کشف اجتماعات در شبکه‌های اجتماعی یک‌لایهٔ پویا با رویکرد بیشینه‌سازی اهمیت

علی سر آبادانی [۱]، سید مرتضی پورنقی [۲]، خیراله رهسپار [۳]
۱- *دانشجوی دکتری ، دانشگاه قم* ، تهران، ایران
۲- استادیار،دانشگاه جامع امام حسین (ع) ، تهران، ایران
۳- عضو هیئت‌علمی گروه مهندسی رایانه و *فناوری اطلاعات*، دانشکده فنی و مهندسی، *دانشگاه قم* ، تهران، ایران
(دریافت: ۱۴۰۲/۱۱/۲۵، بازنگری: ۱۴۰۳/۰۲/۱۳، پذیرش: ۱۴۰۲/۰۲/۲۶، انتشار: ۱۴۰۳/۰۳/۱۳)

4- DOR: https://dorl.net/dor/
-۵

چکیده

دنیای کنونی، عصر شبکه و اینترنت است عصری که در آن با شکل‌گیری شبکه‌های گوناگون اجتماعی، شیوه‌های ارتباطی و اطلاع‌رسانی نوینی به عرصهٔ گسترده ارتباطات اجتماعی معرفی شده‌اند. شبکه‌های اجتماعی از محبوب‌ترین مقاصد سال‌های اخیر کاربران اینترنت شده است[۱]شبکهٔ اجتماعی گروهی از افراد یا سازمان‌های دارای منافع و سلایق مشترک هستند که برای دستیابی به اهداف مشخصی گرد هم می‌آیند. دلایل عمده ایجاد شبکه‌های اجتماعی شامل روابط فردی، روابط کاری، روابط علمی، سلیقه و علایق و تفریحات مشترک و انگیزه‌های اجتماعی سیاسی می‌باشد. آنالیز یا تحلیل شبکه‌های اجتماعی به معنایی مطالعهٔ ویژگی‌های شبکهٔ اجتماعی و روابط بین افراد و بخش‌های یک شبکه با رویکرد نظریه شبکه‌ای یا گراف است. تحلیل شبکه‌های اجتماعی نوعی مطالعه میان‌رشته‌ای در حوزه‌های مختلف است از جمله جامعه‌شناسی، ریاضیات، علوم رایانه و سایبری می‌باشد. یکی از چالش‌های اساسی در تحلیل شبکه‌های اجتماعی ناپایدار بودن این شبکه‌ها است؛ زیرا در هر ثانیه ممکن است افرادی یا سازمانی‌هایی به عضویت این شبکه‌ها درآیند و یا از آن خارج شده و یا رابطه‌های جدیدی شکل گیرد.
ما در این مقاله یک روشی ابتکاری برای کشف اجتماعات پیچیده در شبکه‌های اجتماعی با یک‌لایهٔ پویا و با تمرکز بر بیشینه‌سازی معیار اهمیت ارائه می‌دهیم. سپس با استفاده از مجموعه‌دادهٔ تولید شده توسط LFR Benchmark به بررسی نتایج طرح پیشنهادی خود با سایر طرح‌ها پرداخته و نتایج به‌دست‌آمده حاکی از بهبود عملکرد طرح پیشنهادی ما در دقت شناسایی روابط پیچیده در شبکه‌های اجتماعی و زمان کمتر آن در مقایسه با سایر طرح‌ها است.

کلمات کلیدی

کشف اجتماعات، شبکه‌های اجتماعی پویا، شبکه یک‌لایه، اهمیت، اجتماع، بیشینه‌سازی

۱- مقدمه

به هر ساختار اجتماعی از افراد که بر اساس یک رابطهٔ اجتماعی ایجاد می‌شود، یک شبکهٔ اجتماعی گفته می‌شود؛ بنابراین هر شبکهٔ اجتماعی شامل مجموعه‌ای از انسان‌ها و روابط اجتماعی بین آن‌هاست؛ لذا هر شبکه اجتماعی از دو عنصر موجودیت‌های شرکت‌کننده در ارتباط و ارتباط بین این موجودیت‌ها تشکیل شده است [۲].

شبکه‌های اجتماعی از قرن نوزدهم موردتوجه دانشمندان حوزه جامع‌شناسی و علوم رایانه قرار گرفتند. پژوهش‌ها در این حوزه از دههٔ چهل به بعد با تعریف ابزارهایی چون گراف اجتماعی[۱] شتاب بیشتری گرفت. در سال ۱۹۹۴ میلادی واسرمن[1] با چاپ کتاب تحلیل شبکه‌های اجتماعی[۲] این زمینه از علم را وارد دورهٔ جدیدی کرد، و پس از آن شبکه‌های اجتماعی به‌صورت جدی در زیرمجموعه‌های علوم اجتماعی و ریاضی مورد بررسی قرار گرفت.

رشد شبکه‌های اجتماعی برخط، باعث شده که راهکارهایی جهت استخراج جوامع به وجود بیاید. آنچه که در تشخیص جوامع مدنظر می‌باشد و از اهداف استخراج آن به شمار می‌رود تعاملاتی است که افراد یک جامعه با هم دارند، به‌طوری‌که بیشترین ارتباط درون جامعه و روابط خارج از جامعه آن‌ها کمترین مقدار ارتباط را به خود اختصاص می‌دهد. پس با استفاده از تفکیک شبکه به جوامع گوناگون باتوجه‌به شباهت‌های هر گروه به مطالعه آن‌ها پرداخته می‌شود. هرچند استخراج و مطالعهٔ جوامع در شبکه‌های اجتماعی موضوع جدیدی نیست؛ ولی محدودیت‌ها، روش‌های تحقیق بیشتری را در این زمینه می‌طلبد.

مشکل مهمی که در روش‌های موجود وجود دارد عدم توجه به این نکته است که در یک شبکهٔ اجتماعی ممکن است چندین نوع رابطه بین افراد وجود داشته باشد. برای مثال در یک شبکهٔ اجتماعی علاوه بر رابطهٔ دوستی بین افراد روابطی مانند فامیل بودن نیز وجود دارد. این مسئله در سال‌های اخیر در مبحث کشف اجتماعات در شبکه‌های اجتماعی، موردتوجه پژوهشگران قرار گرفته است.

یکی دیگر از مشکلات مطرح در الگوریتم‌های کشف اجتماعات، پویا بودن شبکه‌هاست که در دنیای واقعی امری متداول است. در واقع نیاز به الگوریتم‌های کارایی برای به‌روزرسانی ساختار اجتماعات بعد از ایجاد تغییرات در روابط المان‌های شبکه ضروری به نظر می‌رسد. ما در این مقاله موفق به ارائه یک روش پویا برای شناسایی روابط پیچیده در شبکه‌های اجتماعی شده‌ایم.

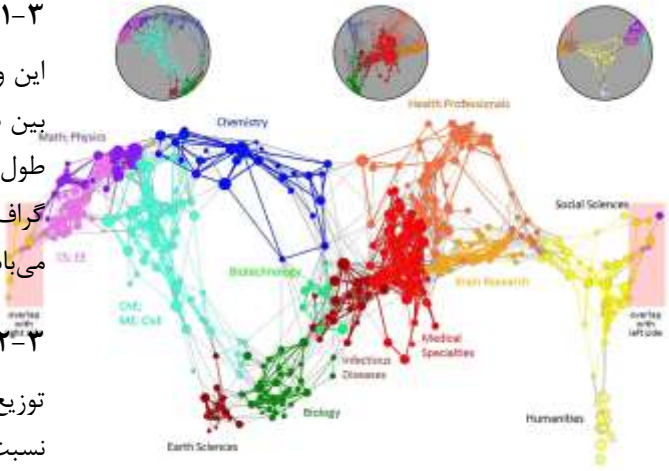
۲- شبکه اجتماعی

در این بخش به بررسی برخی از شبکه‌های پیچیده در دنیای واقعی می‌پردازیم. شبکه‌های رایانه‌ای شامل رایانه‌ها هستند که اتصالاتی مانند کابل یا ماهواره به هم متصل شده‌اند. انواع مختلفی از این شبکه‌ها مانند شبکهٔ تلفن و شبکهٔ حمل‌ونقل وجود دارند. در بسیاری از موارد اتصال بین گره‌ها فیزیکی نیست، برای مثال شبکهٔ WWW یک شبکهٔ جهت‌دار از صفحات وب است که اگر یک صفحه به صفحهٔ دیگر پیوند داشته باشد به هم متصل خواهند بود. مثال دیگر، شبکهٔ پست الکترونیک (رایانامه) است که در آن هر فرد با یک گره نمایش داده می‌شود و یک یال جهت‌دار زمانی وجود دارد که یک فرد پست الکترونیک فرد دیگری را در فهرست نشانی خود داشته باشد. نوع دیگری از شبکه‌ها، گراف تماس بین افراد است که معمولاً توسط اپراتورهای شبکهٔ تلفن ایجاد می‌شود. در این گراف، هر فرد با یک گره نمایش داده می‌شود و یک یال جهت‌دار زمانی وجود دارد که در یک بازهٔ معین یک فرد با فرد دیگر تماس برقرار کرده باشد.

نمونهٔ دیگر از شبکه‌های پیچیده، شبکه‌های اجتماعی هستند که تمرکز ما در این مقاله بر روی این شبکه‌ها است. این شبکه‌ها

[1] Wasserman

بیانگر ارتباط بین افراد مانند رابطهٔ دوستی یا رابطهٔ کاری هستند. برای مثال یکی از این شبکه‌ها شبکهٔ بازیگران است و دو بازیگر زمانی با هم مرتبط هستند که حداقل در یک فیلم مشترک بازی کرده باشند. یک مثال دیگر از شبکه‌های اجتماعی، شبکهٔ همکاری پژوهشگران است. در این شبکه هر پژوهشگر با یک گره نمایش داده می‌شود و دو پژوهشگر زمانی به هم متصل هستند که حداقل یک مقالهٔ مشترک با هم منتشر کرده باشند. یکی دیگر از شبکه‌های اجتماعی، گراف رجوع [۲] است که در آن هر پژوهشگر با یک گره نمایش داده می‌شود و زمانی بین پژوهشگران ارتباط وجود دارد که یک پژوهشگر در مقالهٔ خود به مقالهٔ فرد دیگر ارجاع داده باشد. شکل ۲-۱ نمونه‌ای از این نوع شبکه را نشان می‌دهد.



شکل ۲-۱ شبکهٔ ارجاع مقاله‌ها [۳]

۲-۱- نمایش شبکه‌های اجتماعی

برای مطالعه و تحلیل یک شبکهٔ اجتماعی نیاز به ساختارهایی قابل‌فهم است که افراد شبکه و روابط آن‌ها را نمایش دهد. معمولاً دو شیوه برای نمایش شبکه‌های اجتماعی مورداستفاده قرار می‌گیرد: گراف‌ها و ماتریس‌ها.

۲-۱-۱- ماتریس

می‌توان شبکه‌های اجتماعی را به‌صورت آرایش مربعی از اعداد نشان داد. در ماتریس هر سطر و ستون متعلق به یک فرد در شبکه است. در ماتریس مربوط به یک شبکه، a_{ij} برابر یک است اگر ارتباط موجود باشد، در غیر این صورت این مقدار صفر است.

[2] citation network

ماتریس‌های شبکه‌های اجتماعی لزوماً متقارن نیستند. یعنی ممکن است رابطهٔ بین افراد یک‌طرفه یا دوطرفه باشد.

۲-۱-۲- گراف

گراف‌ها یکی از ابزارهای نمایش بصری و تحلیل شبکه‌های اجتماعی هستند. در گراف هر گره به‌عنوان یک فرد در نظر گرفته می‌شود و وجود یال بین دو گره بیانگر وجود رابطه بین دو فرد متناظر با این دو گره است. گراف مورداستفاده برای یک شبکه بر اساس نوع شبکه می‌تواند گراف جهت‌دار یا بدون جهت باشد.

