
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی کامپیوتر – هوش مصنوعی

آنالیز شبکه‌های اجتماعی برخط برای یافتن انجمن‌های همپوشان

نگارنده: سید محمد مهدی صالحی

استاد راهنما

دکتر علی‌اکبر پویان

استاد مشاور

دکتر حمید حسن‌پور

بهمن ۱۳۹۸

در این صفحه صورت جلسه دفاع را قرار دهید. لازم است پس از صحافی این صفحه مجدداً توسط دانشکده مهر گردد و استاد راهنما با امضای خود اصلاحات پایان‌نامه را تایید کند.

تقدیم به همسر مهربان و صبورم

آرامش زندگی من

و تقدیم به دختران نازنین و عزیزم

شیرینی زندگی من

تشکر و قدردانی

نخستین سپاس از آن حضرت دوست که هر چه دارم از اوست. آن اولین بی پیشین و آن آخرین بی پسین که به من نعمت حیات و فرصت آموختن را ارزانی داشت.

بر خود فرض می دانم که از استاد راهنمای گرانقدرم جناب آقای دکتر علی اکبر پویان صمیمانه بخاطر هدایت و حمایت ایشان در پیشبرد رساله قدردانی کنم. همچنین صمیمانه قدردان استاد مشاور ارجمندم جناب آقای دکتر حمید حسن پور هستم که راهنمایی ها و مساعدتهای متعددشان، حق بسی فراتر از استاد مشاور را از جانب ایشان بر ذمه من قرار داد. از حسن ظن و بذل توجه این دو عزیز نه تنها در پیشبرد و تکمیل رساله بلکه در سایر مراحل زندگی نیز بهره وافر برده ام.

از ایده های راهگشای جناب آقای دکتر محسن رضوانی و سرکار خانم دکتر مهدی مشایخی، و مصاحبت ها و گپ های علمی دوستانم در آزمایشگاه هوش مصنوعی و آزمایشگاه پردازش تصویر خصوصا گپ های دوست ارجمندم جناب آقای دکتر محمود معلم بسیار ممنونم.

فراهم شدن محیط مناسب برای تحقیق علمی را مدیون و مرهون همسرم و والدین ارجمندش هستم که بزرگوارانه شرایط خاص من را تحمل کردند. از آنان صمیمانه تشکر میکنم.

در انتها با یاد روح پرفتوح پدرم قدردان زحمات بی دریغ مادرم، همسرم، خواهرم، برادرانم، و همه عزیزانی هستم که وجودشان تکیه گاهی برای لحظات سخت من، و دعای خیرشان سرمایه و دلیل پیشرفت من به سوی کامیابی بود.

تعهدنامه

اینجانب سید محمد مهدی صالحی دانشجوی دوره دکتری رشته کامپیوتر (هوش مصنوعی) دانشکده کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه: آنالیز شبکه‌های اجتماعی برخط برای یافتن انجمن‌های همپوشان، تحت راهنمائی دکتر علی اکبر پویان، متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می‌باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .

استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

جذابیت‌ها و قابلیت‌های روزافزون فضای مجازی و فراگیر شدن استفاده از وسایل الکترونیکی شخصی، باعث شده تا محیط اینترنت به بستر اصلی بسیاری از تعاملات اشخاص حقیقی و حقوقی تبدیل شود. به عنوان مثالی در این رابطه، می توان به اقبال فراوان به شبکه‌های اجتماعی برخط اشاره نمود. گراف معادل این شبکه‌ها دارای حجم بالایی از داده‌ها ناشی از تعداد زیاد گره‌ها (انسان‌ها) و یال‌ها (دوستی‌ها و ارتباطات میان افراد) می‌باشد، لذا پردازش اطلاعات این شبکه‌های بزرگ و پیچیده، نیازمند استفاده از روش‌های کارآمد است. انجمن‌یابی بعنوان یک کاربرد مهم تحلیل شبکه‌های اجتماعی، به استخراج و دسته‌بندی اطلاعات یک شبکه از طریق تشخیص مولفه‌های تشکیل دهنده‌اش می‌پردازد. کاربردهای انجمن‌یابی در علوم مختلف باعث شده است تا روش‌های متعددی با اهداف و رویکردهای متفاوت برای نیل به این مهم ابداع شود. در ساده‌ترین حالت، مسئله انجمن‌یابی را که شدیدا به نوع شبکه و نحوه مطالعه آن مرتبط است، می‌توان معادل با افراز یک گراف در نظر گرفت. اما با توجه به افزایش روزافزون اندازه و حجم اطلاعات موجود در شبکه‌ها، عضویت گره‌ها در دو یا چند انجمن، اجتناب‌ناپذیر به نظر می‌رسد. گروهی از الگوریتم‌های انجمن‌یابی که هر گره را فقط به یک انجمن منتسب می‌کنند، همپوشانی گره‌ها و یال‌ها در انجمن‌ها را نادیده گرفته و حجم زیادی از اطلاعات را حذف می‌نمایند. در این رساله، با استفاده از بهبود روش‌های رایج بازنمایی گره و توکاری گراف، بهبود نحوه نمایش برداری گره‌ها (مدل‌کردن اطلاعات ورودی به شبکه یادگیری عمیق)، و نیز استفاده از مولفه‌های محاسباتی یادگیری عمیق نظیر کدکننده خودکار پشته‌شده، به روش کاراتری نسبت به روش‌های فعلی با توجه به معیارهای ارزیابی کارایی در انجمن‌یابی همپوشان شبکه‌های اجتماعی رسیده‌ایم. به غیر از مقایسه با توجه به معیارهای ارزیابی، روش پیشنهادی در مقایسه با عمده روش‌های قبلی از پیچیدگی زمانی بسیار مطلوبی برای حالت متوسط برخوردار است، و با حفظ تعداد زیادی از معیارهای مجاورت، ساختار محلی و سراسری گره‌های گراف را به خوبی در انجمن‌یابی انجام‌شده بازتاب می‌دهد.

کلمات کلیدی: تحلیل شبکه‌های اجتماعی، انجمن‌یابی، همپوشانی، یادگیری عمیق، بازنمایی گراف، توکاری گراف

لیست مقالات مستخرج از پایان نامه

Salehi. S. M., Pouyan. A. A. *"A Fast Algorithm for Community Detection Based on Clique Percolation in Social Networks"*. The First International Conference on Signal Processing & Intelligent Systems (SPIS2015).

Salehi. S. M., Pouyan. A. A. *"Detecting Overlapping Communities in Social Networks with Deep Learning"*, International Journal of Engineering (IJE). Accepted

سیدمحمدمهدی صالحی، علی‌اکبر پویان. *«مروری بر روش‌های مدل‌سازی همپوشانی در الگوریتم‌های انجمن‌یابی شبکه‌های اجتماعی»* مجله مدل‌سازی در مهندسی، دوره ۱۷، شماره ٥٦، بهار ۱۳۹۸

سیدمحمدمهدی صالحی، علی‌اکبر پویان. *«انجمن‌یابی در شبکه‌های اجتماعی به کمک یادگیری عمیق».* چهارمین کنفرانس پردازش سیگنال و سیستم‌های هوشمند (ICSPIS2018) .۱۳۹۷، مقاله برگزیده

فهرست مطالب

فهرست جداول

فهرست اشکال

فهرست اختصارات

AE = AutoEncoder
APL = Average Path Length
BBN = Bayesian Belief Networks
BIGCLAM = CLuster Affiliation Model for BIG Networks
CC = Clustering Coefficient
CD = Community Detection
CIS = Connected Iterative Scan
CNM = Clauset, Newman, and Moore
CONGA = Cluster Overlap Newman Girvan Algorithm
COPRA = Community Overlap PRopagation Algorithm
CPM = Clique Percolation Method
CPMd = directed Clique Percolation Method
CPMw = weighted Clique Percolation Method
DEMON = Democratic Estimate of the Modular Organization of a Network
DL =Deep Learning
DOCS = Detecting Overlapping Community Structure
DSE = Dense Subgraph Extraction
EAGLE = agglomerativE hierarchicAl clusterinG based on maximaL cliquE
FOCS = Fast Overlapped Community Search
GCE = Greedy Clique Expansion
GCC = Giant Connected Component
GF = Graph Factorization
GN = Girvan and Newman
GraRep = Graph Representation
HLC = Hierarchical Link Clustering
HOPE = High-Order Proximity preserved Embedding
LC = Link Communities
LFM = Lancichinetti and Fortunato Method
LFR = Lancichinetti, Furtonato, and Radicchi
LMC = Lumped Markov Chains
LPA = Label Propagation Algorithm
MMSB = Mixed membership stochastic blockmodels
MOSES = Model-based Overlapping Seed ExpanSion
NISE = Neighborhood-Inflated Seed Expansion
NMF = Non-negative Matrix Factorization
NMI = Normalized Mutual Information
ODF = Out Degree Fraction
ONMI = Overlapping Normalized Mutual Information
OSLOM = Order Statistics Local Optimization Method

OSN = Online Social Network
PCO = Probabilistic CO-occurrence
PMI = Pointwise Mutual Information
PPMI = Positive Pointwise Mutual Information
PPR = Personalized PageRank
RAK = Raghavan, Albert, and Kumara
SAE = Stacked AutoEncoder
SB = Stochastic Blockmodels
SCP = Sequential Clique Percolation
SDNE = Structural Deep Network Embedding
SN = Social Network
SNA = Social Network Analysis
SNAP = Stanford Network Analysis Project
SLPA = Speaker-Listener Propagation Algorithm
SVM = Support Vector Machine
TPR = Triangle Participation Ratio
UEOC = Unfold and Extract Overlapping Communities

فصل۱ : مقدمه

۱-۱ درآمد

در این فصل به معرفی کلی رساله و اهداف آن پرداخته می‌شود. ابتدا شبکه‌های اجتماعی معرفی شده و تاریخچه مختصری از آنها بیان می‌شود. در ادامه، انجمن‌یابی[۱] با تاکید بر برداشت‌های مختلف از مفهوم انجمن و نیز کاربردهای آن بررسی خواهد شد. مفهوم همپوشانی[۲] و فرض‌های مسئله انجمن‌یابی در ادامه خواهند آمد. در انتهای این فصل نیز ساختار کلی رساله با توضیح در مورد فصل‌های مختلف آن قید شده است.

از آغاز زندگی بشر بر کره خاکی، نیاز به تعامل[۳] او (به عنوان یک موجود اجتماعی) با دیگر هم‌نوعان و نیز عناصر محیط اطرافش امری اجتناب‌ناپذیر بود. پیشرفت‌های فراوان بشر در زمینه‌های مختلف علمی و اجتماعی، هم فرم ارتباطات جدیدی را تعریف می‌نمود و هم شبکه ارتباطات قبلی را پیچیده‌تر می‌کرد. لذا نیاز به یک چارچوب ساده و استاندارد برای تبیین و نیز بررسی‌های بعدی این ارتباطات احساس می‌شد [۱].

برای مدل‌کردن تعاملات میان مجموعه‌ای از عناصر، از تئوری گراف استفاده می‌شود که در آن، یک گره[۴] نشان‌دهنده یک عنصر موجود در شبکه و یک یال[۵] نشان‌دهنده وجود ارتباط مستقیم میان دو گره از آن شبکه است. لذا یک شبکه اجتماعی ساده، با یک گراف G(V,E) مدل می‌شود که در آن، V مجموعه گره‌ها (نودها، رئوس) و E مجموعه یال‌های گراف (اتصالات، لینک‌ها) است. افزایش جمعیت گره‌ها در یک شبکه بزرگ، افزایش ارتباطات میان آنها (تعداد یال‌ها) و به تبع آن، بزرگترشدن گراف شبکه ارتباطی را باعث می‌شود. در شکل۱-۱ مثالی از یک شبکه اجتماعی به نمایش درآمده است. در ساده‌ترین حالت مدل کردن یک شبکه، فرض‌های زیر اعمال می‌گردد:

---

[۱] Community Detection (CD)

[۲] Overlapping

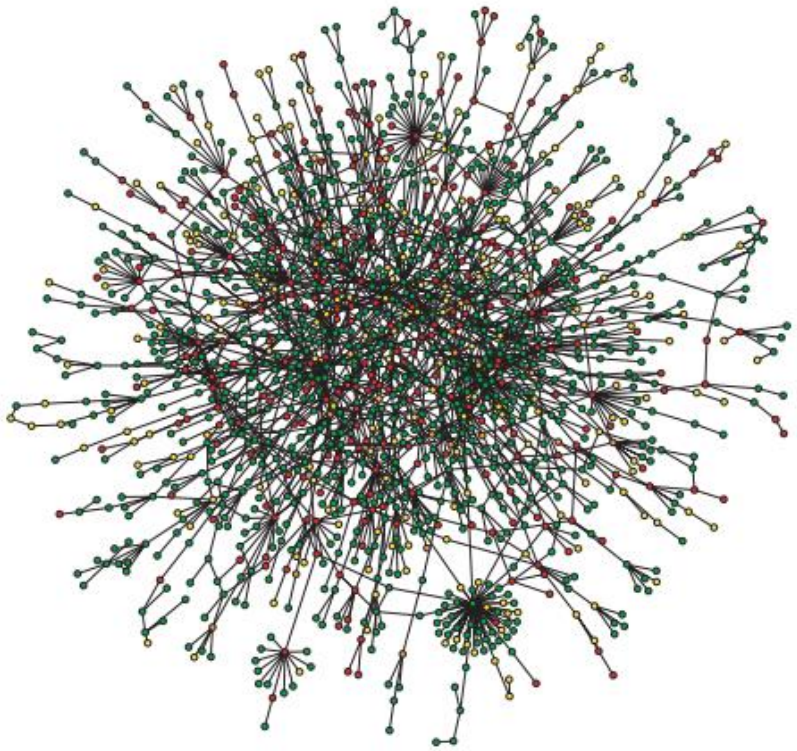
[۳] Interaction

[۴] Node

[۵] Edge

۱. تمام گره‌ها از یک جنس و ماهیت در نظر گرفته می‌شوند.
۲. یال میان دو گره فقط دارای دو وضعیت است (وجود یا عدم وجود)
۳. تمام یال‌ها دارای وزن و اهمیت یکسان می‌باشند.
۴. یال‌ها بدون جهت هستند (ارتباط گره ۱ با گره ۲ به معنای ارتباط گره ۲ با گره ۱ است.)

تعمیم هریک از خواص ذکرشده، باعث افزایش عملکرد و کارایی شبکه‌های پیاده‌سازی‌شده و البته افزایش پیچیدگی‌های محاسباتی شبکه خواهد شد. روشن است که تعداد زیاد گره‌ها و یال‌ها در شبکه‌های پیچیده، باعث ارتباطات آسانتر و کوتاه‌ترشدن مسیر میان اجزای شبکه می‌شود اما دشواری مطالعه و پیش‌بینی رفتار این شبکه‌ها و افزایش احتمال آسیب‌پذیری آنها را هم در پی دارد.



شکل ۱-۱: مثالی از یک شبکه اجتماعی [۲]

رشد سریع فناوری‌های مختلف و فراگیرشدن استفاده از انواع کامپیوترهای شخصی، لپ‌تاپ‌ها، تبلت‌ها، تلفن‌های سلولی چندمنظوره، باعث ایجاد فعالیت‌های جدید در محیط و فضای مجازی شده است که به

عنوان یکی از شاخص‌ترین این موارد می‌توان به «شبکه‌های اجتماعی برخط»[1] اشاره کرد. جذابیت‌های متعدد این شبکه‌ها، استفاده روزافزون از آنها را سبب شده تا آنجاکه افراد بسیاری، استفاده از این شبکه‌های ارتباطی مجازی را به ارتباطات دنیای حقیقی ترجیح می‌دهند[۳]. حجم بسیار بالای داده‌ها در این شبکه‌ها که ناشی از تعداد زیاد گره‌ها (انسان‌ها) و یال‌ها (دوستی‌ها و ارتباطات مختلف میان افراد یا گروه‌ها) است، باعث شده تا پردازش اطلاعات (گردآوری، اعتبارسنجی، اشتراک‌گذاری، بررسی سیر تکاملی زیرشبکه، ...) نیازمند استفاده از روش‌های کارآمدتری نسبت به روش‌های ساده تحلیل گراف باشد.

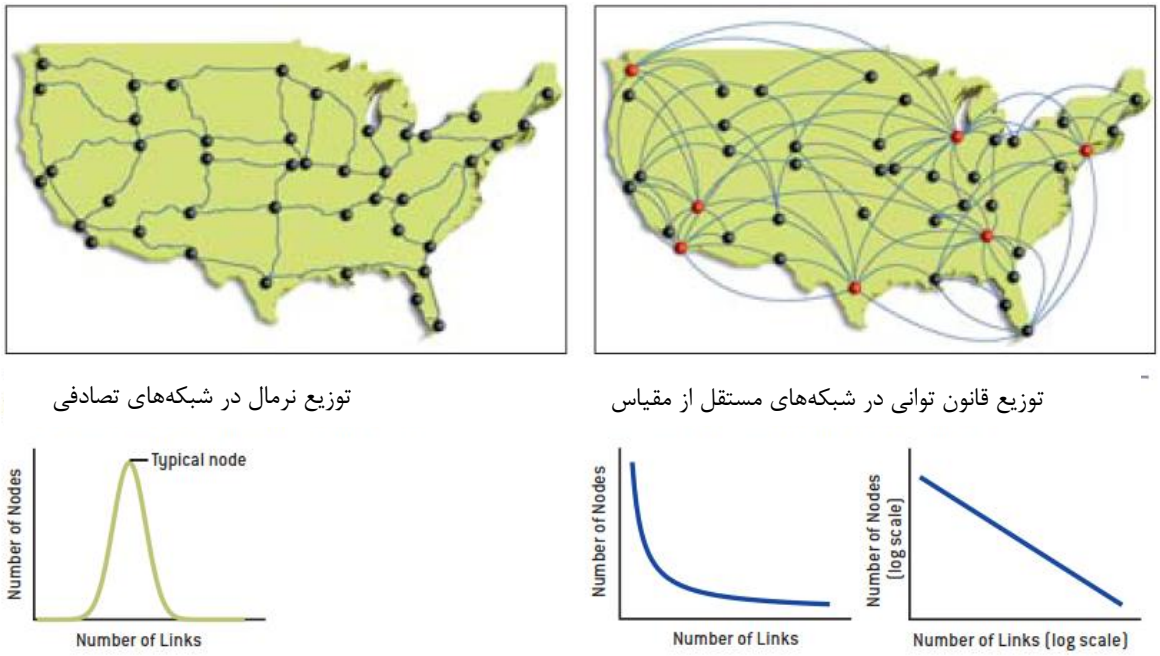
۱-۲ تاریخچه مختصر

شبکه‌های تصادفی نوع خاصی از شبکه‌های پویا[2] هستند که در آنها، افزایش (کاهش) اندازه گراف شبکه ناشی از ایجاد یال‌های جدید (حذف یال‌های قدیم) با متغیرهای تصادفی مدل می‌شود. مطابق شکل ۱-۲، احتمال ایجاد یال میان دو گره در این شبکه‌ها با یک تابع توزیع احتمال (عمدتاً نرمال یا پواسن) بیان می‌شود.

اضافه‌شدن گره‌های جدید به شبکه، مقادیر این احتمال را برای ایجاد هر یال تغییر می‌دهد. مدل‌کردن نحوه ایجاد یال‌ها (یا حذف یال‌های موجود) در زمان رشد یا عمر مفید یک شبکه براساس یک تابع احتمال و متغیر تصادفی، کاربرد زیادی برای توجیه رفتار بسیاری از شبکه‌ها دارد. به عنوان مثالی از این شبکه‌ها می‌توان به شبکه خطوط ریلی یا جاده‌ای اشاره کرد که در آن، شهرها معادل با گره‌ها بوده و خط‌آهن یا جاده‌ها معادل با یال‌ها هستند. در این شبکه‌ها، توزیع یال‌ها میان گره‌ها عادلانه است بدین معنا که تعداد یال‌های متصل به هر گره، تفاوت چندانی با میانگین تعداد یال‌های متصل به گره‌ها ندارد (سمت چپ شکل ۱-۲).

[1] Online Social Networks (OSN)

[2] Dynamic Networks



شکل ۱-۲: مقایسه شبکه‌های تصادفی با شبکه‌های مستقل از مقیاس [۴].

اما با ظهور شبکه‌های جدید و پیچیده‌ترشدن شبکه‌های موجود، به نظر می‌آمد که رفتار بسیاری از شبکه‌ها از این مدل تبعیت نمی‌کند. برای رفع این مشکلات، مدلی برای شبکه‌های مستقل از مقیاس[۱] (مقیاس‌آزاد یا بی‌مقیاس) پیشنهاد شد.

توزیع یال‌ها در شبکه‌های مستقل از مقیاس، از نوع «توزیع قانون توانی»[۲] و به فرم رابطه $x^{-\gamma}$ است که عدد γ برحسب نوع شبکه مورد مطالعه، دارای محدوده تغییرات متفاوتی خواهد بود. عدد γ بیانگر درجه توانی[۳] (ضریب توانی) توزیع بوده و x متغیر توزیع است که مقادیر مختلفی را با توجه به دامنه مسئله (در اینجا: تعداد یال‌ها) اختیار می‌کند. مطابق شکل ۱-۲، این توزیع در مقیاس خطی فرم نمایی میرا دارد و در مقیاس لگاریتمی به فرم خط با شیب γ - نمایش داده می‌شود. به عنوان مثالی از این

[۱] Scale-Free Networks
[۲] Power Law Distribution
[۳] Degree Exponent

گونه شبکه‌ها، می‌توان به شبکه خطوط هوایی اشاره کرد که در آن، بعضی از گره‌ها (فرودگاه‌ها) دارای اهمیت بیشتری نسبت به دیگر گره‌های شبکه هستند.

با توجه به رفتار توزیع توانی، خواص جالبی در شبکه‌های مقیاس‌آزاد قابل مشاهده است. در این نوع از شبکه‌ها، توزیع یال‌ها میان گره‌ها الزاماً عادلانه نیست. تعداد اندکی از گره‌ها (hub) دارای اتصالات زیاد و تعداد زیادی از گره‌ها دارای اتصالات اندکی هستند. به بیان دیگر، این شبکه‌ها برخلاف شبکه‌های تصادفی از مقیاس خاصی پیروی نمی‌کنند. از طرفی، گره‌های جدید تمایل بیشتری دارند تا به hub ها متصل شوند که باعث می‌شود تا درجه گره‌های hub باز هم افزایش پیدا کند[۱]. به خاطر وجود این خاصیت، شبکه‌های مستقل از مقیاس به نام شبکه‌های ترجیحی[۲] نیز معروف هستند. تعداد یال‌های زیاد متصل به hub ها سبب می‌شود تا مختل‌شدن بسیاری از مسیرهای شبکه جبران شود، اما از کار افتادن خود hubها صدمات جبران‌ناپذیری به همراه خواهد داشت. توزیع‌های زیادی برای شبکه‌های مقیاس‌آزاد وجود دارند. مثلاً در توزیع موسوم به zipf[۳]، متغیر x بیانگر فرکانس وقوع یک رویداد در شبکه و رتبه[۴] آن (در مقایسه با دیگر رویدادها) است. مثال‌های متنوعی از شبکه‌های مقیاس‌آزاد در جدول ۱-۱ آورده شده است.

ایده اولیه شبکه‌های اجتماعی برخط (در مقیاس جهانی) از مفهوم جهان کوچک[۵] اقتباس شده است. شبکه‌های جهان کوچک دارای نسبت بالایی از اتصال گره‌ها به یکدیگر بوده و طول مسیر متوسط آنها نیز (با وجود بزرگ بودن خود شبکه) عدد کوچکی است. تعبیر جهان کوچک با عبارت شش درجه جدایی[۶] نیز بیان می‌شود. شش درجه جدایی به این ایده اشاره دارد که اگر فاصله هر فرد را از تمام افرادی که مستقیماً می‌شناسد یک گام (یال) در نظر بگیریم، این فاصله را برای تمام افرادی که با یک

---

۱ Rich gets richer.

۲ Preferential Networks

۳ George Kingsley Zipf

۴ Rank

۵ Small World

۶ Six Degrees of Separation

نفر واسط با آنها آشنایی دارد دو گام (یال) در نظر بگیریم و .... آنگاه میانگین (نه ماگزیمم) فاصله هر دو نفر در کره زمین حداکثر شش گام خواهد بود. صحت این مسئله توسط Duncan Watts در سال ۲۰۰۱ تحقیق شد. او آزمایش خود را با ۴۸۰۰۰ نامه الکترونیکی و ۱۹ گیرنده در ۱۵۷ کشور مختلف دنیا شروع کرد و بعد از پایان آزمایش، به عدد ۶ برای میانگین افراد واسط رسید [۶].

