باسمه تعالی

پروژه‌ی درس آمار و احتمال مهندسی

آشنایی با برخی روش‌های خوشه‌بندی گراف‌ها

فاز دوم

آخرین مهلت تحویل:

۱۴ بهمن ۱۴۰۱

۱ من از دیار حبیبم نه از بلاد غریب![1] (۲)

در بخش پنجم فاز اول پروژه با گراف‌های تصادفی آشنا شدیم. حال در این بخش برخی مدل‌های گراف تصادفی که در شناسایی جوامع استفاده می‌شود را بررسی می‌کنیم.

۱.۱ گراف تصادفی مدل Watts–Strogatz

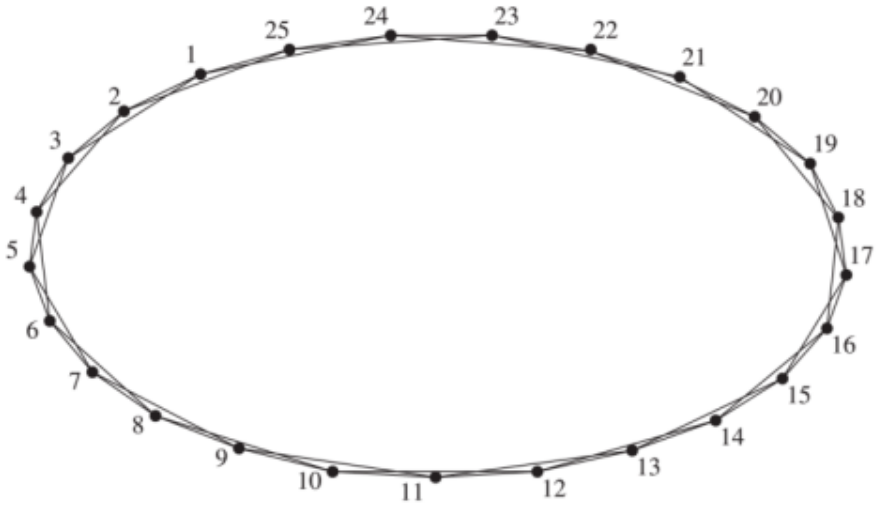
در بخش پنجم فاز اول پروژه با گرافهای تصادفی آشنا شدیم و دیدیم که طبق قانون جهان کوچک، هر دو نفر در دنیا با احتمال نزدیک به یک با حداکثر ۶ واسطه یکدیگر را میشناسند.

تعریف ۱ (فاصله‌ی میانگین). به میانگین تعداد یال‌هایی که در کوتاه‌ترین مسیر بین هر دو رأس موجود در گراف طی می‌شوند، فاصله‌ی میانگین آن گراف گفته می‌شود.

تعریف ۲ (ضریب خوشه‌بندی). ضریب خوشه‌بندی یک رأس، به صورت نسبت تعداد یال‌های بین رأس‌های همسایه‌ی آن رأس به تعداد تمام یال‌هایی که در گراف کامل (یعنی گرافی که همه‌ی یال‌ها در آن وجود دارند) بین همین همسایه‌ها وجود دارد، تعریف می‌شود.

تعریف ۳ (ضریب خوشه‌بندی میانگین). به میانگین ضریب خوشه‌بندی رأس‌های یک گراف، ضریب خوشه‌بندی میانگین می‌گویند.

مدل گرافی زیر را در نظر بگیرید:
n رأس را روی محیط یک دایره قرار می‌دهیم. هر رأس را به نزدیکترین $2m$ رأس اطرافش وصل می‌کنیم. گراف به دست آمده را $\mathcal{G}_{\mathrm{WS}}(n,m)$ می‌نامیم.
به عنوان مثال، برای $n = 25$ گره و پارامتر $m = 2$ به چنین گرافی خواهیم رسید:



شکل ۱: مثالی از مدل Watts-Strogatz

[1] نماز شام غریبان چو گریه آغازم/ به مویه‌های غریبانه قصه پردازم
به یاد یار و دیار آن چنان بگریم زار/ که از جهان ره و رسم سفر براندازم
من از دیار حبیبم نه از بلاد غریب/ مهیمنا به رفیقان خود رسان بازم [حافظ]

پرسش ۱. فاصله میانگین گراف $\mathcal{G}_{\text{WS}}(n,m)$ را بیابید.

پرسش ۲. ثابت کنید در یک گراف دل‌خواه مانند $\mathcal{G}=(V,E)$ که $V=\{v_i\}_{i=1}^{n}$ مجموعه‌ی رأس‌ها و E مجموعه‌ی یال‌های $\mathcal{G}$ هستند، ضریب خوشه‌بندی رأس v_i که آن را با C_i نشان می‌دهیم، از رابطه‌ی زیر به دست می‌آید:

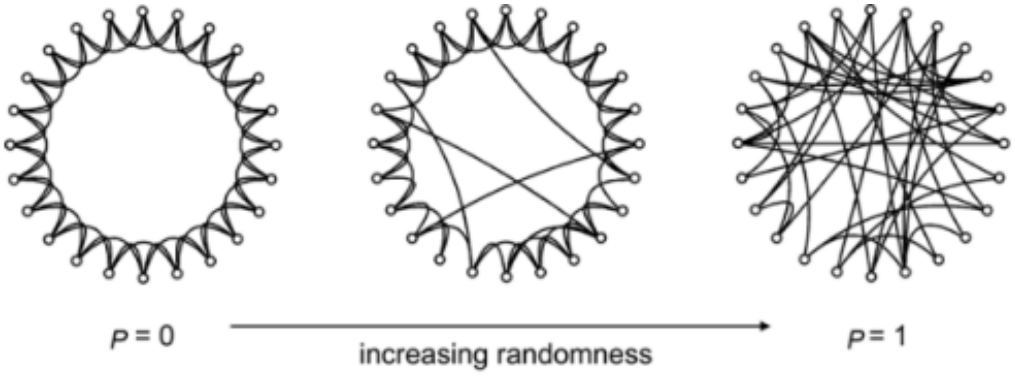
$$C_i=\frac{2|\{e_{ij}:v_i,v_j\in N_i,e_{ij}\in E\}|}{k_i(k_i-1)}.$$

در رابطه‌ی فوق، N_i، مجموعه‌ی رئوس همسایه‌ی رأس v_i، e_{ij}، یالی که رأس‌های v_i,v_j را به هم متّصل می‌کند و k_i تعداد همسایه‌های رأس v_i هستند.

