Asymptotic normality of global clustering coefficient of uniform random intersection graph

Mindaugas Bloznelis*
joint work with Jerzy Jaworski

*Vilnius University, Lithuania

http://www.mif.vu.lt/∼bloznelis

May 16, 2018

Global clustering coefficient

The global clustering coefficient of a finite graph $\mathcal{G}$ is the ratio

$$C_{\mathcal{G}} = 3N_{\Delta}/N_{\vee},$$

N_{Δ} is the number of triangles, $N_{\vee}$ is the number of $2-$paths.

$C_{\mathcal{G}}$ represents the probability that a randomly selected path of length 2 induces triangle in $\mathcal{G}$.

$C_{\mathcal{G}}$ is a commonly used network characteristic, assessing the strength of the statistical association between neighboring adjacency relations.

For example, in a social network the tendency of linking actors which have a common neighbor is reflected by a non-negligible value of the global clustering coefficient.

Random intersection graph, RIG

Clustering in a social network can be explained by an auxiliary bipartite structure: each actor is prescribed a collection of attributes and any two actors sharing a common attribute have high chances of being adjacent, cf. [Newman et all. *Random graphs with arbitrary degree distributions and their applications.* Phys.Rev.E, 2002].

The respective random intersection graph (RIG) on the vertex set $V = \{v_1, \ldots, v_n\}$ and with the auxiliary attribute set $W = \{w_1, \ldots, w_m\}$ defines adjacency relations with the help of a random bipartite graph H linking actors (=vertices) to attributes: two actors are adjacent in RIG if they have a common neighbour in H.

For H drawn uniformly at random from the class of bipartite graphs where each actor $v_i \in V$ has exactly r neighbours in W we obtain the uniform RIG denoted $G = G(n, m, r)$.

Alternative definition: vertices $v_1, \ldots, v_n$ of G are represented by iid random subsets $S_1, \ldots, S_n \subset W$, each of size r. Vertices v_i, v_j are adjacent whenever $S_i \cap S_j \neq \emptyset$.

Uniform random intersection graph $G(n, m, r)$

Vertices $v_1, \ldots, v_n$ of $G = G(n, m, r)$ are represented by iid random subsets $S_1, \ldots, S_n \subset W = \{w_1, \ldots, w_m\}$, each of size r. Vertices v_i, v_j are adjacent whenever $S_i \cap S_j \neq \emptyset$.

The graph has been widely studied in the literature mainly as a model of secure wireless sensor network that uses random predistribution of keys [Eschenauer and Gligor (2002)]: each sensor v_i is prescribed a collection S_i of "secret" keys; two sensors establish a secure communication link if they share a secret key.

We consider large random intersection graphs, where $m, n \to \infty$. For $r^2 = o(m)$ the edge probability is

$$p_e = \mathbf{P}(v_i \sim v_j) = r^2 m^{-1} + O(r^4 m^{-2}).$$

We are interested in sparse graphs with bounded average degree

$$\eta := (n-1)p_e \approx (n-1)r^2 m^{-1}.$$

Paveiksliukas !

Asymptotic normality of $C_{\mathcal{G}} = 3N_{\Delta}/N_{\vee}$ (I)

Let $\mathcal{G}$ be an instance of $G(n, m, r)$. Denote

$$\mu = \frac{3\mathbf{E}N_{\Delta}}{\mathbf{E}N_{\vee}}, \quad N_{\Delta}^{*} = \frac{N_{\Delta} - \mathbf{E}N_{\Delta}}{\mathbf{E}N_{\Delta}}, \quad N_{\vee}^{*} = \frac{N_{\vee} - \mathbf{E}N_{\vee}}{\mathbf{E}N_{\vee}}.$$

We approximate $C_{\mathcal{G}}$ by μ. The approximation error

$$C_{\mathcal{G}} - \mu = \mu\left(\frac{N_{\Delta}^{*} + 1}{N_{\vee}^{*} + 1} - 1\right) = \mu(N_{\Delta}^{*} - N_{\vee}^{*}) - R \qquad (1)$$

has the leading term $\mu(N_{\Delta}^{*} - N_{\vee}^{*})$ and (negligible) remainder

$$R = \mu(N_{\Delta}^{*} - N_{\vee}^{*})\frac{N_{\vee}^{*}}{N_{\vee}^{*} + 1}.$$

Assume that $\mathbf{E}N_{\Delta}$, $\mathbf{E}N_{\vee} \to +\infty$ and properly standardized vector $\left(N_{\Delta}^{*}, N_{\vee}^{*}\right)$ is asymptotically (bivariate) normal. Then (1) implies the asymptotic normality of $C_{\mathcal{G}} - \mu$.

