نشریه علمی "پدافند الکترونیکی و سایبری"

سال هشتم، شماره ۳، پاییز ۱۳۹۹، ص ۱۱-۱

علمی– پژوهشی

شناسایی گره‌های مؤثر در شبکه‌های اجتماعی با ترکیب روش‌های مرکزیت و فعالیت گره

علی کریمی*[۱]، اسماعیل بسطامی[۲]، مهران نعمتی[۳]، محمود صالح اصفهانی[۴]

۱- استادیار، ۲- مربی و پژوهشگر، ۳- دانشجوی کارشناسی ارشد، ۴- استادیار، دانشگاه جامع امام حسین(ع)

(دریافت: ۹۷/۱۰/۲۴، پذیرش: ۹۸/۱۱/۱۲)

چکیده

دنیای امروز تبدیل به یک محیط رقابتی در همه زمینه‌ها مانند سیاست، اقتصاد، مسائل اجتماعی، عقاید و مانند این‌ها شده است که برای پیشبرد اهداف خود از ابزار شبکه های اجتماعی استفاده می‌کنند. محققین برای رسیدن به این اهداف توسط شبکه های اجتماعی از بیشینه‌سازی نفوذ (IM) بهره می‌برند. وظیفه بیشینه‌سازی، شناسایی گره‌های مؤثری است که تحت عنوان گره‌های آغازگر شناخته می‌شوند و یک راهبرد برای رسیدن به بیشترین انتشار اطلاعات و یا کمترین اپیدمی با کمترین هزینه است. بیشینه سازی یک مسئله NP-hard است. محققان برای شناسایی گره‌های مؤثر به دنبال روش‌هایی برای کاهش پیچیدگی و دقت شناسایی قابل قبولی هستند؛ بنابراین در این تحقیق برای فائق آمدن به پیچیدگی مسئله و در عین حال بالا بردن دقت شناسایی، روشی جدید با ترکیب مرکزیت - فعالیت ارائه می‌شود. در این روش به‌صورت سراسری محدودیتی بر روی گراف شبکه برای استخراج گره‌ها توسط روش مرکزیت ایجاد می‌شود در ادامه، تحلیل گراف توسط روش فعالیت بروی گره‌های محلی استخراج شده صورت می‌گیرد. امتیاز تحلیل فعالیت با امتیاز روش مرکزیت ترکیب می‌شود که منتج به نمایش گره‌های مؤثر می‌شود. روش پیشنهادی با روش‌هایی نظیر Rank Page و مرکزیت نزدیکی مقایسه می‌شود و نتایج حاکی از آن است که روش پیشنهادی از نظر دقت در نقاط پایین بهتر از هر دو عمل کرده است و از طرفی توانسته است پیچیدگی پایین‌تری نسبت به هر دو داشته باشد. در آینده برای بالا بردن دقت در نقاط بالا می‌توان در مرحله تحلیل فعالیت از مفاهیم امتیازدهی تکراری استفاده نمود.

کلیدواژه‌ها: تأثیر، گره‌های مؤثر، بیشینه‌سازی نفوذ، مرکزیت، فعالیت

۱- مقدمه

در جهان واقعی موضوعاتی مانند شکل‌گیری روابط اجتماعی، روابط سیاسی، مسائل اقتصادی و مانند این‌ها به سمت پیچیده‌تر شدن حرکت می‌کنند از این رو دستگاه‌هایی را با روابط پیچیده تشکیل می‌دهند. این دستگاه‌ها را می‌توان به‌عنوان شبکه‌های پیچیده در نظر گرفت. این شبکه‌ها دارای خواص مشترکی همچون خاصیت جهان کوچک، ضریب خوشه‌ای بالا و مستقل از مقیاس هستند. به‌عنوان نمونه یکی از شبکه‌هایی پرکاربرد شبکه‌های اجتماعی است که این نوع شبکه‌ها به‌وسیله روابط متعددی مانند دوستی، خانوادگی، شغلی، همکاری و مانند این‌ها شکل می‌گیرد. شبکه‌های اجتماعی به این دلیل که هر رابطه نماینده ارتباط بین دو موجودیت واقعی است اهمیت تحلیلی فراوانی پیدا کرده است. ویژگی‌های شبکه‌های اجتماعی نیز موجب شده است که محققان به روی این نوع شبکه‌ها تحقیقات فراوانی را در زمینه‌های مختلفی همچون تشخیص انجمن، پیش‌بینی پیوند، تحلیل مرکزیت، تحلیل انتشار، گردشگری، شبکه‌های جنایی، مسئله ریزش و بازاریابی انجام دهند [۱, ۲] در این نوع از شبکه‌ها به دلیل ناهمگون بودن ساختار شبکه، نقش‌های مختلفی از گره‌ها را در شبکه پدید می‌آورند [۳]. در شبکه‌های اجتماعی برخی از گره‌ها ساختار و عملکرد شبکه را به میزان بیشتری نسبت به دیگران تحت تأثیر قرار می‌دهند که به گره‌های مؤثر معروف هستند [۴].

یکی از مهم‌ترین کاربردهای شناسایی گره‌های مؤثر بیشینه‌سازی انتشار است که با کمک این گره‌ها در یک شبکه، می‌توان انتشار را به حداکثر رساند. ازاین‌رو، در سال‌های اخیر شناسایی گره های مهم و مؤثر اهمیت پژوهشی فراوانی پیدا کرده است.

روش‌های شناسایی گره‌های مؤثر به دو دسته کلی روش‌های مرکزیت و روش‌های مبتنی بر فعالیت تقسیم می‌شوند. روش‌های مبتنی بر مرکزیت به مانند LR، PR، HITS، شاخص H-Index، بینابینی، درجه، نزدیکی و مانند این‌ها است و روش‌های مبتنی بر فعالیت را می‌توان به انتشار گرما و اندازه گیری شباهت اشاره کرد [۵ و ۶]. فعالیت در هر شبکه می‌تواند تعاریف مختلفی داشته باشد.

* رایانامه نویسنده مسئول: a.karimi@ihu.ac.ir

فعالیت گره‌ها در شبکه‌های اجتماعی مانند پیام‌ها و نظرها در بلاک‌ها، پیام‌ها و نظرها در انجمن‌ها، تحلیل فعالیت - مسافرت برای کمک به گردشگری، تحلیل حرکات مجرمانه، پیام‌ها و نظرها در پست‌ها تصویری، تعداد ویژگی‌های فردی، پیام‌های خصوصی، کلیک کردن، دنبال کننده‌ها، محتواهای منتشر شده، پسندیدن‌ها و مانند این‌ها باشد [۷, ۸]. روش‌هایی مانند مرکزیت درجه، ساده هستند و از پیچیدگی بسیار پایینی برخوردارند ولی به دلیل در نظر نگرفتن ساختار شبکه به صورت سراسری، دید ضعیفی نسبت به شبکه داشته و در نتیجه دقت لازم را ندارد. از طرفی روش‌هایی مانند مرکزیت بینابینی و مرکزیت نزدیکی به دلیل تحلیل سراسری بر روی شبکه دید کلی را از شبکه در اختیار می‌گذارد اما این روش‌ها به دلیل پیچیدگی بالا در شبکه‌های بزرگ قابل اجرا نیستند. البته باید در نظر داشت که روش‌های سراسری به دلیل اینکه ارتباطات میان همسایگان را نادیده می‌گیرند دچار افت دقت شناسایی خواهند شد [۹ و ۱۰]. از دیگر نقاط ضعف روش‌های مرکزیت این است که گره‌هایی را مهم و مؤثر می‌دانند که درون ساختار شبکه جایگاه برتری نسبت به سایر گره ها از جنبه های مختلف داشته باشند این روش‌ها به دلیل در نظر نگرفتن فعالیت‌های گره و همچنین نقش های مختلف گره ها در یک شبکه نمی‌توانند تحلیل مناسبی را بر روی شبکه داشته باشند. از آنجا که فعالیت یک گره با یک ضریب اهمیت به روی ویژگی سایر گره ها تأثیر می‌گذارد در نظر گرفتن موقعیت گره به تنهایی، دقت شناسایی را پایین می‌آورد [۱۱]. در شبکه های اجتماعی بین فعال ترین گره و مؤثرترین گره تفاوت وجود دارد بنابراین تکیه بر یک روش مانند فعالیت نیز برای شناسایی گره های تأثیرگذار موجب افزایش مثبت کاذب می‌شود و دقت شناسایی را پایین می‌آورد [۱۲]. این تحلیل ها از ویژگی‌های مختلفی استفاده می‌کنند که می‌توانند سراسری و محلی باشند. ویژگی محلی، ساده ولی دقت تحلیل آن پایین است در عوض، ویژگی سراسری تمام شبکه را برای تحلیل در برمی‌گیرد به همین دلیل تحلیل دقیق‌تری نسبت به ویژگی محلی دارد [۱۳].