جدول ۱-۱: مثال‌های مختلف از شبکه‌های مقیاس‌آزاد [۲, ۵]

نوع شبکه	برچسب گره‌ها	برچسب یال‌ها
شبکه اجتماعی	افراد	نسبت دوستی، همکاری، وجود علایق مشترک
اینترنت	مسیریابها	اتصالات فیمابین (فیبر نوری، ...)
WWW	صفحات وب	لینک‌های میان صفحات
همکاری‌های دانشگاهی	نویسندگان مقالات	مقالات مشترک
فعالیت‌های سینمایی	هنرپیشه‌ها	فیلم‌های مشترک
متابولیسم سلولی بدن انسان	مولکول‌ها	واکنش‌های بیوشیمیایی (برای سوزاندن غذا و تولید انرژی)
ساختارهای ژنتیکی	پروتئین‌ها	واکنش‌های تعریف‌شده مابین پروتئین‌ها
شبکه‌های خلاف قانون (پولشویی، تروریستی، ...)	افراد یا شرکت‌ها	تراکنش‌های مالی، همکاری در اقدامات مسلحانه
شبکه خطوط هوایی	فرودگاه‌ها	پروازهای مابین دو فرودگاه

در دنیای واقعی، هر شبکه اجتماعی شامل تعدادی از افراد، گروه‌ها و ارتباطات میان آنها است. این ارتباطات (یال‌ها) در محدوده بسیار وسیعی مانند: دوستی، خویشاوندی، انواع تبادلات مالی، علایق هنری یا اجتماعی یا سیاسی، هر صفت مشترک (هم‌شهری، هم‌وطن، هم‌کیش، ...)، تعریف می‌شوند. تحلیل اینگونه شبکه‌ها، بحث مهمی در علوم اجتماعی محسوب می‌شود. با روش‌های تشخیصی می‌توان طبیعت

و الگوی ارتباطات میان اعضای یک گروه خاص را براساس نرم‌های اجتماعی مشخص تحلیل نمود. شبکه‌های اجتماعی برخط، همان شبکه‌های اجتماعی هستند که از اینترنت به عنوان بستر تعاملات میان اعضای خود استفاده می‌کنند. فضای مجازی با امکانات بسیاری که در اختیار شبکه‌های اجتماعی قرار می‌دهد، انعطاف بیشتری را نسبت به شبکه‌های حقیقی ایجاد می‌کند. مواردی مانند[۷]:

- آزاد و بدون هزینه بودن
- ساخت یک پایگاه داده از کاربران و سوابق فعالیت آنان در شبکه
- ارائه خدمات پیام‌های فوری
- امکان اشتراک‌گذاری انواع فایل‌ها (عکس، صوت، ویدئو، ...) و علاقه‌مندی‌ها (دیدگاه‌های سیاسی، فعالیت‌های تجاری، مذهبی، ملیتی یا مبتنی بر تخصص یا هویت خاص)
- ایجاد جوامع و گروه‌های مختلف با جمعیت بسیار بالا (حذف فاصله فیزیکی و جغرافیایی)
- امکان آشنایی و تبادل نظر با کسانی که دیدار رودررو با آنان میسر نیست.
- شکل نوینی از فعالیتهای اجتماعی برای معلولین جسمی و حرکتی
- تعریف سطوح مختلف دوستی و ارتباط با دیگر کاربران (وزن‌دارکردن یال‌های گراف)
- اشتراک در سرویس‌ها و گروه‌ها، بازدید از پروفایل دیگران، follow کردن افراد (گراف جهت‌دار)
- ارتباطات و خدمات مبتنی بر مکان[۱] خصوصاً در شبکه‌های اجتماعی سیار[۲]
- امکان جستجو برای یافتن کاربران دارای علایق یا ویژگیهای مشترک

چنین جذابیتهایی باعث شده تا شبکه‌های اجتماعی برخط، به پدیده‌ای جهانی و مورد اقبال عامه کاربران تبدیل شود. کاربران اینترنتی که حداقل یک‌بار در ماه به سایت یک شبکه اجتماعی مراجعه می‌کنند، از ۴۱٪ کل کاربران فضای مجازی در سال ۲۰۰۸، به ۹۵٪ در سال ۲۰۱۷ رسیده است [۸]. در علوم کامپیوتر، معمولاً ذکر عبارت «شبکه اجتماعی» معادل با «شبکه اجتماعی برخط» در نظر گرفته می‌شود، مگر اینکه خلاف آن تصریح شود.

[۱] Location-Based Services (LBS)

[۲] Mobile Social Networks (MSN)

۱-۳ انجمن‌یابی

یک انجمن (جامعه یا گروه) از یک OSN، مردمی هستند که علایق مشابهی دارند و به تعامل میان خودشان (within) بسیار بیشتر از تعامل با دیگر اعضای شبکه (between) علاقه‌مندند. از دید شبکه، انجمن‌ها را می‌توان مولفه‌های مستقل (به عبارت دقیقتر: نسبتاً مستقل) گراف شبکه دانست که اتصالات (یال‌های) میان گره‌های داخلی این مولفه‌ها زیاد بوده اما اتصالات کمتری با سایر قسمت‌های شبکه اصلی دارند.

در ساده‌ترین حالت، می‌توان مسئله انجمن‌یابی را که شدیداً به نوع شبکه و خروجی ناشی از مطالعه آن مربوط است، با یک مسئله پردازش گراف یا خوشه‌بندی معادل دانست. تئوری گراف بن‌مایه ریاضی بسیار قوی داشته و کاربردهای فراوانی در ریاضیات، فیزیک آماری، مکانیک تحلیلی، و علوم کامپیوتر دارد. لذا مبحث انجمن‌یابی ناهمپوشان معادل با مباحث افراز گراف[۱]، خوشه‌بندی گراف[۲]، و گراف‌کاوی[۳] در نظر گرفته می‌شود.

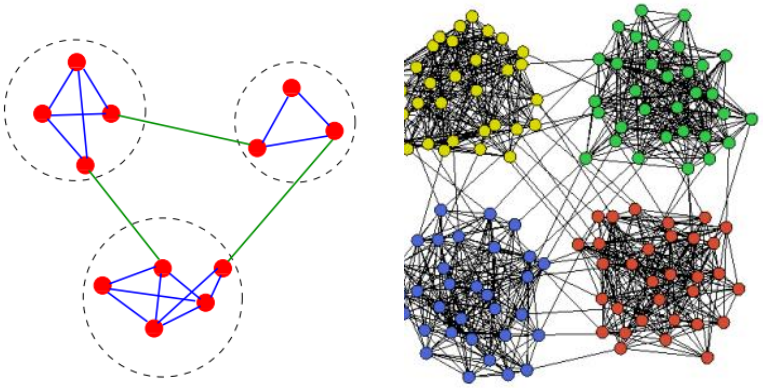
انجمن‌ها بازتاب‌دهنده ویژگی‌ها[۴] و الگوها[۵] مشخص یا مخفی در لایه‌های یک شبکه هستند. لذا بررسی و مطالعه روی انجمن‌های یک شبکه موجبات درک بهتر آن را فراهم می‌کند. کشف و پیش‌بینی روابط جدید (مسیریابی، موارد امنیتی و اقتصادی، امکان‌سنجی‌ها، ...) و داشتن یک دید کلی از نحوه رشد شبکه در آینده، با مطالعه نحوه ایجاد و تکامل انجمن‌های آن میسر می‌شود. یک شبکه بزرگ در سه سطح متمایز قابل بررسی و مطالعه است [۹]:

- در کوچکترین مقیاس ممکن (سطح میکروسکوپی): گره‌های شبکه به عنوان موجودیت
- در بزرگترین مقیاس ممکن (سطح ماکروسکوپی): خود شبکه به عنوان یک موجودیت
- در مقیاس میانی: اجتماع چند گره به عنوان موجودیت مورد مطالعه

---

[۱] Graph Partitioning
[۲] Graph Clustering
[۳] Graph Mining
[۴] Features
[۵] Patterns

روشن است که بررسی انجمن‌های یک شبکه، مطالعات روی شبکه را در مقیاس میانی انجام می‌دهد تا یک دید کلی از شبکه و ماژول‌ها و مولفه‌های عمده آن، فراهم کند. در شکل ۱-۳ نمایشی از انجمن‌های یک شبکه کوچک (سمت چپ) و یک شبکه بزرگ (سمت راست) دیده می‌شود[۱۰].



شکل ۱-۳: انجمن‌ها به عنوان مولفه‌های یک شبکه [۱۰]

مطالعه و تحلیل شبکه‌های اجتماعی به عنوان یک تکنیک کلیدی و یک بحث میان‌رشته‌ای در علوم انسانی (جامعه‌شناسی، روان‌شناسی اجتماعی، مطالعات سازمانی، اقتصاد)، زیست‌شناسی، ریاضی، فیزیک و علوم کامپیوتر (با ظهور OSNها) مطرح شده است. ادعا می‌شود که تعامل افراد در OSNها مدلی از رفتار واقعی آنان است. لذا می‌توان جامعه فعال در OSNها را یک نمونه نسبتاً خوب از جامعه واقعی دانست و روابط موجود در آنها را برای درک بهتر موارد مختلفی از قبیل: کشف ارتباطات، سیر گردش و رشد اطلاعات، ارسال پیامهای مناسب، ... تحلیل نمود.

۱-۳-۱ مفهوم انجمن

بصورت شهودی می‌توان یک انجمن را مجموعه‌ای از موجودیت‌ها دانست که ارتباطات مابین آنها بیشتر (قویتر، نزدیکتر، ...) از ارتباطات با موجودیت‌های خارج انجمن است. نکته مهم، در نحوه تبیین

فرایند این ارتباط و بررسی میزان قوت آن است که منجر به رهیافت‌های[1] متعددی در تعریف و شناسایی انجمن‌ها می‌شود. توصیف یک انجمن توسط این رهیافت‌ها در زیر بیان شده است:

- **نواحی با چگالی بالا در یک گراف:**

تشخیص نواحی پرتراکم در یک گراف و سپس تلاش برای جداسازی این نواحی با حذف یال‌های فیمابین آنها (پل‌ها)، تعمیم روش‌های کلاسیک پردازش گراف

- **بیشینه کردن یک تابع چگالی[2]**

تعریف تابع کیفیت، تعیین مقدار آن براساس اختصاص گره‌ها به انجمن‌ها در یک ساختار سلسله‌مراتبی، تلاش برای افزایش مقدار تابع کیفیت با آرایش مجدد گره‌ها در انجمن‌ها

- **رفتار مشابه در انتشار اطلاعات:**

دسته‌بندی رئوس دارای تاثیر یکسان در انتشار اطلاعات موردنیاز در شبکه، میزان نفوذ گره‌ها روی همسایگانشان، یافتن گره‌های تاثیرگذار

- **یافتن ساختار دقیق و تعریف‌شده از یال‌ها و گره‌های شبکه:**

توجه به ساختار و توپولوژی گره‌ها و یال‌های گراف برای یافتن الگوهای خاص و ترکیب این الگوها با یکدیگر، بی‌توجهی به وظایف و کارکرد گره‌ها در شبکه

- **دسترسی آسان موجودیتها به یکدیگر:**

کم بودن فاصله میان گره‌های یک انجمن در مقایسه با دیگر گره‌های شبکه، هزینه اندک پیمایش گره‌های یک انجمن در مقایسه با گره‌های خارج از آن

- **تعلق رابطه‌ها به انجمن بجای موجودیت‌ها:**

قائل‌شدن اهمیت بیشتر برای یال‌ها، انجمن‌یابی با توجه به رفتار یال‌ها (بجای رئوس)

۱-۳-۲ کاربردها

انجمن‌یابی، به عنوان یکی از مهم‌ترین موارد در تحلیل شبکه‌های اجتماعی به استخراج و دسته‌بندی اطلاعات مهم و ارزشمند از شبکه می‌پردازد. تکنیک‌های انجمن‌یابی (به عنوان نوع خاصی از داده‌کاوی)

[1] Approaches

[2] Density Function

شباهت‌های زیادی با روش‌های خوشه‌بندی داده‌ها[۱] یا قطعه‌بندی تصویر[۲] [۱۱] دارند. موارد زیر تعدادی از مثال‌های موفقی هستند که برای حل مسائل مختلف، انجمن‌یابی را مورد استفاده قرار داده‌اند [۱۲]:

- آسیب‌شناسی شبکه‌های اینترنتی و اینترانتی: مدیریت ترافیک نودهای شبکه
- ناحیه‌بندی مناطق شهری و نحوه ارتباط‌دهی میان آنها به منظور کنترل بهینه ترافیک
- مطالعه شبکه تعاملات افراد یک شرکت و آسیب‌شناسی آن (با توجه به چارت سازمانی)
- مطالعه ارتباطات میان اعضای یک شبکه تروریستی و یافتن افراد موثر در این شبکه‌ها
- افزایش بار در شبکه‌های برق‌رسانی (افت توان شبکه)
- مطالعه نحوه وقوع سقوط آبشاری (Cascading failure)
- روش‌های کارآمد برای وزن‌دهی یال‌های گراف (شدت دوستی‌ها، حجم مبادلات تجاری، میزان داشتن اعتماد به یک فرد یا شرکت خاص، ...)
- مطالعه نحوه آلوده‌سازی یک ویروس مهلک (پیش‌بینی و بررسی اپیدمی‌ها)
- نیت‌کاوی[۳]: پردازش و دسته‌بندی اطلاعات مربوط به نظرات افراد در مورد کالای خاص، عقیده اجتماعی و سیاسی، ... که با توجه به نوع نظر (موافق، مخالف، بی‌تفاوت) و شدت نظر (از شدیداً موافق تا شدیداً مخالف) انجام می‌شود.

۱-۴ همپوشانی

پیش‌فرض مطالعه بسیاری از شبکه‌ها، همپوشانی را مدنظر قرار نمی‌دهد. منظور از همپوشانی این است که یک گره، اجازه تعلق (عضویت) همزمان به دو یا چند انجمن را داشته باشد. هرچند حذف این امر در مطالعه بسیاری از شبکه‌ها خلل جدی ایجاد نمی‌کند اما وجود همپوشانی و پردازش اطلاعات ناشی از مطالعه آن در یک شبکه، یک مزیت مهم در بررسی بسیاری از شبکه‌های پیچیده دنیای واقعی، خصوصاً شبکه‌های اجتماعی محسوب می‌شود. زیرا در این شبکه‌ها اشتراکات زیادی مابین افراد وجود دارد و امکان عضویت آنها در گروه‌های مختلف برحسب علایق مشترک، بسیار محتمل است.

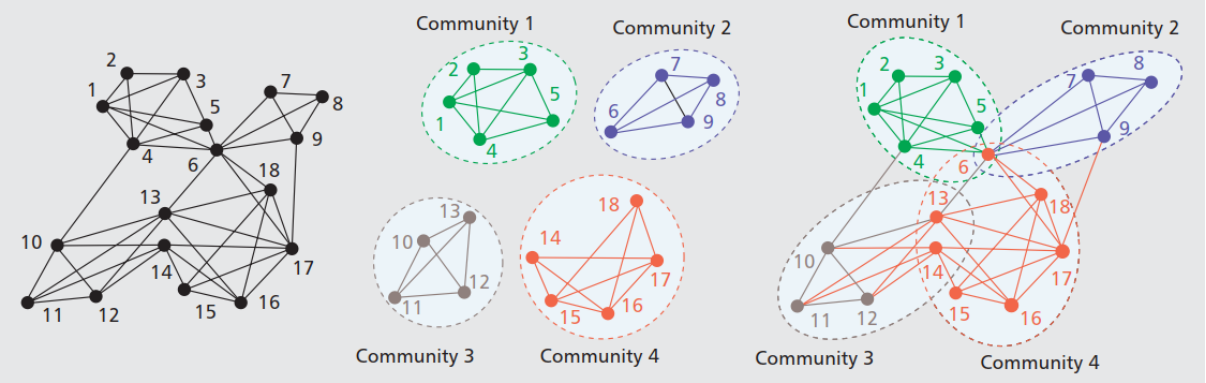
---

[۱] Data Clustering

[۲] Image Segmentation

[۳] Opinion Mining

می‌توان یک شبکه را هم با درنظرگرفتن مفهوم همپوشانی و هم بدون درنظرگرفتن آن مطالعه نمود اما باید توجه داشت که در حالت دوم، حجم زیادی از اطلاعات ذیقیمت را به دلیل بی‌توجهی به یال‌ها و گره‌های مشترک میان انجمن‌ها از دست خواهیم داد. در شکل۱-۴ مقایسه عملکرد میان یک روش بدون همپوشانی با یک روش مدیریت همپوشانی دیده می‌شود [۱۳].



شکل ۱-۴: اهمیت مفهوم همپوشانی در انجمن‌یابی [۱۳]

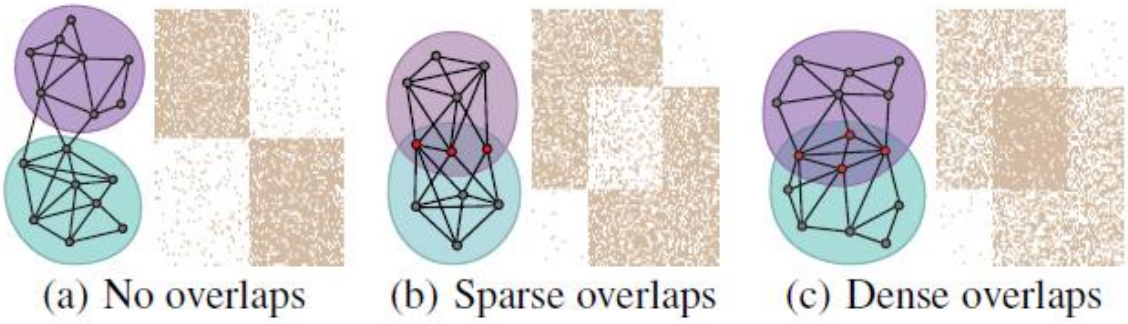
مطابق شکل، رأس ۶ در حالت همپوشان به سه انجمن تخصیص یافته، اما در حالت ناهمپوشان فقط در یک انجمن عضویت دارد، لذا نقش اتصالات این رأس با رئوس۴ و ۵ و ۱۳ و ۱۷ و ۱۸، لحاظ نشده است.

روش‌های مدیریت همپوشانی معمولاً از مفهوم پوشش[۱] به جای افراز (پارتیشن) استفاده می‌کنند. از یک دید کلی می‌توان تمام روش‌ها را به دو دسته همپوشان و ناهمپوشان تقسیم کرد اما روش‌های همپوشان هم ممکن است در تشخیص ناحیه همپوشانی با توجه به میزان تراکم آن، دچار اشتباه شوند[۱۴].

در شکل۱-۵ وجود یا عدم وجود ناحیه همپوشان بین انجمن‌ها و نیز میزان چگال بودن این ناحیه نشان داده شده است. تعدادی از روش‌های انجمن‌یابی، در تشخیص نواحی همپوشان با تراکم اندک

[۱] Cover

مشکلی ندارند اما ناحیه همپوشان پرتراکم را به اشتباه به عنوان یک انجمن مستقل در نظر می‌گیرند. برخی روش‌ها هم بنا بر ماهیت خود، حتی در ناحیه بدون همپوشانی هم دنبال یافتن همپوشانی هستند.



شکل ۱-۵: درجات مختلف همپوشانی: ناهمپوشان، همپوشانی اندک، همپوشانی زیاد. [۱۴]

۱-۵ فرض‌های مسئله

غیر از فرض همپوشانی برای انجمن که مفروض این رساله نیز هست، باید فرض‌های ساده دیگری را در هنگام ایجاد گراف از شبکه موردمطالعه مدنظر قرار داد که در زیر به آنها اشاره می‌شود:

- **Direction**

یال‌های یک گراف ممکن است جهت‌دار یا بدون جهت درنظر گرفته شوند. به عنوان مثال‌هایی از یال‌های جهت‌دار (نامتقارن) می‌توان به لینک یک صفحه وب به صفحه دیگر و دنبال‌کردن افراد در توئیتر اشاره کرد.

- **Weight**

یال‌های یک گراف ممکن است وزن‌دار یا بدون وزن درنظر گرفته شوند. وزن یک یال بیانگر مفاهیمی مانند شدت دوستی میان دو نفر، میزان تعاملات مالی و اقتصادی دو شرکت، دفعات همکاری دو محقق در نوشتن مقاله مشترک، ... است.

- **Dynamism**

پویایی یک شبکه به تکامل آن در طول زمان یعنی ایجاد گره‌ها و یال‌های جدید و نیز حذف گره‌ها و یال‌های قدیمی اشاره دارد. این تغییرات باعث ایجاد انجمن‌های جدید یا حذف و اعمال تغییرات در انجمن‌های قدیمی می‌شود.

- **Input Parameter**
احتمال نیاز به پارامتر ورودی با توجه به نوع روش (عدد پیش‌فرض برای تعداد انجمن‌ها)
- **Multi-Dimensional (Multi-Relational)**
ورودی چندبعدی (بیان روابط غیرهمسان میان رئوس = تعریف چند نوع رابطه در گراف)
- **Multipartite Input**
ورودی چندگانه (شبکه نویسنده – مقاله، شبکه gene-disease، ...)

مطالعات این رساله در مورد شبکه‌های کاربرمحور[۱] انجام می‌شود که نقطه مقابل شبکه‌های محتوامحور[۲] هستند. در شبکه‌های محتوامحور (مانند Flicker و YouTube) اصالت با محتوای مطالب مورد تبادل می‌باشد.

۱-۶ ساختار رساله

فصل۱ شامل معرفی اجمالی شبکه‌های اجتماعی برخط و مفاهیم آن نظیر انجمن، انجمن‌یابی، همپوشانی، و نیز فرض‌های مختلف در تشکیل و یافتن انجمن‌ها است. هریک از این رویکردها به تعریف انجمن، خود منشا ابداع روش‌های نوآورانه و جالبی در انجمن‌یابی شده‌اند که در فصل ۲ مورد بررسی قرار می‌گیرند. این روش‌ها در چند حیطه نظیر انجمن‌یابی براساس طیف گراف، معیارهای مرکزگرایی، تابع چگالی درونی، انتشار برچسب، نزدیکی رئوس، ساختارهای خاص و نیز خوشه‌بندی یال‌ها طبقه‌بندی می‌شوند. فصل ۳ شامل مطالب پیش‌نیاز برای انجمن‌یابی در روش پیشنهادی است. این مطالب شامل معرفی وسیعی از مفهوم بازنمایی و توکاری گراف، مولفه‌های مورد استفاده در یادگیری عمیق، معرفی

[۱] User-based
[۲] Content-based

انواع دادگان‌ها (شبکه‌های دنیای واقعی، محک‌ها، دادگان استخراجی) است. مدل پیشنهادی در فصل ۴ به تفصیل توضیح داده می‌شود. فصل ۵ به تفسیر و مقایسه نتایج حاصله از اعمال روش پیشنهادی و دیگر روش‌های متداول در انجمن‌یابی روی دادگانها براساس معیارهای ارزیابی کارایی می‌پردازد. نهایتا فصل ۶ شامل نتیجه‌گیری و ارائه پیشنهادهایی برای کارهای آتی است.

فصل ۲ : مروری بر کارهای پیشین

۲-۱ مقدمه

همانطور که در فصل اول اشاره شد، انجمن‌یابی بعنوان یک کاربرد تحلیل شبکه‌های اجتماعی، به استخراج و دسته‌بندی اطلاعات مهم یک شبکه از طریق تشخیص مولفه‌های تشکیل‌دهنده و مطالعه آنها می‌پردازد.

رویکردهای متفاوت در یافتن انجمن‌های یک شبکه اجتماعی، باعث استفاده از روش‌های مختلفی شده است که هریک نقاط قوت و ضعف خاص خود را دارند. فصل ۲ به بررسی اجمالی روش‌های مطرح‌شده برای انجمن‌یابی می‌پردازدکه خود شامل محدوده وسیعی از روش‌ها می‌باشد.

دسته‌بندی روش‌های مختلف انجمن‌یابی، مطالعه و مقایسه این روش‌ها را آسانتر می‌کند. قدم اول در این دسته‌بندی، بیان یک تعریف یا برداشت از مفهوم انجمن است تا براساس آن، عمل یافتن انجمن‌ها انجام شود. فی‌المثل در[۱۰, ۱۵] تقسیم‌بندی انجمن‌ها به دو حالت کلی *ساختاری* و *تابعی* مدنظر قرار گرفته است. منظور از ساختار انجمن، آرایش خاصی از گره‌ها است که ارتباطات بین آنها (با توجه به توپولوژی گراف) از ارتباطاتشان با دیگر گره‌های شبکه بیشتر است. در مقابل، عملکرد تابعی به بررسی نقش یا وظیفه گره‌ها در یک شبکه (مثلاً انجام یک وظیفه خاص در شبکه) می‌پردازد تا براساس آن، گره‌های شبکه را در انجمن‌های مختلف دسته‌بندی کند.

در یک مقیاس وسیعتر، چند فرض در تعریف انجمن‌ها توسط coscia بیان شده است [۱۶]. سپس نحوه تبیین انجمن‌ها با هریک از این رویکردها، شکل‌گیری انجمن‌ها، و نهایتاً روش‌های مورد استفاده در انجمن‌یابی براساس هریک از این رویکردها مورد مطالعه قرار گرفته‌اند.

محققان بسیاری [۱۷, ۱۸] به بررسی نقاط قوت و ضعف روش‌های انجمن‌یابی و مقایسه عملکرد آنها با یکدیگر روی شبکه‌های مختلف پرداخته‌اند. در مطالعه روش‌های فعلی، تاکید روی بررسی روش‌های همپوشان (اجازه تعلق عضویت همزمان یک گره به دو یا چند انجمن) بوده است. در این میان، چند

روش ناهمپوشان نیز به دلیل اهمیت فراوان یا پیشگام‌بودن آنها، بررسی شده‌اند. در حال حاضر، بسیاری از روش‌های ناهمپوشان برای حالت همپوشان نیز تعمیم یافته‌اند .

در بخش ۱-۳-۱، فرض‌های مختلفی در تعریف و تشکیل یک انجمن مطرح شده‌اند. این فرض‌ها، الهام‌بخش الگوریتم‌های متعددی در انجمن‌یابی هستند که در بخش‌های۲-۳ تا ۲-۹ مورد مطالعه قرار می‌گیرند. ابتدا پاره‌ای از مفاهیم گراف در بخش ۲-۲ بررسی می‌شود. بخش۲-۳ به روش‌های طیفی و کلاسیک بررسی و افراز گراف و تعمیم آن‌ها می‌پردازد. روش‌های مبتنی بر معیارهای مرکزگرایی در بخش ۲-۴ و روش‌های مبتنی بر چگالی درونی در بخش۲-۵ مطالعه می‌شوند. بخش۲-۶ به روش‌های مبتنی بر نحوه انتشار اطلاعات می‌پردازد. بخش ۲-۷ به مطالعه آن دسته از روش‌هایی اختصاص دارد که اصالت را به ساختار گراف شبکه می‌دهند و انجمن را منطبق با الگوهای ازپیش‌تعیین‌شده تعریف می‌کنند. روش‌های مبتنی بر نزدیکی رئوس به یکدیگر، در بخش۲-۸ مدنظر قرار می‌گیرند و روش‌هایی که به دسته‌بندی یال‌ها (بجای رئوس) می‌پردازند در بخش ۲-۹ مطالعه می‌شوند. بخش ۲-۱۰ به چالش‌های مسئله انجمن‌یابی و جمع‌بندی مطالب این فصل می‌پردازد.

