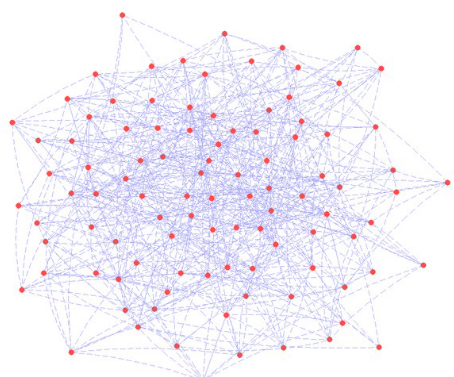
باسمه تعالی

پروژه درس آمار و احتمال مهندسی

آشنایی با برخی روش های خوشه بندی گراف ها

فاز اول



استاد درس:

دکتر محمدحسین یاسائی میبدی

اعضای گروه:

هلیا شاکری (400101389)

ملیکا کردمیل (400101786)

زهرا ملکی (400110009)

۲ ز هرچه رنگ تعلّق پذیرد آزاد است!

پرسش تئوری 1.1

در این حالت به دلیل این که افراد خوشه هاي مختلف با یکدیگر سلیقه مشترکي دارند اگر هر فرد را با یك یال مشخص کنیم و مي توانیم افراد را به دسته هايي تقسیم کنیم که این دسته ها هیچ یالي بین هیچ دو دسته متفاوت آن ها وجود نداشته باشند در این حالت گراف ما باید یك گراف ناهمبند باشد و ماتریس مجاورت آن هم به همین صورت است. همچنین قابل توجه است که درایه هاي قطر ماتریس مجاورت حتما 1 هستند زیرا هرکس با خودش هم سلیقه است. به طور مثال ماتریس زیر با این شرایط سازگار است:

1	1	0
1	1	0
0	0	1

به طور کلي براي حداقل وجود دو خوشه به این صورت مي توان ماتریس مجاورت را نوشت اگر خوشه 1 داراي r گره و خوشه 2 داراي s گره باشند:

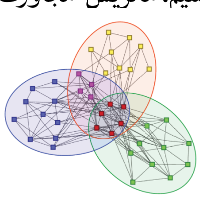
$0_{r\times r}$	B
B^T	$0_{s\times s}$

پرسش تئوری 1.2

تفاوت این بخش با بخش قبل در این است که در این حالت بین دو خوشه هم یال هاي مشترکي وجود دارند. در این صورت میتوان مجموعه هايي از گره ها داشته باشیم به این صورت که براي خوشه A و B به طور مثال مجموعه ها به صورت

$\{B\cap A\}$ $\{A - B\cap A\}$ $\{B - B\cap A\}$ هستند و معادل 3 دسته میشوند و حال براي هر دو خوشه اي این روند را تکرار

میکنیم و گره هاي هر دسته را رنگي جدا میکنیم و آنقدر این کار را ادامه میدهیم تا دیگر قادر به تعیین رنگ جدیدي نباشیم. ماتریس مجاورت این گراف به طور مثال به شکل زیر میشود:



1	1	0	0	0	0
1	1	1	0	0	0
0	1	1	0	0	0
0	0	0	1	1	0
0	0	0	1	1	1
0	0	0	0	1	1

پرسش تئوری 1.3

ین دو فرد در ژانر های ۱ تا k مشترک باشند در آن صورت احتمال اینکه با یکدیگر هم سلیقه باشند برابر است با ۱ منهای احتمال اینکه با یکدیگر هم سلیقه نباشند. احتمال اینکه این دو فرد در با هم هم سلیقه نشوند، با توجه به مستقل بودن این احتمالات، برابر است با احتمال اینکه در هیچ یك از این خوشه ها با هم هم سلیقه نشوند در نتیجه:

$$1 - \prod_{1}^{n}(1 - p_i)$$

اگر دو فرد در خوشه هاي زيادي عضو باشند به این معنا که سلیقه هاي متنوعي داشته باشند در نتیجه خوشه هاي مشترك آنان افزایش مي یابد در نتیجه n افزایش یافته پس $\prod_{1}^{n}(1 - p_i)$ کاهش بافته و $1 - \prod_{1}^{n}(1 - p_i)$ افزایش مي یابد که در نتیجه احتمال هم سلیقه بودنشان افزایش مي یابد که به صورت شهودي هم کاملا ملموس است زماني که دو نفر سلیقه هاي متنوعي دارند ربیشتر سلیقه مشترك پیدا میکنند.

نقض این مدل میتواند در این باشد که احتمال pi ها نباید از هم کاملا مستقل باشند و بین علاقه به ژانر هاي مختلف هم رابطه اي وجود دارد.

پرسش تئوری 1.4

مشابه بخش قبل اگر متمم احتمال هم سلیقگی آن ها را حساب کنیم میبینیم که اگر این دو فرد در خوشه های ۱ تا n مشترک باشند در آن صورت احتمال هم سلیقگی آن ها برابر است با:

$$1 - \prod_{1}^{n} exp(-F_{ui}F_{vi})$$

با توجه به فرمول بالا چون Fvi ها اعدادی مثبت هستند پس توان های e عددی بین ۰ تا ۱ است پس با زیاد تر شدن n حاصل ضرب کاهش و در نتیجه احتمال هم سلیقگی افزایش پیدا میکند.

پرسش تئوری 1.5

اگر محل برخورد سطح u و ستون v در ماتریس A برابر a_{uv} باشد.

$$P[A|F] = \prod_{uv} P[a_{uv}|F]$$

که a_{uv} به احتمال $1 - \prod_{1}^{n} exp(-F_{ui}F_{vi})$ یک و به احتمال $\prod_{1}^{n} exp(-F_{ui}F_{vi})$ برابر صفر است

$$P[a_{uv}|F] = a_{uv} + (-1)^{a_{uv}} \prod_{1}^{n} exp(-F_{ui}F_{vi})$$

$$log(P[A|F]) = \sum_{u,v} log(a_{uv} + (-1)^{a_{uv}} \prod_{1}^{n} exp(-F_{ui}F_{vi}))$$

پرسش تئوری 1.6

این ماتریس به صورت يك تابع از u,v است و ما میخواهیم آن را به صورت بیشینه و تنها تابع u بدست آوریم براي این کار مقادیر ما باید به گونه اي باشد که در راستاي ببشینه کردن این تابع باشند. راستاي گرادیان F در جهت نقطه ماکسیموم است و در نقطه ماکسیموم این مقدار 0 است پس با تکرار این کار تا به جایي برسیم که مقدار گرادیان 0 یا بسیار کم باشد.

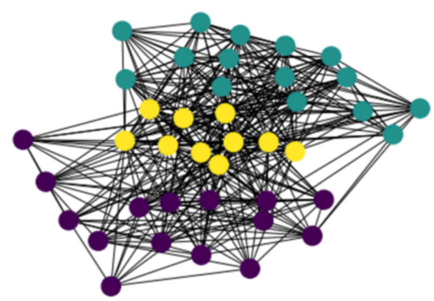
پرسش تئوری 1.7

از رابطه قسمت قبل گرادیان مي گیریم و خواهیم داشت:

$$\nabla_{F_{u,:}} l(F) = \sum_{j=v=1,\, u\neq v}^{n} \frac{\left(\sum_{i=1}^{k}(-F_{vj})\right)(-1)^{a_{vu}} e^{\sum_{i-1}^{k}(-F_{ui}F_{vi})}}{a_{vu}+(-1)^{a_{vu}} e^{\sum_{i-1}^{k}(-F_{ui}F_{vi})}} \hat{v}$$

پرسش شبیه سازي 1

توابع تکمیل شده اند.



پرسش تئوری 1.8

در این حالت اگر احتمال هم سلیقه بودن به شرط اینکه 1 ژانر مشترك داشته باشند را $1 - e^{-F_{uk}F_{vk}}$ بگیریم زماني که r فیلم مشترك دارند به شرطي که s فیلم در آن ژانر وجود داشته باشد با توزیع دو جمله اي آن را مدل مي کنیم. احتمال هم سلیقگی تحت خوشه k اگر به r فیلم علاقه داشته باشند:

$$\binom{s_k}{r}\left(1 - e^{-F_{uk}F_{vk}}\right)^{r}\left(e^{-F_{uk}F_{vk}}\right)^{s_k - r}$$

اگر n ژانر مشترک باشد:

$$P[a_{uv}|F] = 1 - \prod_{k=1}^{n}\left(1 - \binom{s_k}{r}\left(1 - e^{-F_{uk}F_{vk}}\right)^{r}\left(e^{-F_{uk}F_{vk}}\right)^{s_k - r}\right)$$

$$P[A|F] = \prod_{uv} P[a_{uv}|F]$$

$$l(F) = log(P[A|F]) = \sum_{u,v} log(1 - \prod_{k=1}^{n} (1 - \binom{s_k}{r_k}) (1 - e^{-F_{uk}F_{vk}})^{r_k} (e^{-F_{uk}F_{vk}})^{s_k - r_k}))$$

$$\nabla_{F_{u,:}} l(F) = \sum_{j=v=1,\, u\neq v}^{n}$$

$$\frac{\sum_{k0=1}^{n}((\prod_{k=1,k\neq k0}^{n}(1-\binom{s_k}{r_k})(1-e^{-F_{uk}F_{vk}})^{r_k}(e^{-F_{uk}F_{vk}})^{s_k-r_k})\ (1-e^{-F_{uk_0}F_{vk_0}})^{r_{k_0}-1}\ (e^{-F_{uk_0}F_{vk_0}})^{s_{k_0}-r_{k_0}-1}(-F_{vk_0})\,(sk_0(1-e^{-F_{uk_0}F_{vk_0}})-r_{k0})\ )}{1-\prod_{k=1}^{n}(1-\binom{s_k}{r_k})(1-e^{-F_{uk}F_{vk}})^{r_k}(e^{-F_{uk}F_{vk}})^{s_k-r_k})}$$

---

۳ گناه بخت پریشان و دست کوته ماست!