در این تحقیق برای بهبود دقت شناسایی گره های مؤثر از اثر متقابل روش تحلیل مرکزیت و روش تحلیل فعالیت‌های گره با استفاده از دید سراسری و محلی استفاده می‌شود. مبتنی بر روش مرکزیت و فعالیت گره است که بتواند با وجود پیچیدگی پایین دقت مناسبی را ارائه دهد. در واقع تلفیق یک روش با دید سراسری به‌عنوان روش مرکزیت و روشی با دید نیمه محلی که فعالیت گره‌ها را اندازه‌گیری می‌کند. در ابتدا با ایجاد محدودیت در گراف، گره‌هایی را از طریق روش K-shell استخراج کرده و آن‌ها را امتیازدهی می‌کند. شباهت بین همسایگان این گره‌ها نیز به‌عنوان فعالیت، دوباره برای گره‌های استخراج شده امتیازی را در نظر می‌گیرید. امتیاز روش مرکزیت به‌عنوان ضریبی برای روش فعالیت خواهد بود.

روش پیشنهادی با استفاده از محیط جاوا به روی مجموعه داده delicious پیاده‌سازی شده است. از مدل SIR برای ارزیابی روش پیشنهادی استفاده می‌شود. پیاده‌سازی در شرایط مساوی بین روش پیشنهادی با دو روش PageRank و مرکزیت نزدیکی انجام شده است. گره‌های مؤثر در تعداد متفاوتی توسط سه روش بر روی مجموعه داده گفته شده استخراج می‌گردد. روش پیشنهادی توانسته است در اکثر نقاط از جمله در نقاط پایین نمودار بهتر از دو روش دیگر عمل کند. ارزیابی دقت توسط سطح معناداری آزمون مورد بررسی قرار گرفت که نتایج نشان دهنده قابل اعتماد بودن آن است. از نظر پیچیدگی روش مرکزیت نزدیکی بسیار ضعیف عمل کرده است ولی روش PageRank نشان داد که می‌تواند در شبکه‌های به نسبت بزرگ کارایی داشته باشد. روش پیشنهادی با وجود پیچیدگی پایین‌تر نسبت به روش PageRank توانسته در اکثر نقاط مهم نیز دقت بالاتری داشته باشد، بنابراین از این حیث می‌تواند در اغلب شبکه‌ها از جمله شبکه‌های با مقیاس بالا نیز اعمال شود.

۲- ادبیات پژوهش

یک گراف ساده و بدون جهت G با N گره و یال‌های E نشان داده می شود. در این بخش به صورت مختصر چندین سنجه مرکزیت تعریف می‌شود.

۲-۱- سنجه‌های مرکزیت

برای شناسایی گره‌های مؤثر در شبکه‌های مختلف از روش‌های مختلفی استفاده می‌شود. این روش‌ها با توجه به معیارهای مختلفی انتخاب می‌شود. معیارهایی مانند نوع ارتباط بین گره‌ها، نوع هدف، خصوصیات شبکه و مانند این‌ها موجب شده است که روش‌های گوناگونی برای تحلیل شبکه‌ها به وجود آید که آن‌ها را با سنجه‌های مرکزیت می‌شناسند. یکی از این سنجه‌ها، مرکزیت درجه است که می‌تواند نفوذ یک گره را به‌صورت محلی نشان دهد.

۲-۱-۱- مرکزیت درجه

مرکزیت گره یال‌های ورودی یا خروجی یک گره را محاسبه می‌کند. این روش، نفوذ کاربر به‌صورت محلی را نشان می‌دهد که گره با تعداد همسایه های بیشتر در مرکزیت گره به عنوان مؤثرترین گره‌ها در مباحثی همچون انتشار محسوب می‌شوند [۱۴]؛ بنابراین گره هایی که دارای درجه بالایی هستند می‌توانند بر رفتار دیگران در شبکه تأثیر بگذارند [۱۵]. این روش با وجود پیچیدگی پایین، به دلیل استفاده از درجه گره به‌صورت محلی نمی‌تواند در اغلب شبکه‌ها در شناسایی گره‌ها به‌صورت رضایت بخشی عمل کند.

۲-۱-۲- امتیازدهی K-Shell

روش امتیازدهی K-Shell [۱۶] یکی دیگر از روش‌های شناسایی گره‌های مؤثر است که بر اساس درجه گره عمل می‌کند. این روش برخلاف مرکزیت درجه به‌صورت سراسری، گره‌ها را درون لایه هایی با نام پوسته قرار می‌دهد. این روش با نام روش دیگری به نام K-Core نیز شناخته می‌شود با این تفاوت که K-Core به لایه‌های گراف اشاره نمی‌کند و تنها زیر گرافی با حداقل درجه K را بر می‌گرداند. هدف اصلی روش K-Shell رسیدن به لایه اصلی است و سایر لایه ها را در نظر نمی‌گیرد چون معتقد است که گره‌های مهم در مرکز هسته قرار دارند. K-Shell یک عدد صحیح Ks را به هر گره اختصاص می‌دهد که نشان دهنده لایه آن گره است. هر چه Ks بزرگ‌تر باشد گره‌ها به لایه هسته نزدیک‌ترند. فرآیند این روش در شکل (۱) آمده و روال آن به‌صورت زیر است:

- ابتدا با حذف گره‌های با درجه یک شروع می‌کند.
- فرآیند هرس ادامه خواهد یافت تا زمانی که هیچ گره‌ای با درجه‌ای برابر با ۱ وجود نداشته باشد.
- گره‌های حذف شده به‌عنوان لایه اول در نظر گرفته می‌شوند.
- به‌طور مشابه، فرایند هرس برای لایه‌های بعدی اعمال خواهد شد. این فرآیند تا زمانی که هسته شبکه پیدا شود ادامه خواهد یافت.

