باسمه تعالی

پروژه‌ی درس آمار و احتمال مهندسی

آشنایی با برخی روش‌های خوشه‌بندی گراف‌ها

آخرین مهلت تحویل:

۱۴ بهمن ۱۴۰۱

۱ مقدّمه

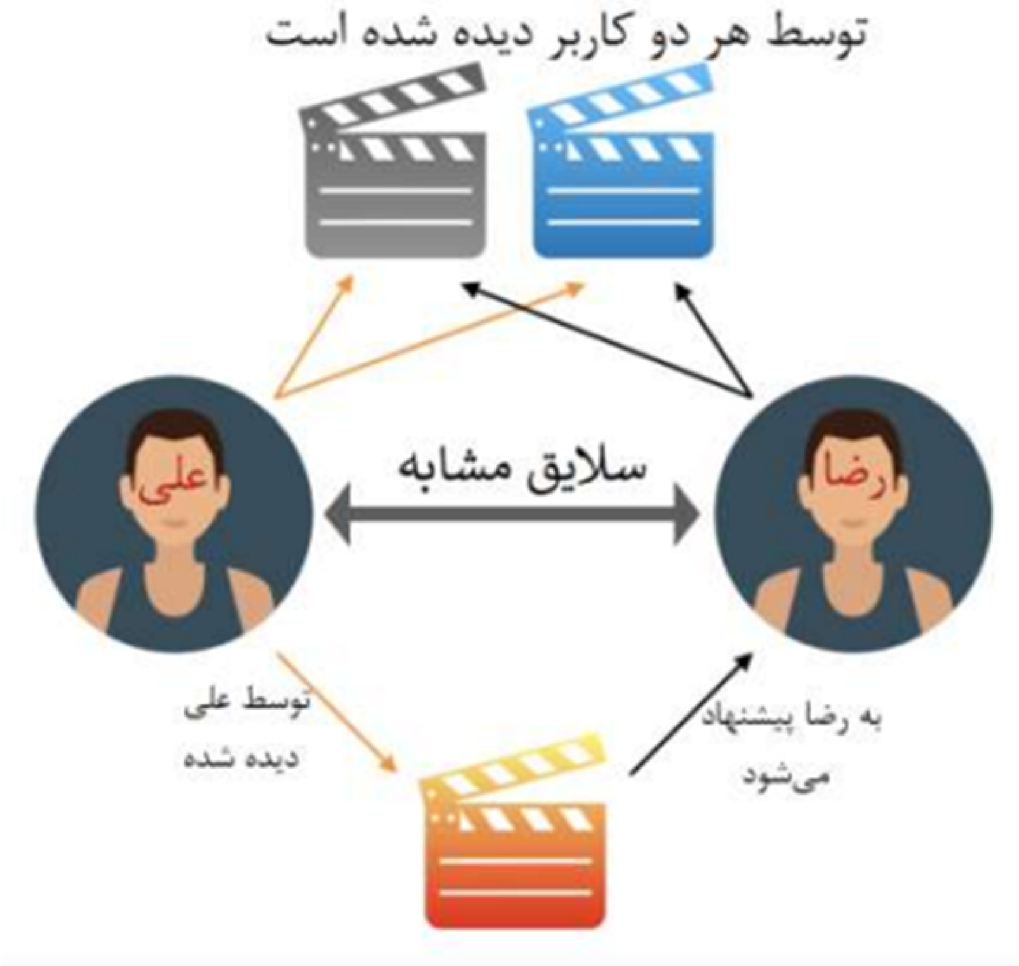
امتحانات تمام شده است و شما با خیالی آسوده می‌خواهید اوقات فراغت خود را بگذرانید تا با انرژی فراوان به سراغ ترم بعدی بروید. یک گزینه‌ی مناسب برای این کار، دیدن فیلم است. شما از طریق نرم‌افزار یا سایت وارد فیلیمو/نماوا/فیلم‌نت می‌شوید و در صفحه‌ی اوّل تعدادی پیشنهاد مشاهده می‌کنید که چندان هم بد نیستند و با سلیقه‌ی شما هم‌خوانی دارند. کم‌کم کنج‌کاو می‌شوید که بدانید این پیشنهادات بر چه اساسی به شما (و به همه‌ی کاربران) ارائه می‌شوند و در این‌جاست که فیلم را قطع می‌کنید و به سراغ پروژه‌ی شیرین درس آمار و احتمال می‌روید...

شکل ۱: برخی از سرویس‌های محتوای تصویری درخواستی

احتمالاً با سرویس‌های محتوای تصویری درخواستی[1] آشنایی دارید. یکی از ارکان مهمّ این سرویس‌ها ارائه‌ی پیشنهادهای مناسب به کاربران مختلف است. فرض کنید شما در یکی از این شرکت‌ها استخدام شده‌اید و می‌خواهید در قسمت پیشنهاد فیلم به کاربران فعالیت کنید. برای این منظور، ابتدا باید گروه‌هایی از کاربران که سلایق مشابه دارند (مثلاً همه‌ی آنها فیلم‌های طنز دوست دارند یا همه‌ی آنها فیلم‌های ترسناک دوست دارند یا ...) را پیدا کنید و سپس بر آن اساس به آن گروه‌ها فیلم‌های یکسان پیشنهاد دهید. در این پروژه می‌خواهیم ابتدا گروه‌های کاربران با سلایق مشابه را پیدا کرده و به هر گروه فیلم متناسب با سلایق آن‌ها را پیشنهاد دهیم. مثال ساده‌ای از این فرآیند در شکل ۲ نشان داده شده است.

یک توصیه‌ی دوستانه در همین ابتدای کار: قبل از هر کاری، بخش ۷ را بخوانید. بعد از آن، لااقل دوبار صورت پروژه را به طور کامل و با دقّت بخوانید، ولی پاسخ هیچ پرسشی را ننویسید. بعد از این که خوانش پروژه را به پایان رساندید، شروع به حلّ پروژه و نوشتن پاسخ پرسش‌ها کنید.

[1]Video On Demand (VOD)



شکل ۲: تأثیر فیلم‌های دیده‌شده توسط کاربران بر پیشنهادات سیستم

۲ ز هرچه رنگ تعلّق پذیرد آزاد است![۲]

ابتدا باید یک مدل ریاضی برای توصیف روابط بین افراد بیابیم. همان‌طور که احتمالاً تا الآن حدس زده‌اید، بهترین مدل ریاضی‌ای که می‌تواند ارتباط بین افراد مختلف در این مسئله را توصیف کند، گراف است. می‌توانیم هر فرد را با یک رأس گراف مدل کنیم و بین هر دو فرد هم‌سلیقه یک یال رسم کنیم. برای هر گراف، می‌توانیم یک ماتریس مجاورت تعریف کنیم.

تعریف ۱. 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:

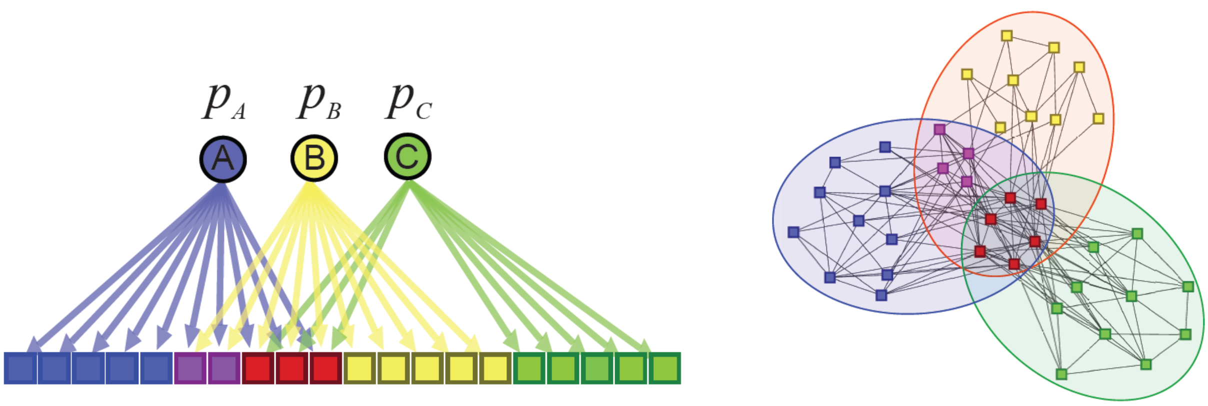
$$A_{i,j} = \begin{cases} ۱ & \text{اگر بین رئوس } i,j \text{ یال وجود داشته باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

هدف ما در این پروژه پیدا کردن گروه هایی از افراد است که سلیقه‌ی مشترک دارند.

پرسش تئوری ۱. اگر جامعه را بتوان به خوشه‌های مجزا از افرادی با سلیقه‌ی یکسان افراز کرد، به طوری که افرادِ یک خوشه با افرادِ خوشه‌ی دیگر هیچ اشتراکی نداشته باشند، آنگاه ماتریس مجاورت این افراد به چه صورت خواهد بود؟

پرسش تئوری ۲. مدلی که در پرسش قبلی به دست آوردید در واقعیت رخ نمی دهد، چرا که ممکن است یک فرد در یک خوشه با فردی در خوشه دیگر سلایق مشترکی (هرچند کم) داشته باشند. همچنین یک فرد می تواند به طور همزمان در چند خوشه عضو باشد. برای مثال یک فرد می تواند در خوشه‌ی افرادی که عموما به ژانر کمدی علاقه دارند باشد و همزمان در خوشه‌ی افرادی که ژانر درام میبینند نیز حضور داشته باشد. در این صورت چه توصیفی از ماتریس مجاورت می توان داشت؟

پرسش تئوری ۳. فرض کنید M ژانر مختلف فیلمِ $۱, ۲, \ldots, M$ داریم. اگر دو نفر به ژانر i–اُم علاقه داشته باشند، به احتمال p_i با یکدیگر هم سلیقه‌اند. حال فرض کنید دو شخص خاص هر کدام به چند ژانر مختلف علاقه دارند. اگر فرض کنیم علاقه به ژانرهای مختلف از هم مستقلند، احتمال اینکه این دو نفر با یکدیگر هم سلیقه باشند چقدر است؟
چرا اگر دو نفر در جوامع خاص زیادی عضو باشند احتمال اینکه با هم در ارتباط باشند بیشتر می‌شود؟ آیا می‌توانید این مسئله را به صورت شهودی هم توجیه کنید؟ به نظر شما این مدل چه نقصی در مدل کردن رابطه‌ی افراد و گروه‌ها دارد؟ شکل ۳ را در این خصوص ببینید.



