فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد زنجان

افزایش دقت شناسایی جوامع همپوشان با استفاده از وزن‌دهی یال‌ها

DOR: 20.1001.1.27832570.1399.1.1.3.7

مقاله پژوهشی

مهدی افضلی*[۱]؛ ایرج تیموری [۲]

۱- مرکز تحقیقات انرژی و سیستم، دانشگاه آزاد اسلامی واحد زنجان، زنجان، ایران afzali@iauz.ac.ir

۲- دانشگاه فرهنگیان، پردیس شهید بهشتی، زنجان، ایران

چکیده: یکی از مهمترین ویژگی‌های شبکه‌های پیچیده وجود ساختارهای اجتماعی می‌باشد. بطور مشخص شناسایی این ساختارها در شبکه‌های پیچیده به تحلیل ویژگی‌های ساختاری شبکه کمک می‌کند. در سال های اخیر الگوریتم‌های متعددی برای کشف اجتماعات در شبکه‌های پیچیده پیشنهاد شده است. با توجه به ویژگی‌های این اجتماعات، یکی از روش‌های موجود برای شناسایی اجتماعات ارائه الگوریتم‌هایی برای وزنه‌دهی یال‌های شبکه است به طوری‌که وزن یال‌های درون اجتماعات افزایش و بطور همزمان وزن یال‌های مابین اجتماعات کاهش یابد تا تمایز میان اجتماعات به سادگی قابل شناسایی باشند. در روش پیشنهادی با استفاده از فرآیند وزن‌دهی به یال‌ها، بین گره‌های که مشابهت بیشتری دارند و گره‌هایی که مشابهت اندکی با هم دارند تمایز قایل می‌شویم. یعنی با اختصاص وزن با استفاده از معیارهای پیشنهادی در برخی الگوریتم‌ها، یال‌هایی که وزن بیشتری دارند نقش بیشتری در تعیین جمعیت خواهند داشت. با توجه به اینکه یک همبستگی مثبت بین ساختارهای جامعه و معیارهای شباهت وجود دارد، نتایج آزمون‌های انجام شده نشان می‌دهد که استفاده از معیارهای مشابهت محلی به عنوان وزن یال‌ها برای برخی از الگوریتم‌ها باعث افزایش دقت تشخیص جوامع می‌شود. این الگوریتم‌ها از درجه گره‌ها به عنوان یکی از ویژگی‌های شبکه برای محاسبه قدرت جذب هسته‌ها برای تشکیل جوامع استفاده می‌کنند. به عنوان نمونه در مورد شبکه‌های واقعی، اجرای الگوریتم WHD-EM روی شبکه High school network، جوامع را با دقت NMI=0.6652 و معیار خلوص purity=0.9845 کشف می‌کند که از بعضی از الگوریتم‌ها مانند CPM، NMF ، GAME ، GCE، OSLOM و LINK از نظر معیار NMI بهتر است.

واژه‌های کلیدی: شبکه‌های پیچیده، شبکه‌های اجتماعی، شناسایی جوامع، وزن‌دهی یال‌ها

Increasing the accuracy of identifying overlapping communities using weighted edges

Mehdi Afzali[1], Iraj Teymouri[2]

1. Research Center of Energy and System, Islamic Azad University, Zanjan Branch, Zanjan, Iran, afzali@iauz.ac.ir

2. Farhangian University, Zanjan Branch, Zanjan, Iran

Abstract: One of the most important features of the complex networks are social structures. More specifically, identifying these structures in complex networks helps to analyze the characteristics of the networks structure. In recent years, several algorithms have been proposed for detection in complex networks. It should be noted that wuth regarding the features of these communities, one of the existing methods for identifying communities, is providing an algorithms for weighting the edges of the network as the weight of the edges of the communities are incresed and at the same time the weight of the edges between communities are reduced. Therefore the distinction between communities could be identified simply. In the proposed method with useing the process of weighting the edges, we distinguish between the nodes that are more similar to each other and the nodes that have slight similarity. i.e. by assigning the weight with using the proposed criteria in some algorithms, the edges with more weight will have a greater role in determining the population. According that there is a positive correlation between similarity measures and community structures, the results of the tests shows that using local similarity measures as the weight of edges for some algorithms, causes an increase in the accuracy of communities recognition. These algorithms use the degree of the nodes as one of the network characteristics for computing the cores absorbing ability for communities' formation. For example, in the case of Real Networks, running WHD-EM algorithm on the High school network, discovers the communities with the NMI=0.6652 and the purity criteria equal to 0.9845. Also it should be noted that some algorithms such as CPM, NMF, GAME, GCE, OSLOM and LINK in terms of NMI criteria, are better.

Keywords: Complex network, Social Networks, Identify communities, Weighing the edges.

تاریخ ارسال مقاله: ۱۳۹۹/۰۷/۱۸

تاریخ پذیرش مقاله: ۱۳۹۹/۰۹/۱۵

*: نویسنده مسئول

۱- مقدمه

بسیاری از سامانه‌های طبیعی، اجتماعی و صنعتی که دربرگیرنده ی بخش‌های برهم‌کنشی بسیار، با توانایی فراوری یک سرشت نو از رفتار گروهی ماکروسکوپی می‌باشند، سامانه‌های پیچیده نامیده می‌شوند[1] یک نمونه از سامانه‌های پیچیده، مغز است. گرچه کارکرد یک نورون یا یک سیناپس به خوبی قابل فهم است ولی سازو کارهایی که برهم کنش تعداد بسیاری نورون به حافظه، یادگیری، خلاقیت و هوشیاری می‌انجامد به طور چشم گیری پوشیده می‌مانند.

پدیده‌های گوناگونی از دنیای واقعی مانند روابط اجتماعی، برهم‌کنش‌های شیمیایی و زیست شناختی، پخش ویروسها و بیماری‌ها، زمین‌لرزه‌ها، اینترنت و شبکه‌ی گسترده‌ی جهانی، پدیده‌های زبانشناختی را می‌توان با شبکه‌های پیچیده مدل‌سازی نمود [2]

شبکه‌های پیچیده گستره‌ی وسیعی از شبکه‌های اجتماعی و شبکه‌های طبیعی را در برگرفته‌اند. ساختار این شبکه‌ها بطور کلی اسپارس یا خلوت است ولی به صورت محلی چگال است. به این معنا که گره‌های این شبکه‌ها، به صورت گروه‌هایی دسته‌بندی می‌شوند که ارتباطات بین گره‌های هم‌گروه بسیار متراکم و بیشتر از ارتباط آن‌ها با گروه‌های دیگر است. به چنین ساختاری community یا اجتماع می-گویند[3]

محققان شبکه‌های اجتماعی و تئوری گراف علاقه‌مند به یافتن ساختار اجتماعات به کمک تحلیل الگوهای ارتباطی بین افراد در شبکه‌های اجتماعی هستند، زیرا تشخیص اجتماع در شبکه‌های اجتماعی یک ابزار قوی برای آشکارسازی بسیاری از خصوصیات پنهان شبکه‌هاست. به عنوان مثال می‌توان به خصوصیات مشترک افراد هم‌گروه اشاره کرد. علاوه بر آن مانیتور کردن این علایق و خصوصیات مشترک می‌تواند در برخی از کاربردهای تجاری مانند بازاریابی هوشمند و امثالهم مفید باشد. ازاین‌رو یافتن و تشخیص اجتماعات نه تنها دلیل تشکیل این گروه‌ها را آشکار می‌کند، بلکه از طرف دیگر به فهم ساختار کلی و خصوصیات شبکه‌های بزرگ هم کمک می‌کند4] .

در شبکه‌های برهم‌کنشی پروتئین پروتئین، جوامع می‌توانند پروتئین‌هایی را که تابع مشخصه یکسانی درون سلول دارند، دسته‌بندی کنند[6 and 5] در گراف شبکه‌ی گسترده جهانی، جوامع می‌توانند متناظر با گروه‌هایی از صفحات با موضوع یکسان و یا مرتبط با یک موضوع ویژه باشند[7]، در شبکه‌های سوخت وسازی[8] جوامع شاید وابسته به حوزه‌های تابعی مانند چرخه‌ها و گذرگاه‌ها باشند، در شبکه-های غذای نیز جوامع می‌توانند دسته‌بندی‌های غذایی را مشخص کنند[3]، و نمونه‌های دیگری مانند اینها.

بسیاری از شبکه‌های موجود در دنیای واقعی می‌توانند توسط یک شبکه پیچیده مدل شوند. بررسی ساختار شبکه‌های پیچیده منجر به معرفی اجتماعات گردیده است. از آنجائی که این اجتماعات دارای ارتباطات داخلی قوی مابین راس‌های همان اجتماع در مقایسه با سایر راس‌ها از اجتماعات دیگر است. ارائه معیاری برای وزن‌دهی و شناخت بهتر این اجتماعات امری ضروری تلقی می‌شود.

راس‌هایی که خواص مشابه دارند می‌توانند گروهی که ارتباطات بیشتر و تعامل بهتری با هم دارند را تشکیل دهند. بنابراین می‌توان یک ارتباط بین ساختارهای جامعه و معیارهای شباهت در شبکه در نظر گرفت، در روش‌های پیشنهادی با استفاده از فرآیند وزن‌دهی به یال‌ها، بین گره‌های نزدیک به هم و گره‌هایی که آشنایی اندکی با هم دارند تمایز قایل می‌شویم. یعنی با اختصاص وزن های مختلف، یال‌هایی که وزن بیشتری دارند نقش بیشتری در تعیین جمعیت خواهند داشت.

۲- بیان مساله

در بررسی سامانه‌های پیچده تلاش برای پی بردن به الگوها و ضوابط برهم‌کنش‌ها انجام می‌گیرد به گونه‌ای است که بخش‌بندی سامانه‌ها به زیرمجموعه‌های تشکیل دهنده‌ی آن‌ها را امکان پذیر می‌کند.

