به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های اجتماعی

محمد ناصری

۸۱۰۱۰۰۴۸۶

تمرین اول

آذر ۱۴۰۰

مقدمه

در پدیده شش گام فاصله ، دو فرد در هرکجای جهان می‌توانند از طریق زنجیره‌ای از آشنایان با طول 6 یا کمتر به هم متصل شوند. این بدان معنی است که اگرچه سارا مستقیما پیتر را نمی‌شناسد، اما رالف را می‌شناسد که او جان را می‌شناسد که درنهایت او پیتر را می‌شناسد. پس سارا 3 گام با پیتر فاصله دارد یا در درجه سوم از پیتر است. مفهوم فاصله شش درجه که به آن ویژگی جهان کوچک هم گفته می‌شود، در ادبیات علم شبکه بدان معنی است که فاصله بین هر دو گره در یک شبکه به‌طور غیرمنتظره‌ای کوچک است..

به زبان علم شبکه، پدیده جهان کوچک بیان می‌کند که فاصله بین هر دو گره دلخواه (تصادفی) در شبکه کوتاه است. لازم است به دو سؤال پاسخ داده شود: معیار کوتاه بودن فاصله چیست و در مقایسه با چه چیزی سنجیده می‌شود؟ چطور می‌توان وجود این فواصل کوتاه را توضیح داد؟

هردو سؤال با محاسبه ساده‌ای پاسخ داده می‌شوند. یک شبکه تصادفی با درجه میانگین را در نظر بگیرید. یک گره در این شبکه به‌طور میانگین دارای:

- $< k >$ گره در فاصله.(d=1)
- $\langle k\rangle^2$ گره در فاصله ،.(d=2)
- $\langle k\rangle^3$ گره در فاصله ،.(d =3)
- ...
- $\langle k\rangle^d$ گره در فاصله .d

است. برای مثال، هر فرد به طور متوسط با 1000 نفر آشنا است، پس اگر $< k > \approx 1000$ در نظر بگیریم، انتظار داریم 10^6 فرد در فاصله 2 از یکدیگر باشند و در حدود یک میلیارد نفر، یعنی تقریباً تمام جمعیت زمین، در فاصله 3 از ما قرار داشته باشند.

به بیانی دقیق‌تر، تعداد گره‌های مورد انتظار با فاصله d از یک گره برابر است با:

$$N(d)\approx 1+\langle k\rangle+\langle k\rangle^2+\ldots+\langle k\rangle^d=(\langle k\rangle^{d+1})/(1\langle k\rangle-1) \quad (١)$$

N(d)نباید از تعداد کل گره‌ها در شبکه، N، فراتر رود. بنابراین فواصل نمی‌توانند هر مقدار دلخواهی بگیرند. ما می‌توانیم بیشترین فاصله، d_{max}، یا قطر شبکه را با مقدار

$$N(d_{max})\approx N \quad (٢)$$

مشخص کنیم. با فرض اینکه $\langle k \rangle > 1$، می‌توانیم (1-) را در صورت و مخرج رابطه (۱) نادیده بگیریم که خواهیم داشت:

$$(۳)\ \langle k\rangle^{d_{max}} \approx N$$

بنابراین قطر یک شبکه تصادفی از عبارت زیر پیروی می‌کند:

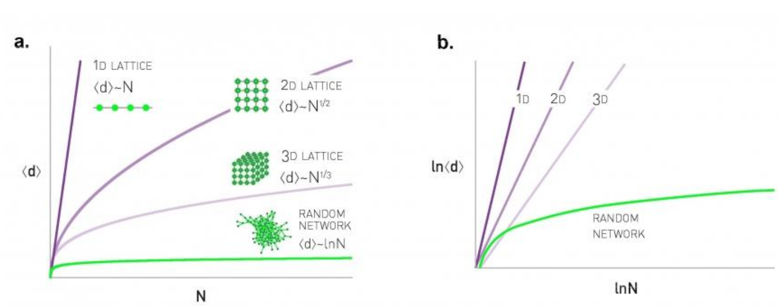
$$(۴)\ d_{max} \approx (\ln N) / (\ln\langle k\rangle)$$