شکل ۳: مثالی از علاقه‌ی افراد مختلف به ژانرهای مختلف

برای رفع مشکلاتی که مدل‌های قبلی داشتند، باید مدلی برای پیدا کردن خوشه‌ها پیشنهاد دهیم که دارای ۳ ویژگی باشد:

[۲] غلام همّت آنم که زیر چرخ کبود/ ز هرچه رنگ تعلّق پذیرد آزاد است [حافظ]

۱. پارامتری برای کمّی‌کردن میزان تمایل افراد به عضویت در یک گروه در نظر می‌گیریم. این پارامتر را «تعلّق» می‌نامیم. تعلّق یک فرد مانند u به خوشه‌ی c عبارت است از میزان تمایل او به عضویت در خوشه‌ی c. این پارامتر را با $F_{uc} \in [۰, \infty)$ نشان می‌دهیم.

۲. هر چه دو فرد تعلّق بیشتری به یک خوشه داشته باشند احتمال هم‌سلیقگی این دو فرد افزایش می‌یابد. برای این کار می‌توانیم احتمال ایجاد ارتباط هم‌سلیقگی بین دو فرد توسط خوشه‌ی c را به صورت $P_{uv}(c) = ۱ - \exp(-F_{uc}F_{vc})$ در نظر بگیریم.

۳. خوشه‌های مختلف به صورت مستقل با احتمال بند قبل بین افراد ارتباط هم‌سلیقگی ایجاد می‌کنند.

اگر تعداد افراد را n و تعداد خوشه‌ها را C فرض کنیم، با کنار هم قرار دادن تعلّق افراد مختلف به خوشه‌های مختلف می‌توانیم ماتریس تعلّق $\mathbf{F} \in \mathbb{R}^{n \times C}$ را تعریف کنیم.
با کمک ماتریس $\mathbf{F}$ می‌توان احتمال وجود ارتباط هم‌سلیقگی میان دو نفر را محاسبه کرد.

پرسش تئوری ۴. با توجه به توضیحات فوق احتمال وجود ارتباط میان دو فرد u, v را برحسب درایه‌های ماتریس تعلّق $\mathbf{F}$ محاسبه کنید. نشان دهید اگر دو فرد در گروه‌های بیشتری اشتراک داشته باشند این احتَمال افزایش می‌یابد.

دقّت کنید که اگر درایه‌های ماتریس $\mathbf{F}$ را داشته باشیم، میزان تمایل هر فرد به هر دسته فیلم را داریم و در نتیجه می‌توانیم به طور بهینه به افراد مختلف فیلم‌های مختلف را پیشنهاد دهیم. ولی در واقعیت به ماتریس $\mathbf{F}$ دسترسی نداریم و باید با استفاده از وجود یا عدم وجود ارتباط بین افراد مختلف آن را تخمین بزنیم. برای این کار از تخمین‌گر بیشترین درست‌نمایی کمک می‌گیریم و مشابه اغلب مسائل تخمین، به جای بیشینه کردن تابع درست‌نمایی، لگاریتم آن را بیشینه می‌کنیم.

پرسش تئوری ۵. می‌دانیم با داشتن ماتریس مجاورت $\mathbf{A}$ می‌توان گراف روابط هم‌سلیقگی بین افراد را به طور یکتا یافت (به تعریف ۱ مراجعه کنید).
تابع log-likelihood که به صورت $l(\mathbf{F}) = \log(\mathbb{P}[\mathbf{A}|\mathbf{F}])$ تعریف می‌شود را محاسبه نمایید.

پیدا کردن بیشینه‌ی تابع فوق در حالت کلّی کار دشواری است. بنابراین باید به صورت عددی آن را بیشینه کنیم. برای این کار، در هر مرحله تنها تعلّق‌های یک شخص خاص از بین n نفر به گروه‌های مختلف را کمی در راستای گرادیان جابجا می‌کنیم. توجّه می‌کنیم که تعلّق‌های شخص u در سطر u–اُم از ماتریس $\mathbf{F}$ که آن را با $F_{u,:}$ نشان می‌دهیم قرار دارد و در نتیجه گرادیان مذکور عبارت است از:

$$\nabla_{F_{u,:}} l(\mathbf{F}) = [\frac{\partial l(\mathbf{F})}{\partial F_{u,۱}}, \frac{\partial l(\mathbf{F})}{\partial F_{u,۲}}, \dots, \frac{\partial l(\mathbf{F})}{\partial F_{u,C}}].$$

پرسش تئوری ۶. به طور شهودی استدلال کنید که چرا این روش می‌تواند ما را به مقدار $\mathbf{F}$ بهینه، یعنی $\mathbf{F}^* = \arg\max_{\mathbf{F}} l(\mathbf{F})$ برساند.

پرسش تئوری ۷. $\nabla_{F_{u,:}} l(\mathbf{F})$ را محاسبه کنید.

به این ترتیب با به‌روز کردن سطرهای مختلف ماتریس $\mathbf{F}$ در راستای گرادیان لگاریتم تابع درست‌نمایی نسبت به همان سطرها در مراحل متعدد می‌توانیم به نقطه‌ی ماکزیمم نزدیک شویم.

پرسش شبیه‌سازی ۱. در نمونه کد زیر الگوریتم تکراری فوق برای تخمین $\mathbf{F}$، در تابع train پیاده سازی شده است که ماتریس مجاورت و تعداد گروه‌ها را به عنوان ورودی دریافت کرده و تخمینی از ماتریس تعلّق $\mathbf{F}$ مانند $\hat{\mathbf{F}}$ را به عنوان خروجی برمی‌گرداند. بخش های مربوط به تابع $l(\mathbf{F})$ و تابع گرادیان را تکمیل کنید.

```
def log_likelihood(F, A):
  #todo
  return log_likelihood

def gradient(F, A, i):
  #todo gradient of log_likelihood respect to person i parameters (F_ic)
  return gradient

def train(A, C, iterations = 200):
  # initialize an F
  N = A.shape[0]
  F = np.random.rand(N,C)

  for n in range(iterations):
    for person in range(N):
      grad = gradient(F, A, person)
      F[person] += 0.005*grad                    # updating F
      F[person] = np.maximum(0.001, F[person]) # F should be nonnegative
    ll = log_likelihood(F, A)
    print('At step %4i logliklihood is %5.4f'%(n,ll))

  return F
```

تعداد تکرارها را می‌توان این گونه تعیین کرد که اختلاف $l(\hat{\mathbf{F}})$ در دو تکرار متوالی کمتر از یک مقدار آستانه مانند $\epsilon = ۰/۰۰۱$ شود. به این ترتیب ما می‌توانیم ماتریس تعلّق $\mathbf{F}$ را تخمین بزنیم.

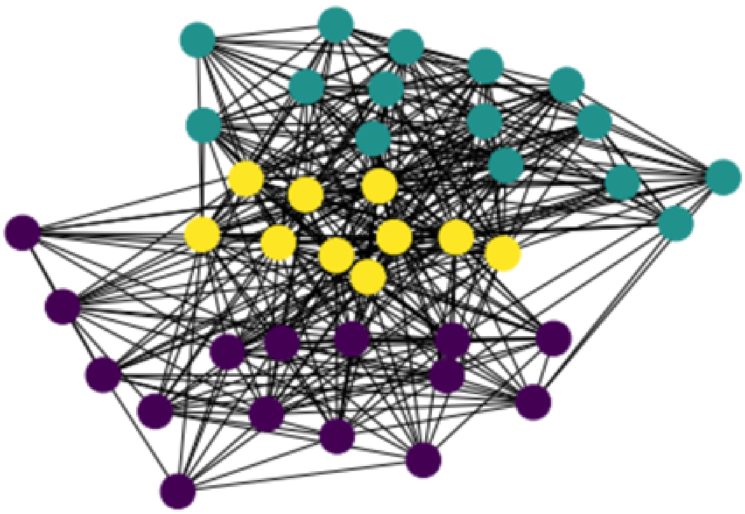
حال می‌خواهیم مشخّص کنیم هر فرد در کدام خوشه‌ها قرار می‌گیرد. برای این کار یک مقدار آستانه مانند δ تعیین می‌کنیم و فرد u را عضو خوشه‌ی c می‌گیریم اگر $F_{uc} > \delta$ باشد. یک روش برای تعیین مقدار δ آن است که آن را طوری تعیین کنیم که احتمال وجود ارتباط بین دو فرد هم‌سلیقه (که حداّقل برابر با $۱ - e^{-\delta^۲}$ است)، از احتمال ارتباط تصادفی آن‌ها، ε، (یعنی ارتباطی که ناشی از هم‌سلیقه‌بودن نباشد) بیشتر باشد. هم‌چنین احتمال ارتباط تصادفی بین دو فرد را می توان به این صورت تعیین کرد: زیرمجموعه‌ای تصادفی از افراد را در نظر می‌گیریم و تعداد نسبی دوتایی‌هایی از افراد که باهم ارتباط دارند را محاسبه می‌کنیم.

پرسش تئوری ۸. حال فرض کنید بخواهیم ارتباط بین دو فرد را از حالت صفر و یکی خارج کرده و کمی دقیق‌تر بررسی کنیم. برای مثال تعداد فیلم‌های مشترک دو فرد که یک عدد صحیح است را به عنوان معیاری از میزان ارتباط آن‌ها در نظر می‌گیریم. با در نظر گرفتن یک توزیع مناسب برای ارتباط دو فرد u, v بر حسب F_{uc}, F_{vc} (به جای توزیع برنولی که تا حالا با آن کار کرده‌ایم) ، تابع درست‌نمایی بازنویسی کنید و مجدّداً $\nabla_{F_{u,:}} l(\mathbf{F})$ را محاسبه کنید.