۳- ویژگی‌های ساختاری شبکه‌های اجتماعی

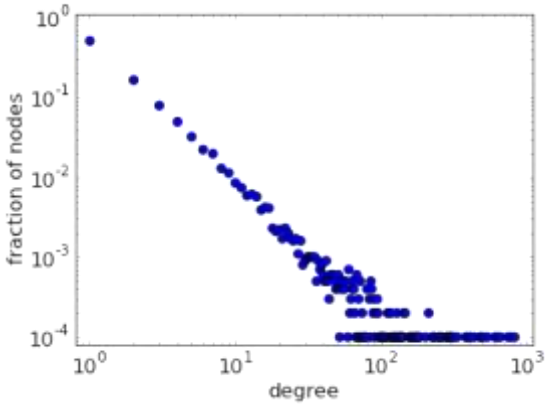
۳-۱- قطر

این ویژگی روی گراف شبکه‌های اجتماعی تعریف می‌شود. مسیر بین دو گره از گراف دنباله‌ای از یال‌های بین این دو گره است و طول مسیر به معنای تعداد یال‌های بین آن دو گره می‌باشد. قطر گراف برابر با طولانی‌ترین مسیر موجود بین دو گره از گراف می‌باشد.

۳-۲- توزیع درجه

توزیع درجه در گراف یک تابع $p(k)$ می‌باشد که نشان‌دهندهٔ نسبت تعداد گره‌های با درجهٔ k نسبت به تعداد کل گره‌های گراف است. توزیع درجه‌یکی از ویژگی‌های مهم برای تحلیل گراف می‌باشد. توزیع درجه رئوس در گراف شبکه‌های اجتماعی متفاوت با گراف‌های تصادفی است. در بسیاری از شبکه‌های واقعی از جمله شبکه‌های اجتماعی، توزیع درجهٔ رئوس گراف متناظر شبکه از توزیع قاعدهٔ توانی [۳] پیروی می‌کند. درصورتی‌که در گراف‌های تصادفی که وجود یال میان هر دو زوج رأس دلخواه دارای احتمال یکسان است، توزیع درجه رئوس از توزیع دوجمله‌ای پیروی می‌کند. در شکل ۲-۲ توزیع درجهٔ یک گراف با توزیع قاعدهٔ توانی نمایش‌داده‌شده است که بیشتر رئوس آن دارای درجهٔ پایین و تعداد محدودی از رئوس دارای درجهٔ بالا هستند. به شبکه‌هایی که در آن‌ها از قاعدهٔ توزیع توانی پیروی می‌شود شبکه‌های مقیاس آزاد [۴] گفته می‌شود.

[3] power law
[4] scale free



شکل ۳-۱ توزیع درجه‌ی یک گراف با قاعده‌ی توانی [۴]

۳-۳- اثر دنیای کوچک[5]

ریل‌بند و میل گرام [۴] در سال ۱۹۶۹ آزمایشی طراحی کردند تا میانگین طول مسیر شبکهٔ اجتماعی انسان در آمریکا را بررسی کنند. در این آزمایش افرادی انتخاب و از آن‌ها خواسته شده بود که نامه‌ای را به فردی در شهر موردنظر با آغاز از یک شهر ارسال کنند. در پایان ۶۴ نامه دریافت شد و میانگین مسیر پیموده شده بین افراد مختلف حدود ۶ بود. در واقع میانگین کمترین فاصلهٔ رئوس در شبکه بسیار پایین بود. این ویژگی که اثر دنیای کوچک نام دارد به این معنی است که بیشتر زوج رئوس حاضر در شبکه‌های واقعی فاصلهٔ بسیار کمی با یکدیگر دارند و این میزان با لگاریتم تعداد رئوس موجود در شبکه رابطه دارد.

۳-۴- ضریب خوشه‌بندی

در نظریهٔ گراف ضریب خوشه‌بندی اندازه‌ای است که بیانگر درجه‌ای است که گره‌ها در یک گراف گرایش به تشکیل خوشه در کنار یکدیگر دارند. شواهد بیانگر این هستند که در بیشتر شبکه‌های موجود در دنیای واقعی و در شبکه‌های اجتماعی خاص گره‌ها تمایل دارند تا به‌صورت قوی گروه‌هایی را تشکیل دهند که از طریق تراکم بالای یال‌ها مشخص و معلوم می‌شوند. در شبکه‌های واقعی این همسایگی از میانگین احتمال این که یک یال به‌صورت تصادفی بین دو گره اتصال برقرار نماید بیشتر است. دو نوع از این اندازه‌ها وجود دارند یکی اندازهٔ کلی و دیگری اندازهٔ محلی است. ضریب کلی خوشه‌بندی بر مبنای گره‌های سه تایی محاسبه می‌شود. یک حالت سه‌گانه از سه گره تشکیل شده است که از طریق دو (سه‌گانهٔ باز) و یا سه (سه‌گانهٔ بسته) یال به هم وصل می‌شوند. یک مثلث از سه‌گانهٔ بسته تشکیل شده است.

[5] small world

ضریب خوشه‌بندی کلی برابر است با نسبت سه‌گانه‌های بسته، یعنی ۳ ضرب‌در تعداد مثلث‌ها بخش بر تعداد کلی سه‌گانه‌ها که هم شامل سه‌گانه‌های بسته و هم سه‌گانه‌های باز می‌باشند. در واقع ضریب خوشه‌بندی نشان‌دهندهٔ میزان اتصال همسایه‌های یک گره و در نتیجه احتمال تشکیل خوشه می‌باشد. این ضریب عددی حقیقی بین صفر و یک است، زمانی که هیچ خوشه‌بندی وجود نداشته باشد این مقدار برابر با صفر و درصورتی‌که گراف، کامل باشد این مقدار برابر با یک می‌شود [۶].

۳-۵- همبندی

در گراف بدون جهت، دو راس را متصل می‌نامیم اگر حداقل یک مسیر بین آنها وجود داشته باشد. اگر دو راس با مسیری به طول یک یال به هم متصل شده باشند، آنگاه این دو راس همسایه نامیده می‌شوند. یک گراف هم‌بند نامیده می‌شود اگر هر دو راس دلخواه آن به یکدیگر متصل باشند. یک مؤلفهٔ هم‌بند یک زیرگراف هم‌بند بیشینه از گراف است. تعداد مؤلفه‌های هم‌بند نشان‌دهندهٔ قابلیت اتصال گراف است. هرچه تعداد مؤلفه‌های هم‌بند گراف کمتر باشد کل گراف دارای اتصالات قوی‌تری است.

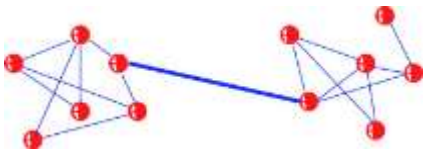
۳-۶- مرکزیت [6]

مرکزیت یکی از معیارهای مهم ساختاری شبکه می‌باشد. مرکزیت معمولاً برای گره‌های گراف تعریف می‌شود و بیانگری میزان اهمیت یک گره از جنبه‌های مختلف می‌باشد. سه جنبهٔ مهم از مرکزیت، درجه، نزدیکی[7] و بینابینی[8] می‌باشد.

- **درجه:** درجه تعداد رئوسی می‌باشد که یک راس به آن متصل است. اگر گراف آشنایان را در نظر بگیریم، مرکزیت درجه برابر با تعداد افرادی می‌باشد که یک فرد خاص می‌شناسد.
- **نزدیکی:** مجموع فاصلهٔ یک گره از گراف تا سایر رئوس شبکه را نزدیکی می‌نامند. زمانی که یک راس امتیاز نزدیکی کمی داشته باشد در واقع مرکزیت بالایی دارد و در جریان اکثر اطلاعات شبکه قرار می‌گیرد و ازآنجایی‌که با بیشتر رئوس فاصلهٔ کمی دارند، اطلاعات را سریع‌تر دریافت می‌کنند.

[6] centrality
[7] closeness
[8] betweenness

- **بینابینی:** بینابینی، تعداد کوتاه‌ترین مسیر بین هر جفت از رئوس گراف است که از یال موردنظر عبور می‌کند. این معیار اهمیت یک یال را در فرایندی مثل انتشار اطلاعات نشان می‌دهد. زیرا در اغلب اوقات اطلاعات از نزدیک‌ترین مسیر انتقال می‌یابند. به طور شهودی قابل‌درک است که بینابینی یال‌های بین دو اجتماع بیش از سایر یال‌هاست. زیرا کوتاه‌ترین مسیرهای بین دو اجتماع از این یال‌ها عبور خواهند کرد. البته این فرض زمانی درست است که تعداد گره‌ها به نسبت برابر بین دو جامعه تقسیم شده باشد. شکل ۳-۲ بیانگر این موضوع است.



شکل ۳-۲ زیاد بودن میزان بینابینی یال بین دو جامعه در یک شبکه [۷]

۴-اجتماع و تعاریف آن

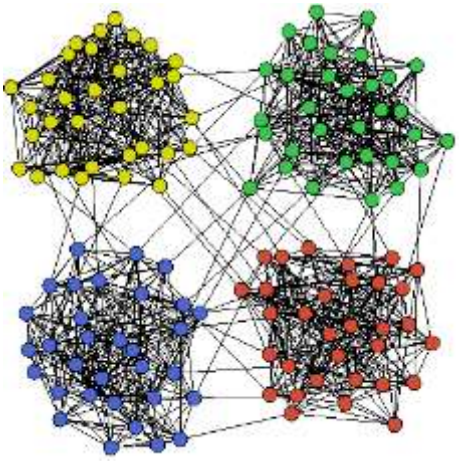
انواع مختلفی از شبکه‌ها در جهان وجود دارد، اما در این نوشته، تمرکز ما روی شبکه‌های اجتماعی است که در واقع بیانگر روابط بین افراد می‌باشد. برای مثال یکی از شبکه‌های اجتماعی، شبکهٔ همکاری [۹] و مقالات مشترک بین تمام محققان است. در این شبکه هر پژوهشگر با یک گره نمایش داده می‌شود و درصورتی‌که دو پژوهشگر حداقل یک مقالهٔ مشترک داشته باشند، بین آنها یک یال وجود دارد. برخی گره‌های شبکه‌های واقعی با هم ارتباط بیشتری دارند و جوامع را تشکیل می‌دهند به‌طوری‌که داخل جوامع یال‌های به‌مراتب بیشتری نسبت به یال‌های بین جوامع وجود دارد. هدف شاخهٔ کشف اجتماعات [۱۰] یافتن جوامع موجود در یک شبکه و احتمالاً ساختار سلسله‌مراتبی بین آنهاست.

کشف اجتماعات کاربردهای مهمی دارد. برای مثال یافتن جوامع در شبکهٔ همکاری پژوهشگران می‌تواند به یافتن گروهی از پژوهشگرانی که در حوزه‌های مشترک کار می‌کنند کمک کند. به‌عنوان مثالی دیگر، در شبکهٔ اجتماعی فیس‌بوک هر کاربر با افراد زیادی رابطهٔ دوستی دارد و دوستی با افراد مختلف باعث اشتراک فرد در جوامع مختلف مانند جامعهٔ خانواده، دانشگاه، محل کار و... می‌شود. اکنون اگر برای مثال فیس‌بوک بخواهد به یک کاربر فهرستی از افراد مشابه برای دوستی را پیشنهاد دهد، کشف اجتماعات می‌تواند به انتخاب افراد مشابه کمک کند.