پرسش ۳. ثابت کنید ضریب خوشه‌بندی میانگین در گراف $\mathcal{G}_{\text{WS}}(n,m)$ برابر با $\frac{1}{2}$ است.

همان‌طور که دیده می‌شود، فاصله‌ی میانگین گراف $\mathcal{G}_{\text{WS}}(n,m)$ زیاد است، برای کاهش مقدار فاصله‌ی میانگین، باید به گراف اوّلیه‌ی $\mathcal{G}_{\text{WS}}(n,m)$ کمی خاصیت تصادفی اضافه کنیم. این کار را به صورت زیر انجام می‌دهیم:

برای هر رأس $v_i\in V$، یال‌هایی که v_i را به m رأس سمت راست آن وصل می‌کند را در نظر می‌گیریم. بنا بر روش ساخت گراف $\mathcal{G}_{\text{WS}}(n,m)$ تعداد این یال‌ها برابر با m است. هرکدام از این یال‌ها را با احتمال p حذف می‌کنیم و به جای آن، به صورت کاملاً تصادفی یالی جدید از رأس v_i به یکی از رئوسی که از طریق یک یال به v_i متصل نیست، رسم می‌کنیم. مقدار p را احتمال بازاتّصال rewiring می‌نامیم.



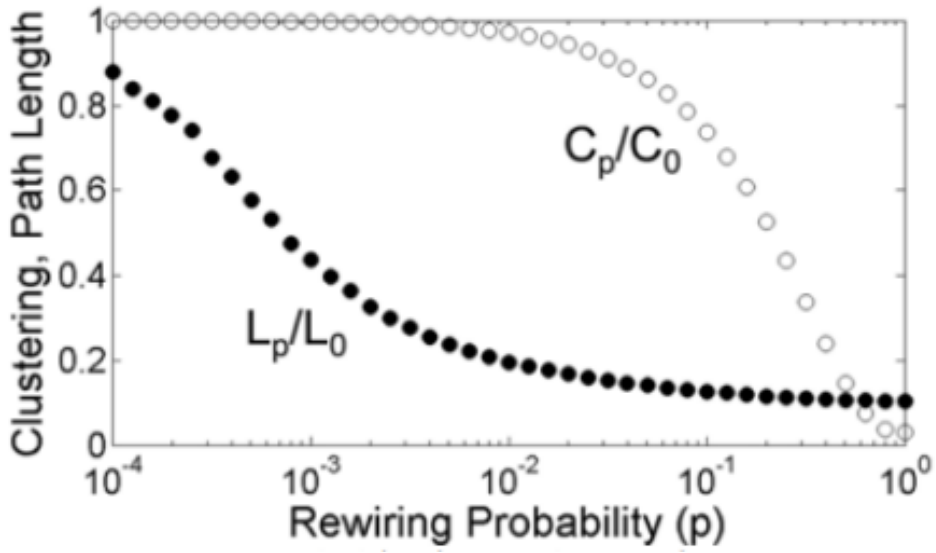
شکل ۲: فرآیند اضافه‌کردن خواص تصادفی به $\mathcal{G}_{\text{WS}}(n,m)$

همانطور که در شکل ۲ دیده می‌شود، با افزایش p این گراف به گراف تصادفی نزدیک می‌شود. با نزدیک شدن گراف به گراف تصادفی، فاصله‌ی میانگین کم می‌شود، ولی به طور همزمان ضریب خوشه‌بندی میانگین هم کاهش می‌یابد. امّا خبر خوش آنست که نرخ کاهش این دو پارامتر باهم متفاوت است.

همانطور که در شکل ۳ می‌بینیم، نرخ کاهش ضریب خوشه‌بندی میانگین کمتر از نرخ کاهش فاصله‌ی میانگین است. در نتیجه اگر مقدار p را طوری تنظیم کنیم که ضریب خوشه‌بندی میانگین چندان کاهش نیافته باشد، ولی فاصله‌ی میانگین به قدر کافی کم شده باشد، توانسته‌ایم قانون جهان کوچک را در مورد گراف‌های $\mathcal{G}_{\text{WS}}(n,m)$ برقرار کنیم. به عنوان مثال در شکل ۳ اگر $p\approx 0/01$ انتخاب شود، تا حدّ خوبی به هدفمان رسیده‌ایم.

پرسش شبیه‌سازی ۱. فرض کنید فیلیمو/نماوا/فیلمنت $n=10000$ کاربر دارند. اگر فرض کنیم $m=400$، گراف $\mathcal{G}_{\text{WS}}(n,m)$ را بسازید.

پرسش شبیه‌سازی ۲. به ازای pهای مختلف در بازه‌ی $[10^{-6},1]$، به گراف پرسش شبیه‌سازی ۱ خاصیت تصادفی اضافه کنید و نموداری مانند شکل ۳ را رسم کنید.



شکل ۳: مثالی از روند کاهش فاصله‌ی میانگین و ضریب خوشه‌بندی میانگین با افزایش خاصیت تصادفی در گراف $\mathcal{G}_{\mathrm{WS}}(n,m)$

پرسش شبیه‌سازی ۳. با کمک نموداری که در پرسش شبیه‌سازی ۲ رسم کردید، به صورت تقریبی مقدار بهینه‌ی p (که آن را با p^* نشان می‌دهیم) را بیابید و برقراری خاصیت جهان کوچک، قبل و بعد از اضافه‌شدن تصادف به گراف $\mathcal{G}_{\mathrm{WS}}(n,m)$ با احتمال بازاتّصال $p=p^*$ را بررسی کنید.