در ادامه یک نمونه از ۴۰ نفر تولید می‌کنیم که ۲۵ نفر اول در یک خوشه بوده و ۲۵ نفر آخر در خوشه‌ی دوم باشند. سپس ماتریس $\mathbf{A}$ را به تابع داده تا خوشه‌ها را تشخیص دهد. مشاهده می‌شود همه‌ی گروه‌ها به درستی تشخیص داده شده‌اند. (افرادی که در دو گروه هستند با احتمال بیشتری با یک‌دیگر ارتباط دارند که در شکل با رنگ زرد مشخص شده‌اند).

```
#testing in two small groups
A=np.random.rand(40,40)
A[0:15,0:25]=A[0:15,0:25]>1- 0.6    # connection prob people with 1 common group
A[0:15,25:40]=A[0:15,25:40]>1-0.1  # connection prob people with no common group
A[15:40,25:40]=A[15:40,25:40]>1-0.7 # connection prob people with 1 common group
A[15:25,15:25]=A[15:25,15:25]>1-0.8 # connection prob people with 2 common group
for i in range(40):
  A[i,i]=0
  for j in range(i):
    A[i,j]=A[j,i]
```

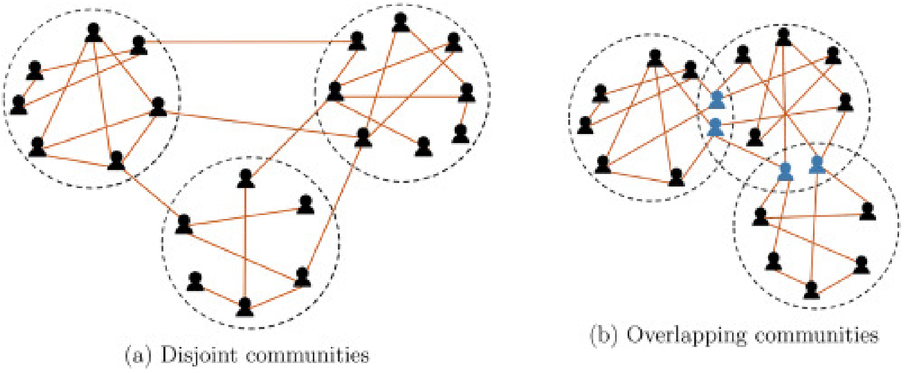
```

import matplotlib.pyplot as plt
import networkx as nx
plt.imshow(A)
delta=np.sqrt(-np.log(1-0.1)) # epsilon=0.1
F=train(A, 2, iterations = 120)
print(F>delta)
G=nx.from_numpy_matrix(A)
C=F>delta    # groups members
nx.draw(G,node_color=10*(C[:,0])+20*(C[:,1]))
```



شکل ۴: خروجی الگوریتم تشخیص خوشه‌ها

۳ گناه بخت پریشان و دست کوته ماست![۳]

در این قسمت مدل احتمالاتی دیگری برای ارتباط میان افراد و دسته‌های آن‌ها در نظر می‌گیریم.
فرض کنید در رستورانی تعدادی میز (محدود) وجود دارد و هر میز، تعداد مشخصی ظرفیت برای نشستن افراد دارد. هر نفر که به رستوران وارد میشود می‌تواند یکی از میزها را انتخاب کند و پشت آن میز بنشیند. واضح است که با وجود محدود بودن تعداد میزها، انتخاب‌های نفرات بعدی به تدریج محدود می‌شود. از طرف دیگر، هر فرد تمایل دارد که سر میزی بنشیند که تعداد بیشتری از دوستانش در آن میز باشند. یعنی هر فرد دوستی بیشتری با افراد سر میز خود دارد، اما به این معنا نخواهد بود که با افراد سایر میزها، هیچ دوستی نداشته باشد.



شکل ۵: مدل احتمالاتی ارتباطات بین افراد

حالا سعی می‌کنیم ارتباطات دوستی بین افراد رستوران را به یک گراف نسبت دهیم. برای این کار فرض‌های زیر را در نظر می‌گیریم:

۱. تعداد افراد حاضر در رستوران را $n = ۱۵$ نفر و تعداد میزها را $k = ۳$ فرض کنید.

۲. فرض می‌کنیم که توزیع افراد در میزهای مختلف همگن است، یعنی تعداد افرادی که برای هر میز اختصاص می‌دهیم مساوی است.

۳. شماره‌ی میزی که هر فرد پشت آن نشسته است را در بردار $\mathbf{z}$ درج می‌کنیم. در این صورت داریم $\mathbf{z} \in \mathbb{R}^n$ و به ازای هر $i \in \{۱, ۲, \ldots, n\}$ داریم $z_i \in \{۱, ۲, \ldots, k\}$.

۴. فرض کنید بردار $\mathbf{z}_۰$ به عنوان نحوه‌ی صحیح خوشه‌بندی در اختیار ماست:

$$\mathbf{z}_۰ = [۳, ۱, ۲, ۱, ۳, ۱, ۲, ۲, ۲, ۳, ۳, ۲, ۱, ۱, ۳]^\top$$

بردار $\mathbf{z}_۰$ برای هر فرد مشخص می‌کند که او پشت کدام میز نشسته است.

۵. حالا باید به روابط دوستی میان افراد برسیم. طبق فرض ابتدایی، در میان افراد سر یک میز، تعداد دوستان بیشتری حضور دارند تا افرادی که سر یک میز نیستند. به زبان احتمالاتی، احتمال دوست بودن هر فرد با شخصی که سر میز خود نشسته، بیشتر از کسی است که سر میز دیگری نشسته باشد.

۶. احتمال دوستی افراد یک میز را $p = ۰/۶$ و احتمال دوستی افرادی که پشت یک میز نیستند را $q = ۰/۱$ در نظر بگیرید.

۷. ماتریس $\mathbf{Q} \in \mathbb{R}^{k \times k}$ را با درایه‌های زیر تشکیل دهید:

$$Q_{i,j} = \begin{cases} p & i = j \\ q & i \neq j \end{cases}$$

که i, j شماره‌ی میزهای مختلف است.

[۳] اگر به زلف دراز تو دست ما نرسد/ گناه بخت پریشان و دست کوته ماست [حافظ]

پرسش تئوری ۹. با توجّه به توضیحات بالا، درایه‌ی $A_{i,j}$ از ماتریس مجاورت (می‌توانید مجدّداً به تعریف ۱ مراجعه کنید!) گراف روابط دوستی افراد یک متغیّر تصادفی است. این متغیّر تصادفی را توصیف کنید و تابع چگالی/جرم احتمال آن را بیابید.

پرسش تئوری ۱۰. آیا ماتریسی که ساخته‌اید توصیف کاملی از روابط دوستی افراد ارائه می‌دهد؟ به عنوان مثال می‌توانید درایه‌های روی قطر اصلی را بررسی کنید یا به استقلال یا وابستگی درایه‌های $A_{i,j}, A_{j,i}$ از یک‌دیگر فکر کنید. ماتریس $\mathbf{A}$ را با یافته‌های جدید توصیف کنید.

پرسش شبیه‌سازی ۲. از ماتریسی که در پرسش تئوری ۱۰ توصیف کردید، ۱۰ نمونه بسازید.

پرسش شبیه‌سازی ۳. از ماتریسی که در پرسش تئوری ۱۰ توصیف کردید، یک نمونه بسازید و گراف روابط دوستی میان افراد را بر اساس آن تشکیل دهید. همچنین روی گراف نمایش داده‌شده، شماره‌ی میز هر فرد را روی رأس مربوط به او با رنگ یا شماره مشخّص کنید. دقّت کنید که گراف حاصل بهتر است هم‌بند باشد (رأس منفرد نداشته باشد)، اگر خلاف این را مشاهده کردید، گراف را از نو بسازید.

حالا ما یک مجموعه از افراد داریم و روابط دوستی میان آن‌ها نیز مشخّص است. هدف نهایی این است که از روی گراف روابط دوستی بین افراد، گروه‌بندی افراد (شماره‌ی میزی که هر فرد پشت آن نشسته است) را تشخیص دهیم. به تعبیر دیگر، جواب نهایی مسئله بردار $\mathbf{z}_0$ است که فرض می‌کنیم آن را در اختیار نداریم و می‌خواهیم از روی درایه‌های گراف $\mathbf{A}$ به تخمینی از $\mathbf{z}_0$، مانند بردار $\hat{\mathbf{Z}}_0 \in \mathbb{R}^k$ برسیم.

ابتدا لازم است معیاری برای سنجش میزان دقّت تخمین معرّفی کنیم. از فاصله‌ی همینگ[4] استفاده می‌کنیم. این فاصله به صورت زیر تعریف می‌شود:

$$d_H : \mathbb{R}^k \times \mathbb{R}^k \to \{1, 2, \ldots, k\}$$
$$d_H(\mathbf{z}_1, \mathbf{z}_2) = \sum_{i=1}^{k} \mathbb{1}\{[\mathbf{z}_1]_i \neq [\mathbf{z}_2]_i\}.$$

که در تعریف بالا، تابع $\mathbb{1}\{.\}$ برای یک پیشامد به صورت زیر تعریف می‌شود:

$$\mathbb{1}\{B\} = \begin{cases} 1 & B \text{ is true} \\ 0 & B \text{ is false} \end{cases}.$$

پرسش شبیه‌سازی ۴. تابعی بنویسید که دو ورودی $\mathbf{z}_1, \mathbf{z}_2$ را بگیرد و فاصله‌ی همینگ میان آن‌ها را محاسبه کند.

معیار فاصله‌ی همینگ یک نقطه‌ی ضعف دارد و آن نقطه‌ضعف، وابستگی این فاصله به نام خوشه‌ها (یا شماره‌ی میزها) است. خوشه بندی باید از نام خوشه‌ها مستقل باشد. تفاوتی وجود ندارد که نام یک خوشه چیست، مهم ساختاری است که به آن رسیده‌ایم. در مثال ما، اگر نام میزها را تغییر دهیم نباید تفاوتی در خوشه بندی ایجاد شود، اگر میزها را به ترتیب ۱، ۲، ۳ نام بگذاریم یا ۱، ۳، ۲، ماهیت خوشه بندی تفاوتی نکرده‌است، تنها شماره‌ای که اختصاص داده‌ایم تفاوت کرده است. در واقع معیار سنجش میزان دقّت تخمین باید نسبت به جایگشت‌های مختلف نامگذاری مستقل باشد و با تغییر آن عوض نشود.

[4] Hamming distance

حال باید کاری کنیم که فاصله‌ی همینگ را از نام خوشه‌ها مستقل کند. برای مثال، در یک جامعه با $n=۶$ عضو و $k=۲$ گروه، فرض کنید خوشه بندی درست به صورت زیر باشد:

$$\mathbf{z}_۰ = [۲,۲,۱,۲,۱,۱]^\top.$$

و ما به تخمین زیر از $\mathbf{z}_۰$ برسیم:

$$\hat{\mathbf{Z}}_۰ = [۱,۱,۲,۱,۲,۲]^\top.$$

اگر از تابعی که در پرسش شبیه‌سازی ۴ نوشتید برای محاسبه‌ی فاصله‌ی این دو بردار استفاده کنیم، خواهیم داشت $d_H(\hat{\mathbf{Z}}_۰,\mathbf{z}_۰)=۶$، ولی اگر جای برچسب ۰ و ۱ را در بردار $\hat{\mathbf{Z}}_۰$ تغییر دهیم و بردار حاصل را $\tilde{\mathbf{Z}}_۰$ بنامیم، داریم $d_H(\tilde{\mathbf{Z}}_۰,\mathbf{z}_۰)=۰$.