یک روش مناسب برای نمایش سامانه‌های پیچیده، استفاده از گراف-ها یا شبکه‌ها می‌باشد. می‌توان اجزای سامانه را با رئوس و برهم‌کنشهای میان اجزای سامانه را با پیوندهای شبکه نشان داد. بنابراین می‌توان زبان ریاضی نظریه‌ی گراف را برای توصیف سامانه های پیچیده، و بررسی ویژگی های توپولوژیکی برهم‌کنش‌هایی که سامانه را تعریف می‌کنند [7 and 10].

هدف از شناسایی جوامع در گراف‌ها، شناخت حوزه‌ها، و اگر امکان پذیر باشد نمایش ساختار سلسله مراتبی آن‌ها، تنها با به کارگیری دانش نهفته در توپولوژی گراف می‌باشد. این مسئله به چندین گونه گفته شده و در الگوها و دستورکارهای گوناگونی آمده است.

این نکته حایز اهمیت است که میان شناسایی جوامع با افراز گراف تمایز قائل شویم. چرا که جوامع می‌توانند همپوشانی داشته باشند، در صورتی که این مورد در افرازبندی به طور کلی منتفی است.

در الگوریتم‌های خوشه‌بندی و تشخیص جوامع، هم به کمک ساختار شبکه و هم از روی خصوصیات گره‌ها می‌توان اجتماعات را تشخیص داد. برخی الگوریتم‌ها، تنها به ویژگی گره‌ها و برخی بر اساس ساختار شبکه به یافتن گروه‌هایی از گره‌ها با ارتباطات قوی اقدام می‌کنند.

با استفاده از اطلاعات ساختاری محلی، می‌توان برای بازسازی ساختاری وزن‌های موجود در شبکه استفاده نمود و روش‌های موجود تشخیص جامعه می‌تواند جوامع را به طور موثرتر در شبکه‌های وزنی نسبت به شبکه‌های اصلی شناسایی کنند[11].

با توجه به مفهوم جامعه در شبکه‌های پیچیده و از آنجا که یک همبستگی مثبت بین ساختارهای جامعه و معیارهای شباهت وجود دارد، معیارهای شباهت ممکن است در تشخیص دقیق‌تر جوامع مفید باشند[12].

با توجه به مفهوم جامعه در شبکه‌های پیچیده که از تعداد زیادی گره که بیشترین ارتباط را با هم دارند و یال‌های بین آنها متراکم تر است، می تواند با گروهی از گره‌ها که خواص مشابهی دارند مطابقت کند

بنابراین معیارهای شباهت ممکن است در تشخیص دقیق تر جوامع مفید باشند.

با این فرض می‌توان از این معیارهای مشابهت محلی به‌عنوان کمیتی برای تخصیص و یا بازسازی وزن یال‌های شبکه استفاده کرد و سپس از الگوریتم‌های شناخته شده در این زمینه برای شناسایی جوامع استفاده نمود.

۳- ادبیات پژوهش

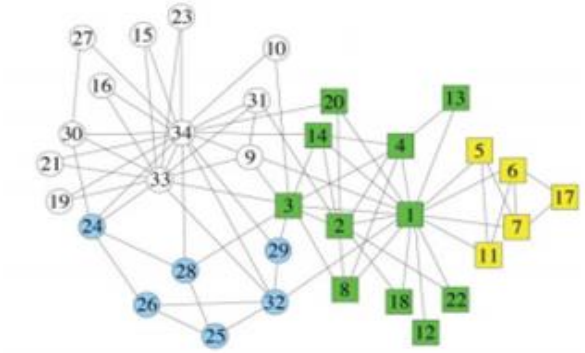
الگوریتم کرنایان-لین یکی از اولین روش‌های پیشنهاد شده می‌باشد که هنوز هم در ترکیب با روشهای دیگر به کار می‌رود[13]. نویسندگان از بخش‌بندی مدارات الکترونیکی روی تخته ی مدار چاپی، ایده گرفته‌اند. گره‌های روی تخته‌های مدار چاپی گوناگون باید با کمترین تعداد پیوندها با هم پیوند داشته باشند.

فراگیرترین الگوریتم جداکننده را گیروان و نیومن پیشنهاد کرده‌اند و از دیدگاه تاریخی ارزشمند است، زیرا گستره‌ی تازه‌ای در زمینه‌ی شناسایی جوامع آشکار نموده و پای فیزیک‌پیشه‌ها را به این پهنه گشوده است[14 and 15]. یال‌ها بر پایه‌ی اندازه‌های مرکزیت یال[1] که ارزش یالها را بر پایه‌ی برخی ویژگی آن‌ها، یا کارکرد آن‌ها در گراف ارزیابی می‌کنند، برگزیده می‌شوند.

نخستین تلاش در بیشینه‌سازی مستقیم ماژولاریتی را نیومن و با به‌کارگیری بیشینه‌سازی حریصانه انجام داد[16]. در آغاز الگوریتم هر گره در یک حوزه جای می‌گیرد. سپس تغییر ماژولاریتی به‌دست‌آمده از یکی شدن هر دو حوزه محاسبه می‌گردد. دو حوزه‌ای که به هم‌پیوستن آن‌ها بیشترین تغییر مثبت در ماژولاریتی را در پی دارد، به هم می-پیوندند. این فرایند تا هنگامی که یک‌اندازه‌ی بیشینه برای ماژولاریتی به دست آید باز انجام می‌شود. این الگوریتم یکی از روش‌های سریع به شمار می‌آید، به‌ویژه هنگامی که شبکه‌ی بررسی‌شونده تنک باشد.

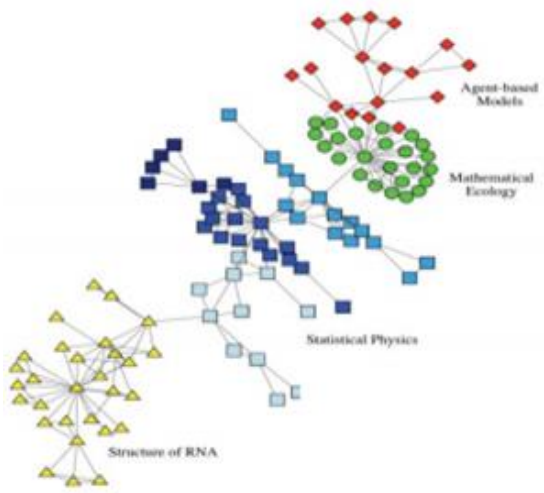
بهینه‌سازی کرانی[2] روشی ابتکاری است که بوچر و پرکاس برای دستیابی به دقتی سنجش‌پذیر با روش بازپخت همانندسازی‌شده ، ولی با سرعتی بالاتر پیشنهاد دادند[17]. این روش ماژولاریتی محلی را که نشان‌دهنده‌ی سهم هر گره در ماژولاریتی سراسری است بهینه می‌کند. برای آغاز گره‌ها به گونه‌ی کاتوره‌ای در یکی از دو جامعه جای می‌گیرند. سپس سهم هر گره در ماژولاریتی، با انتقال گره‌هایی که بدترین ماژولاریتی محلی را دارند به جامعه‌ی دیگر، بهبود می‌یابد. این روند تا هنگامی که ماژولاریتی سراسری به یک اندازه‌ی بیشینه برسد ادامه می-یابد. سپس یال‌های میان دو جامعه برداشته می‌شوند، و همه‌ی مراحل تا هنگامی که ماژولاریتی افزایش می‌یابد به گونه‌ی بازگشتی بازانجام می‌شود.

نمونه‌هایی از شبکه‌های اجتماعی را می‌توان مورد بررسی قرار دارد. نمونه نخست شکل ۱، شبکه‌ی اعضای باشگاه کاراته زاخاری، یک گراف شناخته شده است که بیش‌تر آزمونی برای نشان دادن درستی یا نادرستیِ الگوریتم‌های شناساییِ جوامع می‌باشد[18]. این گراف ۳۴ رأس دارد که اعضای یک باشگاه کاراته در امریکا می‌باشند و زاخاری سه سال آن‌ها را زیر نظر گرفته است. اگر بیرون باشگاه میان افراد ارتباطی باشد، میان رئوس هم ارز آن‌ها یک یال هست. گاهی کشمکش میان رئیس باشگاه و مربی مایه‌ی دودستگیِ اعضای باشگاه می شود و یک دسته از مربی و دسته‌ی دیگر از رئیس باشگاه پشتیبانی می‌کنند که در شکل به ترتیب با مربع‌ها و دوایر نشان داده شده‌اند. پرسش این است که آیا از روی ساختار شبکه‌ی نخستین می‌توانیم به آرایش این دو دسته پی ببریم یا خیر. با یک نگاه به شکل(۱) می‌توان دو انباشتگی، یکی پیرامون رئوس شماره‌ی ۳۳ و ۳۴ (رأس ۳۴ رئیس باشگاه است) و دیگری پیرامون رأس شماره‌ی ۱ (مربی) دید. همچنین چند رأس که میان دو ساختار بزرگ جای دارند، مانند ۹،۳ و ۱۰ به چشم می خورند.



شکل ۱: ساختار اجتماعی شبکه اعضای باشگاه‌ی کاراته زاخاری، یک آزمون استاندارد در شناساییِ جوامع [18]

شکل ۲ بزرگترین اعضای مرتبط یک شبکه‌ی همکاریِ دانشمندان که در مؤسسه‌ی سانتافی کار می‌کنند را نشان می‌دهد. ۱۱۸ رأس داریم که نشان دهنده‌ی دانشمندان مؤسسه و همکاران آن‌ها می باشند. هنگامی میان دانشمندان یالی هست که دست کم یک نوشته‌ی مشترک داشته باشند. در طرح نمایشی، گروه‌های منظم به چشم می‌خورند. در این شبکه گروه‌هایی را می توان دید که در آن‌ها نویسندگان یک مقاله با هم پیوند دارند. ولی میان بیشتر گروه‌ها پیوندهای اندکی است.