رابطه (۴) مقیاس بندی قطر شبکه، d_{max}، را برای سیستمی با اندازه N محاسبه می‌کند. برای بیشتر شبکه‌ها (۴) تخمین بهتری از فاصله میانگین بین دو گره‌ای که بطور تصادفی انتخاب شده‌اند نسبت به d_{max} پیشنهاد می‌کند .این بدان خاطر است که d_{max} معمولاً به واسطه تعداد کمی مسیر تعریف می‌شود، درحالی‌که میانگین بین همه جفت مسیرها است، این فرایند باعث از بین رفتن اختلاف زیاد در تخمین ها می شود ازاین‌رو معمولاً مشخصه دنیای کوچک این‌گونه تعریف می‌شود:

$$(۵)\ \langle d\rangle \approx \ln N / \ln\langle k\rangle$$

که نیازمندی های میانگین فاصله در شبکه ای با پارامترهای N و ر را توصیف می کند.

- به‌طورکلی $\ln(N) \ll N$ است، بنابراین وابستگی $\langle d \rangle$ به $\ln(N)$ نشان‌دهنده آن است که فواصل در یک شبکه تصادفی نسبت به‌اندازه شبکه بسیار کوچک‌تر هستند. درنتیجه منظورمان از کوچک (کوتاه) در پدیده جهان کوچک این است که طول مسیر میانگین یا قطر به‌صورت لگاریتمی به‌اندازه سیستم وابسته است. بنابراین "کوچک" یعنی با $\ln(N)$ متناسب است، نه با N یا توانی از N .
- عبارت $1/\ln\langle k \rangle$ دلالت دارد بر این‌که هرچقدر شبکه متراکم‌تر باشد، فاصله بین گره‌ها کوتاه‌تر است.
- در شبکه‌های واقعی رابطه (۵) اصلاح می‌شود و این واقعیت آشکار می‌شود که تعداد گره‌ها در فاصله d به‌سرعت کم می‌شود.



چرا پدیده جهان کوچک غیرمنتظره است؟

بخش اعظم شهود ما درباره فاصله بر تجربه ما از توری‌های منظم استوار است، که ویژگی جهان کوچک را نشان نمی‌دهد:

توری یک‌بعدی: برای یک توری منظم یک‌بعدی) خطی به طول (N قطر و میانگین طول مسیر بر اساس N به‌صورت خطی رشد می‌کند $N\sim\langle d\rangle\sim d_{max}$:

توری دوبعدی: برای توری منظم مربعی $d_{max}\sim\langle d\rangle\sim N^{1/2}$ است.

توری سه‌بعدی: برای توری منظم مکعبی $d_{max}\sim\langle d\rangle\sim N^{1/3}$ است.

چهاربعدی: به‌طورکلی، برای توری منظم d–بعدی $d_{max}\sim\langle d\rangle\sim N^{1/d}$ است.

این وابستگی‌های نمایی چندجانبه افزایش سریع‌تری برای N نسبت به رابطه (۵) قائل‌اند، که مشخص می‌کند در توری‌های منظم طول مسیرها به‌طور قابل‌ملاحظه‌ای طولانی‌تر از شبکه تصادفی است. برای مثال اگر شبکه‌های اجتماعی یک توری منظم مربعی (2–بعدی) تشکیل می‌داد، که در آن هر شخص تنها همسایگانش را می‌شناخت، فاصله میانگین بین دو فرد دقیقاً $(83{,}666 = (7 * 10^9)^{1/2})$ می‌شد. حتی اگر این واقعیت را هم اصلاح کنیم که هر شخص حدود 1000 آشنا دارد، نه چهارتا، بازهم متوسط جدایی از مرتبه‌ای بزرگ‌تر ازآنچه توسط رابطه (۵) پیش‌بینی‌شده است، به دست می‌آمد.