برای یک بردار $\mathbf{z}$، مجموعه‌ی برداری $\langle \mathbf{z}\rangle$ را مجموعه تمام بردارهایی که با تغییر ترتیب شماره‌ی خوشه‌ها در $\mathbf{z}$ می‌توان به آن‌ها رسید تعریف می‌کنیم. به عنوان مثال در یک جامعه‌ی $n=۶$ عضوی با $k=۳$ خوشه‌ی مختلف، بردار زیر را در نظر بگیرید:

$$\mathbf{z} = [۳,۱,۲,۲,۳,۱]^\top.$$

می‌توانیم جایگشت‌های دیگری از نام خوشه‌ها را در نظر بگیریم و مجموعه‌ی $\langle \mathbf{z}\rangle$ را تشکیل دهیم:

$$\langle \mathbf{z}\rangle = \{[۳,۱,۲,۲,۳,۱]^\top, [۳,۲,۱,۱,۳,۲]^\top, [۱,۳,۲,۲,۱,۳]^\top, \dots\}.$$

پس تابع فاصله‌ی مستقل از نام خوشه‌ها را به صورت زیر تعریف می‌کنیم:

$$\begin{aligned} d &: \mathbb{R}^k \times \mathbb{R}^k \to \{۱,۲,\dots,k\} \\ d(\mathbf{z}_۱,\mathbf{z}_۲) &= \min_{\tilde{\mathbf{z}}_۲\in\langle \mathbf{z}_۲\rangle} d_H(\mathbf{z}_۱,\tilde{\mathbf{z}}_۲). \end{aligned}$$

پرسش شبیه‌سازی ۵. تابعی بنویسید که فاصله‌ی مستقل از نام خوشه‌ها را محاسبه کند. این فاصله را فاصله‌ی همینگ کمینه می‌نامیم.

حالا می‌خواهیم با کمک تخمین بیشترین درست‌نمایی از روی یک تحقّق ماتریس مجاورت، بردار $\mathbf{z}_۰$ را تخمین بزنیم.

پرسش تئوری ۱۱. با توجّه به پرسش‌های تئوری ۹ و ۱۰، در ماتریس $\mathbf{A}$ چند درایه‌ی مستقل از هم وجود دارند؟

پرسش تئوری ۱۲. احتمال تحقّق ماتریس $\mathbf{A}$ به شرط بردار $\mathbf{z}$ را بنویسید. عبارتی که رسیده‌اید همان تابع درست‌نمایی است که آن را با $L(\mathbf{z})=\mathbb{P}[\mathbf{A}|\mathbf{z}]$ نشان می‌دهیم.

پرسش تئوری ۱۳. تابع لگاریتم درست‌نمایی، یعنی $l(\mathbf{z})=\log\left(L(\mathbf{z})\right)$ را محاسبه کنید.

در ادامه تلاش می‌کنیم تا $l(\mathbf{z})$ را بیشینه کنیم. این معادل با آنست که $\tilde{l}(\mathbf{z})=-l(\mathbf{z})$ را کمینه کنیم.

پرسش شبیه‌سازی ۶. تابعی بنویسید که با دریافت $\mathbf{z}$ و $\mathbf{A}$ مقدار عبارت $\tilde{l}(\mathbf{z})=-\log(\mathbb{P}[\mathbf{A}|\mathbf{z}])$ را محاسبه کند.

یافتن پاسخ بهینه‌ی مسئله به صورت تئوری کار راحتی نیست (یک بار دیگر عنوان بخش و پاورقی آن را بخوانید!). در نتیجه باید از روش‌های عددی برای کمینه‌کردن آن استفاده کنیم. ابتدا فرض می‌کنیم که تعداد خوشه‌ها و تعداد اعضای هر خوشه را می‌دانیم. یعنی می‌دانیم که هر درایه از بردار $\mathbf{z}_۰\in\mathbb{R}^n$ عددی از مجموعه‌ی $\{۱,۲,۳\}$ است و دقیقاً ۵ درایه از این بردار برابر با ۱، ۵ درایه برابر با ۲ و ۵ درایه هم برابر با ۳ هستند. در نتیجه تنها چیزی که نمی‌دانیم جایگشت دقیق این درایه‌ها است که می‌خواهیم آن را با کمک ماتریس مجاورت تخمین بزنیم. برای این کار از الگوریتم زیر کمک می‌گیریم:

۱. ابتدا یک تخمین اولّیه از $\mathbf{z}_۰$ به صورت زیر تعریف کنید:

$$\hat{\mathbf{z}}_۰ = [۱, ۱, ۱, ۱, ۱, ۲, ۲, ۲, ۲, ۲, ۳, ۳, ۳, ۳, ۳]$$

۲. مقدار $\tilde{l}(\hat{\mathbf{z}}_۰)$ را محاسبه کنید.

۳. برای $t = ۱, \dots, T$:

(آ) برای $i = ۱, \dots, n$:
جای $[\hat{\mathbf{z}}_{t-۱}]_i$ را با تک‌تک درایه‌های $\hat{\mathbf{z}}_{t-۱}$ عوض کنید. در هر مرحله مقدار تابع $\tilde{l}$ را به ازای بردار جدید به دست آمده محاسبه کنید.

(ب) جابجایی‌ای که بیشترین کاهش در $\tilde{l}(\hat{\mathbf{z}}_{t-۱})$ را نتیجه می‌دهد را روی بردار $\hat{\mathbf{z}}_{t-۱}$ اعمال کنید و به بردار $\hat{\mathbf{z}}_t$ برسید.

۴. بردار $\hat{\mathbf{z}}_T$ را به عنوان خروجی اعلام کنید.

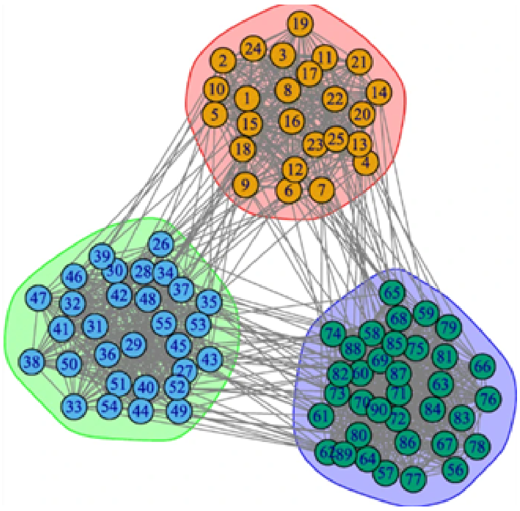
پرسش شبیه‌سازی ۷. الگوریتم بالا را شبیه‌سازی کنید. در هر مرحله میزان $\tilde{l}(\hat{\mathbf{z}}_t)$ و همچنین $d(\hat{\mathbf{z}}_t, \mathbf{z}_۰)$ را ذخیره کنید و در نهایت روی نمودار نمایش دهید. در انتخاب پارامتر T مختارید ولی آن را عددی مناسب انتخاب کنید.

پرسش شبیه‌سازی ۸. الگوریتم را به ازای $N = ۱۰$ نقطه‌ی شروع $(\hat{\mathbf{z}}_۱)$ متفاوت اجرا کنید و نتایج را مقایسه کنید (می‌توانید یک بار دیگر عنوان این بخش و پاورقی آن را بخوانید!).

پرسش شبیه‌سازی ۹. در پرسش قبلی، $N = ۱۰$ خروجی مختلف به دست می‌آورید. این خروجی‌ها را با $\{\hat{\mathbf{z}}_T^{(j)}\}_{j=۱}^N$ نشان می‌دهیم. مقدار $\tilde{l}(\hat{\mathbf{z}}_T^{(j)})$ به ازای jهای مختلف را با $\tilde{l}(\mathbf{z}_۰)$ مقایسه کنید. آیا حالتی وجود دارد که $\tilde{l}(\hat{\mathbf{z}}_T^{(j)}) = \tilde{l}(\mathbf{z}_۰)$ شده باشد؟ در این حالت مقدار فاصله‌ی همینگ کمینه‌ی بردار تخمین و $\mathbf{z}_۰$ چقدر است؟

پرسش شبیه‌سازی ۱۰. آیا jای وجود دارد که $d(\hat{\mathbf{z}}_T^{(j)}, \mathbf{z}_۰) = ۰$ شده باشد؟

پرسش شبیه‌سازی ۱۱. از روی $\mathbf{z}_۰$، دو نمونه‌ی دیگر از ماتریس $\mathbf{A}$ بسازید و در هرکدام به ازای $N = ۱۰$ بردار اولّیه‌ی مختلف، تلاش کنید که $\mathbf{z}_۰$ را تخمین بزنید. بهترین نتیجه را گزارش کنید.



شکل ۶: مثالی از خوشه‌بندی افراد

۴ او نه خیال است و نه طیف![۵]

در این بخش به بررسی الگوریتم دیگری برای یافتن خوشه‌های هم‌سلیقه در یک جامعه می‌پردازیم.

شکل ۷: مثالی از خوشه‌بندی افراد

ابتدا فرض می‌کنیم تنها دو ژانر فیلم وجود دارد و هرکدام از افراد دقیقاً طرفدار یکی از این دو ژانر هستند. همچنین مشابه بخش قبل فرض می‌کنیم احتمال وجود ارتباط (یا همان تشابه سلیقه) بین طرفداران یک ژانر برابر با p و احتمال وجود ارتباط بین طرفداران یک ژانر و ژانر دیگر برابر با q باشد. منطقی است که فرض کنیم $p > q$، یعنی احتمال تشابه سلیقه بین دو فرد که طرفدار یک ژانر خاص هستند از احتمال تشابه سلیقه بین دو فرد که طرفدار ژانرهای مختلف هستند بیشتر است.

با این مفروضات ماتریس مجاوِرت (می‌توانید مجدّداً به تعریف ۱ مراجعه کنید!) یک ماتریس تصادفی است، به این معنا که هرکدام از درایه‌های آن یک متغیّر تصادفی هستند.

پرسش تئوری ۱۴. درایه‌ی $A_{i,j}$ از ماتریس $\mathbf{A}$ چه متغیّر تصادفی‌ای است؟ تابع چگالی/جرم احتمال آن را محاسبه کنید.

پرسش تئوری ۱۵. ماتریس $\mathbf{W} = \mathbb{E}[\mathbf{A}]$ را در نظر بگیرید. درایه‌های این ماتریس به صورت زیر تعریف می‌شوند:

$$W_{i,j} = \mathbb{E}[A_{i,j}].$$

ماتریس $\mathbf{W}$ را محاسبه کنید.

می‌توانیم بنویسیم: $\mathbf{A} = \mathbf{W} + \mathbf{R}$، که $\mathbf{R}$ بیانگر نویز در مشاهده‌ی $\mathbf{W}$ است. ابتدا فرض می‌کنیم که ماتریس $\mathbf{W}$ را در اختیار داریم و تلاش می‌کنیم خوشه‌بندی را با استفاده از این ماتریس انجام دهیم.

پرسش تئوری ۱۶. ابتدا فرض کنید $n = ۴$ است و افراد ۱ و ۲ طرفدار ژانر ۱ و افراد ۳ و ۴ طرفدار ژانر ۲ هستند. در این حالت مقادیر و بردارهای ویژه‌ی ماتریس $\mathbf{W}$ را بیابید.

[۵] خود را چو نمود او نه خیال است و نه طیف/ تو چون و چگونه دانِیَش باشد حیف [اوحدالدّین کرمانی]

پرسش تئوری ۱۷. در حالت کلّی نشان دهید که $\mathbf{W}$ تنها دو مقدارویژه‌ی ناصفر دارد و بردارویژه‌های متناظر با این مقادیر ویژه را محاسبه کنید.