پرسش تئوری ۹.

ماتریس مجاورت A روابط دوستی بین افراد را توصیف می کند و این روابط دوستی به میز افراد بستگی دارد. یعنی برای ماتریس z معین، درایه های $i = j$ ماتریس همان ۱ و درایه های دیگر متغیر تصادفی برنولی با پارامتر $p = 0.6$ اگر دو نفر i و j سر یک میز نشسته باشند و $q = 0.1$ اگر نه.

$$\forall i, j \in \{0, 1, ..., n-1\}: A_{i,j} = 1 \ if \ i = j$$
$$\sim Ber(p = 0.6) \ if \ z_0[i] = z_0[j]$$
$$\sim Ber(q = 0.1) \ if \ z_0[i] \neq z_0[j]$$

$$\mathbb{P}[A] = \prod_{i,j=0}^{n} \mathbb{P}[A_{i,j}]$$

پرسش تئوری ۱۰.

خیر، چون در ماتریس قسمت ۹، درایه های i و j از هم مستقل در نظر گرفته شده اند در حالی که $\mathbb{P}[A_{i,j}|A_{j,i}] = 1 \neq \mathbb{P}[A_{i,j}]$

پس ایندو از هم مستقل نیستند و بهتر است ماتریس A را اینگونه تعریف کنیم:

$$\forall i \in \{0, 1, ..., n-1\}, \forall j \in \{0, 1, ..., i\}:$$
$$A_{i,j} = 1 \ if \ i = j$$
$$\sim Ber(p = 0.6) \ if \ z_0[i] = z_0[j]$$
$$\sim Ber(q = 0.1) \ if \ z_0[i] \neq z_0[j]$$
$$A_{i,j} = A_{j,i}$$
$$\mathbb{P}[A] = \prod_{i,j=0}^{n} \mathbb{P}[A_{i,j}]$$

پرسش شبیه سازی ۲.

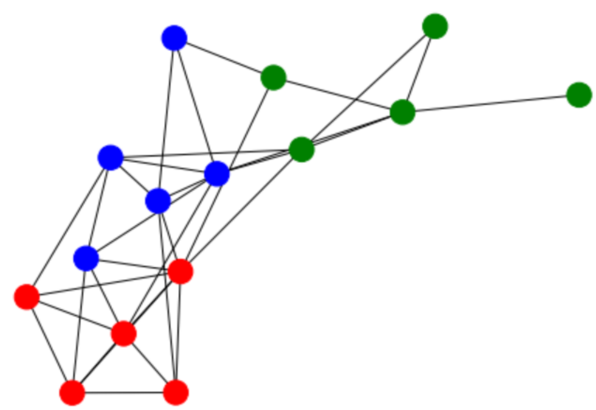
با استفاده از تابع `np.random.binomial` نمونه‌های خواسته شده ساخته شده اند. به دلیل زیاد بودن اینجا آورده نشده اند.

پرسش شبیه سازی ۳.

با استفاده از دستورات کتابخانه `networx` گراف خواسته شده، کشیده شد. تصویر آن ضمیمه شده است:

$$z_0 = [3,\ 1,\ 2,\ 1,\ 3,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 2,\ 1,\ 1,\ 3]$$

```
[[1. 1. 0. 0. 1. 0. 0. 0. 0. 0. 1. 0. 0. 1. 1.]
 [1. 1. 0. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
 [0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
 [0. 1. 0. 1. 0. 1. 0. 0. 0. 1. 0. 0. 1. 1. 0.]
 [1. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1.]
 [0. 1. 0. 1. 0. 1. 0. 0. 0. 1. 0. 0. 1. 1. 1.]
 [0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 1. 1. 0. 1. 1.]
 [0. 0. 1. 0. 0. 0. 1. 1. 1. 0. 0. 1. 0. 0. 1.]
 [0. 0. 0. 0. 1. 0. 0. 1. 1. 0. 0. 0. 0. 1. 0.]
 [0. 0. 0. 1. 0. 1. 0. 0. 0. 1. 1. 0. 0. 1. 1.]
 [1. 0. 0. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0. 1.]
 [0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 1. 0. 0. 0.]
 [0. 0. 0. 1. 0. 1. 0. 0. 0. 0. 1. 0. 1. 1. 0.]
 [1. 1. 0. 1. 0. 1. 1. 0. 1. 1. 0. 0. 1. 1. 0.]
 [1. 0. 0. 0. 1. 1. 1. 1. 0. 1. 1. 0. 0. 0. 1.]]
```



پرسش شبیه سازی ۴ و ۵.

برای پرسش ۴ به راحتی فقط در یک `for`، دو بردار با هم مقایسه می‌شوند و در پرسش ۵ در یک `for`، با استفاده از دستور `permutations` جایگشت های ممکن برای نام میزها در نظر گرفته شده است و سپس با استفاده از تابع بخش ۴ حداقل فاصله همینگ بین دو بردار به دست می‌آید.

پرسش تئوری ۱۱.

اگر درایه‌های قطر ماتریس که به هر حال ۱ هستند نیز مورد نظر باشند، به اندازه درایه‌های قطر و زیر آن متغیر تصادفی مستقل وجود دارد که یعنی

$$number\ of\ independent\ indexes = \frac{n^2-n}{2} + n = \frac{n^2+n}{2}$$

پرسش تئوری ۱۲.

از پرسش تئوری ۱۰ به دست آمد:

$$\forall i \in \{0, 1, ..., n-1\}, \forall j \in \{0, 1, ..., i\}:$$

$$A_{i,j} = 1\ if\ i = j$$

$$\sim Ber(p = 0.6)\ if\ z_0[i] = z_0[j]$$

$$\sim Ber(q = 0.1)\ if\ z_0[i] \neq z_0[j]$$

$$A_{i,j} = A_{j,i}$$

$$L(z) = \mathbb{P}[A|z] = \prod_{i,j=0}^{n} \mathbb{P}[A_{i,j}|z] = p^a(1-p)^b q^c(1-q)^d$$

Assuming: a= the number of people that sit around the same table and are friends

$$(A_{i,j} = 1,\ z[i] = z[j])$$

b= the number of people that sit around the same table but aren't friends

$$(A_{i,j} \neq 1,\ z[i] = z[j])$$

c= the number of people that don't sit around the same table but are friends

$$(A_{i,j} = 1,\ z[i] \neq z[j])$$

d= the number of people that don't sit around the same table and aren't friends

$$(A_{i,j} \neq 1,\ z[i] \neq z[j])$$

پرسش تئوری ۱۳.

در ادامه پرسش تئوری ۱۲ و با همان فرضیات:

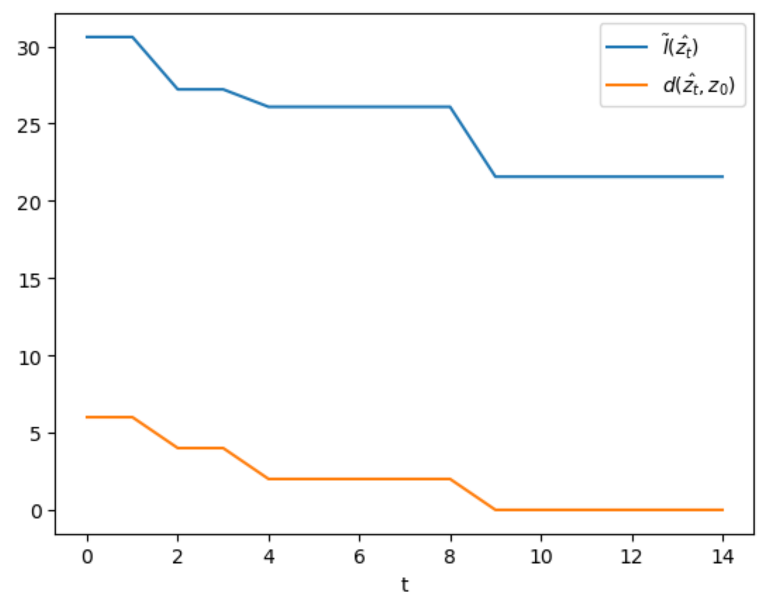
$$l(z) = log(L(z)) = log(p^a(1-p)^b q^c(1-q)^d) =$$
$$= a \cdot log(p) + b \cdot log(1-p) + c \cdot log(q) + d \cdot log(1-q)$$

پرسش شبیه سازی ۶.

چون محاسبه عددی با کامپیوتر راحت است، از رابطه به دست آمده در پرسش تئوری ۱۲ در کد استفاده شده است.

پرسش شبیه سازی ۷.

برای دقت بالا پارامتر $T = 15$ در نظر گرفته شده است. در این کد، از بردار z_0 تعریف شده در ابتدای بخش ۳ استفاده شده است و با استفاده از آن و تابع نوشته شده در پرسش شبیه سازی ۲، نمونه‌ای از ماتریس A تولید شده است و سپس الگوریتم ذکر شده روی ماتریس A اجرا می‌شود. منظور از z_0 در $d(\hat{z}_t, z_0)$ همان بردار مولد ماتریس A فرض شده است. نمودار به دست آمده:



پرسش شبیه سازی ۸.

