Attribute fusion in a latent process model for time series of graphs

Minh Tang, Youngser Park, Nam H. Lee, and Carey E. Priebe

Appendix
Proofs of some stated results

Corollary 2: The maximizer of μ_λ also maximizes

$$\mu_\lambda^2 = \frac{\lambda^T \zeta \zeta^T \lambda}{\lambda^T \xi \lambda}$$

Because ξ is positive definite, there exists a positive definite matrix $\xi^{1/2}$ such that $\xi^{1/2}\xi^{1/2} = \xi$. Letting $\nu = \xi^{1/2}\lambda$, the above expression can be rewritten as

$$\mu_\lambda^2 = \frac{\nu^T \xi^{-1/2} \zeta \zeta^T \xi^{-1/2} \nu}{\nu^T \nu}$$

The claim then follows directly from the Rayleigh-Ritz theorem for Hermitian matrices. □

Lemma 3: $\tau_\lambda(G)$ is a U-statistics with kernel function $h(Y_1, Y_2, Y_3) = Y_1 Y_2 Y_3$. By the theory of U-statistics, we know that

$$\frac{\tau_\lambda(G) - \mathbb{E}[\tau_\lambda^*(G)]}{\sqrt{\mathrm{Var}[\tau_\lambda^*(G)]}} \xrightarrow{\mathrm{d}} N(0,1) \tag{1}$$

provided that $\mathrm{Var}[\tau_\lambda(G) - \tau_\lambda^*(G)] = o(\mathrm{Var}[\tau_\lambda^*(G)])$.

By the independent edge assumption, we have

$$\mathbb{E}[h(Y_i, Y_j, Y_k) = \mathbb{E}[Y_i]\mathbb{E}[Y_j]\mathbb{E}[Y_k] \tag{2}$$

$$\mathbb{E}[h(Y_i, Y_j, Y_k) | Y_i] = Y_i \mathbb{E}[Y_j]\mathbb{E}[Y_k]. \tag{3}$$

Thus, for $t < t^*$, we have $\mathbb{E}[\tau_\lambda^*(G(t))] = \binom{n}{3}\langle \lambda, \pi_{00}\rangle^3$ and

$$\begin{aligned} \mathrm{Var}[\tau_\lambda^*(G(t))] &= \mathrm{Var}\Big[\sum_{\{u,v,w\}} Y_{uv}\mathbb{E}[Y_{uw}]\mathbb{E}[Y_{vw}]\Big] \\ &= (n-2)^2 \langle \lambda, \pi_{00}\rangle^4 \mathrm{Var}\Big[\sum_{\{u,v\}} Y_{uv}\Big] \\ &= (n-2)^2 \langle \lambda, \pi_{00}\rangle^4 \tbinom{n}{2}\langle \lambda, \eta_{00}\lambda\rangle. \end{aligned} \tag{4}$$

We now sketch the derivation of $\mathrm{Var}[\tau_\lambda^*(G(t))]$ for $t = t^*$. We partition the set $\{u,v\} \in \binom{V}{2}$ into the sets

$$\begin{aligned} \mathscr{S}_1 &= \{u, v \in [m]\}, \\ \mathscr{S}_2 &= \{u \in [m], v \in [n] \setminus [m]\}, \\ \mathscr{S}_3 &= \{u, v \in [n] \setminus [m]\}. \end{aligned}$$

We can thus decompose $\mathrm{Var}[\tau_\lambda^*(G(t))]$ as

$$\begin{aligned} \mathrm{Var}[\tau_\lambda^*(G(t))] = S_1^2 \mathrm{Var}[\sum_{\{u,v\}\in\mathscr{S}_1} Y_{uv}] + S_2^2 \mathrm{Var}[\sum_{\{u,v\}\in\mathscr{S}_2} Y_{uv}] \\ + S_3^2 \mathrm{Var}[\sum_{\{u,v\}\in\mathscr{S}_3} Y_{uv}] \end{aligned} \tag{5}$$

Now, for $\{u,v\} \in \mathscr{S}_1$, we have

$$\begin{aligned} S_1 Y_{uv} &= \sum_{w \neq u,v} \mathbb{E}[h(Y_{uv}, Y_{uw}, Y_{vw}) \,|\, Y_{uv}] \\ &= ((m-2)\langle \lambda, \pi_{11}\rangle^2 + (n-m)\langle \lambda, \pi_{10}\rangle^2) Y_{uv}. \end{aligned} \tag{6}$$

