Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

On the clustering property of the random intersection graphs

Xin Yao[1], Jinwen Chen[2] , Changshui Zhang[3] , Yanda Li [3]

submitted: November 3, 2008

[1] Weistrass Institute
for Applied Analysis and Stochastics
Mohrenstrasse 39
10117 Berlin
Germany
E-Mail: yao@wias-berlin.de

[2] Department of Mathematics
Tsinghua University
Chengfu Road, Haidian
100084, Bejing
China

[3] Department of Automation
Tsinghua University
Chengfu Road, Haidian
100084, Bejing
China

No. 1369
Berlin 2008

2000 *Mathematics Subject Classification.* 05C80, 05C75, 05C40.

Key words and phrases. random intersection graphs, clutering coefficient, phase transition.

Research supported by the DFG in the Dutch–German Bilateral Research Group "Mathematics of Random Spatial Models from Physics and Biology".

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

A random intersection graph $\mathcal{G}_{V,W,p}$ is induced from a random bipartite graph $\mathcal{G}^*_{V,W,p}$ with vertices classes V, W and the edges incident between $v \in V$ and $w \in W$ with probability p. Two vertices in V are considered to be connected with each other if both of them connect with some common vertices in W. The clustering properties of the random intersection graph are investigated completely in this article. Suppose that the vertices number be $N = |V|$ and $M = |W|$ and $M = N^\alpha$, $p = N^{-\beta}$, where $\alpha > 0, \beta > 0$, we derive the exact expressions of the clustering coefficient C_v of vertex v in $\mathcal{G}_{V,W,p}$. The results show that if $\alpha < 2\beta$ and $\alpha \neq \beta$, C_v decreases with the increasing of the graph size; if $\alpha = \beta$ or $\alpha \geq 2\beta$, the graph has the constant clustering coefficients, in addition, if $\alpha > 2\beta$, the graph connecChangshui Zhangts almost completely. Therefore, we illustrate the phase transition for the clustering property in the random intersection graphs and give the condition that $\mathcal{G}_{V,W,p}$ being high clustering graph.

There are a lots of collaboration networks in the real world, such as scientists collaboration networks[10, 11], actor collaboration networks[2, 3, 1], metabolism networks[7, 6, 5], et al. In these networks, there are two types of vertices sets V, W and the vertices in V may link to some of the vertices in W. The vertices $v_1, v_2 \in V$ are considered to be connected with each other if both of these two vertices connect to some common vertices in W. For describing such intersection structure mathematically, the *random intersection graph* was introduced firstly in [12, 8]. Let $\mathcal{G}^*_{V,W,p}$ be the random bipartite graph with two vertices classes V and W. That probability that $v \in V$ and $w \in W$ be connected with each other in $\mathcal{G}^*_{V,W,p}$ is p and the connections between the vertices of V and W are independent with each other. The random intersection graph $\mathcal{G}_{V,W,p}$ is a random graph with vertices set V and the connections in $\mathcal{G}_{V,W,p}$ are induced from $\mathcal{G}^*_{V,W,p}$ in this way: the vertices $v_1, v_2 \in V$ are connected with each other if and only if both of them link to some vertex $w \in W$ in $\mathcal{G}^*_{V,W,p}$. The subgraph property and the degree distribution of the random intersection graphs are illustrated in [4] and [13], respectively. The intersection structure in the networks is considered to be the reason that many real world collaboration networks have high clustering coefficients, since all the vertices in V linking to the same $w \in W$ will form a complete subgraph and have very high local clustering coefficients. However, if $p = 0$, it is clear that the $\mathcal{G}_{V,W,p}$ is an empty graph and the clustering coefficient is 0, which means that there may exist a threshold p_c that $\mathcal{G}_{V,W,p}$ is a high clustering graph only when $p > p_c$.

The *clustering coefficient* of the vertex $v \in V$ is defined as

$$C_v = \frac{2c_v}{k_v(k_v-1)},\ C = \frac{1}{N}\sum_{v\in V} C_v, \tag{1}$$

where k_v is the degree of vertex v, c_v, which is called *clustering degree* in [14, 15], is the number of edges that actually exist between these k_v vertices and N is the number of vertices in V. The cluster coefficients of collaboration networks are studied in [9] under the different definition from Eq.(1). However, there are few mathematical results for $\mathcal{G}_{V,W,p}$ to characterize the clustering properties in the form of Eq.(1), which is a more popular definition for clustering coefficients. We studied the clustering property of $\mathcal{G}_{V,W,p}$ thoroughly in this article. The exact expression of C_v is presented and as the result a phase transition phenomenon is discovered.