شکل ۲: ساختار اجتماعی شبکه‌ی همکاریِ دانشمندانی که در مؤسسه‌ی سانتا فی کار می‌کنند (پ). شبکه‌ی دلفین‌های پوزه بطری در نیوزلند[3]

۴- اهداف پژوهش

هر چند مطالعه و استخراج جوامع در شبکه‌ها پیچیده موضوع جدیدی نیست ولی محدودیت های روش‌های استخراج و هم چنین تفسیر نتایج خروجی تحقیق بیشتری در این زمینه را لازم می‌دارد.

حوزه تحقیقاتی استخراج جوامع هنوز یک موضوع چالش برانگیز در شبکه‌های اجتماعی است بنابراین این روش‌ها هنوز هم در حال گسترش و بهبود کارایی هستند. عمده راهکارهای معرفی شده عبارتند از: ۱- استفاده از پارامترهای ارزیاب کیفی و پارامترهای شباهت ۲-استفاده از روش‌های خوشه‌بندی ۳- بخش بندی گراف ۴- روش‌های طیفی ۵- روش‌های بیزین[19]

شناسایی جوامع به دلایل دیگری نیز ارزشمند است؛ آشکارسازی حوزه‌ها و مرزهای آن‌ها اجازه دسته‌بندی رئوس بر پایه‌ی موقعیت ساختاری آن‌ها در حوزه را به ما می‌دهد. با این کار رئوسی که در خوشه-های خود موقعیت مرکزی دارند، یعنی آن‌هایی‌که با تعداد بسیاری از هم‌گروهی‌های خود پیوند دارند، می‌توانند نقش مهمی در کنترل پایداری گروه خود بازی کنند؛ و رئوسی که در مرزها ی حوزه‌ها جای دارند در ارتباط و تبادل میان جوامع، نقش میانجی خواهند داشت. این گونه دسته‌بندی می‌تواند در شبکه‌ای سوخت وسازی[8] و اجتماع انسانی معنی‌دار باشد. یکی از مفاهیم مهم وابسته به ساختار اجتماعی، سازمان-یافتگی سلسله مراتبی است که می‌توان آن را در بسیاری از سامانه‌های شبکه‌ای دنیای واقعی دید. شبکه‌های واقعی بیش‌تر از جوامعی ساخته شده اند که آن‌ها نیز دربرگیرنده‌ی جوامع کوچک‌تر می‌باشند و این جوامع کوچک‌تر نیز به نوبه خود ساخته شده از جوامع کوچک‌تر دیگری می‌باشند و به همین گونه تا پایان. برای نمونه بدن آدمی از اعضا، اعضا از بافت‌ها، و بافت‌ها از سلول‌ها ساخته شده‌اند. نمونه‌ی دیگر را می‌توان در شرکت‌های بازرگانی دارای ساختار هرمی جستجو کرد، که از کارگران آغاز می‌گردد و با سطوح میانی متناظر با گروه‌های کار، بخش‌ها و مدیریت، تا رئیس ادامه می‌یابد. فراوری و دگرگونی یک سامانه که از زیرسامانه‌های پایدار وابسته به هم تشکیل شده است بسیار سریع‌تر از هنگامی است که سامانه بی‌ساختار باشد، چون بسیار ساده‌تر است که نخست بخش‌های کوچکتر ساخته شوند، سپس آن را یکاهای ساختمانی ساختارهای بزرگتر کنیم تا آن که همه‌ی سامانه برپا گردد. در این روش احتمال رخ دادن خطا و تغییرات ناگهان در گامهای گوناگون بسیار کمتر خواهد بود.

در بسیاری از شبکه‌های واقعی رئوس می‌توانند در بیش از یک گروه باشند که به آن‌ها جوامع هم پوشان می‌گوییم. الگوریتم‌های سنتی شناسایی جوامع هر رأس را تنها در یک حوزه جای می‌دهند. با این کار برخی از داده‌های وابسته از میان می‌رود.این امر می‌تواند شامل رئوسی که در بیش از یک جامعه هستند و ممکن است میانجی بخش‌های گوناگون یک گراف باشند، شود.

با در نظر گرفتن جوامع هم‌پوشان، عضویت رئوس در جوامع گوناگون، تعدادپوش‌هایی که می‌توان داشت را در مقایسه با تعداد افرازهای استاندارد بسیار بالا می‌برد.

مساله تشخیص جوامع کاربردهای متعددی دارد که می‌توان به موارد زیر اشاره کرد:

- خوشه‌بندی مشتریان وب که علایق و سلایق مشترکی دارند و یا از لحاظ جغرافیایی نزدیک به هم هستند، ممکن است باعث ارتقا و بهبود کیفیت خدمات ارائه شده به آن‌ها گردد. به این دلیل که هر خوشه‌ای از مشتریان می‌توانند یک کارگزار وقف شده[3] مورد خدمت‌رسانی قرار گیرند.
- تشخیص خوشه‌های مختلف در گراف‌های بزرگ می‌تواند برای ذخیره‌سازی بهینه آن‌ها در حافظه کامپیوتر با استفاده از ساختمان داده‌های کاراتر استفاده شود.
- اجتماع انسانی گونه‌های گسترده‌ای از ساختارهای گروهی را پیشنهاد می‌کند: خانواده‌ها، جمع دوستان و همکاران، اهالی روستاها و شهرها، خویشاوندان و مانند آن‌ها. گسترش اینترنت نیز به پیدایش گروه‌های مجازی بسیاری دامن زده است. جوامع در بسیاری از سامانه‌های شبکه ای دیگر، در حوزه‌هایی چون دانش زیستی، دانش رایانه‌ای، مهندسی اقتصاد و سیاست به چشم می‌خورند.

هدف اصلی، افزایش میزان دقت براساس وزن‌دهی مناسب در فرآیند کشف اجتماعات همپوشان در شبکه‌ها پیچیده می‌باشد.

بسیاری از شبکه‌های موجود در دنیای واقعی می‌توانند توسط یک شبکه پیچیده مدل شوند. بررسی ساختار شبکه‌ها پیچیده منجر به معرفی اجتماعات گردیده است. از آنجا که اجتماعات دارای ارتباطات داخلی قوی بین گره‌های همان اجتماع در مقایسه با سایر گره‌ها از اجتماعات دیگر است ارائه معیاری برای وزن‌دهی و شناخت بهتر این اجتماعات امری ضروری تلقی می‌شود.

در این میان می‌توان معیارهای مختلف برای وزن‌دهی به یال‌های شبکه بدون وزن و یا بازسازی وزن‌های ساختارهای وزن دار پیشنهاد کرد که دقت و کارایی الگوریتم‌های کارآمد موجود را افزایش داد.

با توجه به مفهوم جامعه در شبکه‌های پیچیده که از تعداد زیادی گره که بیشترین ارتباط را با هم دارند و یال‌های بین آن‌ها متراکم‌تر است، می‌تواند با گروهی از گره‌ها که خواص مشابهی دارند مطابقت کند بنابراین معیارهای شباهت ممکن است در تشخیص دقیق‌تر جوامع مفید باشند.

با این فرض می‌توان از این معیارهای مشابهت محلی به عنوان کمیتی برای تخصیص و یا بازسازی وزن یال‌های شبکه استفاده کرد و سپس از الگوریتم‌های شناخته شده در این زمینه برای شناسایی جوامع استفاده نمود.

سوالات این تحقیق عبارتند از:

۱) آیا در نظر گرفتن وزن مناسب برای گراف شبکه پیچیده در دقت شناسایی جوامع موثر است؟
۲) آیا وزن دار کردن یال‌های بدون وزن در میزان دقت شناسایی تاثیرگذار است؟

۵- روش پیشنهادی

تشخیص اجتماعات در یک شبکه می تواند به صورت مسأله خوشه بندی یک مجموعه از گره‌ها در یک سری اجتماعات در نظر گرفته شود که هر گره می تواند همزمان به چند اجتماع متعلق باشد. از آنجا که گره‌های موجود در یک اجتماع، خصوصیات مشترکی دارند و همچنین ارتباطات بسیاری نیز بین آن‌ها وجود دارد، از هر دوی این داده‌ها می توان به عنوان منبع داده ای برای گروه بندی گره‌ها استفاده کرد[20].

در واقع هم به کمک ساختار شبکه و هم از روی خصوصیات گره‌ها می‌توان اجتماعات را تشخیص داد. البته غالب الگوریتم‌ها تنها از یکی از این‌دو داده برای تشخیص اجتماعات استفاده می‌کنند. الگوریتم‌های خوشه بندی تنها به ویژگی گره‌ها اکتفا نموده و از روابط بین آن‌ها صرف نظر می‌کنند. از سوی دیگر الگوریتم‌های تشخیص اجتماع تنها بر اساس ساختار شبکه به یافتن گروه‌هایی از گره‌ها با ارتباطات قوی اقدام می‌کنند و از خصوصیات گره‌ها چشم پوشی می‌کنند.

در[11] نویسندگان نشان دادند که با استفاده از اطلاعات ساختاری محلی می‌توان برای بازسازی ساختاری وزن های موجود در شبکه استفاده نمود و روش‌های موجود تشخیص جامعه می‌تواند جوامع را به طور موثرتر در شبکه‌های وزنی نسبت به شبکه‌های اصلی شناسایی کنند. بنابراین اطلاعات اضافی موجود در وزن لبه‌ها در واقع برای شناسایی جوامع در شبکه‌های پیچیده کمک فراوانی می‌کند.

نکته‌ای که در مورد گراف‌هایی که ذاتا وزن‌دار هستند در نظر بگیریم این است که در روش پیشنهادی منظور وزن‌دهی مجدد به یال‌ها می‌باشد که در مورد گراف‌های بدون وزن مقدار آن برابر ۱ است که در مورد گراف‌های وزن‌دار، وزن آن‌ها تغییر خواهد کرد.

روش وزن‌دهی یال بسیار ساده و موثر است و می‌تواند به طور مستقیم به روش‌های تشخیص جامعه موجود اضافه شود.

