دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

گزارش تکلیف دوم درس الگوریتم‌های شبکه‌های پیچیده

دانشجویان:

سید احمد نقوی نوزاد

ش-د: ۹٤۱۳۱۰٦۰

افشین رودگر

ش-د: ۹٤۱۳۱۰٤٤

استاد:

دکتر امیرحائری

بهار ۹۵

جواب سوال ۱

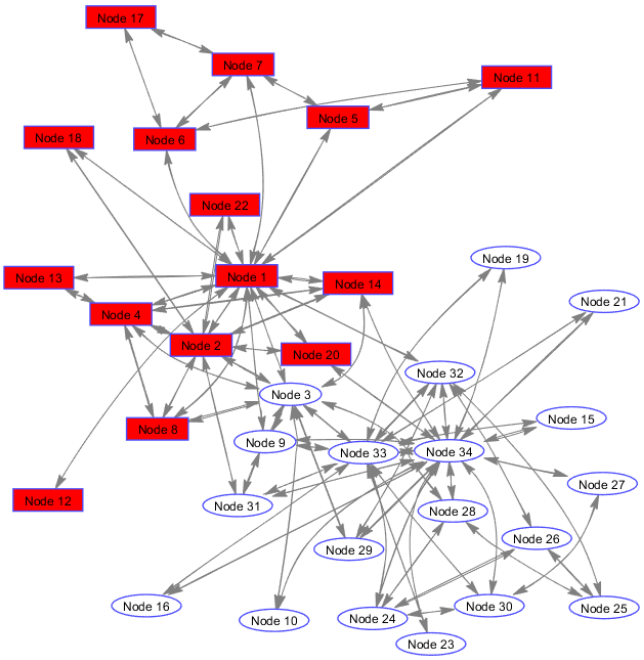
کاربرد رتبه‌صفحه (pageRank) در خوشه‌بندی و پیش‌بینی لینک

الف) کاربرد PageRank در خوشه‌بندی

در این قسمت برای خوشه‌بندی با استفاده از رتبه‌صفحه، از مفهوم **Personalized PageRank** استفاده می‌نمائیم که در واقع حالت خاصی از همان مفهوم **Topic-Sensitive PageRank** می‌باشد. در روش **Topic-Sensitive PageRank** با استفاده از معادله‌ی زیر بردار رتبه‌صفحه را که هر درایه‌ی آن مقدار رتبه‌صفحه را برای گره‌ی مربوطه نشان می‌دهد، محاسبه می‌نمائیم:

$$v' = \beta M v + (1-\beta)e_s/|S| \qquad (1)$$

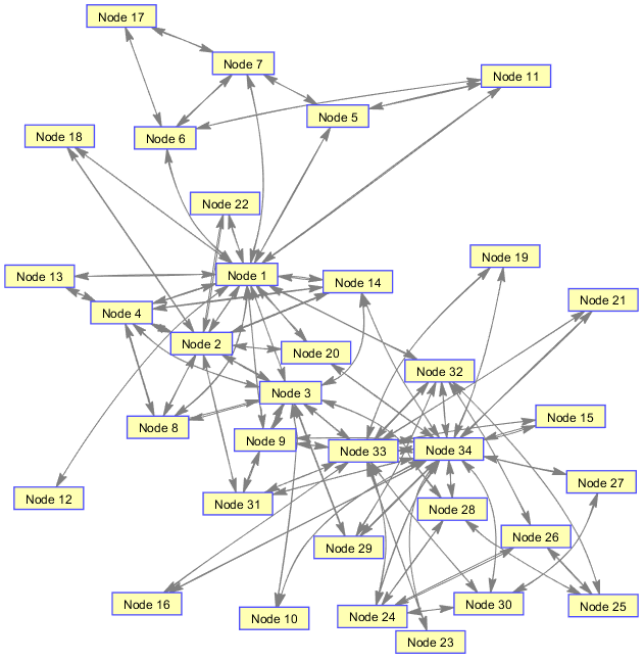
که در آن تنها تفاوتی که با فرمول معمول محاسبه‌ی رتبه‌صفحه مشاهده می‌شود این است که مجموعه‌ی **teleport** (در این‌جا S) تنها شامل تعدادی گره‌ی خاص (**Topic-Sensitive**) می‌شود نه شامل همه‌ی گره‌های گراف، و e_s نیز برداری است که در آن تنها درایه‌های مربوط به تاپیک خاص برابر یک بوده و مابقی درایه‌ها برابر صفر می‌باشند. حال در روش **Personalized PageRank** این مجموعه‌ی **teleport** ، در هر مرحله تنها شامل یک گره‌ی خاص خواهد بود و به عبارتی می‌بایست به تعداد گره‌های شبکه، هر بار یک بردار رتبه‌صفحه را محاسبه نمائیم (که در این بردار ارزش رتبه‌صفحه‌ی مربوط به گره‌ی مربوطه و همسایگان آن گره از بقیه بیشتر خواهد بود) و در نهایت از کنار هم قراردادن این بردارها یک ماتریس ۳٤*۳٤ حاصل خواهد شد، که می‌توان هر ردیف را نماد یک نمونه‌داده (**data sample**) و هر ستون را یک بعد (**feature**) در نظر گرفت و در نهایت با اعمال الگوریتم **K-Means** بر روی این ماتریس و البته اعمال پارامترهای مناسب (از جمله استفاده از معیار فاصله‌ی **cosine** به جای **euclidean**)، عمل خوشه‌بندی را به تعداد خوشه‌های دلخواه (در این مسئله ۲ خوشه) انجام داد. نتایج حاصله به شرح ذیل می‌باشد:



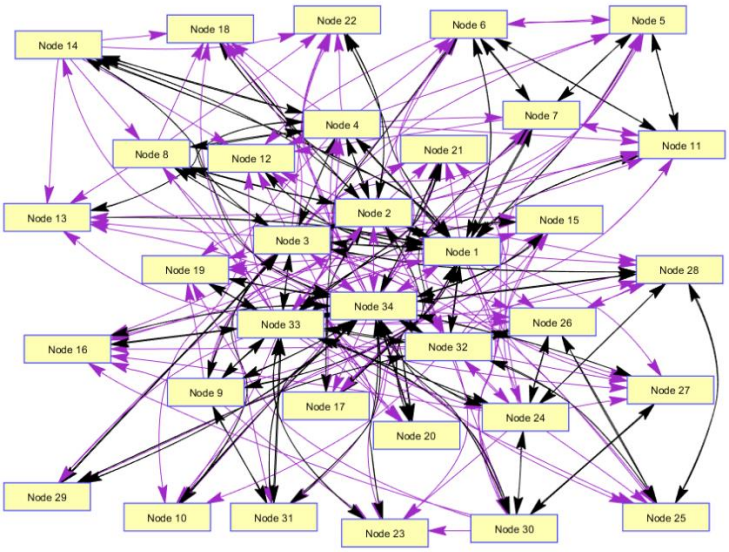
شکل ۱- خوشه‌بندی با استفاده از مفهوم رتبه‌صفحه

ب) کاربرد PageRank در پیش‌بینی لینک

در اینجا نیز در ادامه‌ی قسمت قبلی، تنها کافی است تا ماتریس رتبه‌ی صفحه‌ی ۳٤*۳٤ را مورد استفاده قرار دهیم. به این ترتیب که ابتدا درایه‌های روی قطر اصلی را برابر صفر قرار داده و سپس برای تعیین مقدار حد آستانه‌ی مناسب، از مجموعه‌ی لینک‌های موجود بهره برده و آن‌ها را به دو مجموعه‌ی آموزشی و آزمایشی تقسیم می‌نمائیم. سپس با توجه به کمینه و بیشینه‌ی مجموعه‌ی آموزشی تعدادی حد آستانه تعریف نموده و مقداری را انتخاب می‌نمائیم که به ازای آن صحت تشخیص لینک برای داده‌های آزمایشی بیشینه باشد. در نهایت از مقدار حد آستانه‌ی بهینه برای پیش‌بینی لینک در مورد لینک‌های ناموجود استفاده می‌نمائیم. نتایج حاصله به قرار زیر می‌باشند:



شکل ۲- نمودار گراف اصلی



شکل ۳- نمودار گراف اصلی به همراه لینک‌های پیش‌بینی‌شده که با رنگ بنفش نشان داده شده‌اند

در نهایت از مزایای روش نامبرده برای خوشه‌بندی می‌توان به سهولت پیاده‌سازی آن برای انواع گراف‌ها اسم برده و نیز از معایب آن می‌توان گفت که چون نیاز هست به ازای هر گره یک بردار رتبه‌صفحه محاسبه نمائیم، لذا در مورد گراف‌های با تعداد بسیار زیاد گره، این مسئله شدیدا از لحاظ محاسباتی و زمانی سنگین خواهد بود.

جواب سوال ۲

بیشینه‌سازی تأثیر در گراف‌های شبکه‌های اجتماعی

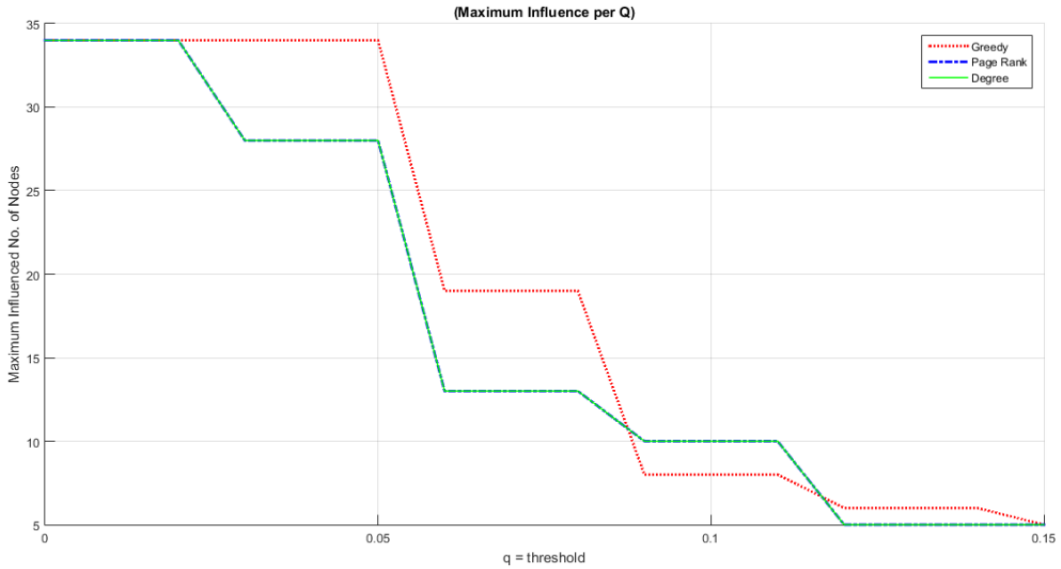
الف) یافتن K فرد با بیشترین تأثیر در شبکه ($K \leq 5$)