Let $N = |V|$, $M = |W|$ be the vertices number and

$$M = N^{\alpha}, \quad p = cN^{-\beta} \tag{2}$$

where $\alpha \geq 0, \beta \geq 0$ and $c > 0$ are the constants. Let W_v be the set of vertices in W which link with $v \in V$ and V_v be the set of vertices in $V\backslash\{v\}$ which connected with v in $\mathcal{G}_{V,W,p}$ by linking with some common vertices in W with v. It can be seen that both W_v and V_v are the random set. Since the connections between vertices of V and W are independent, the probability that $v_1, v_2 \in V$ are connected each other in $\mathcal{G}_{V,W,p}$ is $\hat{p} = 1-(1-p)^M$, which decides the asymptotical property of $\mathcal{G}_{V,W,p}$ under the settings in Eq.(2). When $N \to \infty$, the speeds that $p \to 0$ and $M \to \infty$ will be very different under various α and β, as a result the asymptotical property of $\hat{p}$ will be different.

From above description, it can be seen that the random intersection graph $\mathcal{G}_{V,W,p}$ is in fact a random graph with vertices set V and the connective probability between any two vertices v_1, v_2 is $\hat{p}$. In this case, will all the results of classical random graph presented by Erdös and Renyi be extended to $\mathcal{G}_{V,W,p}$? If so, the random intersection graphs should be a trivial model. However, that is not true. What makes $\mathcal{G}_{V,W,p}$ different is the independence. In the random intersection graphs, the independence of the connections disappears, because

$$P\left(v_2 \in V_{v_3} | v_2 \in V_{v_1}, v_3 \in V_{v_1}\right) \neq P\left(v_2 \in V_{v_3}\right), \tag{3}$$

that is to say, the connection between v_2 and v_3 depends on whether the vertex v_1 being connected with both v_2 and v_3. Therefore, the results of Erdos-Renyi random graphs can not be extended to $\mathcal{G}_{V,W,p}$ directly and we should find the way out for any properties of the random intersection graphs. We will present the analysis of the clustering coefficients of $\mathcal{G}_{V,W,p}$ in the following text.

The local clustering property, which describing the correlation between the clustering degree c_v (or clustering coefficient C_v) of vertex $v \in V$ and the degree k_v, will be investigated at first. Since the connections between the vertices of V result from the links between V and W, the correlation between c_v and $L_v = |W_v|$ should

be studied at first, where the random set W_v is the collection of all the vertices in W that linked with v, therefore L_v is a random variable. With the random set W_v as the condition, the expectation of c_v is

$$\begin{aligned} E\left[c_v|W_v\right] &= \sum_{V_v \subset V\setminus\{v\}} E\left[c_v|W_v, V_v\right] P(V_v|W_v) \\ &= \sum_{i=1}^{N-1} \sum_{|V_v|=i} E\left[c_v|W_v, V_v\right] P(V_v|W_v) \end{aligned} \tag{4}$$

where $E\left[c_v|W_v, V_v\right]$ is the expectation with the random sets W_v, V_v as the conditions. It can be seen that $E\left[c_v|W_v\right]$ may be derived if $E\left[c_v|W_v, V_v\right]$ and $P(V_v|W_v)$ are known.

As V_v be the vertices connected with v, and any two of these vertices connect with each other with probability $\hat{p}$, we can obtain that

$$E\left(c_v|W_v, V_v\right) = \binom{|V_v|}{2} P\left(v_1 \in V_{v_2}, v_1, v_2 \in V_v|W_v, V_v\right) \tag{5}$$

since there are at most $\binom{|V_v|}{2}$ possible connections among the vertices in V_v.