با توجه به مفهوم جامعه در شبکه‌های پیچیده که از تعداد زیادی گره که بیشترین ارتباط را با هم دارند و یال‌های بین آن‌ها متراکم‌تر است، می‌تواند با گروهی از گره‌ها که خواص مشابهی دارند مطابقت کند. از آنجا که یک همبستگی مثبت بین ساختارهای جامعه و معیارهای شباهت وجود دارد، معیارهای شباهت ممکن است در تشخیص دقیق‌تر جوامع مفید باشند [12] .

در این پژوهش از الگوریتم HD [2] که در پیاده‌سازی آن نیز تغییراتی داده شده است به عنوان الگوریتم پایه استفاده شده با تخصیص وزن به یالها با استفاده از معیارهای مشابهت تاثیر آن را در تشخیص جوامع و اندازه گیری دقت تشخیص با ابزارهایی که معرفی خواهیم کرد، بیان می‌کنیم.

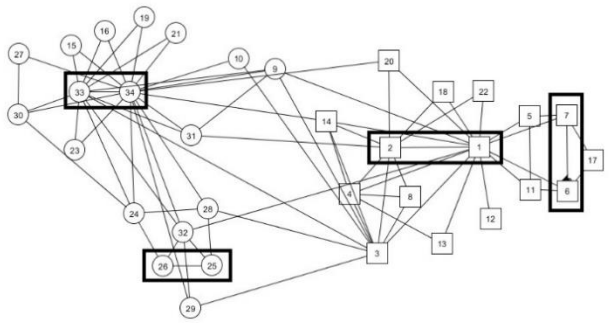
همچنین الگوریتم دیگری که بررسی خواهد شد، الگوریتم EM-BOAD [21] می‌باشد. در این الگوریتم مانند روش قبل از روشی متفاوت با الگوریتم اصلی، هسته جوامع کشف می‌شود. سپس هر هسته جوامع خود را ، توسط درجات مختلف جذب می‌کند. ما در روش خود در این الگوریتم، رویه کشف هسته‌ها را تغییر داده و همچنین با استفاده از برخی معیارهای شباهت در محاسبه درجه جذب کنندگی $A(C_s^x, i)$، سعی در افزایش دقت کشف جوامع داریم.

۵-۱-الگوریتم WHD

علت انتخاب و استفاده از الگوریتم HD این بود که علاوه بر پیچیدگی زمانی خطی آن، روش مناسبی برای شناسایی جوامع همپوشان می‌باشد. همچنین بر خلاف برخی از الگوریتم‌های موجود که تا اندازه ای به گره‌های اولیه که کاتوره ای انتخاب می‌شوند بستگی دارند، این روش مستقل از انتخاب کارتوره ای اولیه است. این الگوریتم به خوبی ساختار سلسله مراتبی را در یک شبکه آشکار می‌کند.

تفکر اساسی در این روش این است که اگر دو گره همسایه بیشترین تعداد همسایه را با هم داشته باشند به احتمال زیاد هم جامعه هستند. بنابراین هسته[4] جامعه به صورت یک جفت گره که با یکدیگر پیوند دارند و هر گره از این مجموعه بیشترین تعداد همسایه مشترک را با گره دیگر این مجموعه را دارد و همچنین تعداد همسایه‌های مشترک میان هر گره هسته و هر گره همسایه آن‌که بیرون هسته قرار دارد، کوچکتر و یا مساوی این مقدار بیشینه است.

برای روشن شدن موضوع، شبکه اعضای باشگاه کاراته زاخاری (شکل ۳) را در نظر بگیرید. به عنوان نمونه گره‌های ۶ و ۷ بیشترین همسایه مشترک نسبت به دیگر گره‌ها دارند(۲ همسایه مشترک) و همچنین این عدد بزرگتر و یا مساوی همسایه های مشترک ۶ و ۷ با دیگر گره‌های خارج از هسته که با این گره‌ها ارتباط دارند، می‌باشد. دیگر هسته‌های یافت شده زوج های (۱و۲) و (۳۴و۳۵) و (۲۵و۲۶) به عنوان هسته می‌باشد.



شکل ۳: هسته های کشف شده در شبکه اعضای باشگاه کاراته زاخاری[18]

در مقاله اصلی، مقایسه بیشترین همسایه مشترک دو گره هسته با تعداد همسایه های مشترک میان هر گره از این مجموعه و هر گره دیگر بیرون از این مجموعه انجام می‌گیرد. در روش پیشنهادی ما، صرفا مقایسه با همسایه‌های مشترک اعضای هسته با گره‌هایی که با آن‌ها

ارتباط دارند انجام می‌گیرد. که این امر سربار محاسباتی را در گرافهای بزرگ به طور قابل ملاحظه‌ای کاهش می‌دهد.

پس از شناسایی هسته‌ها، با استفاده از تابع جذب، که هر هسته گره‌های همسایه خود را نیروهای متفاوتی جذب می‌کند، گسترش یافته و جوامع تشکیل خواهد شد.

در مقاله اصلی، توان جذب[5] معرفی شده به صورت زیر است:

$$f_s = \frac{w_b}{\sum k_s \sum k_n} \quad \text{(فرمول شماره ۱)}$$

که در آن w_b تعداد یالهای میان هسته و همسایه های آن $\sum k_s$ مجموع درجات اعضای هسته s و $\sum k_n$ مجموع درجات اعضای همسایه می‌باشد.

در روش پیشنهادی ما برای هر یال وزنی پیشنهاد می‌شود که برای وزن‌دهی به یالها روش‌های مختلفی وجود دارد که ما در روش خود از معیارهای مشابهت مانند معیار Jaccard[6] یا HP مطابق با جدول ۴-۱ استفاده کردیم. بنابراین توان جذب به صورت زیر معرفی می‌شود:

$$f_s = \frac{w_b \times w_{s,i}}{\sum k_s \sum k_n} \quad \text{(فرمول شماره ۲)}$$

که $w_{s,i}$ وزن یال بین گره i و هسته s می‌باشد.

جدول شماره ۱: تعریف معیارهای شباهت محلی، Γ_x و Γ_y همسایه‌های راس های x و y و k_x و k_y درجات رئوس x و y (Ju et al., 2016)

Name	Definition	Name	Definition
CN	$s_{xy}^{CN} = \lvert\Gamma\ x \cap \Gamma\ y\rvert$	Sørensen	$s_{xy}^{Sorensen} = \frac{\lvert\Gamma\ x \cap \Gamma\ y\rvert}{(k_x+k_y)/2}$
Salton	$s_{xy}^{Salton} = \frac{\lvert\Gamma\ x \cap \Gamma\ y\rvert}{\sqrt{k_x \times k_y}}$	HP	$s_{xy}^{HP} = \frac{\lvert\Gamma\ x \cap \Gamma\ y\rvert}{\min\{k_x,k_y\}}$
Jaccard	$s_{xy}^{Jaccard} = \frac{\lvert\Gamma\ x \cap \Gamma\ y\rvert}{\lvert\Gamma\ x \cup \Gamma\ y\rvert}$	HD	$s_{xy}^{HD} = \frac{\lvert\Gamma\ x \cap \Gamma\ y\rvert}{\max\{k_x,k_y\}}$

در معیار مشابهت محاسبه شده در هر مرحله یکی از گره‌هایی که در نظر گرفته می‌شود هسته همسایه گره‌ها می‌باشد که با گسترش هسته مقادیر آن نسبت به همسایه‌های آن تغییر می‌کند.

برای شناسایی جوامع همپوشان در یک شبکه، پس از یافتن هسته‌ها می‌توانیم به طور جداگانه هر کدام از آن‌ها را به طور مستقل گسترش بدهیم. در این حالت هر یک از هسته‌ها به عنوان یک گره فرض شده و در هر بار پس از محاسبه توان جذب برای هر هسته، بیشینه مقدار آن را محاسبه می‌کنیم. هر هسته گره‌هایی که توان جذب برای آن‌ها بیشینه است را جذب می‌کنند. برای اینکه یک هسته تمامی گره‌ها را در تکرارهای مختلف در خود فرو نبرد، یک مقدار آستانه، α، در نظر می‌گیریم. در صورتی که توان جذب یک هسته روی همه همسایه هایش کم‌تر از α شد، فرایند گسترش آن متوقف می‌شود.

۲-۵-الگوریتم WHD-EM

الگوریتم EM-BOAD که قبلا معرفی شد روشی که براساس پیدا کردن هسته جوامع و توسعه آن به روش دیگری به کشف جوامع می‌انجامد.

در الگوریتم EM-BOAD برای استخراج هسته جوامع، به این صورت عمل می‌شود که اگر برخی از گره‌ها از ویژگی زیر، دو مورد را داشته باشند می‌توانند هسته جامعه باشند:

۱) گره‌هایی که درجه آن‌ها با هم برابر باشد.

۲) وزن راس ار رئوس همسایه خود بزرگتر از وزن راس آن‌ها نیست.

۳) بعضی از رئوس تشکیل زیر گراف کامل می‌دهند.

با توجه به اینکه مساله پیدا کردن بزرگترین زیرگراف کامل در گراف از مسائل NP-hard می‌باشد، در روشی که ارائه دادیم بخش اول الگوریتم EM-BOAD را تغییر داده و از روش یافتن هسته جوامع که در الگوریتم WHD ، معرفی کردیم و پیچیدگی زمان آن خطی است، استفاده خواهیم کرد.

در مرحله دوم الگوریت. که محاسبه درجه جذب کنندگی هر هسته می‌باشد در ابتدا برای هر یال متناسب با گره‌های دو سر آن وزن مناسب تخصیص می‌دهیم. در این الگوریتم ما از انواع معیارهای مشابهت به عنوان وزن یال‌ها بهره می‌گیریم. بنابراین در صورتی که i و j به هم متصل باشند، مقدار پارامتر E_{ij} را برابر w_{ij} ،وزن تخصیص داده شده برای هر یال، در نظر میگیریم. در غیر اینصورت مقدار E_{ij} را صفر در نظر می‌گیریم.