Asymptotic normality of $C_{\mathcal{G}} = 3N_\Delta/N_\vee$ (I)

Let $\mathcal{G}$ be an instance of $G(n, m, r)$. Denote

$$\mu = \frac{3\mathbf{E}N_\Delta}{\mathbf{E}N_\vee}, \quad N_\Delta^* = \frac{N_\Delta - \mathbf{E}N_\Delta}{\mathbf{E}N_\Delta}, \quad N_\vee^* = \frac{N_\vee - \mathbf{E}N_\vee}{\mathbf{E}N_\vee}.$$

We approximate $C_{\mathcal{G}}$ by μ. The approximation error

$$C_{\mathcal{G}} - \mu = \mu\left(\frac{N_\Delta^* + 1}{N_\vee^* + 1} - 1\right) = \mu(N_\Delta^* - N_\vee^*) - R \qquad (1)$$

has the leading term $\mu(N_\Delta^* - N_\vee^*)$ and (negligible) remainder

$$R = \mu(N_\Delta^* - N_\vee^*)\frac{N_\vee^*}{N_\vee^* + 1}.$$

Assume that $\mathbf{E}N_\Delta$, $\mathbf{E}N_\vee \to +\infty$ and properly standardized vector $(N_\Delta^*, N_\vee^*)$ is asymptotically (bivariate) normal. Then (1) implies the asymptotic normality of $C_{\mathcal{G}} - \mu$.

Remark. We have

$\mu = \mathbf{P}(v_2 \sim v_3 | v_1 \sim v_2, v_1 \sim v_3) = \mathbf{P}\left(v_2^* \sim v_3^* \middle| v_1^* \sim v_2^*, v_1^* \sim v_3^*\right)$, where the vertex triple (v_1^*, v_2^*, v_3^*) is drawn at random.

Bivariate asymptotic normality of $(N_\Delta, N_\vee)$.

Denote $\sigma_\Delta^2 = \mathbf{Var} N_\Delta$, $\sigma_\vee^2 = \mathbf{Var} N_\vee$, $\sigma_{\Delta\vee} = \mathbf{Cov}(N_\Delta, N_\vee)$.

Proposition. *Let $m, n \to +\infty$. Assume that $r \geq 2$ and $r = O(1)$. Assume that $\sigma_\Delta^2 \to +\infty$, $\sigma_\vee^2 \to +\infty$. Suppose that the ratio $\sigma_{\Delta\vee}/(\sigma_\Delta \sigma_\vee)$ converges to a limit. We denote the limit ρ. Then the random vector $\left(\sigma_\Delta^{-1}(N_\Delta - \mathbf{E}N_\Delta), \sigma_\vee^{-1}(N_\vee - \mathbf{E}N_\vee)\right)$ converges in distribution to a Gaussian random vector $(\xi_\Delta, \xi_\vee)$, where $\mathbf{E}\xi_j = 0$, $\mathbf{E}\xi_j^2 = 1$, $j = \Delta, \vee$, and $\mathbf{E}\xi_\Delta \xi_\vee = \rho$.*

Remark about the scaling of σ_Δ and $\sigma_\vee$. We have for $r^3 = o(m)$

$$\sigma_\Delta^2 \approx \frac{1}{2r^2} n\eta^3 + \frac{1}{6r} n\eta^2, \quad \sigma_\vee^2 \approx n\eta^3\left(2 + \frac{2}{r} + \frac{1}{2r^2}\right) + n\eta^2\left(\frac{1}{2} + \frac{1}{r}\right),$$

$$\sigma_{\Delta\vee} \approx n\eta^3\left(\frac{1}{2r^2} + \frac{1}{r}\right) + n\eta^2 \frac{1}{2r}.$$