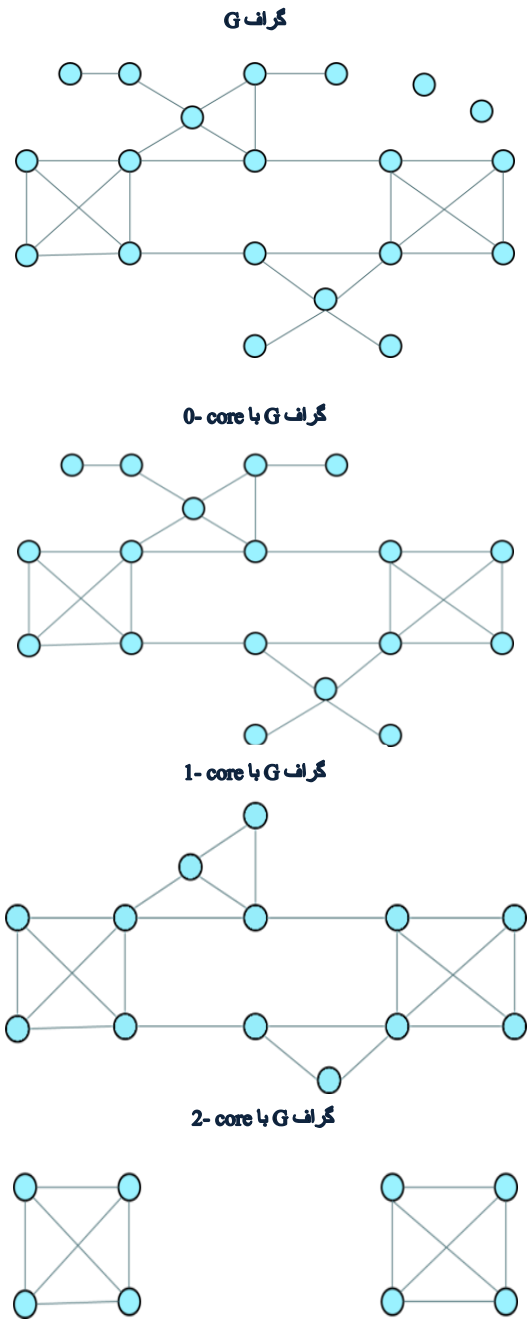
در شکل (۱) ابتدا گره‌های ایزوله حذف می‌شوند. در مرحله بعد گره‌هایی را که دارای درجه یک هستند و گره‌هایی را که دارای درجه یک می‌شوند حذف می‌گردند. در ادامه الگوریتم تا جایی ادامه پیدا می‌کند که گره‌های هسته تعیین شوند.

۲-۱-۳- مرکزیت نزدیکی

یک گره ممکن است به گره‌های زیادی گره خورده باشد بنابراین مرکزیت نزدیکی به فاصله یک گره از دیگران توجه می‌کند و ادعا دارد گره‌هایی که به همه گره‌ها نزدیک باشد به‌عنوان یک گره مرکزی و مؤثر شناخته می‌شود [۱۷, ۱۸]. رابطه (۱) نحوه محاسبه مرکزیت نزدیکی را نشان داده است.

$$C_{closeness}(x) = \left[\frac{1}{\sum_{i \neq x} d(x,i)} \right] \qquad (۱)$$

در رابطه (۱)، d(x) کوتاه‌ترین فاصله بین گره‌های x و i است. اگر n تعداد کل گره‌ها در شبکه باشد بیشترین احتمال (n-1) در این رابطه است. محاسبه مرکزیت نزدیکی برای هر گره نشان دهنده مدت زمان انتشار اطلاعات از یک گره به سایر گره‌هاست یعنی سعی می‌کند گره‌هایی را انتخاب کند که در مرکز شبکه هستند.



شکل (۱): مراحل روش K-Shell [۱۸]

۲-۱-۴- مرکزیت بینابینی

به‌طور خلاصه، روش بینابینی گره‌هایی را مهم می‌داند که به‌عنوان پل عمل می‌کنند. این روش گره‌هایی را شناسایی می‌کند که بیشترین کوتاه‌ترین مسیرها از او عبور کرده باشند. فریمن [۱۹] مرکزیت بینابینی را به عنوان یک روش برای کنترل کننده انسان‌ها در ارتباط با دیگر انسان‌ها در یک شبکه اجتماعی معرفی می‌کند. مرکزیت بینابینی گره های محدودی در شبکه را به عنوان بازیگران مهم در نظر می‌گیرد. محاسبه مرکزیت بینابینی در رابطه (۲) آمده است.

$$C_{betweenness}(x) = \sum_{y,z \neq x, \sigma_{y,z} \neq 0} \frac{\sigma_{y,z}(x)}{\sigma_{y,z}} \quad (۲)$$

که در آن، σy,z تعداد کوتاه‌ترین مسیر بین گره‌های z و y است. عبارت σy,z(x) تعداد کوتاه‌ترین مسیرهای بین z و y است که از x می‌گذرد را نشان می‌دهد. گره‌ها با مرکزیت بینابینی بالا به‌عنوان رابط‌هایی بین جوامع شناخته می‌شوند یعنی اگر این گره ها از شبکه حذف شوند موجب افزایش طول جغرافیایی خواهد شد و یا حتی موجب قطع شدن شبکه می‌شود.

۲-۱-۵- مرکزیت مقادیر ویژه

این روش [۲۰] اهمیت گره‌ها را بر اساس گره‌های مجاور محاسبه می‌کند. این محاسبه در گراف‌های با اتصال قوی اتفاق می‌افتد. اگر گره‌ای به گره‌هایی که دارای اهمیت بالایی هستند متصل باشد، تحت تأثیر آن‌ها اهمیت او نیز بالا می‌رود. این روش به صورت تکراری برای محاسبه گره اهمیت همسایگان را نیز در نظر می‌گیرد. ابتدا به همه گره‌ها یک امتیاز اولیه داده می‌شود. در ادامه به‌صورت زنجیره‌ای تا زمانی که به پایداری برسد این امتیازدهی ادامه می یابد. امتیازدهی در این روش بر اساس این مفهوم است که گره‌های با اتصالات بالا به گره‌های دنبال کننده آن‌ها از نظر امتیاز کمک می‌کنند. روش رتبه‌بندی صفحه از این روش الگوبرداری کرده است[۲۱].

۲-۱-۶- رتبه‌بندی صفحه

می‌توان گفت تا به امروز رایج ترین روش در زمینه تحلیل شبکه‌های اجتماعی روش رتبه‌بندی صفحه است[۲۱]. اکثر روش‌های موجود یا از این روش ایده گرفته‌اند یا گسترش یافته آن هستند. ایده اصلی PR به این صورت است که امتیاز گره ها مبتنی بر جمع کیفیت پیوندهای اشاره شده به آن است. اعتبار هر گره در الگوریتم PR به مانند استناد در مفاهیم استنادات علمی است[۱۸]. اعتبار صفحه p وابسته به تعداد پیوندهای ورودی به او و اعتبار صفحه qای که به او اشاره‌کرده دارد. در شکل (۲) همان طور که مشاهده می‌شود گره b دارای بیشترین پیوند ورودی و گره c نیز به اعتبار گره b به ترتیب دارای بیشترین رتبه در میان سایرین هستند. رتبه‌بندی گره‌ها در الگوریتم PR با توجه به رابطه (۳) محاسبه می‌شود[۲۲].

$$PR_p = (1-d) + \sum_{q \in pa[p]} \frac{PR_q}{h_q} \quad (۳)$$

در رابطه (۳) d ضریب میرایی است و عددی بین ۰ و ۱ خواهد بود و اغلب آن را برابر با عدد ۰/۱۵ می‌گیرند و نشان دهنده احتمال پرش به یک صفحه است. Pa[p] مجموعه گره‌هایی که به p اشاره دارند. PRq رتبه صفحه q است و hq تعداد صفحاتی است که q به آن‌ها اشاره کرده است با این کار امتیاز اعتبار صفحه q برای صفحه p تقسیم می‌شود.

۳- روش پیشنهادی ICA

روش ICA[1] دارای سه مرحله کلی است که در مرحله اول محدودیت داده‌ای روی گراف اعمال می‌شود که موجب کاهش جست‌وجو و محاسبات غیر ضروری می‌شود. در این مرحله گره‌ها توسط روش K-Shell استخراج شده و هر کدام در لایه ای قرار می‌گیرند. امتیازها بر اساس لایه‌ای که گره‌ها درون آن قرار گرفته‌اند به هر کدام از گره‌ها اختصاص می‌یابد.

