ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

ارائه یک الگوریتم سه مرحله‌ای برای تشخیص جوامع مبتنی بر گره‌های ناهنجار مرکزی

سعید تقوی افشرد (نویسنده مسئول)[۱]، محمدحسن محمدباقرزاده[۲]

[۱] عضو هیئت علمی موسسه آموزش عالی رشدیه، شهر تبریز taghavi.afshard@gmail.com

[۲] دانشجوی کارشناسی ارشد موسسه آموزش عالی رشدیه، شهر تبریز mhb.zadeh@gmail

چکیده

تشخیص اجتماع یـک مسـاله پیچیده و مهـم در مطالعـه سـاختار شبکه و فهم ویژگی‌ها، برای استخراج اطلاعات شـبکه اسـت. این مسئله، به کرات مورد توجه پژوهشگران مختلف قرار گرفته است؛ اما هم‌چنان تشخیص جوامع با دقت و کیفیت بالا در شبکه‌های پیچیده، یکی از مسائل باز این حوزه می‌باشد. به طور کلی، تشخیص درست جوامع به شناسایی دقیق گره‌های مرکزی وابسته است. تاکنون در تشخیص گره‌های مرکزی، بیشتر موارد گره‌های با درجه بالا و فاصله بیشتر مد نظر گرفته شده است اما در این پژوهش برای اولین بار، گره‌های مرکزی در شبکه‌های پیچیده بر اساس ناهنجاری گره‌ها شناسایی شدند. منظور از گره‌های ناهنجار، گره‌هایی هستند که رفتار آن‌ها، از رفتار غالب گره‌های شبکه دور بوده و می‌تواند متعلق به یک گره تاثیرگذار در شبکه یا یک اسپمر باشد. در این پژوهش، در گام اول بعد از شناسایی گره‌های مرکزی با بررسی میزان ناهنجاری گره، در گام دوم از الگوریتم انتشار برچسب، برای دسته‌بندی اجتماع‌ها استفاده شده و نهایتاً در گام سوم، پیش اجتماع‌های تشخیص داده شده در گام دوم، با هم ترکیب می‌شوند. آزمایش‌ها و شبیه‌سازی‌ها در محیط متلب نشان می‌دهد که روش پیشنهادی، از نرخ صحت، معیار ماژولاریتی و اطلاعات متقابل نرمال شده بهبود یافته‌تری نسبت به پژوهش‌های پیشین برخوردار است.

واژه‌های کلیدی

تشخیص اجتماع، گره‌های مرکزی، گره‌های ناهنجار، انتشار برچسب، الگوریتم آکوگلو

۱. مقدمه

ما در عصر اطلاعات زندگی می‌کنیم؛ ترابایت‌ها یا پتابایت‌ها داده، هر روز در بعدهای مختلف زندگی ما، تولید می‌شوند. رشد سریع و انفجاری داده، در دسترس بودن گسترده منابع تولید داده مانند شبکه‌های اجتماعی آنلاین، عصر ما را به عصر استخراج اطلاعات و دانش ساخت‌یافته پشت این داده‌ها، تبدیل کرده است [۱]. یکی از انواع داده‌ها، گراف‌ها هستند. از ساختار داده گراف، برای مدل‌سازی شبکه‌های موجود در جهان واقعی استفاده می‌شود. این مدل‌سازی، امکان توصیف سیستم‌های پیچیده‌ی جهان واقعی مانند شبکه-های اجتماعی، شبکه‌ی اینترنت، شبکه‌ی پروتئین و شبکه‌های عصبی را امکان‌پذیر می‌سازد [۷, ۳]. از سوی دیگر آنالیز شبکه‌ها، در حوزه‌های مختلف مانند بیوتکنولوژی، علوم کامپیوتر، علوم اجتماعی و ... دارای کاربردهای فراوان است؛ برای مثال آنالیز ساختار بخش‌های مختلف مغز، دارای وجوه تشابه فراوانی با ساختار گراف‌های شبکه‌های اجتماعی مانند فیس‌بوک دارد [۴]. بنابراین، این گراف‌ها، گراف‌های تصادفی نیستند و الگوهای حاکم بر آن‌ها، قابل تحلیل و کشف است. یکی از این الگوها، پیدا کردن زیرگراف‌هایی است که اعضای آن زیرگراف با هم ارتباط متصل‌تر دارند در حالی که با سایر اعضای گراف ارتباطشان کمتر است. این نوع زیرگراف‌ها، اصطلاحاً جامعه یا اجتماع در گراف نامیده می‌شوند. پیدا کردن چنین جوامعی می‌تواند اطلاعات پنهان در مورد جامعه، ساختار جوامع، افراد تاثیرگذار، روابط بین موجودیت‌ها و غیره را در اختیار ما قرار دهد که در نهایت از این اطلاعات می‌توان به منظور حفظ امنیت اعضای جامعه، اهداف سیاسی، تحلیلهای تجاری و ... استفاده کرد [۶, ۵].

۱،۱. بیان مسئله

در سال‌های اخیر، تشخیص اجتماع در شبکه‌های پیچیده، به کرات مورد توجه قرار گرفته است. به طور کلی، می‌توان این روش‌ها را به دو دسته سراسری و محلی تقسیم‌بندی کرد. روش‌های سراسری در تشخیص اجتماع، کل ساختار گراف را در نظر می‌گیرند. این روش‌ها پیچیدگی زمانی بالا دارند و معمولا الگوریتم‌های مفیدی نیستند. زیرا مجموعه داده‌های واقعی، مانند مجموعه داده‌های توئیتر و فیسبوک، مجموعه داده‌های میلیونی هستند. اما مجموعه روش‌های محلی، بر اساس معیارهای ساختاری محلی گراف به تشخیص اجتماع می‌پردازند؛ این مجموعه روش‌ها، معمولا پیچیدگی زمانی نزدیک به خطی دارند؛ اما اندازه داده‌های مربوط به شبکه‌های پیچیده، رشد روز افزونی دارند و تشخیص احتماع حتی با یک الگوریتم با پیچیدگی زمانی نزدیک به خطی، بر روی یک کامپیوتر واحد امکان‌پذیر نیست. بنابراین ارائه الگوریتم‌های با پیچیدگی زمانی پایین‌تر و نرخ صحت بالا برای تشخیص اجتماع، بسیار موثر واقع می‌شود. تجزیه و تحلیل شبکه کاربردهای زیادی در زمینه بیوتکنولوژی، علوم کامپیوتر، علوم انسانی، علوم طبیعی و اجتماعی و سیستم های مختلف مهندسی دارد. به عنوان مثال، ساختار شبکه ای ارتباطات در بخش های مختلف مغز انسان شباهت‌های ساختاری جالبی با نمودار فیس بوک دارد. همچنین الگوی توزیع بیماری مشابه الگوی انتشار شایعه در شبکه های اجتماعی است. با گسترش سریع اینترنت و شبکه های اجتماعی مجازی، اهمیت شبکه ها در زندگی روزمره بیش از پیش احساس می شود. در عصر داده های بزرگ، نمودارهای اجتماعی در ابعاد و دامنه های وسیع تری مورد بحث قرار می گیرند. شبکه مجموعه‌ای از گره‌ها و پیوندهای متقابل آن‌ها است و بر اساس ماهیت پدیده‌ها را می‌توان به نمودارهای وزن‌دار یا بدون وزن و جهت‌دار یا غیر جهت‌دار تقسیم کرد. جالب است که شبکه های دنیای واقعی مختلف ویژگی های ساختاری و دینامیکی مشابهی دارند [۴] .