سپس برای هر راس گراف وزنی متناسب با E_{ij} به صورت زیر تعریف می‌کنیم:

(فرمول شماره ۳)

$$V_{i(i\epsilon V)} = \sum_{j\epsilon V} E_{ij} \qquad j\ is\ the\ neighboring\ vertices\ of\ i$$

هر هسته C_s^j همسایگان خود، مانند i ، را توسط درجات مختلف جذب می‌کند که رابطه درجه جذب کنندگی، $A(C_s^j, j)$ ،را که در الگوریتم EM-BOAD قبلا ارائه شده بود به صورت زیر استفاده می‌کنیم:

$$A(C_s^j, i) = \frac{\sum_{i\in V,\ j\in C_s^x} E_{ij}}{V_i} \quad \text{(فرمول شماره ۴)}$$

که $\sum_{i\in V,\ j\in C_s^x} E_{ij}$ مجموع وزن تمام یال‌هایی است که بین هسته جوامع C_s^j و رئوس همسایه مانند i وجود دارد و V_i درجه راس i می‌باشد.

پس از محاسبه $A(C_s^x, i)$ برای هر هسته به همراه همسایگان آن هسته، بیشینه مقدار $A(C_s^x, i)$ برای هر راس به ازای هسته های مختلف بدست می‌آوریم. در صورتی که این مقدار بیشینه برای هر راس، از یک مقدار α بیشتر شد، آن راس جذب هسته ای می‌شود $A(C_s^x, i)$ به ازای آن هسته برای آن راس بدست آمده است. در صورتی که این مقدار بیشینه برابر α شد، راس مورد نظر مشترک بین هسته‌ها می‌باشد و درصورتی که کم‌تر از مقدار α باشد. در این مرحله به هیچ هسته‌ای جذب نمی شود.

در تکرار بعد دوباره $A(C_s^x, i)$ برای هسته‌ها و همسایه‌های آن‌ها محاسبه و مراحل قبل دوباره اجرا خواهد شد. در روشی که ارائه داده ایم شرط پایان تکرار را به این صورت در نظر گرفته ایم که هیچ گره ای در این مرحله جذب هسته ای نشود.

در الگوریتم EM-BOAD مقدار $\alpha = 0.5$ و ثابت در نظر گرفته شده است. اما ما در روش پیشنهادی خود، در ابتدا $\alpha = 0.2$ در نظر گرفتیم و در هر باز انجام میانگین $A(C_s^x, i)$ را محاسبه کرده و در صورتی که این میانگین از α قبلی کم‌تر باشد، در مرحله بعد این میانگین را جایگزینα کرده ادامه می‌دهیم. تکرار مراحل تا وقتی ادامه می‌یابد که در مرحله جذب گره‌ها به هسته جوامع، هیچ گره‌ای جذب نشود.

برای روشن‌تر شدن موضوع شبکه باشگاه کاراته زاخاری را در نظر بگیرید. با انجام مرحله اول ۴ هسته برای این شبکه کشف می‌شود. $A(C_s^x, i)$ برای چند گره و هسته های کشف شده در جدول شماره ۲ مشاهده می‌شود. پس از محاسبه $A(C_s^x, i)$ و تشخیص بیشنه آن برای هر راس با توجه به شرط α، جذب یا عدم جذب مشخص می‌شود. برای هسته هایی که با راس مورد نظر همسایه نیستند، مقادیر مشخص نشده است.

جدول شماره ۲: مقادیر $A(C_s^x, i)$ برای چند راس انتخابی برای شبکه زاخاری در باز انجام اول

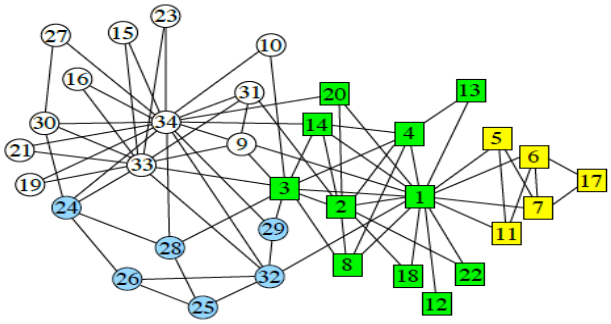
I	C_s^1	C_s^2	C_s^3	C_s^4	وضعیت جذب	جذب	هسته کننده
1	0	0.1109	0	0	جذب نمی شود		-
2	0	0	0	0	جذب نمی شود		-
3	0.3231	0	0	0.0305	جذب	هسته ۱	
4	0.3395	0	0	0	جذب	هسته ۱	
5	0.2429	0.3441	0	0	جذب	هسته ۲	
6	0.137	0	0	0	جذب نمی شود		-
7	0.137	0	0	0	جذب نمی شود		-
8	0.4046	0	0	0	جذب	هسته ۱	
9	0.0622	0	0	0.391	جذب	هسته ۴	

۶- نتایج

هنگامی که یک الگوریتم شناسایی جوامع پیشنهاد می‌شود ، باید کارایی آن آزمایش شده و با الگوریتم‌های دیگری که در این زمینه وجود دارد، مقایسه شود.

۶-۱- داده‌های آزمون

ابتدا برای ارزیابی و تحلیل کارایی روش‌های پیشنهادی، آن را روی دیتا است واقعی شبکه اعضای باشگاه کاراته زاخاری پیاده خواهیم کرد. همان‌طوری که در قبلاً اشاره شد مجموعه Zachary Karate Club، یک مجموعه داده کلاسیک برای تحلیل الگوریتم‌های شبکه‌های اجتماعی است، که بر اساس تعاملات واقعی بین اعضای یک کلوب کاراته در یکی از دانشگاه های آمریکا، توسط وین زاخاری تهیه شده است. این شبکه دارای ۳۴ گره و ۷۸ یال و دو اجتماع است.



شکل 4: باشگاه کاراته زاخاری، یک معیار استاندارد در تشخیص جامعه. رنگ‌ها بر اساس بهترین افرازبندی که با الگوریتم بهینه سازی ماژولاریتی (Newman and Girvan) بدست آمده است[18]

از آنجا که این اجتماعات در دنیای واقعی هویت داشته اند، می‌توانند مبنایی بر صحت تحلیل‌ها باشند از طرفی تعداد بسیار کم اعضای این شبکه‌ها مانع اثبات جامع صحت الگوریتم‌ها می‌باشد، بنابراین در بیشتر مقالاتی که روش تازه‌ای پیشنهاد شده است، بخش آزمون روش ، انجام الگوریتم روی مجموعه‌ی کوچکی از گراف‌های آزمون ساده است که شناسایی ساختار اجتماعی آن‌ها تا اندازه ای ساده می‌باشد.

آزمایش یک الگوریتم به معنی اجرای آن روی یک مسئله‌ی ویژه با پاسخ معلوم و مقایسه این پاسخ با پاسخ به دست آمده از اجرای الگوریتم می‌باشد.

بنابراین باید جوامع به دست‌آمده به روش علمی با جوامع از پیش معلوم مجموعه‌ای از گرافهای آزمون که می‌توان به آن‌ها اعتماد کرد، همخوانی داشته باشند. برای همین گراف‌های تولید شده با رایانه استفاده می‌شوند، که در آن ساختار خوشه‌ای درون آن‌ها را می‌توان به دلخواه طراحی نمود.

یکی از کارهای ارزنده در زمینه‌ی تولید گراف‌های آزمون را لنسچینتی، فورچوناتو و رادیچی دادند که به آزمون LFR معروف می‌باشد[22]. آن‌ها تابع پخش درجه و تابع پخش اندازه‌ی جامعه را به شکل توانی و به ترتیب با نماهای γ و β گرفتند. هر گره $1-\mu$ بخش از یال‌های خود را با بقیه‌ی رئوس جامعه‌ی خود و μ بخش از یال‌های خود را با گره‌های جوامع دیگر به اشتراک می‌گذارد. μ که همواره عددی میان صفر و یک می‌باشد، پارامتر آمیختگی[۷] نامیده می‌شود.

گراف‌ها به شیوه‌ی زیر ساخته می‌شوند:

۱) دنباله‌ای از درجات و دنباله‌ای از اندازه‌های جوامع با تابع پخش توانی تعیین‌شده ساخته می‌شوند.

۲) به هر گره درجه‌ای از دنباله به شکل مجموعه‌ای از نیم یال‌های پیوند داده به آن که کاتوره ای به نیم یال‌های رئوس دیگر می‌پیوندند، واگذار می‌کنیم تا گرافی با دنباله‌ی درجات معلوم را بسازند.

۳) رئوس درون جوامع جای می‌گیرند به یا گونه که دست کم قیود توپولوژیکی برآورده شوند.

۴) کمی بازنویسی و حفظ دنباله‌ی درجات، درجه‌ی درونی هر رأس v در جامعه‌ی خود چنان انتخاب می‌شود که تا جایی که می‌تواند به اندازه $(1-\mu)k_v$ که در آن k_v درجه راس v می‌باشد، نزدیک شود.

ما برای تولید گراف‌های آزمون از نرم‌افزاری که در مقاله [22] معرفی شده استفاده نمودیم.