برای حل این مساله سه روش را مورد بررسی قرار خواهیم داد، که شامل روشی بر پایه درجه‌ی خروجی، روشی بر پایه‌ی رتبه‌صفحه و نیز یک روش حریصانه خواهد بود. روش‌های مبتنی بر درجه‌ی خروجی و رتبه‌صفحه، شباهت بسیاری به یکدیگر داشته و در نهایت بهترین گره‌ها را برای بیشینه‌سازی تاثیر، تا حد زیادی مشابه یکدیگر پیشنهاد خواهند داد. البته هیچ کدام از الگوریتم‌های ارائه شده راه‌حل بهینه را معرفی نخواهند کرد.

روش مبتنی بر درجه‌ی خروجی: در این روش با توجه به این موضوع که هر گره با حداکثر درجه‌ی خروجی، بیشترین تاثیر را روی همسایگان خود در مقایسه با سایر گره‌ها خواهد داشت، تاکید شده است. الگوریتم در ابتدا درجات خروجی هر کدام از گره‌ها را محاسبه کرده و با مرتب کردن آنها بصورت نزولی و انتخاب K گره‌ی اول، مؤثرترین گره‌ها را معرفی خواهد کرد.

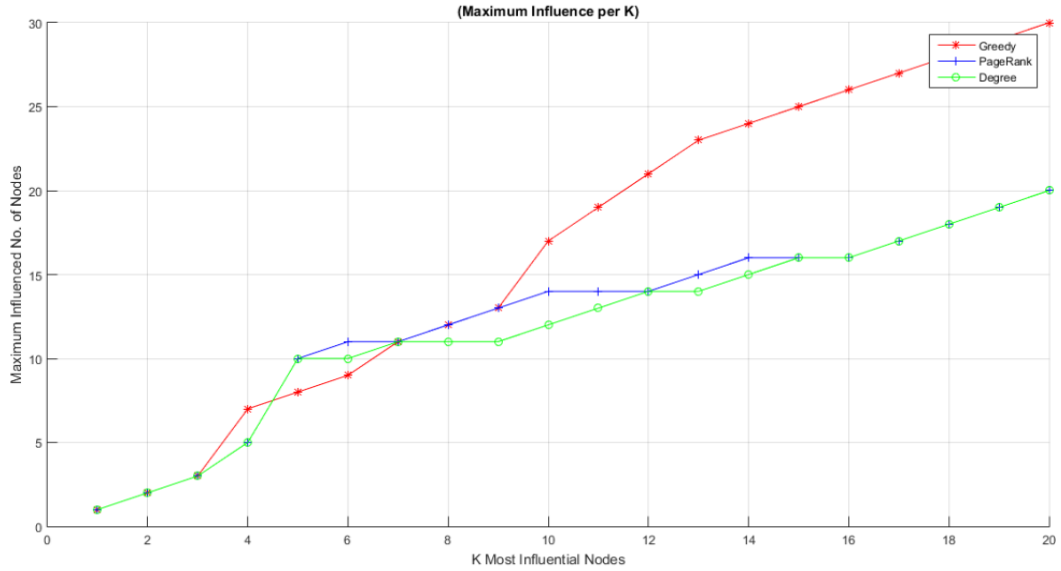
روش مبتنی بر رتبه‌ی صفحه: این روش نیز همانند روش مبتنی بر درجه‌ی خروجی، با مرتب کردن نزولی رتبه‌صفحه‌ی هر کدام از گره‌ها و انتخاب K گره‌ی اول، مؤثرترین گره‌ها را انتخاب می‌کند.

روش حریصانه: در این روش سعی در الگوبرداری از الگوریتم حریصانه مورد استفاده در مساله‌ی Independent Cascade Model و Linear Threshold داریم. ابتدا مجموعه‌ی خالی از گره‌ها را در نظر گرفته و در هر چرخه از الگوریتم بهترین گره‌ای که تاثیر مجموعه‌ی موجود را تا این لحظه افزایش می‌دهد، به آن اضافه خواهیم کرد. مجموعه‌ی گره‌ها در پایان، موثرترین گره‌ها از دید الگوریتم حریصانه خواهند بود.



