به نام خداوند بخشنده مهربان

شبکه‌های اقتصادی و اجتماعی

دانشکده مهندسی کامپیوتر

مریم رمضانی
بهار ۱۴۰۴

گراف‌ها و پیمایش تصادفی، تشخیص جوامع | **تمرین سوم** | تاریخ انتشار: ۱۷ اردیبهشت ۱۴۰۴

۱. سوالات خود در مورد این تمرین را در کوئرا مطرح کنید.

۲. لطفا پاسخ خود را با توضیحات کافی و با رسم شکل نگارش کنید.

سوالات تئوری (۷۰ نمره)

تاریخ تحویل: ۲۸ اردیبهشت ۱۴۰۴؛ با تاخیر: ۳۱ اردیبهشت ۱۴۰۴

پرسش ۱ (۱۰ نمره)

(آ) در مدل Erdös-Renyi با n راس و میانگین درجه λ نشان دهید که قطر گراف برابر است با $\frac{\log n}{\log \lambda}$.

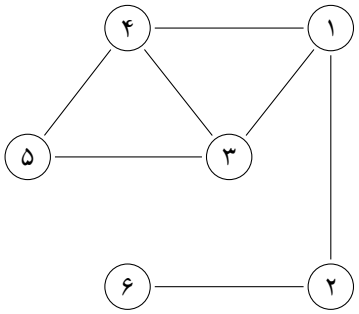
(ب) اگر مقدار λ کمتر از ۱ یا بیشتر از ۱ باشد قطر گراف را تحلیل کنید.

پاسخ

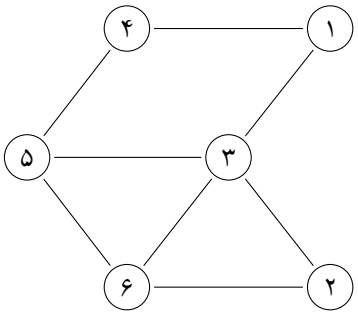
(آ) مدل Erdös-Rényi (ER) گرافی با n گره و میانگین درجه λ را ایجاد می‌کند. در این مدل، درجه هر گره از توزیع باینومیال $Bin(n-۱, p)$ پیروی می‌کند، که در آن p احتمال ایجاد یال بین هر دو گره است. بر اساس نتایج تحلیل‌های گراف تصادفی، قطر گراف در مدل ER تقریباً برابر با $\frac{\log n}{\log \lambda}$ است. این نتیجه به این دلیل حاصل می‌شود که با افزایش تعداد گره‌ها n، فاصله میان گره‌ها به طور نمایی کاهش می‌یابد.
اثبات: قطر گراف D در این مدل به طور تقریبی برابر با $\frac{\log n}{\log \lambda}$ می‌باشد. این رابطه از تحلیل‌های ریاضیاتی در خصوص ارتباطات میان گره‌ها به‌دست می‌آید، که نشان می‌دهد با افزایش تعداد گره‌ها در مدل ER و ثابت بودن میانگین درجه، فاصله‌های میان گره‌ها به طور قابل توجهی کوتاه‌تر می‌شود.

(ب) حال اگر مقدار λ کمتر از ۱ باشد، احتمال برقراری یال‌ها بین گره‌ها کاهش می‌یابد. این امر منجر به کاهش تراکم شبکه و در نتیجه افزایش قطر گراف می‌شود، چرا که گراف به صورت پراکنده‌تر و با پیوندهای ضعیف‌تر به وجود می‌آید. از طرف دیگر، در صورتی که λ بیشتر از ۱ باشد، تعداد یال‌ها به طور قابل توجهی افزایش می‌یابد و گراف به هم پیوسته‌تر می‌شود. در این حالت، قطر گراف کاهش یافته و ارتباطات سریع‌تر و نزدیک‌تر برقرار می‌شود.
نتیجه: تغییرات در مقدار λ تأثیر بسزایی بر قطر گراف دارند. در صورتی که λ کمتر از ۱ باشد، گراف پراکنده‌تر و قطر آن بزرگ‌تر خواهد بود. در مقابل، اگر λ بیشتر از ۱ باشد، گراف به هم پیوسته‌تر شده و قطر آن کاهش می‌یابد.