مرحله دوم گره‌های استخراج شده از مرحله قبل را به وسیله میزان شباهت مورد تحلیل قرار می‌دهد. این مرحله میزان شباهت هر گره با همسایگان مجاور خود را محاسبه می‌کند. در آخرین مرحله امتیاز مرحله اول و امتیاز مرحله دوم با یک ضریب اختیاری بر روی یکدیگر تأثیر خواهند گذاشت و درنهایت k گره اول انتخاب می‌شود. مراحل روش پیشنهادی در زیر آمده است:

۳-۱- مرحله اول

استخراج گره ها، مبتنی بر محدودیت با استفاده از K-Shell عبارت است از:

در روش K-Shell اولین نکته‌ای که باید رعایت شود انتخاب پارامتر kmax است. این پارامتر تعداد لایه های روش K-Shell را مشخص می‌کند. در اینجا دو پارامتر جدید Ks_i و Kn باید تعریف شوند. پارامتر اول امتیازی را مبنی بر Ks که همان شماره لایه گره باشد با عنوان پارامتر Ks_i تعریف می‌کند. این پارامتر برای هر گره در مرحله اول محاسبه شده و در مرحله سوم وارد محاسبات می‌شود. پارامتر دوم به نام Kn، تعداد لایه های از سمت هسته را نشان می‌دهد. این پارامتر به صورت اختیاری است و توسط کاربر تعیین شود.

در مرحله اول به دلیل استفاده از روش مرکزیت گراف G در شکل (۲- الف) بدون ویژگی در نظر گرفته می‌شود. این مرحله ۳ گام دارد و بروی گراف G به‌صورت زیر اعمال می‌شود:

گام اول: تنظیم کردن پارامترها؛ به دلیل چگالی پایین گراف K G برابر با ۳ در نظر گرفته می‌شود، به دلیل چگالی پایین، Kn برابر با ۲ خواهد بود به این معنی که گره های لایه هسته و لایه دوم باید استخراج شوند و سایر گره‌ها نادیده گرفته شود.

گام دوم: ایجاد کردن محدودیت؛ برای استخراج گره‌ها در وهله اول از لایه های بیرونی شروع می‌شود این لایه دارای پارامتر

[1] Integrated Centrality and Activity

Ks برابر ۱ است. محاسبه آن به این صورت است که گره های با درجه یک ابتدا حذف می‌شود. بعد از آن گره های با درجه ۱ دوباره شناسایی می‌شوند و حذف خواهند شد تا جایی که گره‌هایی با درجه ۱ دیگر وجود نداشته باشند. تمام گره‌های حذف شده در لایه اول قرار می‌گیرند. مطابق شکل (۲- ب) گره های زرد رنگ در این لایه قرار می‌گیرند. باید توجه داشت که گره‌های ۱ و ۱۱ به این دلیل در این لایه قرار می‌گیرند که اگر گره‌های برگ را حذف کنیم گره‌های ۱۱ تنها به یک گره دیگر متصل می‌شوند و دارای درجه یک خواهند شد.

شکل (۲- ب) زیر گرافی از G خواهد بود که حداقل دارای درجه یک است. بعد از اینکه گره‌های لایه اول مشخص شدند یال‌های آن‌ها دیگر در نظر گرفته نمی‌شود. برای تشکیل دادن لایه Ks=2 گره هایی در آن قرار می‌گیرند که دارای زیرگرافی با حداقل درجه ۲ باشند. همان طور که در شکل (۳- پ) مشاهده می‌شود گره‌های ۷، ۸ و ۱۵ در لایه دوم قرار گرفته‌اند و با رنگ سبز نمایش داده شده است. بعد از لایه اول و دوم لایه هسته نیز به همین صورت پیش می‌رود. درنهایت با توجه به انتخاب kn=2 تنها گره‌های لایه هسته و لایه ۲ در نظر گرفته می‌شود و سایر گره‌ها حذف می‌گردند. لایه هسته نیز در شکل (۲-ت) آمده است که گره‌های آن با رنگ آبی مشخص شده‌اند.

گام سوم: در این گام مطابق با Kn که عدد ۲ در نظر گرفته شده است گره هایی را که در لایه هسته و لایه دوم قرار دارند استخراج می‌شود و امتیازدهی می‌گردد. در شکل (۴) گره های استخراج شده در مرحله اول را نشان داده است و همچنین امتیازهای آن‌ها در جدول (۱) قابل مشاهده است.

جدول (۱): امتیازدهی مرحله اول به گره‌های لایه هسته و لایه دو

شناسه گره	امتیاز مبتنی بر مرکزیت
۹	۳
۱۱	۳
۱۲	۳
۱۳	۳
۷	۲
۸	۲
۱۵	۲

۳-۲- مرحله دوم

محاسبه شباهت: با استفاده از میزان شباهت گره‌ها و رابطه (۴)، میزان شباهت گره های استخراج شده در جدول (۲) محاسبه شده‌اند. برای محاسبه فعالیت گره‌ها باید میزان شباهت گره با دیگر همسایگانش را به‌دست آورد در نتیجه هر گره با استفاده از مجموع شباهت هایش که با ACT_i نمایش داده می‌شود با همسایگان خود دارای یک امتیاز مبتنی بر فعالیت به‌صورت محلی است. در شکل (۴) گره‌های استخراج شده نمایش داده شده و میزان شباهت آن‌ها نیز بروی یال ارتباطی خود با همسایه‌هایشان درج شده است. در جدول (۲) مجموع امتیازهای گره‌های استخراج که به‌عنوان فعالیت آن‌ها در نظر گرفته می‌شود آمده است.

$$ACT_i = \sum_{j=\text{neighbours}\_i} sim(\text{i}, \text{j}) \qquad (۴)$$

به عنوان مثال برای محاسبه میزان فعالیت گره ۱۳ مانند رابطه (۵) عمل می‌شود.

$$ACT_{13} = \sum_{j=9,11,12,15} sim(13, \text{j}) = (0.21 + 0.17 + 0.28 + 0.36 + 0.30) \qquad (۵)$$

جدول (۲): اندازه‌گیری فعالیت مبتنی بر شباهت گره‌های استخراج شده

شناسه گره	امتیاز مبتنی بر فعالیت
۹	۱/۴۱
۱۱	۰/۶۸
۱۲	۱/۱۲
۱۳	۱/۱۶
۷	۱/۲۵
۸	۰/۸۸
۱۵	۰/۹۰