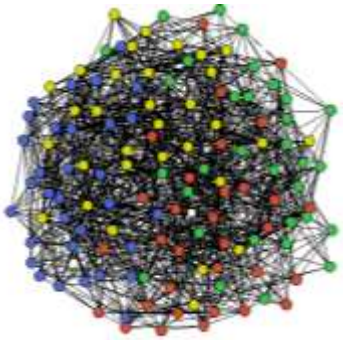
یکی از سخت‌ترین کارهای مربوط به کشف اجتماعات در همان گام نخست روی می‌دهد. اولین مسئله در کشف اجتماعات شبکه‌های پیچیده، یافتن تعریفی کمی از جامعه است. شکل ۲-۴ یک گراف را همراه با ساختار اجتماعات آن نشان می‌دهد. هرچند به لحاظ بصری یافتن اجتماعات این گراف راحت به نظر می‌رسد، اما به لحاظ علم ریاضی و نظریه گراف ارائهٔ تعریف و یافتن این اجتماعات پیچیده می‌نماید. در گراف شکل ۲-۴ توده‌های مجتمع راس‌ها، انجمن‌ها را تشکیل می‌دهند. شکل ۲-۵ یک گراف تصادفی با تعداد راس‌ها و یال‌های مساوی با گراف شکل ۲-۴ است که برای هر یال دو راس را به طور تصادفی انتخاب کرده و به هم وصل می‌کند. همان‌طور که مشخص است به لحاظ شهودی و تصویری گراف‌های مربوط به شبکهٔ اجتماعی از گروه‌های متراکم زیادی تشکیل شده است؛ اما گراف مربوط به گراف تصادفی تنها از یک توده تشکیل شده است. به‌صورت شهودی توزیع یال‌ها در گراف‌های واقعی همانند گراف فیس‌بوک نه به‌صورت سراسری که به‌صورت محلی است. بدین معنی که تعداد یال‌های توزیع شده میان گروهی از راس‌ها بسیار بیشتر از تعداد یال‌های توزیع شده بین این گروه از راس‌ها با بقیه راس‌های گراف است. این ویژگی که در گراف‌های مربوط به داده‌های واقعی می‌توان مشاهده کرد، جامعه خوانده می‌شود. در برخی منابع به جامعه انجمن یا خوشه نیز گفته می‌شود.

از کاربردهای کشف اجتماعات می‌توان به تبلیغات و بازاریابی اشاره کرد. ازآنجایی‌که افراد حاضر در اجتماعات تشکیل شده در یک شبکهٔ اجتماعی به‌احتمال زیاد علایق مشترکی دارند، می‌توان با یافتن علایق آن‌ها از این اطلاعات به‌منظور تبلیغ محصولات خاص استفاده کرد. کاربردهای زیاد دیگری نیز می‌توان برای اجتماعات نام برد. کاربرد فراوان در زمینه‌هایی همچون زیست‌شناسی، مهندسی رایانه، اقتصاد این شاخه از علم شبکه‌های اجتماعی و نظریه گراف را به زمینه‌ای محبوب برای پژوهشگران جهت تحقیق تبدیل کرده است.

[9] collaboration network
[10] community detection



شکل ۴-۱ یک گراف با خوشه‌های معین [۸]



شکل ۴- ۲ یک گراف با ساختار تصادفی [۸]

تعاریف متعددی برای جامعه ارائه شده است. هیچ تعریفی به‌صورت کلی پذیرفته شده نیست. در حقیقت، این تعاریف معمولاً به سیستم خاص در دست بررسی و یا کاربردی که در ذهن برای آن خوشه‌بندی در نظر گرفته شده بستگی دارد. اساس بیشتر تعاریف موجود برای جوامع این نکته است که تعداد یال‌ها در داخل جامعه نسبت به یال‌هایی که رئوس آن جامعه را به بقیهٔ گراف متصل می‌کنند بیشتر است.

۵- روش های کشف اجتماعات

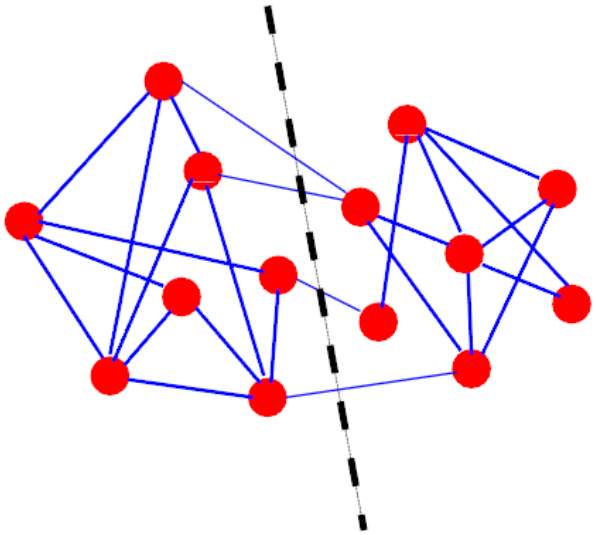
در سال‌های گذشته روش‌های متعددی برای کشف اجتماعات در شبکهٔ های اجتماعی ارائه شده است. در این بخش دسته‌بندی جامعی از این روش‌ها ارائه خواهد شد.

۵-۱- الگوریتم‌های افراز کننده

مسئلهٔ افرازکردن گراف به چند بخش برابر است با تقسیم گره‌ها به گروه‌های با اندازهٔ از پیش تعریف شده به‌طوری‌که تعداد یال‌های بین گروه‌ها حداقل باشد. تعداد یال‌های بین گروه‌ها با عنوان اندازهٔ برشی[11] شناخته می‌شود. شکل ۱-۶ حل این مساله برای یک گراف با ۱۴ راس و ۲ گروه با اندازه‌های مساوی را نشان می‌دهد.

تعیین تعداد خوشه‌های موردنظر برای تقسیم‌بندی ضروری است. زیرا اگر تعداد خوشه‌ها مهم نباشد و هدف کمینه‌کردن مقدار اندازهٔ برشی باشد جواب بدیهی است، تمام یال‌ها باید در یک خوشه قرار بگیرند. مشخص‌بودن اندازهٔ خوشه‌ها نیز ضروری است؛ زیرا در غیر این صورت راه‌حل این است که راس با کمترین درجه را از گراف جدا کنیم که نتیجه مطلوب نخواهد بود. این مشکل با بهینه‌سازی یک معیار دیگر برای خوشه‌بندی که اندازهٔ خوشه را کنترل کند قابل‌حل است.

تعیین تعداد خوشه‌های موردنظر برای تقسیم‌بندی ضروری است.



شکل ۵-۱ تقسیم‌بندی گراف به دو گروه هم‌اندازه

زیرا اگر تعداد خوشه‌ها مهم نباشد و هدف کمینه‌کردن مقدار اندازهٔ برشی باشد جواب بدیهی است، تمام یال‌ها باید در یک خوشه قرار بگیرند. مشخص‌بودن اندازهٔ خوشه‌ها نیز ضروری است؛ زیرا در غیر این صورت راه‌حل این است که راس با کمترین درجه را از گراف جدا کنیم که نتیجه مطلوب نخواهد بود. این مشکل با بهینه‌سازی یک معیار دیگر برای خوشه‌بندی که اندازهٔ خوشه را کنترل کند قابل‌حل است.

مسئلهٔ افرازکردن گراف یک مسئلهٔ NP-Hard است. اگرچه، روش‌های زیادی برای حل این مسئله وجود دارد هرچند ممکن است بهینه نباشند. بسیاری از روش‌های موجود گراف را به دو

[11] cut size

بخش تقسیم می‌کنند و برای ایجاد چند خوشه باید به‌صورت بازگشتی الگوریتم را تکرار کرد.
در این روش‌ها هر گره مختصات معینی دارد و شباهت گره‌ها بر اساس فاصله بین گره‌ها تعریف می‌شود. هدف این است که گره‌ها طوری به چند خوشه تقسیم شوند که یک معیار هدف، بر اساس فاصلهٔ بین گره‌ها کمینه شود. برخی از معیارهایی که برای این کار استفاده می‌شوند عبارت‌اند از:

- **Minimum k-clustering**: تابع هزینه در اینجا قطر خوشه است که به معنی بیشترین فاصله بین دو گره در یک خوشه است. گره‌ها به شکلی خوشه‌بندی می‌شوند که بزرگ‌ترین قطر خوشه‌ها کمترین مقدار ممکن باشد. ایدهٔ این روش این است که خوشه‌ها تا حد ممکن فشرده باشند.

- **k-clustering sum**: این روش مشابه روش minimum k-clustering است با این تفاوت که به‌جای قطر از میانگین فاصله‌های بین گره‌ها استفاده می‌شود.

- **k-center**: برای هر خوشه i یک نقطهٔ مرجع x_i با نام مرکز [۱۲] در نظر گرفته می‌شود و d_i ، بیشترین مقدار فاصله‌ی هر نقطه‌ی خوشه تا مرکز محاسبه می‌شود. خوشه‌ها و مرکزها تا زمانی که به کمترین مقدار d_i برسیم تغییر می‌کنند.

- **k-median**: مشابه k-center است با این تفاوت که به‌جای بیشترین فاصله از مرکز، از فاصلهٔ میانگین تا مرکز استفاده می‌شود.

یکی از محبوب‌ترین روش‌های خوشه‌بندی روش k-means [۹]است. در اینجا تابع هزینه، مجموع فاصله‌های گره‌های داخل هر خوشه است:

$$\sum_{i=1}^{k}\sum_{x_j \in S_i}\left\|x_j - c_i\right\|^2 \qquad (۱-۵)$$

که S_i بیانگر گره‌های داخل خوشه‌ی i ام و c_i مرکز آن خوشه است. مساله‌ی k-means می‌تواند توسط الگوریتم Lloyd حل شود. در این الگوریتم کار با چند مرکز که در فاصله‌ی مناسبی از هم قرار دارند شروع می‌شود. در تکرار اول هر گره به نزدیکترین مرکز تخصیص داده می‌شود. سپس برای هر خوشه مرکز جدید تخمین زده می‌شود و مجموعه مراکز جدید بدست می‌آید و به همین ترتیب مراحل دوباره تکرار می‌شود. بعد از چند تکرار، مراکز تقریباً ثابت می‌مانند و خوشه‌ها بدست می‌آیند. جواب بدست آمده بهینه نیست و خوشه‌های حاصل به مراکز انتخاب شده‌ی اولیه بستگی دارند اما مزیت این روش همگرایی سریع آن است. برای دقیق‌تر کردن جواب می‌توان الگوریتم را چند بار و با مراکز مختلف تکرار کرد و بهترین جواب را انتخاب کرد.
روش‌های دیگری نیز برای استفاده از k-means ارائه شده‌اند که مشکل همهٔ این روش‌ها این است که باید تعداد خوشه‌ها از ابتدا مشخص باشد.