اجازه دهید پیامدهای رابطه ۵ را برای شبکه‌های اجتماعی توضیح دهیم. با در نظر گرفتن $N \approx 7 \times 10^9$ and $\langle k\rangle \approx 10^3$، خواهیم داشت:

$$\langle d\rangle \approx (\ln 7\times10^9) / \ln(103) = 3.28 \quad (۶)$$

بنابرین همه افراد روی زمین باید در فاصله آشنایی 3 یا 4 از یکدیگر قرار داشته باشند. تخمین رابطه (۶) احتمالاً به مقدار واقعی نزدیک تراست تا درجه 6، که بیشتر نقل می‌شود.

بیشتر آن چیزی که ما از ویژگی جهان کوچک در شبکه‌های تصادفی می‌دانیم، که نتیجه رابطه (۵) را هم دربر می‌گیرد، از مقاله کوتاه و معروف به قلم مانفرد کوهن و سولاپول است. در این مقاله مسئله را به‌صورت ریاضی فرموله‌بندی کرده و پیامدهای اجتماعی آن‌را به‌طور عمیق بررسی کردند. این مقاله الهام بخش آزمایش معروف میلگرام است که آن هم به‌نوبه خود الهام بخش عبارت شش گام فاصله است.

معرفی تکنولوژی

کتابخانه NetworkX

NetworkX یک کتابخانه پایتون برای ایجاد، دستکاری و مطالعه ساختار، دینامیک و عملکرد شبکه های پیچیده[1] است.

- ساختارهای داده برای گراف‌ها
- بسیاری از الگوریتم های استاندارد گراف
- ساختار شبکه و اقدامات تجزیه و تحلیل
- ژنراتور برای نمودارهای کلاسیک، نمودارهای تصادفی و شبکه های مصنوعی
- گره ها می توانند "هر چیزی" باشند (به عنوان مثال، متن، تصاویر، رکوردهای XML)
- لبه ها می توانند داده های دلخواه (مانند وزن ها، سری های زمانی) را در خود نگه دارند.
- مجوز منبع باز 3 بند BSD
- به خوبی با پوشش کد بیش از 90٪ تست شده است
- مزایای اضافی پایتون شامل نمونه سازی سریع، آموزش آسان و قابلیت اجرا روی پلتفرم‌های متفاوت

کتابخانه Matplotlib

Matplotlib یک کتابخانه تصویرسازی[2] شگفت انگیز در پایتون برای نمودارهای دو بعدی از آرایه ها است. Matplotlib یک کتابخانه تجسم داده است که بر روی آرایه‌های NumPy ساخته شده و برای کار با پشته گسترده‌تر SciPy طراحی شده است. این کتابخانه توسط جان هانتر در سال 2002 معرفی شد.

یکی از بزرگترین مزایای تصویرسازی این است که به ما امکان دسترسی بصری به حجم عظیمی از داده ها را در تصاویر می دهد. Matplotlib از چندین نمودار مانند خط، نوار، پراکندگی، هیستوگرام و غیره تشکیل شده است.