With the random sets W_v, V_v as the conditions, the probability that any two vertices $v_1, v_1 \in V_v$ being connected with each other can be denoted as

$$\begin{aligned} & P\left(v_1 \in V_{v_2}, v_1, v_2 \in V_v|W_v, V_v\right) \\ = & \frac{1-(1-p^2)^{L_v}}{\left(1-(1-p)^{L_v}\right)^2} + \left(1-p^2\right)^{L_v}\left(1-(1-p^2)^{M-L_v}\right) \end{aligned} \tag{6}$$

The first term of Eq.(6) is the probability, with the random sets W_v, V_v as the conditions, that there exist $w \in W_v$ such that $w \in W_{v_1}$ and $w \in W_{v_2}$. The second term of Eq.(6) is the probability, with the random sets W_v, V_v as the conditions, that v_1, v_2 are connected with no common vertex in W_v but connected with some common vertices in $W \setminus W_v$.

For the vertex v, if the set of its linked vertices in W is W_v, it has the connected vertices set V_v in $\mathcal{G}_{V,W,p}$ with probability

$$P\left(V_v|W_v\right) = \left[1-(1-p)^{L_v}\right]^{|V_v|}(1-p)^{L_v(N-|V_v|-1)} \tag{7}$$

Combining with the Eq.(5)(6)(7), we have

$$\begin{aligned} & E\left[c_v|W_v\right] \\ = & \frac{(N-1)(N-2)}{2}\Big[1-(1-p^2)^{L_v}+(1-p^2)^{L_v} \\ & \cdot\left(1-(1-p)^{L_v}\right)^2 \cdot \left(1-(1-p^2)^{M-L_v}\right)\Big]. \end{aligned} \tag{8}$$

Under the various α and β, the asymptotical property of Eq.(8) will be different as shown below

$$E\left[c_v|W_v\right] \sim \begin{cases} \dfrac{c^2 L_v}{2} N^{2-2\beta}, & \beta \le \alpha < 2\beta \\ \dfrac{\left(1-e^{-L_v p}\right)^3}{2} N^2, & \alpha = 2\beta \\ \dfrac{N^2}{2}, & \alpha > 2\beta \end{cases} \tag{9}$$

Eq.(9) gives the description of the mean clustering degree of v under the condition that W_v is known. If only $L_v = |W_v|$ is known and it is not clear which vertices are in W_v, then $P(W_v|L_v) = 1/\binom{M}{L_v}$ since every set W_v which has L_v elements incidents with the same probability and there are $\binom{M}{L_v}$ such W_v, therefore,

$$\begin{aligned} E\left[c_v|L_v\right] &= \sum_{|W_v|=L_v} E\left[c_v|W_v\right] P\left(W_v|L_v\right) \\ &= \binom{M}{L_v} \frac{E\left[c_v|W_v\right]}{\binom{M}{L_v}} \\ &= E\left[c_v|W_v\right]. \end{aligned} \tag{10}$$

Eq.(9) and (10) tell the following truth: when $\beta \le \alpha < 2\beta$, the vertex clustering degree c_v increases with L_v linearly; when $\alpha = 2\beta$, c_v increases with L_v exponentially; when $\alpha > 2\beta$, c_v is almost $N^2/2$. Considering the connective probability among the vertices of V, $\hat{p} = 1-(1-p^2)^M$, and Eq.(2), we find that the greater α means the greater M and the higher connective probability, similarly, the greater β means the smaller p and the smaller connective probability. For the random intersection graph, higher connective probability means the high clustering degree. Therefore, if α is great enough relative to β, the clustering property of the graph will change. The local clustering property of $\mathcal{G}_{V,W,p}$ under different α, β described by Eq.(9) are illustrated in Fig. 1.

Since

$$E(k_v|L_v) = (N-1)\left(1-(1-p)^L\right), \tag{11}$$

Combining Eq.(1),(9) and (10), we can derive the conditional expectation of clustering coefficients with the asymptotical analysis under different α, β, we have

$$E[k_v|L_v] \sim \begin{cases} cL_v N^{1-\beta}, & \beta \le \alpha < 2\beta \\ N\left(1-e^{-\frac{cL_v}{N^\beta}}\right), & \alpha = 2\beta \\ N, & \alpha > 2\beta. \end{cases} \tag{12}$$

Given the condition L_v, the conditional distribution of k_v is binomial distribution $\mathrm{Bi}(N-1, \hat{p})$. Therefore, from the probability inequalities

$$P\left(k_v \ge E[k_v|L_v] + t\right) \le \exp\left(-\frac{t^2}{2(E[k_v|L_v] + t/3)}\right) \tag{13}$$

$$P\left(k_v \le E[k_v|L_v] - t\right) \ge \exp\left(-\frac{t^2}{2E[k_v|L_v]}\right) \tag{14}$$

