

لطفاً به نکات زیر دقت کنید:

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 ذکر شده 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 مجموعه داده استفاده خواهیم کرد. توضیحات و نحوه‌ی دریافت این دو مجموعه داده در فایل `Dataset.txt` آمده است.

۷. تحویل پروژه به صورت گزارش و کدهای نوشته شده است. گزارش باید شامل پاسخ پرسش‌ها، تصاویر و نمودارها و نتیجه گیری‌های لازم باشد. توجه کنید که قسمت عمده بارم شبیه سازی را گزارش شما و نتیجه ای که از خروجی کد میگیرید دارد. همچنین تمیزی گزارش بسیار مهم است. کدها و گزارش را در یک فایل فشرده شده 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!

موفق باشید!

## ۱ مقدمه

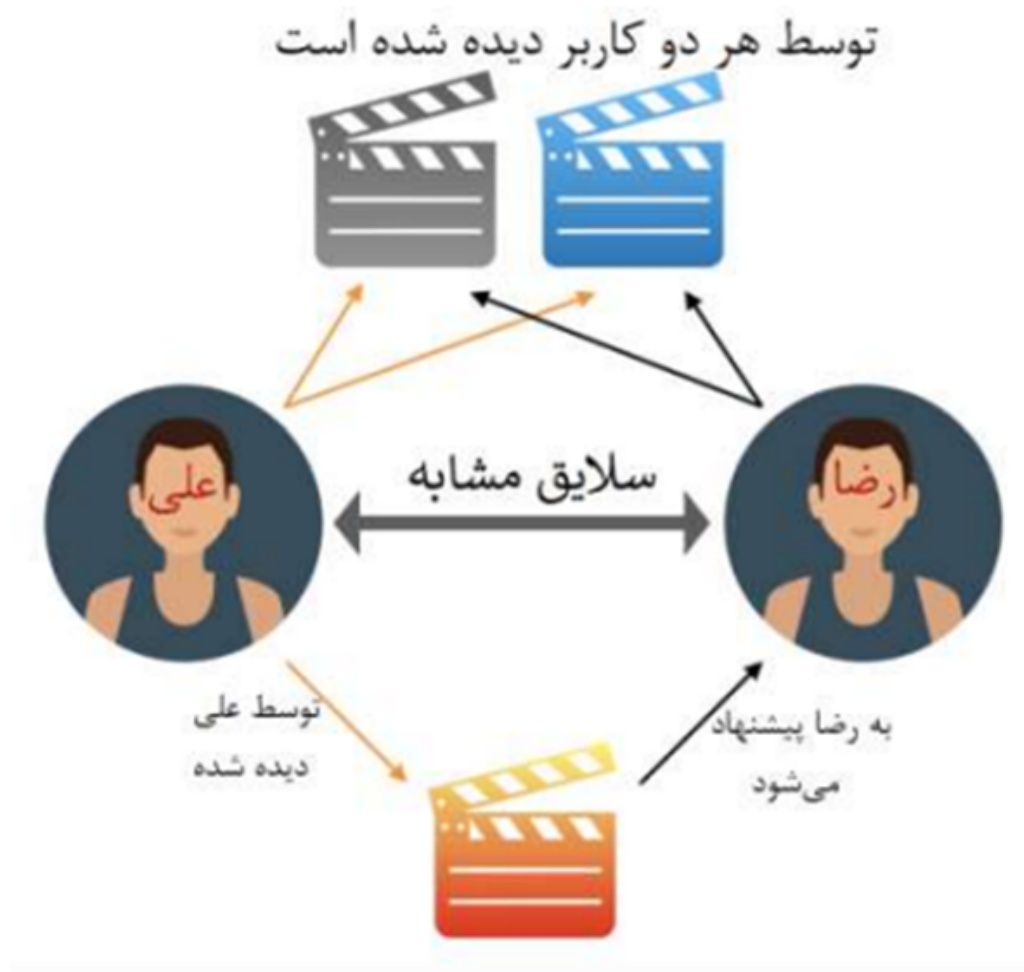
امتحانات تمام شده است و شما با خیالی آسوده می‌خواهید اوقات فراغت خود را بگذرانید تا با انرژی فراوان به سراغ ترم بعدی بروید. یک گزینه‌ی مناسب برای این کار، دیدن فیلم است. شما از طریق نرم‌افزار یا سایت وارد فیلمو/نماوا/فیلم‌نت می‌شوید و در صفحه‌ی اول تعدادی پیشنهاد مشاهده می‌کنید که چندان هم بد نیستند 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 این جاست 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<sup>۱</sup> آشنایی دارید. یکی از ارکان مهم 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 سلیقه 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 سلیقه مشابه را پیدا کرده و به هر گروه فیلم متناسب با سلیقه آن‌ها را پیشنهاد دهیم. مثال ساده‌ای از این فرآیند در شکل ۲ نشان داده شده است.