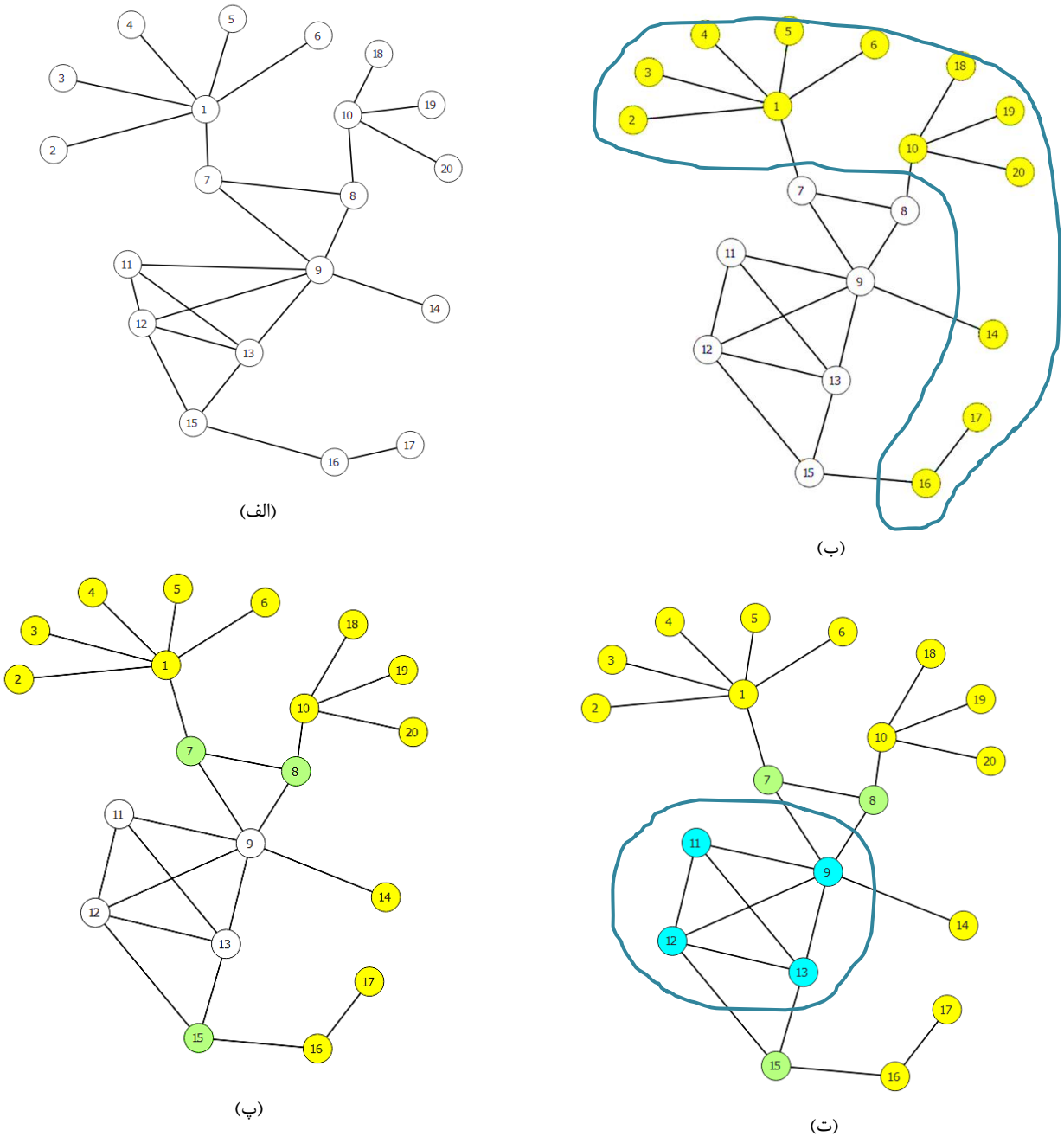
۳-۳- مرحله سوم

این مرحله گره‌های به دست آمده را در فهرستی مطابق جدول (۳) قرار می‌دهد. ۴ گره اول به‌عنوان مؤثرترین گره‌ها در نظر گرفته می‌شود. در جدول (۳) امتیاز مبتنی بر مرکزیت به عنوان ضریبی برای امتیاز مبتنی بر فعالیت خواهد بود. به‌عنوان گره‌های نهایی گره‌های ۱۲،۱۳،۹ از لایه هسته و گره ۷ از لایه دوم به‌عنوان مؤثرترین گره روش پیشنهادی شناخته می‌شوند.

۳-۴- ضرایب

رابطه کلی روش ICA به‌صورت زیر است:

$$ICA_i = \alpha ACT_i + \beta S_{Centrality_i} \qquad (۶)$$



شکل (۳): مراحل روش مرکزیت با استفاده از روش امتیازدهی K-Shell. (الف) گراف ساده G، (ب) لایه Ks=1، (پ) لایه Ks=2، (ت) لایه Ks=3

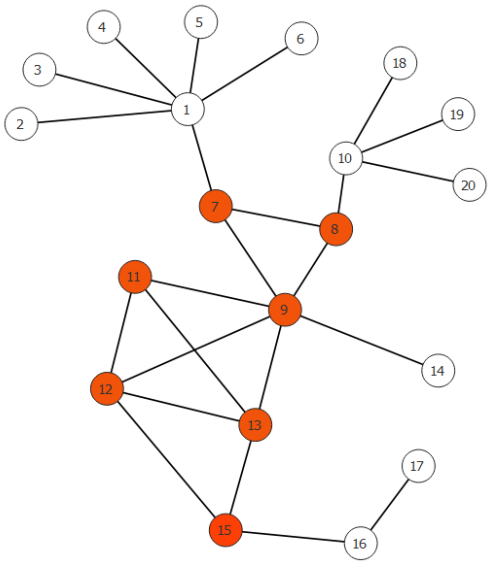
در رابطه (۶)، ACT_i فعالیت گره i و $S_{Centralityi}$ مرکزیت گره i است که در این تحقیق K-Shell در نظر گرفته شده است. ضریب α اهمیت فعالیت را می‌رساند و β اهمیت مرکزیت را عنوان می‌کند. در برخی از شبکه ها ممکن است فعالیت یا جایگاه گره هر کدام اهمیت متفاوتی داشته باشند به همین دلیل این ضرایب کمک می‌کنند که روش پیشنهادی به بهترین وجه عمل کند. برای راحتی تحلیل در اینجا $\alpha + \beta = 1$ در نظر گرفته می‌شود.

۳-۵- نرمال‌سازی

نرمال‌سازی روش ICA مطابق با رابطه (۷) به‌صورت زیر است:

$$ICA' = \frac{ICA_I}{\sum_{I \in N} ICA_I} \qquad (۷)$$

که N تعداد کل گره‌های یک شبکه است.



شکل (۴): گره‌های شناسایی شده در مرحله اول روش پیشنهادی

جدول (۳): امتیاز گره‌های استخراج شده با روش پیشنهادی

شناسه گره	امتیاز مرکزیت	امتیاز فعالیت	امتیاز نهایی	رتبه گره
۹	۳	۱/۴۱	۴/۲۳	۱
۱۱	۳	۰/۶۸	۲/۰۴	۵
۱۲	۳	۱/۱۲	۳/۳۶	۳
۱۳	۳	۱/۱۶	۳/۴۸	۲
۷	۲	۱/۲۵	۲/۵	۴
۸	۲	۰/۸۸	۱/۷۶	۷
۱۵	۲	۰/۹۰	۰/۱۸۰	۶

۴- تجزیه و تحلیل داده‌ها

۴-۱- مجموعه داده

مجموعه داده‌ای به‌عنوان شبکه‌های اجتماعی نشانه‌گذاری Delicious [۲۳] در سال ۲۰۱۱ توسط ACM در دومین کارگاه آموزشی بین‌المللی اطلاعات ناهمگن و فیوژن در سامانه‌های توصیه‌گر منتشر شد. Delicius یک سرویس خدمات نشانه گذاری است که امکان برچسب‌گذاری، ذخیره و به اشتراک‌گذاری تمام صفحات وب در یک محل را ارائه می‌دهد. هر کاربر دارای نشانه‌گذاری، برچسب‌ها، تاپل‌هایی مانند [کاربر، برچسب، نشانه‌گذاری] و روابط ارتباطی در شبکه است. برچسب‌ها به‌عنوان ویژگی برای این مجموعه داده محسوب می‌شود. خلاصه ای از داده‌های این مجموعه داده در جدول (۴) آمده است.

جدول (۴): خلاصه‌ای از ویژگی‌های مجموعه داده

فایل‌ها	تعداد	میانگین/نسبت به	توضیحات
کاربران	۱۸۶۷	-	-
آدرس‌ها	۶۹۲۲۶	-	دامنه‌ها و زیر دامنه‌ها
آدرس‌های خاص	۳۸۵۸۱	-	دامنه‌های منحصربه‌فرد
ارتباطات کاربران	۷۶۶۸- ۱۵۳۲۸	۸,۲۳۶/کاربر	به‌صورت متقارن
برچسب‌ها	۵۳۳۸۸	-	-
تخصیص برچسب(TAS)	۴۳۷۵۹۳	۳۸۳.۲۳۴/آدرس	تاپل [کاربر، برچسب، آدرس]
نشانه‌گذاری‌ها	۱۰۴۷۹۹	۵۶۱۳۲/کاربر	جفت‌های متمایز [کاربر، آدرس]

ارتباطات بین افراد توسط رابطه طرفداری متقابل شکل می‌گیرد. کاربران دارای نشانه‌گذاری‌هایی که از آدرس‌های اینترنتی شکل گرفته هستند و آن‌ها توسط کلیدواژه‌هایی برچسب‌گذاری می‌شود. خصوصیت‌های هر گره - که تعداد آن‌ها ۱۳۵۰ عدد است - نیز به وسیله‌ی یک آرایه دودویی مشخص شده است. خصوصیات با استفاده از مقادیر ۰ و ۱ در آرایه مربوط به هر گره مشخص شده که برای تعیین وجود یا عدم وجود خصوصیت در یک گره استفاده می‌شوند. در واقع برای هر گره یک آرایه ۱۳۵۰ خانه‌ای وجود دارد که برای هر ویژگی یک خانه در نظر گرفته شده است. اگر گره دارای ویژگی مورد نظر بود، خانه مورد نظر برای آن ویژگی مقدار ۱ خواهد داشت، در غیر این صورت مقدار خانه برابر با ۰ خواهد بود. این کار باعث می‌شود که خصوصیات گره‌ها به‌صورت دودویی ذخیره شده تا هنگام محاسبه شباهت، بتوان از شباهت جاکارد به راحتی استفاده نمود. در این تحقیق از این ویژگی‌ها با دید یک فعالیت غیر تعاملی نگاه می‌شود و از آن برای تحلیل فعالیت های اشخاص به‌کار گرفته می‌شود.