۵-۲- الگوریتم‌های سلسله مراتبی

معمولاً اطلاعات کمی در مورد ساختار جوامع گراف در دسترس است و تعداد جوامعی که گراف به آن باید تقسیم شود مشخص نیست. در چنین شرایطی روش‌های خوشه‌بندی کارایی چندانی ندارند و نیاز است که دلیل مناسبی برای تعداد خوشه‌های به‌دست‌آمده ارائه کرد. از طرف دیگر گراف ممکن است ساختار سلسله‌مراتبی داشته باشد، برای مثال ممکن است چند سطح مختلف از گره‌ها وجود داشته باشد و خوشه‌های کوچک‌تر درون خوشه‌های بزرگ‌تر قرار داشته باشند و یک رابطهٔ سلسله‌مراتبی بین خوشه‌ها برقرار باشد. در چنین مواردی می‌توان از روش‌های خوشه‌بندی سلسله‌مراتبی استفاده کرد که سطوح مختلف گراف را مشخص می‌کنند.
گام اول خوشه‌بندی سلسله‌مراتبی تعریف یک معیار شباهت[۱۳] است. بعد از انتخاب معیار شباهت، میزان شباهت بین هر دو گره صرف‌نظر از متصل بودن یا نبودن آن‌ها محاسبه می‌شود. بعدازاین عمل یک ماتریس شامل مقادیر شباهت بین گره‌ها به دست می‌آید.
الگوریتم‌های خوشه‌بندی سلسله‌مراتبی به دودسته تقسیم می‌شوند:

۱. **الگوریتم‌های تراکمی** [۱۴]: خوشه‌های با شباهت بالا به طور تکراری ترکیب می‌شوند.
۲. **الگوریتم‌های تقسیم‌کننده** [۱۵]: خوشه‌ها به طور تکراری با حذف یال‌های بین گره‌های با شباهت کم به خوشه‌های کوچک‌تر تقسیم می‌شوند.

[12] centroid

[13] similarity
[14] agglomerative

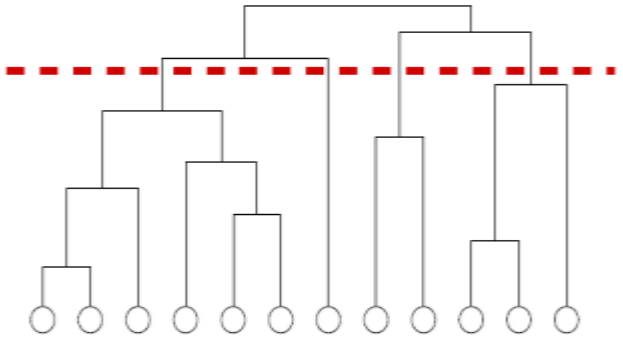
این دو کلاس به دو فرایند متضاد اشاره دارند، الگوریتم‌های تراکمی پایین‌به‌بالا هستند؛ چون در ابتدا هر گره را به‌عنوان یک خوشهٔ مجزا در نظر می‌گیرند و در آخر تمام گره‌های گراف به‌عنوان یک خوشه در نظر گرفته می‌شوند. الگوریتم‌های تقسیم‌کننده رویکرد بالابه‌پایین دارند. این الگوریتم‌ها به‌ندرت استفاده می‌شوند؛ بنابراین از بحث بیشتر روی آن‌ها خودداری می‌کنیم.

از آنجایی که خوشه‌ها بر اساس شباهت مشترک‌شان ترکیب می‌شوند ضروری است که یک معیار برای تخمین میزان شباهت خوشه‌ها از روی ماتریس شباهت گره‌ها تعریف کنیم. در خوشه‌بندی تک پیوندی [16] شباهت بین دو خوشه کمترین مقدار x_{ij} است که گره‌ی i در یک خوشه و گره‌ی j در خوشه‌ی دیگر قرار دارد. در طرف مقابل، بیشترین مقدار x_{ij} برای خوشه‌بندی تمام پیوندی[17] استفاده می‌شود. در خوشه‌بندی میانگین پیوندی [18] از میانگین x_{ij} استفاده می‌شود.

مراحل کار می‌تواند توسط یک دندروگرام مانند شکل ۲-۷ نمایش داده شود. توقف خوشه‌بندی برای انتخاب یک خوشه‌بندی از گراف، بر اساس رسیدن به شرایط خاصی انجام می‌شود. برای مثال به تعداد خوشه‌های موردنظر رسیده‌ایم یا یک معیار کیفیت را بهینه کرده‌ایم.

خوشه‌بندی سلسله‌مراتبی این مزیت را دارد که نیازی به اطلاعات خاصی در مورد گراف و تعداد خوشه‌ها نداریم. نتیجهٔ این خوشه‌بندی به نوع معیار شباهت استفاده شده بستگی دارد. این روش همچنین یک ساختار سلسله‌مراتبی می‌سازد که در بسیاری از مواقع غیرواقعی است، زیرا ممکن است گراف یک ساختار سلسله‌مراتبی نداشته باشد. یکی دیگر از مشکلات این روش این است که یک گرهٔ با تنها یک همسایه، به‌عنوان یک خوشهٔ مجزا در نظر گرفته می‌شود که غیرمعقول است.

زمان اجرای الگوریتم‌های سلسله مراتبی برای حالت تک پیوندی $O(n^2)$ و برای حالات میانگین و کامل، $O(n^2 \log n)$ است.



شکل ۵-۲ نمایش سلسله‌مراتبی خوشه‌ها بوسیله‌ی دندروگرام

۵-۳- روش‌های طیفی

خوشه‌بندی طیفی شامل تمام روش‌هایی است که گراف را با استفاده از بردارهای ویژهٔ ماتریس‌های مربوط به گراف یا ماتریس‌های مشتق شده خوشه‌بندی می‌کند [۱۰].در خوشه‌بندی طیفی مجموعهٔ ابتدایی از اشیا به مجموعه‌ای از نقاطی در فضا تبدیل می‌شوند که مختصات آنها همان عناصر بردار ویژه هستند. سپس مجموعه نقاط به‌دست‌آمده، با روش‌های عادی خوشه‌بندی مانند k-means خوشه‌بندی می‌شوند.

ماتریس لاپلاس[19] یکی از ماتریس‌های متداول برای خوشه‌بندی طیفی است. یک گراف را می‌توان با استفاده از مقادیر ویژه و بردارهای ویژهٔ ماتریس لاپلاس آن گروه‌بندی کرد. کوچک‌ترین مقدار ویژهٔ یک ماتریس لاپلاس، صفر است و بردار ویژهٔ متناظر با آن [۱،۱،...،۱] می‌باشد.

یک راه ساده برای یافتن دومین کوچک‌ترین مقدار ویژه برای هر ماتریس متقارن محاسبهٔ کمترین مقدار $X^T LX$ است که X یک بردار ستونی با n عضو است ([x1, x2, …, xn]) و شرایط زیر برقرار است:

۱. طول بردار X برابر با یک است.

۲. X عمود بر بردار ویژه متناظر با کوچک‌ترین مقدار ویژه است.

کمترین مقدار X برابر با دومین بردار ویژه است. وقتی L ماتریس لاپلاس یک گراف با n گره باشد، ما اطلاعات بیشتری داریم. بردار ویژه متناظر با کوچک‌ترین مقدار ویژه برابر با یک است؛ بنابراین اگر X عمود بر یک باشد باید داشته باشیم:

$$X^T 1 = \sum_{i=1}^{n} x_i = 0 \qquad (۲-۵)$$

علاوه بر این، برای ماتریس لاپلاس معادلهٔ زیر را داریم:

[15] divisive
[16] Single linkage clustering
[17] complete linkage clustering
[18] average link clustering
[19] laplacian

$$L = D - A \qquad (۳-۵)$$

با استفاده از روابط بالا به این نتیجه می‌رسیم که $X^T LX$ برابر با مجموع $(x_i - x_j)^2$ به ازای تمام یال‌های(i,j) ، j از گراف است. همچنین این شرط که X باید بر ۱ عمود باشد نیز باید برقرار باشد، بنابراین مجموع xi ها نمی‌تواند صفر باشد. همچنین این شرط را که $\sum_{i=1}^{n} X_i^2 = 1$ نیز برقرا است، بنابراین تمام xi ها نمی‌توانند صفر باشند. بنابراین X باید شامل هم مقادیر منفی و هم مقادیر مثبت باشد .

یک خوشه‌بندی از گراف را می‌توانیم با جداکردن مجموعه گره‌های متناظر با iهای منفی و مثبت از بردار X به دست بیاوریم. این انتخاب برابر بودن اندازهٔ دو جامعه را تضمین نمی‌کند؛ اما اندازه‌ها نزدیک به هم خواهند بود. همچنین مجموعهٔ برشی شامل یال‌های کمی خواهد بود؛ زیرا مقدار $(x_i - x_j)^2$ زمانی کمینه است که علامت xi و xj یکی باشد، یعنی هر دو گره در یک مجموعه قرار داشته باشند. در اینجا مقدار صفر به‌عنوان مقدار آستانه [۲۰] برای جداسازی دو گروه در نظر گرفته شده است، اما این مقدار می‌تواند مقداری غیر از صفر نیز باشد.

۵-۳-۱- الگوریتم‌های تقسیم‌کننده[۲۱]

یک راه ساده برای کشف اجتماعات در گراف، یافتن و حذف‌کردن یال‌هایی است که جوامع مختلف را به هم وصل می‌کند. با این کار جوامع از یکدیگر مجزا و مشخص خواهند شد. این ایده، فلسفهٔ روش‌های تقسیم‌کننده است. نکتهٔ اصلی در اینجا یافتن ویژگی یال‌های بین جوامع است. یکی از معروف‌ترین روش‌های این دسته، روش ارائه شده توسط نیومن و گیروان است که در ادامه به آن می‌پردازیم.

۵-۴- روش Newman-Girvan

یکی از محبوب‌ترین الگوریتم‌های کشف اجتماعات، روش ارائه شده توسط نیومن و گیروان [۱۱]است. این روش ازاین‌جهت نیز مهم است که آغازگر یک دورهٔ جدید در زمینهٔ کشف اجتماعات در شبکه‌های پیچیده بود. در این روش یال‌ها بر اساس مقدار مرکزیت [۲۲] آن برای حذف انتخاب می‌شود. مراحل این الگوریتم عبارت‌اند از:

۱. محاسبهٔ مرکزیت هر یال
۲. حذف یال با بیشترین مقدار مرکزیت
۳. محاسبه‌ی مجدد مرکزیت یال‌ها
۴. تکرار از مرحلهٔ ۲

نیومن و گیروان روی مفهوم بینابینی تمرکز کردند. بینابینی متغیری برای بیان میزان شرکت یک یال در یک فرایند است. آنها سه فرایند را در نظر گرفتند:

- بینابینی یال‌ها در کوتاه‌ترین مسیرها
- بینابینی یال‌ها در راه‌رفتن تصادفی [۲۳]
- بینابینی یال‌ها در جریان [۲۴]

در ادامه هر کدام از این سه مفهوم را مورد بررسی قرار خواهیم داد. بینابینی یال‌ها برابر با تعداد کوتاهترین مسیرهای بین تمام جفت گره‌ها است که از یال مورد نظر عبور می‌کنند. محاسبه‌ی بینابینی تمام یال‌های یک گراف با استفاده از روش‌ها مبتنی بر جستجوی اول سطح در زمان $O(nm)$ یا $O(n^2)$ در صورت پراکنده بودن ماتریس، قابل محاسبه است. از دیدگاه انتشار اطلاعات این احتمال وجود دارد که اطلاعات به جای کوتاهترین مسیر، از مسیرهای تصادفی حرکت کند. در این حالت بینابینی یک یال برابر با تعداد دفعاتی که یک راه رونده‌ی تصادفی از آن عبور کرده است. یک راه رونده‌ی تصادفی با احتمال مساوی به یال‌های متصل به یک گره حرکت می‌کند. یک جفت گره‌ی تصادفی s و t انتخاب می‌شوند و راه رونده از s شروع می‌کند و وقتی به t رسید متوقف می‌شود. در آخر احتمال عبور راه رونده از هر یال محاسبه می‌شود و این مراحل برای تمام جفت گره‌های ممکن تکرار و میانگین احتمال عبور هر یال محاسبه می‌شود. محاسبه‌ی بینابینی حرکت کننده‌ی تصادفی نیاز به معکوس کردن ماتریس مجاورت برای هر مرحله دارد که زمان آن $O(n^3)$ خواهد بود. برای هر جفت گره نیز میانگین تعداد عبور راه رونده را محاسبه کنیم که زمان آن $O(mn^2)$ خواهد بود. بنابراین مجموع زمان این الگوریتم $O(n^3)$ خواهد بود.