پرسش تئوری ۱۸. ماتریس $\mathbf{D_W}$ را یک ماتریس قطری در نظر می‌گیریم که درایه‌های روی قطر آن به صورت زیر تعریف می‌شوند:

$$[D_{\mathbf{W}}]_{i,i} = \sum_{j=1}^{n} W_{i,j}$$

با فرض $\mathbf{W}$ای که در قسمت‌های قبلی محاسبه کرده‌اید، $\mathbf{D_W}$ را محاسبه کنید.

پرسش تئوری ۱۹. ماتریس $\mathbf{L_W}$ را به صورت $\mathbf{L_W} = \mathbf{D_W} - \mathbf{W}$ تعریف می‌کنیم. مقادیر و بردارهای ویژه‌ی این ماتریس را محاسبه کنید.

پرسش تئوری ۲۰. دو بردارویژه‌ی متناظر با دو مقدارویژه‌ی کوچک‌تر ماتریس را در نظر بگیرید. در همان حالت ساده‌ی $n = 4$ که در پرسش تئوری ۱۶ مطرح شد، داده‌ها را با کمک این دو بردارویژه در یک فضای دوبُعدی نمایش دهید. به تعبیر دیگر اگر دو بردارویژه را با $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$ نشان دهیم، برای فرد i–اُم نقطه‌ی $([u_1]_i, [u_2]_i)$ را در نظر بگیرید. آیا با کمک این نقاط می‌توان افراد را دسته‌بندی کرد؟

الگوریتمی که قدم به قدم در پرسش‌های قبلی مشاهده کردید را خوشه‌بندی طیفی[6] می‌نامند. همان‌طور که مشاهده کردید، اگر ماتریس $\mathbf{W}$ را در اختیار داشته باشیم، می‌توانیم بدون خطا خوشه‌بندی را انجام دهیم. ولی در واقعیت تنها ماتریس $\mathbf{A}$ که نسخه‌ی آغشته به نویز ماتریس $\mathbf{W}$ است در اختیار ماست. اگر بخواهیم فرآیند طی‌شده را دقیقاً برای ماتریس $\mathbf{A}$ تکرار کنیم، یعنی ماتریس‌های $\mathbf{D_A}$ و $\mathbf{L_A}$ را تعریف کنیم و مراحل مطرح‌شده در پرسش‌های قبلی را انجام دهیم، خوشه‌بندی بر اساس بردارهای ویژه‌ی ماتریس $\mathbf{L_A}$ انجام خواهد شد. خوشبختانه می‌توان نشان داد که اگر تعداد اعضای جامعه به قدر کافی بزرگ باشد، بردارویژه‌های ماتریس $\mathbf{L_A}$ به بردارویژه‌های ماتریس $\mathbf{L_W}$ نزدیکند.

پرسش شبیه‌سازی ۱۲. ابتدا تلاش می‌کنیم تا با شبیه‌سازی، عملکرد این الگوریتم را مشاهده کنیم. تعداد افراد را $n = 10000$ در نظر می‌گیریم. همچنین فرض می‌کنیم که $p = 0/1$ و $q = 0/01$. ماتریس‌های $\mathbf{A}$ و $\mathbf{W}$ را تولید کرده و الگوریتم را روی آن‌ها اجرا کنید. نتیجه‌ی دو الگوریتم، یعنی میزان خطای الگوریتم در یافتن دسته‌ها را در دو حالت مشاهده، مقایسه و گزارش کنید. توجّه کنید که تخصیص افراد به گروه‌های مختلف را باید به صورت تصادفی انجام دهید، و بدیهی است که از یک تخصیص واحد باید برای ساختن ماتریس‌های $\mathbf{W}$ و $\mathbf{A}$ استفاده کنید.

پرسش تئوری ۲۱. می‌توان نشان داد که نامساوی زیر دست‌کم با احتمال $1 - 4e^{-n}$ برقرار است:

$$\exists \theta \in \{-1, 1\} : \qquad \sum_{j=1}^{n} |[u_2(\mathbf{L_W})]_j - \theta [u_2(\mathbf{L_A})]_j|^2 \leq \frac{C}{\mu^2}$$

به شرطی که مقادیر ویژه به صورت صعودی مرتّب شده باشند، در عبارت فوق $\mathbf{u}_2(\mathbf{L})$ بردارویژه‌ی متناظر با دومین مقدارویژه‌ی ماتریس $\mathbf{L}$ است، اندیس j عنصر j–اُم بردار را نشان می‌دهد و $\mathbf{L_A} = \mathbf{D_A} - \mathbf{A}$. همچنین $\mu = \min\{q, p - q\}$ یک مقدار ثابت است.

با استفاده از نامساوی فوق نشان دهید الگوریتم خوشه‌بندی طیفی با احتمال بیش از $1 - \epsilon(n)$ (در این‌جا $\epsilon(n)$ یعنی تابعی از n که شما باید آن را محاسبه کنید!) حدّاکثر تعداد ثابتی خطا انجام می دهد. پیچیدگی این الگوریتم را از نظر حداقل تعداد داده‌های مورد نیاز، n، برای دستیابی به احتمال $1 - \epsilon$ به دست آورید.

[6] Spectral Clustering

پرسش شبیه‌سازی ۱۳. پرسش شبیه‌سازی ۱۲ را با چند مقدار مختلف n تکرار کنید و تعداد خطاها را گزارش و مقایسه کنید.

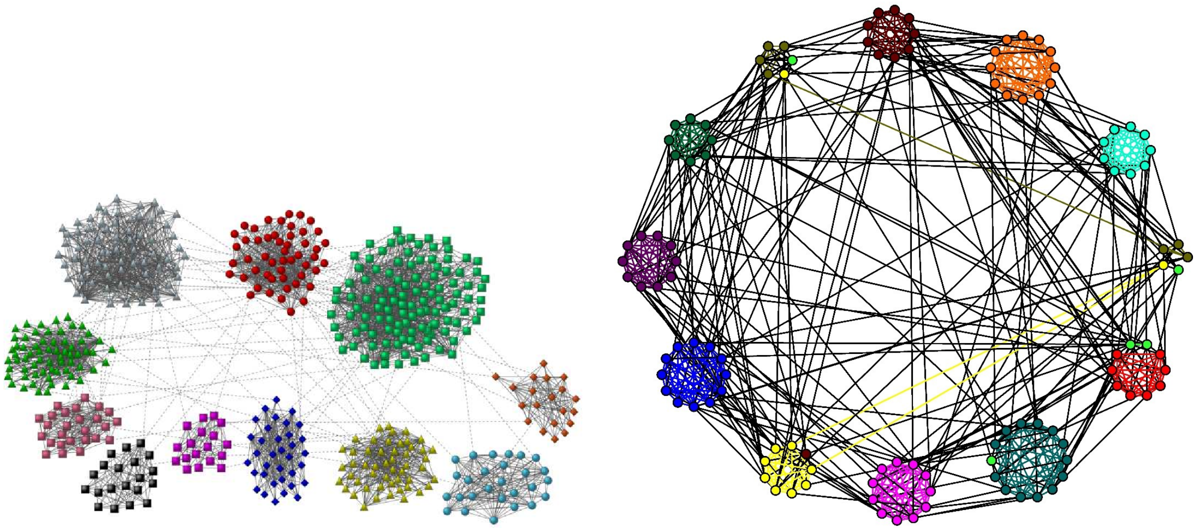
همان‌طور که مشاهده کرده‌اید، الگوریتم خوشه‌بندی طیفی به ما کمک می‌کند که داده‌ها را در یک دستگاه مختصات جدید نمایش دهیم که در این دستگاه، دسته‌های مختلف تا حدّ خوبی از هم جدا شده‌اند. تا کنون تنها حالت دو دسته‌ای را بررسی کرده‌ایم. ولی می‌توان این الگوریتم را به تعداد دسته‌های دلخواه هم تعمیم داد.

پرسش تئوری ۲۲. در رابطه با الگوریتم خوشه‌بندی طیفی در حالت $k \geq ۲$ تحقیق کنید و نتیجه‌ی تحقیقات خود را به صورت خلاصه در گزارش بیاورید. شما باید نمای کلّی الگوریتم و توضیحاتی مختصر در رابطه با تئوری مربوط به آن را بنویسید.

مشاهده کردیم که الگوریتم خوشه‌بندی طیفی با k دسته هم می‌تواند داده‌ها را به یک دستگاه مختصّات جدید و با ابعاد k منتقل کند که در این دستگاه، دسته‌ها نسبتاً از هم جدا شده‌اند. در این حالت نمی‌توانیم داده‌ها را نمایش دهیم (زیرا ابعاد آن‌ها ممکن است از ۳ بیشتر باشد) و باید از الگوریتمی استفاده کنیم که داده‌های ابعاد بالا را بگیرد و خوشه‌ها را استخراج کند. یکی از الگوریتم‌های شناخته‌شده برای این کار، الگوریتم k-means است. توضیحات مربوط به این الگوریتم برای مطالعه‌ی اختیاری شما در فایل درس‌نامه آورده شده است. برای استفاده از این الگوریتم می‌توانید از `sklearn.cluster.KMeans` استفاده کنید. اگر روی عنوان کتاب‌خانه کلیک کنید، به صفحه‌ی راهنمای این کتاب‌خانه هدایت می‌شوید.

پرسش شبیه‌سازی ۱۴. مجموعه داده‌ی Calfornia Housing که در فایلهای پیوست وجود دارد را مشاهده کنید. این مجموعه‌داده، مختصّات تعدادی خانه و میزان درآمد متوسّط ساکنان هر خانه را نشان می‌دهد. می‌خواهیم با کمک درآمد متوسّط که در پارامتر MedInc آمده است، خانه‌ها را خوشه‌بندی کنیم. برای این کار فاصله‌ی بین دو خانه‌ی i, j را به صورت $d(i, j) = |\mathrm{MedInc}_i - \mathrm{MedInc}_j|$ تعریف می‌کنیم. تعداد دسته‌ها را $k = ۳$ در نظر بگیرید و نتیجه‌ی حاصل از خوشه‌بندی را روی نقشه‌ی جغرافیایی خانه‌ها نمایش دهید. برای این کار خانه‌های مربوط به یک دسته را با رنگی متفاوت از دسته‌های دیگر نشان دهید. توجه کنید که برای همگرا شدن الگوریتم لازم است تعداد مراحل تکرار الگوریتم را به حد کافی زیاد بگذارید.

در نتیجه با ترکیب الگوریتم خوشه‌بندی طیفی و الگوریتم k-means می‌توان خوشه‌بندی را برای تعداد دسته‌های دلخواه انجام داد.