۴-۲- محیط آزمایش

پیاده سازی روش پیشنهادی تحت برنامه نویسی جاوا با استفاده از بسته‌های ریاضی به نام JgraphT و JgraphX در محیط Eclipse

Java EE نسخه Oxygen انتشار ۴,۷,۰ انجام شده است. روش‌های نزدیکی و رتبه‌بندی صفحه تحت همین شرایط و تمامی آزمایش‌ها در محیط جدول (۵) انجام شده است.

جدول (۵): محیط آزمایش

مدل	پردازشگر		حافظه موقت	پلتفرم
	مدل	هرتز		
Samsung RC520	Intel i5-2410	2.30 GHz	4GB	Win7.ultimate 64-bit

۴-۳- ارزیابی با مدل SIR

برای شبیه‌سازی روند انتشار، از مدل SIR در مجموعه داده کمک گرفته می‌شود. این مدل برای توصیف گسترش بیماری و نیز انتشار اطلاعات و شایعه در فرایندهای اجتماعی استفاده می‌شود. همان طور که در فصل مرور مفاهیم مبنایی گفته شد مدل SIR دارای حالت‌های مختلفی است و با توجه به شبکه مورد تحلیل، می‌توان حالت های مختلف این مدل را با جهان واقعی تطبیق نمود. با استفاده مدل SIR قرار است نفوذ و تأثیر گره های مؤثری که توسط روش‌ها شناسایی‌شده‌اند ارزیابی شود. هدف از شبیه‌سازی با مدل SIR انطباق روش‌ها تنها با یک مدل در جهان واقعی مانند تبلیغات نیست بلکه ارزیابی توانایی گسترش اطلاعات، ویروس‌ها، شایعه و یا هر فرایندی دیگر توسط گره‌های مؤثر است. به همین دلیل از حالت کلی ارزیابی دقت یعنی مدل SIR استفاده می‌شود که مرسوم ترین روش در محافل علمی برای این امر به شمار می‌رود. فرآیند شبیه سازی انتشار با مدل SIR به‌صورت زیراست:

در ابتدا، تمام گره‌ها وضعیت حساس دارند به‌جز مجموعه‌ای از گره‌های آلوده که به‌عنوان گره‌های آغازگر انتخاب شده‌اند. در هر مرحله دوم، گره آلوده تلاش می‌کند تا یکی از همسایگان خود را با احتمال μ آلوده کند. در همان زمان، هر گره آلوده با احتمال β بهبود خواهد یافت، اگر موفق شد دوباره آلوده نخواهد شد و دیگر گره‌های حساس دیگر را آلوده نخواهد کرد. این روند زمانی متوقف می‌شود که هیچ گره آلوده‌ای در شبکه وجود نداشته باشد. از $\lambda = \mu / \beta$ برای نشان دادن نرخ آلوده کردن استفاده می‌شود که اغلب برای نشان دادن توانایی گسترش گره‌های مؤثر است. در آزمایش‌ها نرخ آلوده کردن را با مقدار همبستگی گره‌ها با یکدیگر که مطابق با شباهت آن‌ها تعیین شده، گرفته شده است. برای اجرای شبیه‌سازی ابتدا باید نتایج پیاده‌سازی را به‌دست آورد. پیاده‌سازی دو روش نزدیکی و رتبه‌بندی صفحه و همچنین روش پیشنهادی در جدول (۶) آمده است.

جدول (۶): ۱۰ گره برتر پیاده‌سازی روش‌ها بروی مجموعه داده

رتبه	شناسه گره		
	Closeness	PageRank	ICA
۱	۲۶۸	۴۵۲	۱۵۷
۲	۴۵۴	۵۱	۳۵
۳	۷۲۰	۲۸۳	۱۰۵
۴	۲۷۲	۴۵۳	۱۰۳
۵	۱۲۸۴	۱۲۵	۳۰۹
۶	۶۱۴	۷۷۷	۳۳۲
۷	۲۵۸	۱۴۵	۱۵
۸	۴۵۳	۶۵۷	۱۷
۹	۸۲۰	۱۵۲	۱۰۴
۱۰	۶۱۳	۵۴۴	۴۲۱

در جدول (۶) ۱۰ گره مؤثر هر سه روش نمایش داده شده است که این گره‌ها به‌عنوان آغاز کننده شبیه‌سازی در مدل SIR خواهند بود. البته برای صحت نتایج شبیه سازی از ۱۰ گره برتر تا ۹۰ گره برتر در مجموعه داده شبیه‌سازی انجام می‌شود.

۴-۴- نتایج شبیه‌سازی

۴-۴-۱- بررسی دقت شناسایی و کارایی

شبیه‌سازی نتایج روش پیشنهادی با دو روش دیگر در شکل (۵) آمده است. نمودار، تعداد گره‌های آلوده طبق گام‌های شبیه‌سازی را نشان می‌دهد. این نمودار کارایی روش‌ها را به این صورت که با زیادتر شدن تعداد گره‌های آغازگر، گره‌های آلوده در تابع اولیه $F(c)$ چه تعداد است. همان‌طور که مشاهده می‌شود روش نزدیکی با توجه به کارایی آن نسبت به سایر روش‌های مرکزیت در اینجا نمره قابل قبولی را به‌دست نیاورده است. روش پیشنهادی ما در گره‌های پایین توانسته کارایی خود را نسبت به PageRank ثابت کند ولی در نقطه ۷۰ روش پیشنهادی از نظر انتشار با روش PageRank به هم می‌رسند. هر چند روش پیشنهادی در نقطه‌های بالا دقت قابل قبولی را دارد اما در آن نقطه PageRank برتری نسبی به روش پیشنهادی پیدا می‌کند. دلیل برتری PageRank این است که روش پیشنهادی از روش محدودیت

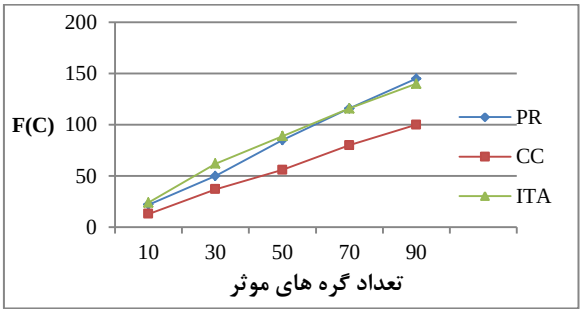
k-Shell استفاده می‌کند و این روش علی‌رغم دقت مناسبی که در شبکه‌های بزرگ دارد در گره‌های بالا برای شبیه‌سازی به مرکز شبکه توجه می‌کند و این موضوع موجب می‌شود که روش پیشنهادی با توجه به برتری آن در بیشتر مراحل شبیه‌سازی به گره‌هایی که در مرکز شبکه نیستند و ممکن است به‌عنوان یک دروازه اطلاعاتی عمل کنند توجه نمی‌نماید. با توجه به شیب روش پیشنهادی می‌توان گفت که در k گره‌های مؤثر پایین بهتر عمل می‌کند و برای روش PageRank نیز با توجه به شیب انتشار اطلاعات آن، در k گره‌های بالاتر بهتر عمل می‌کند. البته در شناسایی گره‌های مؤثر باید توجه داشت که هدف اصلی انتخاب کمترین گره و رسیدن به بیشترین گسترش اطلاعات است بنابراین کارایی روش پیشنهادی برای اهداف انتشار بیشتر از PageRank است.

