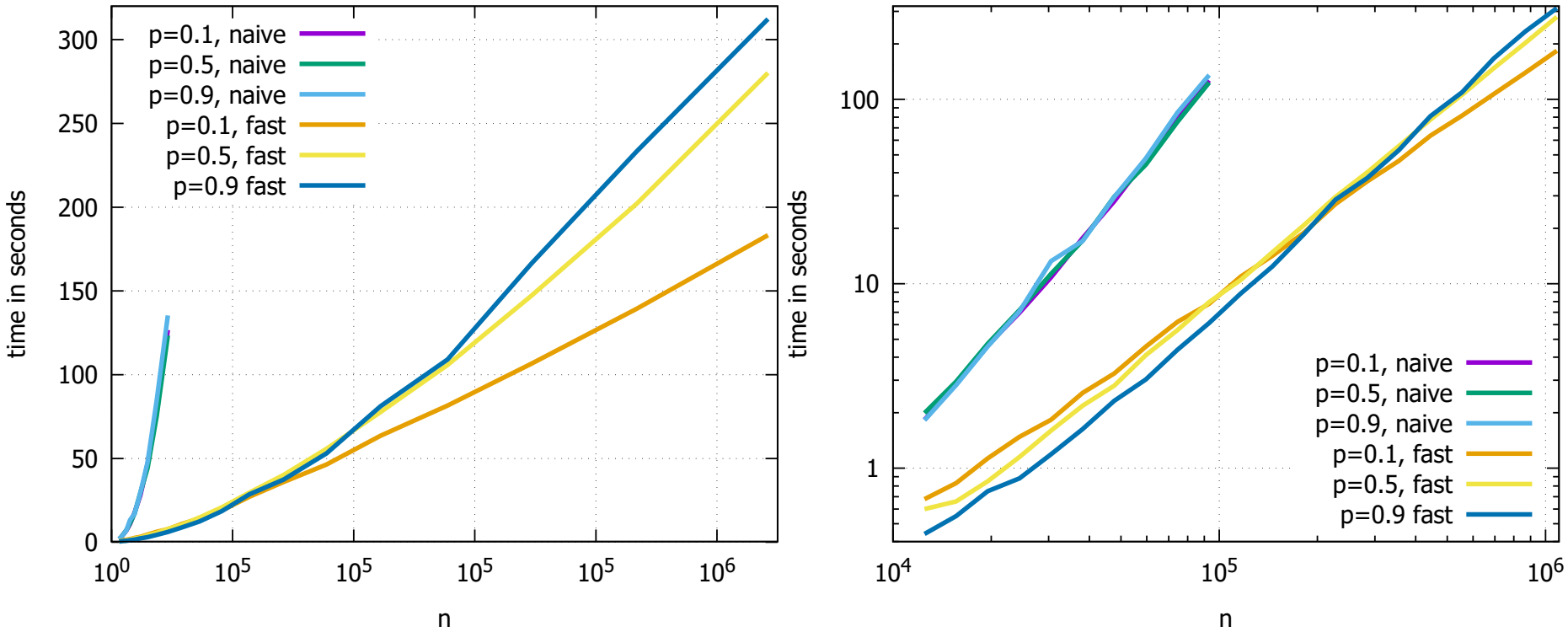


Figure: Running time of the proposed and the naive algorithms.

We also believe that it is possible to achieve $\Theta(n \log n)$ complexity by dividing not into two but into $\log n$ groups with exponentially increasing degree.