پرسش ۲ (۱۰ نمره) شباهت بین دو گراف A و B را با استفاده از تمام Graphlet Kernelsهای سه‌تایی محاسبه کنید.



A (آ)



B (ب)

پاسخ

برای محاسبه شباهت بین دو گراف A و B با استفاده از Graphlet Kernelsهای سه‌تایی، ابتدا باید تعداد و نوع گرافلت‌های سه‌تایی موجود در هر گراف را شمارش کنیم. گرافلت‌های سه‌تایی به تمام زیرگراف‌های ممکن از سه گره گفته می‌شود که در آن گره‌ها می‌توانند به‌طور متفاوتی به هم متصل شوند. در این سوال، گرافلت‌های سه‌تایی مورد نظر عبارتند از گرافلت‌های حلقه‌ای، ستاره‌ای و خطی. پس از شمارش گرافلت‌ها در هر گراف، شباهت بین دو گراف A و B از طریق فرمول Graphlet Kernel محاسبه می‌شود. این فرمول به صورت زیر تعریف می‌شود:

$$\mathrm{Sim}(A,B) = \frac{\sum_g \sqrt{f_A(g) f_B(g)}}{\|f_A\| \|f_B\|}$$

تعداد وقوع گرافلت‌ها در هر گراف به شرح زیر است: f_A(حلقه) $= ۲$، f_B(حلقه) $= ۲$ f_A(ستاره) $= ۳$، f_B(ستاره) $= ۴$ f_A(خطی) $= ۱$، f_B(خطی) $= ۱$

محاسبه نرمال‌سازی:

$$\|f_A\| = \sqrt{۲^۲ + ۳^۲ + ۱^۲} = \sqrt{۱۴}$$
$$\|f_B\| = \sqrt{۲^۲ + ۴^۲ + ۱^۲} = \sqrt{۲۱}$$

حال محاسبه شباهت:

$$\text{Sim}(A, B) = \frac{\sqrt{۲ \times ۲} + \sqrt{۳ \times ۴} + \sqrt{۱ \times ۱}}{\sqrt{۱۴} \times \sqrt{۲۱}}$$
$$\text{Sim}(A, B) = \frac{\sqrt{۴} + \sqrt{۱۲} + \sqrt{۱}}{\sqrt{۲۹۴}} = \frac{۳ + ۲\sqrt{۳}}{\sqrt{۲۹۴}} \approx ۰/۳۷۷$$

بنابراین، شباهت بین گراف‌های A و B برابر با تقریبا ۳۷۷.۰ است.

پرسش ۳ (۱۰ نمره) در یک شبکه اجتماعی، گره‌ها نشان‌دهنده افراد و یال‌ها نشان‌دهنده ارتباطات بین آن‌ها هستند. شما تصمیم دارید که ساختار این شبکه را مدل‌سازی کنید.

(آ) مدل‌های Erdös-Rényi و Watts-Strogatz را مقایسه کنید و توضیح دهید که کدام‌یک بیشتر شبیه یک شبکه اجتماعی واقعی است؟ چرا؟

(ب) مدل Kleinberg's geographical model چگونه می‌تواند توجیهی برای ارتباطات محلی و جهانی در شبکه‌های اجتماعی ارائه دهد؟ توضیح دهید که چگونه این مدل به پدیده "شش درجه جدایی" مرتبط است.

پاسخ

(آ) مدل Erdös-Rényi (ER) گرافی با N گره و احتمال p برای ایجاد یال بین هر دو گره می‌سازد. ویژگی این مدل داشتن توزیع درجه نرمال است، یعنی اکثر گره‌ها درجه مشابهی دارند. اما محدودیت آن این است که Hubهای واقعی که در شبکه‌های اجتماعی دیده می‌شود را تولید نمی‌کند.
مدل Watts-Strogatz (WS) ابتدا یک گراف حلقوی منظم می‌سازد و سپس برخی یال‌ها را تصادفی بازآرایی می‌کند. این مدل ویژگی خوشه‌بندی بالا و مسیرهای کوتاه دارد که مشابه شبکه‌های اجتماعی واقعی است. همچنین پدیده Small-World یعنی اتصالات محلی قوی و میانگین فاصله کوتاه بین گره‌ها را مدل‌سازی می‌کند. با این حال، برخلاف مدل Barabási-Albert، Hubهای واقعی ایجاد نمی‌کند.
نتیجه: مدل Watts-Strogatz به شبکه‌های اجتماعی واقعی شبیه‌تر است، چون هم خوشه‌بندی بالا دارد و هم مسیرهای کوتاه را بازتولید می‌کند، اما اگر بخواهیم هاب‌های قوی را هم مدل کنیم، مدل Barabási-Albert مناسب‌تر است.