شکل ۸: مثالی از خوشه‌بندی افراد با کمک الگوریتم خوشه‌بندی طیفی و روش k–means

پرسش شبیه‌سازی ۱۵. مجموعه‌داده‌ی Zachary's Karate Club را در نظر بگیرید. این مجموعه داده ارتباطات میان افراد در یک باشگاه کاراته را نشان می‌دهد. با کمک الگوریتم خوشه‌بندی طیفی به همراه k-means این شبکه را دسته‌بندی کنید و نتیجه را به ازای تعداد خوشه های $k = ۲, ۳, ۴$ روی سه گراف مجزّا نمایش دهید.

پرسش شبیه‌سازی ۱۶. یکی از معیارهای سنجش برای بررسی کیفیت خوشه بندی و شناسایی جوامع، معیار رسانایی[7] خوشه‌های شناسایی شده در گراف است. تعریف این پارامتر برای یک خوشه به صورت زیر است:

$$f(S) = \frac{c_S}{۲m_S + c_S}$$

که در آن S خوشه‌ی شناسایی شده و m_S و c_S به صورت زیر تعریف می‌شوند:

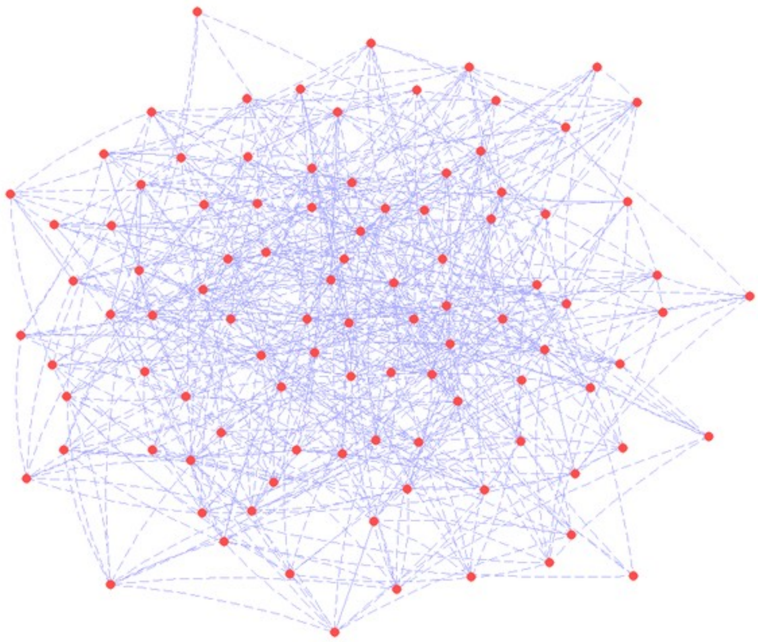
$$c_S = |\{(u, v) \in E : u \in S, v \notin S\}|,$$

$$m_S = |\{(u, v) \in E : u \in S, v \in S\}|.$$

در تعاریف فوق، E مجموعه‌ی تمامی یال‌های گراف است و نماد $|B|$ برای یک مجموعه مانند B به معنای تعداد اعضای مجموعه‌ی B است.
رسانایی میانگین خوشه های شناسایی شده در پرسش شبیه‌سازی ۱۵ را به ازای $k = ۲, ۳, \ldots, ۱۰$ محاسبه نمایید و در یک نمودار رسم کنید. با توجه به نمودار تعداد خوشه‌های مناسب برای این شبکه چند است؟

[7] Conductance

۵ من از دیار حبیبم نه از بلاد غریب![۸]



شکل ۹: مثالی از روابط افراد درون یک خوشه

در بخش قبل دیدیم که می‌توان گراف کاربران فیلیمو/نماوا/فیلم‌نت را با خوشه‌هایی مدل کرد که احتمال تشابه سلیقه بین هر دو نفر داخل یک خوشه برابر p و احتمال تشابه سلیقه بین افراد دو خوشه‌ی متفاوت برابر q است که $q < p$.

همانطور که گفتیم، داخل هر خوشه (مثلاً خوشه‌ی کاربرانی که فیلم‌های کمدی دوست دارند) احتمال تشابه سلیقه‌ی دو نفر برابر با p است. به همین دلیل می‌توان کاربران داخل هر خوشه را با یک گراف تصادفی مدل کرد. حال در این مساله، تمرکز خود را به روابط بین افراد داخل یکی (و تنها یکی) از این خوشه‌ها معطوف می‌کنیم. یک خوشه از جامعه یک گراف با n رأس است که بین هر دو رأس آن با احتمال p یال وجود دارد و با احتمال $1 - p$ یال وجود ندارد.

پرسش تئوری ۲۳. فرض کنید در خوشه‌ی مورد بررسی m رابطه‌ی دوستی بین افراد وجود دارد. ما بر اساس مدل گراف تصادفی که توصیف کردیم، روابط دوستی بین افراد را به صورت تصادفی ایجاد می‌کنیم. چقدر احتمال دارد که تمام روابط هم‌سلیقگی را به درستی تعیین کرده باشیم؟ جواب شما باید بر حسب n, p, m باشد.

پرسش تئوری ۲۴. تنها در این پرسش، فرض کنید مقدار دقیق m را می‌دانیم و در نتیجه m رابطه‌ی هم‌سلیقگی بین این n نفر به صورت تصادفی برقرار می‌کنیم. احتمال اینکه تمام روابط هم‌سلیقگی را به درستی تعیین کرده باشیم بر حسب n, m بیابید.

پرسش تئوری ۲۵. احتمال اینکه ۲۰ درصد از روابط هم‌سلیقگی بین این n نفر را به درستی تعیین کرده باشیم، بیابید.

[۸] 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 شبیه‌سازی ۱۷. برای $n = ۱۰۰۰$، $p = ۰/۰۰۳۴$ و $m = ۳۰۰۰$ برنامه‌ای بنویسید که اختصاص روابط هم‌سلیقگی را به تعداد $N = ۱۰$ بار تکرار کند و هر بار تعداد روابط هم‌سلیقگی را ذخیره کرده و در پایان میانگین تمام مقادیر به دست آمده را محاسبه کند. آیا این مقدار میانگین، تقریباً (با حدّاکثر خطای ۵ درصد) با مقدار m برابر است؟

پرسش تئوری ۲۶. به ازای n و p بیان‌شده در پرسش شبیه‌سازی ۱۷، این مقدار میانگین را به صورت تئوری بدست آورید. در حالت کلی چه رابطه‌ای بین n، p و m باید برقرار باشد تا این مقدار میانگین تقریباً با مقدار m برابر شود؟

احتمالاً شما هم در بین اطرافیانتان کسانی را می‌توانید پیدا کنید که سلیقه‌ی خاص داشته باشند. به این معنا که علی‌رغم علاقه به یک ژانر (مثلاً ژانر طنز) فیلم‌هایی از آن ژانر را دوست داشته باشند که بسیاری از طرفداران آن ژانر به آن‌ها علاقه‌ای ندارند و برعکس. در این پروژه این افراد را رسوا می‌نامیم![9] ابتدا تعریف فرد رسوا را دقیق می‌کنیم.

تعریف ۲ (فرد رسوا و فرد همرنگ). در یک خوشه، اگر هر فرد به طور میانگین L هم‌سلیقه داشته باشد آنگاه یک فرد را رسوا می‌نامیم اگر کم‌تر از L هم‌سلیقه داشته باشد و او را همرنگ می‌نامیم اگر بیش‌تر از L هم‌سلیقه داشته باشد.

پرسش شبیه‌سازی ۱۸. به ازای $n = ۱۰۰۰$ و $p = ۰/۰۰۰۱۶$ برنامه ای بنویسید که با ۱۰ بار تکرار، متوسّط تعداد افراد همرنگ را بیابد.
علاوه بر این، برای اینکه دید بهتری از توزیع تعداد هم‌سلیقه‌های یک فرد داشته باشید، نموداری رسم کنید که محور افقی آن تعداد هم‌سلیقه‌ها و محور عمودی آن متوسّط تعداد افرادی است که آن تعداد هم‌سلیقه دارند.

پرسش تئوری ۲۷. به ازای n و p بیان شده در پرسش شبیه سازی ۱۸، هر فرد به طور میانگین چند هم‌سلیقه دارد؟

پرسش تئوری ۲۸. برای n و p بیان شده در پرسش شبیه سازی ۱۸، اگر یک نفر را به صورت تصادفی انتخاب کنیم، احتمال اینکه همرنگ باشد چقدر است؟ همچنین امید ریاضی تعداد افراد همرنگ را بیابید.

حال می‌خواهیم گروه‌های سه تایی از افراد و روابط هم‌سلیقگی در میان این گروه‌ها در میان خوشه را بررسی کنیم.

تعریف ۳ (خاصیت تراگذری). در رابطه‌ی هم‌سلیقگی بین سه شخص A، B و C خاصیت تراگذری برقرار است به شرطی که اگر A با B هم‌سلیقه باشد و B با C هم‌سلیقه باشد آنگاه میان A و C نیز رابطه‌ی هم‌سلیقگی برقرار باشد.

تعریف ۴ (خاصیت زنجیره‌ای). در رابطه‌ی هم‌سلیقگی بین سه شخص A، B و C خاصیت زنجیره‌ای برقرار است به شرطی که A با B هم‌سلیقه باشد و B با C هم‌سلیقه باشد، امّا میان A و C رابطه‌ی هم‌سلیقگی برقرار نباشد.

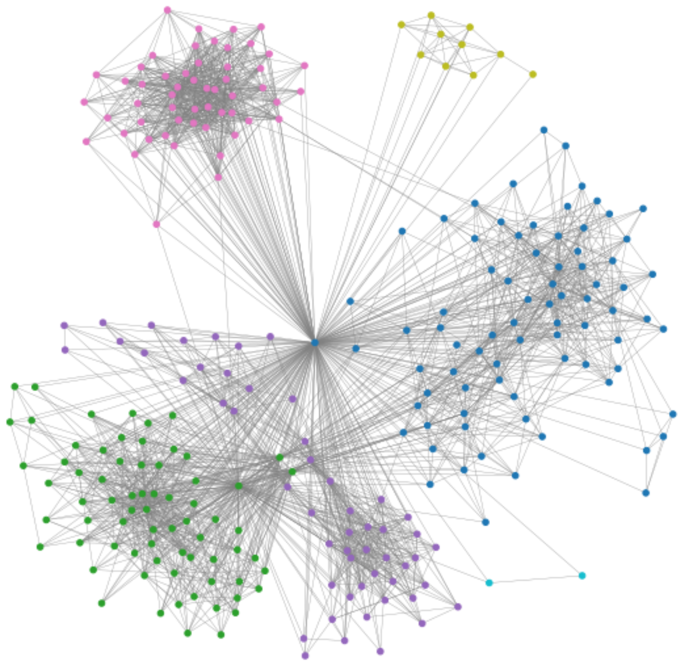
پرسش شبیه‌سازی ۱۹. برنامه ای بنویسید که بعد از ۵ بار اختصاص روابط هم‌سلیقگی به صورت تصادفی بین $n = ۳۰۰۰$ نفر با احتمال $p = ۰/۰۱$ میانگین تعداد روابط هم‌سلیقگی دارای خاصیت تراگذری و میانگین روابط هم‌سلیقگی دارای خاصیت زنجیره‌ای را حساب کند.