۱۰ بار کد بخش قبلی تکرار می‌شود و هر دفعه ماتریس مجاورتی بر اساس جایگشتی از Z_0 درست می‌شود و پس از حدس آرایه مولد نتیجه آن‌ها بر اساس مینیمم فاصله همینگ با آن جابگشت از Z_0 مقایسه می‌شوند که قابل مشاهده است که در هر ورژن از کد، خروجی‌ها یعنی $\hat{Z}_T^j$ فاصله همینگ کمی با Z_0 دارند که به دلیل پارامتر $T = 15$ می‌باشد.

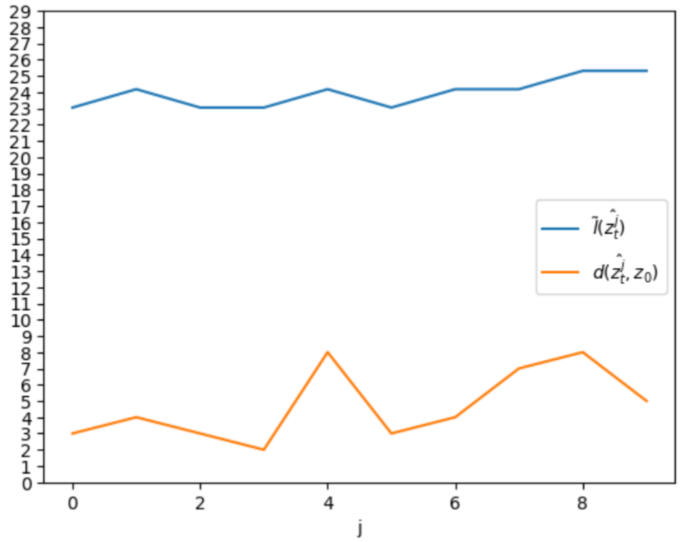
ماتریس ۲ بعدی های حدس زده شده هستند و آرایه آخری مینیمم فاصله همینگ حدس با آرایه مولد ماتریس مجاورت است.

```
[[3 1 2 1 3 1 2 2 2 3 3 2 1 1 3]
 [3 2 3 2 1 1 3 2 3 1 1 3 2 2 1]
 [1 3 2 3 1 3 2 2 2 1 1 2 3 3 1]
 [3 2 1 2 3 2 1 1 1 3 3 1 2 2 3]
 [3 1 2 1 3 1 2 2 2 3 3 2 1 1 3]
 [3 2 3 2 1 1 3 2 3 1 1 3 2 2 1]
 [1 3 2 3 3 1 2 2 2 1 1 2 3 3 1]
 [2 1 3 1 2 3 3 1 3 2 2 3 1 1 2]
 [2 1 1 3 2 1 3 3 3 2 2 3 1 1 2]
 [2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]]
[0. 3. 0. 0. 0. 3. 2. 2. 2. 0.]
```

پرسش شبیه سازی ۹ و ۱۰.

کد این قسمت درست نمایی خروجی‌ها را با درست نمایی Z_0 مقایسه می‌کند و مشاهده می‌شود که اگر و تنها اگر درست نمایی تخمین با Z_0 برابر باشد، فاصله همینگ این دو صفر است. در کد خروجی قابل مشاهده است که معمولا چند خروجی فاصله همینگ با آنها برابر با صفر است که یعنی تخمین Z_0 به درستی انجام شده است.

```
For j=0, 2, 3, 5,  l(zT) is equal to l(z0) and the minimum Hamming Distance is 3.0, 3.0, 2.0, 3.0
```



پرسش شبیه سازی ۱۱.

بهترین نتیجه به دست آمده حالتی است که برای هر دو نمونه تولید شده تخمین به همان بردار اصلی Z_0 رسیده است و حتی اگر به خود آن نرسد به حالت نامگذاری می‌رسد که فاصله همینگ آن با Z_0 برابر با صفر است. در چندین بار انجام کد حالتی که خروجی با فاصله همینگ صفر از Z_0 وجود نداشته باشد دیده نشد.

```
[[3 1 2 1 3 1 2 2 2 3 3 2 1 1 3]
 [1 3 2 3 1 3 2 2 2 1 1 2 3 3 1]
 [1 3 2 3 3 1 2 2 2 1 1 2 1 3 3]
 [2 1 3 1 2 1 3 3 3 2 2 3 1 1 2]
 [3 1 2 1 3 1 2 2 2 3 3 2 1 1 3]
 [2 3 2 3 3 1 1 1 1 2 2 2 1 3 3]
 [2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]
 [2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]
 [1 3 2 3 1 2 2 3 2 1 1 2 3 3 1]
 [1 2 1 2 2 3 3 3 3 1 1 1 3 2 2]]
[[2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]
 [3 1 2 1 3 1 2 2 2 3 3 2 1 1 3]
 [1 3 2 3 1 3 2 2 2 1 1 2 3 3 1]
 [2 1 3 1 3 1 3 2 3 2 2 3 1 1 2]
 [2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]
 [1 3 2 2 1 3 2 3 2 1 1 2 3 3 1]
 [3 2 1 2 3 1 1 2 1 3 3 1 2 2 3]
 [2 1 3 1 1 2 3 3 3 2 2 3 1 1 2]
 [2 3 1 3 2 3 1 1 1 2 2 1 3 3 2]
 [1 2 3 2 1 2 3 3 3 1 1 3 2 2 1]]
For the first sample
I found the original zo!!!
 Z=
[3, 1, 2, 1, 3, 1, 2, 2, 2, 3, 3, 2, 1, 1, 3]


For the second sample:
I found the original zo!!!
 Z=
[3, 1, 2, 1, 3, 1, 2, 2, 2, 3, 3, 2, 1, 1, 3]
```

۴ او نه خیال است و نه طیف!

پرسش تئوری ۱۴.

هر یک از درایه های ماتریس A، یک متغیر تصادفی گسسته (برنولی) هست که تابع جرم احتمال آن به شکل زیر میباشد:

اگر احتمال هم ژانر بودن دو نفر i,j را $\rho_{i,j}$ در نظر بگیریم:

$$P(A_{i,j} = 1) = p\left(\rho_{i,j}\right) + q(1 - \rho_{i,j})$$

$$P(A_{i,j} = 0) = 1 - P(A_{i,j} = 1)$$

$$P(A_{i,j} \neq 0 \cap A_{i,j} \neq 1) = 0$$

پرسش تئوری ۱۵.

$$W_{ij} = E\left[A_{ij}\right] = \Sigma A_{i,j}P(A_{i,j}) = 1 \times P(A_{i,j} = 1) + 0 \times P(A_{i,j} = 0)$$

$$\Rightarrow W_{ij} = P(A_{i,j} = 1)$$

و ماتریس W به فرم زیر بدست میاید:

$$\left[p, \quad (p - q)\rho_{12} + q, \quad (p - q)\rho_{13} + q, \quad (p - q)\rho_{14} + q, \; ......\right]$$

$$\left[(p - q)\rho_{12} + q, \quad p, \quad (p - q)\rho_{23} + q, \quad (p - q)\rho_{24} + q, \; ......\right]$$

$$\left[(p - q)\rho_{13} + q, \quad (p - q)\rho_{23} + q, \quad p, \quad (p - q)\rho_{34} + q, \; ......\right]$$

$$[...............\,, \quad ....\,, \quad ....\,, \quad ....\,, \quad ....]$$

پرسش تئوری ۱۶.

در این بخش ρ برای افرادی که در یک ژانر هستند برابر 1 و برای افراد در دو ژانر مختلف برابر صفر است. پس ماتریس W به این فرم خواهد بود:

W =

$$[p, \ p, \ q, \ q]$$

$$[p, \ p, \ q, \ q]$$

$$[q, \ q, \ p, \ p]$$

$$[q, \ q, \ p, \ p]$$

برای محاسبه مقادیر و بردارهای ویژه ماتریس داریم:

$$Wu = \lambda u \Rightarrow det(W - \lambda I) = 0$$

$$det([p - \lambda, \ p, \ q, \ q; p, \ p - \lambda, \ q, \ q; q, \ q, \ p - \lambda, \ p; q, \ q, \ p, \ p - \lambda]) = 0$$

$$\lambda_1 = 0 \qquad \lambda_2 = 0 \qquad \lambda_3 = 2p - 2q \qquad \lambda_4 = 2p + 2q$$

$$u_1 = [-1; 1; 0; 0] \qquad u_1 = [0; 0; -1; 1] \qquad u_1 = [-1; -1; 1; 1] \qquad u_1 = [1; 1; 1; 1]$$

پرسش تئوری ۱۷.