در انتهای این بخش، (جدول ۲-۳) روش‌های انجمن‌یابی بررسی‌شده در بخشهای قبلی این فصل، با ذکر ویژگی‌های مهم آنها فهرست شده‌اند. به همراه نام اختصاری هر روش، اطلاعاتی چون: مدل مورد استفاده برای تعریف و تشکیل یک انجمن، امکان اعمال روش مربوطه روی گراف‌های جهت‌دار و وزن‌دار، نیاز روش به پارامتر ورودی، پیچیدگی زمانی روش، حمایت از مفهوم همپوشانی، و نهایتاً سال ارائه روش بیان شده است.

۲-۲ مفاهیم گراف

معمولاً ورودی سیستم در تحلیل شبکه‌های بزرگ، داده‌های مربوط به گراف آن شبکه است و هیچ پیش‌فرض خاصی در مورد تعداد انجمن‌های آن شبکه نداریم.

جدول ۲-۱: برخی از مفاهیم مرتبط با گراف شبکه

پارامتر	تعریف
Degree	تعداد یال‌های متصل به یک گره (یک پارامتر محلی است.)
Average Degree	متوسط درجات گره‌ها در یک گراف (یک پارامتر عمومی است و به فرم $2m/n$ بیان می‌شود)
Path Length	مسافت پیموده شده (تعداد یال‌ها) در مسیر بین دو گره یک شبکه
APL[۱]	متوسط طول کوتاهترین مسیر بین هر دو گره یک شبکه
Distance	کوتاهترین مسافت (حداقل تعداد یال‌ها) برای هر دو گره یک شبکه
Diameter	حداکثر تعداد یال ممکن مابین دو گره در شبکه (ماگزیمم Distance ها)
Density	نسبت یال‌های موجود به کل یال‌های ممکن در یک گراف (شاخص میزان تراکم یال‌ها در گراف)
Girth	تعداد یال‌های تشکیل‌دهنده کوتاهترین دور (سیکل) در یک گراف
γ	ضریب ثابت قانون توزیع توانی ($K^{-\gamma}$) برای شبکه‌های بی‌مقیاس
Connectivity	همبند بودن یک گراف، بدین معنا که از هر راس به تمام رئوس دیگر حداقل یک مسیر موجود باشد
GCC[۲]	بزرگترین مولفه همبند یک گراف (دارا بودن بیشترین تعداد رئوس) در گراف‌های کوچک، معمولاً برابر با خود گراف است
Cut point	راسی که با حذف آن، گراف از حالت همبندی خارج شود (به دو یا چند زیرگراف تجزیه شود)
Bridge	یالی که با حذف آن، گراف از حالت همبندی خارج شود (به دو زیرگراف تجزیه شود)
Edges Cut	تعداد یال‌هایی که با حذف آنها ارتباط یک زیرگراف خاص با کل گراف اصلی قطع می‌شود.
ضریب خوشه‌بندی[۳]	نسبت یال‌هایی که رئوس مجاور یگ گره را به هم متصل می‌کنند بر تمام یال‌های ممکن میان همسایگان یک گره. (بررسی احتمال همسایه‌بودن همسایگان یک گره)

[۱] Average Path Length (APL)

[۲] Giant (= Greatest) Connected Component

[۳] Clustering Coefficient (CC)

تمامی روش‌های مطرح‌شده در انجمن‌یابی، از تعاریف و پارامترهای ساده در بیان مشخصات گراف استفاده می‌کنند. تعدادی از این مفاهیم اولیه و ویژگی‌های توپولوژیکی برای یک گراف ساده (با n راس و m یال و بدون حلقه و یال دوبله) در جدول۲-۱ فهرست شده‌اند. از این مفاهیم در روش‌هایی که در ادامه معرفی می‌شوند، استفاده خواهد شد.

هرچند یافتن انجمن‌ها (با توجه به اندازه شبکه‌ها) می‌تواند یک کار طولانی و زمانبر و دارای الگوریتم‌هایی با پیچیدگی بالا باشد، اما تمامی روش‌های مطرح‌شده، از تعاریف و پارامترهای ساده در بیان مشخصات گراف استفاده می‌کنند. هریک از این مفاهیم دارای یک تعبیر شهودی و ساده هستند. مثلاً در مورد درجه یک گره یا متوسط درجات گره‌ها در گراف یا طول کوتاهترین مسیر، می‌توان به این تعبیر رسید: گره‌هایی که درجه بالایی دارند، احتمالاً در مرکز گراف شبکه قرار می‌گیرند و متوسط طول مسیر کوتاه‌تری نسبت به سایر گره‌ها دارند. لذا احتمال تعلق آنها به تعداد زیادی انجمن (شرکت در همپوشانی) بالاست، اما گره‌های موجود در نواحی کناری و حاشیه‌ای گراف از چنین خاصیتی برخوردار نیستند و به احتمال زیاد جزو نواحی ناهمپوشان انجمن‌ها خواهند بود.

۲-۳ انجمن‌یابی مبتنی بر طیف گراف و ماتریس‌های اولیه

باتوجه به مدل‌کردن یک شبکه و عناصر آن با یک گراف، عملاً می‌توان از روش‌های سنتی پردازش و مطالعه یک گراف و نیز تعمیم این روش‌ها برای مبحث انجمن‌یابی نیز استفاده نمود. روش‌های زیادی برای خوشه‌بندی پیشنهاد شده‌اند [۱۹] که با توجه به معیارهای مختلفی به دسته‌بندی داده‌ها می‌پردازند. نقاط قوت و ضعف این روش‌ها باعث شده تا هرکدام در حیطه خاصی از خوشه‌بندی، از کارایی بیشتری برخوردار باشند. یک دسته پرکاربرد از روش‌های خوشه‌بندی، روش‌های سلسله‌مراتبی هستند. این روش‌ها در دو گروه متمایز طبقه‌بندی می‌شوند:

۱. روش‌های تقسیمی[۱]

- شروع کار از خود گراف به عنوان بزرگترین مولفه همبند
- تقسیمات متوالی گراف (با استفاده از معیار عدم‌شباهت) تا رسیدن تعداد مولفه‌های همبند به عدد معین‌شده برای خوشه‌ها
- حداکثر ادامه کار تا زمانی است که هر گره معادل با یک خوشه باشد.
- مشخصاً دارای رویکرد بالا به پایین (top-down) است.

۲. روش‌های تجمعی[۲]

- شروع کار از خود گره‌ها به عنوان کوچکترین مولفه‌های همبند
- ادغام‌های متوالی (با استفاده از معیار شباهت) تا رسیدن تعداد مولفه‌های همبند به عدد معین‌شده برای خوشه‌ها
- حداکثر ادامه کار تا زمانی است که فقط یک خوشه داشته باشیم (خود گراف).
- مشخصاً دارای رویکرد پایین به بالا (bottom-up) است.

روش نمایش عملکرد الگوریتم‌های سلسله مراتبی، استفاده از ساختار درختی دندروگرام[۳] است که قطع آن با یک خط فرضی، خوشه‌های بدست‌آمده تا یک مرحله خاص را مشخص می‌کند.

اطلاعات مربوط به توپولوژی یک شبکه (گراف) از ماتریس مجاورت[۴] آن حاصل می‌شود. این ماتریس در حالت کلی، ماتریسی متقارن (برای گراف غیرجهت‌دار) بوده که مولفه A_{ij} آن، بیانگر وجود یال از راس i به راس j می‌باشد. اساس روش‌های «خوشه‌بندی طیفی[۵]» بر استفاده از ماتریس مجاورت و موارد بدست‌آمده از آن، استوار است. این ماتریسها در ذیل بیان شده‌اند[۱۰]:

- طیف گراف G: معادل با مقادیر ویژه ماتریس A می‌باشد.
- ماتریس R: از تقسیم عناصر هر سطر i ماتریس A بر درجه راس i حاصل می‌شود.
- ماتریس D: ماتریس قطری که در آن، D_{ii} بیانگر درجه راس i است.
- ماتریس T: ماتریس انتقال[۶] نامیده می‌شود و معادل با ترانهاده[۷] ماتریس R است.

---

[۱] Divisive Methods
[۲] Agglomerative Methods
[۳] Dendrogram
[۴] Adjacency
[۵] Spectral Methods
[۶] Transfer
[۷] Transpose

- ماتریس L: ماتریس لاپلاسین (نرمالیزه نشده) می‌باشد و معادل با D-A است.

روش‌های طیفی[۱]، انجمن‌ها را نواحی پرتراکم یک گراف تعریف می‌کنند و این نواحی را به کمک ماتریس مشابهت گراف و ماتریس‌های حاصل از آن تشخیص می‌دهند. هر مؤلفه S_{ij} در ماتریس مشابهت[۲]S، وزن شباهت و ارتباط موجود میان رئوس i و j را در این ماتریس متقارن (در گراف غیرجهت‌دار) بیان می‌کند. روشن است که ماتریس مجاورت یک حالت خاص از ماتریس مشابهت است که فقط مقادیر صفر و یک دارد.

متاسفانه روش‌های طیفی از همپوشانی انجمن‌ها حمایت نمی‌کنند اما تلاش‌های بسیاری برای تعمیم روش‌های مبتنی بر مقادیر و بردارهای ویژه، انجام گرفته است تا حمایت از مفهوم همپوشانی محقق شود. روش Psorakis و همکاران [۲۰] یکی از این موارد شاخص است که از «تجزیه نامنفی ماتریسی»[۳] با کمک روش‌های بیزین، برای انجمن‌یابی همپوشان استفاده کرده است. روش Bayesian-NMF مطرح‌شده، به ارائه یک مدل مولد در چارچوب آماری و دارای پارامترهای مشخص با کمک NMF (بعنوان یک روش low-rank مبتنی بر مقادیر ویژه) می‌پردازد. تکنیک NMF یک روش برای استخراج ویژگی و کاهش ابعاد در مباحث یادگیری ماشین است که در اینجا برای انجمن‌یابی استفاده شده است.

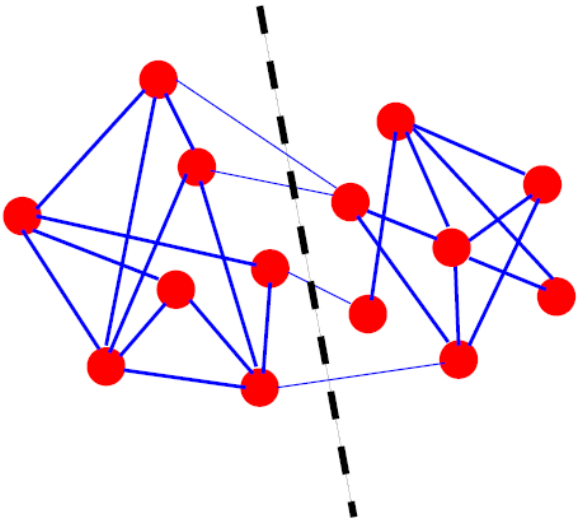
روش دیگری موسوم به BIGCLAM[۴] نیز از ترکیب NMF و خواص مدلهای آماری استفاده می‌کند[۱۴]. در این روش برای نمایش ارتباط موجود میان گره‌ها و انجمن‌ها از یک گراف دوقسمتی به نام «گراف وابستگی انجمنی»[۵] که در [۲۱] معرفی شده است، استفاده می‌شود. این گراف دو نوع گره دارد: انجمن‌ها و گره‌های گراف اولیه. هر یال این گراف که «یال وابستگی»[۶] نامیده می‌شود، معادل با

[۱] Spectral Clustering
[۲] Similarity Matrix
[۳] Non-negative Matrix Factorization (NMF)
[۴] CLuster Affiliation Model for BIG Networks (BIGCLAM)
[۵] Community Affiliation Graph
[۶] Affiliation Edge

یک زوج گره-انجمن بوده، بیانگر وجود و میزان تعلق یک گره به یک انجمن است. این روش همپوشانی دارد، البته باید تعداد انجمن‌ها را بعنوان پارامتر ورودی دریافت کند.

۲-۴ انجمن‌یابی مبتنی بر معیارهای مرکزگرایی

روش‌های مبتنی بر مرکزگرایی نمونه‌ای از روش‌های تقسیمی هستند. یکی از اولین روش‌های تقسیمات در گراف، یک روش تقسیمی موسوم به min-cut است (شکل۲-۱) که در آن، تقسیم یک شبکه به تعداد مشخص و متوازنی از زیرگروه‌ها به کمک حداکثر کاهش (حذف) در تعداد یال‌های بین گروه‌ها انجام می‌شود. توقع می‌رود که هر انجمن مشخص در یک گراف، دارای cut-size کوچکی باشد یعنی تعداد اتصالات اندکی با دیگر قسمت‌های گراف داشته باشد.



شکل ۲-۱: قطع ارتباط برای تقسیم گراف در روش min-cut [۲۲]

روش‌های کلاسیک خوشه‌بندی عمدتاً از مفهوم همپوشانی حمایت نمی‌کنند زیرا برای کسب اطلاعات بسیاری از گراف‌های کوچک، نیازی به بررسی همپوشانی نیست. ضمن اینکه اکثر مباحث پردازش گراف،

مسائل رام‌نشدنی از نوع NP-hard بوده و زمان اجرای نامعقول در گراف‌های بزرگ خواهند داشت. لذا نیاز به روش‌های کارآمدتر یا بهبود روش‌های فعلی احساس می‌شود.

یک ایده جالب در انجمن‌یابی بر این اصل استوار است که انجمن‌ها، نواحی با چگالی بالا در یک گراف هستند. این نواحی به وسیله یال‌های مرزی با یکدیگر در ارتباط می‌باشند (یال مرزی یالی است که دو راس آن به انجمن‌های متفاوتی متعلق باشند). لذا اگر اتصالات مابین قسمتهای چگال یک گراف، شناسایی و حذف شوند، انتظار داریم آنچه باقی می‌ماند انجمن‌های گراف بعنوان تعدادی زیرگراف همبند و مستقل باشند.

بر این اساس، روش ساده و معروف «بینابینی یال»[۱] یا GN[۲] (براساس نام طراحانش) مطرح می‌شود[۲۳]. این روش با استفاده از یک پارامتر مهم مرکزگرایی در گراف، یعنی مفهوم بینابینی[۳] عمل می‌کند. تعریف این مفهوم برای هر یال گراف بدین شرح است: *تعداد کوتاهترین مسیرهای ممکن بین هر دو گره شبکه که بالاجبار از این یال می‌گذرند.*

عدد بینابینی برای هر یال، به‌سادگی با مشخص‌کردن کوتاه‌ترین مسیر میان هر دو گره گراف و سپس شمارش مسیرهایی که قطعاً از یال موردنظر می‌گذرند به دست می‌آید. البته می‌توان این تعریف را با تقسیم بر کل کوتاه‌ترین مسیرهای ممکن، نرمالیزه نمود. تعبیر بینابینی بیانگر میزان قرار گرفتن یک یال بین زوج‌گره‌ها است. همین‌طور می‌توان از آن به میزان عبور ترافیک شبکه از یک گذرگاه (یال) اشاره کرد.

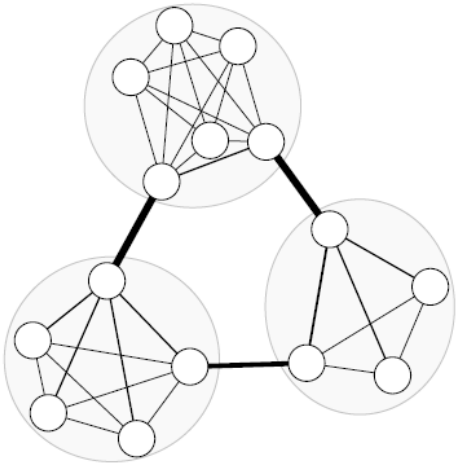
در شکل۲-۲ سه یال دارای بیشترین مقدار بینابینی یال برای یک گراف شامل سه انجمن متمایز، نشان داده شده است.

---

[۱] Edge Betweenness

[۲] Girvan and Newman

[۳] Betweenness



شکل ۲-۲: یال‌های دارای بیشترین درجه بینابینی در یک گراف [۱۶]

الگوریتم این روش که یک مکانیزم تقسیمی دارد به شرح زیر می‌باشد:

۱) مرتب‌سازی تمام یال‌های گراف براساس درجه بینابینی آنها
۲) حذف یال با بزرگترین درجه بینابینی
۳) محاسبه مجدد بینابینی در گره‌های مولفه‌های ایجادشده (حاصل از تقسیم گراف)
۴) تکرار مرحله ۲ تا رسیدن به شرط توقف (تعداد دلخواه انجمن‌ها، ...) یا رسیدن به آخرین یال

روش GN الگوریتم ساده‌ای دارد. اما با توجه به پیچیدگی زمانی نامطلوب آن، برای انجمن‌یابی شبکه‌های بزرگ روش مناسبی نیست[۲۲]. همچنین واضح است که این روش از مفهوم همپوشانی حمایت نمی‌کند.

برای حمایت از مفهوم همپوشانی، تلاش‌های مختلفی برای بهبود و تعمیم روش اولیه ارائه شده است. یکی از این تعمیم‌ها روشی به نام CONGA[۱] است که در کنار معیار بینابینی یال از ایده «بینابینی شکافت[۲]» نیز استفاده می‌کند[۲۴].

---

۱ (CONGA) Cluster Overlap Newman Girvan Algorithm

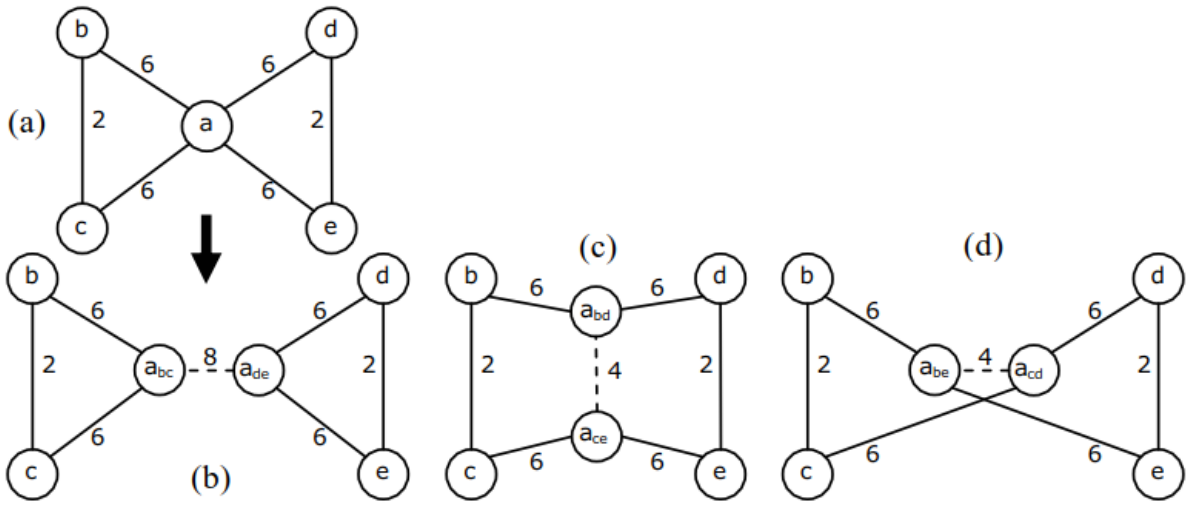
۲ Split Betweenness

معرفی پارامتر بینابینی شکافت، امکان همپوشانی رئوس را فراهم می‌کند. در این روش، یک راس v را به دو راس v_1 و v_2 و یال فرضی مابین آنها می‌شکافیم و سپس بینابینی یال را برای این یال فرضی محاسبه می‌کنیم، یعنی تعداد کوتاه‌ترین مسیرهایی که بالاجبار از این یال فرضی حاصل از شکافت می‌گذرد. بهترین حالت پیش‌آمده را در ارتباط با کلیه حالات ممکن برای شکافت (تقسیم) یک راس با توجه به همسایگانش در نظر می‌گیریم تا مقدار «بینابینی شکافت» را ماگزیموم کند. نکته مهم اینجاست که کدام راس در چه زمانی و چگونه شکافته شود؟

مراحل روش CONGA به شرح زیر است:

۱. محاسبه بینابینی یال برای هر یال و محاسبه بینابینی شکافت برای هر راس
۲. حذف یال با بیشترین مقدار بینابینی یال یا شکافت راس با بیشترین بینابینی شکافت (هرکدام که بزرگتر بود)
۳. محاسبه مجدد بینابینی یال و بینابینی شکافت برای هر یال و راس باقیمانده
۴. تکرار مرحله ۲ تا رسیدن به شرط توقف (تعداد دلخواه انجمن‌ها، ...)

شکل۲-۳ تمام حالات ممکن در شکافت یک راس دلخواه و بهترین انتخاب میان سه حالت موجود را نشان می‌دهد. حالت شکافت راس a به صورت a_{bc} و a_{de} بیشترین مقدار بینابینی شکافت را تولید می‌کند.



شکل ۲-۳: مثالی از شکافت یک راس در روش CONGA [۲۴]

دو مفهوم مهم مرتبط با انجمن‌یابی در [۲۵] مطرح شده است. مفهوم «گریز از مرکز»[۱] برای هر گره v عبارت است از ماگزیمم طول کوتاهترین مسیری که بین گره v و هریک از دیگر گره‌های گراف وجود دارد. از زاویه دید یک گراف، قطر هر گراف برابر با بزرگترین عدد مربوط به کلیه مقادیر گریز از مرکز آن می‌باشد و کوچکترین عدد میان کلیه مقادیر گریز از مرکز، شعاع آن گراف خوانده می‌شود.

این دو مفهوم در کنار سه پارامتر مهم مبتنی بر مرکزگرایی، امکان مطالعات متنوعی را روی انجمن‌های حاصله از روش‌ها، فراهم می‌کنند. این سه پارامتر عبارتند از:

- درجه گره
- بینابینی یال
- نزدیکی[۲] گره: سهولت دستیابی به یک گره خاص از دیگر گره‌های شبکه (کمتربودن مجموع طول مسیر تمام گره‌ها تا آن گره)

یک معیار مهم دیگر در مرکزگرایی ، مفهوم پرکاربرد «تناسب»[۳] است که اشاره به میزان تمایل گره‌های یک شبکه به اتصال با گره‌های هم‌درجه خود دارد. مقدار این پارامتر از عدد 1+ برای شبکه‌های کاملاً متناسب تا عدد 1- برای شبکه‌های کاملاً نامتناسب[۴] متغیر است. فضای وب دارای مقدار منفی برای پارامتر تناسب است. اما شبکه‌های اجتماعی به عنوان مثالی دیگر از شبکه‌های پیچیده و بی‌مقیاس، دارای مقدار مثبت برای این معیار می‌باشند. چون در این‌گونه شبکه‌ها، گره‌های مرکزی (hubs) به اتصال با یکدیگر بیشتر از اتصال با گره‌های پیرامونی تمایل دارند.

روش DOCNet[۵][۲۶] از دو معیار مرکزگرایی شامل درجه و ضریب خوشه‌بندی[۶] (CC) در انجمن‌یابی استفاده می‌کند. ضریب خوشه‌بندی برای گره i، به‌صورت نسبت یال‌های موجود میان گره‌های همسایه

---

[۱] Eccentricity

[۲] Closeness

[۳] Assortativity

[۴] Fully-Disassortative

[۵] Detecting Overlapping Communities in Networks (DOCNet)

[۶] Clustering Coefficient (CC)

با i، به کل یال‌های ممکن میان آن گره‌ها تعریف می‌شود. این ضریب برای هر گره، بیانگر این است که چقدر گره‌های همسایه با آن گره، خود با یکدیگر همسایه هستند. به کمک درجه هر گره و ضریب خوشه‌بندی آن، در روش DOCNet پارامتری به نام «اهمیت گره»[۱] (NI) تعریف می‌شود و مقدارش برای هر گره تعیین می‌گردد. بر اساس آن، یک هسته کاندیدای اولیه بعنوان انجمن مشخص می‌شود. اضافه کردن گره‌های مناسب به هسته، با توجه به NI هر گره انجام خواهد شد.

روش NLA[۲][۲۷] از معیار مرکزگرایی رتبه‌صفحه استفاده می‌کند. معیار رتبه‌صفحه برای یک گره، بیانگر مقدار احتمال ورود به آن گره، با شروع از دیگر گره‌های گراف در یک پیمایش غیرتصادفی است. در شروع، تخصیص مقدار احتمال اولیه به‌طور مساوی برای هر گره گراف (یک عدد کوچک‌تر از ۱ و نزدیک به آن) انجام می‌شود. با انجام مراحل متوالی و پس از رسیدن به یک شرط توقف، رتبه گره‌ها در گراف بر اساس اهمیتشان، مشخص شده است.

در NLA، به هر گره بر اساس رتبه‌صفحه آن، یک عدد جرم اختصاص می‌یابد و کل گراف به یک میدان پتانسیل سه‌بعدی، نگاشته[۳] می‌شود. اختصاص گره‌ها به انجمن بر اساس جرم و مکان قرارگیری گره‌ها در این نگاشت، انجام خواهد شد. این روش در کار با گراف‌های کوچک سرعت زیادی دارد ولی برای گراف‌های بزرگ مناسب نیست.