۴-۴-۲- بررسی پیچیدگی زمانی

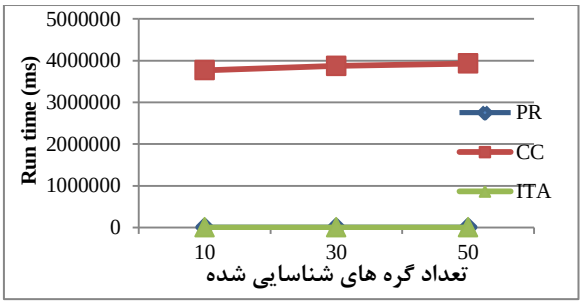
پیچیدگی زمانی یکی از مهم‌ترین مؤلفه‌های انتخاب روش برای تحلیل شبکه‌های اجتماعی در سطح کلان است. روش‌های موجود اغلب پیچیدگی بسیار بالایی دارند و به همین دلیل حتی در شبکه‌های به نسبت بزرگ نیز استفاده نمی‌شوند. یکی از این روش‌ها که در شبکه‌های کوچک کارای مناسبی دارد مرکزیت نزدیکی است که دارای پیچیدگی بسیار زیادی است. پیچیدگی آن دارای O(n3) است که نمی‌تواند روش مناسبی برای تحلیل باشد. پیچیدگی زمانی روش پیشنهادی و PageRank به متغیرهای زیادی وابسته است بنابراین برای مقایسه پیچیدگی از زمان اجرا در محیط مشابه استفاده شده است. در جدول (۷) زمان اجرای روش‌ها و در شکل (۶) پیچیدگی زمانی آن‌ها در محیط گفته شده آورده شده است. همان‌طور که مشاهده می‌شود روش مرکزیت نزدیکی از زمان اجرای بسیار بالاتری نسبت به روش پیشنهادی و PageRank برخوردار است. روش پیشنهادی توانسته از زمان اجرای کمتری نسبت به PageRank مسئله را حل کند که البته در شبکه‌های بزرگ‌تر این زمان اجرا بیشتر خود را نشان می‌دهد.

جدول (۷): پیچیدگی زمانی روش ICA و سایر روش‌ها

روش	زمان اجرا (میلی‌ثانیه)
Closeness	۳۷۶۹۹۳۴
PageRank	۳۴۴۳
روش پیشنهادی (ICA)	۱۰۷



شکل (۵): نتیجه ارزیابی روش ICA و سایر روش‌ها



شکل (۶): مقایسه زمان اجرای روش ICA و سایر روش‌ها

۴-۵- تحلیل نتایج ارزیابی

برای تصادفی یا قطعی بودن نتایج یک آزمایش و یا رد کردن و تأیید کردن یک فرضیه از سطح معناداری آزمون استفاده می‌کنند. سطح معناداری می‌تواند به عنوان یک معیار برای قابل اعتماد بودن همبستگی بین عوامل و قابل اتکاء بودن نتایج حاصل شده از یک آزمایش قرار بگیرد. سطح معناداری را با P-Value نمایش می‌دهند که البته در نرم‌افزارهایی مانند SPSS آن را با Sig که مخفف Significance است نیز نمایش می‌دهند. در آزمایش سطح معناداری پارامتری به نام α وجود دارد و سطح خطا را مشخص می‌کند این سطح خطا به صورت اختیاری است اما در زمینه‌هایی مانند علوم پزشکی که دارای حساسیت بالایی است مقدار ۰/۰۱ را به آن اختصاص می‌دهند. در دیگر علوم از مقادیر ۰/۰۵ و ۰/۱ نیز استفاده می‌کنند. این سطح خطا به این معناست که اگر نتیجه آزمایش کمتر از مقدار α باشد نشان دهنده تفاوت معنی‌دار بین نتایج مختلف شبیه سازی است. در این تحلیل α برابر با ۰/۰۵ قرار می‌گیرد. این آزمون و تحلیل سطح معناداری تحت نرم افزار SPSS نسخه ۱۶ صورت گرفته است. قبل از آزمون، نرمال بودن داده‌ها مورد بررسی قرار گرفته شده است و داده‌ها به‌صورت نرمال هستند. جدول (۸) به‌صورت خلاصه، نتایج سطح معناداری شبیه‌سازی روش ICA و PR را نشان می‌دهد. در این آزمون از روش آزمون t دو نمونه‌ای وابسته کمک گرفته شده است. این آزمون دو جمعیت به یکدیگر وابسته را از جهت سطح معناداری مورد ارزیابی قرار می‌دهد تا ببیند که

آیا اگر جمعیت تغییر کند درنهایت به روی خروجی اثر خواهد گذاشت؟ با توجه به جدول (۸)، تمامی مقادیر ستون مربوط به Sig بیشتر از مقدار α است و این نشان دهنده قطعی بودن نتایج ارزیابی دقت است پس می‌توان نتیجه گرفت که ارزیابی صورت گرفته در تمامی نقاط، نتایجی قطعی را ارائه داده است.

جدول (۸): آزمون t دو نمونه‌ای وابسته برای سطح معناداری آزمون

	Paired Differences					t	df	Sig/ (2-tailed)
	Mean	Std/ Deviation	Std/ Error Mean	95% Confidence Interval of the Difference				
				Lower	Uper			
pr/۱۰ pr/۱۰/ n	۱,۱۳۳۳۳	۴,۲۲۳۵۲	۱,۰۹۰۵۱	-۱,۲۰۵۵	۳,۴۷۲۲۴	۱,۰۳۹	۱۴	۰,۳۱۶
ica/۱۰ ica/۱۰/ n	-۲,۴۶۶۶۷	۶,۰۲۲۱۸	۱,۵۵۴۹۲۱	-۵,۸۰۱۶۴	۰,۸۶۸۳۱	-۱,۵۸۶	۱۴	۰,۱۳۵
pr/۳۰ pr/۳۰/ n	۰,۸۶۶۶	۳,۳۵۶۵۹	۰,۸۶۶۷	-۰,۹۹۱۲۵	۲,۷۲۵۴۸	۱,۰۰۰	۱۴	۰,۳۳۴
ica/۳۰ ica/۳۰ /n	-۰,۹۳۳۳۳	۹,۴۳۷۵۱	۲,۴۳۶۷۶	-۶,۱۵۹۶۵	۴,۲۹۲۹۹	-۰,۳۸۳	۱۴	۰,۷۰۷
pr/۵۰ pr/۵۰/ n	۰,۷۳۳۳	۶,۳۱۸۵۳	۱,۶۳۴۴	-۲,۷۶۵۷۵	۴,۲۳۲۴۲	۰,۴۵۰	۱۴	۰,۶۶۰
ica/۵۰ ica/۵۰ /n	۰,۱۳۳۳	۶,۳۷۹۲۸	۱,۶۴۷۱۲۱	-۳,۳۹۹۴	۳,۶۶۶۰۶	۰,۰۸۱	۱۴	۰,۹۳۷
pr/۷۰ pr/۷۰/ n	۰,۴۶۶۶۷	۶,۸۴۳۸۴	۱,۷۶۷۰۷	-۳,۳۳۳۳	۴,۲۵۶۶	۰,۲۶۴	۱۴	۰,۷۹۷
ica/۷۰ ica/۷۰ /n	۰,۵۳۳۳	۸,۴۱۶۵۴	۲,۱۷۳۱۴	-۴,۱۲۷۵۹	۵,۱۹۴۲۶	۰,۲۴۵	۱۴	۰,۸۱۰
pr/۹۰ pr/۹۰/ n	۰,۱۳۳۳۳	۱۲,۲۲۹۱۶	۳,۱۵۷۵۶	-۶,۶۸۸۹۳	۶,۹۰۵۶۲	۰,۰۴۲	۱۴	۰,۹۶۷
ica/۹۰ ica/۹۰/ n	-۰,۲۰۰۰۰	۴,۴۱۱۰۳	۱,۱۳۸۹۲	-۲,۶۴۲۷۴	۲,۲۴۲۷۴	-۰,۱۷۶	۱۴	۰,۸۶۳