<sup>۱</sup>Video On Demand (VOD)



شکل ۲: تأثیر فیلم‌های دیده‌شده توسط کاربران بر پیشنهادات سیستم

## ۲ ز هرچه رنگ تعلق پذيرد آزاد است!

ابتدا بايد يك مدل رياضي براي توصيف روابط بين افراد بياييم. همان طور كه احتمالاً تا الان حدس زده ايد، بهترين مدل رياضي‌اي كه مي‌تواند ارتباط بين افراد مختلف در اين مسئله را توصيف كند، گراف است. مي‌توانيم هر فرد را با يك رأس گراف مدل كنيم و بين هر دو فرد هم‌سليقه يك يال رسم كنيم. براي هر گراف، مي‌توانيم يك ماتريس مجاورت تعريف كنيم.

تعريف ۱. ماتريس مجاورت براي گراف  $G$  با  $n$  رأس يك ماتريس  $A \in \mathbb{R}^{n \times n}$  است كه درايه‌هاي آن به صورت زير هستند:

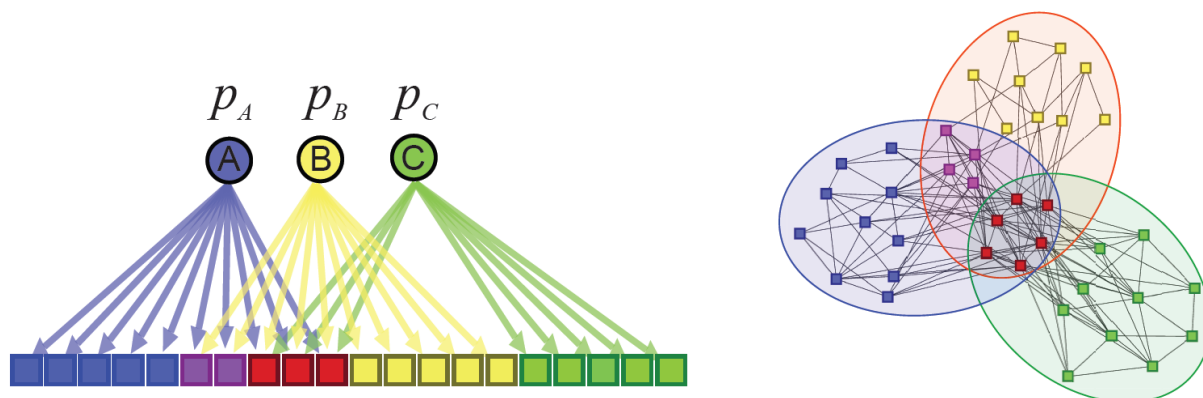
$$A_{i,j} = \begin{cases} 1 & \text{اگر بين رؤس } i, j \text{ يال وجود داشته باشد} \\ 0 & \text{در غير اين صورت} \end{cases}$$

هدف ما در اين پروژه پيدا كردن گروه‌هايي از افراد است كه سليقه‌ي مشترك دارند.