۲-۵ انجمن‌یابی مبتنی بر چگالی درونی

دسته‌ای از روش‌های انجمن‌یابی، انجمن‌ها را موجودیت‌هایی تعریف می‌کنند که مقدار یک تابع خاص را ماگزیمم می‌نمایند. از این تابع، بعنوان تابع چگالی (تابع کیفیت[۴]) نام برده می‌شود. روش‌های مبتنی بر چگالی درونی، به یافتن نواحی چگالتر (متراکم‌تر) براساس تابع کیفیت تعریف‌شده می‌پردازند. تابع

[۱] Node Importance (NI)

[۲] Node Location Analysis (NLA)

[۳] Mapped

[۴] Quality Fuction

کیفیت، برای هر حالتی از افراز گره‌ها (انجمن‌یابی) مقداری را تولید می‌کند. بیشتر بودن مقدار این تابع، به معنای مطلوبتر بودن کیفیت انجمن‌یابی انجام‌شده است.

یکی از متداولترین توابع کیفیت در انجمن‌یابی، «تابع پیمانه‌ای»[۱] (تابع ماژولاریتی) است که اولین بار در کار clauset و همکاران مطرح شد[۲۸]. انجمن‌یابی به کمک ماژولاریتی که به روش CNM[۲] (براساس نام طراحانش) یا fast greedy نیز مشهور است، محبوبیت زیادی دارد و حتی از تابع ماژولاریتی، بعنوان معیار ارزیابی کارایی در دیگر روش‌ها استفاده می‌شود.

روش CNM میزان خوب‌بودن یک حالت انجمن‌یابی پیشنهادی را به کمک تابع ماژولاریتی و در قالب یک الگوریتم سلسله‌مراتبی تجمعی، به صورت کمّی برآورد می‌کند. این برآورد با مقایسه میان چگالی محاسبه‌شده گراف (برای هر حالت انجمن‌یابی) با چگالی یک «مدل مرجع»[۳] انجام می‌شود. منظور از مدل مرجع، یک گراف تصادفی با همان ساختار و دنباله درجات رئوس است.

همانطور که در فصل اول اشاره شد، گراف یک شبکه تصادفی به دلیل توزیع نسبتاً عادلانه یال‌ها ساختار انجمنی ندارد. لذا چگالی قسمت‌های مختلف آن اختلاف زیادی با یکدیگر ندارند. با این تعبیر، تفاوت بیشتر در مقایسه با ساختار یک گراف تصادفی به معنای رسیدن به انجمن‌یابی بهتر و قوی‌تر است. ضابطه تابع ماژولاریتی (Q) مطابق رابطه ۲-۱ تعریف می‌شود:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(c_i, c_j\right) \qquad (۱-۲)$$

که در آن، m تعداد یال‌ها و A_{ij}: تعداد یال‌های میان دو راس v_i و v_j در ماتریس مجاورت و k_i درجه راس v_i است. پارامتر c_i برچسب انجمنی است که راس v_i به آن تعلق دارد و $\delta(c_i,c_j) = 1$ است اگر c_i و c_j متعلق به یک انجمن باشند وگرنه برابر با صفر است.

---

[۱] Modularity Function

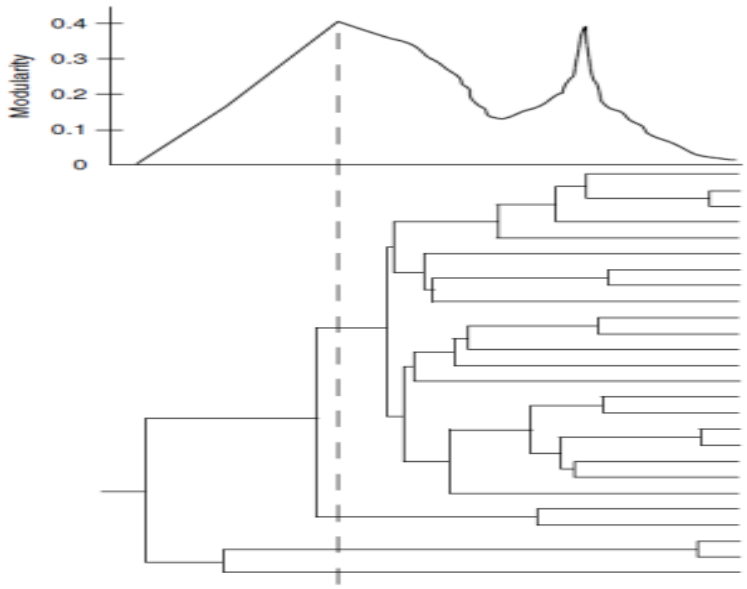
[۲] Clauset, Newman, and Moore

[۳] Null-model

کسر $k_ik_j/2m$ بیانگر احتمال وجود یال میان دو راس در گراف با توزیع تصادفی یکنواخت است که افزایش میزان تفاوت با آن، به افزایش تابع Q کمک می‌کند. حالت $Q < 0$ ممکن است در موارد خاصی نظیر نامتناسب[1] بودن شبکه اتفاق بیافتد، اما تابع Q معمولاً مقداری بین 0 و 1 دارد بدین معنا که روش فعلی به درجاتی از ساختار انجمنی دست یافته است. با افزایش Q، چگالی بیشتر و ساختار انجمنی قویتری خواهیم داشت. بنابراین، هدف ما یافتن حالتی از انجمن‌یابی است که Q را ماگزیموم کند.

شکل۲-۴ مثالی از عملکرد روش CNM را نشان می‌دهد. در شروع کار، هر گره معادل با یک انجمن است (سمت راست دندروگرام). اگر ادغام دو گره باعث افزایش در مقدار تابع ماژولاریتی شود، ادغام صورت می‌گیرد و انجمن جدید تشکیل می‌شود. این کار به تناوب برای مراحل بعدی تکرار خواهد شد. البته معمولاً ادغام در مراحل بعدی، مابین مولفه‌های مرکب (انجمن- انجمن یا گره- انجمن) انجام می‌شود.

مطابق شکل۲-۴، قطع دندروگرام با خط متناظر با بیشترین مقدار ماژولاریتی، بهترین انتخاب برای انجمن‌یابی را (شامل ۴ انجمن برای این مثال) مشخص می‌کند.



شکل ۲-۴: عملکرد تابع ماژولاریتی: قطع دندروگرام و مشاهده مقدار متناظر برای تابع Q [۲۲]

[1] Disassortative

بجز پیچیدگی زمانی نامطلوب، روش‌های مبتنی بر ماژولاریتی مشکل «حد تفکیک»[1] دارند یعنی در خروجی نهایی، انجمن‌های کوچکتر از یک اندازه مشخص را درنظر نمی‌گیرند. این مشکل (خصوصاً برای شبکه‌های بزرگ) از کیفیت روش می‌کاهد. اما مهمترین نکته در اینجاست که این روش پرکاربرد، از همپوشانی حمایت نمی‌کند. البته تلاش‌های بسیاری برای بهبود و تعمیم روش کلاسیک استفاده از ماژولاریتی، انجام‌شده است و روش‌های متعددی با رویکردهای اکتشافی یا حریصانه (برای کاهش فضای جستجو و بهبود عملکرد و کاهش زمان اجرا) پیشنهاد شده‌اند. تابع ماژولاریتی را می‌توان با تغییراتی برای گراف‌های وزن‌دار [۲۹] و گراف‌های جهت‌دار [۳۰] استفاده نمود.

روشی به نام Louvain که توسط Blondel و همکاران ارائه شده است [۳۱]، سریعتر از روش کلاسیک ماژولاریتی عمل می‌کند و زمان پردازش کمتری دارد. در ابتدای این روش، همه رئوس در انجمن‌های مجزا قرار داده می‌شوند (تعداد انجمن‌ها به اندازه تعداد رئوس).

الگوریتم این روش شامل دو گام کلی زیر است: در گام اول، برای گره دلخواه i، مقدار ماژولاریتی وزن‌دار برای حالتی که گره i به هرکدام از همسایگانش (مثلاً گره j) ملحق شود و یک انجمن جدید بسازد را محاسبه می‌کنیم. در صورت مشاهده افزایش، عمل الحاق به گره‌ای که بیشترین مقدار ماژولاریتی برای آن به دست می‌آید، انجام می‌شود تا انجمنی شامل دو گره داشته باشیم.

در گام دوم، گراف وزن‌داری ساخته می‌شود که گره‌های آن، انجمن‌های تولیدشده در گام اول بوده، وزن یال‌های آن نشان‌دهنده میزان قدرت و ارتباط مابین‌انجمنی در انجمن‌های یافته‌شده تا آن زمان است. وجود و میزان ارتباط درون‌انجمنی به وسیله حلقه (یال با مبدا و مقصد یکسان) بیان می‌شود. هر دو گام، به دفعات و تا زمان افزایش در مقدار ماژولاریتی انجام می‌شوند. خروجی روش، بهترین حالت ممکن در انجمن‌یابی (ماگزیموم حاصل برای مقدار ماژولاریتی) خواهد بود.

[1] Resolution Limit

یکی از مهمترین روش‌های پیشنهادی برای اینکه تابع ماژولاریتی از مفهوم همپوشانی حمایت کند، توسط chen و همکاران ارائه شده است[۳۲]. در این روش، یک متغیر احتمالاتی مطابق روابط زیر تعریف می‌شود:

$$a_{i,c} = \frac{\sum_{k \in c} A_{ik}}{\sum_{c' \in C_i} \sum_{k \in c'} A_{ik}} \qquad (۲-۲)$$

$$\forall i \in V, \forall c \in C : 0 \le a_{i,c} \le 1 , \sum_{c \in C} a_{i,c} = 1 \qquad (۳-۲)$$

روشن است که مقادیر ممکن برای این متغیر تصادفی در بازه [0,1] بوده و مجموع مقادیر آن نیز برابر با 1 خواهد بود. در ابتدای کار فرض بر این است که هر گره با میزان احتمال یکسان، به انجمن‌هایی که در آنها عضو است، تعلق داشته باشد.

با توضیحات ارائه‌شده، ضابطه تابع ماژولاریتی برای حالت همپوشان مطابق رابطه ۲-۴ تعریف می‌شود:

$$Q_{ov}^{C} = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} \left[A_{ij} - \frac{k_i k_j}{2m} \right] a_{i,c} a_{j,c} \qquad (۴-۲)$$

مقدار Q_{ov} با تعداد انجمن‌هایی که هر گره می‌تواند به آنها تعلق داشته باشد و سهم (قدرت) عضویت در هر انجمن، نسبت مستقیم دارد.

بقیه روش‌های مطرح‌شده، از توابع چگالی دیگری استفاده می‌کنند. در روشی که توسط Baumes و همکاران ارائه شده است [۳۳]، مفهوم انجمن را معادل با زیرگرافی تعریف کرده‌اند که به صورت محلی یک تابع دلخواه مرتبط با چگالی یال را ماگزیمم می‌کند. مجموعه زیرگراف‌هایی که این تابع تعریف‌شده را ماگزیمم نمایند ممکن است دارای دو یا چند راس مشترک باشند. بنابراین می‌توان مفهوم همپوشانی را پشتیبانی نمود. مطابق این تعبیر، انجمن معادل با مجموعه‌ای تعریف می‌شود که چگالی ارتباطات میان گره‌هایش از یک حدآستانه بیشتر است یعنی اگر یک گره به این مجموعه اضافه شده یا کم شود،

چگالی ارتباطات کاهش می‌یابد. این روش که به نام LA → IS نیز معروف است دارای دو فاز زیر می‌باشد:

در فاز اول یا LA[۱]، راس‌ها براساس مقدار Page Rank آنها بصورت نزولی مرتب می‌شوند. با استفاده الگوریتمRaRe[۲]، رئوس دارای بالاترین رتبه، به تناوب حذف می‌شوند تا فقط خوشه‌های کوچک و مستقلی باقی بمانند. تابع چگالی یال‌ها مطابق رابطه ۲-۵ تعریف می‌شود:

$$f(c) = \frac{w_{in}^{c}}{w_{in}^{c} + w_{out}^{c}} \qquad (۵-۲)$$

که بیانگر تاثیر مجموع وزنهای داخلی انجمنC و مجموع وزنهای خارجی (مرزی) انجمن C می‌باشد. در فاز دوم یا IS[۳] کار را از یک انجمن دلخواه (خروجی مرحله قبل) آغاز کرده، دیگر رئوس را یک‌به‌یک به آن اضافه (یا حذف) می‌کند. هر راس دلخواه بطور قطعی به این انجمن اضافه می‌شود (از آن حذف می‌شود) اگر باعث افزایش در مقدار تابع نظیر آن باشد.

این کار با انتخاب انجمن‌های دیگر به عنوان بذر و تکرار مراحل روش ادامه می‌یابد. خاتمه کار زمانی اعلام می‌شود که در مجموعه رئوس انتخاب‌شده انجمن‌ها (مقادیر متناظر برای تابع) تغییری ایجاد نشود.

چون در اجرای گام دوم این روش، اجازه حذف رئوس در زمان بسط انجمن داده می‌شود، این امکان وجود دارد که مولفه‌های ناهمبند ایجاد شوند. برای رفع مشکل ایجادشده و در تعمیم روش Baumes، روشی به نام CIS[۴] توسط kelley پیشنهاد شد [۳۴] که قبل از هر بار اجرای الگوریتم، همبندی انجمن‌ها را کنترل می‌کند. اگر یک انجمن به دو یا چند مولفه، تجزیه شده بود مولفه‌ای را حفظ می‌کند که بالاترین مقدار چگالی را دارد. تابع ارزیابی هم به فرم رابطه ۲-۶ بهبود یافته است:

$$f(c) = \frac{w_{in}^{c}}{w_{in}^{c} + w_{out}^{c}} + \lambda_{ep} \qquad (۶-۲)$$

---

[۱] Link Aggregate (LA)

[۲] Rank Removal (RaRe)

[۳] Iterative Scan (IS)

[۴] Connected Iterative Scan (CIS)

پارامتر λ رفتار الگوریتم را در نواحی کم‌تراکم شبکه با دخالت دادن «احتمال وجود یال»[۱] کنترل می‌کند. اضافه‌کردن هر راس به انجمن، نیاز به برقراری تعادل میان تغییرات چگالی درجات داخلی و تغییرات چگالی یال دارد.

روش LFM[۲] که توسط Lancichinetti و همکاران ارائه شده است [۳۵] مکانیزمی مشابه با روش‌های Baumes و Kelley دارد. تابع ارزیابی این روش مطابق رابطه ۲-۷ تعریف می‌شود:

$$f(c) = \frac{k_{in}^{c}}{(k_{in}^{c} + k_{out}^{c})^{\alpha}} \qquad (۷-۲)$$

بعد از یافتن یک انجمن، گره‌ای که هنوز به هیچ انجمنی تخصیص نیافته را بطور تصادفی انتخاب کرده و امکان افزودن آن به انجمن فعلی را با توجه به تابع f(c) بررسی می‌کنیم. متغیر α در تابع ارزیابی، بعنوان «پارامتر تفکیک»[۳] معرفی می‌شود تا اندازه انجمن‌های حاصل از این روش را کنترل کند.

رویکرد برخی روش‌ها به مسئله انجمن‌یابی، تلاش برای مدل‌کردن آن در قالب مسئله آماری (شبکه‌های باور بیزین[۴] ، مدل‌های استنتاجی، ....) است. روش‌های زیادی در این حیطه مطرح شده‌اند که از جمله آنها می‌توان به روش OSLOM[۵] توسط Lancichinetti و همکاران اشاره کرد [۳۶].

روش OSLOM انجمن را ساختاری تعریف می‌کند که در مقایسه با یک مدل مرجع، از لحاظ آماری معنادار[۶] باشد. منظور از مدل مرجع، گراف تصادفی متناظر با گراف شبکه فعلی است (مشابه تعریف تابع ماژولاریتی) و منظور از معناداری آماری، مقدار احتمال یافتن ساختار مشابه با انجمن مورد بررسی (در تعداد رئوس، دنباله درجات، اتصالات داخلی انجمن) در مدل مرجع می‌باشد.

---

۱ Edge Probability

۲ Lancichinetti and Fortunato Method (LFM)

۳ Resolution Paremeter

۴ Bayesian Belief Networks (BBN)

۵ Order Statistics Local Optimization Method (OSLOM)

۶ Significant

فرایند کار شامل شروع از یک گره و بررسی امکان بسط دادن آن بصورت حریصانه با کنترل میزان معناداری است. پارامتر r برای هر گره، احتمال تجمعی دارابودن تعداد مساوی یا بیشتر از یال‌های داخلی در مقایسه با تعداد اتصالات از یک همسایه در همین انجمن در مدل مرجع را نشان می‌دهد. ابتدا مقادیر r برای تمام همسایگان یک گره محاسبه می‌شوند و به ترتیب صعودی مرتب خواهندشد. اگر کوچکترین مقدار r از یک حدآستانه کمتر باشد، یعنی معنادار است و گره مربوط به آن مقدار، به گره (انجمن) فعلی اضافه می‌شود. اگر این شرط برقرار نبود به سراغ مقدار بعدی r خواهیم رفت. با توجه به ماهیت آماری کار، کل این فرایند چند بار تکرار می‌شود تا حالت پایدار حاصل شود. روش OSLOM منعطف و کارآمد است. این روش قابلیت کار با گراف‌های وزن‌دار و جهت‌دار را داراست و انجمن‌های همپوشان و سلسله‌مراتبی را بخوبی ردیابی می‌کند.

روش همپوشان FOCS[۱] جستجو در شبکه‌های بزرگ را براساس امتیازات معیارهایی که بصورت محلی محاسبه می‌شوند، انجام می‌دهد[۳۷]. این روش دو معیار مهم را معرفی می‌کند که خواص انجمن‌های دنیای واقعی را بازتاب می‌دهند.

معیار «همبندی انجمنی»[۲] برای هر گره، بیانگر میزان تعلق آن گره به انجمنی است که در آن عضو می‌باشد و معادل با نسبت تعداد همسایگان آن گره که عضو انجمن هستند به تمام گره‌های انجمن، تعریف می‌شود. گره‌های با بیشترین میزان تعلق، قطعاً در انجمن فعلی می‌مانند.

معیار «همبندی همسایگی»[۳] برای هر گره، بیانگر میزان دلبستگی یک گره به عضویت در یک انجمن دیگر (با حفظ عضویت در انجمن فعلی) است و معادل با نسبت تعداد همسایگان آن گره که عضو انجمن هستند به تمام همسایگان آن گره، تعریف می‌شود. گره‌های با بیشترین میزان دلبستگی برای پیوستن به یک انجمن دیگر، حتماً جزو گره‌های همپوشان هستند.

---

[۱] Fast Overlapped Community Search (FOCS)

[۲] Community Connectedness

[۳] Neighborhood Connectedness

برای انجمن‌یابی، می‌توان از الگوریتم‌های تکاملی ایده گرفت که مبتنی بر شبیه‌سازی فرایندهای بیولوژیکی انجام‌شده توسط یک توده یا کولونی از موجودات زنده برای ادامه بقا (اصل بقای اصلح[۱]) هستند که در طول سال‌ها به تکامل رسیده‌اند. این روش‌های بهینه‌سازی دارای رویکرد اکتشافی[۲] یا اتفاقی[۳] هستند. برای همه مسائل مورد بررسی جوابی قابل‌قبول (نه الزاماً بهنیه) را پیدا می‌کنند. هرچند خیلی دقیق نیستند و در هر بار اجرا جواب متفاوتی می‌دهند و معمولاً باید میانگین نتایج تعداد دفعات زیادی از اجرای الگوریتم را در نظر گرفت، اما وقتی فضای جستجوی مسئله خیلی وسیع است، انتخاب مناسبی برای حل مسئله به شمار می‌روند.

با بهینه‌کردن تابع برازش که روی پارامترهای توصیفگر مشخصات توپولوژیکی محلی اعضای یک شبکه، تعریف شده‌اند می‌توان مسئله انجمن‌یابی شبکه‌های اجتماعی را با یک مسئله بهینه‌سازی محلی مدل کرد.

در الگوریتم ژنتیک[۴] که پراستفاده‌ترین روش تکاملی محسوب می‌شود، با شروع از یک جمعیت اولیه، مقدار تعیین‌شده توسط یک تابع برازش برای هرکدام از اعضای جمعیت اولیه را می‌یابیم. ایجاد تغییرهای پیش‌بینی‌شده روی عناصر مناسب در نسل فعلی برای رسیدن به نسل بعدی، شامل بازتولید با انتخاب والدهای مناسب از نسل فعلی و عملگرهای برش[۵] و جهش[۶] خواهد بود. نهایتاً تکرار این عملیات به مقدار لازم (رسیدن به حد قابل‌قبولی از برازش یا عدم وقوع تغییر بعد از چند تکرار) انجام می‌شود.

برای مدل‌کردن مسئله انجمن‌یابی، هر گره شبکه معادل با یک عنصر از جمعیت اولیه در نظر گرفته می‌شود و تابع برازش در نقش تابع کیفیت عمل می‌کند. روشی به نام GA-Net[۷] [۳۸] و نیز روش

---

[۱] Survival of the Fittest

[۲] Heuristic

[۳] Stochastic

[۴] Genetic Algorithm

[۵] Crossover

[۶] Mutation

[۷] Genetic Algorithm for Networks (GA-Net)

تعمیم‌یافته آن به نام MOGA-Net[1] [۳۹] از ویژگی‌های روش‌های تکاملی برای انجمن‌یابی استفاده می‌کنند.

تلاش‌هایی در تلفیق مفاهیم «تئوری بازی»[2] با روش‌های تئوری گراف انجام شده است. مدل‌کردن انجمن‌یابی همپوشان در شبکه‌های اجتماعی با یک بازی استراتژیک، معادل با این است که هر گره را یک عامل حریص بدانیم که ملحق شدن به یک انجمن یا خارج شدن از آن را با میزان سودمندی این عمل می‌سنجد. سودمندی معادل با بیشینه‌سازی مقدار یک تابع بهره[3] (سود) در نظر گرفته می‌شود

در روش پیشنهادی [۴۰] با عنوان Game، متناظر با هر گره یک تابع بهره و یک تابع اتلاف[4]، تعریف شده است. هر گره (عامل) برای رسیدن به نقطه تعادل میان مقدار بیشینه انتخاب‌شده برای تابع بهره و مقدار کمینه انتخاب‌شده برای تابع اتلاف، می‌تواند در دو یا چند انجمن عضو باشد. در این روش، تابع کیفیت (چگالی) همان تابع بهره یا تابع اتلاف است.

۲-۶ انجمن‌یابی مبتنی بر مکانیزم انتشار

انتشار اطلاعات و اخبار میان افرادی که عضو یک گروه هستند، بسیار سریعتر از انتشار اخبار میان افرادی که گروه مشترکی ندارند انجام می‌شود. با استفاده از این ایده ساده، می‌توان انجمن‌ها را با ردیابی نحوه انتشار اطلاعات در شبکه یافت. بدین معنا که یک انجمن را شامل رئوسی دانست که از انتشار اطلاعات یا خاصیت تعریف‌شده در شبکه، تاثیر یکسان می‌گیرند و رفتار مشابهی بروز می‌دهند. کاربردهای مهمی

[1] Multi-Objective GA-Net (MOGA-Net)
[2] Game Theory
[3] Gain (Utility) Function
[4] Loss Function

نظیر: یافتن افراد و «گره‌های تاثیرگذار»[۱] در انتشار اطلاعات خاص، «پخش اطلاعات»[۲]، «ماگزیمم‌سازی جریان»[۳]، ... مرتبط با عملکرد این گروه از روش‌ها هستند.

کاربردهای فراوانی برای این گروه از روش‌ها وجود دارد که می‌توان به مواردی چون: بررسی اثر «دهان‌به‌دهان»[۴] در انتشار اطلاعات و رفتار، نیت‌کاوی[۵] (پردازش اطلاعات مربوط به نظرات افراد)، رفتار و میزان برون‌گرایی اعضای شبکه اجتماعی، شناسایی افراد خبره و «گره‌های تأثیرگذار»[۶] در انتشار اطلاعات خاص، انواع بازاریابی خصوصاً «بازاریابی ویروسی»[۷] اشاره کرد.

معمولاً نمی‌توان برای گراف متراکمی که انجمن‌های آن به سادگی قابل تشخیص نیستند از روش‌های ساده (مثلا حذف پل‌های میان انجمن‌ها) استفاده کرد. در اینگونه موارد، روش‌های مبتنی بر انتشار عملکرد بسیار خوبی دارند.

یک روش پرکاربرد در گروه روش‌های مبتنی بر انتشار، «الگوریتم انتشار برچسب»[۸] نام دارد که توسط Raghavan و همکاران ارائه شد[۴۱]. این روش سریع و ساده، براساس نام طراحانش به روش RAK[۹] نیز مشهور است. ایده این روش مبتنی بر توزیع اولیه تعدادی برچسب میان گره‌های شبکه است. در ادامه، هر گره برچسبی را اخذ می‌کند که بیشترین تکرار میان همسایگانش را دارد. قسمت‌های چگال شبکه تا حد امکان بسط می‌یابند و به برچسب‌های مشترک می‌رسند. در پایان کار، هر انجمن شامل تعدادی گره با برچسب یکسان (متفاوت با برچسب دیگر انجمن‌ها) خواهد بود.

روش انتشار برچسب نیازی به پارامتر ورودی ندارد. وضعیت گراف شبکه به صورت ماتریس مجاورت یا لیست مجاورت به الگوریتم داده می‌شود. مراحل این روش در زیر بیان شده است:

---

[۱] Infulential Nodes
[۲] Information Spread
[۳] Flow Maximization
[۴] Word of Mouth
[۵] Opinion Mining
[۶] Influential Nodes
[۷] Viral Marketing
[۸] Label Propagation Algorithm (LPA)
[۹] RAK (Raghavan, Albert, and Kumara)

۱. قرار دادن t=1 و اختصاص اولیه یک برچسب منحصربفرد به هر گره گراف:

قراردادن $C_x(0) = x$ که $C_x(t)$ بیانگر برچسب گره x در زمان t است. همسایه‌های گره x با $x_1, x_2, \ldots, x_k$ نشان داده می‌شوند. تعداد برچسبها در ابتدا برابر با تعداد گره‌ها خواهد بود اما در پایان کار، به تعداد واقعی انجمن‌ها کاهش می‌یابد.