شکل ۴- شمار حداکثر گره‌های متأثر با توجه به مقادیر مختلف حد آستانه‌ی q

در تصویر بالا حداکثر انتشار پیش‌آمده حاصل از هر کدام از الگوریتم‌های معرفی شده، به ازای qهای مختلف و $K = 5$ را مشاهده می‌کنید. همانطور که در این تصویر مشاهده می‌شود، نتایج بدست آمده از الگوریتم‌های مبتنی بر درجه‌ی خروجی و الگوریتم رتبه‌صفحه با توجه به این نمودار، یکی بوده و البته می‌توان کاهش ملایم شمار بیشینه‌ی گره‌های متأثر در شبکه نسبت به مقدار حد آستانه، ناشی از الگوریتم حریصانه را در مقایسه با دو الگوریتم دیگر مشاهده کرد، که این مطلب خود یک حُسن به حساب آمده و نشان از آن دارد که عملکرد الگوریتم حریصانه با افزایش مقدار حد آستانه به طور ناگهانی افت نکرده و تأثیرگذاری مجموعه گره‌های اولیه همچنان قابل توجه است. اما با توجه به تصویری که در ادامه می‌آید مشخص می‌شود که به ازای برخی از مقادیر K، الگوریتم مبتنی بر رتبه‌ی صفحه، نتایج نسبتاً بهتری را در مقایسه با الگوریتم مبتنی بر درجه‌ی خروجی ارائه می‌دهد. این در حالی است که نتیجه‌ی بدست آمده برای الگوریتم حریصانه در اغلب موارد بهتر از دیگر الگوریتم‌ها بوده است.



شکل ۵- شمار حداکثر گره‌های متأثر با توجه به مقادیر مختلف K و q=.1

در تصویر بالا نمودار حداکثر گره‌های متاثر حاصل از هر کدام از الگوریتم‌های پیشنهادی را به ازای Kهای مختلف و $q = 0.1$ مشاهده می‌شود. گرچه با ازای Kهای پایین، همه‌ی الگوریتم‌ها تقریبا در یک سطح عمل می‌کنند، اما با بالا رفتن مقدار K ، الگوریتم حریصانه شایستگی خود را به نمایش می‌گذارد.

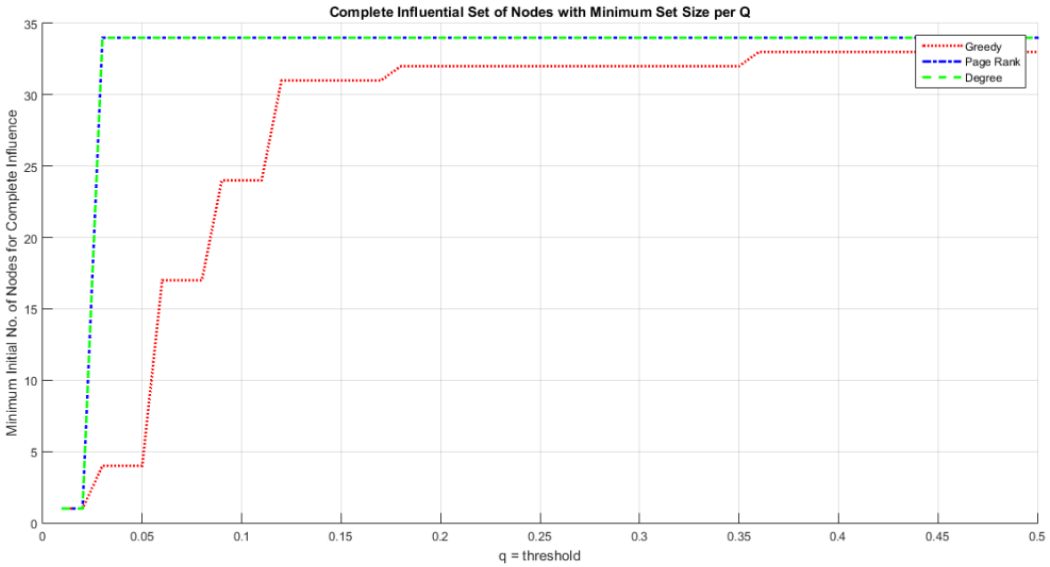
```
Average time spent on each algorithm (seconds) :

Greedy influence maximization: 0.0049136
PageRank based influence maximization: 0.00016035
Degree based influence maximization: 1.656e-05
```

در بالا زمان متوسط استفاده‌ی هر کدام از الگوریتم‌ها از CPU را مشاهده می‌کنید. پیچیدگی زمانی الگوریتم حریصانه بیشتر از دو الگوریتم دیگر می‌باشد، اما نتایج حاصل از آن بهتر است. لذا حین استفاده از این الگوریتم در کنار جواب بهینه‌ی آن باید پیچیدگی زمانی آن را نیز مد نظر قرار داد. از آنجایی که زمان مصرفی الگوریتم مبتنی بر درجه‌ی خروجی تقریبا یک دهم الگوریتم مبتنی بر رتبه‌ی صفحه می‌باشد و نتایج آن دو نیز تقریبا مشابه یکدیگر است،

می‌توان اظهار کرد که استفاده از الگوریتم مبتنی بر رتبه‌ی صفحه توجیهی نداشته و در هر حالت بهتر آن است تا از الگوریتم مبتنی بر درجه‌ی خروجی استفاده کنیم.