پرسش تئوري ۱. اگر جامعه را بتوان به خوشه‌هاي مجزا از افرادي با سليقه‌ي يکسان افراز كرد، به طوري كه افراد يك خوشه با افراد خوشه‌ي ديگر هيچ اشتراكي نداشته باشند، آنگاه ماتريس مجاورت اين افراد به چه صورت خواهد بود؟

پرسش تئوري ۲. مدلي كه در پرسش قبلي به دست آورديد در واقعيت رخ نمي‌دهد، چرا كه ممكن است يك فرد در يك خوشه با فردی در خوشه ديگر سلايق مشتركی (هرچند كم) داشته باشند. همچنين يك فرد مي‌تواند به طور همزمان در چند خوشه عضو باشد. براي مثال يك فرد مي‌تواند در خوشه‌ي افرادي كه عموماً به ژانر كمدي علاقه دارند باشد و همزمان در خوشه‌ي افرادي كه ژانر درام ميپسندد نيز حضور داشته باشد. در اين صورت چه توصيفي از ماتريس مجاورت مي‌توان داشت؟

پرسش تئوري ۳. فرض كنيد  $M$  ژانر مختلف فيلم  $1, 2, \dots, M$  داريم. اگر دو نفر به ژانر  $i$ -ام علاقه داشته باشند، به احتمال  $p_i$  با يكديگر هم سلیقه‌اند. حال فرض كنيد دو شخص خاص هر کدام به چند ژانر مختلف علاقه دارند. اگر فرض كنيم علاقه به ژانرهای مختلف از هم مستقلند، احتمال اينكه اين دو نفر با يكديگر هم سلیقه باشند چقدر است؟ چرا اگر دو نفر در جوامع خاص زيادي عضو باشند احتمال اينكه با هم در ارتباط باشند بيشتر مي‌شود؟ آيا مي‌توانيد اين مسئله را به صورت شهودي هم توجيه كنيد؟ به نظر شما اين مدل چه نقصي در مدل كردن رابطه‌ي افراد و گروه‌ها دارد؟ شكل ۳ را در اين خصوص ببينيد.



شكل ۳: مثالی از علاقه‌ي افراد مختلف به ژانرهای مختلف

آغلام همت آنم كه زير چرخ كبود/ ز هرچه رنگ تعلق پذيرد آزاد است [حافظ]

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

۱. پارامتری برای کمی کردن میزان تمایل افراد به عضویت در یک گروه در نظر می‌گیریم. این پارامتر را «تعلق» می‌نامیم. تعلق یک فرد مانند  $u$  به خوشه‌ی  $c$  عبارت است از میزان تمایل او به عضویت در خوشه‌ی  $c$ . این پارامتر را با  $F_{uc} \in [0, \infty)$  نشان می‌دهیم.

۲. هر چه دو فرد تعلق 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  $P_{uv}(c) = 1 - \exp(-F_{uc}F_{vc})$  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 تعلق افراد مختلف به خوشه‌های مختلف می‌توانیم ماتریس تعلق  $\mathbf{F} \in \mathbb{R}^{n \times C}$  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 تعلق  $\mathbf{F}$  محاسبه کنید. نشان دهید اگر دو فرد در گروه‌های بیشتری اشتراک داشته باشند این احتمال افزایش می‌یابد.