(ب) مدل Kleinberg ترکیبی از اتصالات محلی و جهانی Long-range links ارائه می‌دهد. در این مدل، ابتدا یک گراف شبکه‌ای دوبعدی ساخته می‌شود که هر گره به همسایگان نزدیکش وصل است، سپس برخی اتصالات تصادفی طبق قانون توانی اضافه می‌شوند.
ارتباط با "شش درجه جدایی": این مدل نشان می‌دهد که اگر اتصالات تصادفی به‌طور هوشمندانه اضافه شوند، میانگین مسیر بین دو گره کاهش می‌یابد. بنابراین در شبکه‌های اجتماعی واقعی، تنها با چند مرحله (معمولاً حدود ۶ مرحله) می‌توان به هر فردی رسید.

پرسش ۴ (۱۰ نمره) در بسیاری از شبکه‌های واقعی، توزیع درجه گره‌ها از نوع distribution heavy-tailed است.

(آ) تفاوت بین توزیع نرمال و Power-law Distribution در شبکه‌ها چیست؟ چگونه می‌توان فهمید که یک شبکه از Power-law Distribution پیروی می‌کند؟

(ب) مدل‌های Barabási-Albert و Forest Fire را مقایسه کنید. چگونه هر یک از این مدل‌ها می‌توانند هسته‌های بزرگ در شبکه‌ها را توضیح دهند؟

(ج) چگونه مدل Zipf & Pareto می‌تواند توزیع ارتباطات در شبکه‌های اجتماعی را توجیه کند؟

پاسخ

(آ) توزیع نرمال (Normal Distribution) توزیعی است که بیشتر داده‌ها نزدیک به مقدار میانگین قرار دارند. شبکه‌هایی که از مدل Erdös-Rényi پیروی می‌کنند، درجه گره‌هایشان معمولاً دارای توزیع نرمال است.
در مقابل، توزیع قانون توانی (Power-law Distribution) به حالتی اشاره دارد که تعداد کمی از گره‌ها دارای درجه بسیار بالا (Hub) هستند و بیشتر گره‌ها درجه پایینی دارند. در شبکه‌های اجتماعی، اینترنت و شبکه‌های استنادی، معمولاً درجه گره‌ها از توزیع Power-law پیروی می‌کند.
برای تشخیص توزیع Power-law، نمودار Degree Distribution رسم می‌شود؛ اگر این نمودار در مقیاس لگاریتمی‌ـ‌لگاریتمی خطی باشد، نشان‌دهنده پیروی از قانون توانی است.

(ب) مدل Barabási-Albert (BA) بر اساس رشد تدریجی گراف و مکانیزم Preferential Attachment ساخته می‌شود؛ یعنی گره‌های جدید ترجیح می‌دهند به گره‌های پراتصال متصل شوند. این مدل Hubهای قوی ایجاد کرده و منجر به توزیع Power-law می‌شود.
مدل Forest Fire انتشار یال‌ها را مانند یک آتش‌سوزی شبیه‌سازی می‌کند، یعنی یک گره تصادفی انتخاب می‌شود و به احتمال خاصی به همسایگانش متصل می‌شود. این مدل مشابه BA است، اما امکان تشکیل جوامع محلی قوی‌تر را فراهم می‌کند.
نتیجه: مدل BA برای مدل‌سازی Hubهای قوی مناسب‌تر است، در حالی که مدل Forest Fire ساختارهای خوشه‌ای و اجتماع‌های شبکه‌ای را بهتر نمایش می‌دهد.

(ج) قانون Zipf's Law و اصل Pareto Principle (قانون ۲۰/۸۰) توضیح می‌دهند که در بسیاری از شبکه‌های اجتماعی تعداد کمی از کاربران بسیار فعال‌اند، در حالی‌که بیشتر کاربران فعالیت کمی دارند. برای مثال، در Twitter، ۲۰٪ کاربران حدود ۸۰٪ محتوا را تولید می‌کنند.
نتیجه: این اصل نشان می‌دهد که چرا برخی Influencerها در شبکه‌های اجتماعی بسیار تأثیرگذارند و نقش کلیدی در انتشار اطلاعات دارند.