۲.۱ گراف تصادفی مدل Configuration

در این بخش مدل دیگری از گراف‌های تصادفی را بررسی می‌کنیم. فرض کنید دنباله‌ی $\mathbf{d}=(d_1,d_2,...,d_n)$ دنباله‌ی درجات رأس‌های یک گراف n رأسی باشد. دنباله‌ی زیر از رأس‌ها را تشکیل می‌دهیم:

$$\mathbf{a_d}=(\underbrace{1,1,\dots,1}_{d_1\text{ entries}},\ \underbrace{2,2,\dots,2}_{d_2\text{ entries}},\ \dots,\ \underbrace{n,n,\dots,n}_{d_n\text{ entries}})$$

گراف $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$ را به صورت زیر تشکیل می‌دهیم:

از دنباله‌ی $\mathbf{a_d}$ به صورت کاملاً تصادفی و بدون جایگذاری، دو عدد را انتخاب کرده و رأس‌های متناظر با آن‌ها را به هم وصل می‌کنیم. این کار را تا جایی ادامه می‌دهیم که اعضای دنباله‌ی $\mathbf{a_d}$ تمام شوند.

مثلا فرض کنید $\mathbf{d}=(3,4,3)$. در این صورت داریم:

$$\mathbf{a_d}=(1,1,1,2,2,2,2,3,3,3).$$

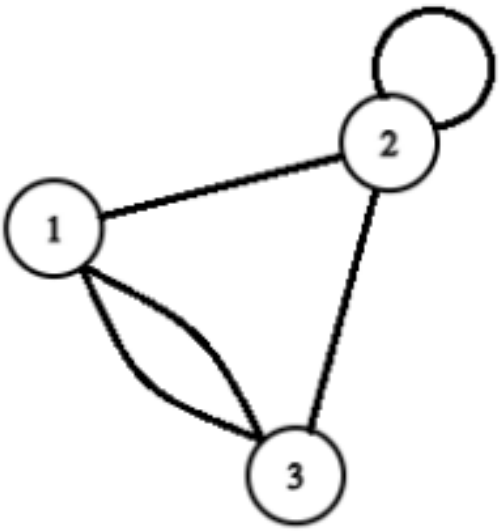
حال فرض کنید اعداد تصادفی‌ای که انتخاب می‌کنیم این اعداد باشند:

$$(1,2),(2,3),(1,3),(2,2),(3,1).$$

در این صورت گراف $\mathcal{G}_{\mathrm{C}}(n,\mathbf{d})$ به صورت شکل ۴ خواهد شد.

تعریف ۴ (طوقه). یالی که از یک رأس به خودش وصل شده باشد را طوقه می‌نامیم.

تعریف ۵ (اَبَریال). اگر بین دو رأس بیشتر از یک یال وجود داشته باشد، آن‌گاه می‌گوییم بین آن دو رأس، ابریال وجود دارد.



شکل ۴: مثالی از گراف $\mathcal{G}_{\mathrm{C}}(n, \mathbf{d})$

تعریف ماتریس مجاورت را در فاز اوّل پروژه دیده‌اید. برای گراف‌هایی که طوقه و ابریال داشته باشند، ماتریس مجاورت به صورت زیر تعریف می‌شود:

تعریف ۶ (ماتریس مجاورت برای گراف‌های دارای طوقه و ابریال). ماتریس مجاورت برای گراف $\mathcal{G}$ با n رأس یک ماتریس $\mathbf{A} \in \mathbb{R}^{n \times n}$ است که درایه‌های آن به صورت زیر هستند:

$A_{i,j} =$ تعداد یال‌های بین دو رأس i, j.

از تعریف ۶ واضح است که $A_{i,i}$ برابر با تعداد طوقه‌های رأس i است.

پرسش ۴. ثابت کنید:

$$d_i = A_{i,i} + \sum_{j=1}^{n} A_{i,j}.$$

پرسش ۵. اگر تعداد یال‌های گراف $\mathcal{G}_{\mathrm{C}}(n, \mathbf{d})$ را با m نشان دهیم، ثابت کنید که $m = \frac{1}{2} \sum_{i=1}^{n} d_i$.

پرسش ۶. نشان دهید احتمال اگر دنباله‌ی $\mathbf{d}$ به ما داده شده باشد، احتمال مشاهده‌ی گراف $\mathcal{G}_{\mathrm{C}}(n, \mathbf{d})$ با ماتریس مجاورت $\tilde{\mathbf{A}}$ برابر است با:

$$\mathbb{P}[\mathbf{A} = \tilde{\mathbf{A}}] = \frac{1}{\prod_{i=1}^{m}(2i-1)} \frac{\prod_{i=1}^{n}(d_i!)}{\left(\prod_{i=1}^{n} 2^{\tilde{A}_{i,i}}\right)\left(\prod_{1 \le i < j \le n}(\tilde{A}_{i,j}!)\right)}$$

پرسش شبیه‌سازی ۴. فرض کنید $n = 20000$ کاربر به ژانر درام علاقه‌مندند و هر کس دقیقاً با $k = 100$ کاربر دیگر هم‌سلیقه است. فرض کنید با مدل Configuration کاربران این ژانر را مدل می‌کنیم. اگر یک یال را به تصادف انتخاب کنیم احتمال اینکه طوقه یا ابریال نباشد چقدر است؟
در حقیقت یکی از ضعف‌های مدل Configuration وجود طوقه یا ابریال است. آیا می‌توان از این نقطه ضعف صرف‌نظر کرد؟

۳.۱ گراف تصادفی مدل Expected Degree

مشابه مدل قبل، فرض کنید $\mathbf{d} = (d_1, d_2, ..., d_n)$ دنباله‌ی درجات یک گراف باشد.
اگر فرض کنیم $(\max_{1 \le i \le n} d_i)^2 < \sum_{k=1}^{n} d_k$، در این صورت تعریف می‌کنیم:

$$p_{i,j} = \frac{d_i d_j}{\sum_{k=1}^{n} d_k}$$

در این مدل یال بین دو رأس v_i, v_j با احتمال $p_{i,j}$ وجود دارد و با احتمال $1 - p_{i,j}$ وجود ندارد. همچنین وجود و عدم وجود یال‌های مختلف از هم مستقل است. دقّت کنید که در این مدل طوقه هم می‌توانیم داشته باشیم. گراف به دست آمده از این روش را $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ می‌نامیم.

پرسش ۷. دقّت کنید که $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ یک گراف تصادفی است و در نتیجه مشخّصات آن مانند درجه‌ی رئوس متغیّرهای تصادفی هستند. اگر متغیّر تصادفی درجه‌ی رأس v_i را با D_i نشان دهیم، ثابت کنید $\mathbb{E}[D_i] = d_i$.