۲-۶- سنجش افرازها

بررسی درستی اجرای یک الگوریتم برای یافتن جوامع در یک گراف بستگی به تعریف معیاری دارد که نشان می‌دهد افراز به دست آمده از الگوریتم تا چه اندازه با افراز مورد انتظار، یکسان است. معیارهای بسیاری برای همسانی افرازها وجود دارد. در اینجا به معرفی یکی از فراگیرترین آن‌ها که لئون دنون[۸] و همکارانش پیشنهاد کرده‌اند می‌پردازیم(Leon, Albert, Jordi, & Alex, 2005). پیشنهاد آن‌ها به‌کارگیری معیار-ی اطلاعات متقابل نرمال شده[۹] می‌باشد که NMI نامیده می‌شود. تعریف NMI بر پایه‌ی ماتریس اغتشاش[۱۰] N، استوار است که سطرهای آن با جوامع واقعی و ستونهای آن با جوامع به‌دست آمده همخوانی دارد. اعضای N ، که آن‌ها را با $N_{i\,j}$ نمایش می‌دهیم. تعداد گره‌های جامعه ی واقعی i هستند که در جامعه‌ی به‌دست آمده j آمده‌اند. تعداد جوامع واقعی را با c_A و تعداد جوامع به‌دست آمده را با c_B نشان می‌دهیم. اگر N تعداد کل گره‌ها باشد، یک معیار برای همسانی افرازها، بر پایه نظریه اطلاعات از رابطه زیر بدست می‌آید:

(۷)

$$NMI(A,B)=\frac{-2\sum_{I-1}^{c_A}\sum_{j=1}^{c_B}N_{i\,j}log(\frac{N_{i\,j}\,N}{N_i\,N_j})}{\sum_{i=1}^{c_A}N_i\,log(\frac{N_i}{N})\;+\;\sum_{j=1}^{c_B}N_j\,log(\frac{N_j}{N})}$$

اگر جوامع به دست آمده با جوامع واقعی یکسان باشند NMI بیشترین اندازه‌ی خود یعنی عدد یک را به خود می‌گیرد و اگر جوامع به دست آمده از الگوریتم روی هم رفته مستقل از جوامع واقعی باشند، برای نمونه اگر کل گراف به شکل یک جامعه به دست آید، MNI برابر صفر خواهد بود. این معیار، معیار خوبی برای اندازه‌گیری دقت اجرای یک روش می-باشد و اگر اجرای الگوریتمی مشکوک باشد این معیار می‌تواند کیفیت آن را مشخص کند.

معیار خلوص[۱۱] به عنوان یکی از معیارها برای ارزیابی کیفیت اجتماعات استفاده شده است[23]. این معیار برای هر جامعه یک کلاس در نظر می‌گیرد و در هر کلاسی بیشترین تعداد راس هایی که به درستی در آن کلاس قرار دارد را محاسبه می‌کند. اگر $G=\{G_1,G_2,\dots,G_s\}$ اجتماعات واقعی شبکه و $C=\{C_1,C_2,\dots,C_k\}$ جوامع کشف شده توسط الگوریتم مورد نظر باشد. خلوص جامعه C_i به صورت زیر تعریف می‌شود:

(فرمول شماره ۵) $Purity(C_i)=\frac{1}{|C_i|}max_j\{C_i\cap G_j\}$

معیار خلوص جوامع C به صورت زیر تعریف می‌شود:

(فرمول شماره ۶) $Purity(C)=\frac{1}{k}\sum_{i=1}^{k}Purity(C_i)$

با این که دقت، مسئله‌ی مهمی هنگام انتخاب یک الگوریتم می‌باشد، پیچیدگی زمانی آن نیز به همان اندازه مهم است. هنگام گزینش یک الگوریتم باید عوامل بسیاری را در نظر بگیریم و در پایان شایسته‌ترین روش را برای مسئله‌ی داده شده برگزینیم. بیشتر مواقع سازشی میان دقت و زمان اجرا انجام می‌گیرد، به‌ویژه هنگامی‌که شبکه‌ی بررسی‌شونده بزرگ باشد. یافتن روشهایی که جوامع یک شبکه را با دقت خوب شناسایی کنند.

برای بررسی روش پیشنهادی ان را روی شبکه‌های آزمون LFR با پارامترهای مشخص می‌سنجیم. همچنین دقت روش پیشنهادی خود را با بک بکار گیری سنجه NMI بررسی خواهیم کرد.

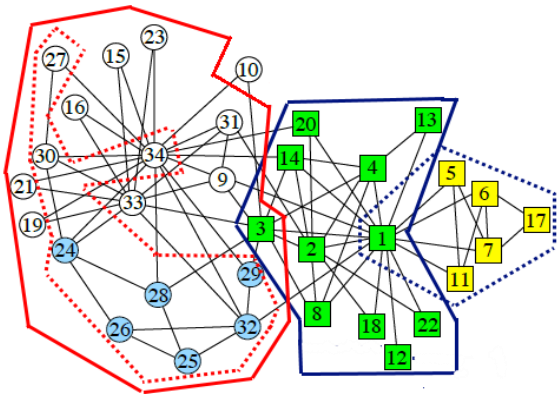
۳-۶- استفاده از معیار مشابهت به عنوان وزن یال در الگوریتم WHD

با به کارگیری الگوریتم WHD با استفاده از معیار وزن‌دهی بر اساس معیار مشابهت Jaccard روی شبکه اعضای باشگاه زاخاری که دارای ۳۴ گره و ۷۸ یال و دو اجتماع اصلی می‌باشد، به چهار جامعه که از ترکیب آن‌ها دو جامعه اصلی کشف می‌شود، رسیدیم. جوامع کشف شده که به صورت

$C_1=\{1,2,18,22,12,20,8,4,13,14,3,6,7,17,5,11,1\}$
$C_2=\{6,7,17,5,11,1\}$
$C_3=\{25,26,28,32,24,29,30,27,34\}$
$C_4=\{33,34,15,16,19,21,23,30,27,24,28,32,25,26,29,9,31,3\}$

می‌باشد. همان‌طور که در شکل 5 مشاهده می‌شود جوامع C_3 و C_4 به صورت سلسله مراتبی هستند که جامعه بزرگتر یعنی C_4 ، یکی از جوامع اصلی جامعه واقعی می‌باشد و ترکیب جوامع کشف شده، C_2 و C_1 هم یکی دیگر از جوامع اصلی شبکه واقعی را تشکیل می‌دهند. همچنین همپوشانی بین جوامع C_1 و C_4 وجود دارد و گره ۳ به صورت مشترک متعلق به این جوامع کشف شده، وجود دارد. نتایج حاصل به جوامع کشف شده توسط الگوریتم MAGA [24] بسیار نزدیک است. معیار خلوص محاسبه شده برای این حالت 0.9861 می‌باشد.

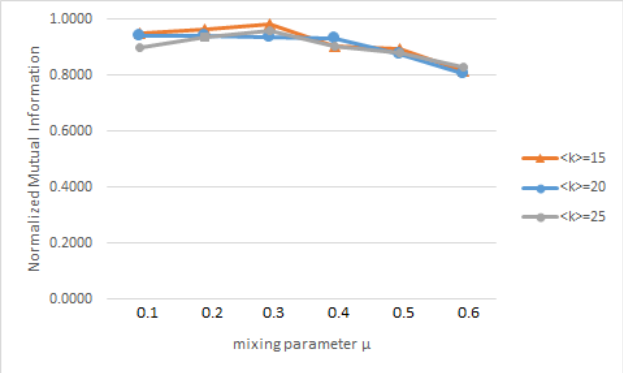
دلیل جذب نشدن راس ۱۰ نداشتن همسایه مشترک با هیچ یک از هسته‌ها می‌باشد. گره ۱۰ عامل اتصال دو جامعه به هم است.



شکل 5: جوامع بدست آمده برای شبکه اعضای باشگاه کاراته زاخاری با به کارگیری الگوریتم WHD با $\alpha=.0035$ با معیار شباهت Jaccard

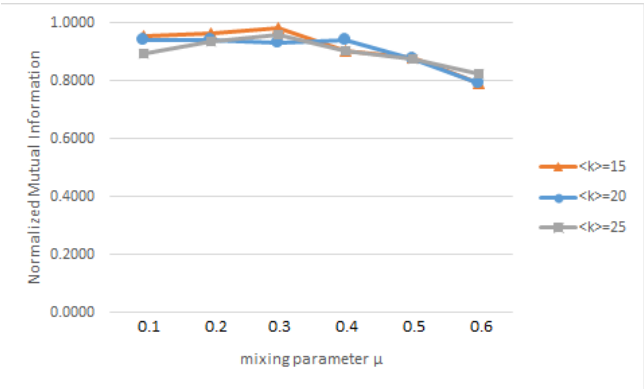
با به کارگیری معیار NMI می‌توان کیفیت جوامع به دست آمده با روش پیشنهادی را بررسی کرد. ابتدا روش پیشنهادی WHD با استفاده از معیار شباهت Jaccard، روی شبکه LFR با ۱۰۰۰ گره و به ازای

پارامترهای آمیختگی مختلف که عددی بین صفر و ۱ می‌باشد و نشان دهنده درصد یال‌هایی است که بین جوامع مشترک است و میانگین درجه گره‌های مختلف (k) پیاده سازی کردیم. شکل 6، نمودار NMIبر حسب پارامتر آمیختگی می‌باشد که نشان دهنده دقت بالای شناسایی جوامع این شبکه می‌باشد.



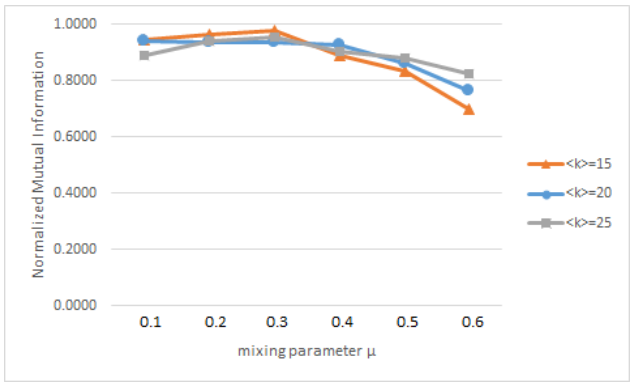
شکل 6: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD با استفاده از وزن‌دهی با معیار مشابهت Jaccard بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.

شکل 7 نمودار NMI بر حسب پارامتر آمیختگی نتیجه اجرای همان الگوریتم با معیار شباهت HP بر روی همان شبکه آزمون می‌باشد. مقایسه نمودار شکل 8 و نمودار شکل 7 نشان می‌دهد که انتخاب معیار برای وزن‌دهی در دقت تشخیص جوامع موثر است. که در الگوریتم WHD، انتخاب معیار مشابهت Jaccard خصوصا در $\mu \geq 0.5$ نتیجه بهتری را ایجاد می‌کند.



شکل 7: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD با استفاده از وزن‌دهی با معیار مشابهت HP بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.