۲،۱. اهمیت و ضرورت انجام تحقیق

تشخیص اجتماع یک مساله سخت و مهم در مطالعه ساختار شبکه و فهم ویژگی‌ها، برای استخراج اطلاعات شبکه است [۱]. می‌توان از تشخیص اجتماع در شبکه‌های پیچیده، تحت عنوان خوشه‌بندی این شبکه‌ها نام برد. یعنی در تشخیص اجتماع، نهایتاً گره‌های با ویژگی یکسان دور هم جمع می‌گردند. با شناسایی این اجتماع‌ها و افراد متعلق به چند اجتماع، می‌توان سرعت انتشار اطلاعات در شبکه را افزایش داد [۱, ۸]. هم‌چنین بررسی چند جامعه، می‌تواند منجر به دست آوردن دید کلی نسبت به کل شبکه گردد و می‌توان روابط بین گره‌های شبکه پیچیده را تحلیل کرد. انتخاب معیار مناسب برای سنجش میزان تشابه و هم‌چنین نحوه تعیین اهمیت گره در کارایی الگوریتم تشخیص اجتماع، بسیار موثر است. در پژوهش‌های پیشین، با در نظر گرفتن سناریوهای مختلف اجتماعی، مانند این‌که "افراد دور گره‌های مهم در شبکه جمع می‌گردند" [۳]، تلاش می‌شود تا گره‌های با درجه بالا یا گره‌هایی با بیش‌ترین فاصله را [۳] که

اصطلاحاً گره‌های پراهمیت می‌نامند، شناسایی کنند. اما اکثریت الگوریتم‌های ارائه شده با مشکل پیچیدگی زمان بالا و دقت پایین روبرو هستند. از سوی دیگر، با در نظر گرفتن این واقعیت که حجم داده‌ها، به‌صورت مداوم در حال افزایش بوده و جامعه امروزی نیازمند پاسخ سریع در پردازش داده‌ها است، باید الگوریتم‌هایی با پیچیدگی زمانی کم و دقت بالا ارائه گردد.

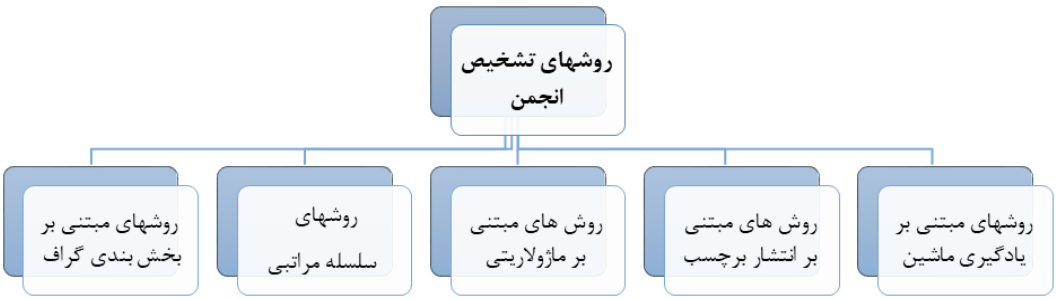
کیفیت جامعه تشخیص داده شده، به شدت به کیفیت گره‌های مرکزی شناسایی شده وابسته است. در روش‌های پیشین، همواره گره‌های با درجه بالا، فاصله بیشتر و غیره را به عنوان گره‌های مرکزی در نظر می‌گرفتند، در حالی‌که اکثر روش‌های موجود دارای معایبی مانند دقت پایین، مقیاس‌پذیری ضعیف و بی‌ثباتی در نتایج هستند [۴]. در این مقاله برای اولین بار، تعریف گره‌های پراهمیت در شبکه‌های پیچیده بر اساس گره‌های ناهنجار خواهد بود. یعنی گره‌هایی به عنوان گره مرکزی شناخته خواهند شد که رفتار آن‌ها، از رفتار غالب گره‌های شبکه دور است.

۲. مروری بر تحقیقات پیشین

مفهوم شبکه اجتماعی بعد از به شهرت رسیدن وب‌سایت‌هایی مانند فیس‌بوک، اینستاگرام، توئیتر و ... مورد توجه قرار گرفت. دو ویژگی اصلی شبکه‌های اجتماعی، وجود موجودیت‌های متعدد و ارتباطات بین این موجودیت‌ها است. در این شبکه‌ها، موجودیت‌ها می‌توانند افراد و ارتباطات بین این موجودیت‌ها می‌توانند دوستی بین افراد باشند. هم‌چنین موجودیت‌ها می‌توانند وب‌سایت‌ها و سازمان‌ها و ارتباطات بین آن‌ها می‌توانند تعاملات یا خرید و فروش بین این موجودیت‌ها باشند [۹, ۱۰]. به طور کلی از ساختار داده گراف، برای مدل‌سازی شبکه‌های اجتماعی استفاده می‌شود. موجودیت‌ها در این مدل‌سازی به عنوان رئوس گراف و ارتباطات بین موجودیت‌ها به عنوان یال‌های گراف شناخته می‌شوند. از این مدل‌سازی در پدیده‌های دنیای واقعی تحت عنوان شبکه‌های پیچیده نام برده می‌شود که در ادامه شرح داده خواهد شد [۱۰].

یک گراف، شامل مجموعه‌ای ناتهی از رئوس و یال‌ها است. در علوم رایانه اغلب از این ساختار داده برای مدل‌سازی شبکه‌های موجود در جهان واقعی مانند شبکه‌های اجتماعی، شبکه‌ی اینترنت، شبکه‌ی پروتئین، شبکه‌های عصبی و غیره استفاده می‌شود [۱۱]. ساختار داده گراف استفاده شده برای مدل‌سازی پدیده‌های دنیا واقعی، شبکه‌های پیچیده نامیده می‌شود. این دسته شبکه‌ها از لحاظ توپولوژیکی دارای ویژگی‌هایی مانند بالا بودن ضریب خوشه‌بندی، پیروی درجات گره‌ها از قانون توانی و ... هستند (در ادامه توضیح داده می‌شود)؛ در حالی‌که دسته دیگر گراف‌ها یعنی گراف‌های تصادفی دارای درجات مشابه و توزیع یال همگون هستند [۱۲].

الگوریتم‌های کشف جوامع، طیف گسترده‌ای دارند. در شکل ۱ یک بخش‌بندی از این روش‌ها، آورده شده است.