ب) یافتن مجموعه گره‌های اولیه با حداقل عضو که موجب انتشار کامل خواهند شد

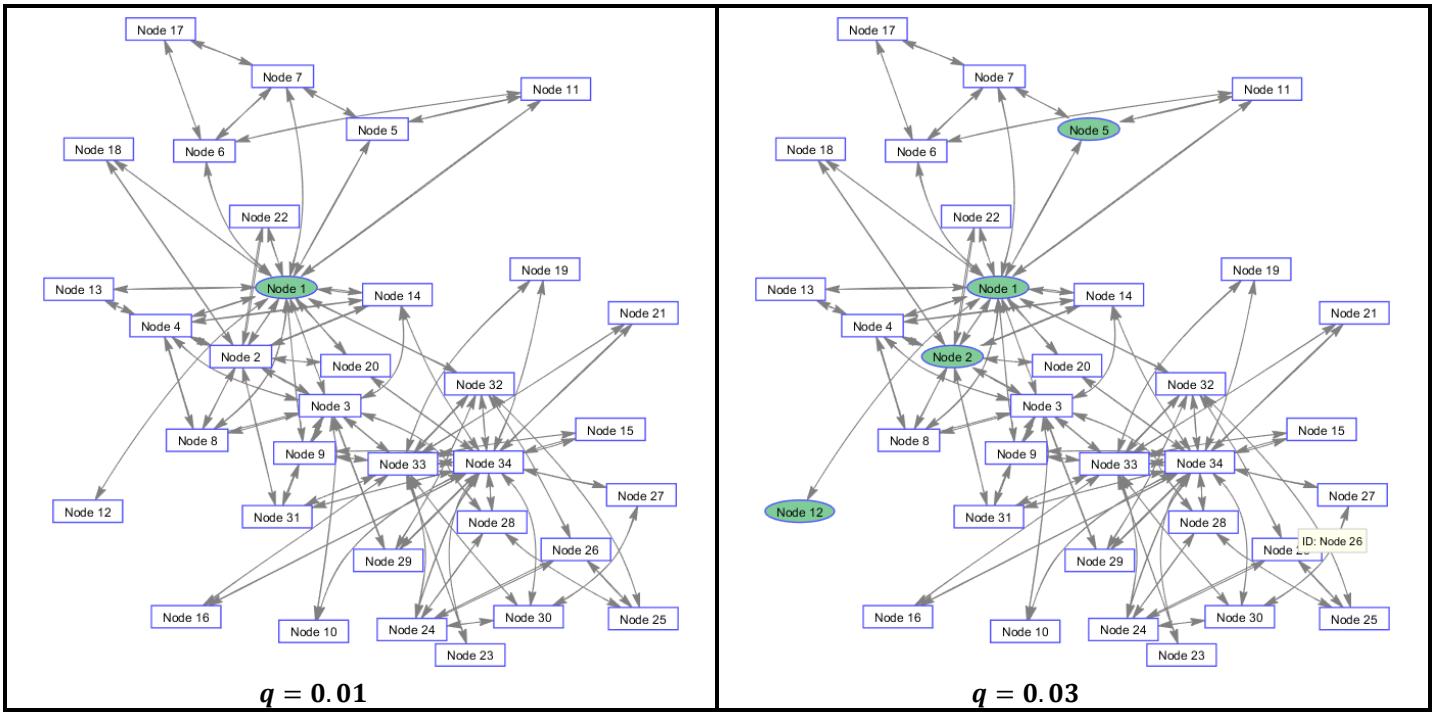
برای یافتن مجموعه‌ی گره‌های اولیه با حداقل عضو که انتشار کامل را موجب خواهد شد، از الگوریتم‌هایی که در بالا معرفی شد با مقداری تغییر، استفاده خواهیم کرد. در هر کدام از الگوریتم‌ها از روش‌های ذکرشده برای انتخاب بهترین گره استفاده کرده و اضافه کردن گره را تا جایی که به مجموعه‌ی مورد نظر برسیم، ادامه خواهیم داد.