در بینابینی جریان گراف به‌عنوان یک شبکه از مقاومت‌ها در نظر گرفته می‌شود و یال‌ها دارای مقاومت واحد هستند. اگر یک اختلاف ولتاژ بین هر دو گره ایجاد شود در هر یال یک جریان به

[20] threshold
[21] divisive

[22] centrality
[23] random walk edge betweenness
[24] current flow edge betweenness

وجود می‌آید که مقدار آن با قانون کیرشهف [۲۵] قابل‌محاسبه است. این مراحل برای همهٔ جفت گره‌ها تکرار می‌شود. بینابینی جریان برای هر یال برابر با میانگین جریان عبوری از آن یال است. می‌توان نشان داد که این معیار برابر با همان معیار راه روندهٔ تصادفی است، زیرا اختلاف ولتاژ و راه رونده به یک‌شکل روی یال‌ها حرکت می‌کنند؛ بنابراین پیچیدگی زمانی این روش نیز مانند راه‌رفتن تصادفی $O(n^3)$ است.

همان‌طور که مشاهده شد محاسبهٔ بینابینی یال‌ها پیچیدگی محاسباتی کمتری نسبت به دو روش دیگر دارد. همچنین در کاربردهای عملی الگوریتم نیومن - گیروان [۲۶] با استفاده از بینابینی یال‌ها نتایج بهتری نسبت به سایر معیارها می‌دهد. در گراف‌های دارای جوامع که گراف به‌سرعت به چند بخش تقسیم می‌شود، مرحلهٔ محاسبهٔ مجدد مقادیر مرکزیت باید بین مؤلفه‌های هم‌بندی از گراف انجام شود که توسط یالی که در این مرحله برای حذف انتخاب شده است انجام شود و مقدار مرکزیت سایر یال‌ها تغییر نمی‌کند. الگوریتم نیومن - گیروان بسیار کند است و تنها برای گراف پراکنده با کمتر از ۱۰۰۰۰ گره قابل اجراست.

۶- کشف اجتماعات در شبکه‌های اجتماعی یک لایه پویا با رویکرد بیشینه‌سازی اهمیت

۶-۱- معیار اهمیت

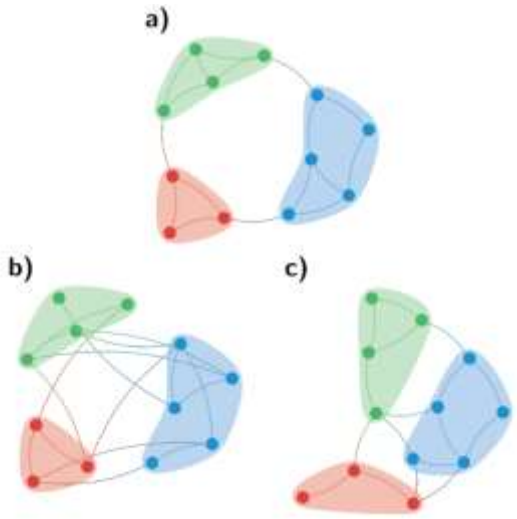
اگرچه معیار پیمانگی تعداد یال‌های یک اجتماع را با یک گراف تصادفی مقایسه می‌کند، اما این معیار هیچ اطلاعاتی در مورد اهمیت یک خوشه‌بندی در اختیار قرار نمی‌دهد، زیرا یک گراف تصادفی و یک گراف پراکنده[۲۷] بدون داشتن ساختار اجتماع نیز می‌توانند پیمانگی بالایی داشته باشند. روش‌هایی برای اندازه‌گیری میزان اهمیت یک خوشه‌بندی ارائه شده‌اند. یکی از این روش‌ها، تحیّر[۲۸] است که روی احتمال یافتن E یال در یک گراف تصادفی تمرکز دارد [۱۲].یکی دیگر از معیارهای محلی نیز وجود دارد که درجات را ثابت نگه می‌دارد و به دنبال احتمال اتصال یال‌ها به یک اجتماع می‌گردد[۱۳]که در روش OSLOM [۱۱]استفاده شده است. روش دیگری نیز وجود دارد که روی راست‌نمایی[۲۹] تولید یک گراف با خوشه‌بندی و توزیع درجهٔ مشخص تمرکز دارد [۱۴]که به‌عنوان روش SBM [۳۰] شناخته می‌شود.

اما این روش‌ها معمولاً به اهمیت واقعی یک خوشه‌بندی نمی‌پردازند و روشی اشتباه را در پیش می‌گیرند. ما دنبال یافتن احتمال یک خوشه‌بندی ثابت که شامل E یال داخلی می‌باشد نیستیم، بلکه در پی یافتن خوشه‌ای با حداقل E یال داخلی در یک گراف تصادفی هستیم.

در واقع روش‌های قبلی خوشه‌ها را ثابت در نظر می‌گرفتند و یال‌ها را به‌صورت تصادفی در این خوشه‌ها توزیع می‌کردند، درحالی‌که معیار اهمیت به دنبال یافتن خوشه‌ها در یک گراف تصادفی می‌باشد. به‌عبارت‌دیگر این روش‌ها این موضوع که یک جای‌گشت از گره‌ها نیز می‌تواند یک خوشه را تشکیل دهد نادیده گرفته‌اند درحالی‌که در معیار اهمیت، به این موضوع پرداخته می‌شود [۱۵].

تفاوت دو روش در شکل ۶-۱ بیان شده است.

خوشه‌بندی قسمت a را در نظر بگیرید. هدف تخمین‌زدن احتمال وجود خوشه در یک گراف تصادفی با درنظرگرفتن اهمیت خوشه است. در قسمت b و c نیز همان گراف نمایش‌داده‌شده است، با این تفاوت که در قسمت b گره‌ها در همان



شکل ۶-۱ تفاوت خوشه‌بندی با تمرکز بر بیشینه‌سازی اهمیت و سایر روش‌های مشابه [۱۵]

خوشه‌های قسمت a قرار گرفته‌اند؛ اما در قسمت c خوشه‌ها یال‌های داخلی بیشتری دارند. در روش‌های موجود، خوشه‌ها ثابت در نظر گرفته می‌شوند و احتمال قرارگیری یال‌ها در هر

[25] Kirchof's law
[26] Girvan-Newman
[27] Sparse
[28] Surprise
[29] Likelihood
[30] Stochastic Block Model

خوشه محاسبه می‌شود، این مطلب در قسمت b نمایش‌داده‌شده است. این روش‌ها احتمال وجود خوشه‌های با یال‌های داخلی بیشتر را در نظر نمی‌گیرند، درحالی‌که معیار اهمیت روی احتمال یافتن یک خوشه‌بندی متراکم در یک گراف تصادفی تمرکز دارد. بااین‌وجود، روش‌های موجود ممکن است نتایج خوبی داشته باشند، برای مثال محاسبهٔ احتمال وجود E یال داخلی به نظر می‌رسد که به نتایج مناسبی منجر شود [۱۵].

اگر احتمال یافتن یال‌های داخلی زیادی در یک خوشه‌بندی زیاد باشد، مطمئناً یافتن یک خوشه‌بندی با یال‌های زیاد نیز محتمل‌تر می‌باشد. از طرف دیگر، اگر احتمال یافتن یال‌های داخلی زیاد در یک خوشه‌بندی کم باشد، مطمئناً احتمال یافتن یک خوشه‌بندی با یال‌های زیاد نیز کم است. بین این دو مطلب فضای مبهمی وجود دارد که برای فهم دقیق آن نیاز به بررسی و تحلیل بیشتر می‌باشد.

در معیار اهمیت، روی احتمال یافتن خوشه‌هایی با تراکم بالا در یک گراف تصادفی تمرکز می‌شود. اندازه‌ی خوشه‌ها متناسب با اندازه‌ی گراف با n گره یعنی sn و چگالی q در نظر گرفته می‌شود که در آن $1>s>0$ می‌باشد.

احتمال یافتن یک زیرگراف با اندازه‌ی n_c و چگالی q در یک گراف با اندازه‌ی n و چگالی p برابر است با:

$$\Pr\left(S\left(n_c,q\right)\right)\subseteq G\left(n,p\right)=e^{\Theta\left(-\binom{n_c}{2}D(q\|p)\right)} \qquad (۱-۶)$$

که در این فرمول $D\left(q\|p\right)$ برابر با واگرایی کولبک-لیبلر[31] می‌باشد و از طریق فرمول (۵-۲) قابل‌محاسبه است.

$$D\left(q\|p\right)=q\log\frac{q}{p}+\left(1-q\right)\log\frac{1-q}{1-p} \qquad (۲-۶)$$

برای مقادیر $p\neq q$ این احتمال یک توزیع گاوسی[32] با نرخ متناسب با فاصله‌ی بین q و p می‌باشد که با فرمول واگرایی کولبک-لیبلر بیان شده است. طبق این فرمول هرچه اندازه‌ی زیرگراف بیشتر باشد، احتمال یافتن زیرگراف با چگالی متفاوت با p کمتر می‌شود.

با ترکیب این احتمال‌ها، فرمول (۵-۳) برای محاسبه‌ی احتمال یافتن یک خوشه‌بندی در یک گراف تصادفی به دست می‌آید.

$$\Pr\left(\sigma\right)=\prod_c\exp\left(-\binom{n_c}{2}D\left(p_c\|p\right)\right) \qquad (۳-۶)$$

که در آن p_c چگالی اجتماع c می‌باشد؛ بنابراین معیار اهمیت به‌صورت زیر تعریف می‌شود:

$$S\left(\sigma\right)=-\log\Pr\left(\sigma\right)=\sum_c\binom{n_c}{2}D\left(p_c\|p\right) \qquad (۴-۶)$$

توجه به این نکته ضروری است که دو خوشه‌بندی بدیهی یعنی خوشه‌بندی که تمام گره‌ها در یک دسته قرار بگیرند و خوشه‌بندی که هر گره در یک اجتماع مجزا قرار بگیرد، دارای اهمیت با مقدار صفر می‌باشد؛ بنابراین معیار اهمیت همواره غیرصفر خواهد بود.

فرمول واگرایی کولبک-لیبلر با استفاده از مفهوم تفاوت آنتروپی[33] قابل‌تفسیر است:

$$D\left(q\|p\right)=H\left(q,p\right)-H\left(q\right) \qquad (۵-۶)$$

که در آن $H\left(q\right)$ برابر است با:

$$H\left(q\right)=-q\log q-\left(1-q\right)\log\left(1-q\right) \qquad (۶-۶)$$

$$H\left(q,p\right)=-q\log p-\left(1-q\right)\log\left(1-p\right) \qquad (۷-۶)$$

۶-۲- تعمیم معیار اهمیت به شبکه‌های پویا

فرمول ارائه شده برای معیار اهمیت، در شبکه‌های ایستا کارایی دارد. برای استفاده از این معیار در شبکه‌های پویا در زمان نیاز است تغییراتی در آن اعمال شود. در این بخش به معرفی معیار اهمیت پویا، با استفاده از تعمیم معیار اهمیت ایستا می‌پردازیم. همین‌طور با استفاده از روش ابتکاری لو وین[34] روشی جهت کشف اجتماعات در شبکه‌های اجتماعی یک‌لایهٔ پویا ارائه خواهد شد.

۶-۲-۱- معیار اهمیت پویا

برای کشف اجتماعات در شبکه‌های پویا، نیاز است در معیار اهمیت تغییراتی ایجاد شود تا اضافه و یا کم‌شدن یال‌ها و گره‌ها نیز در نظر گرفته شود. با اضافه‌شدن یال به شبکه، ۴ عمل زیر قابل انجام است:

- ساختار اجتماعات را دست‌نخورده و بدون تغییر نگه داریم.
- دو اجتماع را به یک اجتماع ادغام کنیم.
- یک گره را به یک اجتماع موجود تخصیص دهیم.