---

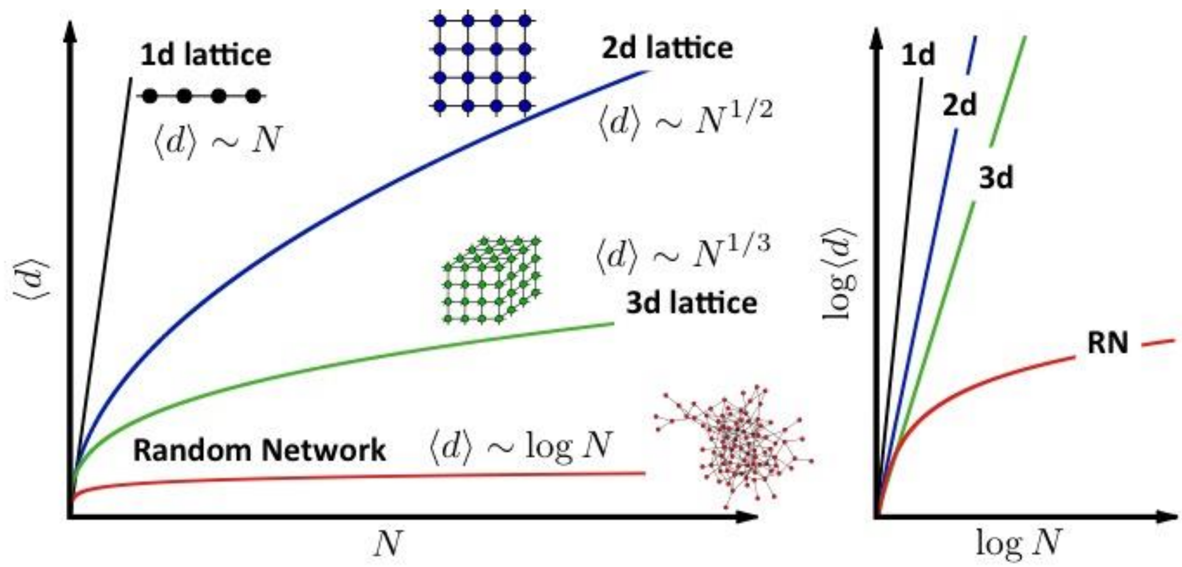
[1] Complex networks

[2] Visualization

شرح مساله

در این تمرین سعی شده با شبیه سازی کامپیوتری نشان داده شود که متوسط فاصله نودها <d> در یگ گراف تصادفی G(N,P) متناسب با log N/log<k> افزایش پیدا میکند که N تعداد نودها و <k> متوسط درجه نودهاست در حالی که متوسط فاصله در گرافهای Lattice متناسب با $N^{(1/D)}$ رشد میکند که D تعداد بُعد lattice است(D=1,2,3). شبیه سازی به این صورت انجام میگیرد که تعدادی نمونه گراف تصادفی با تعداد N نود و احتمال وجود یال p تولید شده و متوسط فاصله نودها روی آنها حساب میشود. به ازای هر N و p ثابت تعداد S گراف تصادفی بعنوان نمونه تولید شده و متوسط فاصله نودها روی آنها متوسط گیری میشود. درنهایت بررسی میکنیم که با افزایش N و افزایش p متوسط فاصله چگونه رشد میکند.

برای پیاده‌سازی توابع و انجام محاسبات این تمرین از زبان پایتون و کتابخانه‌های NetworkX و Matplotlib استفاده شده که در ادامه توضیحات مختصری پیرامون آنها آورده شده است.



شرح آزمایش

مقداردهی متغیرها

اولین مساله در این آزمایش پیدا کردن مقادیر مناسب برای متغیرهای N و P و S می‌باشد. در ادامه به بررسی چگونگی مقداردهی متغیرها و دلایل انتخاب آنها میپردازیم.

مقداردهی N

برای مقادیر N در تولید گراف تصادفی در این مساله چون فرض ما این است که رشد تابع بصورت لگاریتمی خواهد بود، میتوان N را بصورت لگاریتمی بالا برد بنا براین برای تولید گراف تصادفی از مجموعه زیر استفاده شده است:

$$N = [2, 4, 8 ,16, 32, 64, ...] = 2^n$$

همچنین برای مقادیر N در تولید گراف‌های شبکه‌ای یا توری با توجه به تعداد بُعد گراف مقادیر را بصورت $N = X^d$ در نظر میگیرم تا گراف‌های منظم تولید شوند.

مقداردهی P

برای مقداردهی متغیر P که احتمال وجود یال در گراف است از مقدار threshold (connected regime) در مدل Phase Transition و One Links Threshold استفاده میشود تا سعی شود که گراف‌های همبندی تولید شود تا بتوان فاصله گره‌ها از یکدیگر را محسابه کرد. (در گراف غیر متصل، فاصله دو راس که در دو کامپوننت جدا هستند برابر $+\infty$ در نظر گرفته میشود که در محسابات ما جای نمیگیرد)