۲. مرتب‌کردن گره‌ها با یک آرایش (ترتیب) تصادفی و قرار دادن آن در X

۳. برای هر $x \in X$ که از ترتیب بالا انتخاب می‌شود: $C_x(t)$ را مطابق تابع انتشار update کن (تابع انتشار، برچسب بالاترین فرکانس تکرار بین همسایگان یک گره را مشخص می‌کند.)

۴. اگر برچسب هر نود با پرتکرارترین برچسب میان همسایگان یکسان بود: اعلام خاتمه کار

در غیر این صورت: $t = t+1$ و تکرار الگوریتم از مرحله ۲ انجام شود.

تغییر[۱] برچسب یک گره در زمان t به دو طریق امکان‌پذیر است. در *تغییر همزمان*، فقط به برچسب تمام همسایگان در زمان قبلی توجه می‌شود:

$$C_x(t) = f\left(C_{x_1}(t-1), C_{x_2}(t-1), \ldots, C_{x_k}(t-1)\right) \qquad (۸-۲)$$

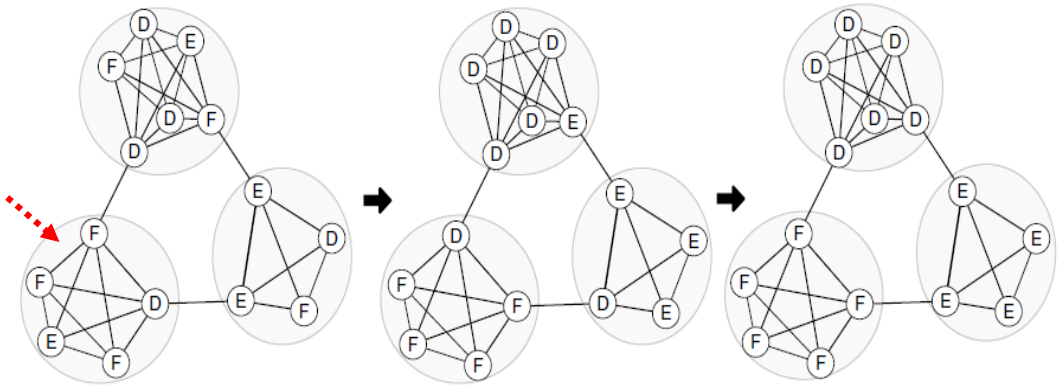
در حالت *تغییر ناهمزمان*، فرض بر این است که برچسب m گره از همسایگان، مربوط به زمان فعلی (انجام آخرین تغییرات برچسب آنها) و برچسب k-m گره مربوط به زمان قبلی باشد:

$$C_x(t) = f\left(C_{x_1}(t), C_{x_m}(t), C_{x_{m+1}}(t-1), \ldots, C_{x_k}(t-1)\right) \qquad (۹-۲)$$

معمولاً بیشتر از روش تغییر (به‌هنگام‌سازی) همزمان استفاده می‌شود. یکی از چالشهای مهم در انتشار برچسب، حالتی است که دو یا چند گره از همسایگان دارای برچسب با بیشترین فرکانس تکرار باشند. ساده‌ترین برخورد با این مشکل، مکانیزم tie-breaking است بدین معنا که برچسب یکی از این همسایگان بطور تصادفی انتخاب شود و مراحل الگوریتم ادامه یابد.

[۱] Update

شکل۲-۵ اجرای الگوریتم انتشار برچسب را نشان می‌دهد. راس مشخص‌شده در این شکل، در نهایت به برچسب F خواهد رسید، حتی اگر در مراحل میانی برچسب دیگری برای آن انتخاب شود.



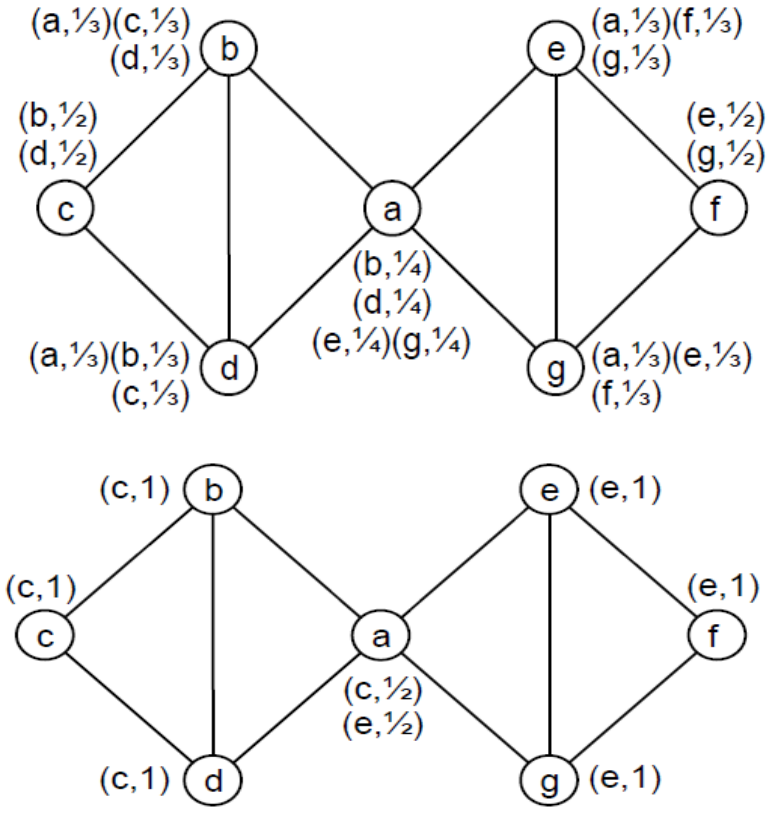
شکل ۲-۵: مثالی از انتشار برچسب در LPA [۱۶]

امکان اخذ برچسبهای متفاوت درحین اجرای روش، باعث می‌شود تا الگوریتم انتشار برچسب یک روش غیرقطعی باشد. این روش خروجی یکتا ندارد و ممکن است به جوابهای متعدد از شرایط اولیه یکسان برسیم. اجرای چندباره الگوریتم (مشابه روش‌های تکاملی) یکی از راهکارهایی است که برای این مورد توصیه می‌شود.

مزیت مهم این روش در پیچیدگی زمانی آن است. برچسب‌دهی اولیه به گره‌ها با O(n) انجام می‌شود و تکرار مراحل الگوریتم از مرتبه O(m) است. در نهایت، این روش پیچیدگی زمانی بسیار مطلوب و تقریباًخطی O(m + n) را خواهد داشت تا به عنوان یکی از مناسب‌ترین روش‌ها برای کار با گراف‌های خلوت و کم‌تراکم مطرح شود.

این روش مهم و پرکاربرد انجمن‌یابی از همپوشانی حمایت نمی‌کند زیرا هر گره می‌تواند فقط یک برچسب داشته باشد و به یک انجمن تعلق یابد. تعمیم‌های متعددی روی روش کلاسیک LPA پیشنهادشده تا انجمن‌یابی همپوشان را نیز شامل شود.

یکی از این تعمیمها روش COPRA[۱] است که توسط Gregory مطرح شد[۴۲]. در این روش یک گره می‌تواند دو یا چند برچسب را به‌طور همزمان اخذ نماید. الگوریتم این روش مشابه روش کلاسیک است یعنی در آغاز، به هر گره یک برچسب منحصربه‌فرد اختصاص می‌یابد و در ادامه کار، هر گره برچسب با بیشترین تکرار میان همسایگانش را اخذ می‌کند. برای تحقق همپوشانی، هر گره با یک یا چند زوج مرتب (c,b) توصیف می‌شود که در آن، پارامترc بیانگر شناسه یک انجمن خاص بوده و پارامتر b ضریب تعلق[۲] به آن انجمن را (بین0 و 1) بیان می‌کند. نیمه بالایی شکل۲-۶، این تخصیص را برای یک مثال خاص نشان می‌دهد.



شکل ۲-۶: مثالی از تخصیص‌های اولیه و خروجی نهایی در روش COPRA [۴۲]

[۱] Community Overlap PRopagation Algorithm (COPRA)

[۲] Belonging Coefficient

در هر مرحله اجرا، ضرایب تعلق هر گره براساس میانگین ضرایب همسایگانش تغییر می‌کند. اجرای روش تا رسیدن به همگرایی و عدم تغییر برچسبها ادامه می‌یابد. خروجی این روش نیز به دلیل استفاده از مکانیزم انتخاب تصادفی، غیرقطعی بوده و نیاز به تکرار در دفعات اجرا دارد.

پیچیدگی زمانی روش برابر با $O(vm \log(vm/n))$ است. پارامتر v عدد کوچکی بوده و بیانگر حداکثر تعداد انجمن‌هایی است که یک گره می‌تواند عضو آنها باشد. خروجی مثال مطرح‌شده (نیمه پایینی شکل۲-۶) مقدار v را برابر با ۲ نشان می‌دهد.

دیگر روش همپوشان در انتشار برچسب، روش SLPA[۱] است که توسط Xie و همکاران [۴۳] با ایده‌گرفتن از نحوه مکالمه افراد با یکدیگر (شنیدن و تصمیم‌گیری براساس سابقه ذهنی) ارائه شده است. در این روش، هر گره حافظه‌ای دارد که سابقه برچسب‌ها را نگهداری می‌کند. این روش همچنین دارای دو پارامتر ورودی می‌باشد که باید مقداردهی اولیه شوند: T بیانگر حداکثر تعداد مراحل اجرا و r نشان‌دهنده حدآستانه مرحله نهایی (پس‌پردازش[۲]) است. مراحل الگوریتم به شرح زیر خواهد بود:

۱. مقداردهی اولیه حافظه هر گره با یک برچسب منحصربه‌فرد
۲. تکرار این مرحله تا رسیدن به شرط توقف (مقدار T):
   a. انتخاب یک گره بعنوان شنونده
   b. هر یک از همسایگان گره شنونده، بصورت تصادفی یک برچسب با احتمال متناسب با فرکانس وقوع آن در حافظه خودش را برای گره شنونده ارسال می‌کند.
   c. گره شنونده، پرتکرارترین برچسب دریافت‌شده را به حافظه خود اضافه می‌کند.
۳. مرحله پس‌پردازش: تولید انجمن‌ها در خروجی با اعمال حدآستانه r

در مرحله پس‌پردازش، محتویات حافظه هر گره به یک توزیع احتمالاتی از برچسب‌ها تبدیل می‌شود. در ادامه، برچسب‌هایی که میزان تعلق آنها کمتر از حدآستانه r است را حذف می‌کنیم. عدد r از بازه $[0, 0.5]$ انتخاب می‌شود. اکنون گره‌های همبندی که دارای برچسب‌های یکسان باشند را به عنوان یک انجمن تلقی می‌کنیم. اگر گره‌ای بیش از یک برچسب داشته باشد، یک گره همپوشان خواهد بود. روشن

[۱] Speaker- Listener Propagation Algorithm (SLPA)

[۲] Post-Processing

است که مقادیر $r > 0.5$ انجمن‌های ناهمپوشان تولید خواهد کرد. روش SLPA پیچیدگی زمانی مطلوبی برابر با $O(Tm)$ دارد که نسبت به تعداد یال‌ها خطی است[۴۴].

روشی به نام DEMON[۱] که توسط Coscia ارائه شده است[۴۵] همپوشانی را در انجمن‌یابی شبکه‌های مستقل از مقیاس، مد نظر قرار می‌دهد. دید کلی از شبکه‌های بسیار بزرگ اطلاعات خاصی را به بیننده منتقل نمی‌کند بنابراین رویکردهای محلی‌تری برای بررسی این شبکه‌ها مورد نیاز است. ایده کلی این روش به این شرح است: هر گره دید[۲] محدودی از کل شبکه دارد که شامل انجمن‌های عضو در آنها و نیز انجمن‌های مجاور (محلی) می‌باشد. با ادغام این نقطه‌نظرها در یک رویکرد دموکراتیک می‌توان به ساختارهای همپوشان رسید یعنی عملاً خود گره‌ها در مورد ساختارهای انجمنی موجود قضاوت می‌کنند.

گروهی از روش‌ها، انتشار برچسب را به یک مسئله استنتاج آماری مدل می‌کنند. تابع تعلق این روش‌ها بجای یک عدد ساده، مشتمل بر برداری از احتمالات است. $P(X,Y)$ احتمال وجود یال میان گرهی از انجمن X با گرهی از انجمن Y را بیان می‌کند و به کمک میانگین وزنی $P(X,X)$ و $P(Y,Y)$ بدست می‌آید.

از فرآیندهای اتفاقی[۳] و اجازه تعلق مخلوط[۴] در روش MMSB[۵] [۴۶] استفاده شده است. در شروع کار، هر گره با یک احتمال مشخص به هر انجمن تعلق دارد و در ادامه، بردار احتمال متعلق به هرگره از بردارهای سایر گره‌ها تاثیر می‌پذیرد (لفظ مخلوط به همین منظور آمده است). این روش همپوشان، عمدتاً برای یافتن انجمن‌های خیلی بزرگ بکار می‌رود. چون احتمال وقوع یال بین دو گره که در چند انجمن با یکدیگر اشتراک دارند (با توجه به نحوه عملکرد میانگین‌گیری وزن‌دار) نمی‌تواند از احتمال وقوع یال بین دو گره واقع در یک انجمن بیشتر باشد، این روش نمی‌تواند نواحی پرتراکم همپوشان

---

[۱] Democratic Estimate of the Modular Organization of a Network (DEMON)

[۲] Perspective

[۳] Stochastic

[۴] Mixed Membership

[۵] Mixed Membership Stochastic Blockmodel (MMSB)

(نواحی که گره‌های میان دو یا چند انجمن، تعداد زیادی یال مابین خود دارند) را به‌خوبی تشخیص دهد.

در روش MOSES[۱] که توسط McDaid و همکاران ارائه شده است[۴۷]، هر گره با یک بردار پنهان[۲] که شامل k متغیر بولی مستقل است، مدل می‌شود. بسط انجمن‌ها در این روش، با اضافه کردن یال‌ها از طریق یک تابع بهینه‌ساز محلی بصورت حریصانه انجام می‌شود.

تلاشهای متعددی در تلفیق مفاهیم «تئوری بازی»[۳] (بررسی *رفتار* شبکه) با روش‌های تئوری گراف برای بهبود کیفیت انجمن‌یابی صورت گرفته است[۴۸, ۴۹]. مدل‌کردن انجمن‌یابی همپوشان در شبکه‌های اجتماعی با یک بازی استراتژیک، معادل با این است که هر گره را یک عامل حریص بدانیم که ملحق شدن به یک انجمن یا خارج شدن از آن انجمن را با میزان سودمندی[۴] این عمل می‌سنجد. سودمندی معمولاً معادل با ماگزیمم‌کردن یک تابع بهره[۵] یا مینیمم‌کردن یک تابع اتلاف[۶] قلمداد می‌شود. این روش‌ها کاربرد زیادی در بررسی تاثیر تبلیغات و بازاریابی محصولات در شبکه‌های اجتماعی دارند [۵۰].

Chen و همکاران چارچوبی مبتنی بر تئوری بازیها برای انجمن‌یابی پیشنهاد داده‌اند[۴۰]. در این روش، متناظر با هر گره یک تابع بهره و یک تابع اتلاف تعریف شده است. هرگره (عامل) برای رسیدن به نقطه تعادل میان مقدار ماگزیمم انتخاب‌شده برای تابع بهره و مقدار مینیمم انتخاب‌شده برای تابع اتلاف، می‌تواند در بیش از یک انجمن نیز عضو باشد.

---

[۱] Model-based Overlapping Seed ExpanSion (MOSES)

[۲] Latent Vector

[۳] Game Theory

[۴] Utility

[۵] Gain Function

[۶] Loss Function

در یک روش جدید به نام NISE[۱] از ایده مؤثر تورم[۲] همسایه‌ها استفاده می‌شود[۵۱]. در این روش انتشار، تأکید فقط روی یک گره انتخاب‌شده به‌عنوان نقطه شروع (بذر) نیست بلکه همسایگان گره نیز در فرایند انتشار دخیل هستند و یک ناحیه بذر را تشکیل می‌دهند.

۲-۷ انجمن‌یابی مبتنی بر ساختار خاص

گروهی از روش‌ها، انجمن‌یابی را با بررسی ساختار گراف شبکه انجام می‌دهند. انجمن در این روش‌ها، یک ساختار دقیق و تعریف‌شده از یال‌ها است که از ترکیب زیرگراف‌های کوچکتر شبکه با یکدیگر حاصل می‌شود. به بیان دیگر، یک انجمن مجموعه‌ای از گره‌ها با تعداد مشخصی یال مابین آنهاست که از یک توپولوژی واضح و مشخص مطابق مفاهیم ریاضی گراف شبکه، نتیجه می‌شود.

اساس کار این روش‌ها بر مفهوم ساده‌ای به نام کلیک استوار است. منظور از عبارت k-clique در گراف یک شبکه، یک زیرگراف کامل k راسی است. یک k-clique، با فرض ساده و همبندبودن، دقیقاً دارای k*(k-1)/2 یال خواهد بود. با فرض k=1 یک k-clique، بیانگر خود رئوس است و برای k=2 نشان‌دهنده یال‌ها بوده و اطلاعات مهم و خاصی را منتقل نمی‌کند. انتخاب k=3 کوچکترین کلیک غیربدیهی (مثلث) را بیان می‌کند. انتخاب دیگر مقادیر برای k، تلاش برای یافتن کلیک‌های بزرگتر و ساختار متراکم‌تر را باعث می‌شود. مجازبودن تعلق همزمان راس‌ها به دو یا چند کلیک متفاوت، امکان مدل‌کردن همپوشانی را فراهم می‌کند. ایده اصلی روش‌های همپوشانی مبتنی بر مفهوم کلیک، از Percolation theory در ریاضیات و فیزیک نشأت می‌گیرد که به مطالعه رفتار خوشه‌های همبند یک گراف تصادفی می‌پردازد.

---

[۱] Neighborhood-Inflated Seed Expansion (NISE)

[۲] Inflation

روش ساده CPM[1] که توسط Palla و همکاران ارائه شده، انجمن‌های همپوشان یک گراف را به کمک کلیک‌ها بدست می‌آورد[۵۲]. این روش از مفاهیم روی کلیک‌ها (برای مقدار دلخواه k) نظیر همسایگی، اجتماع، زنجیره، ارتباط، انجمن، و کلیک‌ماگزیموم‌ها سود می‌برد. این مفاهیم اولیه و دیگر تعاریف متداول در جدول۲-۲ بیان شده‌اند. ایده کلی این است که کلیک‌ها را شناسایی کنیم (مقدار اولیه برای k) و مطابق با تعریف مجاورت دو کلیک، کلیک‌های همسایه را یافته و از اجتماع k-clique های مجاور یکدیگر، بزرگترین مولفه ممکن یعنی یک انجمن را تولید کنیم.

در گام اول الگوریتم CPM، تمام ماگزیموم‌کلیکهای گراف (کلیکهایی که خود زیرگراف هیچ کلیک دیگری نیستند) تعیین می‌شوند. لذا هیچ راسی با درجه کمتر از k-1 در نظر گرفته نمی‌شود. یافتن ماگزیموم‌کلیکهای یک گراف، یک مسئله رام‌نشدنی از نوع NP-Complete است و پیچیدگی زمانی از مرتبه نمایی (در بدترین حالت) دارد. در گام دوم، یک گراف کلیک متناظر (ماتریس clique-clique) ساخته می‌شود. این ماتریس متقارن، ارتباط میان ماگزیموم‌کلیکهای گراف را نشان می‌دهد. هر مولفه روی قطر اصلی ماتریس، بیانگر تعداد یال‌های یک ماگزیموم‌کلیک بوده و دیگر مولفه‌های ماتریس، یال‌های مشترک بین دو ماگزیموم‌کلیک را بیان می‌کنند. در گام سوم، برای یک مقدار تعیین‌شده از k، مولفه‌های کوچکتر از این مقدار در ماتریسclique-clique را صفر کرده و بقیه مقادیر را برابر با یک قرار می‌دهیم. هریک از مولفه‌های همبند ماتریس فعلی، یک انجمن در خروجی نهایی مسئله خواهد بود.

این روش اگرچه نوآوری زیادی داشت و از مفهوم همپوشانی حمایت می‌کرد، اما از معایبی هم برخوردار بود مانند اینکه فقط برای گراف‌های غیرجهت‌دار و بدون وزن کار می‌کرد.

در [۵۴] روش CPM برای گراف‌های وزن‌دار[2] ارائه شده است که یک k-clique وزن‌دار را به این صورت تعریف می‌کند: اولاً زیرگراف کاملی از k راس باشد (مشابه تعریف k-clique در حالت کلاسیک)

[1] Clique Percolation Method (CPM)

[2] Weighted Clique Percolation Method (CPMw)

و ثانیاً میانگین هندسی وزن تمام k*(k-1)/2 یال آن از یک عدد حدآستانه بیشتر باشد. روشن است که در این حالت بار محاسباتی بیشتری نسبت به روش CPM ایجاد می‌شود.

جدول ۲-۲: مفاهیم و پارامترهای مرتبط با کلیک [۵۲, ۵۳]

پارامتر	توضیح
k-clique neighbor	دو k-clique مجاور یکدیگر هستند اگر در k-1 راس با یکدیگر مشترک باشند (فقط در یک راس تفاوت داشته باشند).
k-clique chain	اجتماع k-clique های مجاور یک زنجیره k-clique را تشکیل می‌دهد.
k-clique connectivity	دو k-clique را مرتبط می‌نامیم اگر هر دو متعلق به یک زنجیره k-clique باشند.
maximal clique	بزرگترین کلیک ممکن در یک گراف است. یعنی اگر هر راس مجاور یک کلیک (و غیر متعلق به آن کلیک) به آن اضافه شود، کلیک جدیدی (با k بزرگتر) حاصل نشود.
k-clique community	اجتماع ماگزیمال تمام k-cliqueهای مجاور را یک انجمن k-clique می‌نامند.
Clique number	عدد کلیک گراف، عدد k متعلق به بزرگترین کلیک ممکن در یک گراف است.
k-core	بزرگترین زیرگرافی که در آن هر گره با حداقل k گره متعلق به زیرگراف، همسایه است. (کوچکترین درجه هر گره در زیرگراف، برابر با k است)
k-plex	بزرگترین زیرگرافی از گراف اصلی که هر راس آن با تمام رئوس زیرگراف، بجز حداکثر k راس، مجاور است.
k-clan	زیرگرافی که طول کوتاهترین مسیر بین هر دو گره دلخواه آن (قطر گراف)، حداکثر برابر با k باشد.
k-club	زیرگرافی که طول کوتاهترین مسیر بین هر دو گره دلخواه آن (قطر گراف)، حداکثر برابر با k باشد و ضمنا بزرگترین زیرگراف ممکن با قطر k باشد.
TPR[1]	نرخ (تعداد) گره‌هایی از گراف که مثلث می‌سازند. (در ساخت ساده‌ترین کلیک غیربدیهی k=3 مشارکت دارند)

[1] Triangle Participation Ratio (TPR)

پیاده‌سازی روش CPM برای گراف‌های جهت‌دار[۱] در [۵۵] بررسی شده است که در آن، یک انجمن با اجتماع تمام k-clique های جهت‌دار و مجاور یکدیگر مشخص می‌شود. یک k-clique جهت‌دار، زیرگراف کاملی شامل k راس است که در آن، جهت تمام یال‌ها از سمت راس با اولویت بیشتر به سمت راس با اولویت کمتر می‌باشد. اولویت‌بندی[۲] برای هر راس با اختصاص یک عدد به آن راس صورت می‌گیرد که عبارت از تعداد یال‌های خروجی از آن راس به دیگر یال‌های کلیک تقسیم بر مجموع درجات خروجی راسهای کلیک جهت‌دار است. کلیک برای گراف‌های دوبخشی[۳] (دو نوع متمایز از رئوس) نیز قابل تعریف است [۵۶] که با تعمیم کلیک به مفهوم جدیدی به نام bi-clique امکان‌پذیر است.

برای بهبود پیچیدگی زمانی و افزایش سرعت روش کلاسیک، Kumpula و همکاران روشی ترتیبی را ارائه کردند که به SCP[۴] معروف شده است [۵۷]. در این روش از گراف دوبخشی (دو نوع گره) برای نمایش کلیک‌ها استفاده می‌شود. فاز اول این روش به دنبال یافتن k-cliques است. این مهم با کنترل همه k-2 cliques در همسایگان مشترک دو نقطه پایانی وقتی که یال‌ها به صورت ترتیبی به گراف شبکه اضافه می‌شوند، انجام خواهدشد. فاز دوم به ردیابی k-communities اختصاص دارد. این مهم نیز با یافتن مولفه‌های همبند در تصویر[۵] (k-1)-cliques انجام می‌شود. جستجوی (k-1)-cliques جدیداً ایجادشده، در یک محدوده کوچک (شامل گرهی که یال جدید به آن اضافه شده و دیگر همسایگان آن گره) صورت می‌گیرد. در نهایت، گراف دوبخشی به دو گراف ساده (دارای گره‌های یکسان) تصویر[۶] می‌شود. چون هر k-clique دقیقاً دوبار پردازش می‌شود، پیچیدگی زمانی این روش بصورت یک تابع خطی از تعداد کلیکهای گراف خواهد بود و صرفه‌جویی زمانی خوبی نسبت به روش CPM خواهد داشت.