[31] Kullback-Leibler divergence
[32] Gaussian
[33] Entropy
[34] Louvain Heuristic

- یک اجتماع جدید ایجاد کنیم و دو گره را به آن تخصیص دهیم.

یال‌های اضافه شده به شبکه می‌توانند یکی از انواع زیر باشند:

- یال‌های داخل اجتماع: هر دو گرهٔ این یال از قبل در شبکه موجود هستند و هر دو در یک اجتماع مشترک قرار دارند.
- یال‌های بین اجتماع: دو گرهٔ این نوع یال از قبل در شبکه وجود دارند، اما در دو اجتماع مختلف قرار دارند.
- یال‌های نیمه جدید: یکی از گره‌های این نوع یال، به‌تازگی به شبکه اضافه شده است.
- یال‌های جدید: هر دو گرهٔ این یال به‌تازگی به شبکه اضافه شده‌اند.

در ادامه به بررسی هر نوع یال و اعمال قابل‌انجام در مورد هر حالت می‌پردازیم.

۶-۲-۱-۱- یال‌های داخل اجتماع

در این حالت، دو گرهٔ یال اضافه شده به لایهٔ $k^{'}$، از قبل در شبکه موجود بوده و هر دو گره در یک اجتماع یکسان قرار دارند. در نتیجه تراکم یال‌های داخل اجتماع افزایش خواهد یافت و باعث افزایش معیار اهمیت شبکه خواهد شد؛ بنابراین در این حالت بهترین انتخاب این است که ساختار اجتماعات دست‌نخورده باقی بماند.

۶-۲-۱-۲- یال‌های بین دو اجتماع

در این حالت، دو گره‌ی یال اضافه شده به لایه‌ی $k^{'}$، از قبل در شبکه موجود بوده و دو گره متعلق به اجتماعات مختلفی هستند. این یال را به صورت (i, j, w_{ij}) در نظر می‌گیریم که گره‌های i و j به ترتیب متعلق به اجتماعات c_i و c_j می‌باشند و وزن این یال برابر با w_{ij} می‌باشد.

در این حالت دو عمل قابل‌انجام است:

۱. ساختار اجتماعات بدون تغییر باقی بماند.

۲. دو اجتماع با یکدیگر ادغام شوند.

در صورت انتخاب عمل ۱ اهمیت جدید برابر با مقدار زیر خواهد بود:

(۸-۶)

$$S_1(\sigma)=\sum_{c}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)=\sum_{c\neq c_i,c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_i}}{2}D\left(p_{c_i} \parallel p^{'}\right)+\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n}{2}}$$

در صورت انتخاب عمل ۲ مقدار اهمیت جدید برابر با مقدار زیر خواهد بود:

(۹-۶)

$$S_2(\sigma)=\sum_{c\neq c_i,c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_i}+n_{c_j}}{2}D\left(p_{c_{i,j}} \parallel p^{'}\right)$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n}{2}},\ p_{c_{i,j}}=\frac{m_{c_{i,j}}+w_{ij}}{\binom{n_i+n_j}{2}}$$

تفاوت میزان اهمیت دو عمل بالا برابر است با:

(۱۰-۶)

$$\Delta S(\sigma)=\binom{n_{c_i}+n_{c_j}}{2}D\left(p_{c_{i,j}} \parallel p^{'}\right)-\binom{n_{c_i}}{2}D\left(p_{c_i} \parallel p^{'}\right)-\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)$$

اگر $\Delta S(\sigma)>0$ باشد عمل ۲ انتخاب می‌شود، در غیر این صورت عمل ۱ انتخاب خواهد شد.

۶-۲-۱-۳- یال‌های نیمه‌جدید

در این حالت، یکی از گره‌های متصل به یال جدید، از قبل در شبکه موجود بوده و یکی از گره‌ها جدید است. این یال را به صورت (i, j, w_{ij}) در نظر می‌گیریم که گره‌ی i گره‌ی جدید، گره‌ی j به ترتیب متعلق به اجتماع c_j و وزن این یال برابر با w_{ij} می‌باشد.

در این حالت دو عمل قابل‌انجام است:

۱. گره‌ی i به اجتماع مربوط به گره‌ی j اضافه شود.

۲. یک اجتماع جدید برای گره‌ی i ایجاد شود.

در صورت انتخاب عمل ۳ مقدار اهمیت جدید برابر با مقدار زیر خواهد بود:

(۱۱-۶)

$$S_3(\sigma)=\sum_{c\neq c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_j}+1}{2}D\left(p_{c_j}^{'} \parallel p^{'}\right)$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n+1}{2}},\ p_{c_j}^{'}=\frac{m_{c_j}+w_{ij}}{\binom{n_j+1}{2}}$$

در صورت انتخاب عمل ۴ مقدار اهمیت جدید برابر با مقدار زیر خواهد بود:

(۱۲-۶)

$$S_4(\sigma)=\sum_{c\neq c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)+0$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n+1}{2}},\ p_{c_j}=\frac{m_{c_j}+w_{ij}}{\binom{n_j}{2}}$$

تفاوت مقدار اهمیت دو عمل ذکر شده برابر با مقدار زیر است:

(۵-۱۳)

$$\Delta S(\sigma)=\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)-\binom{n_{c_j}+1}{2}D\left(p_{c_{j1}} \parallel p^{'}\right)$$

اگر $\Delta S(\sigma)>0$ باشد عمل ۴ انتخاب می‌شود، در غیر این صورت عمل ۳ انتخاب خواهد شد.

۶-۲-۱-۴- یال‌های جدید

در این حالت، هر دو گره‌ی یال جدید از قبل در شبکه موجود نبوده و همراه با یال به شبکه اضافه می‌شوند. این یال را به صورت (i,j,w_{ij}) در نظر می‌گیریم که گره‌های i و j گره‌های جدید می‌باشند و وزن این یال برابر با w_{ij} می‌باشد.

در این حالت دو عمل قابل‌انجام است:

۱. دو گره به یکی از اجتماعات موجود تخصیص داده شوند.

۲. یک اجتماع جدید ایجاد شود و دو گره به آن تخصیص داده شوند.

در صورت انتخاب عمل ۳ مقدار جدید اهمیت برابر با مقدار زیر خواهد بود:

(۱۴-۶)

$$S_3(\sigma)=\sum_{c\neq c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_j}+2}{2}D\left(p_{c_j}^{'} \parallel p^{'}\right)$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n+2}{2}},\ p_{c_j}^{'}=\frac{m_{c_j}+w_{ij}}{\binom{n_j+2}{2}}$$

در صورت انتخاب عمل ۴ مقدار جدید اهمیت برابر با مقدار زیر خواهد بود:

(۱۵-۶)

$$S_4(\sigma)=\sum_{c\neq c_j}\binom{n_c}{2}D\left(p_c \parallel p^{'}\right)+\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)+\binom{2}{2}D\left(1 \parallel p^{'}\right)$$

$$p^{'}=\frac{m+w_{ij}}{\binom{n+2}{2}},\ p_{c_j}=\frac{m_{c_j}}{\binom{n_j}{2}}$$

تفاوت مقدار اهمیت دو عمل برابر خواهد بود با:

(۱۶-۶)

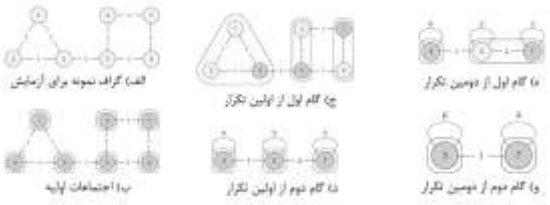
$$\Delta S(\sigma)=\binom{n_{c_j}}{2}D\left(p_{c_j} \parallel p^{'}\right)+\binom{2}{2}D\left(1 \parallel p^{'}\right)-\binom{n_{c_j}+2}{2}D\left(p_{c_j}^{'} \parallel p^{'}\right)$$

اگر $\Delta S(\sigma)>0$ باشد عمل ۴ انتخاب می‌شود، در غیر این صورت عمل ۳ انتخاب خواهد شد.

در ادامه روش ابتکاری لو وین بررسی و تشریح خواهد شد و سپس روش مدنظر در این نوشته ارائه خواهد شد.

۶-۳- روش ابتکاری لووین

با استفاده از یک مثال به بررسی روش لو وین می‌پردازیم. گراف مورداستفاده در این مثال در شکل ۶-۲ نمایش داده شده است. الگوریتم لووین از یک روش حریصانه برای بهینه سازی پیمانگی خوشه‌بندی استفاده می‌کند [۱۶]. این روش یک روش غیرمعیّن



شکل ۶-۲ مراحل روش لووین

[۳۵] است، به این معنی که در اجراهای مختلف این الگوریتم، ممکن است خروجی‌ها متفاوت باشد. در این روش اجتماعات با بیشینه‌سازی یک تابع هدف در دو گام و به‌صورت تکراری کشف می‌شوند. در گام اول اجتماعات کوچکی از طریق بهینه‌سازی محلی پیمانگی تشکیل می‌شوند. در این گام تنها تغییرات محلی اجتماعات در نظر گرفته می‌شوند. در گام دوم، گره‌های متعلق به یک اجتماع با یکدیگر ادغام شده و یک گرهٔ واحد را تشکیل می‌دهند. خروجی گام دوم یک گراف ادغام شده است که به‌عنوان ورودی گام اول تکرار بعدی الگوریتم استفاده می‌شود.

[35] Non-deterministic

این دو گام تا زمانی که بهبودی در پیمانگی مشاهده نشود تکرار می‌شوند.

۶-۴- الگوریتم پیشنهادی برای کشف اجتماعات شبکه‌های یک‌لایه‌ی پویا با روش ابتکاری مبتنی بر معیار اهمیت

در این بخش روشی ابتکاری برای کشف اجتماعات در شبکه‌های یک‌لایهٔ پویا با تمرکز بر بیشینه‌سازی معیار اهمیت ارائه خواهد شد. در ادامهٔ این نوشته این روش تحت عنوان DySig شناخته خواهد شد.

روش ابتکاری ارائه شده مانند روش لووین از دو گام تکراری تشکیل شده است. برای پیشتیبانی از شبکه‌های یک‌لایه‌ی پویا تغییراتی در معیار اهمیت داده شد تا بتوان پس از اضافه و حذف یال از شبکه ساختار اجتماعات را با هزینه‌ای اندک به روز رسانی کرد. در واقع میزان تغییر در معیار اهمیت پس از انتقال محلی گره‌ها بین اجتماعات محاسبه شده و اجتماع مناسب برای انتقال گره انتخاب می‌شود. انتقال مجموعه گره‌ی i با n_i گره و e_{ir} یال در اجتماع r و e_{is} یال در اجتماع s از اجتماع r به اجتماع s تفاوت اهمیتی مطابق با فرمول ۵-۱۷ ایجاد می‌کند.

بعد از اضافه‌کردن یال جدید به گراف، باتوجه‌به نوع یال ۴ عمل زیر قابل انجام می‌باشد:

۱- ساختار اجتماعات بدون تغییر باقی بماند.

۲- دو اجتماع با یکدیگر ادغام شوند.

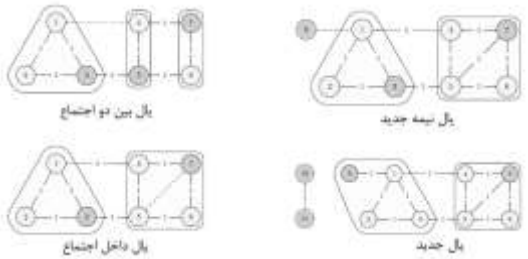
۳- گره‌های جدید به یک اجتماع موجود تخصیص داده شوند.