Simulations: clustering coefficient

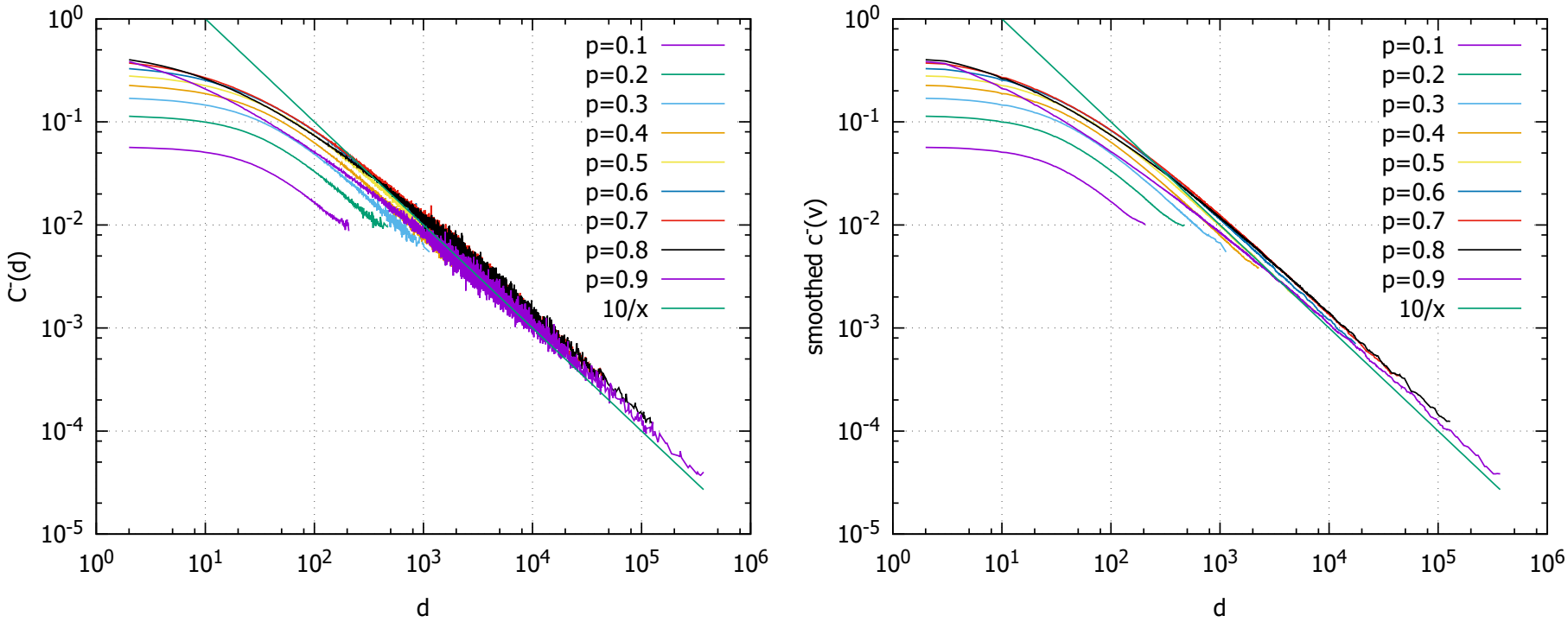


Figure: Average local clustering coefficient and smoothed within ε-interval.

Our statement for $c^{av}_{\varepsilon,d}(G_n) = \Theta(1/d)$ requires $d > 10^{20}$, which is irrelevant and pessimistic, but anyway, we see that it holds even for much smaller d.