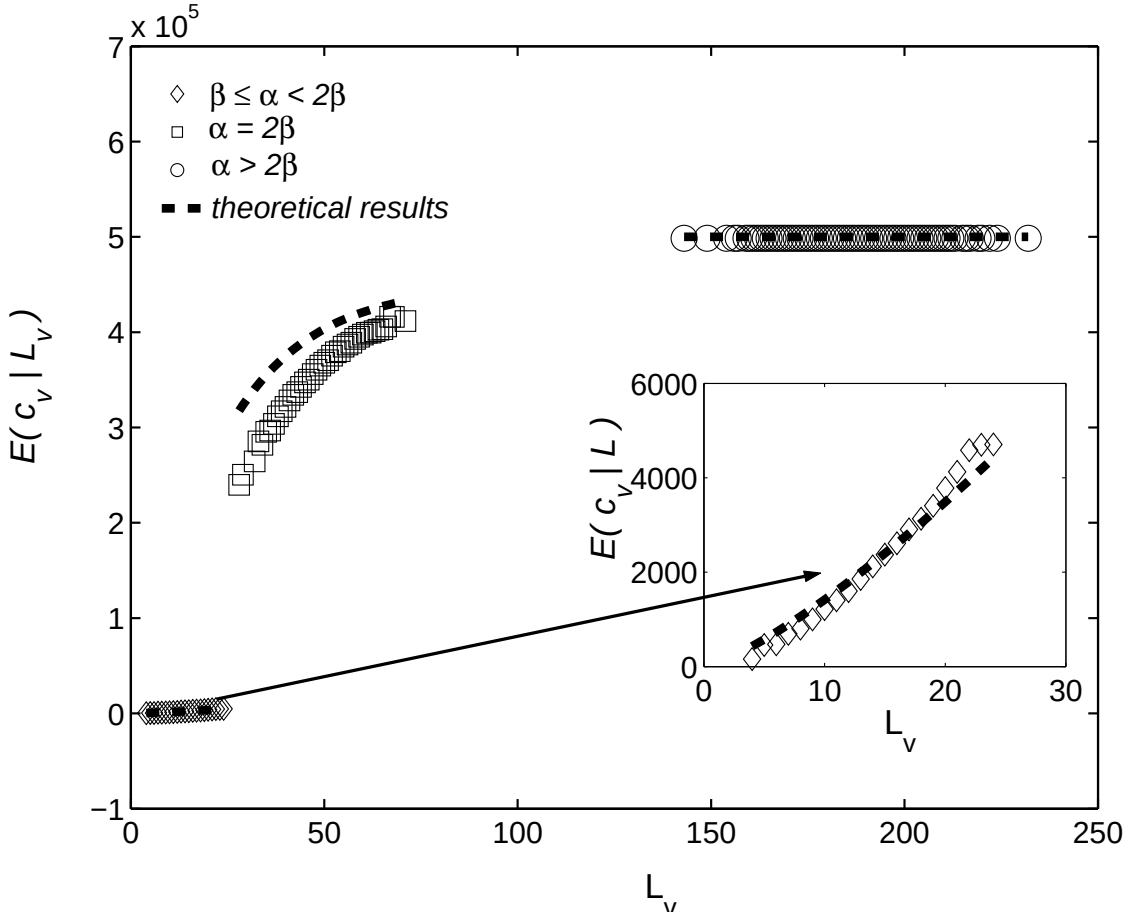

Figure 1: The conditional expectation of the clustering degree of v given L_v . The imbedded graph are the magnification of the case $\beta \leq \alpha < 2\beta$. Given the size of the networks, N=1000, it can be seen that when $\alpha = 1.0, \beta = 0.7$ ($\diamondsuit$), $E(c_v|L_v)$ increase with L_v linearly; when $\alpha = 1.0, \beta = 0.5$ ($\square$), $E(c_v|L_v)$ increase with L_v exponentially; when $\alpha = 1.0, \beta = 0.3$ ($\bigcirc$), $E(c_v|L_v) \sim N^2/2$. The dash lines are the theoretical results.