پرسش شبیه‌سازی ۵. فرض کنید ژانر طنز n کاربر دارد و دنباله‌ی درجات را به صورت $\mathbf{d} = (d_1, d_2, ..., d_n)$ در نظر بگیرید. دو مقدار «مناسب» برای n و $\mathbf{d}$ پیشنهاد بدهید و با در نظر گرفتن مقادیر پیشنهادی خودتان، $N = 100$ بار گراف $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ را بسازید. سپس بررسی کنید که آیا میانگین دنباله‌های درجات رئوسِ این N گراف، با دنباله‌ی $\mathbf{d}$ برابر است یا نه؟

پرسش ۸. فرض کنید فیلیمو/نماوا/فیلمنت n کاربر دارد که به ژانر طنز علاقه‌مندند. همچنین فرض کنید هر کاربر با k کاربر دیگر هم‌سلیقه باشد. در این حالت بدیهی است که $\mathbf{d} = (\underbrace{k, k, \ldots, k}_{n \text{ entries}})$ خواهد شد. حال نشان دهید احتمال آن‌که تعداد روابط هم‌سلیقگی هر کاربر در گراف تصادفی $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ برابر با k شود، برابر است با: $\frac{e^{-k} k^k}{k!}$. سپس ثابت کنید که این مقدار از $\frac{1}{2}$ کمتر است که نشان‌دهنده‌ی یک ضعف این مدل است. دقّت کنید در مدل قبلی درجه‌ی رأس v_i همواره d_i بود.

پرسش ۹. نشان دهید احتمال آن‌که درگراف تصادفی $\mathcal{G}_{\text{ED}}(n, \mathbf{d})$ درجه‌ی رأس v_i برابر با صفر شود کرانی به صورت زیر دارد:

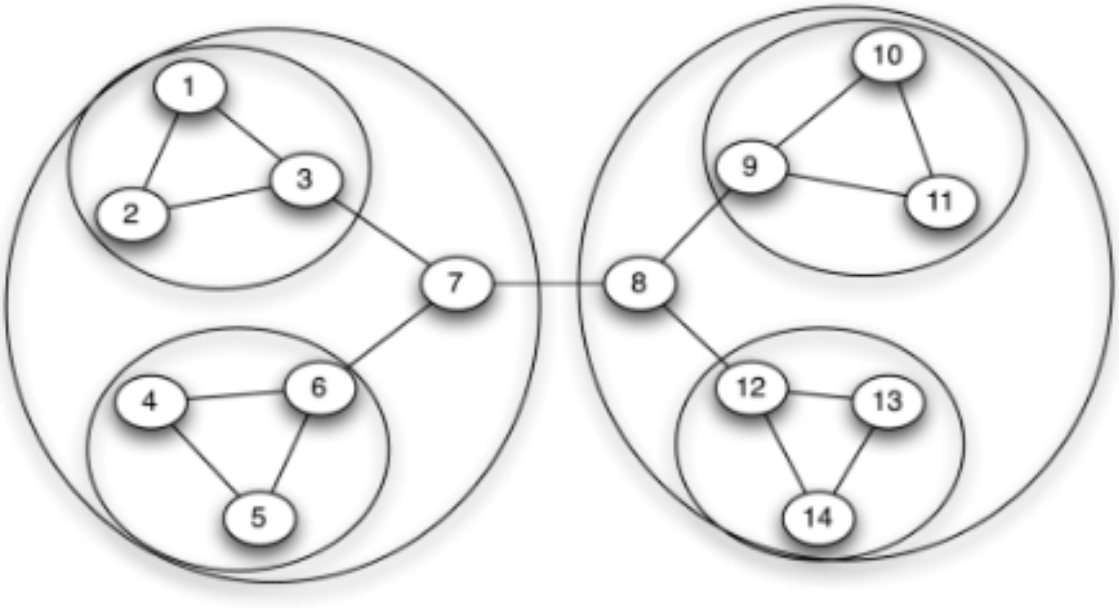
$$\mathbb{P}[D_i = 0] \le e^{-d_i}.$$

پرسش ۱۰. ثابت کنید که اگر $\sum_{i=1}^{n} e^{-d_i} \le \epsilon$، آنگاه با احتمال بیشتر از $1 - \epsilon$، درجه‌ی تمام رئوس بزرگ‌تر یا مساوی با ۱ است.

---

۲ فهمِ ضعیفْ رایْ فضولی چرا کند؟[۲]

فرض کنید کاربران ۱ تا ۷ به ژانر طنز علاقه دارند و کاربران ۸ تا ۱۴ به ژانر درام علاقه دارند. داخل هر یک از این خوشه‌ها نیز زیرخوشه‌های مختلفی وجود دارند که علایق اعضای هرکدام از زیرخوشه‌ها به هم تشابه دقیق‌تر و ظریف‌تری دارد. مثلا از بین کاربران ۱ تا ۷ که به ژانر طنز علاقه دارند، کاربران ۱ تا ۳ به سریال‌های ژانر طنز علاقه دارند، امّا کاربران ۴ تا ۶ فقط به فیلم سینمایی با ژانر طنز علاقه دارند.



شکل ۵: مثالی از خوشه‌ها و زیرخوشه‌ها در یک جامعه

از طرف دیگر، همان‌طور که در شکل ۵ دیده می‌شود، کاربر ۷ در ژانر طنز، با کاربر ۸ در ژانر درام هم‌سلیقه است. با توجّه به گراف روابط، می‌توان حدس زد که سلیقه‌ی کاربر ۷ احتمالاً به کاربران ۳ و ۶ که در خوشه‌ی علاقه‌مندان ژانر طنز هستند، نزدیک‌تر است تا به کاربر ۸ که در خوشه‌ی دیگری قرار دارد. برای نشان دادن این نکته، می‌توانیم یک برچسب «ضعیف» روی یال مربوط به رابطه‌ی هم‌سلیقگی بین کاربر ۷ و ۸ بچسبانیم. دقّت کنید که برچسب یال هم‌سلیقگی بین کاربر ۳ و ۷ و بین کاربر ۶ و ۷ می‌تواند ضعیف یا قوی باشد اما بین کاربر ۷ و ۸ (یعنی یال بین دو خوشه) را ضعیف در نظر می‌گیریم.