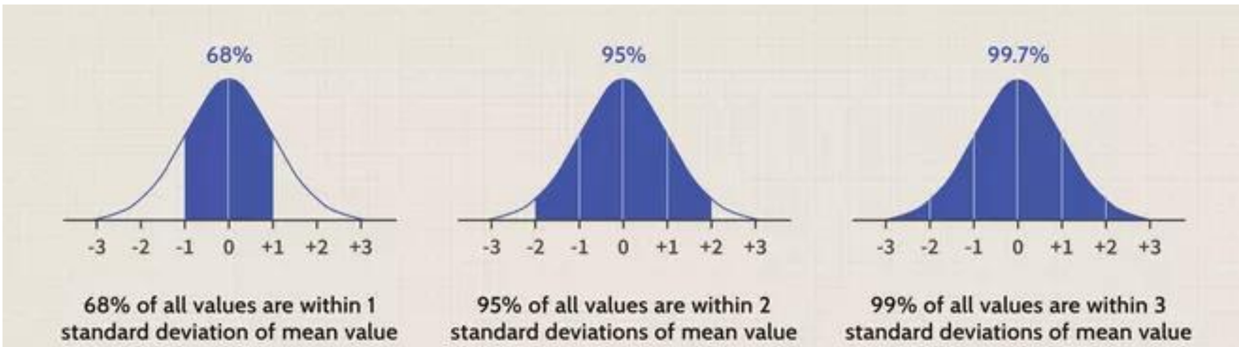
مقداردهی S

برای مقداردهی متغیر S نکته اول این است که این مقدار را برای گراف‌های مشبک برابر یک قرار میدهیم. زیرا در این گراف‌ها به علت منظم بودن تغییری در فواصل نخواهیم داشت.

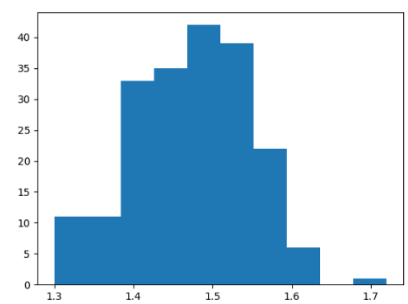
اما برای گراف تصادفی نیاز هست تا برای هر N,P ثابت تعدادی گراف نمونه تولید شده و از آنها میانگین گرفته شود.

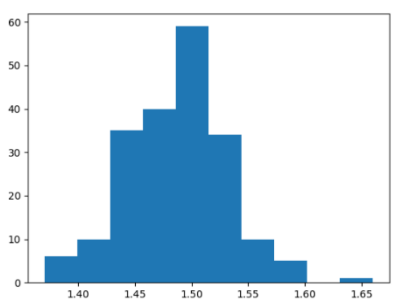
برای اینکار و بدست آوردن تعداد مناسب نمونه از Central Limit Theorem استفاده میکنیم. بر اساس این تئوری با توجه به حجم نمونه به اندازه کافی بزرگ از جمعیتی با سطح واریانس محدود، میانگین همه متغیرهای نمونه از همان جامعه تقریباً برابر با میانگین کل جامعه خواهد بود. علاوه بر این، طبق قانون اعداد بزرگ، این نمونه‌ها

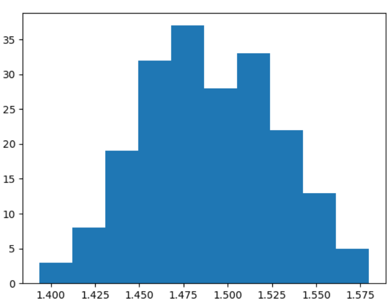
تقریباً یک توزیع نرمال دارند، و واریانس‌های آنها تقریباً برابر با واریانس جامعه‌ای است که اندازه نمونه بزرگتر می‌شود. به عنوان یک قاعده کلی، اندازه نمونه حدود 30-50 برای نگهداری CLT کافی تلقی می شود. به این معنی که توزیع میانگین نمونه نسبتاً نرمال توزیع شده است. بنابراین، هر چه تعداد نمونه های بیشتری گرفته شود، نتایج نمودار شده بیشتر شکل توزیع نرمال را به خود می گیرند