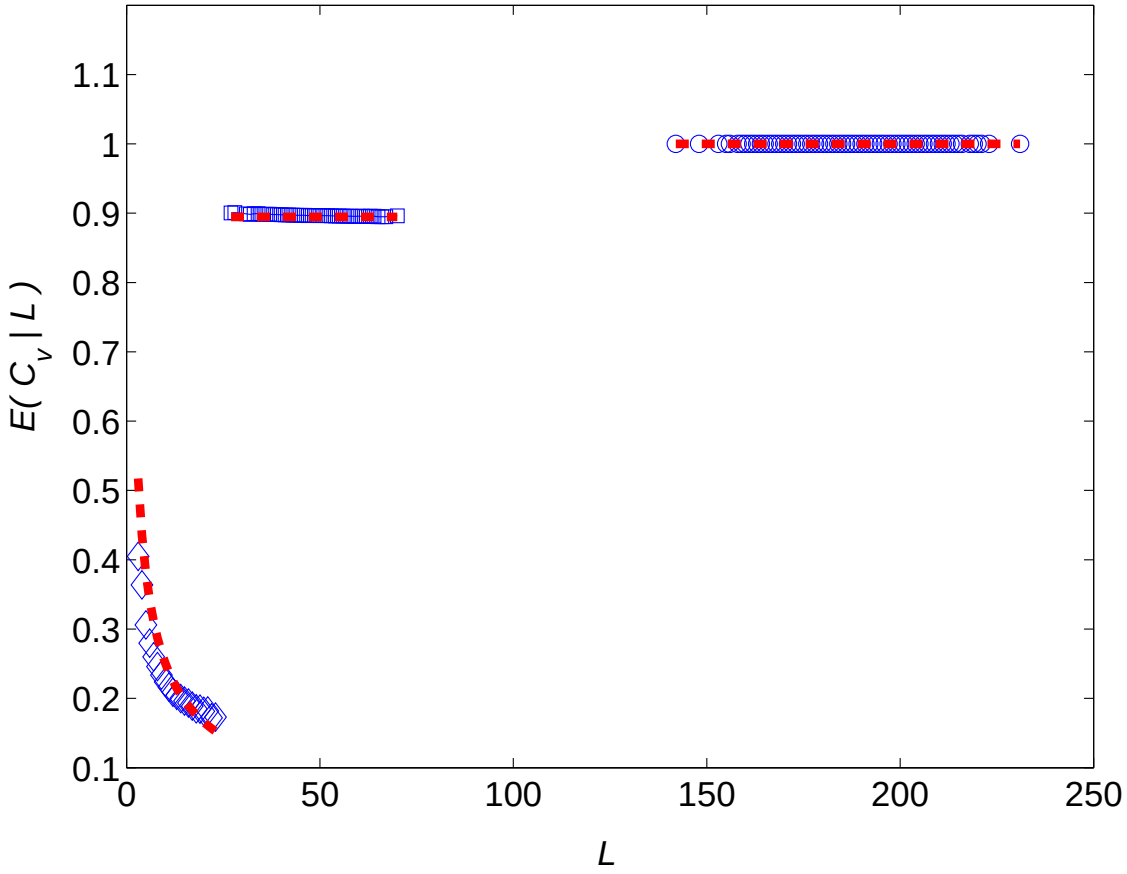

Figure 2: The dependence of the clustering coefficient of v on the number of the W links of v . Given the size of the networks, it can be seen that when $\alpha = 1.0, \beta = 0.7$ ($\diamondsuit$), $E(C_v|L_v)$ decrease with L_v in hyperbolic way; when $\alpha = 1.0, \beta = 0.5$ ($\square$), $E(C_v|L_v)$ is near to 1 and almost a constant; when $\alpha = 1.0, \beta = 0.3$ ($\bigcirc$), $E(C_v|L_v) = 1.$. The dash lines are the theoretical results.

we have

$$k_v \sim \begin{cases} cL_v N^{1-\beta}, & \beta \leq \alpha < 2\beta \\ N\left(1 - e^{-\frac{cL_v}{N^\beta}}\right), & \alpha = 2\beta \\ N, & \alpha > 2\beta. \end{cases} \tag{15}$$

asymptotically almost surely with $N \to \infty$, which means that k_v concentrates highly to its mean value so that the probability that k_v apart from $E[k_v|L_v]$ goes to zero when N goes to infinity. So the clustering coefficient follows directly from the Eq.(1) as

$$E\left[C_v|L_v\right] \sim \begin{cases} \dfrac{1}{L_v}, & \beta \leq \alpha < 2\beta \\ 1 - e^{-\frac{cL_v}{N^\beta}}, & \alpha = 2\beta \\ 1, & \alpha > 2\beta. \end{cases} \tag{16}$$

and the simulation results are illustrated in Fig. 2.