۴- یک اجتماع جدید ساخته شود و گره‌های جدید به آن اختصاص داده شوند.

شکل ۶-۳ چهار نوع یال ممکن را نمایش می‌دهد.

(۶-۱۷)

$$\Delta S(\sigma)=\binom{n_r}{2}D(q_r \parallel p)-\binom{n_r-n_i}{2}D(q_r^{'} \parallel p)$$
$$-\binom{n_s}{2}D(q_s \parallel p)+\binom{n_s+n_i}{2}D(q_s^{'} \parallel p),$$
$$q_r^{'}=\frac{m_r-e_{ir}}{\binom{n_r-n_i}{2}}, q_s^{'}=\frac{m_s-e_{is}}{\binom{n_s+n_i}{2}}$$



شکل ۶-۳ انواع یال‌های افزایشی به گراف

برای پشتیبانی کامل از شبکه‌های پویا، نیاز است حذف یال‌ها نیز در نظر گرفته شود. باتوجه‌به نوع یال حذف شده، ۴ عمل زیر قابل انجام است:

۵- ساختار اجتماعات بدون تغییر باقی بماند.

۶- یک اجتماع به دو یا بیشتر اجتماع دیگر شکسته شود.

۷- یک یال لبه از یک اجتماع حذف شود.

۸- اجتماعی که با دو گره‌ی یال حذف شده ایجاد شده است حذف شود.

درواقع عمل‌های ۲ و ۸ تعداد اجتماعات را کاهش می‌دهند، در حالی که عمل‌های ۴ و ۶ تعداد اجتماعات را افزایش می‌دهند. عمل‌های ۳، ۵ و ۷ نیز تغییری در تعداد اجتماعات ایجاد نمی‌کنند.

الگوریتم ۶-۱ شبه کد روش DySig را نمایش می‌دهد. ورودی الگوریتم گراف $G(V,E)$ و لیست یال‌های اضافه شده به گراف یعنی A و یا کم شده از گراف یعنی R می‌باشد. برای ذخیره‌ی ساختار اجتماعات، از دو گراف استفاده شده است، C_{ll} که ساختار گراف اصلی را نگه‌داری می‌کند و C_{ul} که ساختار گراف حاصل از ادغام اجتماعات را نگه‌داری می‌کند.

۶-۵- تشریح مراحل الگوریتم پیشنهادی

در ابتدا خوشه‌های گراف اولیه با استفاده از روش ابتکاری مبتنی بر اهمیت تعیین می‌شود. سپس یال‌های موردنظر برای حذف و

الگوریتم ۶-۱ شبه کد DySig

```
1: V ← {u1, u2, .., un} , E ← {(i1, j1), (i2, j2), .., (im, jm)}
2: A ← array{(i1, j1), .., (im, jm)}
3: R ← array{(i1, j1), .., (im, jm))}
4: procedure MAIN(G ← (V, E), A, R)
5:     C_ll ← {C1, C2, .., Cn}, C_ul ← {}, C_new ← C_ll
6:     INITPARTITION(C_new)
7:     sig ← Significance (C_new), old_sig ← 0
8:     m ← 1, n ← 1
9:     while ( sig ≥ old_sig ∨ m ≤ |A| ∨ e ≤ |R|) do
10:        C_new ← ONELEVEL(C_new)
11:        (n, c) ← COMMUNITYCHANGEDNODES(C_ll, C_new)
12:        C_ll ← UPDATECOMMUNITIES(C_ll, n, c)
13:        old_sig ← sig , sig ← Significance (C_ll)
14:        C_ul ← PARTITIONTOGRAPH(C_ll)
15:        if m ≤ |A| then
16:            (src, dest) ← A[m]
17:            a_nodes ← AFFECTEDBYADDITION(src, dest, C_ll)
18:            C_ll ← ADDEDGE(src, dest, C_ll)
19:            C_ll ← DISBANDCOMMUNITIES(C_ll, a_nodes)
20:            C_ul ← SYNCCOMMUNITIES(C_ll, C_ul, a_nodes)
21:        end if
22:        if n ≤ |R| then
23:            (src, dest) ← R[n]
24:            a_nodes ← AFFECTEDBYREMOVAL(src, dest, C_ll)
25:            C_ll ← REMOVEEDGE(src, dest, C_ll)
26:            C_ll ← DISBANDCOMMUNITIES(C_ll, a_nodes)
27:            C_ul ← SYNCCOMMUNITIES(C_ll, C_ul, a_nodes)
28:        end if
29:        C_new ← C_ul, m ← m + 1, n ← n + 1
30:    end while
31: end procedure
```

اضافه شدن به صورت متوالی در شبکه اعمال می‌شوند. با هر تغییر در مجموعه یال‌های گراف، اجتماعاتی که مربوط به یال‌های تغییر یافته هستند از گراف C_{ul} استخراج می‌شوند و با استفاده از گراف C_{ll} به گره‌های سازنده تجزیه می‌شوند. سپس بعد از حذف یا اضافه کردن یال مدنظر، الگوریتم DySig تنها روی گره‌های حاصل از تجزیه‌ی اجتماعات تغییریافته اجرا می‌شود و ساختار جدید اجتماعات شکل می‌گیرد و در نهایت، گراف‌های C_{ul} و C_{ll} به روز رسانی می‌شوند.

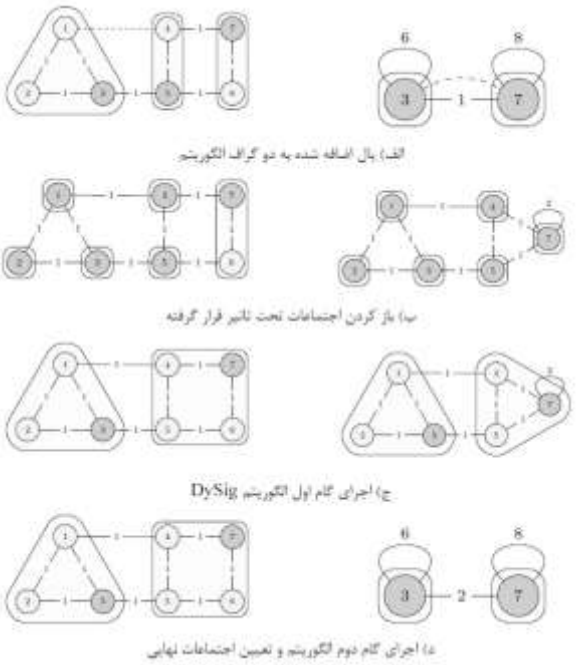
شکل ۶-۴ تمام مراحل الگوریتم DySig را بعد از اضافه شدن یک یال بین اجتماع به گراف نمایش می‌دهد.

روش ابتکاری لو وین نسبت خطی با تعداد یال‌ها دارد و از مرتبهٔ $O(m)$ یعنی تعداد یال‌ها می‌باشد. در حالت پویا، تنها یک یا دو اجتماعی که تحت‌تأثیر حذف و یا اضافه‌کردن یال قرار گرفته‌اند مجدداً خوشه‌بندی می‌شوند که این عمل نیز در بدترین حالت از مرتبهٔ $O(m)$ خواهد بود. البته در عمل اندازهٔ یک اجتماع بسیار کوچک‌تر از اندازهٔ کل گراف خواهد بود و بنابراین زمان اجرا کمتر از حالت حدی خواهد بود.

۶-۶- ارزیابی الگوریتم

۶-۶-۱- ارزیابی روی مجموعه داده‌های ساختگی

در این بخش با استفاده از مجموعه‌دادهٔ تولید شده توسط LFR Benchmark به بررسی نتایج حاصل از الگوریتم DySig می‌پردازیم.



شکل ۶-۴ مراحل اجرای الگوریتم DySig بعد از اضافه شدن یک یال بین اجتماع

برای بررسی میزان صحت تشخیص اجتماعات، از معیار NMI استفاده شده است.

معیار NMI (Normalized Mutual Information) یک معیاری است که برای اندازه‌گیری تطابق بین دو تقسیم‌بندی یا اجتماع استفاده می‌شود. این معیار بر پایه اطلاعات مشترک بین دو تقسیم‌بندی می‌تواند میزان همبستگی و تشابه بین آن‌ها را سنجیده و ارزیابی کند. NMI به دو عامل اصلی بر می‌گردد: آنتروپی و اطلاعات مشترک. آنتروپی یک مفهوم از حافظهٔ تصادفی است و نشان می‌دهد که چقدر می‌توانیم انتظار داشته باشیم که یک رویداد خاص در یک تقسیم‌بندی رخ دهد. آنتروپی برابر با صفر است درصورتی‌که تقسیم‌بندی دقیقاً یک اجتماع را تشکیل دهد، و بیشترین مقدار خود را در صورتی به دست می‌آورد که تقسیم‌بندی هیچ الگویی نداشته باشد و به طور کامل تصادفی باشد.

اطلاعات مشترک میزان تطابق بین دو تقسیم‌بندی را نشان می‌دهد و بر اساس اندازه و جهش‌های مشترک بین دو تقسیم‌بندی محاسبه می‌شود. NMI با استفاده از این دو عامل، یک مقدار نرمال‌شده را بر می‌گرداند که بین ۰ و ۱ واقع می‌شود. این مقدار به ما می‌گوید که تقسیم‌بندی‌ها چقدر با یکدیگر همخوانی دارند، به این صورت که مقدار صفر به معنای هیچ تطابق و مقدار یک به معنای تطابق کامل است؛ بنابراین، با استفاده از معیار NMI می‌توانیم تشابه و همبستگی بین دو

تقسیم‌بندی یا اجتماع را اندازه‌گیری کنیم و معیار NMI برای بررسی صحت تشخیص اجتماعات می‌تواند به ما اطلاعات مفیدی درباره تطابق بین دو تقسیم‌بندی یا اجتماعات مختلف در دست دهد. با محاسبه NMI بین دو تقسیم‌بندی، می‌توانیم بفهمیم که آیا اجتماعات در هر دو تقسیم‌بندی به طور مشابه شناسایی شده‌اند یا خیر.

اگر NMI برابر با ۱ باشد، این نشانگر تطابق کامل بین دو تقسیم‌بندی است، به این معنی که هر دو تقسیم‌بندی دقیقاً یکسان هستند. اگر مقدار NMI بسیار نزدیک به ۱ باشد، نشان‌دهنده تطابق قوی بین دو تقسیم‌بندی است. در مقابل، اگر NMI به سمت صفر حرکت کند، تطابق کم و به طور تصادفی بین دو تقسیم‌بندی نشان می‌دهد. به‌عنوان‌مثال، فرض کنید دو روش مختلف برای تقسیم‌بندی یک مجموعه‌داده را ارائه داده‌ایم. با محاسبه معیار NMI می‌توانیم بفهمیم که این دو تقسیم‌بندی چقدر به یکدیگر نزدیک هستند و در چه میزان اجتماعات مشابهی را شناسایی کرده‌اند. این معیار می‌تواند ما را در انتخاب بهترین تقسیم‌بندی برای مجموعه‌داده‌ها یا ارزیابی دقت الگوریتم‌ها کمک کند.