پرسش تئوری ۲۹. امید ریاضی تعداد روابط هم‌سلیقگی دارای خاصیت تراگذری و امید ریاضی تعداد روابط هم‌سلیقگی دارای خاصیت زنجیره‌ای را محاسبه کنید.

پرسش تئوری ۳۰. در یک خوشه، چه کسری از کلّ روابط هم‌سلیقگی بین سه نفر، با فرض آن‌که هر کدام از این سه نفر حداقل با یکی از دو نفر دیگر هم‌سلیقه باشد، دارای خاصیت تراگذری هستند؟ آیا شبیه‌سازی‌های شما با عددی که از محاسبه‌ی تئوری به دست می آورید هم‌خوانی دارند؟ نتیجه را تحلیل کنید.

[9] خواهی نشوی رسوا، همرنگ جماعت شو!

پرسش شبیه‌سازی ۲۰. برای خوشه‌ای با $n = ۱۰۰۰$ و $p = ۰/۰۰۳$ میانگین تعداد روابط هم‌سلیقگی میان هم‌سلیقه‌های یک شخص را با کمک شبیه‌سازی محاسبه کنید.



شکل ۱۰: مثالی از روابط افراد درون یک خوشه

پرسش تئوری ۳۱. امید ریاضی تعداد روابط هم‌سلیقگی میان هم‌سلیقه‌های یک شخص در خوشه‌ای با n رأس و احتمال p را بیابید. (جواب بسته لازم است!)

تا کنون خواصی از گراف متناظر با روابط هم‌سلیقگی در یک خوشه از افراد، یعنی افرادی که به یک ژانر خاص از فیلم‌ها علاقه دارند را بررسی کرده‌ایم. این گراف که n رأس دارد و هر دو رأس آن با احتمال p به هم متّصل هستند را با $\mathcal{G}(n,p)$ نشان می‌دهیم.

طبق قانون «جهان کوچک»، هر دو نفر در دنیا با احتمال نزدیک به یک با حدّاکثر ۶ واسطه هم‌دیگر را می‌شناسند. یعنی به طور مثال اگر دو فرد A و B را به صورت تصادفی انتخاب کنیم و مجموعه‌ی دوستانِ A، دوستانِ دوستانِ A، دوستانِ دوستانِ دوستانِ A و ... را بررسی کنیم، حدّاکثر بعد از ۶ مرحله به فرد B می‌رسیم.

در این بخش، می‌خواهیم وجود این ویژگی را در یک خوشه (که روابط هم‌سلیقگی در آن، برخلاف جهان واقعی، به صورت تصادفی چیده شده‌اند) تحقیق کنیم.

ابتدا به بررسی میانگین فاصله‌ی دو شخص در یک خوشه می‌پردازیم. لازم به ذکر است که فاصله‌ی دو شخص، حدّاقلّ تعداد روابط هم‌سلیقگی‌ای است که آن‌ها را به یک‌دیگر متّصل می‌کند.

پرسش شبیه‌سازی ۲۱. برنامه‌ای بنویسید که میانگین فاصله‌ی دو شخص در $\mathcal{G}(n,p)$ را به ازای $n = ۱۰۰۰$ و $p = ۰/۰۰۳۳$ محاسبه کند.

حال به بررسی حدّاکثر فاصله‌ی دو شخص در $\mathcal{G}(n,p)$ (که آن را قطر گراف می‌نامیم) می‌پردازیم.

پرسش شبیه‌سازی ۲۲. با فرض $n = ۵۰$ و $p = ۰/۳۴$، $\mathcal{G}(n,p)$ را ۱۰۰ بار به طور تصادفی تولید کنید. هر بار جفت کاربری را پیدا کنید که بیشترین فاصله را از هم در گراف دارند. میانگین بیشترین فاصله بین دو کاربر روی این ۱۰۰ گراف را به دست آورید.

پرسش شبیه‌سازی ۲۳. با ثابت (و برابر با مقدار بیان‌شده در پرسش شبیه‌سازی ۲۲) نگه داشتن p تعداد رأس‌ها (n) را در بازه‌ی $[۱۰, ۲۰۰]$ با گام ۱۰ واحد تغییر دهید و برای هر n پرسش شبیه‌سازی ۲۲ را تکرار کنید. در نهایت نمودار میانگین حدّاکثر فاصله بین جفت افراد (که میانگین روی ۱۰۰ نمونه‌ی مختلف $\mathcal{G}(n,p)$ با مشخّصات مشابه گرفته می‌شود) را به صورت تابعی از n رسم کنید. این نمودار چه فرمی دارد؟ با افزایش n رفتار این نمودار به چه صورتی است؟

پرسش تئوری ۳۲. برای دو رأس u و v از $\mathcal{G}(n,p)$ متغیّر تصادفی برنولی $I_{u,v}$ را به این صورت تعریف می‌کنیم:

$$I_{u,v} = \begin{cases} ۰ & \text{دو رأس } u,v \text{ همسایه‌ی مشترک داشته باشند} \\ ۱ & \text{دو رأس } u,v \text{ همسایه‌ی مشترک نداشته باشند} \end{cases}.$$

$\mathbb{P}[I_{u,v} = ۱]$ را محاسبه کنید.

پرسش تئوری ۳۳. برای گراف $\mathcal{G}(n,p)$ متغیّر تصادفی X_n را به صورت «تعداد جفت راس هایی از گراف که همسایه مشترکی ندارند» تعریف می‌کنیم. $\mathbb{E}[X_n]$ را بیابید.

پرسش تئوری ۳۴. با استفاده از نامساوی مارکف کران بالایی برای $\mathbb{P}[X_n \geq ۱]$ بیابید. سپس با میل دادن n به سمت بی‌نهایت رفتار این کران را بررسی کنید.

پرسش تئوری ۳۵. نتیجه بگیرید که وقتی n عدد خیلی بزرگی باشد، قطر $\mathcal{G}(n,p)$ با احتمال بالا یک کران بالا دارد و همچنین مقدار این کران بالا را نیز مشخص کنید. آیا قطر گراف برای nهای بزرگ به p وابسته است؟ آیا نتیجه‌ای که از اثبات تئوری گرفتید با نتیجه‌ی شبیه‌سازی تطابق دارد؟

پرسش شبیه‌سازی ۲۴. گراف $\mathcal{G}(n,p)$ با $n = ۱۰۰$ و $p = ۰/۳۴$ را ۱۰۰ بار به طور تصادفی تولید کنید و هر بار تعداد حلقه‌های دوستی ۳ نفره را در آن بشمارید. میانگین تعداد حلقه‌های دوستی در این ۱۰۰ گراف را به دست آورید.

پرسش شبیه‌سازی ۲۵. تعداد رأس ها (n) را در بازه‌ی $[۱۰, ۱۰۰]$ و با گام ۱۰ تغییر دهید و برای هر n، p را به صورت تابعی از n و برابر

$$p(n) = \frac{۶۰}{n^۲}$$

قرار دهید. برای هر n میانگین تعداد حلقه‌های دوستی ۳ نفره را به روش پرسش شبیه‌سازی ۲۴ حساب کنید و این میانگین را در یک نمودار بر حسب n رسم کنید.
آیا با افزایش n میانگین به عدد خاصی میل می‌کند؟ این رفتار را چگونه توجیه می‌کنید؟

پرسش شبیه‌سازی ۲۶. پرسش شبیه‌سازی ۲۵ را با $p = ۰/۳۴$ تکرار کنید. آیا میانگین تعداد حلقه‌های دوستی با افزایش n به عدد خاصی میل می‌کند؟

پرسش شبیه‌سازی ۲۷. این بار از

$$p = \frac{۱}{n}$$

استفاده کنید. n را در بازه‌ی $[۵۰, ۱۲۰۰]$ با گام ۵۰ تغییر دهید. مجدّداً نمودار میانگین تعداد حلقه‌های دوستی را بر حسب n رسم کنید. همچنین میانگین تجمّعی[۱۰] این نمودار را رسم کنید. آیا به عدد خاصی میل می‌کند؟

[10] cumulative mean

۶ سَلِ المَصانعَ رَکباً تَهیمُ فِی الفَلَوات![۱۱]

در این قسمت با الگوریتم قدم زدن تصادفی برای پیدا کردن گروه‌های با سلایق مشابه آشنا می‌شویم.
گراف n–رأسیِ G و ماتریس مجاورت $\mathbf{A} \in \mathbb{R}^{n\times n}$ آن را در نظر می‌گیریم. درجه‌ی یک رأس از گراف $\mathcal{G}$ مانند i، که آن را با $d(i)$ نشان می‌دهیم، برابر است با تعداد یال‌های متّصل به رأس i. در نتیجه بدیهی است که: $d(i) = \sum_j A_{ij}$.
در فرآیند قدم‌زدن تصادفی[۱۲] در هر مرحله به صورت تصادفی و یکنواخت (عنوان این بخش را مجدّداً بخوانید!) از یک رأس به رأس دیگری در همسایگی آن می‌رویم.

پرسش تئوری ۳۶. احتمال انتقال از رأس i به رأس j در یک مرحله چقدر است؟ این احتمال را با $P_{i,j}$ نشان می‌دهیم.

می‌توانیم $P_{i,j}$هایی که محاسبه کردید را کنار هم قرار دهیم و به ماتریس انتقال $\mathbf{P} \in \mathbb{R}^{n\times n}$ برسیم. همچنین ماتریس درجه را با یک ماتریس قطری با درایه‌های قطری $D_{ii} = d(i)$ در نظر می‌گیریم. $\mathbf{D} \in \mathbb{R}^{n\times n}$

پرسش تئوری ۳۷. ماتریس $\mathbf{P}$ را برحسب ماتریس‌های $\mathbf{A}, \mathbf{D}$ بنویسید.

پرسش تئوری ۳۸. با توجّه به رابطه‌ی ضرب ماتریس‌ها، درایه‌ی i, j ماتریس $\mathbf{P}^۲$ که آن را با $[\mathbf{P}^۲]_{i,j}$ نشان می‌دهیم را برحسب درایه‌های ماتریس $\mathbf{P}$ بنویسید. آیا می‌توانید یک تعبیر احتمالی از درایه‌های ماتریس $\mathbf{P}^۲$ ارائه دهید؟

پرسش تئوری ۳۹. عبارتی برای احتمال رسیدن از رأس i به رأس j در t مرحله که آن را با $P^{(t)}_{i,j}$ نشان می‌دهیم، به دست آورید.