Recall that $\eta = (n-1)p_e \approx nr^2 m^{-1}$ denotes the average degree. For fixed r and bounded $\eta \approx \eta_0 > 0$ we have

$$\rho \approx \frac{\eta_0(1+2r) + r}{\sqrt{(\eta_0 + 3^{-1}r)(\eta_0(4r^2+4r+1) + r^2 + 2r)}}.$$

Asymptotic normality of $C_{\mathcal{G}} = 3N_{\Delta}/N_{\vee}$ (II)

The bivariate asymptotic normality (Proposition)

$$\Big(\frac{N_{\Delta} - \mathbf{E}N_{\Delta}}{\sigma_{\Delta}^{-1}}, \frac{N_{\vee} - \mathbf{E}N_{\vee}}{\sigma_{\vee}^{-1}}\Big) \to (\xi_{\Delta}, \xi_{\vee})$$

implies the approximation (in distribution)

$$C_{\mathcal{G}} \approx \mu\Big(\frac{N_{\Delta} - \mathbf{E}N_{\Delta}}{\mathbf{E}N_{\Delta}} - \frac{N_{\vee} - \mathbf{E}N_{\vee}}{\mathbf{E}N_{\vee}}\Big) \approx \mu\Big(\frac{\sigma_{\Delta}}{\mathbf{E}N_{\Delta}}\xi_{\Delta} - \frac{\sigma_{\vee}}{\mathbf{E}N_{\vee}}\xi_{\vee}\Big) \quad (*)$$

Hence the asymptotic normality of $C_{\mathcal{G}}$.

Comment on the contribution of N_{Δ} and $N_{\vee}$ to (*):

• for $r = O(1)$ the coefficients $\sigma_{\Delta}/\mathbf{E}N_{\Delta}$ and $\sigma_{\vee}/\mathbf{E}N_{\vee}$ are of the same order of magnitude, thus, N_{Δ} and $N_{\vee}$ enter on equal terms;

• for $r \to +\infty$, $\eta = o(r)$ we have $\sigma_{\vee}/\mathbf{E}N_{\vee} = o\big(\sigma_{\Delta}/\mathbf{E}N_{\Delta}\big)$, that is, the triangle count N_{Δ} outplays $N_{\vee}$.

These observations follow easily from the relations

$$\mathbf{Var}\xi_{\Delta} = \mathbf{Var}\xi_{\vee} = 1, \quad \mu = 3\mathbf{E}N_{\Delta}/\mathbf{E}N_{\vee} \approx r^{-1},$$

$$\sigma_{\Delta}^2 \approx \frac{1}{2r^2}n\eta^3 + \frac{1}{6r}n\eta^2, \quad \sigma_{\vee}^2 \approx n\eta^3\Big(2 + \frac{2}{r} + \frac{1}{2r^2}\Big) + n\eta^2\Big(\frac{1}{2} + \frac{1}{r}\Big).$$

Approach to the asymptotic normality of $(N_\Delta, N_\vee)$.

It suffices to show that for any reals a, b the random variables

$$X_{n,m} = a\,\frac{N_\Delta - \mathbf{E}N_\Delta}{\sigma_\Delta} + b\,\frac{N_\vee - \mathbf{E}N_\vee}{\sigma_\vee}$$

converge in distribution to $a\xi_\Delta + b\xi_\vee$ as $n, m \to +\infty$.
To this aim we use Stein's method. In our case this approach leads to awkward calculations. To make calculations feasible we "preprocess" the subgraph counts

$$N_\Delta = N_\Delta(S_1, \ldots, S_n) = \sum_{\{i,j,k\}\subset[n]} \Delta_{i,j,k}, \tag{1}$$

$$N_\vee = N_\vee(S_1, \ldots, S_n) = \sum_{\{i,j,k\}\subset[n]} \left(\vee_{ijk} + \vee_{jki} + \vee_{kij}\right).$$

Namely, we decompose them into series of uncorrelated "polynomials" of increasing orded, called Hoeffding's decomposition. The decomposition helps to tackle statistical dependencies. In particular, we obtain reasonably simple expressions of $\mathbf{Var}(N_\Delta)$, $\mathbf{Var}(N_\vee)$ and $\mathbf{Cov}(N_\Delta, N_\vee)$.