Simulations: old and new neighbours

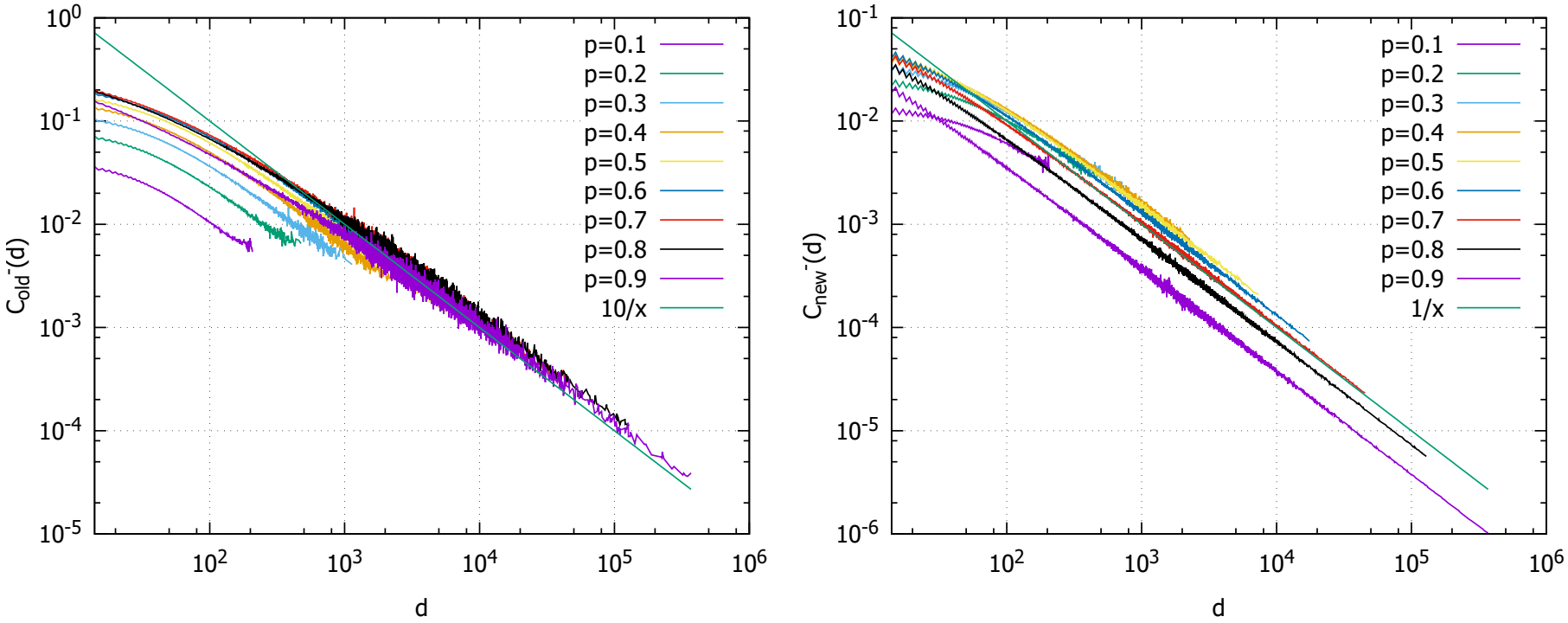


Figure: Comparison of "new" and "old" parts of the average local clustering coefficient.

Simulations: individual coefficient

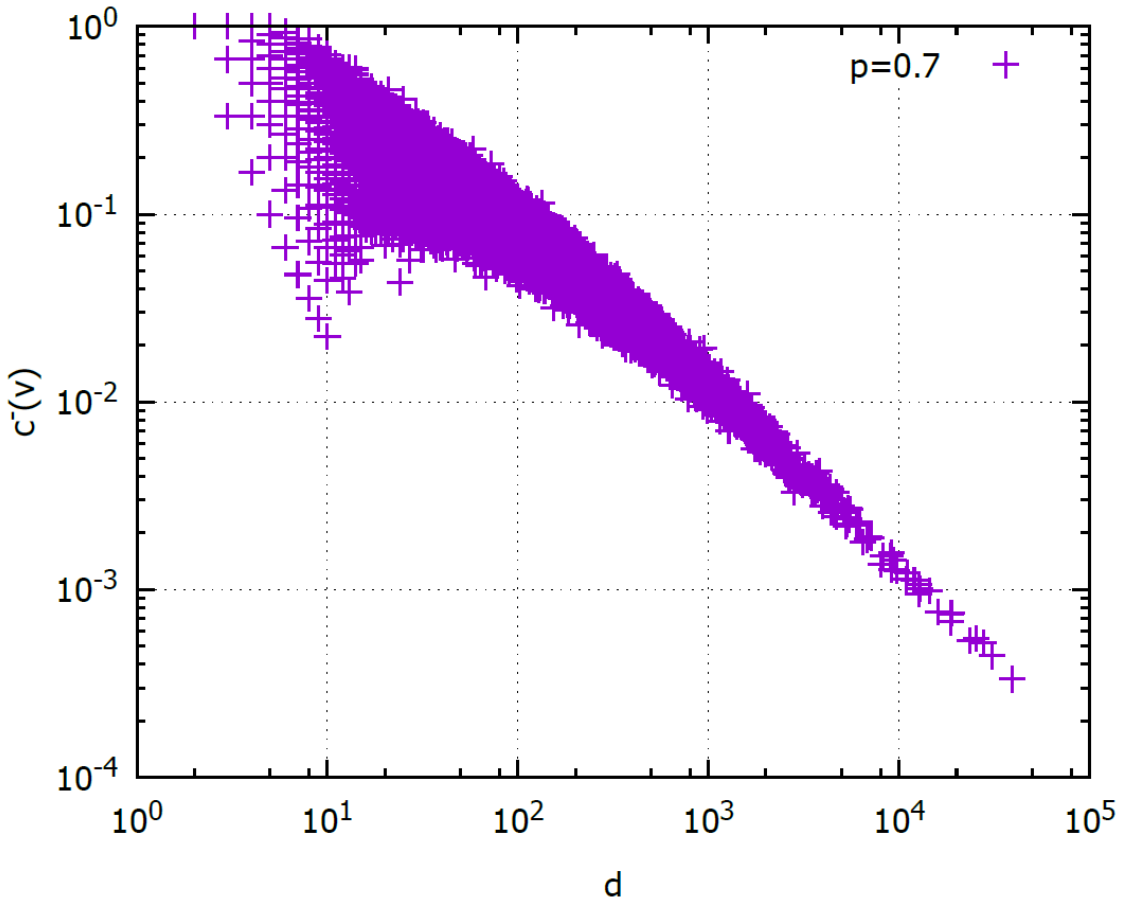


Figure: The distribution of individual local clustering coefficients.

Thanks for the attention!