بنام خدا

تکلیف دوم عملی درس گراف کاوی

زمستان ۱۴۰۴

برای این تکلیف از داده‌های موجود در گیت‌هاب درس در بخش Intermediate استفاده کنید.

بخش اول: رفتار آبشاری (Cascading Behavior)

در مدل آستانه(Linear Threshold) ، هر گره یک آستانه مشخص دارد و زمانی فعال می‌شود که بخش معینی از همسایگان فعال او (مثلاً بیش از ٪۵۰) فعال باشند. در مدل انتشار مستقل (Independent Cascade) نیز هر گره تازه فعال یک‌بار فرصت دارد تا با یک احتمال ثابت همسایگان خود را فعال کند .این دو مدل پایه‌های اصلی رفتار آبشاری در شبکه‌های اجتماعی را تشکیل می‌دهند. در این بخش باید این مدل‌ها را در پایتون پیاده‌سازی و با داده‌های واقعی (مثلاً زیرمجموعه‌ای از شبکه توییتر هیگز یا هر شبکه اجتماعی دیگری) بررسی کنید. نتایج شبیه‌سازی را تحلیل کنید و نمودارهای مناسب را رسم نمایید.

- **مسئله ۱ :**یک شبکه اجتماعی (Twitter Higgs) را در نظر بگیرید. مدل آستانه خطی را پیاده‌سازی کنید به این صورت که هر گره یک آستانه فعال‌سازی (٪۵۰) داشته باشد. با تعیین چند گره فعال اولیه، شبیه‌سازی کنید چگونه فعال‌سازی به تدریج در شبکه پخش می‌شود. تعداد گره‌های فعال را در هر گام تکرار (iteration) ثبت و نمودار آن را رسم کنید. وضعیت نهایی (تعداد نهایی فعال‌ها برای هر دسته) را گزارش دهید. بررسی کنید آیا با افزایش آستانه گره‌ها سرعت انتشار تغییر می‌کند یا خیر.

- **مسئله ۲ :**اکنون مدل انتشار مستقل را پیاده‌سازی کنید. چند گره را به عنوان منابع اولیه فعال در شبکه Twitter Higgs در نظر بگیرید. احتمال فعال‌سازی هر لبه را (٪۱۰) تنظیم کنید و شبیه‌سازی‌های تصادفی متعددی انجام دهید. اندازه خوشه نفوذ (یعنی تعداد گره‌های فعال شده در نهایت) را در هر شبیه‌سازی محاسبه و میانگین آن را گزارش کنید. نمودار رشد تعداد فعال‌ها را ترسیم نمایید. تفاوت نتایج این مدل با مدل آستانه را تحلیل کنید؛ برای نمونه چرا در مدل انتشار مستقل ممکن است انتشار آهسته‌تر ولی گسترده‌تر باشد و در مدل آستانه انتشار سریع‌تر اما محدودتر؟

بخش دوم: قوانین توانی و مدل اتصال امتیازی (Power Laws and Preferential Attachment)

مدل اتصال امتیازی باراباسی–آلبرت (Barabási–Albert) می‌تواند توزیع دم‌بلند (heavy-tailed degree distributions) ایجاد کند. به بیان دیگر، گره‌های با درجه زیاد تمایل دارند گره‌های جدید بیشتری جذب کنند (پول‌دارها پول‌دارتر می‌شوند). در این بخش با تولید و تحلیل گراف‌های متناسب با این مدل و مقایسه با داده‌های واقعی آشنا می‌شوید.

- **مسئله ۱ :**یک گراف باراباسی–آلبرت تولید کنید، با ۵۰۰۰ گره و هر گره جدید با ۳ یال به گره‌های قبلی متصل می‌شود. توزیع درجات گراف را در مقیاس لگاریتمی-لگاریتمی (log-log) رسم کنید. برازش خطی (linear fitting) برروی داده‌های لگاریتمی انجام دهید تا نمای قانون توانی (power-law exponent) را بیابید. نمای به‌دست‌آمده را گزارش

کنید. بررسی کنید که آیا توزیع درجات گراف تولید شده، در نمودار لگاریتم-لگاریتم، خطی است یا خیر (این امر نشان‌دهنده قانون توانی بودن توزیع است).

- **مسئله ۲ :** شبکه واقعی DisGeNET (شبکه ژن-بیماری) را تحلیل کنید. توزیع درجات آن را محاسبه و در هر دو مقیاس خطی و لگاریتم-لگاریتم رسم کنید. آیا این توزیع با یک قانون توانی سازگار است یا خیر. نمای استخراج‌شده را با نمای گراف باراباسی-آلبرت مقایسه نمایید و درباره علت شباهت یا تفاوت‌ها به صورت خلاصه بحث کنید.

بخش سوم: قدم زدن تصادفی (Random Walk)

در این قسمت هدف پیاده‌سازی الگوریتم PageRank و Random Walk with Restart (RWR) است.