برای اثبات بیشتر این مساله توزیع میانگین فاصله یک راس از باقی رئوس را در چند گراف با جامعه ۲۰۰ عضوی مورد بررسی قرار دادیم:


نمودار 3 توزیع با ۵۰ گره


نمودار 2 توزیع با ۱۰۰ گره


نمودار 1 توزیع با ۱۵۰ گره

همانطور که از نتایج بالا مشخص است و بر طبق تئوری CLT و توزیع شدن مقادیر به صورت نرمال، ۹۵ درصد مقادیر در حوالی ۳۰ قرار دارند پس برای بدست آوردن میانگین فواصل بهتر است تعداد جامعه نمونه برابر ۳۰ باشد. در این آزمایش مقدار پیش فرض S برابر ۳۰ قرار داده شده این درحالیست که با افزایش تعداد گره، برای صرفه جویی در مصرف منابع، این مقدار کاهش داده میشود.

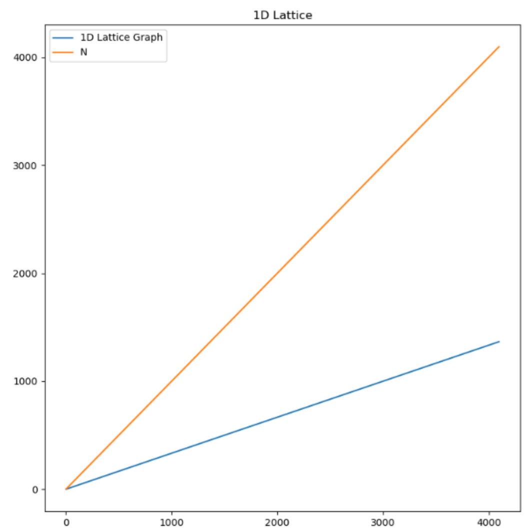
پیاده سازی

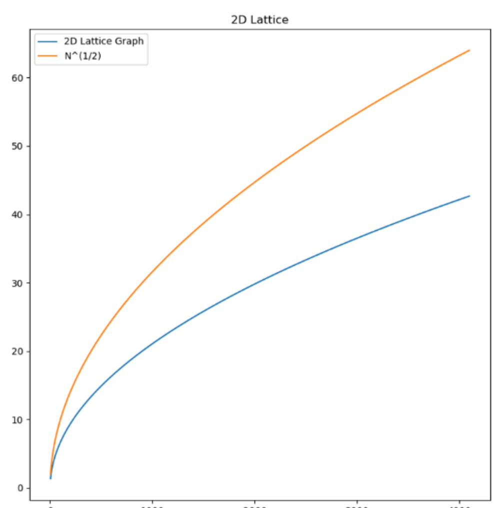
در برنامه پیاده‌سازی شده برای این آزمایش ۲ عملیات صورت میگیرد:

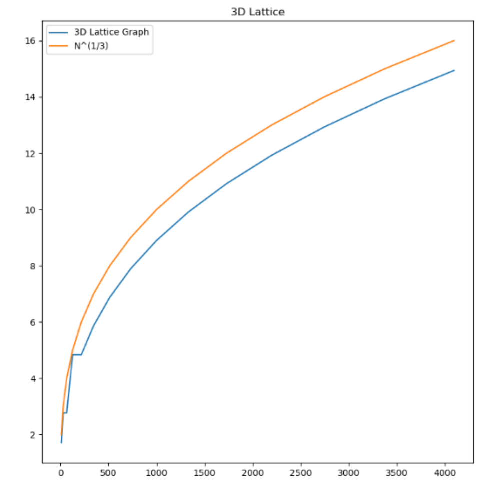
1. نمایش نمودار توزیع میانگین فاصله یک راس از رئوس دیگر در جامعه نمونه
2. نمایش نمودارهای مربوط به مقایسه میانگین فاصله در گراف‌های مذکور

مورد یک در قسمت قبل توضیح داده شد و در این قسمت به بررسی مورد دوم میپردازیم.