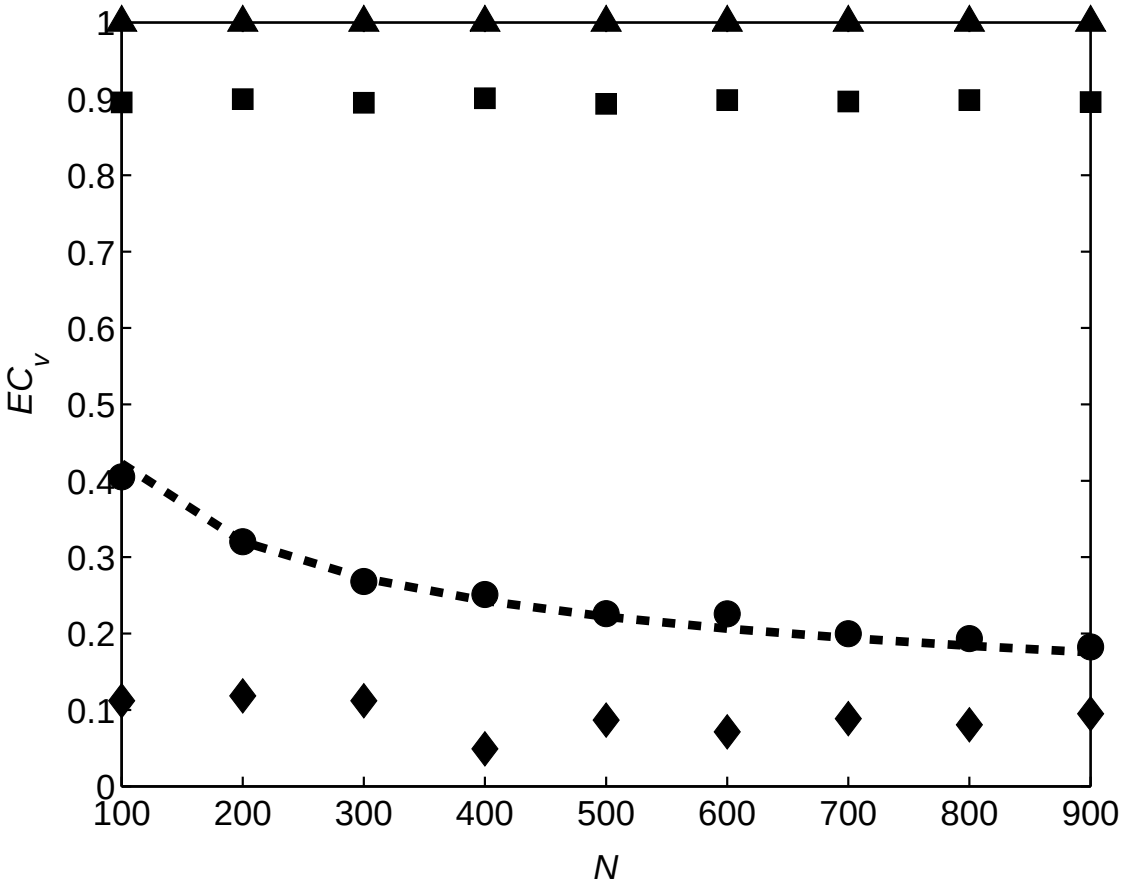

Figure 3: The dependence of the clustering coefficient of v on the vertices number in $\mathcal{G}_{V,W,p}$. It can be seen from the figure that when $\alpha = \beta = 1.0$ (♦), $E(C_v)$ is a small constant; when $\alpha = 1.0, \beta = 0.7$ (•), $E(C_v)$ decreases with the growth of the network size; when $\alpha = 1.0, \beta = 0.5$ (■), EC_v is a constant near to 1; when $\alpha = 1.0, \beta = 0.3$ (▲), $EC_v = 1$.