The above expression is reasoned as follows. If $w \in [m]$, then $\mathbb{E}[Y_{uw}] = \mathbb{E}[Y_{vw}] = \langle \lambda, \pi_{11}\rangle$ and there are $m-2$ possible choices for $w \in [m]$ different from u and v. If $w \in [n] \setminus [m]$, then $\mathbb{E}[Y_{vw}] = \mathbb{E}[Y_{uw}] = \langle \lambda, \pi_{10}\rangle$ and there are $n-m$ possible choices for w. Analogous reasoning gives the expressions for S_2 and S_3 in the statement of the lemma.

We also have

$$\mathrm{Var}[Y_{uv}] = \begin{cases} \langle \lambda, \eta_{00}\lambda\rangle & \text{if } \{u,v\} \in \mathscr{S}_1 \\ \langle \lambda, \eta_{01}\lambda\rangle & \text{if } \{u,v\} \in \mathscr{S}_2 \\ \langle \lambda, \eta_{11}\lambda\rangle & \text{if } \{u,v\} \in \mathscr{S}_3 \end{cases}. \tag{7}$$

and thus

$$\begin{aligned} \mathrm{Var}[\tau_\lambda^*(G(t))] = \tbinom{m}{2}\langle \lambda, \eta_{00}\lambda\rangle S_1^2 + m(n-m)\langle \lambda, \eta_{01}\lambda\rangle S_2^2 \\ + \tbinom{n-m}{2}\langle \lambda, \eta_{00}\lambda\rangle S_3^2 \end{aligned}$$

as desired. To complete the proof one must show that $\mathrm{Var}[\tau_\lambda(G) - \tau_\lambda^*(G)] = o(\mathrm{Var}[\tau_\lambda^*(G)])$ and this follows directly from the argument in [1] or [2]. □

Proposition 5: Let $v \in V(t)$ and denote by $d_\lambda(v;t)$ the (fused) degree of vertex v, i.e.,

$$d_\lambda(v;t) = \sum_{w \in N(v)} \langle \lambda, \Gamma_{vw}\rangle.$$

For $t < t^*$, each of the Γ_{vw} is a multinomial trial with probability vector π_{00}. The following statements are made as $n \to \infty$ for fixed K. By the central limit theorem, we have

$$\frac{d_\lambda(v;t) - (n-1)\langle \lambda, \pi_{00}\rangle}{\sqrt{(n-1)\langle \lambda, \eta_{00}\lambda\rangle}} \xrightarrow{\mathrm{d}} \mathscr{N}(0,1). \tag{8}$$

We can thus consider the degree sequence of $G(t)$ for $t < t^*$ as a sequence of *dependent* normally distributed random variables. By an argument analogous to the argument for Erdös-Renyi random graphs in [3, §III.1] we can show that the dependency among the $\{d_\lambda(v;t)\}_{v \in V(t)}$ can be ignored. Another way of doing this is to note that the covariance between X_u and X_v, where X_u and X_v are the ratio in Eq. (8) for vertices u and v, is given by

$$r = \mathrm{Cov}(X_u, X_v) = \frac{3\langle \lambda, \pi_{00}\rangle}{\sqrt{(n-1)\langle \lambda, \eta_{00}\lambda\rangle}}. \tag{9}$$

Because $r \log n \to 0$ as $n \to \infty$, the sample maximum of the X_u converges to the sample maximum of a sequence of *independent* $\mathcal{N}(0,1)$ random variables. $d_\lambda(v;t)$, can thus be considered as a sequence of independent random variables from a normal distribution. It is well known that the sample maximum of standard normal random variables converges weakly to a Gumbel distribution [4, §2.3]. It is, however, not clear whether the convergence of $\Delta_\lambda(t)$ to a Gumbel distribution continues to hold under the composition of weak convergence as outlined above. We avoid this problem by showing directly that

$$\mathbb{P}\Big(\tfrac{\Delta_\lambda(t)-(n-1)\langle\lambda,\pi_{00}\rangle}{\sqrt{(n-1)\langle\lambda,\eta_{00}\lambda\rangle}} \leq a_n + b_n x\Big) \to e^{-e^{-x}}. \tag{10}$$