[۱] Directed Clique Percolation Method (CPMd)
[۲] Ranking
[۳] Bipartitie
[۴] Sequential Clique Percolation (SCP)
[۵] Projection
[۶] Projection

در روش EAGLE[۱] توسط Shen و همکاران، استفاده از ایده روش‌های انتشار (در بخش۲-۴) و بهینه‌سازی محلی مدنظر قرار گرفته است[۵۸]. با این تفاوت که بجای استفاده از گره‌ها بعنوان نقاط شروع و بذرهای اولیه، از کلیک‌ها برای این کار استفاده می‌شود. در نواحی چگال یک شبکه، به احتمال زیاد می‌توان سراغی از یک کلیک بزرگ (به عنوان بذر یا هسته یک انجمن) گرفت.

این روش ابتدا تمام کلیک‌ماگزیمومهای موجود در گراف را می‌یابد و برای هر زوج دلخواه از این کلیک‌ماگزیممها، به بررسی میزان شباهت می‌پردازد. این بررسی به کمک یک تابع ماژولاریتی تعمیم‌یافته انجام می‌شود که همپوشانی را حمایت می‌کند. اگر ماژولاریتی از حدآستانه بیشتر بود آن دو کلیک‌ماگزیموم را ادغام می‌کند و مراحل را تا رسیدن به یک انجمن (ساختن دندوگرام) ادامه می‌دهد.

با قطع دندوگرام در ماگزیمم مقدار ماژولاریتی، انجمن‌یابی با کیفیت خواسته‌شده انجام می‌شود. تابع ماژولاریتی استفاده‌شده در این روش مطابق رابطه ۲-۱۰ تعریف می‌شود که در آن، O_i تعداد انجمن‌هایی است که گره i در آنها عضو است.

$$Q_{ov}^{E} = \frac{1}{2m}\sum_{c}\sum_{i,j\in c}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\frac{1}{O_i O_j} \qquad (۱۰-۲)$$

Lee و همکاران روش GCE[۲] را مطرح کرده‌اند که مشابه روش EAGLE از کلیک‌ماگزیموم‌ها به عنوان بذر استفاده می‌کند[۵۹]. بسط بذرها در این روش همپوشان، با بهینه‌سازی حریصانه یک تابع برازش محلی انجام می‌شود. در این روش، انجمن‌های شبیه به انجمن‌های قبلاًکشف‌شده حذف می‌شوند و اگر فاصله بین دو انجمن از یک حدآستانه کمتر باشد، آن دو را مشابه درنظر می‌گیریم. تابع فاصله دو انجمن با رابطه ۲-۱۱ تعریف می‌شود که در آن، $|c_1|$ تعداد گره‌های انجمن c_1 است.

---

[۱] hierarchicAl clusterinG based on maximaL cliquE (EAGLE) agglomerativE

[۲] Greedy Clique Expansion (GCE)

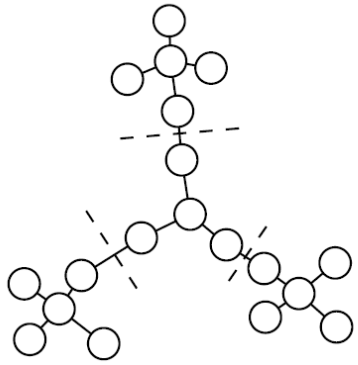
$$1-\frac{\left|c_1 \cap c_2\right|}{\min(\left|c_1\right|,\left|c_2\right|)} \qquad (۱۱-۲)$$

پیچیدگی زمانی این روش، متناسب با تعداد کلیک‌ها و یال‌ها بوده که قسمت عمده آن، مربوط به بسط حریصانه است. این روش، نواحی همپوشان متراکم را تشخیص نمی‌دهد و دو انجمن با ناحیه همپوشان متراکم را عملاً یک انجمن در نظر می‌گیرد[۶۰].

۲-۸ انجمن‌یابی مبتنی بر نزدیکی رئوس

در این روش‌ها، انجمن‌یابی به معنای دسترسی آسان‌تر گروهی از موجودیت‌ها (گره‌های گراف) به یکدیگر است. اعضای انجمن به یکدیگر (در مقایسه با دیگر گره‌های گراف) نزدیکتر بوده و متوسط فاصله بین آنها مقدار کوچکی است. مطابق شکل۲-۷ و براساس مفهوم «جهان کوچک» هرچه فاصله میان دو گره کمتر باشد (همسایگان، همسایگان همسایگان، ...) احتمال تعلق آنها به یک انجمن بیشتر است.

مفهوم «گام تصادفی»[۱] در این روش‌ها کاربرد فراوانی دارد که به حرکت تصادفی یک گام‌زننده از راس موردنظر به یکی از رئوس همجوار با اعمال چندین گام تصادفی و خوشه‌بندی گره‌هایی که اکثراً در یک گام با یکدیگر ظاهر می‌شوند، اشاره دارد.



شکل ۲-۷: مفهوم جهان کوچک در گراف شبکه‌ها [۱۶]

[۱] Random Walk

در روش موسوم به WalkTrap تلاش می‌شود تا گام‌زننده در نواحی با چگالی بالا (انجمن‌ها) به دام بیافتد[۶۱]. این روش یک تابع فاصله بین دو راس تعریف می‌کند که مقدار آن براساس اطلاعات بدست‌آمده از گام‌زدن تصادفی در گراف، محاسبه می‌شود. طبیعتاً هرچه این مقدار کمتر باشد به معنای نگاه یکسان دو راس به شبکه است و احتمال حضور آنها در یک انجمن را قویتر می‌کند. تابع فاصله مطابق رابطه ۲-۱۲ تعریف می‌شود تا احتمال رفتن از i به j با t گام تصادفی را مدل کند:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^{t} - P_{jk}^{t}\right)^2}{d(k)}} \qquad (۱۲-۲)$$

اگر گره‌های i و j متعلق به یک انجمن باشند، میزان احتمال رسیدن از هریک از این دو گره با گام تصادفی به گره k واقع در همان انجمن، نباید دارای تفاوت معناداری باشد.

انتخاب طول گام تصادفی هم مسئله مهمی است زیرا باید به اندازه کافی بزرگ باشد تا بتواند اطلاعات مناسب را گردآوری کند. طول آن معمولاً براساس توپولوژی شبکه تعیین می‌شود. این روش، همپوشانی ندارد.

در روش DOCS[۱] که از همپوشانی حمایت می‌کند، از ماژولاریتی به عنوان یک تابع برازش و نیز برای بسط گراف استفاده می‌شود[۶۲]. شروع کار با حدس اولیه در مورد انجمن‌هاست. سپس در مورد بسط یا حذف از انجمن مفروض براساس اطلاعات بعدی ناشی از انجام چندباره گام تصادفی، تصمیم‌گیری می‌شود.

روش پرکاربرد Infomap که توسط Rosvall و همکاران مطرح شده است[۶۳]، از ترکیب تکنیک‌های مبتنی بر تئوری اطلاعات با معلومات ناشی از اعمال گام‌های تصادفی سود می‌برد. این روش یافتن بهترین ساختار انجمنی را معادل با مینیموم کردن کمیت اطلاعات موردنیاز برای بیان گام تصادفی در شبکه می‌داند.

[۱] Detecting Overlapping Community Structure (DOCS)

کشف ساختار گراف با تعدادی گام تصادفی با طول‌های متفاوت و با احتمال پرش به یک گره تصادفی دیگر انجام خواهد شد. برای هر گره انجمن یک نام منحصربه فرد انتخاب می‌شود. گره‌های داخلی انجمن (گره‌های پرتکرارتر) دارای نام کوتاه‌تری هستند. فشرده‌سازی مناسب با کمک کد هافمن دوسطحی انجام خواهد شد که یک سطح آن برای تشخیص انجمن‌ها در گراف شبکه و سطح دیگر، برای تشخیص گره‌های درون یک انجمن است. بهترین حالت انجمن‌یابی، کوتاه‌ترین طول نام‌گذاری (توصیف) را در میان گام‌های تصادفی خواهد داشت. این روش بسیار دقیق است و در میان روش‌های ناهمپوشان یکی از بهترینها است.

روش UEOC[۱] که همپوشانی را حمایت می‌کند[۶۴] در تشخیص انجمن یک گره، براساس احتمال گذر L-گامی یک گام تصادفی مارکوف روی شبکه تبرید[۲] متناظر با شبکه گراف اصلی عمل می‌کند. شبکه تبرید، یک شبکه تصادفی است که با یک مکانیزم تجمیعی[۳]، درجات گره‌ها را حفظ می‌کند.

در روش LMC[۴]، گام‌زننده تصادفی با یک زنجیره مارکوف مدل می‌گردد[۶۵]. در این روش معیاری به نام احتمال ماندگاری[۵] برای هر انجمن بررسی می‌شود. اگر مقدار این معیار از α بیشتر باشد یک α- community خواهیم داشت و نهایتاً با ترکیب این مولفه‌ها به α–partition خواهیم رسید. این روش همپوشانی ندارد.

[۱] Unfold and Extract Overlapping Communities (UEOC)

[۲] Annealed

[۳] Ensembeling

[۴] Lumped Markov Chains (LMC)

[۵] Persistence

۲-۹ انجمن‌یابی مبتنی بر خوشه‌بندی یال‌ها

تمامی روش‌هایی که تاکنون بررسی شده‌اند، انجمن را مجموعه‌ای از گره‌ها فرض کرده‌اند. در این بخش روش‌هایی بررسی می‌شوند که انجمن را افرازهایی از یال‌ها فرض می‌کنند. مفهوم انجمن در این روش‌ها، قسمت‌هایی از شبکه است که یال‌های داخلی آن بیشتر از یال‌های خارجی آن باشد.

مکانیزم این روش‌ها، مبتنی بر استفاده از «گراف خط»[۱] یا L(G) است که متناظر با هر گراف G(V,E) به فرم گرافی تعریف می‌شود که هر گره آن نظیر یک یال از گراف اصلی بوده و هر یال آن، بیانگر وجود راس مشترک بین هر دو یال از گراف اصلی می‌باشد. مثلاً گراف خط متناظر با یک گراف ستاره، به فرم یک کلیک به دست می‌آید. هر گره در گراف اصلی یک گره همپوشان تلقی می‌شود اگر یال‌های متصل به آن در بیش از یک خوشه از گراف خط نظیر گراف اصلی، عضویت داشته باشند. هر افرازی در گراف خط، معادل با یک انجمن‌یابی همپوشان در گراف اصلی می‌باشد.

روش موسوم به LC[۲] برای یافتن زیرگراف‌های همپوشان از گام تصادفی کمک می‌گیرد[۶۶]. این روش به پایش مقدار تابع ماژولاریتی در حین گام زدن تصادفی روی گراف و تولید دنباله‌های یال- یال و یال- راس- یال می‌پردازد. یک گراف خط وزن‌دار تولید می‌شود تا براساس مقادیر تابع ماژولاریتی در مورد افزار یال‌های آن تصمیم‌گیری شود.

در روش HLC[۳] که توسط Ahn و همکاران پیشنهاد شده است، تعلق گره به دو یا چند انجمن، با خوشه‌بندی یال‌ها مدل می‌شود[۶۷]. در این روش، یک معیار شباهت یال تعریف می‌شود تا برای مجموعه‌ای از همسایگان مشترک گره‌ها محاسبه شود. شباهت یال برای دو یال e_{ik} و e_{jk} که در راس k مشترک هستند مطابق رابطه ۲-۱۳ تعریف می‌شود:

---

[۱] Line Graph

[۲] Link Communities (LC)

[۳] Hierarchical Link Clustering (HLC)

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \qquad (۱۳-۲)$$

معیار S به صورت تعداد همسایگان مشترک میان دو راس، تقسیم بر تعداد کل همسایگان هر دو راس، تعریف می‌شود. مجموعه همسایگان گره i با $n_+(i)$ نشان داده می‌شود و e_{ik} معرف یال موجود بین گره i و گره k می‌باشد. حال یک روش کلاسیک خوشه‌بندی سلسله‌مراتبی برای ساخت دندوگرام یال اعمال می‌شود تا قطع کردن دندوگرام حاصل در مقدار مناسب، انجمن‌یابی براساس یال‌ها را محقق کند.

۲-۱۰ چالش‌های مسئله

روش‌های انجمن‌یابی بررسی‌شده در بخشهای قبلی این فصل، با ذکر ویژگی‌های مهم آنها در جدول ۲-۳ فهرست شده‌اند. اطلاعاتی شامل نام اختصاری هر روش، مدل مورد استفاده برای تعریف و تشکیل یک انجمن، امکان اعمال روش مربوطه روی گراف‌های جهت‌دار و وزن‌دار، نیاز روش به پارامتر ورودی، پیچیدگی زمانی روش، حمایت از مفهوم همپوشانی، و نهایتاً سال ارائه روش بیان شده است.

پیچیدگی زمانی هر روش برای بدترین حالت آن ذکر شده است. در عمده موارد، پیچیدگی زمانی روش‌ها برای حالت متوسط و بدترین حالت آن، یکسان می‌باشد. یادآوری می‌شود که مقدار m برای گراف‌های کم‌تراکم، متناسب با n و برای گراف‌های متراکم، متناسب با n^2 است.

در بررسی قابلیت روش برای کار با گراف‌های جهت‌دار و وزن‌دار، ویژگی‌های رایج‌ترین نسخه یک روش مدنظر بوده است. توضیح اینکه، معمولاً تعمیم‌های متعددی برای روش‌های معروف ارائه می‌شود تا بعضی قابلیت‌ها را به نسخه اولیه آن اضافه کند. عمدتاً رسیدن به این قابلیت‌ها، با پیچیدگی زمانی بالاتر یا کم‌رنگ شدن دیگر نقاط قوت روش همراه است.

عدم نیاز به پارامتر ورودی، به معنای سادگی یک روش است و مزیتی برای آن تلقی می‌شود. در روش‌های بدون پارامتر ورودی، فقط لازم است تا گراف ورودی به‌صورت ماتریس مجاورت یا لیست

مجاورت به روش مربوطه داده شود. به‌عنوان مثال‌های شاخص از پارامتر ورودی می‌توان به متغیرهای توزیع آماری یا مقدار k در تعیین اندازه کلیک، اشاره کرد.

با نگاه به جدول۲-۳ که بسیاری از مطالب مطرح‌شده را تجمیع و خلاصه کرده است، یک نکته مهم مشخص می‌شود. برای همین خاصیت‌های مطرح‌شده نیز هیچ روش انجمن‌یابی وجود ندارد که هم از همپوشانی روی گراف‌های جهت‌دار و وزن‌دار حمایت کند، هم پارامتر ورودی نداشته باشد، و هم پیچیدگی زمانی آن مطلوب باشد. هریک از روش‌ها با توجه به مکانیزم خود، دارای نقاط قوت و ضعف هستند.

برخی از روش‌ها با گراف‌های خلوت[۱] و کم‌تراکم بهتر کار می‌کنند که پیچیدگی یال‌ها در آنها به صورت $O(n)$ است، اما مابقی روش‌ها روی گراف‌های متراکم که پیچیدگی یال‌ها در آنها از مرتبه $O(n^2)$ است نتایج بهتری می‌گیرند.

برخی از روش‌ها روی برخی از شبکه‌ها (گراف متناظر آنها) دارای عملکرد بهتری است. این امر از تفاوت ماهوی شبکه‌ها با یکدیگر ناشی می‌شود. لذا لازم است تا در انجمن‌یابی به تفاوت‌ها و موارد متمایزکننده میان شبکه‌های اجتماعی برخط با دیگر شبکه‌ها توجه نمود. عمده روش‌ها از رویکرد حریصانه[۲] استفاده می‌کنند. البته ممکن است خود این رویکردها نیز با یکدیگر اشتراک داشته باشند. مثلاً در روش انتشار، از یک تابع چگالی یا یک الگوی ساختاری شبکه برای انجام و نیز تخمین کیفیت انجمن‌یابی استفاده شود. ضمن اینکه ایده‌های به‌کاررفته در شکل‌گیری یک انجمن نیز ممکن است دارای اشتراکاتی باشند. مثلاً روشی که کلیک را (بجای گره) به‌عنوان هسته اولیه برای انتشار انجمن‌ها مدنظر قرار می‌دهد، هم از مفاهیم بخش ساختار و هم از مفاهیم بخش انتشار استفاده می‌کند.

امکان موازی‌سازی اجرای الگوریتم‌ها نقش به‌سزایی در کاهش زمان اجرای آنها دارد. گاهی اوقات توازی در ذات خود روش مورداستفاده وجود دارد و می‌تواند بصورت محاسبات مستقل‌ازهم تلقی شده و

---

[۱] Sparse

[۲] Greedy

در چارچوب MapReduce یا تکنیکهای داده‌کاوی انجام شود. بنا بر ماهیت روش انجمن‌یابی، می‌توان امکان موازی‌سازی اجرای الگوریتم را با قابلیتهای سخت‌افزاری و امکانات برنامه‌نویسی به روش افزود [۶۸].

عملاً هیچ روشی نمی‌تواند به عنوان بهترین روش انجمن‌یابی، خصوصاً برای شبکه‌های دنیای واقعی مطرح شود. زیرا هرکدام از روش‌ها با توجه به فرض‌های اولیه، روی شبکه خاصی (با توجه به ویژگی‌ها و توپولوژی آن) نتیجه بهتری می‌گیرند.

جدول ۲-۳: مشخصات برخی الگوریتم‌های شاخص در انجمن‌یابی

نام روش	مدل	پیچیدگی	جهت‌دار بودن گراف ورودی	وزن‌دار بودن گراف ورودی	نیاز به پارامتر ورودی	همپوشانی	سال
Bayesian-NMF	طیفی	$O(cn^2)$	×	✓	دارد	✓	2011
BIG CLAM	طیفی	$O(cn + m)$	×	✓	دارد	✓	2013
GN	مرکزگرایی	$O(m^2n)$	✓	✓	دارد	×	2002
CONGA	مرکزگرایی	Best: $(nlogn)$ Worst: $O(m^3)$	×	✓	دارد	✓	2007
DOCNet	مرکزگرایی	$O(n^2)$	×	×	ندارد	✓	2014
NLA	مرکزگرایی	$O(n^2)$	✓	✓	دارد	✓	2016
CNM	تابع کیفیت	$O(nlog^2n)$	×	✓	ندارد	×	2004
Louvain	تابع کیفیت	$O(m)$	×	✓	ندارد	×	2008
CIS	تابع کیفیت	$O(n^2)$	×	✓	دارد	✓	2009
LFM	تابع کیفیت	$O(n^2)$	×	✓	دارد	✓	2009
OSLOM	تابع کیفیت	$O(n^2)$	✓	✓	دارد	✓	2011
FOCS	تابع کیفیت	$O(m + n)$	×	×	دارد	✓	2015
GA-Net	تابع کیفیت	$O(n^2)$	×	×	دارد	✓	2008
Game	تابع کیفیت	$O(m^2)$	✓	✓	دارد	✓	2010
LPA	انتشار	$O(m + n)$	×	✓	ندارد	×	2007
COPRA	انتشار	$O(mlog(m/n))$	×	✓	ندارد	✓	2010
SLPA	انتشار	$O(Tm)$	✓	✓	دارد	✓	2012
DEMON	انتشار	$O(nK^{3-\alpha})$	×	×	ندارد	✓	2012
MMSB	انتشار	$O(\|c\|.n)$	✓	×	دارد	✓	2007
MOSES	انتشار	$O(n^2)$	×	×	ندارد	✓	2010
NISE	انتشار	$O(m + n)$	×	✓	دارد	✓	2016
CPM	ساختار	$O(m^{\log m/10})$	×	×	دارد	✓	2005
EAGLE	ساختار	$O(n^2 +(h+n)p)$	✓	✓	دارد	✓	2009
GCE	ساختار	$O(mh)$	✓	✓	دارد	✓	2010
WalkTrap	نزدیکی	$O(mn^2)$	×	✓	دارد	×	2006
Infomap	نزدیکی	$O(nlogn)$	✓	✓	دارد	×	2008
UEOC	نزدیکی	$O(ln^2)$	×	×	دارد	✓	2011
LinkCom	یال	$O(m.\|c\|.logn)$	×	✓	دارد	✓	2009
HLC	یال	$O(nK^2)$	×	✓	دارد	✓	2010

(n = تعداد رئوس، m = تعداد یال‌ها، K = بیشینه درجات رئوس، α = ضریب توانی، $|c|$= تعداد انجمن‌ها، T = تعداد دفعات اجرای الگوریتم، h = تعداد ماکزیمال‌کلیک‌های یک گراف، p = تعداد زوج‌های ماکزیمال‌کلیک مجاور یکدیگر، l = طول گام تصادف

فصل ۳ : پیش نیازها

۳-۱ مقدمه

در این فصل، ملزومات موردنیاز مدل پیشنهادی به‌طور مجزا بررسی خواهند شد. بخش ۳-۲، به معرفی و دسته‌بندی کوتاهی از کاربردهای بازنمایی موجودیت‌ها (کلمات، گره‌ها، ....) و توکاری[۱] گراف اختصاص دارد. در بخش ۳-۳ مبانی یادگیری عمیق و مولفه‌های پرکاربرد آن بررسی می‌شود. بخش۳-۴ به معرفی ابزارها و انواع دادگان[۲] مورداستفاده برای انجام تجربیات مختلف در ارزیابی عملکرد روش‌های انجمن‌یابی می‌پردازد.

۳-۲ توکاری گراف

تکنیک‌های یادگیری ماشین در کاهش ابعاد داده‌های مسئله در سطوح مختلف تجرید (انتزاع) از قابلیت-های بسیاری برخوردار هستند. کارکرد کلی این روشها در استخراج و بازنمایی را می‌توان مطابق زیر بیان کرد:

- یادگیری چندسطحی از بازنمایی مختلف به کمک ساختار سلسله‌مراتبی روی ویژگی‌های سطح پایین
- یادگیری بازنمایی به عنوان یک رویکرد خودکار در استخراج ویژگی‌ها از ورودی‌های سطح پایین
- کشف چندین سطح از بازنمایی‌ها (نمایش‌های) توزیع‌شده از داده ورودی
- استخراج ویژگی‌های سطح بالا از داده‌های آموزشی برچسب نخورده (عمدتا در شبکه‌های باور عمیق)
- افزایش قدرت تمایز بین دسته‌های مختلف داده‌ها

بازنمایی و توکاری گراف کاربردهای متعددی دارد و در مراجع زیادی به آنها پرداخته شده است[۶۹, ۷۰]. ما در این رساله از توکاری گراف بعنوان بخش از مدل انجمن‌یابی خود استفاده کرده‌ایم. هدف

---

[۱] Embedding

[۲] Dataset

اصلی در توکاری گراف[1] این است که ضمن حفظ ساختار گراف آن را به ابعاد کوچکتر نگاشت بدهیم و و گره‌ها را در قالب بردارهایی از اعداد حقیقی ارائه نماییم. در پایان عملیات توکاری، هر گره گراف به شکل یک بردار بیان خواهد شد. عمده تلاش روش‌های توکاری بر این نکته است که تا سرحد امکان، ویژگیهای ساختاری گراف‌ها (خصوصا برای شبکه‌های بزرگ) را حفظ نمایند. این مهم با بررسی نحوه حفظ معیارهای مجاورت مرتبه اول و دوم و بالاتر امکان‌پذیر خواهد بود. روش‌های مبتنی بر کار روی ماتریس‌ها نظیر روش Graph Factorization (GF) [۷۱] و نیز روش Node2vec [۷۲] تمرکز خود را روی حفظ معیارهای مجاورت مرتبه پایین (اول و دوم) در گراف گذاشته‌اند. برخی روش‌ها مانند روش HOPE[2] نیز مرتبه‌های بالاتر مجاورت را حفظ می‌کنند [۷۳]

باید اذعان داشت که ایده اصلی نگاشت گراف و گره‌های آن به ابعاد پایین‌تر، از پردازش زبان‌های طبیعی و مواردی چون مدل‌سازی زبانی و متن‌کاوی نشات می‌گیرد[۷۴]. زیرا پیشگامان عرصه توکاری گراف، پیشرفت‌هایی که در حوزه مدل‌سازی زبانی (به‌عنوان رشته‌هایی از کلمات) رخ داده بود را به مقوله گراف (به‌عنوان گام‌هایی معادل با رشته‌هایی از گره‌ها) تعمیم دادند. روش Word2vec [۷۵] که توسط Mikolov و همکاران ارائه شده است یک روش مهم و پیشگام در این امر محسوب می‌شود.

الگوریتم DeepWalk نیز با در نظر گرفتن این تشابه، روشی را برای تعمیم مدل‌سازی زبان به نظریه گراف ارائه کرده است. این تعمیم بر اساس مجموعه‌ای از گام‌زنی‌ها یا پیمایش‌های تصادفی در گراف ایجاد شده است. زیرا هر یک از پیمایش‌های یادشده را می‌توان به عنوان یک عبارت یا جمله کوتاه در یک زبان خاص در نظر گرفت (زبانی که گره‌های گراف، کلمات آن را تشکیل می‌دهند). روش Deep Walk با گردآوری مجموعه محدودی از پیمایش‌های تصادفی، متن پیکره‌ای را ایجاد می‌کند که در واقع معرف گراف مربوطه است و بردارهای توکاری گره‌ها به عنوان کلمات این متن از درون آن استخراج

[1] Graph Embedding

[2] High-Order Proximity Ppreserved Embedding (HOPE)

می‌شوند. این روش به دلیل وجود یک فرآیند نمونه‌برداری مفصل کند است. به علاوه تعیین ابرپارامترهای[۱] آن مثل طول گام و تعداد دفعات گام‌زنی برای هر گره، کار آسانی نیست.

۳-۳ یادگیری عمیق

در کنار روش‌های مطرح‌شده، روش‌های مبتنی بر یادگیری عمیق[۲] هم خود را به روش‌های مفید و کارآمد در انجمن‌یابی اضافه کرده‌اند. نقطه قوت این دسته از روش‌ها را می‌توان در استخراج ویژگی‌ها[۳] به صورت خودکار دانست [۷۶]. این بخش به معرفی یادگیری عمیق با تاکید بر مولفه‌های مورداستفاده آن از جمله کدکننده خودکار، در این رساله خواهد پرداخت.