این ارتباط خاص بین رئوس ۷ و ۸ برای یک کمپانی VOD مهم است زیرا می‌توان فیلم‌هایی را که به کاربر ۸ پیشنهاد می‌شود، به کاربر ۷ که با او هم‌سلیقه، ولی از خوشه‌ای متفاوت است هم پیشنهاد داد و اگر کاربر ۷ آن فیلم‌ها را دوست داشت به افراد دیگرخوشه‌ی ژانر طنز که با کاربر ۷ هم‌سلیقه هستند هم آن فیلم‌ها را معرفی کرد. با این روش ممکن است بتوانیم پیشنهادات جدیدی به کاَربران ارائه کنیم.

پرسش ۱۱. با توجه به توضیحاتی که در بالا داده شد، به نظر شما در روابط اجتماعی واقعی، احتمال این‌که یک فرصت شغلی یا خبر جدید را از دوستان صمیمی خود به دست آوریم بیشتر است یا از آشنایان دور؟ چرا؟

فرض کنید کاربر A با کاربر B, C هم‌سلیقه است. امّا به دلیل این‌که به کاربران B, C فیلم‌های مشترک زیادی پیشنهاد نشده است، فعلا نمی‌دانیم که آیا B, C هم‌سلیقه هستند یا نه، و باید به آن‌ها چند فیلم پیشنهاد داده شود تا این موضوع بررسی شود.

[۲] در کارخانه‌ای که رَهِ عقل و فضل نیست/ فهمِ ضعیفْ رایْ فضولی چرا کند؟ [حافظ]

پرسش ۱۲. توضیح دهید که چرا احتمال هم‌سلیقه بودن B, C در حالتی که هردو با A هم‌سلیقه هستند، نسبت به حالتی که هم‌سلیقه‌ی مشترک ندارند بیشتر است؟ به این خاصیت Triadic Closure گفته می‌شود. سپس نتیجه بگیرید که تعداد یال‌های گراف مربوط به خوشه‌ی علاقه‌مندان یک ژانر به مرور بیشتر می‌شود.

حال فرض کنید برای هر سه کاربر دلخواه A, B, C، اگر بدانیم کاربران B, C با کاربر A هم‌سلیقه هستند، احتمال ایجاد Triadic Closure و هم‌سلیقه بودن B, C برابر با p باشد.

پرسش ۱۳. اگر بدانیم کاربران B, C با کاربران $A_۱, A_۲, \ldots, A_k$ هم‌سلیقه هستند، آنگاه احتمال هم‌سلیقه بودن B, C را بیابید. رفتار احتمال را با افزایش k توجیه کنید.

۳ فرض ایزد بگزاریم و به کس بد نکنیم![۳]

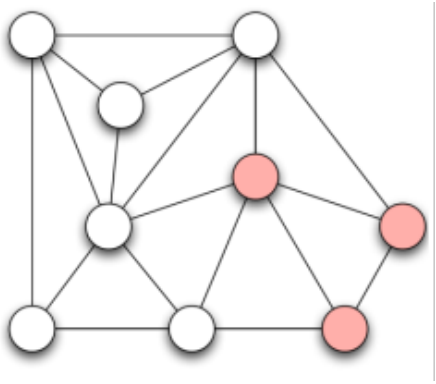
در این بخش می‌خواهیم درستی یا نادرستی فرضیه‌ی زیر را بررسی کنیم.
فرضیه: به طور کلّی سلیقه یک خانم به یک خانم شبیه‌تر است تا یک آقا.
گراف روابط هم‌سلیقگی بین افراد را به صورت $\mathcal{G} = \mathcal{G}_۱ \cup \mathcal{G}_۲$ در نظر می‌گیریم که $\mathcal{G}_۱ = (V_۱, E_۱)$ تنها شامل کاربران خانم و $\mathcal{G}_۲ = (V_۲, E_۲)$ تنها شامل کاربران آقا است. در نتیجه $\mathcal{G}_۱ \cap \mathcal{G}_۲ = \emptyset$. همچنین فرض می‌کنیم $n_۱$ کاربر خانم، $n_۲$ کاربر آقا و $n = n_۱ + n_۲$ کاربر در کل داریم. گراف $\mathcal{G}_۱$ را یک گراف تصادفی می‌گیریم که بین هر دو رأس آن با احتمال $p_۱$ و مستقل از رئوس دیگر یک یال وجود دارد. گراف $\mathcal{G}_۲$ را یک گراف تصادفی می‌گیریم که بین هر دو رأس آن با احتمال $p_۲$ و مستقل از رئوس دیگر یک یال وجود دارد. همچنین بین هر دو رأس مانند u, v که $u \in V_۱$ و $v \in V_۲$ با احتمال $p_{۱۲}$ و مستقل از یال‌های دیگر یک یال وجود دارد.

پرسش ۱۴. فرض صفر و فرض مقابل را برحسب پارامترهای فوق بنویسید.

فرض کنید تعداد روابط هم‌سلیقگی بین خانم‌ها $m_۱$، تعداد روابط هم‌سلیقگی بین آقایان $m_۲$ و تعداد روابط هم‌سلیقگی بین یک خانم و یک آقا $m_{۱۲}$ باشد.

پرسش ۱۵. یک آزمون فرض برای آزمودن فرضیه‌ی مطرح شده طرّاحی کنید.