دقت 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 کلی کار دشواری است. بنابراین باید به صورت عددی آن را بیشینه کنیم. برای این کار، در هر مرحله تنها تعلق‌های یک شخص خاص از بین  $n$  نفر به گروه‌های مختلف را کمی در راستای گرادیان جابجا می‌کنیم. توجه می‌کنیم که تعلق‌های شخص  $u$  در سطر  $u$ -ام از ماتریس  $\mathbf{F}$  که آن را با  $F_{u,:}$  نشان می‌دهیم قرار دارد و در نتیجه گرادیان مذکور عبارت است از:

$$\nabla_{F_{u,:}} l(\mathbf{F}) = \left[ \frac{\partial l(\mathbf{F})}{\partial F_{u,1}}, \frac{\partial l(\mathbf{F})}{\partial F_{u,2}}, \dots, \frac{\partial l(\mathbf{F})}{\partial F_{u,C}} \right].$$

**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 سطرهاى 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 ۰.۱ در نمونه کد زیر الگوریتم تکراری فوق برای تخمین  $F$ ، در تابع `train` پیاده سازی شده است که ماتریس مجاورت و تعداد گروه‌ها را به عنوان ورودی دریافت کرده و تخمینی از ماتریس تعلق  $F$  مانند  $\hat{F}$  را به عنوان خروجی برمی‌گرداند. بخش‌های مربوط به تابع  $l(F)$  و تابع گرادیان را تکمیل کنید.

```

1 def log_likelihood(F, A):
2     #todo
3     return log_likelihood
4
5 def gradient(F, A, i):
6     #todo gradient of log_likelihood respect to person i parameters (F_ic)
7     return gradient
8
9 def train(A, C, iterations = 200):
10    # initialize an F
11    N = A.shape[0]
12    F = np.random.rand(N,C)
13
14    for n in range(iterations):
15        for person in range(N):
16            grad = gradient(F, A, person)
17            F[person] += 0.005*grad # updating F
18            F[person] = np.maximum(0.001, F[person]) # F should be nonnegative
19            ll = log_likelihood(F, A)
20            print('At step %4i logliklihood is %5.4f'%(n,ll))
21
22    return F

```

تعداد تکرارها را می‌توان این گونه تعیین کرد که اختلاف  $l(\hat{F})$  در دو تکرار متوالی کمتر از یک مقدار آستانه مانند  $\epsilon = 0.001$  شود. به این ترتیب ما می‌توانیم ماتریس تعلق  $F$  را تخمین بزنیم.

حال می‌خواهیم مشخص کنیم هر فرد در کدام خوشه‌ها قرار می‌گیرد. برای این کار یک مقدار آستانه مانند  $\delta$  تعیین می‌کنیم و فرد  $u$  را عضو خوشه‌ی  $c$  می‌گیریم اگر  $F_{uc} > \delta$  باشد. یک روش برای تعیین مقدار  $\delta$  آن است که آن را طوری تعیین کنیم که احتمال وجود ارتباط بین دو فرد هم‌سلیقه (که حداقل برابر با  $1 - e^{-\delta^2}$  است)، از احتمال ارتباط تصادفی آن‌ها،  $\epsilon$ ، (یعنی ارتباطی که ناشی از هم‌سلیقه‌بودن نباشد) بیشتر باشد. هم‌چنین احتمال ارتباط تصادفی بین دو فرد را می‌توان 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 مجدداً  $\nabla_{F_{u,:}} l(F)$  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  $A$  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).

```

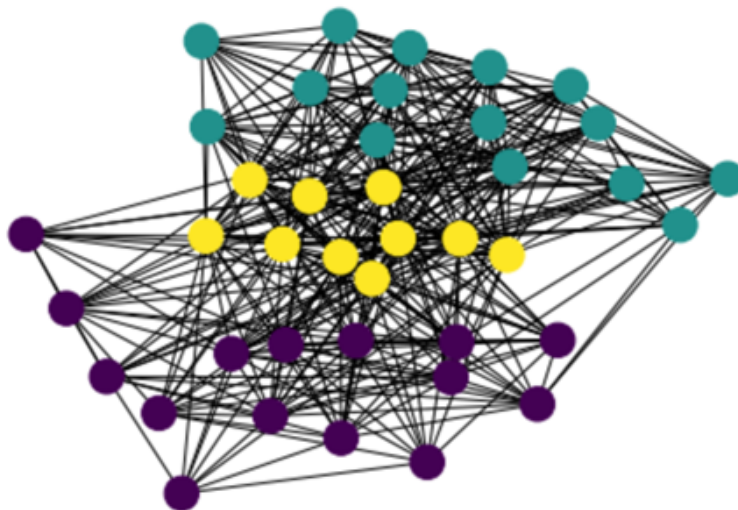
1 #testing in two small groups
2 A=np.random.rand(40,40)
3 A[0:15,0:25]=A[0:15,0:25]>1-0.6 # connection prob people with 1 common group
4 A[0:15,25:40]=A[0:15,25:40]>1-0.1 # connection prob people with no common group
5 A[15:40,25:40]=A[15:40,25:40]>1-0.7 # connection prob people with 1 common group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