شکل ۱. بخش‌بندی پژوهش‌های پیشین در حوزه تشخیص اجتماع در شبکه‌های اجتماعی

۱،۲. روش‌های سلسله مراتبی [۱۸]

خوشه‌بندی سلسله مراتبی یکی از رایج‌ترین روش‌های تشخیص اجتماع برای شبکه‌های اجتماعی است. در این روش مرتباً یال‌های بین دو گره به صورت تصادفی انتخاب می‌شوند تا در نهایت، خوشه‌ها تشکیل شوند، اما این روش دارای معایبی است. بـدین ترتیـب کـه در مـورد گره‌های واسط بین اجتماع‌ها، احتمال ترکیب وجود دارد در حالیکه این دو گره در دو خوشه کاملاً متمایز هستند. برای حل این مشکل در خوشه‌بندی با الگوریتم سلسله مراتبی، معمولاً چندین مرتبه این الگوریتم را اجرا می‌کنند و تکراری که بیش‌تـرین خوشـه‌های منسـجم را ارائه دهد، به عنوان پاسخ برگردانده می‌شود. البته روش دیگری نیز هست که در آن، اهمیت گره با استفاده از معیارهایی مانند نیومن و یـا معیار ماژولاریتی سنجیده می‌شود و خوشه‌بندی سلسله مراتبی با توجه به این معیارها انجام می‌گیرد. برای مثال در این تحقیق، با استفاده

از معیار سورنسن، گره‌های مرکزی شناسایی شده و با استفاده از الگوریتم خوشه‌بندی سلسله مراتبی، گره‌های اطراف گره مرکزی بررسی می‌شوند، با این تفاوت که انتخاب یال‌ها تصادفی نخواهد بود، بلکه بر اساس معیار فاصله، اهمیت یال‌ها محاسبه می‌شود و در نهایت اگر اندازه خوشه‌های به دست آمده کم‌تر از حد آستانه باشد، خوشه‌ها ترکیب خواهند شد.

۲،۲. روش بخش‌بندی گراف [۲۴]

در این روش سعی می‌شود گراف به k قسمت از پیش تعیین شده با اندازه و تعداد گره‌های معلوم تقسیم‌بندی شود، به طوری‌که تعداد یال‌های بین جوامع کمینه باشد [۲۳]. به طورکلی، روش‌های بخش‌بندی گراف در دو دسته کلی جای می‌گیرند؛ روش‌های بخش‌بندی طیفی که از ماتریس لاپلاسین گراف و بردار ویژه آن برای بخش‌بندی استفاده می‌کند؛ و الگوریتم کرنیگان-لین که تلاش می‌کند ساختار جامعه را بر روی بخش‌بندی اولیه گراف به شکل حریصانه بهینه کند [۲۴]. تابع سود در این الگوریتم برابر با مجموع یال‌های درون جامعه منهای مجموع یال‌های بین جوامع است. ضعف اصلی الگوریتم کرنیگان-لین، وابسته بودن به اندازه و تعداد اولیه جوامع در مرحله ابتدایی الگوریتم است. این مشکل به خصوص در مجموعه داده‌های واقعی، رخ می‌نماید. بعدها، توسعه‌هایی در این الگوریتم داده شد تا زمانی‌که تعداد و اندازه جوامع مشخص نباشد نیز با جابه‌جایی یک گره واحد به طور همزمان به سایر جوامع، تشخیص جوامع انجام گردد اما این الگوریتم هم‌چنان دارای کاستی‌هایی است که معمولاً در زمان و تشخیص جوامع، ضعیف عمل می‌کند.

۳،۲. روش‌های مبتنی بر ماژولاریتی [۷]

اگر چه الگوریتم‌های زیادی برای خوشه‌بندی ارائه شده است، اما این الگوریتم‌ها بدون معیاری برای محاسبه کیفیت اجتماع‌ها، ناقص هستند. در سال ۲۰۰۴، نیومن-گیروان [۷] معیار ماژولاریتی را در این راستا معرفی کردند که در آن ماژولاریتی برابر با تعداد یال‌ها در آن اجتماع منهای تعداد یال مورد نظر مطابق قانون توانی است و در بازه ۰،۵- تا ۱ قرار دارد [۲۶]. روش‌های مبتنی بر ماژولاریتی، یک روش حریصانه در پیش می‌گیرند که تنها هدف آن‌ها، افزایش میزان ماژولاریتی حاصل از ادغام و یا انتقال گره‌ها در جوامع است. از معایب این روش، گیر کردن در بهینه محلی و اهمیت ندادن این دسته روش‌ها به نقش آن گره در جوامع است؛ برای مثال، مرکزی یا حاشیه‌ای بودن آن گره در روشهای مبتنی بر ماژولاریتی اهمیتی ندارد. برای حل این مشکل در پژوهش ابتدا گره‌های مرکزی بر اساس یک معیار شباهت مشخص می‌شود و در نهایت بر اساس معیار ماژولاریتی، پیش‌اجتماع‌ها ادغام می‌گردند.

۴،۲. روش‌های مبتنی بر انتشار برچسب [۱۹]

این روش یک الگوریتم محلی برای شناسایی اجتماعات ارائه می‌کند، به طوریکه از ساختار محلی و توپولوژیکی خود شبکه کمک می‌گیرد. در الگوریتم انتشار برچسب، در ابتدا هر راس با یک مقدار واحد مقداردهی اولیه می‌شود و در هر چرخه از تکرار الگوریتم، به هر راس برچسبی اختصاص می‌یابد به گونه‌ای که اکثر همسایگان آن راس نیز به آن برچسب اختصاص یافته‌اند. الگوریتم انتشار برچسب مطابق با توضیحات مذکور ادامه می‌یابد تا یک مجموعه متراکم با برچسب یکسان شکل گیرد. در پایان الگوریتم، رئوس با برچسب‌های یکسان با یکدیگر ادغام می‌شوند. لازم به ذکر است که در این الگوریتم نیازی به اطلاع از تعداد و اندازه اجتماعات از قبل نیست و در پایان الگوریتم، تعداد و اندازه اجتماعات بدست می‌آیند [۲۷].