پرسش شبیه‌سازی ۶. فرض کنید روابط هم‌سلیقگی بین افراد به صورت شکل ۶ باشد، (گره‌های صورتی نشان‌دهنده خانم‌ها و گره‌های سفید نشان‌دهنده آقایان هستند). برنامه‌ای بنویسید که مقدار خطای نوع اوّل (α) را از ورودی بگیرد و در خروجی چاپ کند که به ازای این مقدار α فرض صفر رد می‌شود یا رد نمی‌شود؟ در چه مقادیری از α فرض صفر رد می‌شود؟



شکل ۶: مثالی از روابط هم‌سلیقگی بین خانم‌ها و آقایان

---

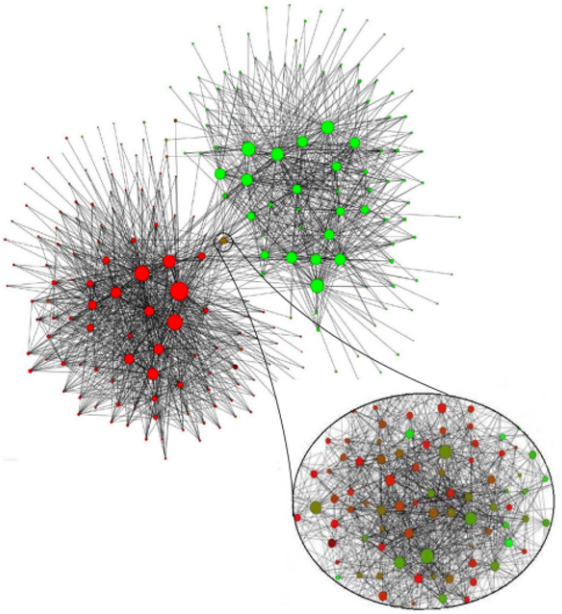
[۳] فرض ایزد بگزاریم و به کس بد نکنیم/ وان چه گویند روا نیست، نگوییم رواست [حافظ]

۴ اگر مرگ نبود، زندگی، زندگی نبود![۴]

فرض کنید فیلیمو/نماوا/فیلمنت n کاربر دارد. یکی از نکاتی که می‌تواند ما را در یافتن خوشه‌ها یاری کند، تعداد فیلم‌های مشترک موردعلاقه بین دو کاربر است. به تعبیر دیگر، اگر دو کاربر تعداد زیادی فیلم را هم‌زمان دوست داشته باشند، به احتمال بالا عضو یک خوشه هستند. تعداد فیلم‌های مشترک موردعلاقه بین دو کاربر را می‌توان با وزن یال بین رأس‌های متناظر با آن دو کاربر در گراف نشان داد. در نتیجه تعریف ماتریس مجاورت عوض می‌شود:

تعریف ۷ (ماتریس مجاورت برای گراف وزن‌دار). ماتریس مجاورت برای گراف وزن‌دارِ $\mathcal{G}$ با n رأس یک ماتریس $\mathbf{A} \in \mathbb{R}^{n\times n}$ است که درایه‌های آن به صورت زیر هستند:

$$A_{i,j} = \text{وزن یالِ بین دو رأس } i,j.$$



شکل ۷: مثالی از روابط هم‌سلیقگی در یک گراف

در این قسمت می‌خواهیم روش دیگری برای شناسایی خوشه‌ها را بررسی کنیم. این الگوریتم Louvain[۵] نام دارد. در این روش ابتدا برای یک گراف وزن‌دار با ماتریس مجاورت $\mathbf{A} \in \mathbb{R}^{n\times n}$ مقدار modularity را به صورت زیر تعریف می‌کنیم:

$$Q = \frac{۱}{۲m}\sum_{i,j=۱}^{n}\left[A_{i,j} - \frac{k_i k_j}{۲m}\right]\delta(c_i, c_j) \qquad (۱)$$

در رابطه‌ی (۱)، $A_{i,j}$ وزن یال بین رئوس v_i, v_j، m مجموع وزن تمام یال‌های گراف، k_i مجموع وزن یال‌هایی که به رأس v_i متّصل هستند و $\delta(c_i, c_j)$ به صورت $\delta(c_i, c_j) = \mathbb{۱}\{c_i = c_j\}$ تعریف می‌شود. هدف ما در این الگوریتم افزایش مقدار Q است.

در الگوریتم Louvain ابتدا هر رأس از گراف را یک خوشه‌ی جداگانه در نظر می‌گیریم (یا معادلاً، هر شخص را علاقه‌مند به یک ژانر جداگانه در نظر می‌گیریم، به طوری که هیچ دو نفری به ژانر یکسانی علاقه نداشته باشند). سپس در یک حلقه تلاش می‌کنیم مقدار Q را در هر دور کمی افزایش دهیم. این کار را تا زمانی انجام می‌دهیم که مقدار Q افزایش قابل توجّهی نیابد.

[۴] از موریس مترلینک، نویسنده، شاعر و فیلسوف بلژیکی
[۵] شهری در بلژیک!

الگوریتم به صورت زیر است:

۱. هر رأس از گراف را یک خوشه‌ی مجزّا در نظر بگیرید.

۲. برای $t = ۱, ۲, \ldots, T$:

(آ) برای $i = ۱, ۲, \ldots, |V|$:

(i) تمام همسایه‌های رأس v_i را در نظر بگیرید، یعنی تمام رأس‌هایی مانند v_j که $A_{i,j} > ۰$. این همسایه‌ها را در مجموعه‌ای مانند N_i قرار می‌دهیم.

(ii) برای هر $v_j \in N_i$:

- رأس v_i را از خوشه‌ی خودش خارج کنید و به خوشه‌ای که v_j در آن است، اضافه کنید. میزان تغییرات Q با این جابجایی را ثبت کنید. این مقدار را $\Delta Q_{i,j}$ می‌نامیم.

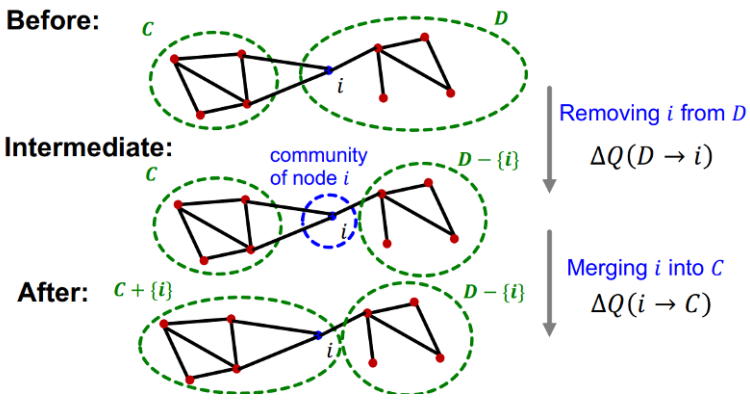
(iii) تعریف کنید $\Delta Q_i^{(t)} = \max_{j:v_j \in N_i} \Delta Q_{i,j}$ و $j^* = \arg\max_{j:v_j \in N_i} \Delta Q_{i,j}$.