Let $\zeta_v = \frac{d_\lambda(v;t)-(n-1)\langle\lambda,\pi_{00}\rangle}{\sqrt{(n-1)\langle\lambda,\eta_{00}\lambda\rangle}}$ and $F_n(u) = \mathbb{P}(\zeta_v \leq u)$. If $n \to \infty$ and $u = O(\sqrt{\log n})$, we have the following moderate deviations result [5], [6, Theorem 2, §XVI.7].

$$\frac{1-F_n(u)}{1-\Phi(u)} = \Big[1 + (C\tfrac{u^3}{\sqrt{n}}) + O(\tfrac{u^6}{n})\Big] \tag{11}$$

for some constant C. Letting $u_n = a_n + b_n x$ in Eq. (11), we have

$$\begin{aligned}
F_n(u_n) &= 1 - (1-\Phi(u_n))(1 + C\tfrac{u_n^3}{\sqrt{n}} + O(\tfrac{u_n^6}{n})) \\
&= \Phi(u_n) + (1-\Phi(u_n))(C\tfrac{u_n^3}{\sqrt{n}} + O(\tfrac{u_n^6}{n})) \\
&= \Phi(u_n) + O(\tfrac{1}{u_n n^{1-\delta}})(C\tfrac{u_n^3}{\sqrt{n}} + O(\tfrac{u_n^6}{n})) \\
&= \Phi(u_n) + O(\tfrac{u_n^5}{n^{3/2-\delta}})
\end{aligned}$$

for some sufficiently small $\delta > 0$. We therefore have

$$\begin{aligned}
\mathbb{P}(\max_{v\in[n]} \zeta(v) \leq u_n) &= (F_n(u_n))^n \\
&= \Big[\Phi(u_n) + O(\tfrac{u_n^5}{n^{3/2-\delta}})\Big]^n \\
&= (\Phi(u_n))^n + O(\tfrac{u_n^5}{n^{1/2-\delta}}) \\
&\to e^{-e^{-x}}.
\end{aligned} \tag{12}$$

Eq. (10) is established and we obtain the limiting Gumbel distribution for $\Delta_\lambda(t)$ for $t < t^*$.

The case when $t = t^*$ can be derive in a similar manner. We first show that if $m = \Omega(\sqrt{n \log n})$ then $\Delta_\lambda(v;t^*) \xrightarrow{\mathrm{d}} \max_{v\in[m]} d_\lambda(v;t^*)$ [7, Lemma 3.1]. We then show, again by the central limit theorem, that for $v \in [m]$, $\frac{d_\lambda(v;t^*)-\mu_2}{\sigma_2} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. It then follows, similar to our previous reasoning for the case where $t < t^*$, that $\max_{v\in[m]} \frac{d_\lambda(v;t^*)-\mu_2}{\sigma_2} \xrightarrow{\mathrm{d}} \mathcal{G}(a_m, b_m)$ and we obtain the limiting Gumbel distribution for $\Delta_\lambda(t)$ for $t = t^*$. □

Theorem 6: Let $X \sim \mathcal{G}(\alpha,\beta)$. We consider the normalization $\frac{X-\mu}{\sigma}$. We have

$$\begin{aligned}
\mathbb{P}\Big[\tfrac{X-\mu}{\sigma} \leq z\Big] &= \mathbb{P}[X \leq z\sigma + \mu] = e^{-e^{-(z\sigma+\mu-\alpha)/\beta}} \\
&= e^{-e^{-(z-(\alpha-\mu)/\sigma)/(\beta/\sigma)}}.
\end{aligned}$$

Thus, $\frac{X-\mu}{\sigma} \sim \mathcal{G}(\frac{\alpha-\mu}{\sigma}, \frac{\beta}{\sigma})$. Because the sample mean and the sample variance are consistent estimators, the claim follows after an application of Slutsky's theorem. □

Lemma 7: Let $\phi_\lambda(v;t) = \psi_\lambda(v;t) - d_\lambda(v;t)$ be the (fused) locality statistics for vertex v at time t not including the (fused) degree of v, i.e.,

$$\phi_\lambda(v;t) = \sum_{\substack{u\,w\in N(v)\\ u,w\neq v}} \langle \lambda, \Gamma_{uw}\rangle. \tag{13}$$

The following statements are conditional on $|N(v)| = l$. First of all, we have

$$\phi_\lambda(v;t) = \sum_{k=1}^{K} \lambda_k z_k$$

where the $(z_1,\dots,z_K)$ are distributed as

$$(z_1, z_2, \dots, z_K) \sim \mathrm{multinomial}\Big(\tbinom{l}{2}, \pi_{00}\Big).$$