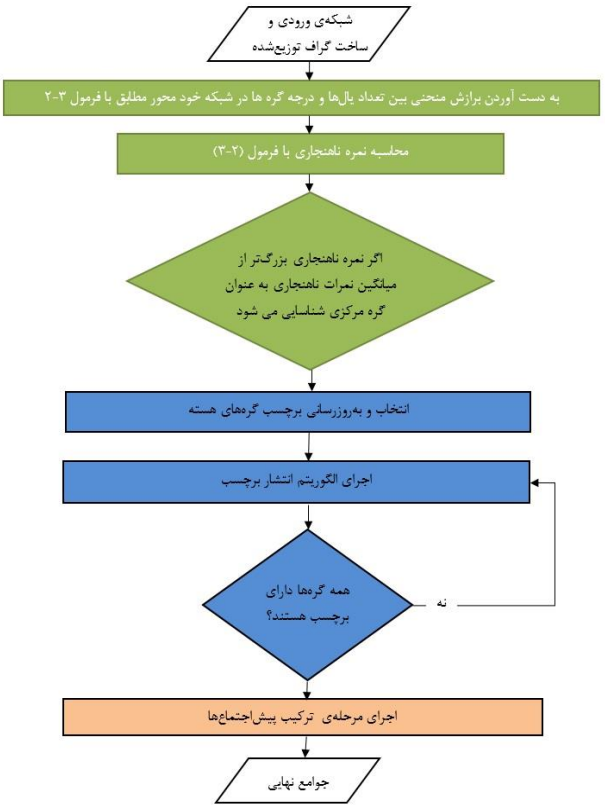
۵،۲. روشهای مبتنی بر یادگیری ماشین [۲۰]

این حوزه در تشخیص اجتماع، حوزه‌ای نو پدید با چشم‌اندازی روشن به حساب می‌آید. توانایی یادگیری عمیق در آموزش ویژگی‌های داده‌های با ابعاد بالا، این امکان را فراهم می‌آورد که این دسته از روش‌ها، در یادگیری الگوهای گراف مانند همسایه‌ها، گره‌ها و غیره نیز موفق عمل کنند [۲۸]. در پژوهش صالحی و همکارانش [۲]، از یک رویکرد سه گامی، با استفاده از یادگیری عمیق به شناسایی اجتماع‌ها در شبکه‌های اجتماعی می‌پردازند. گام‌های این پژوهش بدین صورت است که در گام اول یعنی گشت تصادفی یک مکانیزم برای نمایش گره‌های شبکه به عنوان ورودی قسمت یادگیری عمیق معرفی می‌شود، یعنی در این گام، با شروع از گره i، احتمال مشاهده دیگر گره‌ها به عنوان گام بعدی محاسبه می‌شود و به عنوان ورودی به شبکه یادگیری عمیق داده می‌شود. به منظور حل توسط شبکه یادگیری عمیق نیاز است تا وزن‌های مناسب هر لایه و تابع تحریک و بهینه‌ساز تعیین گردند. در این پژوهش از تابع Relu و تابع گرادیان کاهشی

تصادفی به عنوان تابع بهینه‌ساز استفاده شده است و در نهایت بردار ویژگی‌های استخراج شده با اسـتفاده از دو الگـوریتم خوشـه‌بندی k-mean و DBSCAN خوشه بندی شده‌اند [۲].

۳. روش پیشنهادی

این روش دارای سه گام مهم است که در زیربخش‌های مربوط به هر گام، به جزئیات پرداخته شده است. در گام اول، به شناسایی گره‌های مرکزی خواهیم پرداخت. در گام دوم از الگوریتم انتشار برچسب، استفاده خواهیم کرد و نهایتاً در گام سوم، پـیش اجتماع‌هـای تشـخیص داده شده در گام دوم، با هم ترکیب خواهند شد. در شکل ۲ فلوچارت کلی روش پیشنهادی آورده شده است.



شکل ۲. فلوچارت روش پیشنهادی

گره‌های هسته نقش بسیار اساسی در شکل‌گیری درست جوامع می‌توانند داشته باشند، زیرا اگر یک ترتیب صـحیح از گـره‌هـا بـه عنـوان گره‌های هسته انتخاب شود، در مراحل بعدی الگوریتم برچسب جوامع با صحت و اطمینان بیشتری انتخاب خواهـد شـد و گـره‌هـای کـم اهمیت‌تر برچسب درستی دریافت خواهند کرد؛ به طور معمول، روند تشخیص این گره‌ها در شبکه‌های پیچیده بر اساس اهمیت گره‌هـا در این شبکه‌هاست. تاکنون تعریف این اهمیت بر اساس مفاهیمی مانند درجه بالا، فاصـله بیش‌تـر، معیـار سورنسـن بالـاتر و ... بـود. در ایـن پژوهش برای اولین بار، تعریف گره‌های پراهمیت در شبکه‌های پیچیده بر اساس گره‌های ناهنجار خواهد بود. یعنی گره‌هایی، به عنوان گره مرکزی شناخته خواهند شد که رفتار آن‌ها، از رفتار غالب گره‌های شبکه دور است. دو دسته از گره‌ها، متفاوت با روال عادی هسـتند یـک گره‌های تاثیرگذار (اینفلوئنسرها) و دوم گره‌هایی که منزوی، خرابکار و خاص هستند. اگر در تشخیص گره‌های مرکزی، هر دو دسـته ایـن گره‌ها در نظر گرفته شود؛ یعنی هم گره‌های با درجه بالا و هم گره‌هایی که دوره خاص خود دارند؛ احتمال افزایش تشـخیص اجتماع‌هـای باکیفیت افزایش می‌یابد؛ زیرا در مورد گره‌های اینفلوئنسر، فلسفه معمول تشخیص اجتماع صـادق اسـت؛ امـا اگـر در کنـار ایـن گره‌هـای اینفولئنسر، گره‌های خاص مانند گره‌های خلافکار نیز شناخته شوند، مرز تشخیص گره‌ها، بیش‌تر ممیز خواهد بود.

۱،۳. گام اول: تشخیص گره‌های مرکزی

گره‌های مرکزی، مجموعه‌ای از گره‌های مهم و با اهمیت هستند که ویژگی‌های متفاوتی نسبت به سایر گره‌های کم اهمیت دارنـد. در ایـن پژوهش برای اولین بار، تعریف گره‌های پراهمیت در شبکه‌های پیچیده بر اساس گره‌های ناهنجار خواهد بود و گره‌هایی که رفتار آن‌هـا، از رفتار غالب گره‌های شبکه دور باشد، به عنوان گره‌های مرکزی شناخته می‌شود؛ این گره‌های ناهنجار می‌توانند متعلق بـه یـک سـلبریتی، اینفلئونسر یا یک کلاهبردار مالی باشند.

در این پژوهش، شناسایی گره‌های مرکزی بر اساس روش آکوگلو [۴۴] که یکی از پژوهش‌هـای پایـه تشـخیص ناهنجـاری اسـت، انجـام خواهد گرفت. در پژوهش آکوگلو، یک دایره اطراف یک گره تعریف می‌گردد که شامل خود گره، همسایه‌های مستقیم و تمـامی ارتباطـات ما بین این گره‌ها است. این دایره اصطلاحاً شبکه خودمحور نامیده می‌شود و نشان می‌دهند که بین تعداد یال‌ها و رأس‌هـا در یـک شـبکه خودمحور، رابطه نمایی مطابق رابطه (۱) برقرار است:

$$E_i \varphi N_i^{\alpha} \qquad 1<=\alpha<=2 \qquad (۱)$$

که در آن E_i، تعداد یال‌های شبکه خودمحور گره i و N_i تعداد گره‌های شبکه خودمحور گره i است.