تعداد r نفر را در ژانر اول و n-r نفر را در ژانر دوم در نظر میگیریم. با استفاده از عملیات سطری مقدماتی w را به 'w به فرم زیر در می آوریم:

$$W' = PWP^T = [P_{r\times r} \quad q_{r\times(n-r)}; \ q_{(n-r)\times r} \quad P_{(n-r)\times(n-r)}]$$

در ماتریس های p و q تمام درایه ها، p و q هستند.

$$W'x = y \rightarrow \left\{ y_i = \sum_{j=1}^{r} px_j + \sum_{j=1+r}^{n} qx_j \qquad 1 \le i \le r \right\}$$

$$\rightarrow \left\{ y_i = \sum_{j=1}^{r} px_j + \sum_{j=1+r}^{n} qx_j \qquad r + 1 \le i \le n \right\}$$

با توجه به اینکه y_i های $1 \le i \le r$ مقادیر یکسان و y_iهای $r + 1 \le i \le n$ مقادیر ثابتی هستند، در صورتی که x بردار ویژه

باشد، r درایه ی اول آن باید برابر باشند (a) و n-r درایه ی بعد نیز باید یکسان باشند(b). پس با جایگذاری در رابطه ی بالا خواهیم داشت:

$$\lambda_a = rpa + (n - r)qb \;\rightarrow\; (\lambda - rp)a = (n - r)qb$$

$$\lambda_b = rqa + (n - r)pb \;\rightarrow\; b(\lambda - (n - r)p) = rqb$$

$$\Rightarrow\; \lambda^2 - np\lambda + (n - r)r\left[p^2 - q^2\right] = 0$$

ریشه های دستگاه فوق، مقادیر ویژه ناصفر 'w هستند . همینطور میدانیم با اعمال عملیات ستونی مقدماتی، مقادیر ویژه w تغییر نمیکند. پس ریشه ها

ی دستگاه فوق مقادیر ویژه ی w نیز هستند.برای محاسبه ی بردار ویژه ها، با بدست اوردن λ میتوان a,b را محاسبه کرد.

$$W'x = \lambda x \;\rightarrow P^T WPP^T x = \lambda P^T x$$

$\Leftarrow P^T x$ بردار ویژه متناظر ماتریس W هست.

پرسش تئوری ۱۸.

طبق w بدست آمده از قسمت ۱۶داریم:

$$D_W =$$

$$\begin{bmatrix} 2p + 2q & 0 & 0 & 0 \\ 0 & 2p + 2q & 0 & 0 \\ 0 & 0 & 2p + 2q & 0 \\ 0 & 0 & 0 & 2p + 2q \end{bmatrix}$$

پرسش تئوری ۱۹.

$L_W =$

$$[p + 2q \quad -p \quad -q \quad -q]$$

$$[-p \quad p + 2q \quad -q \quad -q]$$

$$[-q \quad -q \quad p + 2q \quad -p]$$

$$[-q \quad -q \quad -p \quad p + 2q]$$

$det(L_W - \lambda I) = 0 \rightarrow$

$(p + 2q - \lambda)(4q^3 - 10\lambda q^2 + 8pq^2 + 6\lambda^2 q + 4p^2 q - 12\lambda pq - \lambda^3 - 2\lambda p^2$

$+ 3\lambda^2 p) + p(-4q^3 + 2\lambda q^2 - 8pq^2 - 4p^2 q + 4\lambda pq + 2\lambda p^2 - \lambda^2 p)$

$- q(4q^3 - 4\lambda q^2 + 8pq^2 + \lambda^2 q + 4p^2 q - 4\lambda pq) + q(-4q^3 + 4\lambda q^2 - 8pq^2$

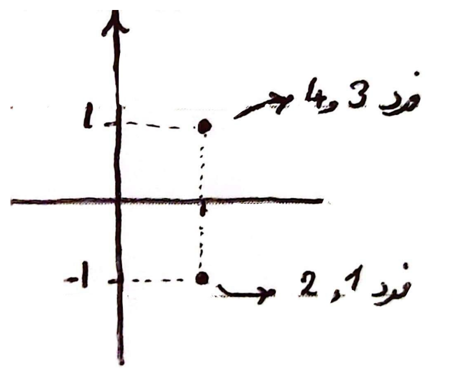
$- \lambda^2 q - 4p^2 q + 4\lambda pq) = 0 \qquad \Rightarrow \lambda = \{0, 4q, 2p + 2q, 2p + 2q\}$

$\lambda = 0 \Rightarrow u = [1; 1; 1; 1] \qquad \lambda = 4q \Rightarrow u = [-1; -1; 1; 1]$

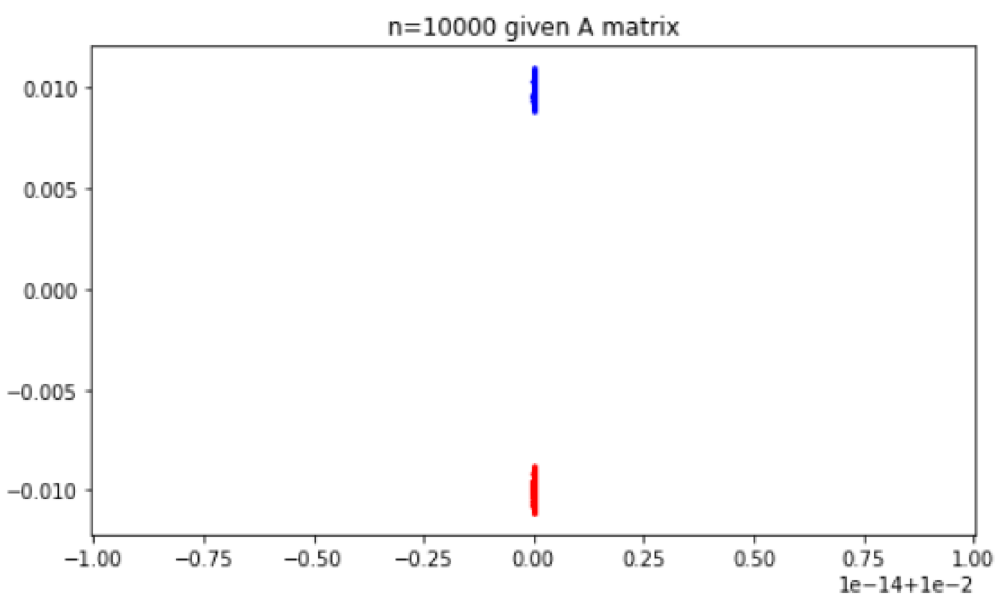
$\lambda = 2p + 2q \Rightarrow u = [-1; 1; 0; 0] \qquad \lambda = 2p + 2q \Rightarrow u = [0; 0; -1; 1]$

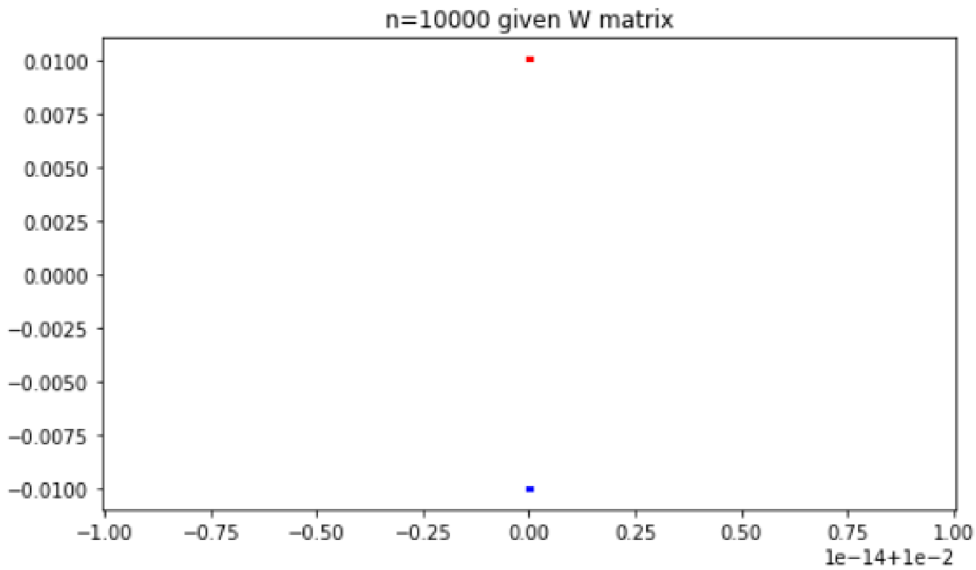
پرسش تئوری ۲۰.

با توجه به اینکه دو مقدار ویژه ی کوچکتر 0 و 4q هستند، باید هریک از درایه های نظیر به نظیر بردار ویژه های آنها را به عنوان x و y یک نقطه(به نمایندگی از فرد متناظر با آن درایه) در نظر بگیریم. در اینصورت میتوان افراد را به خوبی دسته بندی کرد:



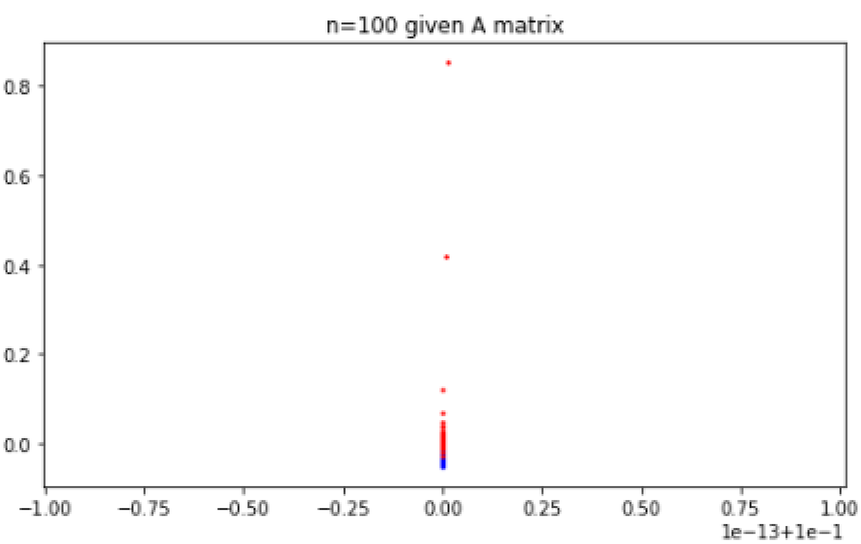
پرسش شبیه سازی ۱۲.