By the central limit theorem, we have

$$\frac{\phi_\lambda(v;t) - \binom{l}{2}\langle\lambda,\pi_{00}\rangle}{\sqrt{\binom{l}{2}\langle\lambda,\eta_{00}\lambda\rangle}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Let $\lambda^{(2)}$ be the element-wise square of λ. Define C_{00} and p_{00} to be

$$C_{00} = \frac{\langle\lambda^{(2)},\pi_{00}\rangle}{\langle\lambda,\pi_{00}\rangle}, \quad p_{00} = \frac{(\langle\lambda,\pi_{00}\rangle)^2}{\langle\lambda^{(2)},\pi_{00}\rangle}. \tag{14}$$

We note that $p_{00} \in [0,1]$. Now let $Y_l = C_{00}\mathrm{Bin}(\binom{l}{2}, p_{00})$. Then $\mathbb{E}[Y_l] = \binom{l}{2}\langle\lambda,\pi_{00}\rangle$ and $\mathrm{Var}[Y_l] = \binom{l}{2}\langle\lambda,\eta_{00}\lambda\rangle$ and again by the central limit theorem, we have

$$\frac{\psi_\lambda(v;t) - \binom{l}{2}\langle\lambda,\pi_{00}\rangle}{\sqrt{\binom{l}{2}\langle\lambda,\eta_{00}\lambda\rangle}} \xrightarrow{\mathrm{d}} \frac{Y_l - \binom{l}{2}\langle\lambda,\pi_{00}\rangle}{\sqrt{\binom{l}{2}\langle\lambda,\eta_{00}\lambda\rangle}}. \tag{15}$$

Eq. (15) states that the locality statistics for our attributed random graphs model with $t < t^*$ can be approximated by the locality statistics for an Erdös-Renyi graph with edge probability p_{00}. The lemma then follows from Theorem 1.1 in [7]. □

Lemma 8: For ease of exposition we drop the index t^* from our discussion. Let $\phi_\lambda(v) = \psi_\lambda(v) - d_\lambda(v)$. Let $M(v)$ be the number of neighbors of v that lies in $[m]$ and $W(v)$ be the number of neighbors of v that lies in $[n]\setminus[m]$. The following statements are conditional on $M(v) = l_\zeta$ and $W(v) = l_\xi$. We have

$$\phi_\lambda(v) = \sum_{k=1}^{K} \lambda_k (y_k^{(\zeta)} + y_k^{(\xi)} + y_k^{(\omega)}) \tag{16}$$

where $(y_1^{(\zeta)},\dots,y_K^{(\zeta)})$, $(y_1^{(\xi)},\dots,y_K^{(\xi)})$, $(y_1^{(\omega)},\dots,y_K^{(\omega)})$ are distributed as

$$(y_1^{(\zeta)},\dots,y_K^{(\zeta)}) \sim \text{multinomial}\Big(\tbinom{l_\zeta}{2},\pi_{11}\Big)$$
$$(y_1^{(\xi)},\dots,y_K^{(\xi)}) \sim \text{multinomial}\Big(\tbinom{l_\xi}{2},\pi_{00}\Big)$$
$$(y_1^{(\omega)},\dots,y_m^{(\omega)}) \sim \text{multinomial}\Big(l_\zeta l_\xi,\pi_{10}\Big).$$

Let ρ and ς be defined as

$$\rho = \langle \lambda, \tbinom{l_\zeta}{2}\pi_{11} + \tbinom{l_\xi}{2}\pi_{00} + l_\zeta l_\xi \pi_{10}\rangle$$
$$\varsigma = \langle \lambda, \Big(\tbinom{l_\zeta}{2}\eta_{11} + \tbinom{l_\xi}{2}\eta_{00} + l_\zeta l_\xi \eta_{10}\Big)\lambda.\rangle$$

By the central limit theorem, as $l_\zeta \to \infty$ and $l_\xi \to \infty$

$$\frac{\phi_\lambda(v)-\rho}{\varsigma} \xrightarrow{\text{d}} \mathcal{N}(0,1) \tag{17}$$