در این روش برازش منحنی، بین تعداد یال‌ها در شبکه خودمحور و درجه گره‌ها به دست می‌آید و برای هر گره i در رگرسیون توانی نمره «دور بودن» را مطابق با رابطه (۲) محاسبه می‌شود:

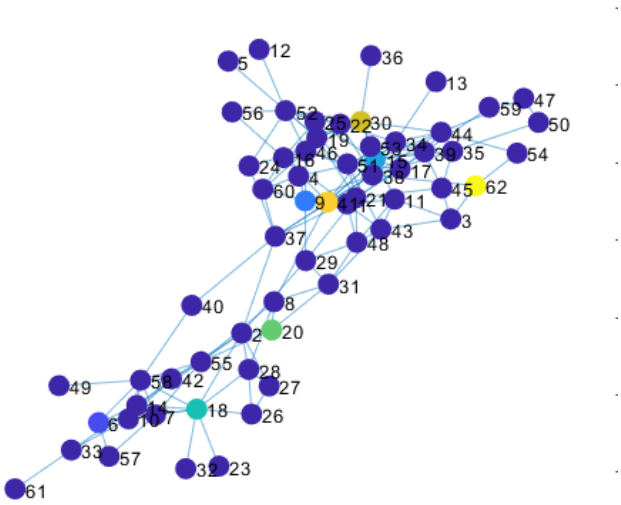
$$\text{outline}(i) = \frac{\max(y_i . C_{x_i}^{\theta})}{min(y_i . C_{x_i}^{\theta})} * \log|y_i - C_{x_i}^{\theta}| + 1 \qquad (۲)$$

که در ان y_i مقدار حقیقی معیار، $C_{x_i}^{\theta}$ مقدار محاسبه شده معیـار توسـط رگرسـیون، $\max(y_i . C_{x_i}^{\theta})$ مقـدار محاسـبه شـده توسـط بـازه حداکثری رگرسیون (رگرسیون در دو بازه حداکثری و حداقلی تخمین زده می‌شود) و $min(y_i . C_{x_i}^{\theta})$ مقدار محاسبه شـده توسـط بـازه‌ی حداقلی رگرسیون می‌باشد.

در اصل رابطه (۲) فاصله تا خط رگرسیون به دست آمده را تخمین می‌زند و در نهایت از الگوریتم عامل دور افتاده محلی بـرای تشـخیص ناهنجاری استفاده شده و گره‌ها بر اساس نمره ناهنجاری مرتب می‌شوند. N نقطه که نمره ناهنجاری آن‌ها از میانگین نمره ناهنجاری کـل گره‌ها بیش‌تر باشد به عنوان گره‌های مرکزی شناسایی خواهند شد.

الگوریتم عامل دورافتاده محلی بر مبنای مفهوم چگالی محلی بنا شده و در آن محلی بـودن بـر اسـاس k نزدیـک‌ترین همسـایگی تعیـین می‌شود که فاصله آن‌ها برای تخمین چگالی مورد استفاده قرار می‌گیرد. با مقایسه چگالی محلی یک شی بـا چگالی‌هـای همسـایه‌های آن می‌توان نواحی دارای چگالی مشابه و نقاطی که اساسا چگالی کمتری نسبت به همسایه‌های خود دارند را تعیین کرد. این موارد بـه عنـوان دورافتادگی (داده پرت) در نظر گرفته می‌شوند. چگالی محلی به وسـیله فاصـله معمـولی کـه یـک نقطـه داده توسـط همسـایه‌های خـود «دسترسی‌پذیر» است، تخمین زده می‌شود.

با یک مثال بر روی مجموعه داده واقعی دلفین گره‌های مرکزی شناسایی می‌شوند. در شکل ۳ گراف ارتباطـات ایـن دیتاسـت و گره‌هـای مرکزی بر اساس نمره ناهنجاری آورده شده است. با استفاده از این روش، هشت گره مرکزی مطابق با جدول ۱ شناسایی می‌شود. حال آن که با استفاده از روش‌هایی مانند بیش‌ترین درجه و فاصله، ۴۱ گره مرکزی شناسایی خواهد شد.



شکل ۳. گره‌های مرکزی شناسایی شده بر اساس روش اکوگلو و نمره ناهنجاری

در این دیتاست، گره‌های {۶۲،۴۱،۳۰،۲۰،۱۸،۱۵،۹،۶} به عنوان گره‌های مرکزی شناسایی شدند. در جدول ۱ مشخصات این نقـاط آورده شده است.

جدول ۱. مشخصات گره‌های مرکزی

گره	تعداد یال در شبکه خودمحور	تعداد گره در شبکه خودمحور
۶	۷	۴
۹	۱۰	۶
۱۵	۲۹	۱۲
۱۸	۱۷	۹
۲۰	۷	۴
۳۰	۱۸	۹
۴۱	۱۵	۸
۶۲	۳	۳

۲،۳. گام دوم: انتشار برچسب

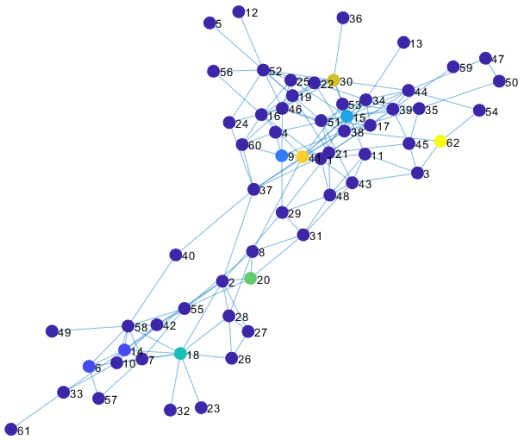
الگوریتم انتشار برچسب سنتی به این صورت است که در این الگوریتم، هر گره را با یک برچسب منحصر به فرد اولیه مقداردهی کرده و به برچسب اجازه انتشار از طریق شبکه را می‌دهد؛ یک گره به تصادف انتخاب می‌شود، سپس گره انتخابی، برچسب خود را مطابق با برچسب با بالاترین تعداد رخداد در میان همسایگان، قرار خواهند داد. اگر تعداد رخداد برچسب در بین همسـایگان بـه تعـداد مسـاوی باشـد یـک برچسب به تصادف انتخاب خواهد شد.

اما ما در این پژوهش، با تغییراتی در الگوریتم انتشار برچسب، از یک الگـوریتم جدیـد اسـتفاده خـواهیم کـرد. بـدین صـورت کـه بعـد از شناسایی گره‌های مرکزی، هر گره مرکزی برچسب یا رنگ متمایز بـه خـود می‌گیـرد؛ سـپس در گره‌هـای همسـایگان گره‌هـای مرکـزی، بیشترین میزان شباهت با امتیاز آنومالی محاسبه می‌شود، برای مثال در دیتاست مثال قبل معیارهای شباهت محاسبه شده برای چند گره با بشترین امتیاز از قرار جدول ۲ است.