If we consider the property of L_v under different α, β , we can obtain the following results

$$L_v \sim cN^{\alpha-\beta} \tag{17}$$

asymptotically almost surely with $N \to \infty$ and we notice that $E(C_v|L_v) \sim 1-e^{-c^2}$, which is a constant. Since $E[C_v] = \sum_l E[C_v|L_v]P(L_v = l)$, we have

$$E\left[C_v\right] \approx \begin{cases} c' & \alpha = \beta \\ \frac{1}{c}N^{-(\alpha-\beta)}, & \beta < \alpha < 2\beta \\ 1 - e^{-c^2}, & \alpha = 2\beta \\ 1, & \alpha > 2\beta. \end{cases} \tag{18}$$

and the simulation results are illustrated by Fig. 3.

Eq.(18) describes the whole clustering property of the random intersection graph $\mathcal{G}_{V,W,p}$. It can be seen that the clustering properties of $\mathcal{G}_{V,W,p}$ are very different for

different parameter α, β . The decisive facts are not the value of the parameters, but the relative relation of the two parameters. In fact, increasing β will reduce the connective probability and so the clustering coefficient, however, increasing α will enhance the connective probability and clustering coefficient so as to counteract the effect from the increase of β .

The graphs or networks with high clustering coefficients are the research focus now and the reasons of the high clustering are interesting problems. Some evolutionary mechanisms have been studied for such problem. However, the random intersection graphs are easier to be high clustering graphs. All the vertices in V which link to the same $w \in W$ will be connected with each other. Therefore, $\mathcal{G}_{V,W,p}$ is constituted from a lots of complete subgraphs. Eq.(18) tells that $\mathcal{G}_{V,W,p}$ will have high clustering coefficient when $\alpha \geq 2\beta$. In other words, $\alpha = 2\beta$ is the critical point for the clustering property in $\mathcal{G}_{V,W,p}$.

Moreover, considering Eq.(2) and the critical condition expressed by α, β , we can obtain easily the critical probability, $p_c = c/\sqrt{M}$, that $\mathcal{G}_{V,W,p}$ will be a high clustering graph if $p \geq p_c$. In addition, the local clustering property, which described by $E[C_v|L_v]$, only varies with L_v at critical point, $p = p_c$. When $p > p_c$, $\mathcal{G}_{V,W,p}$ is almost a complete graph. This means the connectivity property of $\mathcal{G}_{V,W,p}$ varies sharply at the critical point.

As we have given the complete analysis of the clustering property of $\mathcal{G}_{V,W,p}$, the condition that $\mathcal{G}_{V,W,p}$ be a high clustering graph and the characteristics of $\mathcal{G}_{V,W,p}$ at critical point has been revealed. What may be more interesting and important are the analysis of the clustering properties of scale-free random intersection graphs, which will be our future work.

References

[1] R. Albert and A.-L. Barabási. Topology of evolving networks: Local events and universality. *Phys. Rev. Lett.*, 85:5234–5237, 2000.

[2] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proc. Nat. Acad. Sci. U. S. A.*, 97:11149, 2000.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen. Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $g(n, m, p)$ and $g(n, p)$ models. *Random Structures and Algorithm*, 16:156–176, 2000.

[5] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell. Relevant cycles in chemical reaction networks. *Advances in Complex Systems*, 4:207–226, 2001.

[6] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41, 2001.

[7] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651, 2000.

[8] M. Karoński, E. R. Scheinerman, and K. Singer-Cohen. On random intersection graphs: the subgraph problem. *Combin. Probab. Comput.*, 08:131–159, 1999.

[9] M. E. J. Neuman, S. H. Strogatz, and D. J. Watt. Random graph with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[10] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131, 2001.

[11] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, 2001.

[12] K. Singer-Cohen. *Random intersection graphs.* PhD thesis, The Johns Hopkins University, Baltimore, Maryland, 1995.

[13] D. Stark. The vertex degree distribution of random intersetion graphs. *Random Structures and Algorithm*, 24:249–258, 2004.

[14] X. Yao, C. S. Zhang, J. W. Chen, and Y. D. Li. On the formation of degree and cluster-degree correlations in scale-free networks. *Physica A*, 353:661–673, Aug. 2005.

[15] X. Yao, C. S. Zhang, J. W. Chen, and Y. D. Li. On the scale-free intersection graphs. In *Lecture Notes in Computer Science (ICCSA2005)*, pages 1217–1224, Singapore, May 2005.