Let $\lambda^{(2)}$ be the element-wise square of λ. Define C_{00}, C_{01}, C_{11} and p_{00}, p_{01}, p_{11} to be

$$C_{00} = \tfrac{\langle\lambda^{(2)},\pi_{00}\rangle}{\langle\lambda,\pi_{00}\rangle}; \quad p_{00} = \tfrac{(\langle\lambda,\pi_{00}\rangle)^2}{\langle\lambda^{(2)},\pi_{00}\rangle} \tag{18}$$
$$C_{11} = \tfrac{\langle\lambda^{(2)},\pi_{11}\rangle}{\langle\lambda,\pi_{11}\rangle}; \quad p_{11} = \tfrac{(\langle\lambda,\pi_{11}\rangle)^2}{\langle\lambda^{(2)},\pi_{11}\rangle} \tag{19}$$
$$C_{10} = \tfrac{\langle\lambda^{(2)},\pi_{11}\rangle}{\langle\lambda,\pi_{10}\rangle}; \quad p_{10} = \tfrac{(\langle\lambda,\pi_{10}\rangle)^2}{\langle\lambda^{(2)},\pi_{10}\rangle} \tag{20}$$

We note that p_{00}, p_{01}, and p_{11} are all elements of $[0,1]$. Now let $Y_\zeta \sim C_{11}\text{Bin}\Big(\tbinom{l_\zeta}{2},p_{11}\Big)$, $Y_\xi \sim C_{00}\text{Bin}\Big(\tbinom{l_\xi}{2},p_{00}\Big)$ and $Y_\omega \sim C_{10}\text{Bin}\Big(l_\zeta l_\xi, p_{10}\Big)$. We also set $Y = Y_\zeta + Y_\xi + Y_\omega$. By the central limit theorem, we have

$$\frac{\phi_\lambda(v)-\rho}{\varsigma} \xrightarrow{\text{d}} \frac{Y-\rho}{\varsigma}. \tag{21}$$

Eq. (21) states that the locality statistics $\phi_\lambda(v)$ for our attributed random graphs model at time $t = t^*$ can be approximated by the locality statistics $Y(v)$ for an unattributed kidney and egg model. The limiting distribution for the scan statistics in unattributed kidney-egg graphs had previously been considered in [7]. We provided a sketch of the arguments from [7] below, along with some minor changes to handle the case where the probability of kidney-kidney and kidney-egg connections are different.

Let G be an instance of $\kappa(n,m,p_{11},p_{10},p_{00})$, an unattributed kidney-egg graph with the probability of egg-egg, egg-kidney, and kidney-kidney connections being p_{11}, p_{10}, and p_{00}, respectively. $D(v) = M(v) + W(v)$ is then the degree of v in G. We now show two inequalities relating the tail distribution of $\Delta(G)$ and $\Upsilon(G) = \max_{v\in V(G)} Y(v)$.

$$\limsup \mathbb{P}(\Upsilon(G) \geq a_{n,m}) \leq \lim \mathbb{P}(\Delta(G) \geq N_\kappa), \tag{22}$$
$$\liminf \mathbb{P}(\Upsilon(G) \geq a_{n,m}) \geq \lim \mathbb{P}(\Delta(G) \geq N_\kappa). \tag{23}$$

Eq. (22): Let $C^* = \max\{C_{11}, C_{10}, C_{00}\}$ and $d_{n,m} = \sqrt{2a_{n,m}/C^*}$. We first note that

$$\Upsilon(G) \geq a_{n,m} \Rightarrow C^*\tbinom{D(v)}{2} \geq a_{n,m} \Rightarrow D(v) \geq d_{n,m}.$$

Let us define $h(v) = \mathbb{E}[Y(v)]$, i.e.,

$$\begin{aligned} h(v) = C_{00}p_{00}\tbinom{D(v)}{2} &+ (C_{11}p_{11} - C_{00}p_{00})\tbinom{M(v)}{2} \\ &+ (C_{10}p_{10} - C_{00}p_{00})M(v)W(v). \end{aligned}$$