جدول ۲. محاسبه امتیاز آنومالی دیتاست دلفین

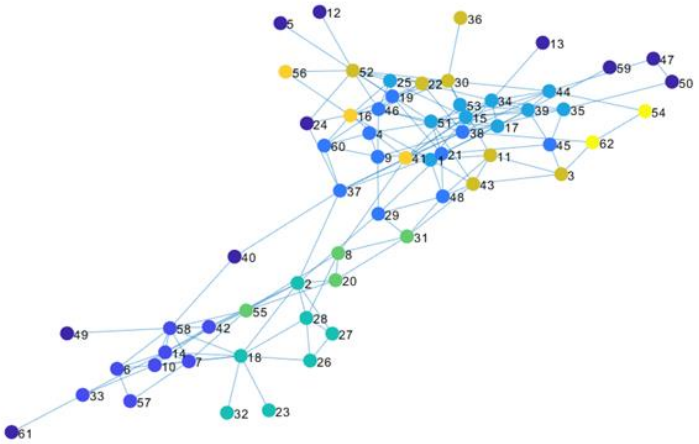
گره	امتیاز
۱۴	۱۸۰۰
۵۷	۱۷۰۰
۵۸	۱۵۰۰
۱۰	۱۴۰۰

و در نهایت مطابق شکل گره ۱۴ به عنوان برنده اولین مرحله برچسب‌زنی خواهد بود.



شکل ۴. گره ۱۴ برنده اولین مرحله انتشار برچسب بر اساس نمره شباهت ناهنجاری

دوباره همین روال را، برای گره‌های بدون برچسب تکرار می‌کنیم، با این تفاوت که الان گره ۱۴ برچسب آبی دارد. با چندین مرحله تکـرار، در نهایت تمامی گره‌ها، مطابق شکل ۵ برچسب می‌خوردند. بعد از این مرحله، وارد گام سوم یعنی ترکیب پیش اجتماع‌ها می‌شویم.



شکل ۵. برچسب‌های اولیه گره‌ها در مجموعه دیتاست دلفین

۳،۳. گام سوم: ترکیب پیش‌اجتماع‌ها بر اساس بهینه‌سازی معیار ماژولاریتی

رویکرد گام سوم بدین صورت است که پیش‌اجتماع‌ها را با هم ترکیب شده و معیار ماژولـاریتی محاسـبه می‌شـود و تـا جـایی کـه معیـار ماژولاریتی افزایش پیدا کند، این روند ادامه می‌یابد. هدف، به دست آوردن بهترین مقدار، برای معیار ماژولاریتی است.

ماژولاریتی، معیاری برای ارزیابی کیفیت خوشه‌ها در گراف است. این معیار به دلیل سادگی و کارایی، جزء معیارهای پرکاربرد در ارزیابی کیفیت خوشه‌ها است. متد لوین یک روش نسبتا سریع برای پیدا کردن گروه‌ها در شبکه‌های بزرگ است. این روش با استفاده از روش‌های حریصانه بر روی ماژولاریتی، آن را بهینه کرده و به خوشه‌بندی راس‌های گراف می‌پردازد. این بهینه‌سازی در دو مرحله انجام می‌شود:

مرحله ۱- با بهینه‌سازی محلی، متد به دنبال گروه‌های کوچک می‌گردد.

مرحله ۲- سپس با ادغام گروه‌های کوچک که توانایی ایجاد گروه‌های بزرگ‌تر را دارند، خوشه‌بندی را ادامه می‌دهد. این مراحل مرتباً تکرار می‌شود تا مقدار بیشینه ماژولاریتی حاصل شود.

معیار ماژولاریتی در اصل، تفاضل بین تعداد یال‌ها در خوشه از تعداد یال‌های مورد نظر متناسب با تعداد گره‌های خوشه مطابق قانون توانی را محاسبه می‌کند و همواره در بازه ۰٫۵- تا ۱ قرار دارد.

۴. آزمایشات و نتایج

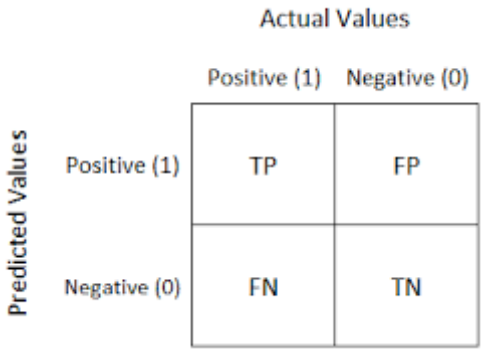
در این بخش، به ارزیابی دقت و کارایی رویکرد معرفی شده برای تشخیص اجتماع در شبکه های پیچیده خواهیم پرداخت. این ارزیابی‌ها با استفاده از معیارهایی خواهد بود که در بخش بعد به تفصیل توضیح داده خواهد شد. آزمایشات و پیاده‌سازی‌ها در محیط متلب ۲۰۲۱ و بر روی سیستمی با مشخصات سیستم عامل ویندوز ۱۰، حافظه ۸ گیگا بایت و پردازنده Intel Core i5 انجام گرفته است.

۱،۴. معیارهای ارزیابی

در این پژوهش، برای مقایسه¬ی دقت و عملکرد رویکرد پیشنهادی، از معیارهای F-measure، نرخ صحت، ماژولاریتی و معیار NMI استفاده خواهیم کرد.

۱،۱،۴. محاسبه نرخ صحت

معیارهای نرخ صحت و F-measure، معیارها مبتنی بر ماتریس پریشانی هستند. ماتریس پریشانی یک ابزار مناسب برای تحلیل کارکرد مدل تمایزکننده است.



شکل ۶. ماتریس پریشانی دو کلاسه با دو کلاس صفر و یک

که در آن TP و TN نشان‌دهنده زمانی است که مدل به درستی پیش‌بینی کرده و FP و FN هم نشان‌دهنده زمانی است که مدل به اشتباه دسته‌بندی شده است. نرخ صحت، نشان دهنده این است که چه قدر از داده‌ها برچسب درست خورده‌اند و F-measure از درصد داده‌های به درستی پیش‌بینی شده خبر می‌دهد و محاسبه آن نیازمنده محاسبه نرخ فراخوانی است. این مقدار طبق فرمول (۳) محاسبه خواهند شد:

$$\text{نرخ صحت} = \frac{TP + TN}{TP + FP + FN} \quad (۳)$$

۲،۱،۴. محاسبه ماژولاریتی

در سال ۲۰۰۴، نیومن-گیروان معیار ماژولاریتی را معرفی کرده‌اند. به طور کلی، اساس ماژولاریتی بر پیروی شـبکه‌های پیچیـده واقعـی از قانون توانی استوار است و برابر با تعداد یال‌ها در اون اجتماع منهای تعداد یال مورد نظر مطابق قانون توانی است و در بازه ۰٫۵- تا ۱ قـرار دارد. معیار ماژولاریتی بر اساس فرمول (۴) محاسبه می‌شود.