در ابتدا برای بررسی رشد گراف‌های مشبک، نمودار رشد میانگین فاصله در این گراف‌ها در کنار توابع متمایل آنها را در زیر مشاهده میکنید.

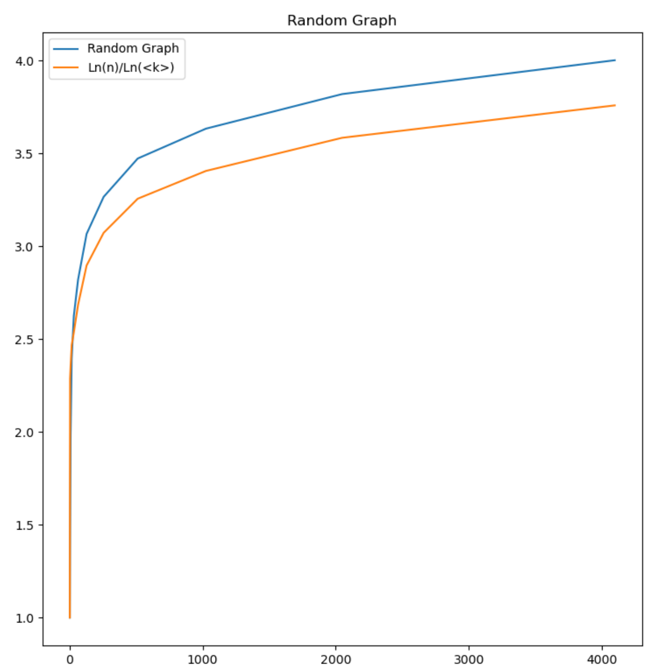




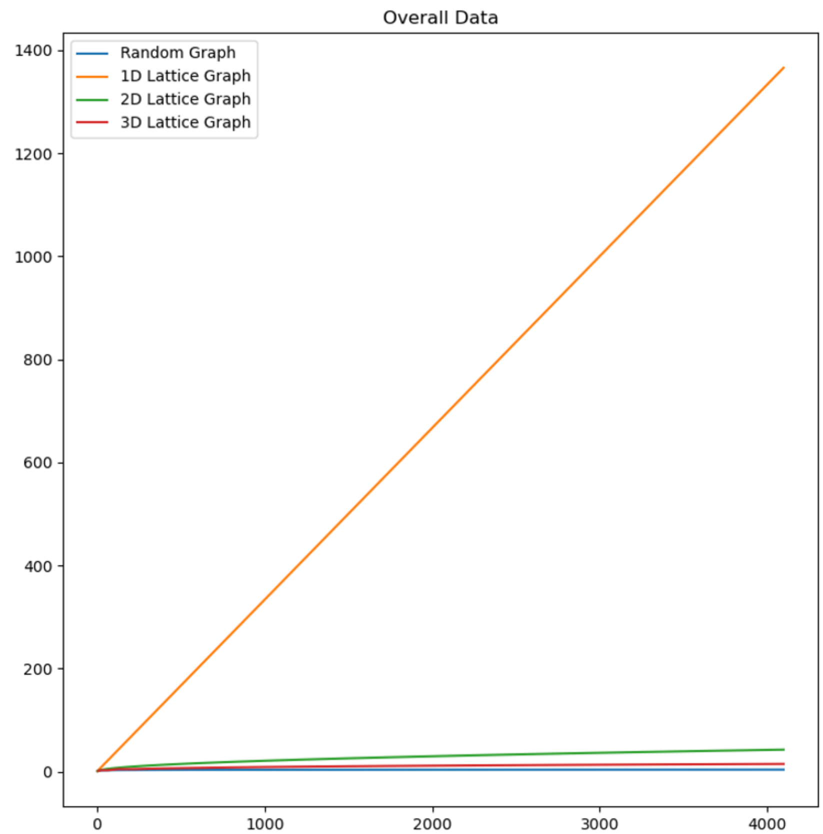


همانطور که پیش‌بینی میشد رشد نمودار گراف‌های مشبک با نسبت اندکی به سمت نمودارهای $N^{(1/d)}$ تمایل دارند. (نمودارها برابر نیستند و فقط رشد تقریبا یکسانی دارند)