۳-۳-۱ معرفی کلی

شبکه‌های عصبی مصنوعی، یک روش قدرتمند محاسباتی با الهام از عملکرد سلول‌های عصبی مغز است و یادگیری عمیق به‌عنوان نوع خاصی از این شبکه‌ها، با حجم زیادی از داده و در لایه‌های متعدد، آموزش می‌بینند. هرچند عمده شهرت یادگیری عمیق ناشی از عملکرد عالی آن در حیطه پردازش و تفسیر تصاویر است اما یادگیری عمیق در زمینه‌هایی چون درک و پردازش صوت، ترجمه ماشینی، و پردازش زبان‌های طبیعی هم موفق عمل کرده است [۷۷].

از سوی دیگر، یادگیری عمیق با مفهوم بازنمایی[۴] ارتباط نزدیکی دارد آنچه در یادگیری ماشین سنتی مرسوم است این است که ما باید برای تدوین یک روش یادگیری، مجموعه‌ای از ویژگی‌ها را به‌صورت دستی استخراج کنیم و سعی نماییم با استفاده از نگاشت‌های مختلف (خطی یا غیرخطی) مسیری را

---

[۱] Hyper Parameters
[۲] Deep Learning
[۳] Feature Extraction
[۴] Representation

برای رسیدن از این ویژگی‌ها به خروجی‌های صحیح و موردنظر بیابیم. استخراج و انتخاب ویژگی‌های مناسب و کارآمد (به‌صورت دستی) وابستگی زیادی به مهارت و خبرگی افراد دارد و می‌تواند به کاری دشوار و زمان‌بر تبدیل شود. به‌علاوه ویژگی‌هایی که در حل یک مسئله خاص بسیار خوب عمل می‌کنند، الزاماً برای حل دیگر مسائل، مفید و موثر نخواهند بود.

بنابراین انتخاب شیوه مناسب برای بازنمایی داده‌های ورودی (استخراج و انتخاب ویژگی‌های مناسب از میان انبوه ورودی‌ها) می‌تواند در کیفیت و کارآیی روش یادگیری تأثیر فراوانی داشته باشد.

در یادگیری عمیق، ابتدا اولین لایه بعد از ورودی، ساده‌ترین ویژگی‌ها را استخراج می‌کند. با حرکت به سمت بالا، هر لایه ویژگی‌های اخذشده از لایه‌های پایین‌تر را دریافت کرده و با فیلترینگ و تبدیلات عمدتا غیرخطی آنها را به ویژگی‌های سطح بالاتری نگاشت می‌دهد. به‌عنوان مثال، روش SDNE با استفاده از مفاهیم یادگیری عمیق، تلاش در توکاری شبکه برای کاربردهای خاص دارد [۷۸].

۳-۳-۲ کدکننده خودکار

یک ساختار پرکاربرد در یادگیری عمیق برای کاهش ابعاد مسئله و استخراج ویژگی‌های مناسب، استفاده از مولفه‌ای به نام کدکننده خودکار[۱] (خودرمزگذار) است. کدکننده خودکار شامل دو قسمت کدگذار[۲] و کدگشا[۳] است. کدکننده خودکار (AE) فقط یک لایه پنهان دارد اما کدکننده خودکار پشته‌شده[۴]، مطابق شکل۳-۱ مکانیزم AE را تکرار می‌کند و دارای چند لایه پنهان می‌باشد[۷۹].

در صورتی که بازنمایی‌های پیچیده‌تری مد نظر باشد می‌توان تعدادی کدکننده ساده را به هم متصل کرده و یک کدکننده خودکار پشته‌شده را ایجاد نمود. ساختارهای شامل SAE، برخلاف اکثر شبکه‌های عصبی، در حالت بدون ناظر آموزش داده می‌شوند. این ساختارها تلاش می‌کنند تا ورودی را در خروجی

---

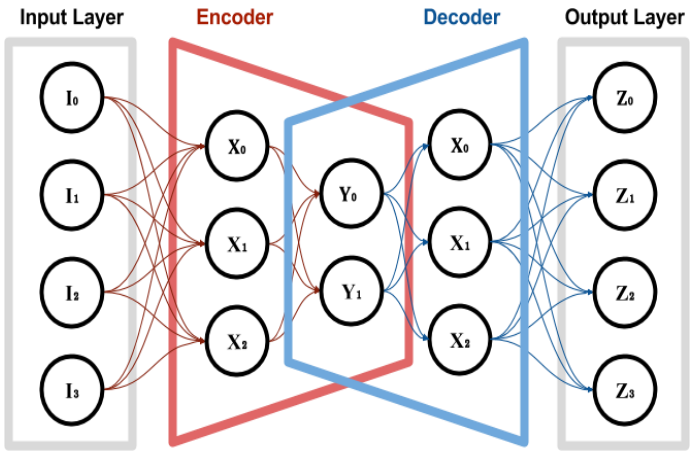
[۱] AutoEncoder (AE)

[۲] Encoder

[۳] Decoder

[۴] Stacked Auto-Encoder (SAE)

خود تکرار و بازتولید نمایند. برای امکان بازسازی ورودی در قسمت خروجی، باید تعداد نورون‌های لایه ورودی با نورون‌های لایه خروجی یکسان باشد. تفاوت مشاهده‌شده مابین ورودی و خروجی یک SAE، به خطای بازسازی[1] تعبیر می‌شود. تصمیم‌گیری روی ویژگی‌های یافته‌شده (با توجه به خطای بازسازی)، مبنای کار بسیاری از روش‌های مبتنی بر SAE است.



شکل ۳-۱: مثالی از کدکننده خودکار پشته‌شده با سه لایه پنهان.

بعد از اتمام فرایند آموزش، استخراج ویژگی‌ها در لایه‌های پنهان SAE انجام می‌شود. لایه‌های پایین‌تر وظیفه استخراج ویژگی‌های سطح پایین را بر عهده دارند و لایه‌های بالاتر (لایه‌های پایانی SAE) ویژگی‌های سطح بالا را استخراج می‌کنند. لایه‌های پنهان دارای نورون‌های کمتری نسبت به لایه ورودی هستند. در واقع، نورون‌های لایه‌های پنهان و وزن‌های حاصل برای اتصالات شبکه، تفسیرکننده ویژگی‌های داده ورودی هستند.

آموزش یک AE شامل دو مرحله کدگذاری و کدگشایی است. در مرحله اول، کد کننده خودکار ورودی خود را در قالب یک بردار دریافت می‌کند و با عبور دادن این بردار از فیلتر لایه پنهان آن را به یک فضای ویژگی جدید مطابق رابطه ۳-۱ نگاشت می‌نماید:

[1] Reconstruction Error

$$f_{\theta_1}(x) = \sigma(W_1 x + b_1) \quad (۱-۳)$$

نتیجه رابطه بالا کد[۱]، متغیر پنهان[۲]، یا بازنمایی پنهان[۳] نامیده می‌شود. $\theta_1 = (W_1, b_1)$ شامل پارامترهایی است که در فرایند کدگذاری موثر هستند، W_1 معرف وزن ها بوده و b_1 به معنای بردار بایاس می‌باشد. برای مدل سازی روابط غیرخطی بین فضای اصلی و فضای پنهان، معمولاً از یک تابع فعال سازی $\sigma(.)$ استفاده می کنند. ما برای این بخش از کار، تابع نیمه خطی زیر را به کار گرفته‌ایم:

$$f(x) = \max(x, 0) \quad (۲-۳)$$

این تابع که در ادبیات یادگیری عمیق واحد یکسوساز خطی یا ReLU[۴] نامیده می‌شود در شبکه‌هایی که از لایه‌های متعدد تشکیل شده‌اند فرایند یادگیری را با سرعت قابل‌توجهی انجام می‌دهد.

اصلی‌ترین مولفه مرحله کدگشایی، تابع بازسازی $g(.)$ است. این تابع سعی می‌کند داده‌های اصلی را از درون فضای پنهان استخراج کرده و بازسازی نماید. تابع یاد شده در رابطه ۳-۳ نمایش داده شده است:

$$g_{\theta_2} = \sigma(W_2 y + b_2) \quad (۳-۳)$$

در رابطه فوق y خروجی حاصل از مرحله کدگذاری است و θ_2 پارامترهای شبکه کدگشا یعنی ماتریس اوزان (W_2) و بردار بایاس (b_2) را خاطر نشان می‌کند.

آموزش کامل AE یعنی تعیین مقادیر بردارهای W و W’ و متغیرهای b و b’ که معرف وزن‌ها و بایاس‌های شبکه هستند و با کمینه‌سازی خطای بازسازی محقق می‌گردد. معیار اندازه‌گیری خطای یادشده به توزیع احتمالاتی داده‌های ورودی بستگی کامل دارد. مثلا درحالی‌که این داده‌ها از یک توزیع گاوسی پیروی می‌کنند، بهتر است از تابع مرسوم مربعات خطا مطابق رابطه ۴-۳ استفاده شود.

$$L(x, y) = \|x - y\|^2 \quad (۴-۳)$$

---

[۱] Code

[۲] Latent Variable

[۳] Latent Representation

[۴] Rectified Linear Unit (ReLU)

اما درصورتی‌که داده‌های ورودی بردارهای بیتی یا بردارهایی از احتمالات وقوع بیت‌ها باشند، تابع آنتروپی متقابل[۱] که غالباً در رده‌بندی مورداستفاده قرار می‌گیرد نتایج دقیق‌تری را تولید می‌کند. این تابع در رابطه ۳-۵ نمایش داده‌شده است:

$$L(x, y) = \sum_{k=1}^{d} (x_k * \log y_k + (1 - x_k) * \log(1 - y_k)) \quad (۵-۳)$$

کد کننده خودکار با PCA ارتباط نزدیکی دارد. درواقع اگر تنها یک‌لایه پنهان متشکل از k نورون با تابع فعال‌سازی خطی وجود داشته باشد، و برای محاسبه خطا از روش MSE استفاده کنیم، آنگاه نورون‌های لایه پنهان، k مؤلفه اصلی[۲] داده‌های موردنظر را تولید می‌کنند. با به‌کارگیری نگاشت‌های غیرخطی مثل tanh یا تابع سیگموئید، کد کننده خودکار وجوه پیچیده‌تری از داده‌های ورودی را استخراج خواهد کرد. معمولاً برای اینکه AE به دام نگاشت همانی نیافتد (افتادن در دام جواب بدیهی) تعداد نورون‌های لایه پنهان باید کمتر از ابعاد داده‌های ورودی باشد. بدین ترتیب AE را می‌توان به‌عنوان یک فشرده‌ساز دارای خطا در نظر گرفت که هیچ‌وقت ورودی خود را به‌طور کامل و مطلق بازسازی نمی‌کند.

۳-۴ ابزار و دادگان‌ها

در این بخش، معرفی مختصری درباره ابزارها و دادگان مورد استفاده در این رساله انجام می‌شود. برای انجام تجربیات و مقایسه مدل پیشنهادی با دیگر روش‌های انجمن‌یابی، امکان استفاده از دادگان دنیای واقعی (استاندارد، استخراجی) و نیز دادگان شبیه‌سازی‌شده (محک[۳]) وجود دارد که در ادامه توضیح داده می‌شوند.

---

[۱] Cross Entropy

[۲] Principal Component

[۳] Benchmark

۳-۴-۱ ابزارها

در این رساله از ابزار Gephi 8.2b[۱] که یک بسته نرم‌افزاری به زبان جاوا روی پلتفرم NetBeans است برای مطالعه و ویرایش و دستکاری[۲] گراف انواع شبکه‌ها (پیچیده، پویا، سلسله‌مراتبی، جهان کوچک، ...) استفاده شده است. علت انتخاب را می‌توان پشتیبانی از تمام فرمت‌های نمایش گراف‌ها، رایگان و متن باز بودن، نمایش بلادرنگ[۳] تغییرات شبکه‌ها، سهولت کار با مجموعه داده‌های بزرگ، انجام پرس‌وجو های متعدد روی گراف‌ها دانست [۸۰]. از دیگر ابزارهای مورداستفاده می‌توان به اپلیکیشنها و افزونه‌های[۴] رایگان موجود در شبکه‌های اجتماعی مانند FQL[۵]، نرم‌افزار Cfinder[۶] برای بررسی عملکرد روش CPM، امکانات و بسته‌های زبان پایتون مانند NetworkX، و نیز امکانات متنوع Keras که کتابخانه زبان پایتون برای یادگیری عمیق است اشاره نمود. این ابزارها قابلیت‌های متعددی در کار با دادگان موجود را فراهم می‌کند. [۸۱]

نرم‌افزارهای تجزیه و تحلیل شبکه‌ها جهت شناسایی، تجزیه و تحلیل، تجسم و شبیه‌سازی گره‌ها و یال‌ها از انواع داده‌های ورودی به کار می‌روند. ابزار تجزیه و تحلیل شبکه به محققان اجازه می‌دهد تا شبکه‌هایی با اندازه‌های مختلف را بررسی کرده و رویه‌های ریاضی و آماری را بر مدل‌های شبکه اعمال کنند. وجود نمایشهای بصری مختلف به درک بهتر نتایج کمک زیادی می‌کند.

۳-۴-۲ دادگان استاندارد

قسمت مهمی از تجربیات این رساله روی انواع دادگان دنیای واقعی انجام می‌شود که در مقایسه با یکدیگر از تنوع خوبی در اندازه و چگالی و دیگر مشخصات گراف برخوردار هستند. داده‌های بسیاری از

---

[۱] http://www.gephi.org

[۲] Manipulation

[۳] Real-time

[۴] Plug-in

[۵] Facebook Query Language

[۶] http://www.cfinder.org

انواع شبکه‌ها با جزییات کامل آنها (تعداد گره‌ها، یال‌ها، طرز گردآوری، مطالعات انجام‌شده، ...) در رده‌بندی‌های[۱] مختلف، گردآوری شده‌اند تا مورد استفاده محققین قرار بگیرند و امکان بررسی و مقایسه عملکرد روش‌های متفاوت را روی دادگان یکسان و استاندارد فراهم سازند.

مجموعه دادگان گردآوری‌شده در SNAP[۲] [۸۲] از دانشگاه استنفورد مجموعه دقیقی است که شبکه‌ها را بخوبی Categorized نموده و یک توضیح کوتاه برای هر category و نیز توضیحات مفصلی برای هر دادگان ارائه شده است. این مجموعه عملا یک کتابخانه چندمنظوره برای تحلیل شبکه و گراف‌کاوی است. البته چند شبکه متفاوت دارای مرجع (Ground Truth) هم دارد که بزرگ و حجیم هستند و گرفتن نتایج روی آنها زمان زیادی می‌برد. شکل۳-۱ مثال‌هایی از دادگان گردآوری‌شده در مجموعه SNAP دانشگاه استنفورد به همراه مشخصات آنها را نشان می‌دهد.

Social networks

Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter
soc-Epinions1	Directed	75,879	508,837	Who-trusts-whom network of Epinions.com
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
soc-Pokec	Directed	1,632,803	30,622,564	Pokec online social network
soc-Slashdot0811	Directed	77,360	905,468	Slashdot social network from November 2008
soc-Slashdot0922	Directed	82,168	948,464	Slashdot social network from February 2009
wiki-Vote	Directed	7,115	103,689	Wikipedia who-votes-on-whom network
wiki-RfA	Directed, Signed	10,835	159,388	Wikipedia Requests for Adminship (with text)

Networks with ground-truth communities

Name	Type	Nodes	Edges	Communities	Description
com-LiveJournal	Undirected, Communities	3,997,962	34,681,189	287,512	LiveJournal online social network
com-Friendster	Undirected, Communities	65,608,366	1,806,067,135	957,154	Friendster online social network
com-Orkut	Undirected, Communities	3,072,441	117,185,083	6,288,363	Orkut online social network
com-Youtube	Undirected, Communities	1,134,890	2,987,624	8,385	Youtube online social network
com-DBLP	Undirected, Communities	317,080	1,049,866	13,477	DBLP collaboration network
com-Amazon	Undirected, Communities	334,863	925,872	151,037	Amazon product network

شکل ۳-۲: نمونه‌هایی از دادگان استاندارد (گردآوری‌شده توسط دانشگاه استنفورد)

[۱] Categories

[۲] Stanford Network Analysis Project (SNAP)

یکی از بهترین دسته‌بندی‌های دادگان استاندارد هم توسط یک دانشگاه آلمانی با معرفی پایگاه Konect انجام شده است [۸۳]. دادگان‌های ارائه‌شده توسط طراحان نرم‌افزار Gephi هم خوب است. اما عمدتا شبکه‌های کوچک و کلاسیک را شامل می‌شود. چند نمونه از دیگر مجموعه‌های داده عبارتند از:

- دادگان فهرست‌شده در سایت دانشگاه استنفورد[۱]
- دادگان فهرست‌شده در دانشگاه میشیگان[۲]
- دادگان فهرست‌شده در سایت موسسه ماکس پلانک[۳]

عمده دادگان مورد استفاده از پایگاه SNAP [۸۲] انتخاب شده است.

۳-۴-۳ دادگان سنتزشده (محک)

بخشی از دادگان مورداستفاده در رساله حاضر به‌وسیله محک LFR ایجادشده است. مجموعه LFR نام خود را از سه حرف اول نام خانوادگی مؤلفین خود یعنی Lancichinetti و Fortunatoو Radicchi اخذ کرده است. این سه محقق در ۲۰۰۹ در یک مقاله پر مراجعه الگوریتم خود و برنامه‌های مربوطه را معرفی کردند [۸۴].

مزیت و سهولت در کار با گراف‌های محک در این است که تمام پارامترها معلوم هستند، چون توسط خودمان تشکیل و تعیین می‌شوند. لذا حتما مرجع مقایسه یا ground truth داریم و از معیارهای بیشتری نیز می‌توان برای ارزیابی کارایی استفاده کرد. در جدول۳-۱ پارامترهای مورد نیاز در تعیین قطعی مشخصات گراف شبیه‌سازی‌شده توسط کاربر، فهرست شده‌اند.

[۱] http://snap.stanford.edu/data/index.html

[۲] http://www-personal.umich.edu/~mejn/netdata/

[۳] http://socialnetworks.mpi-sws.org/datasets.html

الگوریتم LFR به کاربر اجازه می‌دهد که از طریق پارامترهای متعدد، ساختار گراف حاصله ازجمله تعداد و اندازه انجمن‌ها را معین کند. بدین ترتیب کاربر علاوه بر گراف، داده مرجع را نیز در اختیار دارد و به همین دلیل امکان ارزیابی دقیق الگوریتم‌های انجمن یابی برای او فراهم می‌شود.

تفاوت عمده این مجموعه با مجموعه‌های مشابه در اینجاست که در گراف‌های خروجی الگوریتم LFR تعداد گره‌ها و اندازه انجمن‌ها بر اساس توزیع توانی پراکنده‌شده‌اند به همین دلیل این گراف‌ها به گراف‌های دنیای واقعی بسیار شبیه هستند.

جدول ۳-۱: پارامترهای محک LFR

نام	توضیح
n	تعداد گره‌ها
k_{avg}	متوسط درجه گره‌ها
k_{max}	حداکثر درجه یک گره
μ	ضریب اختلاط برای گره‌ها
min_c	حداقل اندازه هر انجمن
max_c	حداکثر اندازه هر انجمن
O_n	نسبت گره‌های همپوشان
O_m	تعداد انجمن‌هایی که گره‌های همپوشان در آن‌ها عضو هستند
t_1	نمای توزیع توانی درجه گره‌ها
t_2	نمای توزیع توانی اندازه انجمن‌ها

درجه گره‌ها و اندازه انجمن‌ها نیز بر اساس دو توزیع توانی متفاوت با نماهای γ و β تعیین می‌شوند. پارامتری به نام ضریب اختلاط با μ نمایش داده می‌شود. ضریب اختلاط معرف درصد گره‌هایی است که در همسایگی یک گره مشخص قرار دارند و در هیچ‌یک از انجمن‌هایی که گره مذکور در آن‌ها عضو است، عضو نیستند. این پارامتر درواقع کسری از لبه‌ها که در میان انجمن‌ها قرار می‌گیرند را کنترل

می‌کند و به همین دلیل ضمناً بیانگر میزان اعوجاج گراف نیز هست. اگر مقدار ضریب اختلاط برابر صفر باشد ($\mu=0$) همه اتصالات در میان انجمن‌ها قرار می‌گیرند، و اگر این مقدار برابر با یک تنظیم شود ($\mu=1$) همه لبه‌ها در میان گره‌هایی واقع می‌شوند که عضو انجمن‌های متفاوت می‌باشند.

مقادیر مرزی برای حداقل و حداکثر درجات گره، به ترتیبی انتخاب می‌شوند که متوسط درجه گره‌ها برابر با $<k>$ دربیاید. گام سوم: اندازه انجمن‌ها از یک توزیع توانی با نمای β اخذ می‌شود. مجموع این اندازه‌ها باید برابر با تعداد گره‌های گراف یعنی N باشد. حداقل و حداکثر اندازه انجمن‌ها، باید به ترتیبی انتخاب شود که سایر محدودیت‌های موردنظر کاربر در ساختار گراف لحاظ گردد. الگوریتم LFR طوری طراحی و پیاده‌سازی شده است که به‌سرعت همگرا می‌شود. پیچیدگی زمانی این الگوریتم به‌حسب تعداد گره‌ها و تعداد اتصالات خطی است.

۳-۴-۴ دادگان استخراج‌شده

به غیر از موارد بکارگیری دادگان استاندارد، استفاده از روش‌های متعدد نمونه‌گیری[۱] از شبکه‌های اجتماعی (crawling در شبکه) با توجه به کمبود منابع پردازشی و تلاش برای کاهش زمان محاسبات، امری اجتناب‌ناپذیر است [۸۵]. البته باید با تکنیکهای متعددی اعلام کرد که نمونه گرفته‌شده حتماً مبیّن[۲] است.

این نوع از دادگان به کمک خزشگرها[۳] حاصل می‌شوند. الزاما استاندارد نیستند و حتی ممکن است حالت خصوصی داشته باشند و عملا استفاده از آنها نیاز به مجوز داشته باشد. دادگان با نام Fb180، گراف دوستان یک کاربر شبکه فیس‌بوک (نویسنده رساله) بوده که با نرم‌افزار Netvizz از این شبکه استخراج شده است.

---

۱ Sampling

۲ Representative

۳ Crawlers

فصل ۴ : مدل پیشنهادی

۴-۱ مقدمه

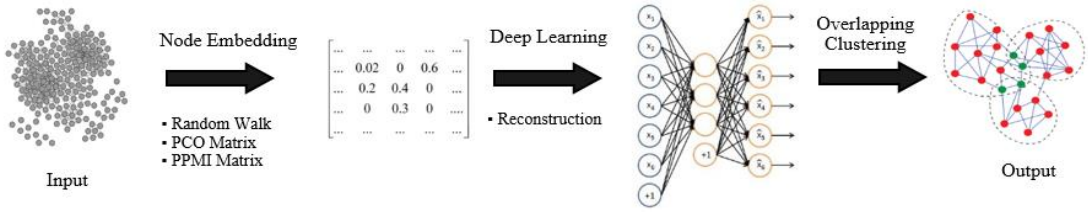
در این فصل، کلیات مراحل مدل پیشنهادی مطرح شده و سپس جزییات این مراحل به تفصیل شرح داده می‌شوند. عمده دانش اولیه برای هر مرحله کلی (بازنمایی و توکاری گراف، معماری یادگیری عمیق، ...) در فصول قبل بالاخص در مطالب فصل ۳ بعنوان پیش‌نیاز ذکر شده‌اند. بخش ۴-۲ به دید کلی از پیکربندی مدل پیشنهادی اختصاص دارد. در این بخش، ابتدا نحوه توکاری با الهام از روشهای پردازش متن شرح داده می‌شود سپس به معماری مورد استفاده از مولفه‌های یادگیری عمیق پرداخته خواهد شد و نهایتا نحوه خوشه‌بندی همپوشان بررسی می‌شود.

۴-۲ پیکربندی مدل پیشنهادی

در این بخش، مراحل مدل پیشنهادی پس از یک دید کلی شرح داده می‌شود.

۴-۲-۱ دید کلی

ساختار کلی مدل پیشنهادی در شکل ۴-۱ مشاهده می‌شود.



شکل ۴-۱: ساختار کلی مدل پیشنهادی

ورودی روش پیشنهادی، گراف متناظر با شبکه ورودی بوده و خروجی روش، انجمن‌های یافته‌شده برای آن شبکه می‌باشد.

روش پیشنهادی که از SAE به عنوان استخراج‌کننده ویژگی‌ها و بازسازی شبکه در خروجی خود استفاده می‌کند، شامل سه مرحله کلی است:

۱- Random Surfing

مکانیزم تعیین بردار نمایش گره‌های شبکه به عنوان ورودی به قسمت یادگیری عمیق: در اولین مرحله از روش پیشنهادی، بردارهای نمایش نظیر هر گره تولید می‌شوند. الگوریتم پیشنهادی ما برای اجتناب از فرآیند نمونه برداری زمان بر الگوریتم Deep Walk ماتریسی به نام ماتریس توارد یا هم رخدادی احتمالی یا PCO[۱] را مستقیماً از گراف استخراج می‌کند.

۲- Deep Learning

در مرحله دوم، از نمایش بدست‌آمده به عنوان ورودی یادگیری عمیق استفاده می‌کنیم. به منظور حل مسئله توسط مدل شبکه عصبی، نیاز است تا وزن‌های مناسب برای هر لایه تعیین شوند. پس از تعیین پارامترهای مدل (تعداد لایه‌های پنهان، تعداد نورون‌های هر لایه، نرخ یادگیری، تابع فعال‌سازی، ...) وزن‌های بهینه مرتبط با هر لایه به کمک دادگان معین‌شده برای آموزش، محاسبه می شوند. معیار دقت کدکننده خودکار خطای بازسازی است.