```

```

11
12 import matplotlib.pyplot as plt
13 import networkx as nx
14 plt.imshow(A)
15 delta=np.sqrt(-np.log(1-0.1)) # epsilon=0.1
16 F=train(A, 2, iterations = 120)
17 print(F>delta)
18 G=nx.from_numpy_matrix(A)
19 C=F>delta # groups members
20 nx.draw(G,node_color=10*(C[:,0])+20*(C[:,1]))

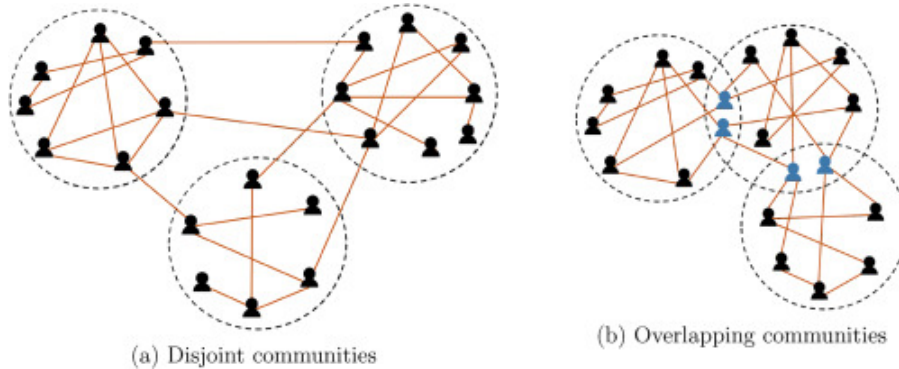
```



شکل ۴: خروجی الگوریتم تشخیص خوشه‌ها

### ۳ گناه بخت پریشان و دست کوتاه ماست!<sup>۳</sup>

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  $i \in \{1, 2, \dots, n\}$  داریم  $z_i \in \{1, 2, \dots, k\}$ .
  ۴. فرض کنید بردار  $\mathbf{z}_0$  به عنوان نحوه‌ی صحیح خوشه‌بندی در اختیار ماست:
- $$\mathbf{z}_0 = [3, 1, 2, 1, 3, 1, 2, 2, 2, 3, 3, 2, 1, 1, 3]^T$$
- بردار  $\mathbf{z}$  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.

<sup>۳</sup> اگر به زلف دراز تو دست ما نرسد/ گناه بخت پریشان و دست کوتاه ماست [حافظ]



پرسش تئوری ۰۹. با توجه به توضیحات بالا، درایه‌ی  $A_{i,j}$  از ماتریس مجاورت (می‌توانید مجدداً به تعریف ۱ مراجعه کنید!) گراف روابط دوستی افراد یک متغیر تصادفی است. این متغیر تصادفی را توصیف کنید و تابع چگالی/جرم احتمال آن را بیابید.

پرسش تئوری ۰۱۰. 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  $A$  را با یافته‌های جدید توصیف کنید.