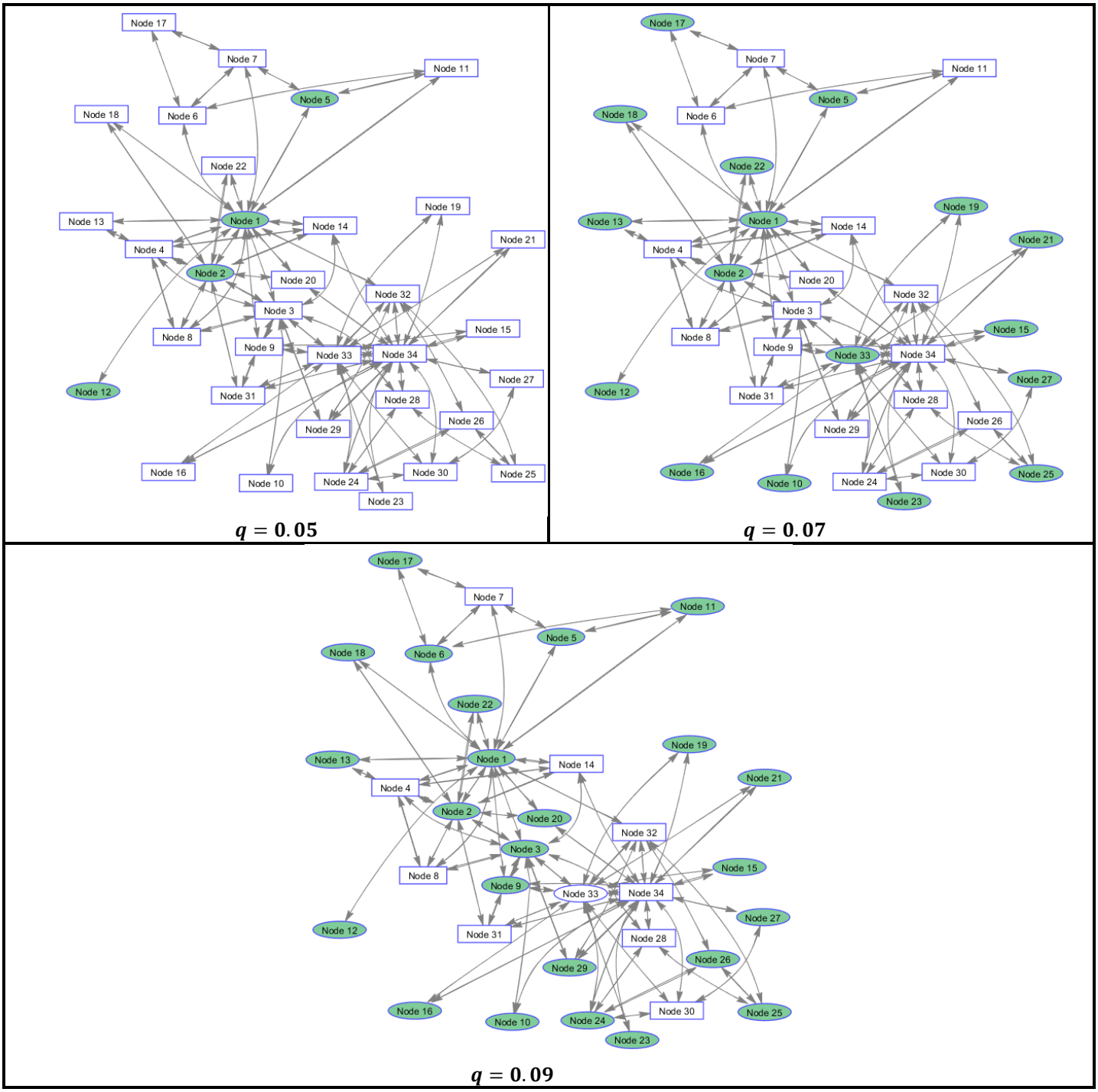
شکل ۶- اندازه‌ی مجموعه‌ی کمینه‌ی لازم برای انتشار کامل در شبکه به ازای حد آستانه‌های مختلفِ q

نمودار بالا تعداد اعضای مجموعه‌ی کمینه‌ی لازم برای انتشار کامل، حاصل از هر کدام از الگوریتم‌های ذکر شده را نشان می‌دهد. همانطور که مشاهده می‌شود، الگوریتم مبتنی بر درجه‌ی خروجی و الگوریتم مبتنی بر رتبه‌صفحه هر دو کاملا مشابه یکدیگر عمل کرده و با بالا رفتن مقدار q به سرعت تعداد اعضای این مجموعه را افزایش می‌دهند، به این معنی که تقریبا تمامی گره‌های گراف را جهت انتشار کامل نیاز خواهند داشت، و البته این مسئله نشانه‌ی ضعف آن‌ها در امر انتشار کامل می‌باشد. در حالی که الگوریتم حریصانه این‌گونه نبوده و با یک شیب فزاینده‌ی نسبتاً نرمال، با بالارفتن مقدار حد آستانه، اندازه‌ی مجموعه‌ی کمینه‌ی نامبرده جهت انتشار کامل را افزایش می‌دهد و البته که این مسئله خود یک نقطه‌ی قوت در مقایسه با دو الگوریتم دیگر می‌باشد.

در ادامه تصویر مجموعه‌ی کمینه‌ی گره‌های انتخابی برای انتشار کامل توسط الگوریتم حریصانه را به ازای qهای متفاوت مشاهده می‌کنید.



$q = 0.01$ $q = 0.03$



شکل ۷- مجموعه‌ی کمینه‌ی گره‌های انتخابی برای انتشار کامل توسط الگوریتم حریصانه را به ازای qهای مختلف

جواب سوال ۳

پیش‌بینی لینک با استفاده از روش‌های مبتنی بر بیشینه‌سازی تأثیر

در این قسمت برای امر پیش‌بینی لینک با استفاده از روش‌های بیشینه‌سازی تأثیر، از روشی موسوم به **Stochastic Block Model** استفاده می‌نمائیم. در این روش، گره‌های گراف را به تعداد دفعات زیاد به زیرمجموعه‌های متعدد افراز می‌نمائیم به گونه‌ای که اشتراک این مجموعه‌ها تهی می‌باشد، و احتمال اتصال دو گره در این وضعیت تنها به گروه‌هایی که گره‌ها به آن‌ها تعلق دارند وابسته است. حال اگر یک افراز از مجموعه گره‌ها را با $\mathcal{M}$ نشان دهیم به گونه‌ای که هر گره تنها به یک گروه تعلق دارد، و نیز احتمال اتصال دو گره که به ترتیب به گروه‌های α و β تعلق دارند را با $Q_{\alpha\beta}$ نمایش دهیم (بدیهی است که $Q_{\alpha\alpha}$ احتمال اتصال دو گره‌ی متعلق به یک گروه خاص را نمایش می‌دهد)، می‌توان درست‌نمائی ساختار افراز نامبرده را به صورت زیر نمایش داد:

$$\mathcal{L}(A|\mathcal{M}) = \prod_{\alpha \le \beta} Q_{\alpha\beta}^{l_{\alpha\beta}} (1 - Q_{\alpha\beta})^{r_{\alpha\beta} - l_{\alpha\beta}} \qquad (2)$$