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i.C_j) \quad (۴)$$

که در آن A ماتری مجاورتی گراف، m تعداد یالهای گراف، P ماتریس مورد انتظار از نظر تعـداد یال‌هـای بـین گـره i و گـره j و δ تـابع بررسی در یک اجتماع بودن دو گره i و j است. اگر دو گره i وj در یک جامعه باشند، خروجی این تابع برابر بـا یـک اسـت و در غیـر ایـن صورت برابر با صفر است.

۲،۱،۴. محاسبه اطلاعات متقابل نرمال شده

معیار اطلاعات متقابل نرمال شده یا NMI، تشابه آماری بین خوشه‌های ایجاد شده و برچسب‌های از پیش تعیین شده را انـدازه می‌گیـرد. هرچه کیفیت خوشه‌بندی بهتر باشد، مقدار آنتروپی عددی کمتر و NMI عـددی بیشـتر خواهـد بـود و مطـابق بـا فرمـول (۵) محاسـبه می‌شود.

$$NMI(x,y) = \frac{2I(x.y)}{H(x) + H(y)} \quad (۵)$$

که در آن $I(x.y)$ اطلاعات مشترک دو اجتماع x و y و $H(x)$ و $H(y)$ آنتروپی دو اجتماع x و y را محاسبه می‌کند.

۲،۴. نتایج آزمایشات

هدف از انجام آزمایش¬ها، بررسی و مقایسه عملکرد الگوریتم پیشنهادی از جنبه¬های مختلفی مانند، نرخ صحت؛ اطلاعات متقابل نرمـال شده و زمان عملیات با الگوریتم تشخیص اجتماع بهبود یافته در منبع یک و روش مورد قبول و عموميیت یافته LPA است.

۱،۲،۴. بررسی اطلاعات متقابل نرمال شده

در جدول ۳ نتایج ارزیابی‌ها بر روی دیتاست‌های دنیای واقعی آورده شده و با روش معروف LPA و روش TS [۱] مقایسه شده است.

جدول ۳. ارزیابی و مقایسه روش پیشنهادی بر اساس معیار اطلاعات متقابل نرمال شده

نام دیتاست	روش پیشنهادی	الگوریتم TS [۱]	روش LPA
باشگاه کاراته زاکاری	۰/۴۴	۰/۴۲	۰/۵۱
دلفین‌ها	۰/۴۶	۰/۳۷	۰/۶۹
کتاب‌های سیاسی آمریکا	۰/۵۷	۰/۵۶	۰/۵۳
باشگاه‌های فوتبال	۰/۶۹	۰/۸۵	۰/۷۹

در سه دیتاست باشگاه کاراته، دلفین‌ها و کتاب‌های سیاسی آمریکا روش پیشنهادی ما، عملکرد بهتری نسبت به الگوریتم TS داشته است. در حالیکه در دیتاست فوتبال، الگوریتم LPA و TS نسبت به رویکرد پیشین عملکرد بهتری دارند.

۲،۲،۴. بررسی معیار ماژولاریتی

در جدول ۴ ارزیابی‌ها بر روی دیتاست‌های دنیای واقعی آورده شده و با روش معروف LPA و روش TS مقایسه شده است.

جدول ۴. ارزیابی و مقایسه روش پیشنهادی بر اساس معیار ماژولاریتی

نام دیتاست	روش پیشنهادی	الگوریتم TS [۳]	روش LPA
باشگاه کاراته زاکاری	۰/۴۰	۰/۳۸	۰/۳۴
دلفین‌ها	۰/۴۶	۰/۱۹	۰/۴۷
کتاب‌های سیاسی آمریکا	۰/۴۰	۰/۴۰	۰/۴۹
باشگاه‌های فوتبال	۰/۵۰	۰/۶۶	۰/۵۸

عقب بودن روش پیشنهادی در دیتاست فوتبال نسبت به روش LPA و روش TS در مورد معیار ماژولاریتی نیز صدق می‌کنند.

۳،۲،۴. بررسی و ارزیابی نرخ صحت

علاوه بر معیارهای ماژولاریتی و اطلاعات متقابل نرمال شده، در روش‌های عنوان شده برای تشخیص اجتماع، پیش‌بینی تعلق درست گره‌ها به اجتماع، نیز از اهمیت بالایی برخوردار است. بنابراین روش پیشنهادی و روش TS را از لحاظ نرخ صحت، مقایسه کردیم. لازم به ذکر است در ارزیابی این معیار به دلیل عدم همگرایی و تصادفی بودن نتایج نهایی روش LPA لحاظ نشده است.

جدول ۵. ارزیابی و مقایسه روش پیشنهادی بر اساس نرخ صحت

نام دیتاست	روش پیشنهادی	الگوریتم TS [۳]
باشگاه کاراته زاکاری	۰/۵۸	۰/۳۹
دلفین‌ها	۰/۶۷	۰/۳۲
کتاب‌های سیاسی آمریکا	۰/۷۲	۰/۶۲
باشگاه‌های فوتبال	۰/۳۸	۰/۲۵

در هر ۴ دیتاست مذکور، روش پیشنهادی از عملکرد بهتری نسبت به روش مقاله پایه برخوردار است.

۴،۲،۴. بررسی و ارزیابی سرعت عملیات

زمان اجرای الگوریتم، یکی از فاکتورهای مهم الگوریتم‌های تشخیص اجتماع است. هم‌چنان رشد سریع و روزافزون داده‌ها در دنیای امروز، اهمیت تحلیل داده‌ها در زمان کم‌تر را بیش‌تر نمایان می‌کند. بنابراین در جدول ۶ مدت زمان اجرای روش پیشنهادی، با روش TS و روش LPA مقایسه شده است.

جدول ۶. ارزیابی و مقایسه روش پیشنهادی بر اساس مدت زمان انجام عملیات (واحد ثانیه)

نام دیتاست	روش پیشنهادی	الگوریتم TS [۳]	روش LPA
باشگاه کاراته زاکاری	۱۱	۱۴/۳	۰/۰۶۱
دلفین‌ها	۱۷	۱۸/۷۸	۰/۰۹۲
کتاب‌های سیاسی آمریکا	۴۱/۳۳	۴۸/۲۴	۰/۱۱
باشگاه‌های فوتبال	۵۰/۳۲	۱۱۲/۹	۰/۱۲

روش LPA از سریع‌ترین الگوریتم‌های تشخیص اجتماع است اما این الگوریتم به دلیل تصادفی بودن و عدم همگرایی، چندان روش قابل اطمینانی نیست. در حالیکه روش پیشنهادی در کنار زمان پایین‌تر نسبت به الگوریتم TS، از صحت نرخ بالاتری نیز برخوردار است.