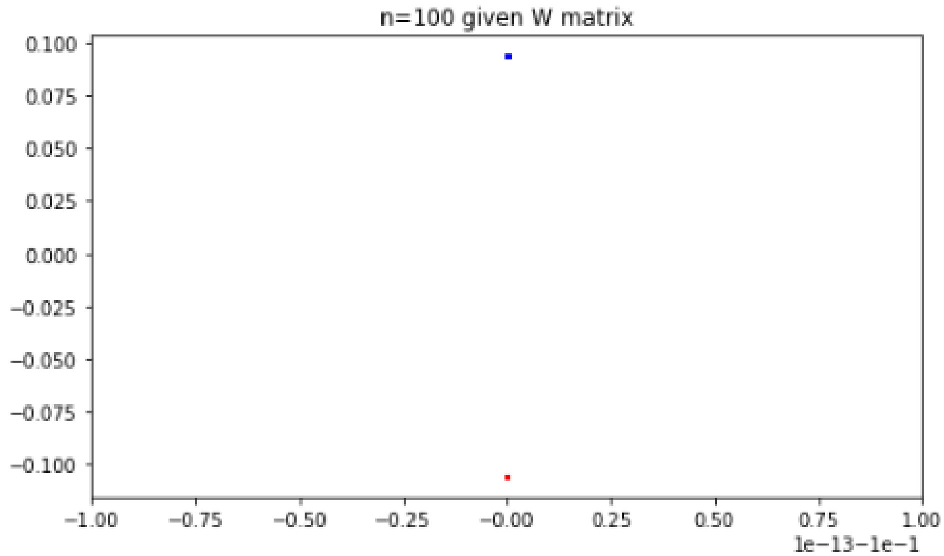


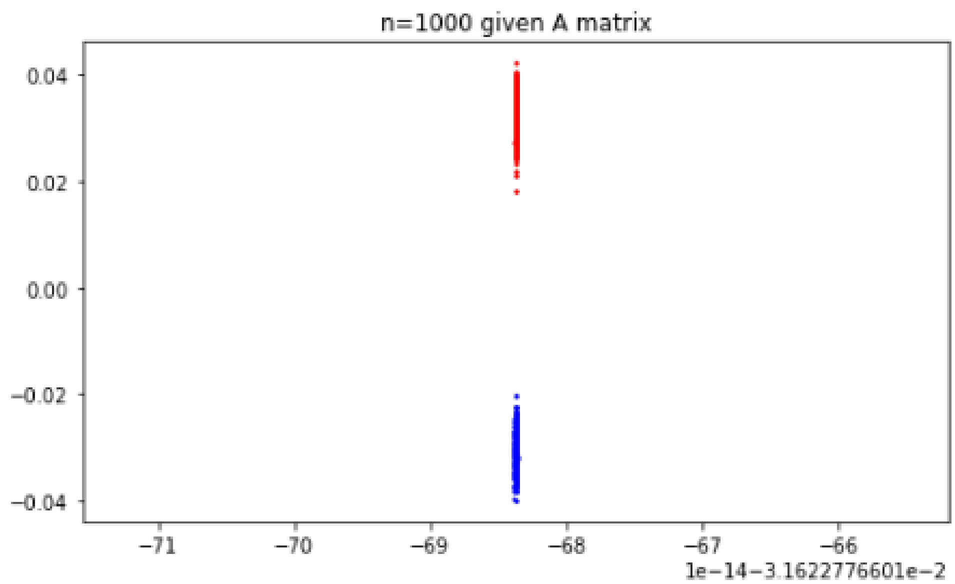


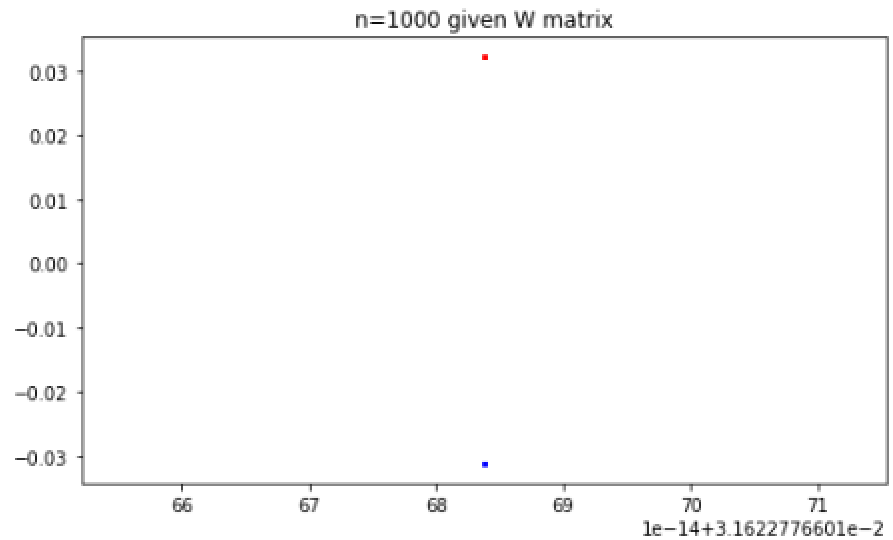
دو گروه 5000تایی از افراد را به طور رندوم ایجاد کرده و همانطور که مشخص است میزان خطای الگوریتم در یافتن دسته ها در A بسیار بیشتر از W میباشد و با توجه به افزایش تعداد n میزان خطای W کاهش بیشتری یافته است.

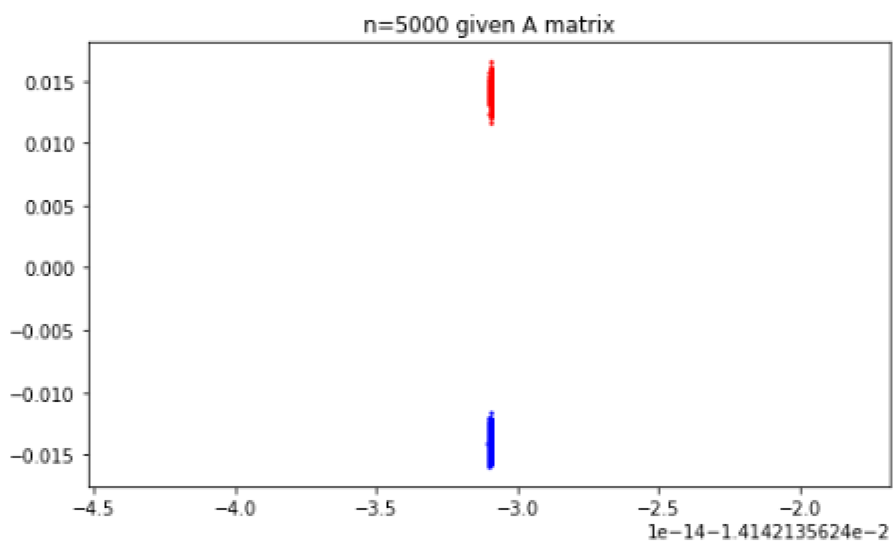
پرسش شبیه سازی ۱۳.

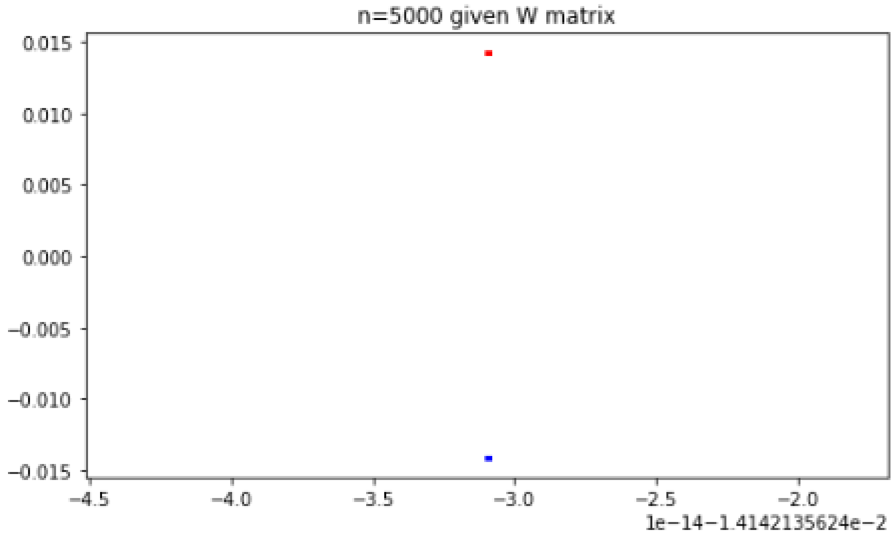



n=100 given W matrix
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
−1.00
−0.75
−0.50
−0.25
−0.00
0.25
0.50
0.75
1.00
1e−13−1e−1


n=1000 given A matrix
0.04
0.02
0.00
−0.02
−0.04
−71
−70
−69
−68
−67
−66
1e−14−3.1622776601e−2


n=1000 given W matrix
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
66
67
68
69
70
71
1e−14+3.1622776601e−2


n=5000 given A matrix
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−4.5
−4.0
−3.5
−3.0
−2.5
−2.0
1e−14−1.4142135624e−2