پرسش ۵ (۱۰ نمره) شما مسئول طراحی یک الگوریتم برای جستجوی اطلاعات در یک شبکه استنادی علمی هستید، که در آن گره‌ها مقالات علمی و یال‌ها استنادات هستند.

(آ) توضیح دهید که چگونه می‌توان از الگوریتم PageRank یا HITS برای یافتن مقالات مهم در این شبکه استفاده کرد؟ تفاوت این دو الگوریتم چیست؟

(ب) مدل Kronecker Graph Model چگونه می‌تواند برای پیش‌بینی ارتباطات آینده در یک شبکه مورد استفاده قرار گیرد؟

(ج) در شبکه‌های اطلاعاتی، مسیرهای کوتاه‌ترین و انتشار اطلاعات چه نقشی در بازیابی اطلاعات دارند؟ مثالی از یک کاربرد عملی ارائه دهید.

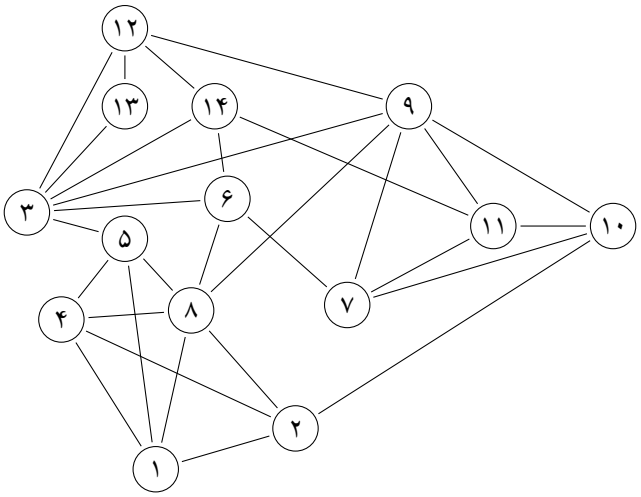
پاسخ

(آ) الگوریتم PageRank به هر گره مقداری رتبه اختصاص می‌دهد که به‌تدریج به سایر گره‌ها منتقل می‌شود. گره‌هایی که لینک‌های بیشتری از گره‌های مهم دریافت می‌کنند، رتبه بالاتری خواهند داشت.
الگوریتم HITS (Hyperlink-Induced Topic Search) دو نوع گره معرفی می‌کند: Authorities (صفحات معتبر) و Hubs (صفحاتی که به بسیاری از صفحات معتبر لینک می‌دهند).
نتیجه: الگوریتم PageRank برای رتبه‌بندی صفحات وب و شبکه‌های عمومی مناسب‌تر است، در حالی‌که HITS در جستجوهای موضوعی عملکرد بهتری دارد.

(ب) مدل Kronecker برای توسعه و رشد شبکه‌ها طراحی شده است. این مدل با استفاده از یک ماتریس پایه کوچک و انجام محاسبات Kronecker، شبکه‌های بزرگ‌تر و پیچیده‌تری تولید می‌کند.
کاربرد در پیش‌بینی ارتباطات: این مدل می‌تواند الگوهای احتمالی ارتباطات آینده بین گره‌ها را شبیه‌سازی کند.

(ج) مفهوم Shortest Path (کوتاه‌ترین مسیر) برای یافتن سریع‌ترین مسیر بین دو گره در شبکه‌های اطلاعاتی اهمیت دارد. مثلاً الگوریتم Dijkstra برای یافتن مسیرهای بهینه در شبکه‌های ارتباطی استفاده می‌شود.
انتشار اطلاعات: مدل‌هایی مانند Forest Fire و PageRank به شبیه‌سازی و تحلیل انتشار اطلاعات در شبکه‌های گرافی کمک می‌کنند. برای مثال، در شبکه‌های اجتماعی، تشخیص سریع مسیرهای بهینه برای پخش یک خبر یا تبلیغ می‌تواند بسیار مؤثر باشد.

پرسش ۶ (۱۰ نمره) باتوجه به گراف زیر به سؤالات پاسخ دهید.