فرمول محاسبه معیار NMI بر اساس آنتروپی و اطلاعات مشترک به‌صورت زیر است:

$$NMI = \frac{2 \times I(X;Y)}{H(X)+H(Y)}$$

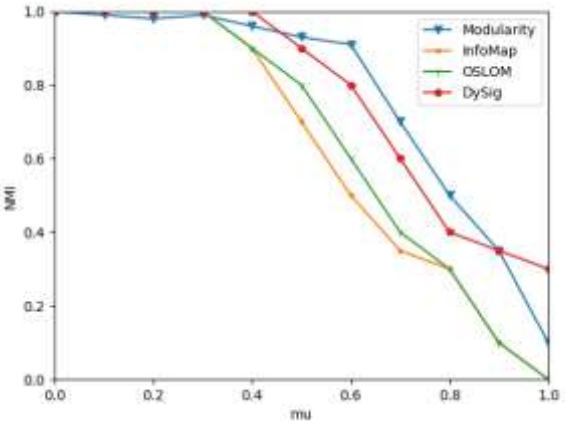
که در آن X و Y دو تقسیم‌بندی یا اجتماع مورد بررسی هستند.

I(X;Y) اطلاعات مشترک بین X و Y است که برابر است با آنتروپی X منهای آنتروپی شرطی X به‌شرط Y (بالعکس).

H(X) و H(Y) به ترتیب آنتروپی تقسیم‌بندی X و تقسیم‌بندی Y هستند.

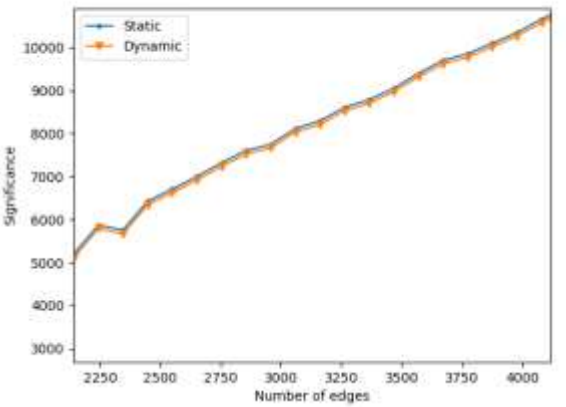
برای این منظور مجموعه‌داده‌های مختلفی با پارامترهای مختلف μ تولید شد و اجتماعات آن توسط الگوریتم‌های مختلفی استخراج شد. برای مقایسهٔ عملکرد الگوریتم DySig از الگوریتم‌های Modularity، Infomap و OSLOM استفاده شده است. نتیجهٔ اجرای این الگوریتم‌ها در شکل ۶-۵ قابل‌مشاهده است.

همان‌طور که مشاهده می‌شود برای پارامترهای وضوح مختلف عملکرد الگوریتم DySig بهتر از Infomap [۲۲] و OSLOM [۲۳] و قابل‌مقایسه با Modularity [۱۱] می‌باشد.



شکل ۶-۵ مقایسه‌ی NMI الگوریتم‌های Modularity، Infomap، OSLOM و DySig روی LFR Benchmark

برای بررسی کیفیت پویایی الگوریتم DySig نیز از مجموعه‌دادهٔ تولید شده توسط LFR Benchmark استفاده شد. گراف تولید شده شامل ۵۰۰ گره و ۴۱۱۸ یال می‌باشد. در ابتدا نیمی از یال‌ها به عنوان گراف پایه خوشه‌بندی شد و سپس در ۲۰ بازه‌ی زمانی مختلف یال‌ها به صورت تصادفی به گراف اضافه شدند و الگوریتم پویا و ایستا روی آن اجرا شدند. میزان تفاوت در معیار اهمیت در دو حالت ایستا و پویا بیانگر کیفیت الگوریتم پویا می‌باشد. نتیجه‌ی این آزمایش در شکل ۶-۶ نمایش داده شده است. مطابق انتظار، الگوریتم پویا تفاوتی با الگوریتم ایستا ندارد و دو نمودار منطبق بر یکدیگر هستند.



شکل ۶-۶ مقایسه‌ی الگوریتم پویا و ایستای مبتنی بر معیار اهمیت

۶-۶-۲- ارزیابی روی مجموعه داده‌ی حقیقی

برای بررسی زمان اجرای روش DySig ابتدا روش را روی مجموعه‌دادهٔ شبکه استناد نظریه فیزیک انرژی‌های بالا [۳۶] که یک شبکهٔ بدون جهت با ۲۷۷۷۰ گره و ۳۵۲۸۰۷ یال می‌باشد

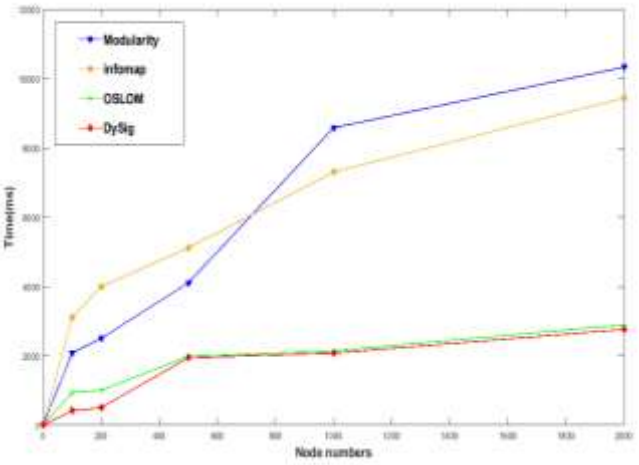
[36] High-energy physics theory citation network

آزمایش می‌کنیم. نتایج بعد از اضافه کردن ۱، ۲، ۵، ۱۰ و ۲۰ دسته یال ۱۰۰ تایی برای هر افزایش، محاسبه شده اند.

جدول ۶-۱ نتایج حاصل از مقایسه‌ی روش DySig با روش‌های Modularity، Infomap و OSLOM را نمایش می‌دهد. در این آزمایش پس از اضافه شدن ۱۰۰، ۲۰۰، ۵۰۰، ۱۰۰۰، و ۲۰۰۰ گره و اجرای الگوریتم‌ها زمان اجرا بر حسب میلی ثانیه اندازه‌گیری شده است.

جدول ۶-۱ مقایسه‌ی زمان اجرای الگوریتم‌های مختلف

	۱۰۰	۲۰۰	۵۰۰	۱۰۰۰	۲۰۰۰
Modularity	۲۰۸۶	۲۵۰۱	۴۱۰۳	۸۵۷۴	۱۰۳۲۴
Infomap	۳۱۰۷	۳۹۹۴	۵۱۱۴	۷۳۰۱	۹۴۳۳
OSLOM	۹۴۶	۱۰۰۳	۱۹۸۴	۲۱۴۰	۲۸۸۸
DySig	۴۱۱	۵۰۴	۱۹۴۲	۲۰۷۴	۲۷۵۹



شکل ۶-۷ مقایسه‌ی زمان اجرای الگوریتم‌های مختلف

۶-۷- جمع بندی و نتیجه گیری

در این مقاله سعی بر این بود تا رویکردی جدید مشابه معیارهای مشابهت اجتماع در شبکه‌های اجتماعی به‌منظور بهبود این روش‌ها ارائه گردد. برای حل مشکلاتی مانند محدودیت وضوح معیار پیمانگی، از معیار اهمیت استفاده شد. همچنین دو روش قبل تنها برای شبکه‌های افزایشی کاربرد داشتند، اما روش ارائه شده در این مقاله از تمام حالات مختلف از جمله حذف یال پشتیبانی می‌کند.

همچنین اجتماعات کشف شده با این روش کاملاً منطبق بر اجتماعات بهینه هستند و معیار اهمیت در هر زمان، برابر با مقدار بهینه می‌باشد. روش ارائه شده در این مقاله نیز مانند دو روش قبلی تنها تغییرات شبکه را در نظر گرفته و ساختار اجتماعات بخش تغییریافته را به‌روزرسانی می‌کند و در نتیجه از سرعت بالایی برای کشف اجتماعات برخوردار می‌باشد.

۷- منابع

[1] Kim, Jungeun, and Jae-Gil Lee. "Community detection in multi-layer graphs: A survey." *ACM SIGMOD Record* 44, no. 3 (2015): 37-48.

[2] Berlingerio, Michele, Michele Coscia, and Fosca Giannotti. "Finding and characterizing communities in multidimensional networks." In *2011 international conference on advances in social networks analysis and mining*, pp. 490-494. IEEE, 2011.

[3] Börner, Katy, Michael Conlon, Jon Corson-Rikert, and Ying Ding. "VIVO: A semantic approach to scholarly networking and discovery." *Synthesis lectures on the Semantic Web: theory and technology* 7, no. 1 (2012): 1-178.

[4] Fortunato, Santo, and Darko Hric. "Community detection in networks: A user guide." *Physics reports* 659 (2016): 1-44.

[5] Newman, Mark EJ. "Spectral methods for community detection and graph partitioning." *Physical Review E* 88, no. 4 (2013): 042822.

[6] Loe, Chuan Wen, and Henrik Jeldtoft Jensen. "Comparison of communities detection algorithms for multiplex." *Physica A: Statistical Mechanics and its Applications* 431 (2015): 29-45.

[7] Whang, Joyce Jiyoung, David F. Gleich, and Inderjit S. Dhillon. "Overlapping community detection using neighborhood-inflated seed expansion." *IEEE Transactions on Knowledge and Data Engineering* 28, no. 5 (2016): 1272-1284.

[8] Leskovec, Jure, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. "Statistical properties of community structure in large social and information networks." In *Proceedings of the 17th international conference on World Wide Web*, pp. 695-704. 2008.

[9] MacQueen, J. "Classification and analysis of multivariate observations." In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281-297. 1967.

[10] Newman, Mark EJ. "Spectral methods for community detection and graph partitioning." *Physical Review E* 88, no. 4 (2013): 042822.

[11] Newman, Mark EJ. "Modularity and community structure in networks." *Proceedings of the national academy of sciences* 103, no. 23 (2006): 8577-8582.

[12] Aldecoa, Rodrigo, and Ignacio Marín. "Surprise maximization reveals the community structure of complex networks." *Scientific reports* 3, no. 1 (2013)

[13] Lancichinetti, Andrea, Filippo Radicchi, and José J. Ramasco. "Statistical significance of communities in networks." *Physical Review E* 81, no. 4 (2010): 046110.
[14] Dubik, Mikael. "A comparative evaluation of state-of-the-art community detection algorithms for multiplex networks." (2017).
[15] Traag, Vincent A., Gautier Krings, and Paul Van Dooren. "Significant scales in community structure." *Scientific reports* 3, no. 1 (2013): 1-10.
[16] Dubik, Mikael. "A comparative evaluation of state-of-the-art community detection algorithms for multiplex networks." (2017).
[17] Fernandes, Andreia, Patrícia CT Gonçalves, Pedro Campos, and Catarina Delgado. "Centrality and community detection: a co-marketing multilayer network." Journal of Business & Industrial Marketing (2019).
[18] Bakhthemmat, Ali, and Mohammad Izadi. "Communities detection for advertising by futuristic greedy method with clustering approach." Big Data 9, no. 1 (2021): 22-40.
[19] Su, Yan. "Accurate Marketing Algorithm of Network Video Based on User Big Data Analysis." *Mathematical Problems in Engineering* 2022 (2022).
[20] Leiva, Fabiola Herrera, Romina Torres, Orietta Nicolis, and Rodrigo Salas. "Characterization of the chilean public procurement ecosystem using social network analysis." *IEEE Access* 8 (2020): 138846-138858.
[21] Sharma, Prem Sagar, Divakar Yadav, and R. N. Thakur. "Web Page Ranking using Web Mining Techniques: A comprehensive survey." *Mobile Information Systems* 2022 (2022).
[22] Yu-Liang, Lu, Tian Jie, Guo Hao, and Wang Yu. "Infomap based community detection in weibo following graph." In 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control, pp. 1220-1222. IEEE, (2012).
[23] Lancichinetti, Andrea, Filippo Radicchi, José J. Ramasco, and Santo Fortunato. "Finding statistically significant communities in networks." PloS one 6, no. 4 (2011): e18961.