برای مقایسه دقت الگوریتم در حالتی که وزن به یال‌ها تخصیص داده می‌شود با حالتی که وزن برای یال‌ها در نظر گرفته نمی شود. الگوریتم پایه بدون در نظر گرفتن وزن بر روی همان داده‌ها اجرا شد. نتیجه معیار NMI در نمودار شکل 8 مشاهده می‌شود.

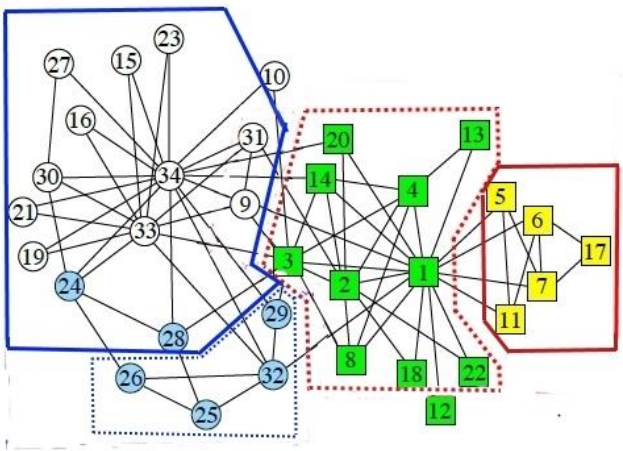


شکل 8: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD بدون تخصیص وزن بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.

مقایسه شکل 6 و 7 که در آن روش پیشنهادی WHD با تخصیص وزن به یال‌ها اجرا شده، با نمودار شکل 8 که همان الگوریتم بدون در نظر گرفتن وزن مناسب برای یال‌ها روی همان داده‌ها پیاده شده، نشان می‌دهد که در نظر گرفتن وزن برای یال‌ها با یکی از معیارهای مشابهت، دقت شناسایی جوامع را افزایش می‌دهد.

۴-۶- استفاده از معیار مشابهت به عنوان وزن یال در الگوریتم WHD-EM

اجرای الگوریتم WHD-EM با استفاده از معیار مشابهت HD به عنوان وزن یال‌ها در شبکه باشگاه کاراته زاخاری با $\alpha = 0.3$ ، چهار جامعه کشف گردید که با نتایج با الگوریتم بهینه‌سازی ماژولاریتی (Newman and Girvan) سازگاری زیادی داشت. در صورتی که اگر جوامع واقعی در این الگوریتم جوامع یافت شده شکل 4 فرض کنیم، شاخص NMI=0.8926 و معیار خلوص purity=0.9615 می‌باشد که نشان دهنده دقت شناسایی بالای جوامع در این شبکه را دارد. اما با این ضریب α همپوشانی بین جوامع وجود ندارد.(شکل 9)

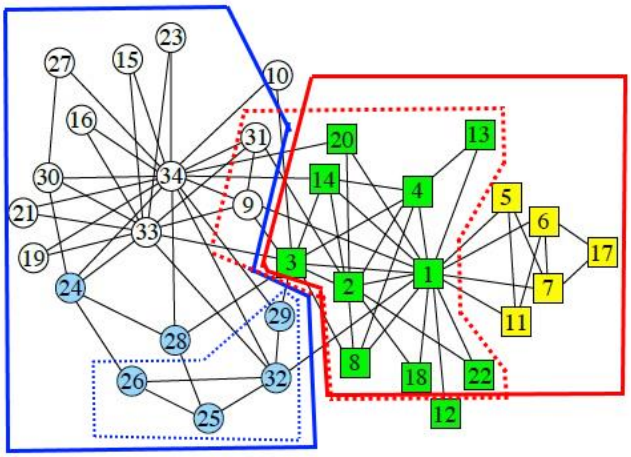


شکل 9: جوامع بدست آمده برای شبکه اعضای باشگاه کاراته زاخاری با به کارگیری الگوریتم WHD-EM با $\alpha = 0.3$ با معیار شباهت HP

برای افزایش جذب هسته‌ها و تشکیل جوامع همپوشان مقدار αرا کاهش می‌دهیم. با $\alpha = 0.23$ دو جامعه واقعی که در مقاله زاخاری معرفی شده بود (راس‌های مربع و دایره‌ای) با معیار خلوص purity=.9289 شناسایی

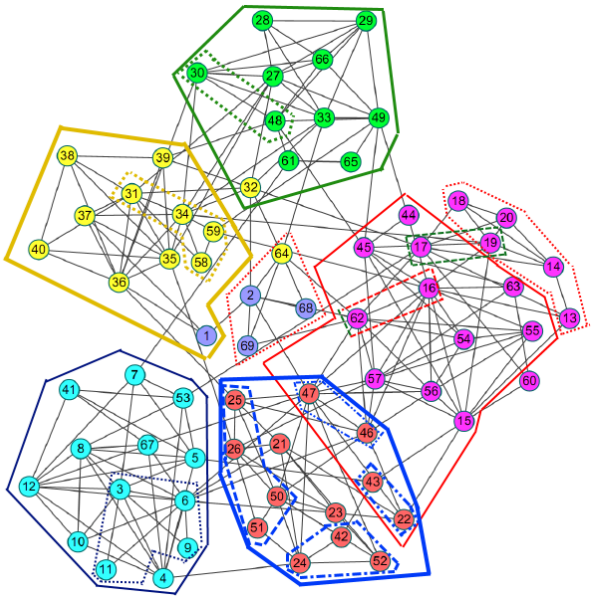
می‌شود، که درون هر یک از آن‌ها جوامع کوچکتر که سازمان‌یافتگی سلسله مراتبی را دارند تشکیل شده اند(شکل ۷). همان‌طور که در شکل دیده می‌شود، دو جامعه اصلی شبکه باشگاه کاراته زاخاری به جز گره شماره ۱۰ و ۱۲ به درستی شناسایی شده است. دلیل عضو نشدن، این اعضا در هیچ جامعه‌ای این است که چون آن‌ها با هیچ یک از اعضای هسته‌ها همسایه مشترک ندارند بنابراین در وزن تخصیصی که بر اساس همسایه‌های مشترک است نیروی جاذبه هسته‌ها برای این گره‌ها صفر باشد.

مقایسه جوامع به‌دست آمده با جوامع واقعی نشان می‌دهد اولا جوامع کشف شده با یکدیگر همپوشانی دارند و خاصیت سلسله مراتبی در جوامع کوچکتر وجود دارد.



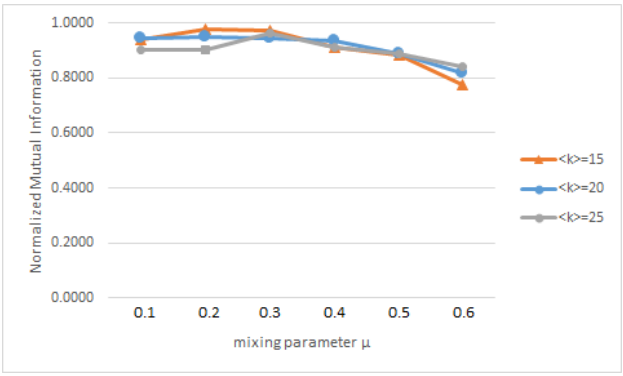
شکل 10: جوامع بدست آمده برای شبکه اعضای باشگاه کاراته زاخاری با به کارگیری الگوریتم WHD-EM با $\alpha = 0.23$ با معیار شباهت HP

یکی دیگر از جوامع واقعی که در اغلب مقالات بررسی می‌شود، شبکه دوستی High school network می‌باشد که توسط تحقیقات موسسه ملی بهداشت کودکان و توسعه انسانی براساس گزارش دانش آموزان‌که به ویژگی‌های نمره، نژاد و جنسیت توجه شده تهیه شده است. این شبکه دارای ۶۹ راس می‌باشد که با اجرای الگوریتم WHD-EM با مقدار $\alpha = 0.42$ جوامع را با دقت NMI=0.6652 و معیار خلوص purity=0.9845 کشف می‌کند که از بعضی از الگوریتم‌ها مانند NMF، CPM، GAME، GCE، OSLOM و LINK که در مقاله[24] بررسی شده، از نظر معیار NMI بهتر است.

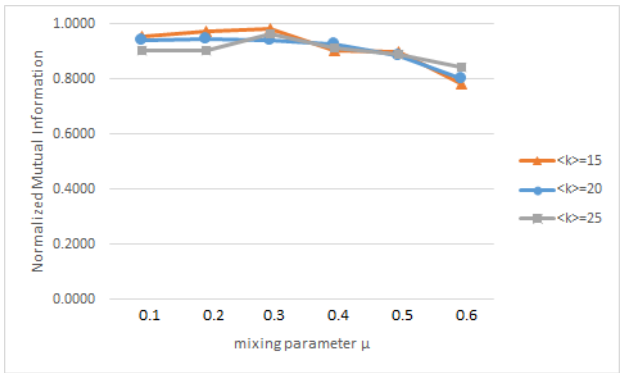


شکل ۱۱: جوامع بدست آمده برای شبکه High school network با به کارگیری الگوریتم WHD-EM با $\alpha = 0.3$ با معیار شباهت HP

نتایج اجرای الگوریتم WHD-EM با استفاده از وزن‌دهی بر اساس معیارهای Jaccard و HP، بر روی داده‌های شبکه مصنوعی LFR با ۱۰۰۰ راس با میانگین درجات متفاوت و اندازه‌گیری NMI بر حسب پارامتر آمیختگی (μ) در نمودارهای شکل‌های ۱۲ و ۱۳ مشاهده می‌شود.