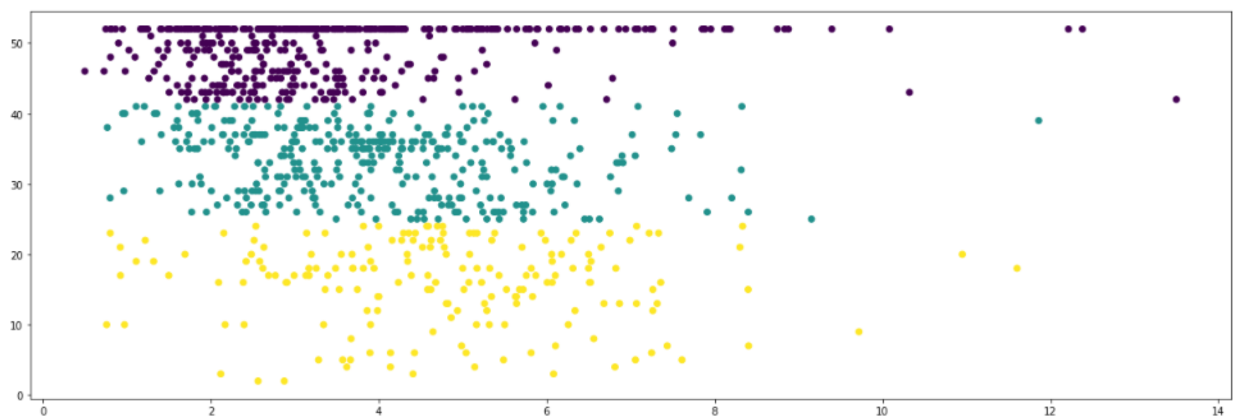


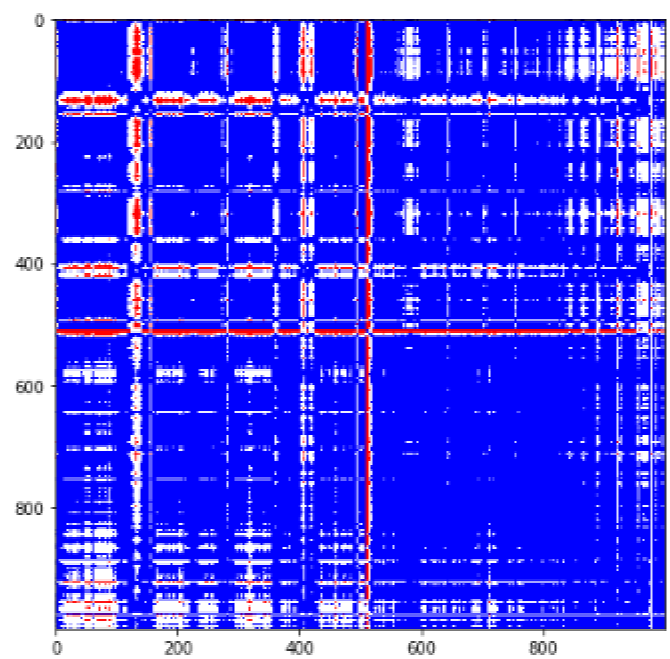
همانطور که در شبیه سازی ۱۲ گفته شد، با افزایش n در هر دو حالت به دسته ها نزدیکتر میشویم اما همچنان خطای A بیشتر از W است و پراکندگی بیشتری دارد.

پرسش تئوری ۲۲.

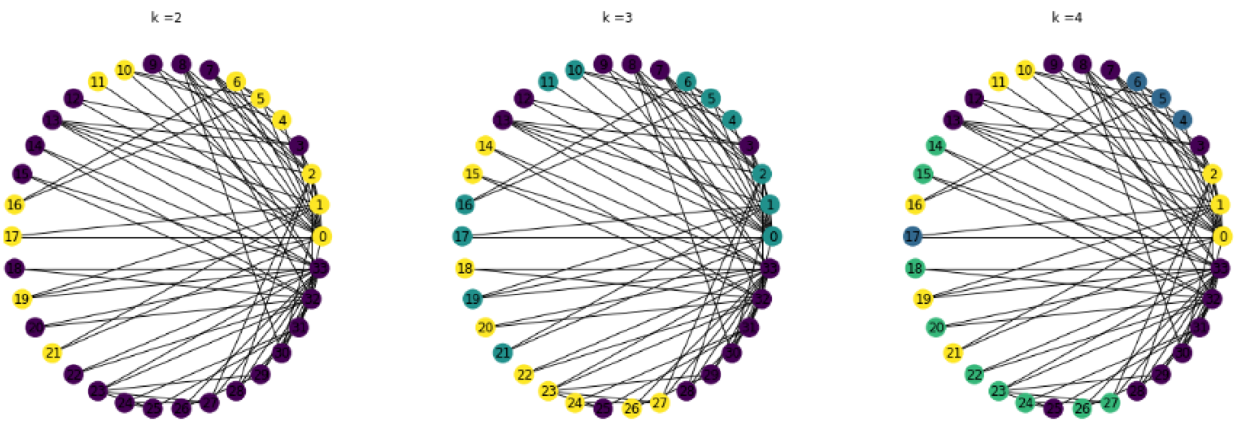
روش حل برای بیش از 2 دسته همانند روشی است که در K=2 بررسی کردیم و باید ماتریس لاپلاسین را همانند الگوریتم به کارگرفته شده در پرسش های قبلی بدست اورده و بردارهای ویژه آن را محاسبه کنیم. حال k تا از کوچکترین ضرایب ویژه را در نظر گرفته و به کمک بردار ویژه آنها داده ها در دستگاه k بعدی خوشه بندی میکنیم. با زیاد شدن تعداد دسته ها تصور آنها در دستگاه k بعدی سخت تر شده و معمولا در ادامه از روش k-means استفاده میکنیم .

پرسش شبیه سازی ۱۴.





پرسش شبیه سازی ۱۵.



5 من از دیار حبیبم نه از بلاد غریب!

پرسش تئوری ۲۳.

هر یال این گراف یک متغیر تصادفی برنولی با پارامتر p است و تعداد یال‌های یک گراف کامل $C(n,\ 2) = \frac{n(n-1)}{2}$ است پس هر گراف یک متغیر دو جمله ای با پرامتر های $\frac{n(n-1)}{2}$ و p است.

احتمال اینکه در چینش تصادفی این گراف، گراف مورد نظر با m یال به دست آید برابر است با:

$$\mathbb{P}[desired\ graph]\ = m^{p}(\frac{n(n-1)}{2} -\ m)^{1-p}$$

پرسش تئوری ۲۴.

تعداد یال‌های یک گراف کامل $C(n,\ 2) = \frac{n(n-1)}{2}$ است پس هنگام انتخاب m رابطه هم‌سلیقگی، باید از $\frac{n(n-1)}{2}$ یال m یال انتخاب شود پس تعداد حالات ممکن $C(\frac{n(n-1)}{2}\ ,\ m)$ است پس احتمال انتخاب گراف مورد نظر:

$$\mathbb{P}[desired\ graph]\ = \frac{1}{C(\frac{n(n-1)}{2}\ , m)} = \frac{m!\ (\frac{n(n-1)}{2} - m)!}{(\frac{n(n-1)}{2})!}$$

پرسش تئوری ۲۵.

برای حدس درست ۲۰٪ از m یال گراف، باید $\lceil \frac{m}{5} \rceil$ یال درست حدس زده شود و $m\ -\ \lceil \frac{m}{5} \rceil$ یال غلط و بقیه یال‌ها می‌توانند باشند یا نباشند:

$$\mathbb{P}[\ \lceil \frac{m}{5} \rceil\ edges\ correct]\ = (\ \lceil \frac{m}{5} \rceil)^{p}(m\ -\ \lceil \frac{m}{5} \rceil)^{1-p}$$

پرسش شبیه سازی ۱۷.

همانطور که در پرسش تئوری ۲۳ گفته شد، تعداد یال‌های این گراف یک متغیر تصادفی $Bin(\frac{n(n-1)}{2},\ p)$ است پس برای شبیه سازی ۱۷ از دستور `np.random.binomial` استفاده شده است. درصد خطا در حدود $(40\ -\ 45)\%$ به دست آمد و میانگین با m برابر نیست.

پرسش تئوری ۲۶.