پرسش شبیه‌سازی ۰۲. از ماتریسی که در پرسش تئوری ۰۱۰ توصیف کردید، ۱۰ نمونه بسازید.

پرسش شبیه‌سازی ۰۳. از ماتریسی که در پرسش تئوری ۰۱۰ 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 مشخص کنید. دقت 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 مشخص 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  $\mathbf{z} \in \mathbb{R}^k$  است که فرض می‌کنیم آن را در اختیار نداریم و می‌خواهیم از روی درایه‌های گراف  $A$  به تخمینی از  $\mathbf{z}$ ، مانند بردار  $\hat{\mathbf{z}}$  برسیم.

ابتدا لازم است معیاری برای سنجش میزان دقت تخمین معرفی کنیم. از فاصله‌ی همینگ<sup>۴</sup> استفاده می‌کنیم. این فاصله به صورت زیر تعریف می‌شود:

$$d_H: \mathbb{R}^k \times \mathbb{R}^k \rightarrow \{1, 2, \dots, k\}$$

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

که در تعریف بالا، تابع  $\mathbb{1}\{\cdot\}$  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 ۰۴. 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 نقطه‌ی ضعف، 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، ۲، ۳، ماهیت خوشه بندی تفاوتی نکرده‌است، تنها شماره‌ای که اختصاص داده‌ایم تفاوت کرده است. در واقع معیار سنجش میزان دقت تخمین باید نسبت به جایگشت‌های مختلف نامگذاری مستقل باشد و با تغییر آن عوض نشود.

<sup>4</sup>Hamming distance

حال باید کاری کنیم که فاصله‌ی همینگ را از نام خوشه‌ها مستقل کند. برای مثال، در یک جامعه با  $n = 6$  عضو و  $k = 2$  گروه، فرض کنید خوشه بندی درست به صورت زیر باشد:

$$\mathbf{z}_0 = [2, 2, 1, 2, 1, 1]^T.$$

و ما به تخمین زیر از  $\mathbf{z}_0$  برسیم:

$$\hat{\mathbf{z}}_0 = [1, 1, 2, 1, 2, 2]^T.$$

اگر از تابعی که در پرسش شبیه‌سازی ۴ نوشتید برای محاسبه‌ی فاصله‌ی این دو بردار استفاده کنیم، خواهیم داشت  $d_H(\hat{\mathbf{z}}_0, \mathbf{z}_0) = 6$  ولی اگر جای برچسب ۰ و ۱ را در بردار  $\hat{\mathbf{z}}_0$  تغییر دهیم و بردار حاصل را  $\tilde{\mathbf{z}}_0$  بنامیم، داریم  $d_H(\tilde{\mathbf{z}}_0, \mathbf{z}_0) = 0$ . 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  $n = 6$  عضوی با  $k = 3$  خوشه‌ی مختلف، بردار زیر را در نظر بگیرید:

$$\mathbf{z} = [3, 1, 2, 2, 3, 1]^T.$$

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

$$\langle \mathbf{z} \rangle = \{[3, 1, 2, 2, 3, 1]^T, [3, 2, 1, 1, 3, 2]^T, [1, 3, 2, 2, 1, 3]^T, \dots\}.$$

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

$$d: \mathbb{R}^k \times \mathbb{R}^k \rightarrow \{1, 2, \dots, k\}$$

$$d(\mathbf{z}_1, \mathbf{z}_2) = \min_{\tilde{\mathbf{z}}_2 \in \langle \mathbf{z}_2 \rangle} d_H(\mathbf{z}_1, \tilde{\mathbf{z}}_2).$$

**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 تحقق ماتریس مجاورت، بردار  $\mathbf{z}_0$  را تخمین بزنیم.

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

**پرسش تئوری ۱۲.** احتمال تحقق 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  $l(\mathbf{z}) = \log(L(\mathbf{z}))$  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  $\mathbf{z}_0 \in \mathbb{R}^n$  عددی از مجموعه‌ی  $\{1, 2, 3\}$  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 اولیه از  $\mathbf{z}_0$  به صورت زیر تعریف کنید:

$$\hat{\mathbf{z}}_0 = [1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3]$$

۲. مقدار  $\tilde{l}(\hat{\mathbf{z}}_0)$  را محاسبه کنید.