به طوری که $l_{\alpha\beta}$، معرف تعداد یال‌های فعلی موجود مابین دو گروه α و β بوده و $r_{\alpha\beta}$ نیز بیانگر حداکثر تعداد یال‌های ممکن میان دو گروه (و به عبارتی شمار جفت گره‌هایی که یک گره متعلق به گروه α و دیگری متعلق به گروه β باشد) مذکور می‌باشد. حال در این‌جا اگر بخواهیم مقدار درست‌نمائی مزبور بیشینه گردد می‌بایست $Q_{\alpha\beta}$ را به صورت زیر مقداردهی نمائیم:

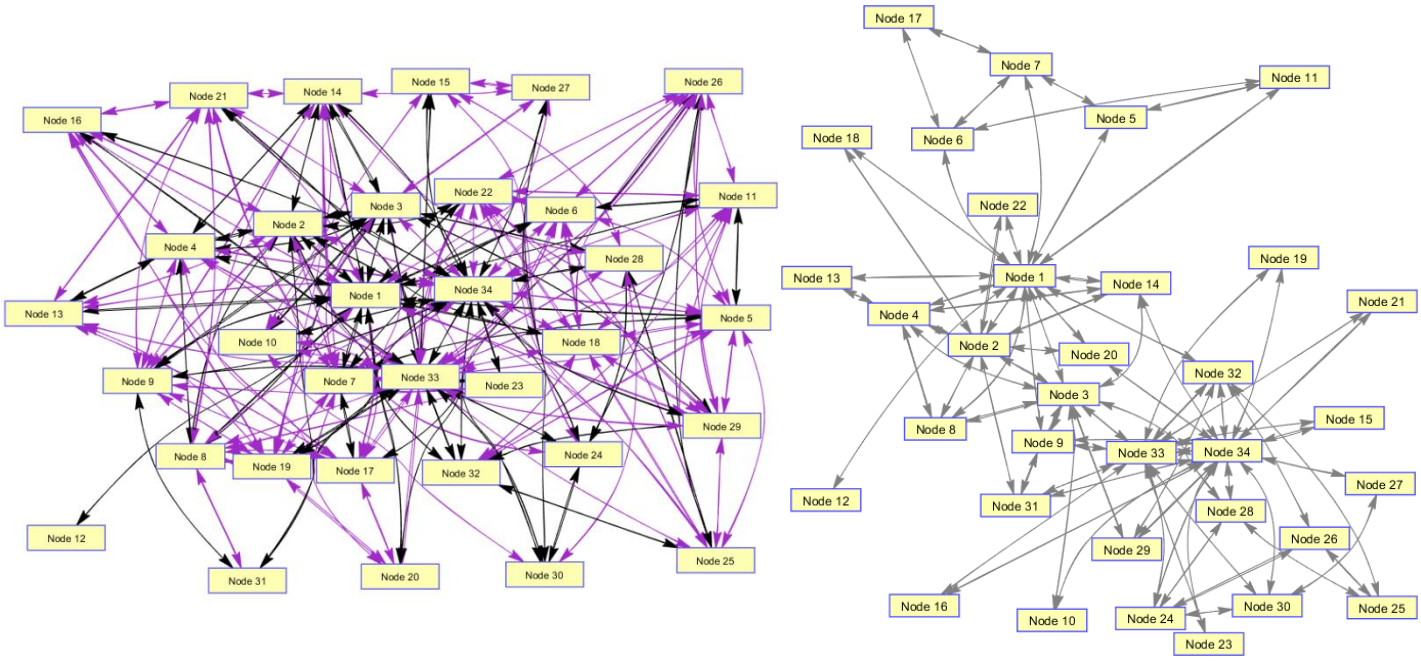
$$Q^*_{\alpha\beta} = \frac{l_{\alpha\beta}}{r_{\alpha\beta}} \qquad (3)$$

حال در مورد شبکه‌ی مورد بحث در این تکلیف نیز، در ابتدا گره‌های گراف را به تعداد دفعات زیاد به تعداد زیرمجموعه‌های متعدد افراز می‌نمائیم (که هر دوی این مقادیر از کاربر دریافت می‌شوند و در مورد این مسئله ما با مقادیر تعداد تکرار برابر ۱۰۰۰ و حداکثر تعداد زیرمجموعه‌های حاصل از افراز برابر ۱۲کار می‌کنیم). سپس یک ماتریس سه‌بعدی را با نام ***qMat*** درست می‌کنیم، به طوری که دو بُعد اول آن برابر ابعاد ماتریس مجاورت و بُعد سوم نیز برابر تعداد دفعات نمونه‌برداری (افرازبندی) می‌باشد. درایه‌ی ***ij*** این ماتریس برابر احتمال اتصال گره‌ی ***i***–اُم به گره‌ی ***j***–اُم بوده و با توجه به مقدار $Q^*_{\alpha\beta}$ که در بالا صحبت شد، به دست می‌آید، و به عبارتی برای تمامی گره‌های ***i*** و ***j*** که به ترتیب متعلق به گروه‌های α و β می‌باشند، یک مقدار یکسان در ماتریس ***qMat*** ثبت می‌گردد. علاوه بر ماتریس ***qMat*** نیاز به مقادیر درست‌نمائی برای هر کدام از این افرازها نیز احساس می‌شود، که در این‌جا این مقادیر را در برداری با نام ***likeHoodVec*** ذخیره نموده و با استفاده از معادله‌ی (۲) آن‌ها را به ازای هر افراز به دست می‌آوریم.