برای یک متغیر دوجمله‌ای با پارامترهای n و p امید ریاضی برابر است با np پس برای متغیر دوجمله‌ای تعداد یال‌ها، میانگین برابر است با:

$$\mathbb{E}[N]\ =\frac{n(n-1)}{2}\cdot\ p\ =\ 1698.3$$

در حالت کلی برای برابر شدن میانگین تعداد یال‌ها و :

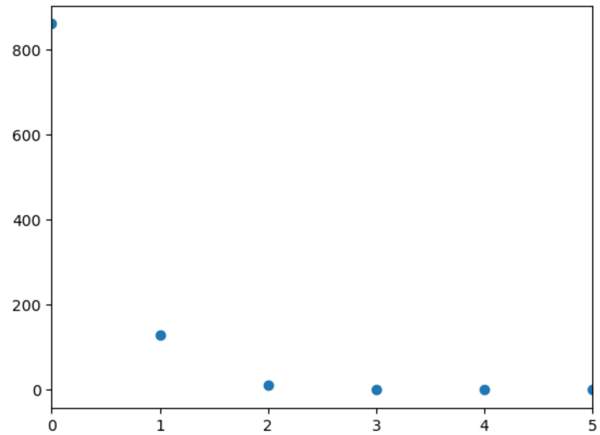
$$\mathbb{E}[N]\ =\frac{n(n-1)}{2}\cdot\ p\ =\ m\ \rightarrow\ n(n\ -\ 1)\ =\ 2mp$$

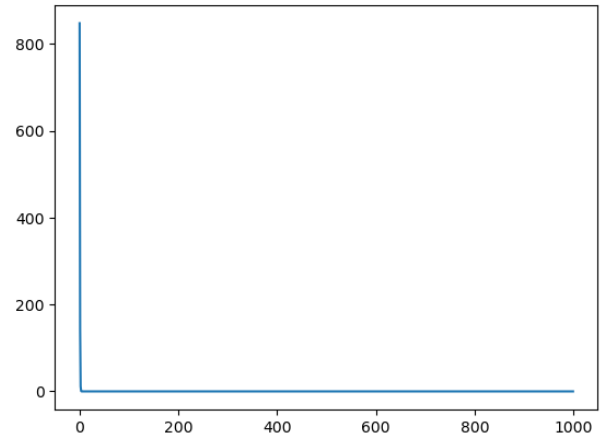
پرسش شبیه سازی ۱۸.

برای ایجاد گراف، از ماتریس مجاورت استفاده شده است به طوری که:

$$A_{i,j}=A_{j,\,i}\quad,\quad A_{i,\,i}=\ 0\quad,\quad A_{i,\,j}\sim\ Ber(p\ =\ 0.00016)$$

سپس با استفاده از این ماتریس، دو آرایه، یکی برای تعداد یال‌های هر رأس و یکی برای تعداد رأس‌هایی که تعداد معینی یال دارند ایجاد شده است. با استفاده از آرایه اولی پارامتر L با میانگین گیری و سپس متوسط تعداد افراد هم‌رنگ به دست آمد. از آرایه دوم نیز پس از میانگین گیری بر محور عمودی، برای کشیدن نمودار استفاده شد. شکل نمودار به این دلیل است که L در واقع بسیار کوچک است و هر فرد، هم سلیقه‌های کمی دارد. (نمودار سمت چپ زوم شده سمت راست به حالت نقطه‌ای است)





پرسش تئوری ۲۷.

هر فرد (رأس) تعداد افراد هم‌سلیقه (یال) یک متغیر دوجمله‌ای با پارامتر p و $n - 1$ است پس امید ریاضی (میانگین) تعداد یال‌های هر رأس برابر است با:

$$L = \mathbb{E}[N_i] = p(n-1) = 0.15984$$

پرسش تئوری ۲۸.

هر فرد (رأس) به طور میانگین ۰.۱۵ نفر هم‌سلیقه (یال) دارد که یعنی هر فرد (رأس) اگر هم‌سلیقه‌ای (یالی) داشته باشد پس هم‌رنگ است:

H:تعداد افراد هم‌رنگ , H_i: همرنگ بودن هر فرد $\rightarrow$ $H = \sum_{i=1}^{n} H_i$

$$\mathbb{P}[H_i] = 1 - \mathbb{P}[isolated\,node] = 1 - (1-p)^{(n-1)} = 0.1477$$

$$\mathbb{E}[H] = \sum_{i=1}^{n} \mathbb{E}[H_i] = n \cdot \mathbb{E}[H_i] = n\mathbb{P}[H_i] = 147.7$$

پرسش شبیه سازی ۱۹.

با بررسی تک به تک آرایه‌های واتریس مجاورت گراف به دست آمد: (عدد آخر برای استفاده در پرسش تئوری ۳۰ چاپ شده است.)

```
The mean of Transitive relationships is 4473.0.

The mean of Chain relationships is 446304.8.

0.009922848906933748
```

پرسش تئوری ۲۹.

با اعداد پرسش تئوری ۱۹:

T: تعداد روابط تراگذر , T_i: تراگذر بودن ۳ یال $\rightarrow$ $T = \sum_{i=1}^{C(n,3)} T_i$

$$\mathbb{P}[T_i] = p^3 \rightarrow \mathbb{E}[T] = \sum_{i=1}^{C(n,3)} \mathbb{E}[T_i] = \frac{n(n-1)(n-2)}{6} \cdot p^3 = 4495.501$$

C: تعداد روابط زنجیره‌ای , C_i: زنجیره‌ای بودن ۳ یال $\rightarrow$ $C = \sum_{i=1}^{C(n,3)} C_i$

$$\mathbb{P}[C_i] = p^2(1-p) \rightarrow \mathbb{E}[C] = \sum_{i=1}^{C(n,3)} \mathbb{E}[C_i] = \frac{n(n-1)(n-2)}{6} \cdot p^2(1-p) = 445,054.599$$

که این اعداد با اعداد پرسش تئوری ۱۹ همخوانی دارند.

پرسش تئوری ۳۰.

$$\mathbb{P}[\Delta | At\ least\ 2\ sides\ to\ a\ triangle] = \frac{\mathbb{P}[\Delta \cap 2\ sides]}{\mathbb{P}[At\ least\ 2\ sides\ to\ a\ triangle]}$$

$$= \frac{p^3}{p^2\binom{3}{2}+p^3} = \frac{p}{3+p} = 0.0033$$

برای مقایسه نتیجه با شبیه‌سازی، شرط هر کدام از این سه نفر حداقل با یکی از دو نفر دیگر هم سلیقه باشد، یعنی باید کسر تعداد روابط تراگذر به روی جمع تعداد روابط تراگذر و زنجیره‌ای انجام شود که با نتایج درون گزارش، برابر است با ۰.۰۰۹ پس نتیجه تقریبا با تئوری همخوانی دارد.

پرسش شبیه سازی ۲۰.

برای بررسی روابط درون هم‌سلیقه‌های هر شخص، هم‌سلیقه‌های هر شخص در آرایه‌ای جمع شده اند و سپس با استفاده از این آرایه ماتریس مجاورت گراف به ردیف‌ها و ستون‌های هم‌سلیقه‌های هر شخص کاهش یافته است و سپس با استفاده از آن میانگین روابط به دست آمده است.

```
The mean of edges between one nodes's connected nodes is 0.009.
```

پرسش تئوری ۳۱.

امید ریاضی روابط هم‌سلیقگی درون هم‌سلیقه‌های یک شخص همان امید ریاضی روابط هم‌سلیقگی در گرافی با تعداد رأس به اندازه امید ریاضی هم‌سلیقه‌های یک شخص است پس:

$$\mathbb{E}[N_{N_i}] = p(\mathbb{E}[N_i] - 1) = p((n-1)p - 1) = 0.005991$$

پرسش شبیه سازی 21.

با استفاده از nx.average_shortest_path_length داریم:

```
3.884292292292924
```

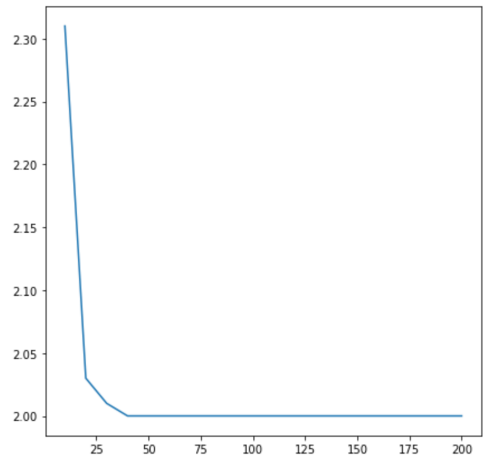
پرسش شبیه سازی 22.

زوج ها با بیشترین فاصله در ژوپیتر آمده است و اینجا داریم:

```
mean longest path is: 2.0 nodes
```

پرسش شبیه سازی 23.

در اینجا به دلیل طولانی بودن دیگر لیست ها را چاپ نکردیم و تنها نمودار را رسم میکنیم



این نمودار با زیاد شدن n به مقدار 200 میل میکند.

پرسش تئوری ۳۲.

احتمال اینکه دو رأس u و v همسایه مشترک نداشته باشند همانند آن است که هیچ راهی بین این دو رأس نباشد که طبق [این سایت] برابر است با: (i مجموعه همسایه‌های رأس u و j مجموعه همسایه‌های رأس v و خود رئوس و k رئوس نامتصل به هردو هستند.)

$$\mathbb{P}[I_{u,v}] = \sum_{i+j+k=n} \binom{n-2}{i-1,\, j-1,\, k} f(i,\ p) f(j,\ p)(1\ -\ p)^{ij+ik+jk}$$

$$= \sum_{i+j+k=n} \binom{n-2}{i-1,\, j-1,\, k} p^{i+j}(1\ -\ p)^{ij+ik+jk} = \sum_{i+j+k=n} \frac{(n-2)!}{(i-1)!(j-1)!k!} p^{i+j}(1\ -\ p)^{ij+ik+jk}$$

پرسش تئوری ۳۳.

$$\mathbb{E}[X_n] = \sum_{0\le u,\, v\le n} \mathbb{E}[I_{u,v}] = \binom{n}{2}\mathbb{P}[I_{u,v}] = \frac{n(n-1)}{2} \sum_{i+j+k=n} \frac{(n-2)!}{(i-1)!(j-1)!k!} p^{i+j}(1-p)^{ij+ik+jk}$$

پرسش تئوری ۳۴ و ۳۵.

$$\mathbb{P}[X_n \ge 1] \le \frac{\mathbb{E}[X_n]}{1} = \frac{n(n-1)}{2} \sum_{i+j+k=n} \frac{(n-2)!}{(i-1)!(j-1)!k!} p^{i+j}(1-p)^{ij+ik+jk}$$

من نتوانستم رفتار حدی این عبارت را به دست بیاورم ولی طبق سایت رفرنس شده:

> Asymptotically, for $p \ge \frac{1+\epsilon}{n}$, the random graph has a giant component of linear size, and all remaining components are sublinear. So up to lower-order terms, two vertices u, v are connected with about the same probability as that they're both in the giant component. If $p \sim \frac{c}{n}$, then $(x + o(1))n$ vertices are in the giant component w.h.p., where x satisfies $1 - x = e^{-cx}$, and the probability that u and v are not connected is therefore $(1-x)^2 + o(1)$.
>
> In particular, if $p \gg \frac{1}{n}$, the probability that u and v are not connected tends to 0 as $n \to \infty$.
>
> If $p \le \frac{1}{n}$, there is no giant component, and the probability that u and v are not connected tends to 1 as $n \to \infty$.