۳. برای  $t = 1, \dots, T$ :

(آ) برای  $i = 1, \dots, n$ :

جای  $[\hat{\mathbf{z}}_{t-1}]_i$  را با تک تک درایه‌های  $\hat{\mathbf{z}}_{t-1}$  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  $\tilde{l}(\hat{\mathbf{z}}_{t-1})$  را نتیجه می‌دهد را روی بردار  $\hat{\mathbf{z}}_{t-1}$  اعمال کنید و به بردار  $\hat{\mathbf{z}}_t$  برسید.

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

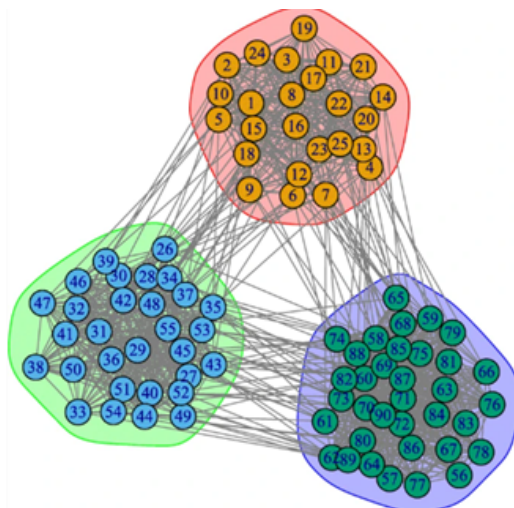
پرسش شبیه‌سازی ۷. الگوریتم بالا را شبیه‌سازی کنید. در هر مرحله میزان  $\tilde{l}(\hat{\mathbf{z}}_t)$  و همچنین  $d(\hat{\mathbf{z}}_t, \mathbf{z}_0)$  را ذخیره کنید و در نهایت روی نمودار نمایش دهید. در انتخاب پارامتر  $T$  مختارید ولی آن را عددی مناسب انتخاب کنید.

پرسش شبیه‌سازی ۸. الگوریتم را به ازای  $N = 10$  نقطه‌ای شروع ( $\hat{\mathbf{z}}_1$ ) 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،  $N = 10$  خروجی مختلف به دست می‌آورد. این خروجی‌ها را با  $\{\hat{\mathbf{z}}_T^{(j)}\}_{j=1}^N$  نشان می‌دهیم. مقدار  $\tilde{l}(\hat{\mathbf{z}}_T^{(j)})$  به ازای زهای مختلف را با  $\tilde{l}(\mathbf{z}_0)$  مقایسه کنید. آیا حالتی وجود دارد که  $\tilde{l}(\hat{\mathbf{z}}_T^{(j)}) = \tilde{l}(\mathbf{z}_0)$  شده باشد؟ در این حالت مقدار فاصله‌ی همینگ کمینه‌ی بردار تخمین و  $\mathbf{z}_0$  چقدر است؟

پرسش شبیه‌سازی ۱۰. آیا زای وجود دارد که  $d(\hat{\mathbf{z}}_T^{(j)}, \mathbf{z}_0) = 0$  شده باشد؟

پرسش شبیه‌سازی ۱۱. از روی  $\mathbf{z}_0$ ، دو نمونه‌ی دیگر از ماتریس  $\mathbf{A}$  بسازید و در هرکدام به ازای  $N = 10$  بردار اولیه‌ی مختلف، تلاش کنید که  $\mathbf{z}_0$  را تخمین بزنید. بهترین نتیجه را گزارش کنید.



شکل ۶: مثالی از خوشه‌بندی افراد