We then have

$$\begin{aligned} \mathbb{P}(\Upsilon(G) \geq a_{n,m}) &= \mathbb{P}\Big(\bigcup_{v\in V(G)} Y(v) \geq a_{n,m}\Big) \\ &= \mathbb{P}\Big(\bigcup_{v\in V(G)} Y(v) \geq a_{n,m}, D(v) \geq d_{n,m}\Big) \\ &\leq P_1 + P_2 \end{aligned}$$

where

$$\vartheta_n = C_{00}\Big[\tbinom{n}{2}p_{00}(1-p_{00})\Big]^{1/2}\log n$$
$$P_1 = \mathbb{P}(\bigcup_{v\in V(G)} D(v) \geq d_{n,m}, h(v) \geq a_{n,m} - \vartheta_n)$$
$$P_2 = \mathbb{P}(\bigcup_{v\in V(G)} D(v) \geq d_{n,m}, Y(v) - h(v) \geq \vartheta_n).$$

We now show that P_2 is negligible as $n \to \infty$. To proceed, let A be the event $\{M(v) = e, W(v) = f\}$ and let $p_{e,f} = \mathbb{P}(A)$. P_2 can then be bounded as follows

$$\begin{aligned} \frac{P_2}{n} &\leq \sum_{e+f\geq d_{n,m}} \mathbb{P}(Y(v) - h(v) \geq \vartheta_n \,|\, A)p_{e,f} \\ &= \sum_{e+f\geq d_{n,m}} \mathbb{P}\Big(\tfrac{Y(v)-h(v)}{\text{Var}[Y(v)]^{1/2}} \geq \tfrac{\vartheta_n}{\text{Var}[Y(v)]^{1/2}} \Big| A\Big)p_{e,f} \\ &\leq \sum_{e+f\geq d_{n,m}} (1+o(1))\mathbb{P}(Z \geq \Theta(\log n))p_{e,f} \\ &= o(n^{-1}). \end{aligned}$$

We now consider P_1. We note that $P_1 \leq R_1 + R_2$ where

$$R_1 = \mathbb{P}\Big(\bigcup_{v\in[m]} D(v) \geq d_{n,m}, h(v) \geq a_{n,m} - \vartheta_n\Big),$$
$$R_2 = \mathbb{P}\Big(\bigcup_{v\in[n]\setminus[m]} D(v) \geq d_{n,m}, h(v) \geq a_{n,m} - \vartheta_n\Big).$$

Let us define $g(v) = h(v) - C_{00}p_{00}\tbinom{D(v)}{2}$. R_1 is then bounded as follows

$$\begin{aligned} R_1 &\leq \mathbb{P}\Big(\bigcup_{v\in[m]} h(v) \geq a_{n,m} - \vartheta_n\Big) \\ &\leq \mathbb{P}\Big(\bigcup_{v\in[m]} D(v) \geq \sqrt{\tfrac{2(a_{n,m}-\vartheta_n-g(v))}{C_{00}p_{00}}}\Big). \end{aligned} \tag{24}$$

We now consider the term $a_{n,m} - g(v)$. We have

$$\begin{aligned} a_{n,m} - g(v) = C_{00}p_{00}\tbinom{N_\kappa}{2} &+ (C_{11}p_{11} - C_{00}p_{00})(\tbinom{\mu_E}{2} - \tbinom{M(v)}{2}) \\ &+ (C_{10}p_{10} - C_{00}p_{00})(\mu_E\mu_F - M(v)W(v)). \end{aligned}$$

Let $\mathfrak{E}$ and $\mathfrak{F}$ be sets of vertices defined by

$$\mathfrak{E} = \{v : |M(v) - \mu_E| \leq \sigma_E \log m\} \tag{25}$$
$$\mathfrak{F} = \{v : |W(v) - \mu_F| \leq \sigma_F \log(n-m)\}. \tag{26}$$

Then we have, for $v \in \mathfrak{E}\cap\mathfrak{F}$

$$\begin{aligned} a_{n,m} - g(v) = C_{00}p_{00}\tbinom{N_\kappa}{2} &+ \Theta(m^{3/2}\log m) \\ &+ \Theta(m\sqrt{n-m}) \end{aligned} \tag{27}$$