(آ) یک گراف ۳ منتظم همبند با ۸ رأس را پیدا کنید.

(ب) یک گراف ۳ منتظم با ۸ رأس را پیدا کنید و نشان دهید که نسبت به اولی هم‌ریخت نیست.

(ج) چهار گراف ۳ منتظم دیگر با ۸ رأس را بیابید و نشان دهید هیچ دو شکلی هم‌ریخت نیستند.

پاسخ

(آ) یک گراف ۳ منتظم همبند با ۸ رأس به‌صورت زیر معرفی می‌شود:

- **رئوس:** $\{۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸\}$
- **اتصالات (یال‌ها):**
  - ۱ به ۲، ۳، ۴ متصل است.
  - ۲ به ۱، ۵، ۶ متصل است.
  - ۳ به ۱، ۷، ۸ متصل است.
  - ۴ به ۱، ۵، ۷ متصل است.
  - ۵ به ۲، ۴، ۶ متصل است.
  - ۶ به ۲، ۵، ۸ متصل است.

- ۷ به ۳، ۴، ۸ متصل است.
- ۸ به ۳، ۶، ۷ متصل است.

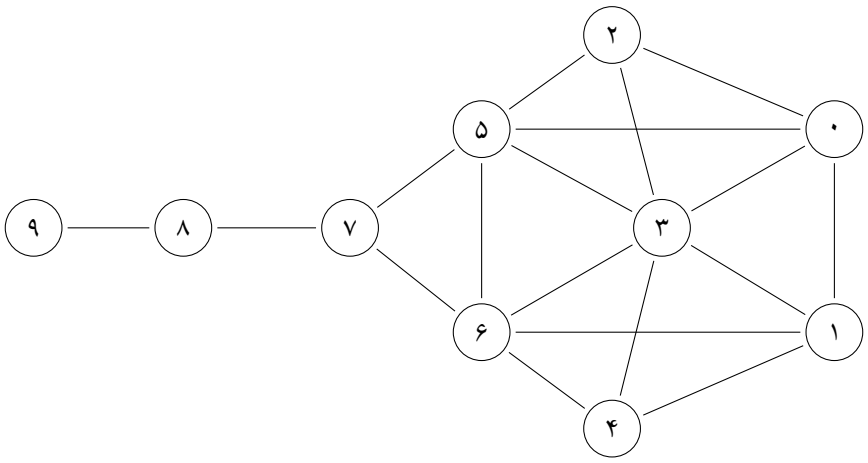
(ب) گراف ۳ منتظم دیگری که نسبت به گراف اول هم‌ریخت نیست:

- **رئوس**: $\{۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸\}$
- **اتصالات (یال‌ها)**:
  - ۱ به ۲، ۵، ۶ متصل است.
  - ۲ به ۱، ۳، ۷ متصل است.
  - ۳ به ۲، ۴، ۸ متصل است.
  - ۴ به ۳، ۵، ۶ متصل است.
  - ۵ به ۱، ۴، ۸ متصل است.
  - ۶ به ۱، ۴، ۷ متصل است.
  - ۷ به ۲، ۵، ۸ متصل است.
  - ۸ به ۳، ۶، ۷ متصل است.

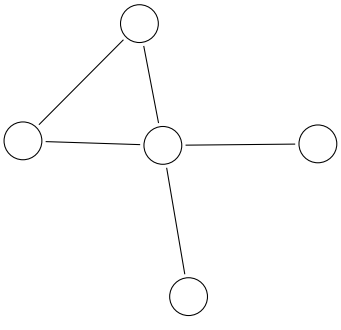
(ج) چهار گراف ۳ منتظم دیگر با ۸ رأس که هیچ دو شکلی هم‌ریخت نیستند:

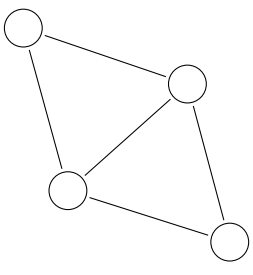
- **رئوس**: $\{۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸\}$
- **اتصالات (یال‌ها)**:
  - ۱ به ۲، ۳، ۴ متصل است.
  - ۲ به ۱، ۵، ۶ متصل است.
  - ۳ به ۲، ۴، ۷ متصل است.
  - ۴ به ۱، ۳، ۸ متصل است.
  - ۵ به ۲، ۶، ۷ متصل است.
  - ۶ به ۵، ۲، ۸ متصل است.
  - ۷ به ۳، ۵، ۸ متصل است.
  - ۸ به ۴، ۶، ۷ متصل است.