(iv) اگر $\Delta Q_i^{(t)} > ۰$ است، رأس v_i را از خوشه‌ی خودش خارج کنید و به خوشه‌ی شامل v_{j^*} اضافه کنید. در غیر این صورت v_i را در خوشه‌ی خودش نگه دارید.

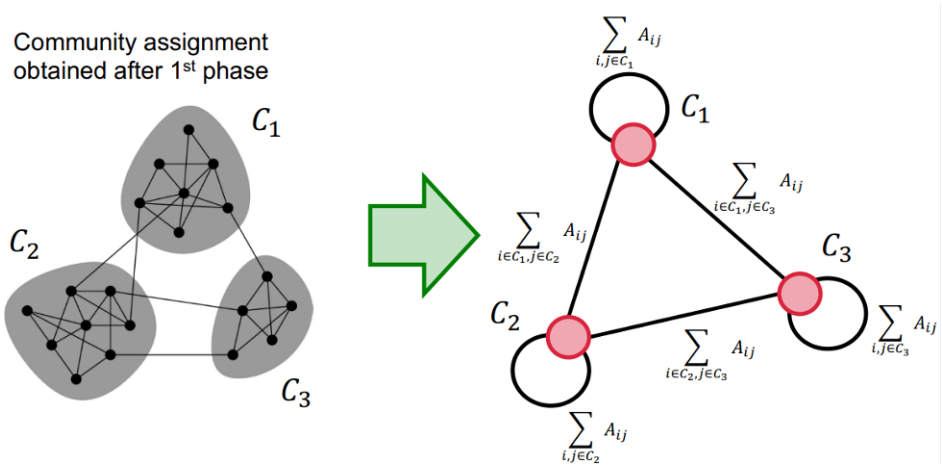
(ب) هر خوشه را به صورت یک آبررأس درآورید. این آبررأس یک یال به دور خودش (طوقه) دارد که وزن این یال برابر است با مجموع تعداد یال‌هایی که از هر رأس، به رأسی داخل همین خوشه وصل می‌شود. همچنین وزن یال بین این آبررأس‌ها برابر است با مجموع وزن یال‌های بین رأس‌های دو خوشه.

(ج) مجموعه‌ی رئوس گراف (V) را به‌روز کنید.

در شکل‌های ۸ و ۹ دو مرحله‌ی اصلی این الگوریتم را مشاهده می‌کنیم.

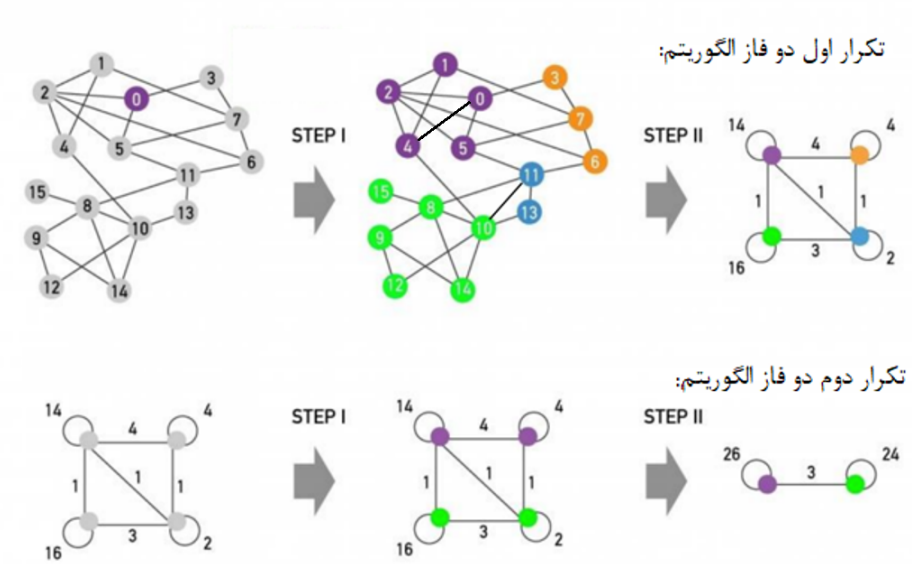


شکل ۸: مثالی از جابجایی رئوس بین خوشه‌ها



شکل ۹: مثالی از تبدیل خوشه‌ها به اَبَررأس

پرسش ۱۶. الگوریتم Louvain را روی گراف شکل ۱۰ به صورت دستی اجرا کنید. آیا به همین خروجی‌ها می‌رسید؟ (در مرحله‌ی ۱ رأس‌ها را به ترتیب شماره‌های شکل بررسی کنید.)



شکل ۱۰: مثالی از الگوریتم Louvain

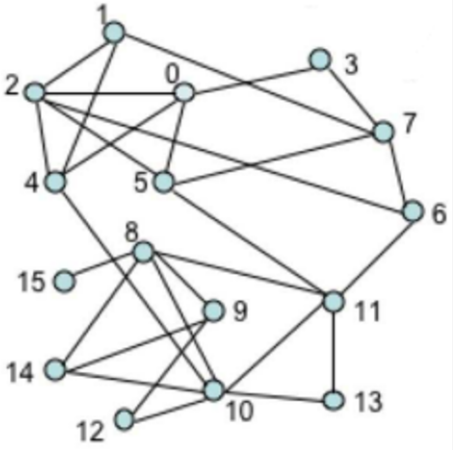
پرسش ۱۷. با استفاده از رابطه (۱) و الگوریتم Louvain، توضیح دهید که چرا با افزایش (و نه کاهش) Q، خوشه‌ها را بهتر می‌توانیم تشخیص دهیم؟

پرسش ۱۸. فرض کنید رأس v_i می‌خواهد به خوشه‌ی حاوی رأس v_j اضافه شود، این خوشه را با c نشان می‌دهیم. نشان دهید مقدار تغییرات Q با این جابجایی (فقط مرحله‌ی اضافه‌شدن را در نظر می‌گیریم) برابر است با:

$$\Delta Q_{i,j} = \left[\frac{\Sigma_{\text{in}} + k_{i,\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

که در آن Σ_{in} مجموع وزن تمام یال‌های درون خوشه‌ی c، Σ_{tot} مجموع وزن تمام یال‌های متّصل به رئوس خوشه‌ی c، k_i مجموع وزن تمام یال‌های که به رأس v_i وصل هستند و $k_{i,\text{in}}$ مجموع وزن یال‌هایی که v_i را به یکی از اعضای خوشه‌ی c وصل می‌کند، هستند.