۵. نتیجه‌گیری

در این پژوهش برای اولین بار، تعریف گره‌های پراهمیت در شبکه‌های پیچیده، بر اساس گره‌های ناهنجار تعریف شد. یعنی گره‌هایی به عنوان گره مرکزی شناخته شدند که رفتار آن‌ها، از رفتار غالب گره‌های شبکه دور باشد. این گره‌های ناهنجار می‌توانند گره‌هایی متعلق به

یک سلبریتی؛ اینفلئونسر یا یک کلاهبردار مالی باشد. روش پیشنهادی، از زمان اجرای کم‌تری نسبت به الگوریتم‌های هم‌تراز خود برخوردار است، در حالیکه نسبت به الگوریتم‌هایی مشهوری مانند LPA کندتر عمل می‌کند ولی همگرایی، نرخ صحت و ماژولاریتی بالاتر این روش باعث می‌شود تا این روش نسبت به روش‌های پیشین، قابل قبول‌تر باشد.

۶. منابع

[1] حسینی, م. و ا.ا. مه آبادی, روش توزیعی تشخیص انجمن در شبکه‌های اجتماعی بزرگ بر اساس انتشار برچسب ۲۰,۱۰۰۱,۱,۲۳۲۲۴۳۴۷,۱۳۹۹,۸,۴,۱,۷ : . ۱۳۹۹.

[2] صالحی, س.م.م. and ع.ا. پویان, انجمن یابی در شبکه‌های اجتماعی به کمک یادگیری عمیق, چهارمین کنفرانس پردازش سیگنال و سیستم‌های هوشمند. ۱۳۹۷

[3] You, X., Y. Ma, and Z. Liu, A three-stage algorithm on community detection in social networks. Knowledge-Based Systems, 2020. 187: p. 104822.

[4] S. Aghaalizadeh, S. T. Afshord, A. Bouyer, and B. Anari, "A three-stage algorithm for local community detection based on the high node importance ranking in social networks," Physica A: Statistical Mechanics and its Applications, vol. 563, p. 125420, 2021/02/01/ 2021.

[5] Freeman, L., The Development of Social Network Analysis. 2004.

[6] Moody, J. and D. White, White, D.: Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. American Sociological Review 68, 103-127. American Sociological Review, 2003. 68: p. 103-127.

[7] Fortunato, S., Community detection in graphs. Physics Reports, 2010. 486(3): p. 75-174.3.You, X., Y. Ma, and Z. Liu, A three-stage algorithm on community detection in social

[8] Aynaud, T. and J.-L. Guillaume. Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks. in Fifth SNA-KDD Workshop Social Network Mining and Analysis, in conjunction with the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011). 2011. San Diego, CA, United States.

[9] Scott, J., Social Network Analysis. Sociology, 1988. 22(1): p. 109-127.

[10] Wasserman, S. and K. Faust, Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. 1994, Cambridge: Cambridge University Press.

[11] Ravasz, E. and A.-L. Barabási, Hierarchical organization in complex networks. Physical Review E, 2003. 67(2): p. 026112.

[12] Boccaletti, S., et al., Complex networks: Structure and dynamics. Physics Reports, 2006. 424(4): p. 175-308.

[13] Ostroumova Prokhorenkova, L. and E. Samosvat. Global Clustering Coefficient in Scale-Free Networks. 2014. Cham: Springer International Publishing.

[14] Csányi, G. and B. Szendrői, Structure of a large social network. Physical Review E, 2004. 69(3): p. 036131.

[15] Jaccard, P., Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat, 1901. 37: p. 547-579.

[16] T. Sørensen et al., "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," 1948.

[17] J. Scripps, P. Tan, and A. Esfahanian, "Exploration of Link Structure and Community-Based Node Roles in Network Analysis," in Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 649-654.

[18] M. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," Physical review. E, Statistical, nonlinear, and soft matter physics, vol. 69, p. 026113, 03/01 2004.

[19] N. Raghavan, R. Albert, and S. Kumara, "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks," Physical review. E, Statistical, nonlinear, and soft matter physics, vol. 76, p. 036106, 10/01 2007.

[20] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," (in eng), Kdd, vol. 2016, pp. 855-864, Aug 2016.

[21] M. E. J. Newman, "Detecting community structure in networks," The European Physical Journal B, vol. 38, no. 2, pp. 321-330, 2004/03/01 2004.

[22] M. E. J. Newman, "The Structure and Function of Complex Networks," Computer Physics Communications, vol. 147, pp. 40-45, 03/15 2003.

[23] C. Ma and J. Wu, Data Clustering: Theory, Algorithms, and Applications. 2007.

[24] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," The Bell System Technical Journal, vol. 49, no. 2, pp. 291-307, 1970.

[25] S. Dutt, "New faster Kernighan-Lin-type graph-partitioning algorithms," in Proceedings of 1993 International Conference on Computer Aided Design (ICCAD), 1993, pp. 370-377.

[26] M. Chen, K. Kuzmin, and B. K. Szymanski, "Community Detection via Maximization of Modularity and Its Variants," IEEE Transactions on Computational Social Systems, vol. 1, no. 1, pp. 46-65, 2014.

[27] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven Label Propagation Algorithm," in 2011 IEEE Network Science Workshop, 2011, pp. 188-195.

[28] J. Cao, D. Jin, L. Yang, and J. Dang, "Incorporating network structure with node contents for community detection on large networks using deep learning," Neurocomputing, vol. 297, 02/01 2018.

[29] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical review E, vol. 78, no. 4, p. 046110, 2008.
[30] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of anthropological research, vol. 33, no. 4, pp. 452-473, 1977.
[31] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology, vol. 54, no. 4, pp. 396-405, 2003.
[32] M. E. Newman, "Modularity and community structure in networks," Proceedings of the national academy of sciences, vol. 103, no. 23, pp. 8577-8582, 2006.
[33] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821-7826, 2002.
[34] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," Physical review E, vol. 68, no. 6, p. 065103, 2003.
[35] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical review E, vol. 74, no. 3, p. 036104, 2006.
[36] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," nature, vol. 393, no. 6684, p. 440, 1998.
[37] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, p. 2, 2007.
[38] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," Physical review E, vol. 72, no. 2, p. 027104, 2005.
[39] M. Boguná, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," Physical review E, vol. 70, no. 5, p. 056122, 2004.
[40] B. Rozemberczki, C. Allen, and R. Sarkar, "Multi-scale Attributed Node Embedding," arXiv preprint arXiv:1909.13021, 2019.
[41] M. E. Newman, "Fast algorithm for detecting community structure in networks," Physical review E, vol. 69, no. 6, p. 066133, 2004.
[42] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," Internet Mathematics, vol. 6, no. 1, pp. 29-123, 2009.
[43] M. E. Newman, "The structure of scientific collaboration networks," Proceedings of the national academy of sciences, vol. 98, no. 2, pp. 404-409, 2001.
[44] L. Akoglu, M. McGlohon, and C. Faloutsos, "oddball: Spotting Anomalies in Weighted Graphs," in Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg, 2010, pp. 410-421: Springer Berlin Heidelberg.