When $m = \Omega(\sqrt{n \log n})$, Eq. (27) gives

$$a_{n,m} - g(v) = N_\kappa^2\Big(\tfrac{C_{00}p_{00}}{2} + O(n^{-1/2-a}\log n)\Big). \quad (28)$$

for some $a > 0$. The set $\{v \in [m]\}$ can be partition into $\{v \in [m] \cap (\mathfrak{E} \cap \mathfrak{F})\}$ and $\{v \in [m] \setminus (\mathfrak{E} \cap \mathfrak{F})\}$. We can show that $\mathbb{P}\{v \in [m] \setminus (\mathfrak{E} \cap \mathfrak{F})\} = o(1)$ by using a concentration inequality, e.g., Hoeffding's bound. We thus have

$$\begin{aligned} R_1 &\leq \mathbb{P}\Big(\bigcup_{\substack{v\in[m]\\ v\in\mathfrak{E}\cap\mathfrak{F}}} D(v) \geq N_\kappa\sqrt{1 + O(\tfrac{\log n}{n^{1/2+a}})}\Big) + o(n^{-1}) \\ &= \mathbb{P}\Big(\bigcup_{v\in[m]} D(v) \geq N_\kappa + O(n^{1/2-a}\log n)\Big) + o(n^{-1}) \quad (29) \\ &= \mathbb{P}\Big(\Delta \geq \mu_{E+F} + \sigma_{E+F}(z_m + O(\tfrac{\log n}{n^a}))\Big) + o(n^{-1}) \\ &\to \mathbb{P}(\Delta \geq N_\kappa). \end{aligned}$$

The same argument can be applied to R_2 to show that

$$R_2 \leq \mathbb{P}\Big(\bigcup_{v\in[n]\setminus[m]} D(v) \geq N_\kappa(1 + o(1))\Big) = o(1). \quad (30)$$

Eq. (22) is therefore established. □

Eq. (23)*:* We start by noting that

$$\begin{aligned} \mathbb{P}(\Upsilon(G) \geq a_{n,m}) = \mathbb{P}\Big(\bigcup_{v\in[n]} Y(v) \geq a_{n,m}\Big) &\geq \mathbb{P}\Big(\bigcup_{v\in[m]} Y(v) \geq a_{n,m}, D(v) \geq N_\kappa\Big) \\ &\geq \mathbb{P}\Big(\bigcup_{v\in[m]} D(v) \geq N_\kappa\Big) \\ &\quad - \mathbb{P}\Big(\bigcup_{v\in[m]} Y(v) < a_{n,m}, D(v) \geq N_\kappa\Big). \end{aligned}$$

We now show that $\mathbb{P}(\cup_{v\in[m]} Y(v) < a_{n,m}, D(v) \geq N_\kappa) \to 0$ as $n \to \infty$. Let $v \in [m]$ be arbitrary. It is then sufficient to show that $m\mathbb{P}(Y(v) < a_{n,m}, D(v) \geq N_\kappa) = o(1)$. We note that $\mathbb{P}(Y(v) < a_{n,m}, D(v) \geq N_\kappa)$ can be rewritten as

$$\sum_{e+f\geq N_\kappa} \mathbb{P}(Y(v) \leq a_{n,m} \,|\, M(v) = e, W(v) = f)p_{e,f}. \quad (31)$$

We now split the indices set $e + f \geq N_\kappa$ in Eq. (31) into three parts S_1, S_2 and S_3, namely

$$S_1 = \{e \geq \mu_E + \sigma_E \log m\} \quad (32)$$
$$S_2 = \{e \leq \mu_E + \sigma_E \log m, e + f \leq N_\kappa + \varphi(n)\} \quad (33)$$
$$S_3 = \{e \leq \mu_E + \sigma_E \log m, e + f \geq N_\kappa + \varphi(n)\} \quad (34)$$

where $\varphi(n) = \Theta(n^{1/2-a})$ for some $a > 0$. We can then show that $m\mathbb{P}(M(v) = e, W(v) = f, \{e,f\} \in S_1) = o(1)$ by applying a concentration inequality. Similarly, $e + f \geq N_\kappa$ and $e \leq \mu_E + \sigma_E \log m$ implies that

$$f \geq \mu_F + (z_m - o(1))\sigma_F \quad (35)$$

and once again, by a concentration inequality, we can show that $m\mathbb{P}(M(v) = e, W(v) = f, \{e,f\} \in S_2) = o(1)$. As for S_3, from the fact that $e + f \geq N_\kappa + \varphi(n)$, we have the bound

$$\begin{aligned} a_{n,m} - h(v) \leq (C_{11}p_{11} - C_{10}p_{10})[m\sigma_E \log m + \tbinom{\log m}{2}] \\ - C_{00}p_{00}N_\kappa\varphi(n). \end{aligned} \quad (36)$$