۵- نتیجه‌گیری

روش پیشنهادی با ترکیب روش مرکزیت و فعالیت گره با ایجاد محدودیت بروی گراف ارائه شد. روش پیشنهادی دارای دو مرحله اصلی است که در مرحله اول سعی می‌شود پردازش را در مرحله دوم کاهش دهد تا علاوه بر افزایش دقت شناسایی پیچیدگی نیز کاهش پیدا کند. البته در روش پیشنهادی برخلاف روش پایه‌ای K-shell که تنها به گره‌های هسته توجه دارد، گره‌های لایه‌های دیگر را نیز بر اساس چگالی و قطر گراف در نظر می‌گیرد. روش پیشنهادی برای این که بتواند از گره‌های استخراج شده مؤثرترین گره‌ها را شناسایی کند روش مبتنی بر فعالیت را به‌کارگیری کرده و به صورت محلی و تأثیرات نیمه محلی عمل می‌کند. با ادغام روش‌های مرکزیت و فعالیت که به‌صورت ضریبی برای یکدیگر محسوب می‌شوند در واقع موجب ادغام تأثیر گره به‌صورت سراسری و نیمه محلی است. به دلیل ذات روش های مبتنی بر مرکزیت، آن‌ها نمی‌توانند در همه نقاط بهترین باشند که در شکل (۵) قابل ملاحظه است. با توجه به انتخاب روش K-Shell به‌عنوان روش مرکزیت بدیهی است به گره‌هایی که در خارج از مرکز شبکه هستند توجهی نشود. البته با تغییرات بروی K-Shell سعی شد این نقطه ضعف کمرنگ شود ولی این مسئله موجب می‌شود که روش پیشنهادی در نقاط بالا نسبت به روش PageRank ضعیف‌تر عمل کند. روش PageRank به دلیل توجه داشتن به تمامی شبکه، در برخی نقاط بالا توانسته از روش پیشنهادی بهتر عمل کند. از نظر پیچیدگی روش مرکزیت نزدیکی حرفی برای گفتن ندارد ولی روش PageRank نشان داد که می‌تواند در شبکه های نسبتاً بزرگ کارایی داشته باشد. روش پیشنهادی با توجه به اینکه توانسته پیچیدگی پایین تری از PageRank داشته باشد در اکثر نقاط مهم نیز بهتر از این روش عمل کرده است لذا از این حیث می‌تواند در اغلب شبکه‌ها از جمله شبکه‌های بزرگ اعمال شود.

۶- مراجع

[1] Z. Sun, et al., "Identifying influential nodes in complex networks based on weighted formal concept analysis," IEEE Access, vol. 5, pp. 3777-3789, 2017.

[2] K. Taha and P. D. Yoo, "Using the spanning tree of a criminal network for identifying its leaders," IEEE Transactions on Information Forensics and Security, vol. 12, no. 2, pp. 445-453, 2016.

[3] A.-L. Barabási, "Scale-free networks: a decade and beyond," science, vol. 325(5939), pp. 412-413, 2009.

[4] Y. Sun, et al, "Key nodes discovery in large-scale logistics network based on MapReduce," In 2015 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2015.

[5] S. Singh, N. Mishra, and S. Sharma, "Survey of various techniques for determining influential users in social networks," In 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), IEEE, 2013.

[6] R. Rabade, N. Mishra, and S. Sharma, "Survey of influential user identification techniques in online social networks," In Recent advances in intelligent informatics, Springer, pp. 359-370, 2014.

[7] G. Spadon, et al., "Behavioral Characterization of Criminality Spread in Cities," In ICCS, 2017.

[8] J. Kim, S. Rasouli, and H. J. Timmermans, "Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence," Transport Reviews, vol. 38, no. 4, pp. 499-523, 2018.

[9] T. Bian, J. Hu, and Y. Deng, "Identifying influential nodes in complex networks based on AHP," Physica A: Statistical Mechanics and its Applications, vol. 479, pp. 422-436, 2017.

[10] F. Kazmi, W. H. Butt, and A. Saeed, "Evaluation of Role Detection Approaches in Terrorist Networks. in Proceedings of the 2018 2nd International Conference on Management Engineering," Software Engineering and Service Sciences, 2018.

[11] M. Doo and L. Liu, "Extracting top-k most influential nodes by activity analysis," In Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), IEEE, 2014.

[12] M. Nouh and J. R. Nurse, "Identifying key-players in online activist groups on the facebook social network," In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE, 2015.

[13] P. Hu and T. Mei, "Ranking influential nodes in complex networks with structural holes," Physica A: Statistical Mechanics and its Applications, vol. 490, pp. 624-631, 2018.

[14] R. C. Gunasekara, "Identification of key players in networks using multi-objective optimization and its applications," 2016.

[15] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks. Reviews of modern physics," vol. 74, no. 1, p. 47, 2002.

[16] A. Bickle, "The k-cores of a graph," 2010.

[17] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," Physical review letters, vol. 87, no. 19, p. 198701, 2001.

[18] A. N. Langville and C. D. Meyer, "Google's PageRank and beyond: The science of search engine rankings," Princeton university press, 2011.

[19] L. C. Freeman, "A set of measures of centrality based on betweenness," Sociometry, pp. 35-41, 1977.

[20] P. Bonacich, "Some unique properties of eigenvector centrality," Social networks, vol. 29, no. 4, pp. 555-564, 2007.

[21] Z.-M. Han, et al., "Ranking key nodes in complex networks by considering structural holes, Acta physica sinica," vol. 64, no. 5, p. 058902, 2015.

[22] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," 1998.

[23] Dataset: http://www.delicious.com, ACM, New York, NY, USA, 2011.

Journal of Electronical & Cyber Defence
Vol. 8, No. 3, 2020, Serial No. 31

Identifying Influential Nodes in Social Networks by Integrating the Centrality Method and Node Activity

A. Karimi*, E. Bastami,M. Nemati, M. Saleh Esfehani,

*Imam Hossein Comprehensive University

(Received: 14/01/2019, Accepted: 01/02/2020)

ABSTRACT

Nowadays, social networks have become a strong tool among researchers in addition to their social functions. This tool has many applications in identifying crimes, criminals and terrorists, solving epidemic problems, successful marketing and other topics in various fields. The researchers are using the influence maximization (IM) to achieve these goals. The task of maximization is to identify the influential nodes that are known as the seed nodes. It is a strategy to achieve the maximum information diffusion or minimum epidemy with minimal cost. Since maximization is an NP-hard problem, researchers are looking for ways to reduce the complexity and acceptable identification accuracy by identifying influential nodes. Therefore, to overcome the complexity and increase the identification accuracy, in this research a new method with activity-centrality combination is proposed. In this approach, to extract nodes by the centrality method a total constraint is constructed on the network graph in order to proceed to the local nodes extracted from the node activity analysis. The results of analyzing the activity of each node are combined with its centrality method score which ultimately leads to the identification of influential nodes. The proposed method is compared with other methods such as PageRank and Closeness Centrality methods, and the evaluation results show that whilst having a lower complexity, the proposed method is better than both in terms of accuracy. In the future, the concepts of repetitive scoring can be used to further enhance the accuracy of the activity analysis phase.

Keywords: Influence, Influential Nodes, Influence Maximization (IM), Centrality, Activity

*Corresponding Author Email: a.karimi@ihu.ac.ir