پرسش ۷ (۱۰ نمره) با توجه به گراف زیر به سؤالات پاسخ دهید.



(آ) مشخص کنید هر کدام از موتیف‌های زیر چند بار تکرار شده است (با رسم شکل).


Motif E (آ)


Motif R (ب)

(ب) تعداد وقوع هر موتیف از ۱۰ گراف تصادفی آورده شده است. بیان کنید که آیا هر یک از این موتیف‌ها از نظر آماری over-represented شده‌اند یا خیر. اگر مقدار z در z-score بیشتر از ۱/۱۵۰ باشد، از نظر آماری over-represented در نظر گرفته می‌شود.

- Motif E: [9, 8, 1, 6, 8, 13, 6, 8, 3, 9]
- Motif R: [7, 5, 7, 6, 8, 5, 0, 7, 9, 7]

پاسخ

(آ) در گراف داده‌شده، دو موتیف مشخص شده‌اند:

- **E Motif**: یک مثلث متصل با ۳ رأس و ۳ یال.
- **R Motif**: گراف "T" شکل شامل ۴ رأس و ۴ یال.

با بررسی دقیق ساختار گراف، تعداد وقوع این موتیف‌ها به صورت زیر به‌دست آمد:

- **E Motif**: ۶ بار
- **R Motif**: ۵ بار

(ب) برای بررسی آماری اینکه آیا این موتیف‌ها در گراف داده‌شده over-represented هستند یا نه، از معیار $z$score- استفاده می‌کنیم:

$$z = \frac{X - \mu}{\sigma}$$

که در آن:

- X: تعداد وقوع در گراف اصلی
- μ: میانگین وقوع در گراف‌های تصادفی
- σ: انحراف معیار

الف) محاسبات مربوط به E Motif

$$\text{Occurrences} = [۹, ۸, ۱, ۶, ۸, ۱۳, ۶, ۸, ۳, ۹]$$

$$\mu_E = \frac{۹ + ۸ + ۱ + ۶ + ۸ + ۱۳ + ۶ + ۸ + ۳ + ۹}{۱۰} = ۷/۵$$

$$\begin{aligned}\sigma_E^۲ &= \frac{(۹ - ۷/۵)^۲ + (۸ - ۷/۵)^۲ + (۱ - ۷/۵)^۲ + (۶ - ۷/۵)^۲ + (۸ - ۷/۵)^۲}{۱۰} \\ &\quad + \frac{(۱۳ - ۷/۵)^۲ + (۶ - ۷/۵)^۲ + (۸ - ۷/۵)^۲ + (۳ - ۷/۵)^۲ + (۹ - ۷/۵)^۲}{۱۰} \\ &= \frac{۲/۲۵ + ۰/۲۵ + ۴۲/۲۵ + ۲/۲۵ + ۰/۲۵ + ۳۰/۲۵ + ۲/۲۵ + ۰/۲۵ + ۲۰/۲۵ + ۲/۲۵}{۱۰} \\ &= ۱۰/۴۵ \quad \Rightarrow \quad \sigma_E = \sqrt{۱۰/۴۵} \approx ۳/۲۳ \end{aligned} \tag{۱}$$

$$z_E = \frac{۶ - ۷/۵}{۳/۲۳} \approx -۰/۴۶۴$$

از آنجا که $z_E < ۱/۱۵۰$، بنابراین E Motif از نظر آماری over-represented نیست.