حال اگر مجموعه‌ی تمامی افرازهای انجام‌شده را با Ω نشان دهیم، با استفاده از تئوری ***Bayes***، می‌توان شدت اتکاءپذیری (***reliability***) هر لینک را با استفاده از معادله‌ی زیر محاسبه نمود:

$$R_{xy} = \mathcal{L}(A_{xy} = 1|A) = \frac{\int_{\Omega} \mathcal{L}(A_{xy} = 1|\mathcal{M})\mathcal{L}(A|\mathcal{M})p(\mathcal{M})d\mathcal{M}}{\int_{\Omega} \mathcal{L}(A|\mathcal{M}')p(\mathcal{M}')d\mathcal{M}'} \qquad (4)$$

در نهایت پس از حصول ماتریس R، درایه‌های روی قطر اصلی آن را صفر کرده و سپس مانند سوال اول، با تقسیم مجموعه لینک‌های موجود به دو مجموعه‌ی آموزشی و آزمایشی، مقدار بهینه برای حد آستانه را به دست آورده و سرانجام در مورد لینک‌های ناموجود تصمیم‌گیری می‌نمائیم. نتایج حاصله به قرار زیر می‌باشند:

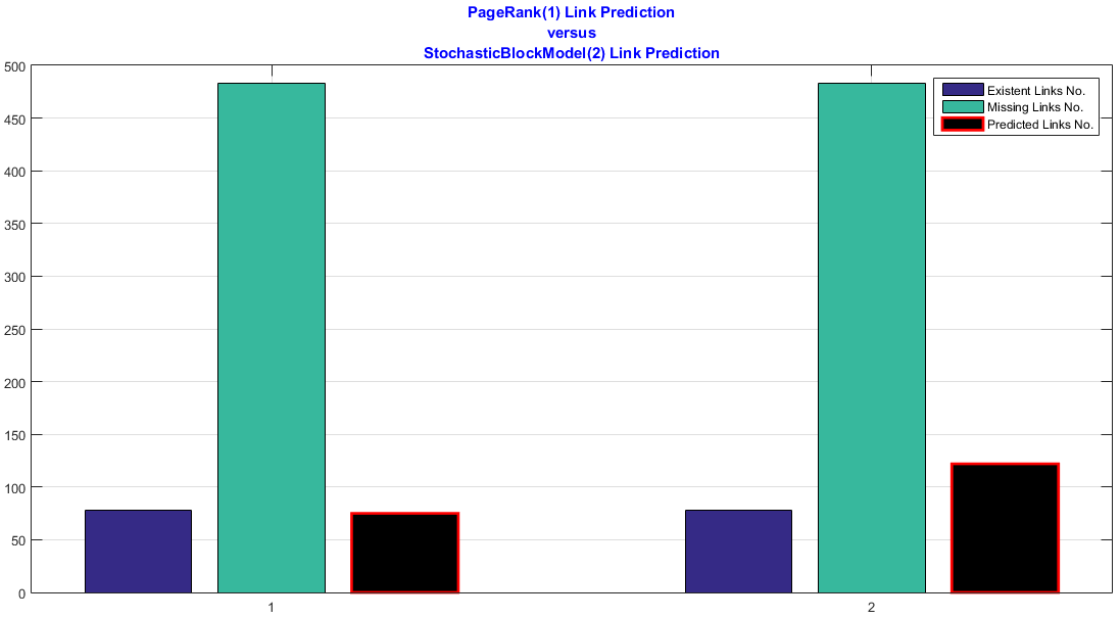


شکل ۸- نمودار گراف اصلی

شکل ۹- نمودار گراف اصلی به همراه لینک‌های پیش‌بینی‌شده که با رنگ بنفش نشان داده شده‌اند

در پایان نتایج حاصله از الگوریتم **Stochastic Block Model** مورد استفاده در این سوال جهت پیش‌بینی لینک را با نتایج حاصل از روش مبتنی بر **PageRank** مورد استفاده در سوال اول مقایسه می‌نمائیم که جزئیات آن در ادامه می‌آید:

```
1) PageRank Predicted Links No.:        75
2) StochasticBlockModel Predicted Links No.:        122
Max predicted algorithm:        "2) Stochastic Block Model"
Prediction coincidence no.:     57
```



شکل ۱۰- مقایسه‌ی روش مبتنی بر PageRank با روش StochasticBlockModel

همانطور که از نمودار bar-graph بالا و نتایج خروجی ماقبل آن قابل مشاهده است، شمار لینک‌های پیش‌بینی‌شده توسط هر دو الگوریتم تقریبا با یکدیگر برابری نموده و البته الگوریتم StochasticBlockModel در این‌جا موفق به پیش‌بینی اتصالات بیشتری گشته است. شمار تداخل در پیش‌بینی لینک برای دو الگوریتم نامبرده نیز رقم قابل توجهی می‌باشد که می‌توان آن را نشانی از صحت عملکرد دو الگوریتم دانست.