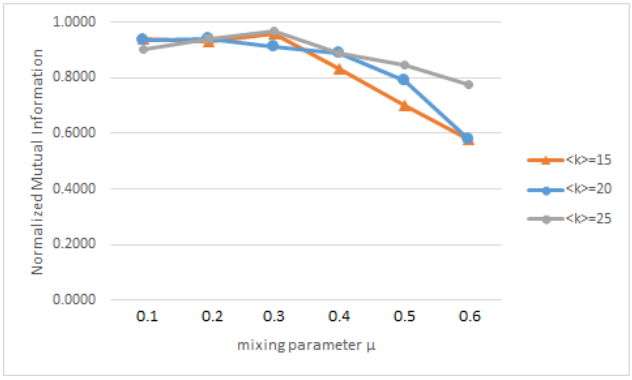
شکل ۱۲: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD-EM با استفاده از وزن‌دهی با معیار مشابهت Jaccard بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.



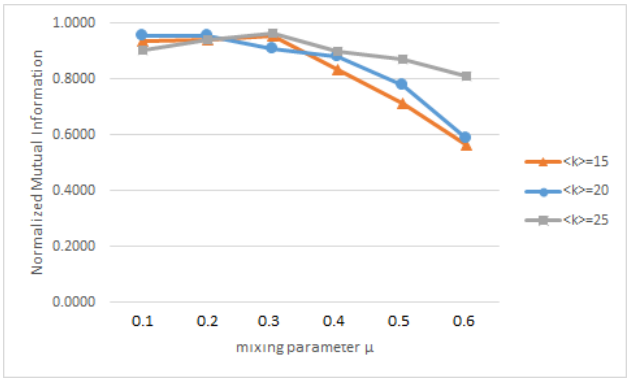
شکل ۱۳: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD-EM با استفاده از وزن‌دهی با معیار مشابهت HP بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.

با توجه به استفاده از داده‌های مشابه با الگوریتم WHD، مشاهده می‌شود، دقت محاسبه شده برای داده‌های آزمون با استفاده از الگوریتم EM-WHD مشابه نتایج الگوریتم WHD می‌باشد.

با کاهش مقدار α برای یافتن جوامع هم‌پوشان بیشتر در این الگوریتم، برای $\mu < 0.5$ دقت قابل قبول می‌باشد اما برای $\mu \geq 0.5$ دقت اندکی کاهش یافته و برای این داده‌ها دقت الگوریتم WHD بهتر است. (نمودار ۵-۶).



شکل ۱۴: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD-EM با α کاهش یافته برای افزایش تعداد جوامع همپوشان با استفاده از معیار Jaccard بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000می‌باشد.



شکل ۱۵: بررسی کیفیت جوامع بدست آمده با بکار گیری روش پیشنهادی WHD-EM با α کاهش یافته برای افزایش تعداد جوامع همپوشان با استفاده از معیار HP بر روی شبکه‌های آزمون LFR. تعداد گره‌های هر شبکه N=1000 می‌باشد.

۷- نتیجه‌گیری

در این پژوهش ابتدا الگوریتم HD که پیچیدگی زمانی آن خطی است و نیازی به انتخاب گره‌های اولیه و تعداد جوامع نیست انتخاب کرده، علاوه بر تغییراتی که ساختار انتخاب هسته جوامع در نظر گرفتیم، با تخصیص (یا بازسازی) وزن یال‌های گراف شبکه، دقت انتخاب جوامع را افزایش دادیم. اجرای این الگوریتم روی شبکه‌های واقعی و همچنین مقایسه دقت کشف جوامع با استفاده از معیار NMI، برای حالت وزن دار و غیر وزن‌دار، نشان می‌دهد که انتخاب وزن برای گراف باعث افزایش دقت شناسایی جوامع می‌شود.

سپس الگوریتم دیگری با عنوان EM-BOAD برگزیده و علاوه بر تعریف وزن برای یال‌های گراف شبکه، تغییراتی در الگوریتم پایه ایجاد نمودیم که نتایج آزمون‌ها در شبکه واقعی، نشان دهنده افزایش معیار خلوص و معیار NMI در شبکه باشگاه کاراته زاخاری و شبکه High school network بود به گونه‌ای که علاوه بر پیدا کردن جوامع همپوشان، معیار کیفیت NMI از برخی الگوریتم‌های مطرح بهتر شده است.

بررسی کیفیت جوامع بدست آمده در شبکه‌های مصنوعی، با به‌کارگیری روش پیشنهادی WHD-EM که از معیار های مشابهت مختلف استفاده می‌کنند، نشان می‌دهد که انتخاب روش و تابع وزن‌دهی به یال‌ها در دقت کشف جوامع موثر است.

در مجموع نتایج آزمون‌ها نشان می‌دهد که استفاده از معیارهای مشابهت محلی به عنوان وزن یال‌ها برای الگوریتم‌هایی که از درجه گره‌ها به عنوان یکی از ویژگی‌های شبکه برای محاسبه قدرت جذب هسته‌ها برای تشکیل جوامع استفاده می‌کنند، اثر مثبت دارد.

با توجه به اینکه یک همبستگی مثبت بین ساختارهای جامعه و معیارهای شباهت وجود دارد، نتایج آزمون‌های انجام شده نشان می‌دهد که استفاده از معیارهای مشابهت محلی به عنوان وزن یال‌ها برای برخی از الگوریتم‌ها باعث افزایش دقت تشخیص جوامع می‌شود. این الگوریتم‌ها از درجه گره‌ها به عنوان یکی از ویژگی‌های شبکه برای محاسبه قدرت جذب هسته‌ها برای تشکیل جوامع استفاده می‌کنند. به عنوان نمونه در مورد شبکه‌های واقعی، اجرای الگوریتم WHD-EM روی شبکه High school network، جوامع را با دقت NMI=0.6652 و معیار خلوص purity=0.9845 کشف می‌کند که از بعضی از الگوریتم‌ها مانند CPM، NMF، GAME، GCE، OSLOM و LINK از نظر معیار NMI بهتر است.

مراجع

[1]Gros, C. *Complex and adaptive dynamical systems : a primer, with 98 figures and 10 tables*. Berlin: Springer, 2008.

[2]Hakami Zanjani, A., & Darooneh, A. "Finding communities in linear time by developing the seeds". *Physical Review E, 84*(3), 2011: 036109.

[3]Fortunato, S. "Community detection in graphs". *Physics Reports, 486*(3–5), 2010: 75-174.

[4]Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. "Uncovering the overlapping community structure of complex networks in nature and society". *Nature, 435*(7043), 2005: 814-818.

[5]Rives, A. W., & Galitski, T. "Modular organization of cellular networks". *Proc Natl Acad Sci U S A, 100*(3), 2003: 1128-1133.

[6]Chen, J., & Yuan, B. "Detecting functional modules in the yeast protein-protein interaction network". *Bioinformatics, 22*(18), 2006: 2283-2290.

[7]Caldarelli, G., & Vespignani, A. *Large scale structure and dynamics of complex networks : from information technology to finance and natural science*. Singapore ; Hackensack, NJ: World Scientific, 2007.

[8]Guimera, R., & Nunes Amaral, L. A. " Functional cartography of complex metabolic networks". *Nature, 433*(7028), 2005: 895-900.

[9]Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., & Taylor, W. W. "Compartments revealed in food-web structure". *Nature, 426*(6964), 2003: 282-285.

[10]Reichardt, J. *Structure in complex networks*. Berlin: Springer. 2009.

[11]Xiang, J., Hu, K., & Tang, Y. “A class of improved algorithms for detecting communities in complex networks”. *Physica A: Statistical Mechanics and its Applications, 387*(13), 2008: 3327-3334.
[12]Ju, X., et. Al. “Enhancing community detection by using local structural information. *Journal of Statistical Mechanics: Theory and Experiment, 2016*(3), 20016: 033405.
[13]Kernighan, B. W., & Lin, S. “An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal, 49*(2), 1970: 291 - 307.
[14]Girvan, M., & Newman, M. E. “Community structure in social and biological networks”. *Proc Natl Acad Sci U S A, 99*(12), 2002: 7821-7826.
[15]Newman, M. E. J., & Girvan, M. “Finding and evaluating community structure in networks”. *Phys. Rev. E, 69*(2). 2004.
[16]Newman, M. E. J. “Fast algorithm for detecting community structure in networks. *Phys. Rev, 69*(6), 2004:
[17]Boettcher, S., & Percus, A. G. “Optimization with Extremal Dynamics”. *Physical Review Letters, 86*(23), 2001: 5211-5214.
[18]Zachary, W. W. “An Information Flow Model for Conflict and Fission in Small Groups”. *Journal of Anthropological Research, 33*(4), 1977: 452-473.
[19]Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. “Community detection in Social Media. *Data Mining and Knowledge Discovery, 24*(3), 2012: 515-554.
[20]Yang, J., McAuley, J., & Leskovec, J. “*Community Detection in Networks with Node Attributes”.* Paper presented at the 2013 IEEE 13th International Conference on Data Mining.
[21]Li, J., Wang, X., & Eustace, J. Detecting overlapping communities by seed community in weighted complex networks. *Physica A: Statistical Mechanics and its Applications, 392*(23), 2013: 6125-6134.
[22]Lancichinetti, A., Fortunato, S., & Radicchi, F. “Benchmark graphs for testing community detection algorithms”. *Physical Review E, 78*(4), 2008: 046110.
[23]Gui-Lan, S., & Xiao-Ping, Y. “ A Topic Community Detection Method for Information Network based on Improved Label” Propagation. *International Journal of Hybrid Information Technology, 9*(2), 2016: 299-310.
[24]Zhan, W., Guan, J., Chen, H., Niu, J., & Jin, G. “Identifying overlapping communities in networks using evolutionary method”. *Physica A: Statistical Mechanics and its Applications, 442*, 2016: 182-192.

پانویس

1. edge centrality
2. Extremal optimization
3. Dedicated server
4. Seed of the community
5. Absorption power

۶. اندیس ژاکار یا ضریب شباهت معیاری برای مقایسه شباهت یا تفاوت مجموعه نمونه‌های آماری است. میزان شباهت دو مجموعه نمونه با توجه به اندیس ژاکار، از تقسیم تعداد اشتراک دو مجموعه بر تعداد اجتماع دو مجموعه به دست می‌آید.

7. mixing parameter
8. Leon Danon
9. Normalized Mutual Information
10. confusion matrix
11. purity