ب) محاسبات مربوط به R Motif

$$\text{Occurrences} = [۷, ۵, ۷, ۶, ۸, ۵, ۰, ۷, ۹, ۷]$$

$$\mu_R = \frac{۷ + ۵ + ۷ + ۶ + ۸ + ۵ + ۰ + ۷ + ۹ + ۷}{۱۰} = ۶/۱$$

$$\begin{aligned}\sigma_R^۲ &= \frac{(۷ - ۶/۱)^۲ + (۵ - ۶/۱)^۲ + (۷ - ۶/۱)^۲ + (۶ - ۶/۱)^۲ + (۸ - ۶/۱)^۲}{۱۰} \\ &\quad + \frac{(۵ - ۶/۱)^۲ + (۰ - ۶/۱)^۲ + (۷ - ۶/۱)^۲ + (۹ - ۶/۱)^۲ + (۷ - ۶/۱)^۲}{۱۰} \\ &= \frac{۰/۸۱ + ۱/۲۱ + ۰/۸۱ + ۰/۰۱ + ۳/۶۱ + ۱/۲۱ + ۳۷/۲۱ + ۰/۸۱ + ۸/۴۱ + ۰/۸۱}{۱۰} \\ &= ۵/۴۹ \quad \Rightarrow \quad \sigma_R = \sqrt{۵/۴۹} \approx ۲/۳۴ \end{aligned} \tag{۲}$$

$$z_R = \frac{۵ - ۶/۱}{۲/۳۴} \approx -۰/۴۷$$

از آنجا که $z_R < ۱/۱۵۰$، بنابراین R Motif نیز از نظر آماری over-represented نیست.

سوالات عملی (۳۰ نمره)

تاریخ تحویل: ۲۸ اردیبهشت ۱۴۰۴؛ با تاخیر: ۳۱ اردیبهشت ۱۴۰۴

پرسش ۱ (۱۵ نمره) برای حل سوال عملی اول به دفترچه ژوپیتر ضمیمه این تمرین مراجعه کنید.

پاسخ

پرسش ۲ (۱۵ نمره) در این تمرین، هدف شما درک، پیاده‌سازی و مقایسه‌ی الگوریتم‌های مختلف تشخیص اجتماع[۱] در یک شبکه‌ی واقعی از مقالات علمی[۲] است. همچنین در قسمت دوم این سوال نیاز است تا شما با استفاده از ساختار شبکه، ویژگی‌های رئوس و برچسب‌های واقعی هر رأس استفاده کنید تا کیفیت خوشه‌بندی‌هایی که تعیین کردید را بررسی نمایید. داده مورد بررسی در این تمرین، داده Cora شامل ۲۷۰۸ مقاله علمی است که در یک شبکه استنادی با ۵۴۲۹ یال ارجاع‌دهی شده‌اند. همچنین این مقالات به یکی از ۷ کلاس مختلف موضوعی زیر تعلق دارند:

- Case Based
- Genetic Algorithms
- Neural Networks
- Probabilistic Methods
- Reinforcement Learning
- Rule Learning
- Theory

این داده شامل دو فایل اصلی است.

- **cora.content** که اطلاعات مربوط به ویژگی‌های هر مقاله به صورت بردارهای باینری از کلمات موجود در لغت‌نامه نشان داده و همچنین شامل موضوع اصلی هر مقاله است.
- **cora.cites** گراف استنادی که در آن هر سطر نشان‌دهنده یک ارتباط استنادی (ارجاع‌دهی) بین دو مقاله است.

به هر یک از بخش‌های زیر در یک سلول یک دفترچه ژوپیتر پاسخ دهید.

(آ) داده‌های موجود در فایل‌های cora.content و cora.cites را بارگذاری کنید.

(ب) تعداد مقالات در هر یک از دسته‌های موضوعی را با نمودار مناسب نشان دهید.

(ج) گراف استنادها را بر اساس داده‌های cora.cites بسازید.

(د) توزیع درجات این گراف را نمودار کنید. میانگین این درجات چقدر است؟

(ه) تعداد و اندازه مولفه‌های هم‌بندی را مشخص کنید.

(و) میانگین کوتاه‌ترین فواصل در مولفه‌های هم‌بندی این گراف چند است؟

(ز) با استفاده از الگوریتم‌های Girvan Newman و Louvain و Lukes جوامع موجود در این گراف را تشخیص دهید. می‌توانید از پیاده‌سازی‌های موجود در کتاب‌خانه networkx استفاده کنید.

(ح) با توجه به غیر نظارتی بودن مساله تشخیص جوامع، توضیح دهید چگونه می‌توان از معیار F1 برای این مساله استفاده کرد. آن را پیاده‌سازی کنید و روش‌های فوق را با استفاده از ان مقایسه کنید.

(ط) با استفاده از معیار Normalized Mutual Information عملکرد این روش‌ها را بررسی کنید.

پاسخ

[۱] community detection
[۲] citation network