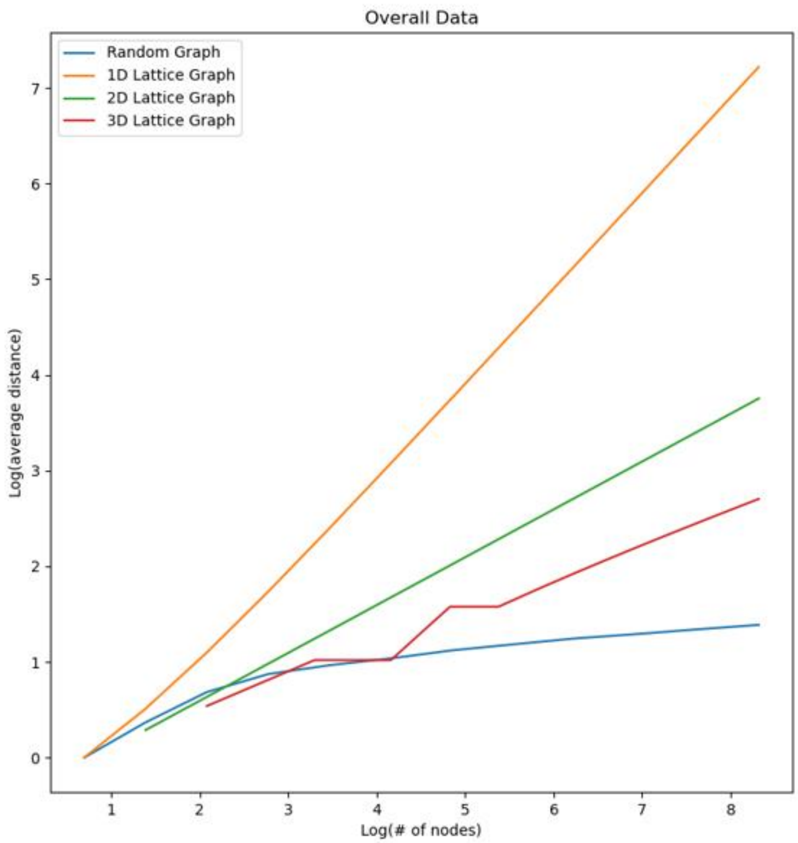
در ادامه با توجه به مقادیری که برای N,P,S انتخاب کردیم، گراف‌های تصادفی را تولید میکنیم و میانگین فاصله آنها را به روی نمودار می‌آوریم که در شکل زیر مشاهده میشود



باز هم طبق پیش‌بینی مشاهده میشود که دو تابع میانگین فاصله در گراف تصادفی با تابع logN / log<k> دارای رشد نسبی تقریبا یکسانی هستند. برای مقایسه تمامی این توابع، کنار هم در یک نمودار در شکل زیر آورده شده‌اند:



با توجه به نتایج حاصل شده، نمودار حاصل با نمودار پیش‌بینی شده تطابق دارد. البته با توجه به مقیاس، ظاهر لگاریتمی نمودارها به طور کامل مشخص نیس. برای نمایش بهتر تفاوت‌ها از ابعاد نمودار لگاریتم میگیریم.



تفاوت رشد لگاریتمی نمودارها به وضوح در این تصویر قابل مشاهده است.

نتیجەگیری

با توجه به شبیه‌سازی‌ها و آزمایش‌های انجام شده، درستی روابط و نسبت‌های ذکر شده به صورت بصری قابل مشاهده و اندازه گیری بودند. از این رو میتوان به درستی پدیده جهان کوچک پی برد. پدیده جهان کوچک ویژگی در گراف است که طول مسیر یا قطر متوسط آن از نظر لگاریتمی، به اندازه سیستم بستگی دارد و نه بصورت مستقیم. یعنی اندازه <d> به logN بستگی دارد و نه N.