- **مسئله ۱ :** شبکه پروتئین-پروتئین STRING PPI را درنظر بگیرید. با استفاده از NetworkX یا کتابخانه مشابه، امتیاز PageRank را برای هر نود محاسبه کنید. پنج نود با بالاترین امتیاز PageRank را گزارش کنید و بررسی کنید آیا این نودها با نودهای با درجه بالاتر یا نقش‌های خاصی در شبکه متناظر هستند یا خیر.
- **مسئله ۲ :** اکنون قدم‌زدن تصادفی با شروع مجدد را پیاده‌سازی کنید. یک نود هدف (seed) در نظر بگیرید، (یک پروتئین کلیدی در شبکه STRING PPI). با احتمال بازگشت مشخص (۱۵٪) از آن گره شروع کنید و RWR را اجرا کنید تا امتیاز شباهت هر نود به نود هدف به‌دست آید. پنج نود با بیشترین امتیاز شباهت (شبیه‌ترین‌ها به نود هدف) را گزارش کنید و تحلیل کنید که آیا این نودها به نوعی به نود هدف مرتبط هستند یا خیر. با تغییر درصد احتمال بازگشت، تحلیل کنید رفتار الگوریتم چه تغییری می‌کند.

بخش چهارم: قدرت پیوندهای ضعیف و ساختار جوامع (Strength of Weak Ties and Community Structure)

بر اساس تئوری پیوندهای ضعیف گرانواتر (Granovetter’s theory)، پیوندهای ضعیف معمولاً نقش پل بین خوشه‌ها یا جوامع مختلف در شبکه را دارند. به عبارت دیگر، پیوندهای ضعیف ارتباط ما را با شبکه‌های خارج از دایره نزدیک اطراف خود فراهم می‌کنند.

- **مسئله ۱:** در شبکه‌ی DisGeNET شاخصی برای سنجش قدرت پیوند تعریف کنید، مانند نسبت همسایگان مشترک بین دو گره (Jaccard) یا وزن یال. یال‌هایی که کمترین مقدار این شاخص را دارند (پیوندهای ضعیف) را شناسایی کنید. بررسی کنید آیا این یال‌ها اغلب بین جوامع مجزایی قرار دارند یا نه. برای این کار ابتدا خوشه‌های جامعه را (با هر روش ساده‌ای) استخراج کرده و سپس یال‌های بین جوامع را با یال‌های درون‌جامعه‌ای مقایسه کنید.
- **مسئله ۲:** تاثیر حذف پیوندهای ضعیف را بررسی کنید. همان شبکه را در نظر بگیرید و پیوندهای شناسایی‌شده در قسمت قبل را حذف کنید. سپس ساختار جوامع حاصل از گراف جدید را با ساختار اولیه مقایسه کنید. معیارهایی مانند قطر خوشه‌ها یا میانگین طول کوتاه‌ترین مسیر (یا هر معیار دیگری که برای تحلیل مناسب است) را محاسبه کرده و تحلیل کنید که حذف پیوندهای ضعیف چگونه به تفکیک جوامع کمک کرده است.

بخش پنجم تشخیص جوامع (Community detection) – الگوریتمGirvan–Newman

الگوریتم Girvan–Newman یک روش سلسله‌مراتبی برای کشف جامعه‌ها در شبکه است. این الگوریتم با حذف متوالی یال‌هایی که بالاترین Edge-betweenness را دارند، شبکه را تفکیک می‌کند. به عبارت دیگر، یال‌هایی که بیشترین مسیرهای کوتاه بین جوامع را در خود دارند شناسایی شده و حذف می‌شوند تا اجزای جداگانه (جامعه‌ها) مشخص شوند.

- مسئله : یک شبکه کوچک شناخته‌شده مانند Karate Club را در نظر بگیرید. مراحل الگوریتم Girvan–Newman را اجرا نمایید (مراحل الگوریتم را بدون استفاده از کتابخانه‌های موجود در پایتون پیاده‌سازی کنید). یال‌هایی که در هر گام دارای بالاترین betweenness هستند را حذف کنید و پس از هر حذف، مؤلفه‌های اتصال‌یافته جدید (جامعه‌ها) را مشخص کنید. نموداری به صورت درخت دودویی (dendrogram) رسم کنید که جداسازی جوامع در هر مرحله را نشان دهد. در هر مرحله ماژولاریتی در نهایت، تقسیم نهایی به چند جامعه را با معیارهایی مانند ماژولاریتی (modularity) ارزیابی و مقایسه کنید.

فایل خروجی، یک فایل نوتبوک با دسته‌بندی مناسب، فرمت مناسب Markdown برای پرسش‌های تحلیل و توصیف های مورد نیاز باشد.

توجه داشته باشید، در فایل نوتبوک آپلود شده، خروجی‌ها پاک نشده باشند.

در تمامی بخش‌ها (بجز بخش پنجم)، می‌توانید از کتابخانه‌ها و الگوریتم‌های موجود در پایتون استفاده کنید.