As $\mathrm{Var}[Y(v)] = \Theta(N_\kappa)$ for $\{M(v), W(v)\} \in S_3$, we have

$$\begin{aligned} p_{S_3} &= \sum_{\{e,f\}\in S_3} \mathbb{P}(Y(v) < a_{n,m})p_{e,f} \\ &\leq \sum_{\{e,f\}\in S_3} \mathbb{P}\Big(\tfrac{Y(v)-h(v)}{\mathrm{Var}[Y(v)]^{1/2}} \leq \tfrac{a_{n,m}-h(v)}{\mathrm{Var}[Y(v)]^{1/2}}\Big)p_{e,f} \quad (37) \\ &\leq \sum_{\{e,f\}\in S_3} \mathbb{P}\Big[Z \leq O\big(\tfrac{m^{3/2}\log m}{N_\kappa} - \varphi(n)\big)\Big]p_{e,f}. \end{aligned}$$

We now set $a = \frac{1}{2(k+1)}$. Then for $m = O(n^{k/(k+1)})$ and $\varphi(n) = O(n^{1/2-a})$ we have

$$\frac{m^{3/2}\log m}{N_\kappa} - \varphi(n) = -O(n^{k/2(k+1))}) \quad (38)$$

which then implies

$$mp_{S_3} \leq m\sum_{\{e,f\}\in S_3} \mathbb{P}\Big[Z \leq -O(n^{k/2(k+1)}\Big]p_{e,f} = o(1). \quad (39)$$

Thus $\mathbb{P}(Y(v) < a_{n,m}, D(v) \geq N_\kappa) \to 0$ as desired. □

From Eq. (22) and Eq. (23), we have

$$\lim \mathbb{P}(\Upsilon(G) \geq a_{n,m}) = \lim \mathbb{P}(\Delta(G) \geq N_\kappa). \quad (40)$$

Let $N_{\kappa,y} = N_\kappa + y\frac{\sigma_{E+F}}{\sqrt{2\log m}}$. We now define $a_{n,m,y}$ as

$$\langle\lambda, \pi_{00}\rangle\tbinom{N_{\kappa,y}}{2} + \langle\lambda, \pi_{11} - \pi_{00}\rangle\tbinom{\mu_E}{2} + \langle\lambda, \pi_{10} - \pi_{00}\rangle\mu_E\mu_F.$$

The above expression is equal to

$$a_{n,m} + \langle\lambda, \pi_{00}\rangle y\frac{\sigma_{E+F}}{\sqrt{2\log m}}\Big(N_\kappa + y\frac{\sigma_{E+F}^2}{2\sqrt{2\log m}} + O(1)\Big). \quad (41)$$

We thus have

$$a_{n,m,y} = a_{n,m} + (y + o(1))b_{n,m}.$$

We therefore have

$$\begin{aligned} \lim \mathbb{P}(\Upsilon(G) \geq a_{n,m,y}) &= \lim \mathbb{P}\Big(\frac{\Upsilon(G) - a_{n,m}}{b_{n,m}} \geq y\Big) \\ &= \lim \mathbb{P}(\Delta(G) \geq N_{\kappa,y}) \\ &= \lim \mathbb{P}\Big(\frac{\Delta(G) - N_\kappa}{\sigma_{E+F}} \geq \frac{y}{\sqrt{2\log m}}\Big). \end{aligned}$$

Because $\Delta(G)$ converges weakly to a Gumbel distribution in the limit ([3], [7]), we have

$$\mathbb{P}\Big(\frac{\Upsilon(G) - a_{n,m}}{b_{n,m}} \leq y\Big) \to e^{-e^{-y}}. \quad (42)$$

□

References

[1] K. Nowicki and J. C. Wierman, "Subgraph counts in random graphs using incomplete U-statistics methods," *Discrete Mathematics*, vol. 72, pp. 299–310, 1988. 1

[2] A. Rukhin, "Asymptotic analysis of various statistics for random graph inference," Ph.D. dissertation, Johns Hopkins University, 2009. 1

[3] B. Bollobás, *Random Graphs*. Academic Press, 1985. 1, 4

[4] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*. John Wiley & Sons, 1987. 2

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. John Wiley and Sons, 1971. 2

[6] H. Rubin and J. Sethuraman, "Probabilities of moderate deviations," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 27, pp. 325–346, 1965. 2

[7] A. Rukhin and C. E. Priebe, "On the limiting distribution of a graph scan statistic," *Communications in Statistics: Theory and Methods* (in press). 2, 3, 4