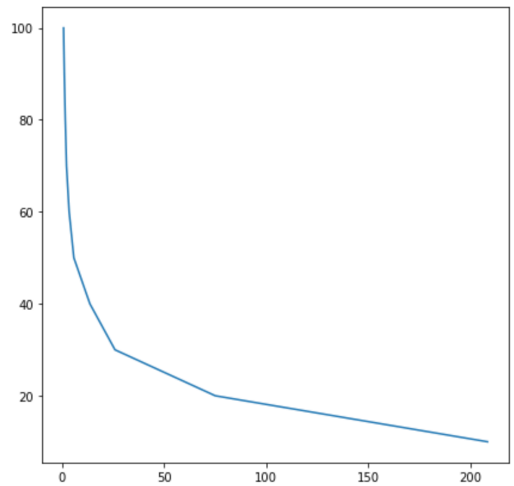
رفتار حدی این رابطه به p بستگی دارد به طوری که اگر $p \ge \frac{1}{n}$ باشد، گراف یک قسمت بزرگ متصل دارد که باعث می‌شود احتمال اینکه u و v همسایه مشترک داشته باشند برابر شود با احتمال اینکه این دو رأس در این قسمت بزرگ متصل باشند یا خیر. اگر $p \gg \frac{1}{n}$ احتمال اینکه u و v متصل نباشند به صفر می‌رود و اگر $p \le \frac{1}{n}$ دیگر قسمت بزرگ متصلی وجود ندارد و احتمال اینکه u و v متصل نباشند به 1 می‌رود.

پرسش شبیه سازی 24.

با استفاده از دستور nx.triangles داریم:

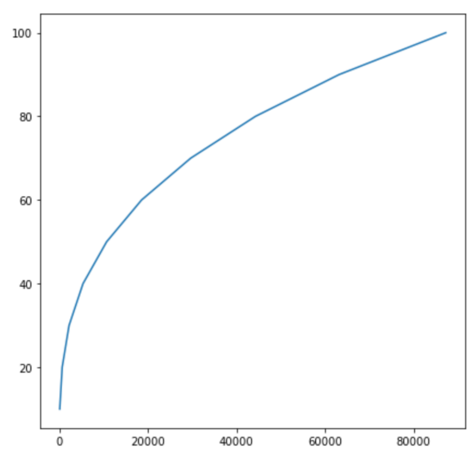
```
10528.92
```

پرسش شبیه سازی 25.



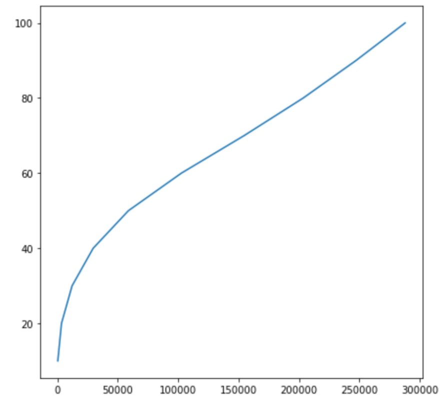
برای اینکه بفهمیم به چه عددی میل میکند n را بسیار زیاد میکنیم و مشاهده میکنیم که به 0 میل میکند دلیل اسن است که p با سرعت بیشتری به 0 میل میکند ویال های کمتری ایجاد می شود در نتیجه تعداد روابط 3 نفری هم کاهش میابد.

پرسش شبیه سازی 26.



در اینجا در حال زیاد شدن با شیب کمی است و به نظر میرسد ممکن است به مقادیر بالایی میل کند.

پرسش شبیه سازی 27.



این جا هم تا جایی که p از حدی بیشتر بود در حال همگرایی بود ولی از جایی که p افت میکند میانگین هم واگرا میشود.

---

6 سل المصانع رکباً تَهیم فی الفَلَوات!

پرسش تئوری 6.36

در یک حرکت به اینصورت احتمال محاسبه میشود:

چون حرکت یکنواخت است اگر j در همسایگی i باشد احتمال آن $\frac{1}{d_i}$ و اگر نباشد 0 میشود.

$$P_{ij} = \frac{1}{d_{ij}} \quad A_{ij} = 1\ ,\ \ 0 \quad A_{ij} = 0$$

پرسش تئوری 6.37

با امتحان کردن حالات و نوشتن ماتریس احتمال به این نتیجه میرسیم که:

$$P\ D = A$$

$$P = (A\,D^{-1})$$

پرسش تئوری 6.38

با نوشتن رابطه ضرب ماتریس ها به سادگی به این نتیجه میرسیم که

$$[P^2]_{ij} = \sum_k P_{ik}P_{kj}$$

تعبیر احتمالی به گونه ای است که اگر بخواهیم از i به j در 2 مرحله برویم احتمال ما چگونه میشود. منطقا باید تمام احتمال هایی که حالاتی که از i به k های مختلف و از k به j میرویم را باهم جمع کنیم.

پرسش تئوری 6.39

با تعمیم نتیجه رسیده در قسمت قبل خواهیم داشت:

$$P^{(t)}{}_{ij} = [P^t]_{ij} = \prod_t P_{ij}$$

پرسش تئوری 6.40

به صورت شهودی که کاملا واضح است که تمام مسیر از i به j و مسیر از j به i یکی بوده و تنها تفاوت در حرکت اولیه است به م در حالت اول یک احتمال $\frac{1}{d_i}$ و در حالت دوم یک احتمال $\frac{1}{d_j}$ میدهد. به صورت مثال برای دو مرحله:

$$[P^2]_{ij} = \sum_k P_{ik}P_{kj}$$

$$[P^2]_{ji} = \sum_k P_{jk}P_{ki}$$

چون ماتریس p یک ماتریس متقارن و ماتریس D^{-1} و D ماتریس های قطری هستند داریم:

$$D\,X\,D^{-1} = X\,D\,D^{-1} = X = P^t = (P^t)^T$$

اگر برای هر درایه این رابطه را مورد بررسی قرار دهیم داریم:

$$d(i)P^t{}_{ij} = d(j)p^t{}_{ji}$$

در نتیجه برای نسبت خواهیم داشت:

$$\frac{P^t{}_{ij}}{p^t{}_{ji}} = \frac{d(j)}{d(i)}$$

پرسش تئوری 6.41

اگر این کاربر ها هم سلیقه باشند یعنی یک یال بین آنان وجود دارد و با احتمال $\frac{1}{d_i}$ با یک حرکت میرسیم پس میتوانیم بگوییم که احتمال با t مرحله از $\frac{1}{d_i}$ بیشتر است.

در این حالت چون i و j هم سلیقه هستند پس دو احتمالی که از i به k و یا از j به k برویم برابر خواهند بود.

پرسش تئوری 6.42

چون هر کسی از دسته c به k برود احتمال برابری دارد پس می توانیم بگوییم اگر i از دسته c1 و j از دسته c2 باشد:

$$r_{c_1c_2}=\sqrt{\sum_{k=1}^{n}\frac{(\sum_{c_1}P_{ik}\ /c_1-\sum_{c2}P_{jK}/c_2)^2}{d(k)}}$$

پرسش تئوری 6.43

این الگوریتم را تا جایی باید ادامه دهیم که اختلاف سلیقه ها بیشینه شود.
به عنوان معیاری از بیشینه شدن ماتریس r میتوانیم به درایه های ماتریس نگاه کنیم به صورتی که اگر بیشتر از 80 درصد درایه ها در حال رشد بودند به دسته بندی ادامه میدهیم و بعد از آن دیگر دسته بندی را ادامه نمی دهیم.

پرسش شبیه سازی 6.28

با استفاده از داده ها گراف را ایجاد کرده و الگوریتم را پیاده سازی می کنیم

ماتریس C ماتریس دسته بندی ما است و دسته ها در آن آپدیت می شوند بر اساس ماتریس R که ماتریس اختلاف سلیقه دسته های موجود است. اندیس آخر دسته هایی که حذف می شوند -2 است.

پرسش شبیه سازی 6.29

مانند قسمت قبل خواهد بود.

References

Zachary's Karate Club graph

Data file from: http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm

Math stackexchange: Probability of not having a path between two certain nodes, in a random graph: https://math.stackexchange.com/q/2818773