پرسش تئوری ۴۰. $P^{(t)}_{i,j}$ و $P^{(t)}_{j,i}$ چه ارتباطی با یکدیگر دارند؟ ثابت کنید نسبت این دو احتمال تنها به درجه‌ی رئوس i و j بستگی دارد.

ایده‌ی اصلی روش قدم‌زدن تصادفی آنست که اگر یک قدم‌زن تصادفی روی گراف شبکه شروع به حرکت کند، بعد از مدّتی با احتمال بالا انتقال فقط بین زیرمجموعه‌ای از گره‌ها صورت می‌گیرد که همان خوشه‌های هم‌سلیقگی موردنظر ماست. به تعبیری قدم‌زن تصادفی در یک خوشه به دام می‌افتد، زیرا تعداد یال‌های بین اعضای یک خوشه نسبت به یال‌های بین اعضای یک خوشه و رئوس دیگر گراف بسیار بیشتر است.

پرسش تئوری ۴۱. به طور شهودی توضیح دهید که اگر کاربر i و کاربر j هم‌سلیقه باشند، $P^{(t)}_{ij}$ چه ویژگی‌هایی باید داشته باشد؟ در این صورت در خصوص احتمالات $P^{(t)}_{ik}$ و $P^{(t)}_{jk}$ که کاربر k یک فرد دلخواه در شبکه است، چه می‌توان گفت؟

پرسش تئوری ۴۲. برای اینکه بتوانیم کاربران هم سلیقه را دسته بندی کنیم، نیاز به یک معیار هم سلیقگی داریم. «اختلاف سلیقه‌ی دو کاربر i و j» را بصورت زیر تعریف می کنیم:

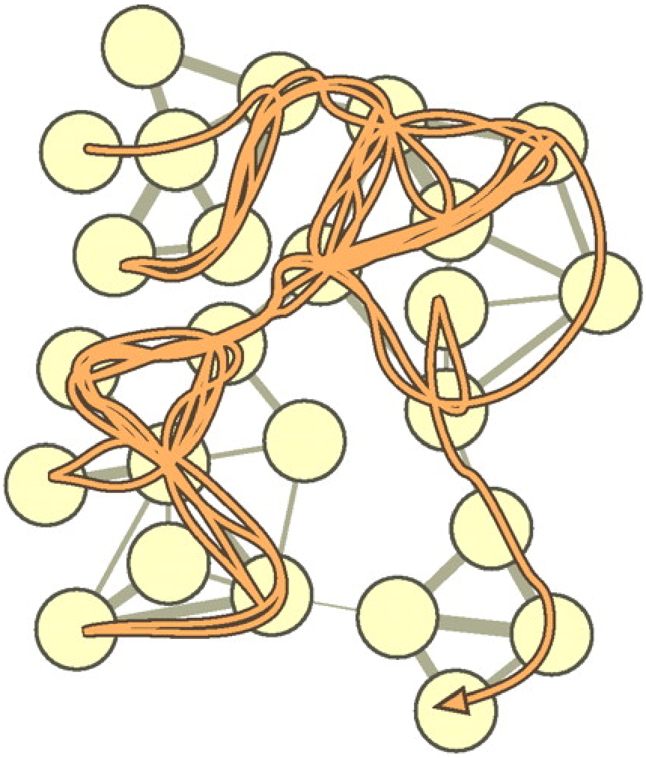
$$r_{ij} = \sqrt{\sum_{k=۱}^{n} \frac{(P^{(t)}_{ik} - P^{(t)}_{jk})^۲}{d(k)}}$$

اکنون یک دسته از کاربران هم سلیقه را درنظر گرفته و C می نامیم. احتمال آن که یک قدم‌زن تصادفی از یکی از رئوس مرتبط با کاربران این دسته شروع کند و بعد از t مرحله به کاربر دلخواه k برسد را محاسبه کنید. سپس مشابه رابطه‌ی فوق، «اختلاف سلیقه‌ی بین دو دسته‌ی $C_۱$ و $C_۲$» را تعریف کنید.

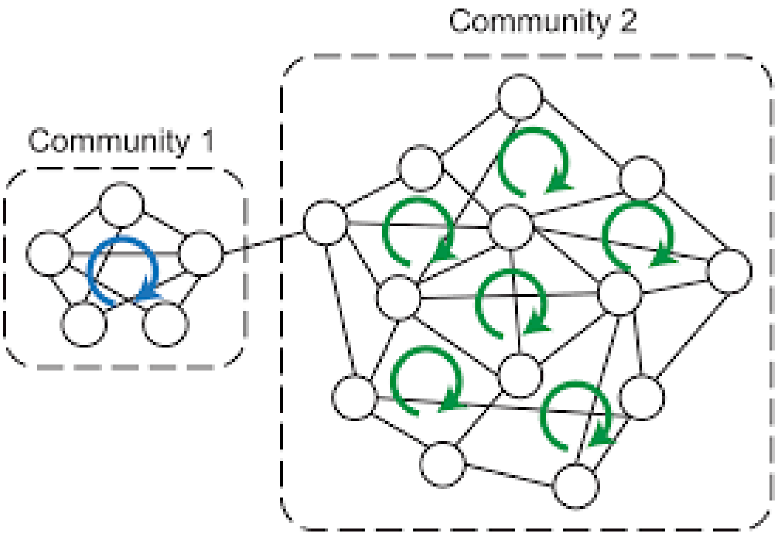
---

[۱۱] سَلِ المَصانعَ رَکباً تَهیمُ فی الفَلَوات/ تو قدرِ آب چه دانی که در کنارِ فراتی؟ [سعدی]
معنای مصرع اول: از سواری که در بیابان سرگردان است درباره‌ی [آب] چشمه‌ها بپرس

[12] Random Walk



شکل ۱۱: مثالی از قدم‌زدن تصادفی



شکل ۱۲: به دام افتادن قدم‌زن تصادفی درخوشه‌ها

حالا می‌توانیم الگوریتمی برای دسته بندی کاربران هم‌سلیقه ارائه کنیم:

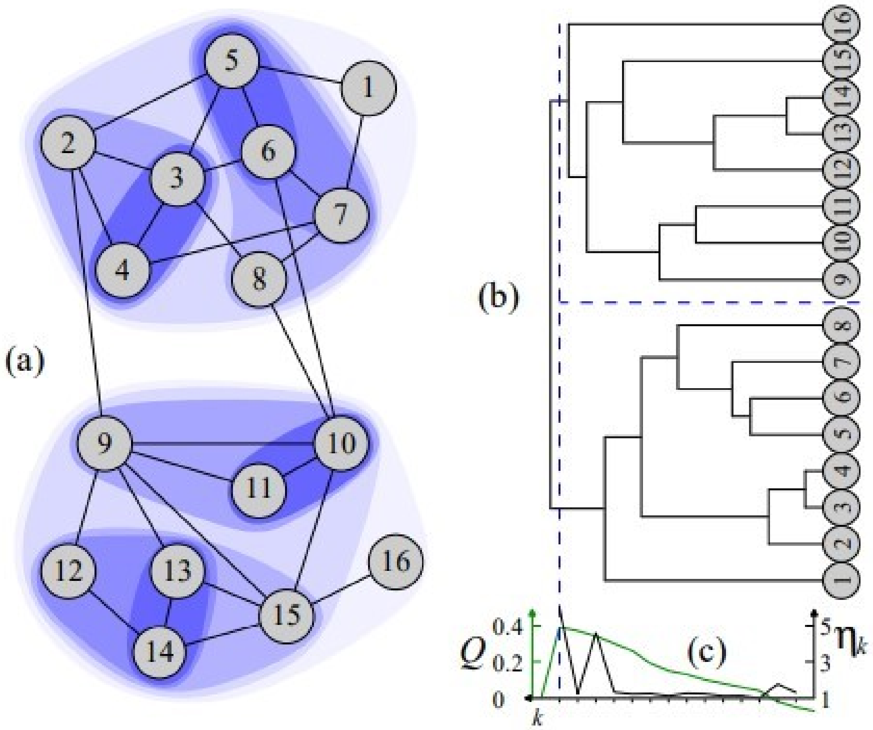
۱. هر کاربر را یک دسته درنظر گرفته و اختلاف سلیقه‌ی هر دو کاربر را محاسبه کنید.

۲. تا هنگامی که تعداد دسته‌ها بزرگ‌تر از ۱ است:

(آ) دو دسته با کمترین اختلاف سلیقه را انتخاب کنید.

(ب) آنها را با هم ترکیب کنید ($C_3 = C_1 \cup C_2$).

(ج) اختلاف سلیقه‌ی بین دسته‌ها را به روز کنید.



شکل ۱۳: الگوریتم خوشه‌بندی افراد بر اساس قدم‌زدن تصادفی

پرسش تئوری ۴۳. معیاری ارائه دهید که نشان دهد در کدام مرحله از الگوریتم، دسته‌بندی به بهترین شکل انجام شده است.

پرسش شبیه‌سازی ۲۸. الگوریتم توصیف‌شده را روی مجموعه‌داده‌ی Zachary’s Karate Club پیاده‌سازی کنید. تعداد گام‌ها در تعریف فاصله‌ها را $t = ۲$ در نظر بگیرید.

پرسش شبیه‌سازی ۲۹. تعداد گام‌ها در تعریف فاصله‌ها را $t = ۵$ در نظر بگیرید و پرسش شبیه‌سازی ۲۸ را تکرار کنید.

۷ نکات مهم!

لطفاً به نکات زیر دقّت کنید:

۱. این پروژه بخشی از نمره‌ی شما در این درس را تشکیل خواهد داد.

۲. می‌توانید پروژه را در قالب گروه‌های ۲ یا ۳ نفره انجام دهید. فرمی برای ثبت گروه‌ها در اختیار شما قرار خواهد گرفت. دقّت داشته باشید که در هنگام تحویل پروژه باید تمامی اعضای گروه به تمامی بخش‌ها مسلّط باشند و در نهایت همه‌ی اعضای یک گروه نمره‌ی واحدی را دریافت خواهند کرد.

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

۴. 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.

۵. در این پروژه از دو مجموعه‌داده استفاده خواهیم کرد. توضیحات و نحوه‌ی دریافت این دو مجموعه داده در فایل Dataset.txt آمده است.

۶. در فایل kmeans.pdf، توضیحاتی در مورد الگوریتم k–means برای مطالعه‌ی اختیاری شما آمده است. همچنین می‌توانید فایل linear_algebra_prerequisites.pdf را برای آشنایی بیشتر با جبرخطّی بخوانید. البته در این پروژه نیازی به جبرخطّی پیشرفته نخواهید داشت و تنها در حدّ ضرب ماتریس‌ها و محاسبه‌ی مقادیرو بردارهای ویژه‌ی یک ماتریس کافی خواهد بود.

۷. 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.

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

۹. نوشتن گزارش کار با $\LaTeX$ نمره‌ی امتیازی دارد.

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

۱۱. 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!

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

موفّق باشید!