پرسش شبیه‌سازی ۷. در یک مدل ساده، اگر تعداد فیلم‌هایی که دو کاربر به صورت مشترک دیده‌اند از یک مقدار آستانه بالاتر باشد، بین آن دو یک یال با وزن ۱ رسم می‌شود و در غیر این صورت بین آن دو کاربر یالی رسم نمی‌شود. فرض کنید روابط هم‌سلیقگی بین ۱۶ کاربر یک VOD را به صورتی که گفته شد بررسی کرده‌ایم و گراف شکل ۱۱ حاصل شده است. برنامه‌ای بنویسید که



شکل ۱۱: مثالی از روابط هم‌سلیقگی بین افراد در یک VOD

الگوریتم Louvain را روی گراف شکل ۱۱ اجرا کند و خروجی تکرار اوّل تا پنجم را نمایش دهد.

به جای رابطه‌ی (۱) می‌توان از تعاریف دیگری هم برای Q استفاده کرد، به عنوان مثال:

$$Q = \frac{۱}{۲m}\sum_{c}(e_c - \gamma\frac{{K_c}^۲}{۲m}). \qquad (۲)$$

در رابطه‌ی (۲)، e_c تعداد یال‌های خوشه‌ی c و K_c مجموع درجات رئوسِ خوشه‌ی c هستند. همچنین m مجموع وزن تمام یال‌های گراف است.

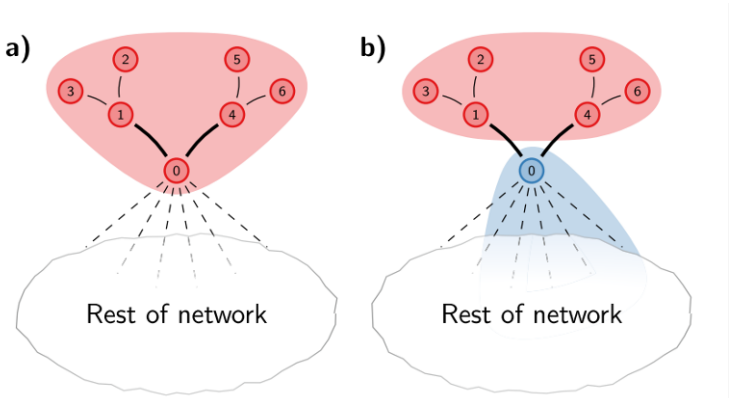
پرسش شبیه‌سازی ۸. پرسش شبیه‌سازی ۷ را با تعریفی از Q که در رابطه‌ی (۲) بیان شده است، تکرار کنید. مقدار $\gamma > ۰$ را چند مقدار مختلف قرار دهید و تأثیر آن را گزارش کنید.

پرسش شبیه‌سازی ۹. برای مجموعه‌داده‌ی Zachary's Karate Club که در میان فایل‌های پیوست مربوط به فاز اوّل پروژه به شما داده شده است، خوشه‌ها را با استفاده از الگوریتم Louvain به دست آورید. گراف اصلی و گراف خوشه‌بندی شده و modularity را به دست آورید. برای این سوال می‌توانید از کتاب‌خانه‌های آماده استفاده کنید.

پرسش شبیه‌سازی ۱۰. در مورد مشکلات حذف یال‌های داخل خوشه‌ها[۶] و همچنین توقّف زود هنگام الگوریتم Louvain جست‌وجو و نتایج آن را بیان کنید. چرا معیار modularity می‌تواند باعث چنین مشکلاتی بشود؟ شکل ۱۲ را در این خصوص ببینید.

---

[6] internally disconnected community



شکل ۱۲: مثالی از مشکلات الگوریتم Louvain

۵ نکات مهم!

لطفاً به نکات زیر دقّت کنید:

۱. عنوان بخش‌های مختلف پروژه از آثار شعرا و بزرگان ادبیات ایران و جهان انتخاب شده است. این اشعار بی‌ربط به مفاهیمی که در هر بخش با آن‌ها برخورد می‌کنید نیستند.

۲. تمامی شبیه‌سازی‌ها باید با کمک زبان Python انجام شود. شما تنها مجاز به استفاده از کتاب‌خانه‌های `networkx`، `numpy`، `scipy`، `random` و `matplotlib` هستید. همچنین تنها در مواردی که ذکرشده استفاده از کتاب‌خانه‌ی `scikit-learn` مجاز است. اگر روی عنوان هر کتاب‌خانه کلیک کنید، به راهنمای آن کتاب‌خانه هدایت می‌شوید.

۳. تحویل پروژه به صورت گزارش و کدهای نوشته‌شده است. گزارش باید شامل پاسخ پرسش‌ها، تصاویر و نمودارها و نتیجه‌گیری‌های لازم باشد. توجه کنید که قسمت عمده بارم شبیه سازی را گزارش شما و نتیجه‌ای که از خروجی کد میگیرید دارد. همچنین تمیزی گزارش بسیار مهم است. کدها و گزارش را در یک فایل فشرده‌شده در سامانه‌ی درس‌افزار آپلود کنید.

۴. اگر برای پاسخ به پرسش‌ها، از منبعی (کتاب، مقاله، سایت و...) کمک گرفته‌اید، حتماً به آن ارجاع دهید.

۵. نوشتن گزارش کار با LaTeX نمره‌ی امتیازی دارد.

۶. پرسش‌های شبیه‌سازی با رنگ سبز و پرسش‌های تئوری با رنگ آبی مشخص شده‌اند.

۷. بخش‌های تئوری گزارش که در قالب پرسش‌ها طرح شده‌اند را می‌توانید روی کاغذ بنویسید و تصویر آن‌ها را در گزارش خود بیاورید، ولی توصیه‌ی برادرانه می‌کنم که این کار را نکنید!

۸. درصورت مشاهده‌ی تقلّب، نمره‌ی هردو فرد صفر منظور خواهد شد.

موفّق باشید!