۳- Clustering

استفاده از الگوریتم خوشه‌بندی و تعیین پارامترهای آن: ویژگی‌های استخراج شده را به عنوان ورودی به یک دسته‌بند (خوشه‌بند) می‌دهیم. با توجه به خروجی یادگیری عمیق، آن دسته از گره‌ها که قدر مطلق خطای بازسازی آنها، نزدیک به یکدیگر است در یک انجمن قرار می‌گیرند. البته هر خوشه‌بند ملاحظاتی در مورد پارامترها و نحوه کار خود دارد که باید آنها را در نظر گرفت.

[۱] Probabilistic Co-Occurrence

۴-۲-۲ توکاری گراف

پیشگامان عرصه توکاری گراف پیشرفت‌هایی که در حوزه مدل‌سازی زبان (به عنوان رشته‌هایی از کلمات) رخ داده بود را به مقوله گراف (به عنوان رشته‌هایی از گره‌ها) تعمیم دادند. کلمات و پیکره[۱] ورودیهای اصلی هر الگوریتم مدل‌سازی زبان را تشکیل می‌دهند. مدل‌سازی زبان تلاش می‌کند تا احتمال وقوع پیاپی سلسله‌ای از کلمات را در یک متن پیکره تخمین بزند. اگر رشته‌ای از کلمات را مطابق رابطه ۴-۱ تعریف کنیم :

$$W_1^n = (w_1, w_2, \dots, w_n) \quad (۱-۴)$$

آنگاه هدف مدلسازی زبان این است که احتمال موجود در رابطه ۴-۲ را در سراسر متن پیکره به حداکثر برساند:

$$\Pr(w_n \mid w_1, w_2, \dots, w_{n-1}) \quad (۲-۴)$$

به شکل مشابه، می‌توانیم احتمال ملاقات گره v_i را در یک قدم زدن تصادفی[۲] به فرم زیر تعریف کنیم:

$$\Pr(v_i \mid v_1, v_2, \dots, v_{n-1}) \quad (۳-۴)$$

الگوریتم پیشنهادی ما برای اجتناب از فرآیند نمونه برداری زمان بر الگوریتم DeepWalk ماتریسی به نام ماتریس توارد یا هم رخدادی احتمالی یا PCO[۳] را مستقیماً از گراف استخراج می‌کند. این ماتریس یک ماتریس مربعی است که ابعاد آن برابر تعداد کلمات گراف می‌باشد. درایه مربوط به هر زوج از کلمات در ماتریس مذکور، بیانگر تعداد دفعاتی است که این کلمات در کنار هم دیده شده‌اند.

---

[۱] Corpus

[۲] Random Walking

[۳] Probabilistic Co-Occurrence

برای رفع مشکل تعیین مقادیر ابر پارامترها از روشی برای پیمایش تصادفی گراف استفاده می‌کنیم که از الگوریتم PageRank گوگل برداشت شده است. در ابتدای این روش همه رئوس به ترتیبی تصادفی مرتب می‌شوند. فرض می‌کنیم که راس فعلی راس iام است و ماتریسی به نام T وجود دارد که احتمال انتقال از هر یک از رئوس گراف به راس‌های دیگر در آن مشخص شده است. این ماتریس در ابتدا با توجه به محتویات ماتریس مجاورت (یا مشابهت) گراف، مقداردهی اولیه می‌شود.

هر سطر از ماتریس PCO برداری به شکل p_k است. مقدار درایه jام بردار p_k مشخص می‌کند که احتمال رسیدن به راس jام گراف بعد از یک پیمایش k مرحله‌ای چقدر است. مقدار اولیه این بردار یعنی p_0 در واقع یک بردار تک‌روشن[۱] است. یعنی مقدار آن به ازاء راس jام برابر یک و برای اعضا سایر گره‌ها برابر صفر است. اکنون ما یک مدل پیمایش تصادفی را با احتمال و امکان شروع مجدد[۲] در نظر می‌گیریم.

در هر مرحله از این مدل، پیمایش به احتمال α ادامه پیدا کرده و به گره بعدی منتقل می‌شود؛ و به احتمال $1-\alpha$ به گره آغازین انتقال پیدا کرده و بازنشانی خواهد گردید. بدین ترتیب احتمال رسیدن به هر یک از گره های گراف بعد از k مرحله انتقال بر اساس رابطه بازگشتی زیر محاسبه می‌شود:

$$p_k = \alpha . p_{k-1} . T + (1-\alpha) . p_0 \qquad (۴-۴)$$

چنانکه گفته شد بردار p_0 در واقع بردار تک روشن گره i است. در صورتی که بازنشانی در این روند وجود نداشته باشد $(\alpha = 1)$، سطرهای ماتریس PCO از رابطه ۴-۵ محاسبه خواهند شد:

$$p_k^* \triangleq p_k|_{\alpha=1} = p_{k-1}^* . T = p_0 . T^k \qquad (۵-۴)$$

پس از ایجاد ماتریس PCO بازنمایی راس iام گراف، یعنی r، بر اساس رابطه ۴-۶ ارائه می گردد:

$$r = \sum_{k=1}^{K} f(k) . p_k^* \qquad (۶-۴)$$

در این رابطه $f(.)$ یک تابع وزن‌دهی نزولی است $(f(t-1) > f(t))$. وزنی که این تابع به هر یک از گره‌های زمینه‌ای راس i اختصاص می‌دهد، تابع فاصله این گره از راس مذکور می‌باشد. در پیاده‌سازی روش

---

[۱] One-Hot

[۲] Restart

پیشنهادی از تابع $f(x) = {}^{1}/_{x}$ که برای وزن دهی در اینگونه موارد بسیار معمول و متداول است ،استفاده کرده‌ایم [۸۶].

در مدل سازی زبان برای تخمین احتمال وقوع لغت w -که درصدد ارائه آن هستیم - از زمینه[۱] این لغت یعنی c استفاده می‌کنند. لغات زمینه به لغت‌هایی اطلاق می شود که در جملات متفاوت در اطراف لغت مورد نظر مشاهده می گردند. در صورتی که ما اندازه پنجره مورد بررسی را برابر k در نظر بگیریم، هر یک از جملات باید حاوی k لغت قبل و k لغت بعد از واژه w باشند.

با ایده از این روش ماتریس PCO را برای محاسبه ماتریس اطلاعات متقابل نقطه‌ای یا PMI[۲] مورد استفاده قرار داده‌ایم. این ماتریس را می‌توان به عنوان حاصل ضرب ماتریس ارائه لغات و ماتریس زمینه در نظر گرفت. فرمول ریاضی ماتریس PMI در رابطه ۴-۷ نمایش داده شده است [۸۷]:

$$\mathrm{PMI}_{w,c} = \log\left(\frac{\#(w,c) \times |L|}{\#(w) \times \#(c)}\right) \qquad (۴\text{-}۷)$$

که در این رابطه مولفه $|L|$ براساس رابطه ۴-۸ محاسبه می‌شود:

$$|L| = \sum_{w}\sum_{c} \#(w,c) \qquad (۴\text{-}۸)$$

با انتقال همین مفاهیم به حوزه توکاری گراف، رابطه ۴-۹ برای محاسبه ماتریس PMI پیشنهاد می‌شود. در این رابطه v_i به گره جاری اشاره می‌کند و W_{v_i} معرف راس‌هایی است که در پیمایش‌های تصادفی از مبدا v_i مشاهده شده‌اند:

$$\mathrm{PMI}_{v_i,w_{v_i}} = \log\left(\frac{\#(v_i,w_{v_i}) \times |L|}{\#(v_i) \times \#(w_{v_i})}\right) \qquad (۴\text{-}۹)$$

برای افزایش کارایی روش پیشنهادی به جای ماتریس PMI از نسخه مثبت آن یعنی PPMI[۳] استفاده کرده‌ایم. یعنی برای محاسبه این ماتریس تمام درایه های منفی ماتریس PMI با صفر جایگزین شده‌اند:

$$\mathrm{PPMI}_{v_i,w_{v_i}} = \max(\mathrm{PMI}_{v_i,w_{v_i}}, 0) \qquad (۴\text{-}۱۰)$$

---

[۱] Context

[۲] Pointwise Mutual Information

[۳] Positive PMI

هر سطر ماتریس PPMI اوضاع عمومی همسایگی گره مربوطه را بیان می‌کند. در واقع برای هر دو گره دلخواه، درایه مندرج در ماتریس یادشده به این پرسش پاسخ می‌دهد که آیا این دو گره بیشتر در اطراف یکدیگر واقع می‌شوند یا از هم بیگانه و دور هستند.

روش Deep walk برای یادگیری توکاری گراف، به جای استفاده از ماتریس‌های پیشنهادی ما از پیمایش تصادفی و Skip-gram استفاده می‌کند. Skip-gram که در روش Word2vec به کار گرفته شده تلاش داردتا برای کلمات مشابه بردارهای مشابهی را براساس زمینه این کلمات وضع کند. نتایج تجربیات نشان می‌دهد که روش پیشنهادی ما از این روش بهتر عمل می‌کند.

در برخی از روش‌های متن‌کاوی پس از محاسبه ماتریس PPMI روش تجزیه مقادیر تکین یا SVD[۱] برای کاهش ابعاد این ماتریس برای یک متن پیکره به کار گرفته شده است [۸۸]. این روش، ماتریس ورودی خود را به سه ماتریس شامل دو ماتریس متعامد و یک ماتریس قطری تجزیه می‌کند. SVD روش قدرتمندی محسوب می شود اما تنها روابط خطی بین دو بردار را لحاظ می‌نماید. ما برای پوشش روابط غیرخطی موجود در میان ماتریس‌های مورداستفاده، به جای روش یادشده از شبکه‌های عصبی عمیق استفاده کرده‌ایم. روش GraRep از SVD در بازنمایی گراف استفاده کرده است [۸۹].

۴-۲-۳ معماری یادگیری عمیق

بکارگیری کدکننده‌های خودکار در کاربردهای گراف بی‌سابقه نیست. مثلاً [41] ماتریس لاپلاسی[۲] یک گراف را به عنوان ورودی در اختیار یک کدکننده خودکار گذاشته و بدین وسیله گراف مربوطه را خوشه‌بندی کرده است. ما ماتریس PPMI مرحله قبل را به عنوان ورودی به یک کدکننده خودکار پشته‌ای می‌دهیم. هر یک از گره‌های گراف در این ماتریس به شکل یک بردار $n \times 1$ نمایش داده شده

[۱] Singular Value Decomposition

[۲] Laplacian Matrix

است. این ماتریس در اختیار کدکننده خودکار قرار می‌گیرد تا در مرحله کدگذاری ابعاد بردار مذکور به $1 \times d$ $(d \ll n)$ کاهش پیدا کند.

در مرحله دوم کار، از نمایش بدست‌آمده به عنوان ورودی یادگیری عمیق استفاده می‌کنیم. به منظور حل مسئله توسط مدل شبکه عصبی، نیاز است تا وزن‌های مناسب برای هر لایه تعیین شوند. پس از تعیین پارامترهای مدل (تعداد لایه‌های پنهان، تعداد نورون‌های هر لایه، نرخ یادگیری، تابع فعال‌سازی، ...) وزن‌های بهینه مرتبط با هر لایه به کمک دادگان معین‌شده برای آموزش، محاسبه می شوند.

معیار دقت کدکننده خودکار خطای بازسازی است. این خطا که بیانگر اختلاف داده‌های اصلی (ورودی) و داده‌های بازسازی‌شده (خروجی) می‌باشد در رابطه ۴-۱۱ نشان داده شده است:

$$\sum_{i} L\left(x^{(i)}, g_{\theta_2}\left(f_{\theta_1}\left(x^{(i)}\right)\right)\right) \qquad (۱۱-۴)$$

در این رابطه، L یکی از توابع هزینه معمول مثل آنتروپی متقابل[۱] یا خطای میانگین مربعات[۲] است. هدف کلی در فرایند آموزش کننده خودکار کمینه‌کردن خطای بازسازی از طریق تعیین مقادیر مناسب برای θ_1 و θ_2 میباشد. البته برای داشتن یک کدکننده خودکار خوب و رسیدن به یک بازسازی دقیق علاوه بر مقادیر پارامترهای یادشده، ابرپارامترهایی مثل تعداد لایه‌ها و تعداد نورونهای هر لایه نیز باید به شکل مناسبی تنظیم و مقداردهی بشوند. متاسفانه برای تعیین مقادیر این ابرپارامترها هیچ قاعده عمومی و دقیقی وجود ندارد و مقداردهی آنها معمولاً با توجه به نوع مسئله و داده‌های مورد نظر و پس از پیاده سازی با سعی و خطا صورت می‌پذیرد. قدر مسلم این که تعداد نورون‌های لایه‌های ورودی و خروجی برابر با ابعاد داده‌های ورودی است؛ و صرف نظر از تعداد لایه‌ها و تعداد نورون‌های هر لایه، بخش‌های کدگذار و کدگشا معمولاً به صورت قرینه تعریف می‌شوند.

---

[۱] Cross Entropy

[۲] MSE

همچنین برای تحقق روش انتشار پسگرد[۱] شیوه گرادیان کاهشی تصادفی یا SGD[۲] را اختیار کرده‌ایم. در پیاده سازی روش مذکور نرخ یادگیری از مقدار ۰/۱ آغاز شده و به تدریج در خلال فرآیند آموزش کاهش پیدا می‌کند. همچنین از شیوه حذف تصادفی[۳] با نرخ حذف ۰/۱ استفاده کرده‌ایم. حذف تصادفی یک روش تنظیم مدل برای کاهش بیش برازش[۴] در شبکه های عصبی است. در این روش درصدی از وزنهای شبکه عصبی هر بار به صورت تصادفی از روند یادگیری حذف می‌شود. برنامه‌های مورد نیاز در زبان پایتون و با استفاده از کتابخانه یادگیری عمیق Keras نوشته شده‌اند.

تابع تحریک برای تمام لایه‌ها بجز لایه آخر، تابع ReLU[۵] است که برای هر x به فرم max(0,x) تعریف می‌شود. برای تابع تحریک لایه آخر از Softmax استفاده می‌شود. نرخ یادگیری برابر با ۰,۰۱ درنظر گرفته شده است. ما در پیاده سازی روش پیشنهادی به جای شروع از وزن‌های تصادفی روش گلوروت[۶] (یا خاویر[۷]) [۹۰] را به کار گرفته‌ایم. این روش مقادیر اولیه وزن‌ها را تعیین می‌کند و در بسیاری از کاربردهای شبکه‌های عصبی کارایی خوبی را از خود نشان داده است. از گرادیان کاهشی تصادفی[۸] به عنوان تابع بهینه‌ساز شبکه عمیق برای آموزش استفاده می‌کنیم. همچنین نقطه آغازین برای شروع تنظیم وزن‌ها نیز مهم است. زیرا اگر تنظیمات آغازین شبکه به اندازه کافی مناسب نباشند، ممکن است بهینه‌سازی وزن‌ها در یک بهینه محلی متوقف شود.

---

۱ Backpropagation
۲ Stochastic Gradient Descent
۳ Dropout
۴ Overfitting
۵ Rectified Linear Unit (ReLU)
۶ Glorot
۷ Xavier
۸ Stochastic Gradient Descent (SGD)

۴-۲-۴ خوشه‌بندی همپوشان

خروجی واحد کدگذار کدکننده خودکار یعنی همان بازنمایی پنهان به عنوان ورودی در اختیار الگوریتم OKM گذاشته شده است. این الگوریتم گسترشی از الگوریتم مشهور Kmeans به شمار می‌آید. گسترشی که برخلاف نسخه اصلی خود داده‌های ورودی را به خوشه‌های همپوشان جدا می‌کند [۹۱]. از آنجا که بازنمایی پنهان کدکننده خودکار در واقع معرف گراف مورد بررسی است، خوشه‌بندی آن به خوشه‌های همپوشان معادل با تقسیم گراف یادشده به انجمن‌های همپوشان می‌باشد.

اگر $\chi = \{x_i\}_{i=1}^{n}$ دادگانی در $\mathbb{R}^p$ باشد، تابع هدفی که الگوریتم OKM برای کمینه‌کردن آن تلاش می‌کند را می‌توان با رابطه ۴-۱۲ بیان کرد:

$$Q\left(\{\pi_c\}_{c=1}^{k}\right) = \sum_{x_i \in \chi} \|x_i - \phi(x_i)\|^2 \qquad (۱۲-۴)$$

در این رابطه $\{\pi_c\}_{c=1}^{k}$ بیانگر k خوشه همپوشان است و $\phi(x_i)$ معرف و نماینده جایگاه x_i در شمای خوشه‌بندی می‌باشد و تصویر x_i نامیده می‌شود. تصویر x_i از ترکیب مراکز خوشه‌هایی که x_i به آن‌ها تعلق دارد حاصل می‌شود. ما هر یک از این مراکز را m_c می‌نامیم و مجموعه آن‌ها را با $\{m_c\}_{c=1}^{k}$ نشان می‌دهیم.

پیشنهاد الگوریتم برای تعریف تصویر یادشده به شرح رابطه زیر است:

$$\phi(x_i) = \frac{\sum_{m_c \in A_i} m_c}{|A_i|} \qquad (۱۳-۴)$$

در رابطه فوق $A_i = \{m_c \,|x_i \in \pi_c\}$ مجموعه مراکز خوشه‌هایی است که x_i در آن‌ها ظاهر می‌شود. چارچوب کلی الگوریتم OKM و تابع Assign در آن، در شکل ۴-۲ نشان داده‌شده است.

در این الگوریتم مثل الگوریتم *Kmeans* تعداد خوشه‌ها باید از پیش به‌وسیله کاربر تعیین‌شده باشد. الگوریتم با انتخاب K نقطه تصادفی به‌عنوان مراکز اولیه خوشه‌ها آغاز می‌شود. سپس فاصله نقاط داده با این مراکز ارزیابی‌شده و هر نقطه به نزدیک‌ترین خوشه منتسب می‌گردد. بعد از این، الگوریتم وارد یک فرآیند تکراری می‌شود، در هر تکرار مراکز خوشه‌ها بر اساس قواعدی مجدداً تعیین می‌گردند و نقاط

داده‌ای با توجه به فاصله‌ای که از این مراکز جدید دارند در خوشه قبلی خود باقی‌مانده یا به نزدیک‌ترین خوشه انتقال پیدا می‌کنند. این کار تا جایی ادامه می‌یابد که مجموع مربعات فواصل نقاط از مراکز دیگر کمتر نشود و یا هیچ جابجایی جدیدی در میان خوشه‌ها انجام نپذیرد.

Initialisation: t=0
choose arbitrary k centers $Z^t = \{z_1^t, z_2^t, \ldots, z_k^t\}$ from X,
For each $x_i \in X$: Assign(x_i,Z^t) (*build* $W_{.,i}^t$),
Build a first coverage $\mathcal{R}^t = (W^t, Z^t)$.
Do
t=t+1
- Update(Z^{t-1},W^{t-1}) (*build* Z^t),
- For each $x_i \in X$: Assign(x_i,Z^t) (*build* $W_{.,i}^t$),

While ($W^t \neq W^{t-1}$ or $\tilde{Q}(W^{t-1}, Z^{t-1}) - \tilde{Q}(W^t, Z^t) < \epsilon$)

Assign(x_i,Z):
Initialisation :
Let z^* be the nearest center from x_i in Z ($\forall z_j \in Z,\ d(x_i, z^*) \leq d(x_i, z_j)$):
$A = \{z^*\}$ (with A the list of assignments for x_i),
$Z = Z \setminus \{z^*\}$.
Do
Let $\overline{x_i}^A$ denoting the center of gravity of A:
Let z^* be the nearest center from x_i in Z,
if $d(x_i, \overline{x_i}^{A \cup \{z^*\}}) < d(x_i, \overline{x_i}^A)$ then $A \leftarrow A \cup \{z^*\}$ and $Z = Z \setminus \{z^*\}$
While a new assignment is performed
Final decision:
Let A' be the old assignments for x_i,
if $d(x_i, \overline{x_i}^A) < d(x_i, \overline{x_i}^{A'})$ then assign x_i to the centers from A,
else keep the old assignment A'.

شکل ۴-۲: چارچوب کلی الگوریتم خوشه‌بندی همپوشان و تابع نسبت‌دهی آن [۹۱]

تفاوت اصلی الگوریتم *OKM* با الگوریتم *Kmeans* در این است که *OKM* هر نقطه را به یک یا چند خوشه نسبت می‌دهد و خوشه‌های همپوشان ایجاد می‌کند. به‌علاوه *OKM* برای تعیین مراکز جدید از روابط متفاوتی استفاده می‌کند.

برای نمایش مفهوم تصویر x_i یا $\phi(x_i)$ از نماد $\overline{x_i}$ استفاده‌شده است. مجموعه $Z = \{z_1, z_2, \ldots, z_k\}$ حاوی نقاط مرکزی خوشه‌هاست و x_i داده‌ای است که باید به یک یا چند خوشه نسبت داده شود. کد

برای تعیین خوشه‌های داده x_i، مجموعه Z را از نزدیک‌ترین تا دورترین مرکز به x_i پیمایش می‌کند و هر مرکز را تنها به شرطی به تصویر x_i می‌افزاید که اضافه کردن آن مرکز موجب بهبود تصویر x_i بشود، یعنی به شرطی که فاصله x_i با تصویرش یا $d(x_i, \overline{x}_i)$ کاهش پیدا کند. در پایان کار انتسابات جدید به شرطی حفظ خواهند شد که بهتر از انتسابات قبلی باشند. این موضوع ضمانت می‌کند که مقدار تابع هدف در اثر فرایند انتساب افزایش پیدا نکرده و الگوریتم واگرا نخواهد شد.

الگوریتم OKM در گام بعدی مراکز جدیدی را برای خوشه‌ها تعیین می‌کند. در این گام مرکز خوشه R_j که آن را z_j می‌نامیم بر اساس رابطه زیر به‌هنگام می‌شود :

$$z_{j,v} = \frac{1}{\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} \times \hat{x}_{i_v}^j \qquad (۱۴-۴)$$

در این رابطه $z_{j,v}$ معرف vامین مؤلفه بردار z_j است، δ_i تعداد خوشه‌هایی است که داده x_i به آن‌ها تعلق دارد و $\hat{x}_{i_v}^j$ معرف مؤلفه v ام بردار مرکز مطلوب خوشه R_j برای x_i است. مرکز مطلوب مرکزی است که بر اساس آن فاصله x_i و تصویرش $\overline{x}_i$ صفر هست، یعنی $d(x_i, \overline{x}_i) = 0$. این مرکز مطلوب از رابطه زیر به دست می‌آید:

$$\hat{x}_{i_v}^j = \delta_i \times x_{i,v} - (\delta_i - 1) \times \overline{x}_{i,v}^{A-\{z_j\}} \qquad (۱۵-۴)$$

در رابطه بالا A مجموعه خوشه‌هایی است که x_i به آن‌ها منتسب شده است.

برای ارائه یک تعریف شهودی‌تر از مقدار جدید z_j برای خوشه R_j باید توجه داشت که z_j درواقع مرکز خوشه زیر است :

$$\hat{R}_j = \left\{ \left(\hat{x}_i^j , \frac{1}{\delta^2} \right) \mid x_i \in R_j \right\} \qquad (۱۶-۴)$$

این خوشه مجموعه نقاط ایده‌آلی است که براساس تعداد خوشه‌هایشان وزن‌دهی شده‌اند. هرچه تعداد خوشه‌های یک نقطه بیشتر باشد، آن نقطه در تعیین مرکز جدید خوشه R_j نقش ضعیف‌تری را ایفا

می‌کند. می‌توان ثابت کرد که این شیوه بهنگام‌سازی نه‌تنها باعث کاهش تابع هدف می‌شود، بلکه امکان همگرا شدن این تابع به نقاط کمینه خود را نیز فراهم می‌کند.

الگوریتم OKM پشتوانه نظری خوبی دارد و جنبه‌های کاربردی آن مثل زمان اجرا و دقت نتایج حاصله نیز قابل‌قبول می‌باشد.

فصل ۵ : نتایج حاصل

۵-۱ مقدمه

در این فصل، پس از معرفی چند معیار ارزیابی کارایی انجمن‌یابی، نتایج حاصل از اعمال روش پیشنهادی روی دادگانهای متنوع با نتایج دیگر روش‌های متداول در انجمن‌یابی مقایسه شده و مورد بحث و بررسی تفصیلی قرار می‌گیرد.

۵-۲ معیارهای ارزیابی کارایی

معیارهای زیادی در بررسی عملکرد روش‌های انجمن‌یابی وجود دارند. برخی از این معیارها فقط در حالت ناهمپوشان و برخی فقط در حالت همپوشان استفاده می‌شوند. گروهی از معیارها، قابلیت استفاده در هر دو حالت را دارند.

یک معیار ساده، تعداد انجمن‌های یافته‌شده توسط روش انجمن‌یابی است که با |c| بیان می‌شود. به غیر از مواردی که تعداد انجمن‌ها به‌عنوان یک پارامتر ورودی مسئله داده می‌شود، یافتن مقدار بهینه (یا مناسب) برای |c| به‌شدت به تعریف مسئله و توپولوژی گراف وابسته است.

معیار Flake-ODF [۹۲] برآورد می‌کند که چه کسری از گره‌های انجمن S، دارای تعداد یال‌های داخلی کمتر از یال‌های مرزی هستند (یال‌های داخلی آن گره از نصف درجه آن گره کمتر است). حالت مطلوب، در کمتر شدن مقدار این معیار است. این کاهش به معنای وجود شباهت بیشتر در ساختار گره‌های در حال بررسی با ساختار یک انجمن است. این معیار در رابطه ۵-۱ برای انجمن S با n_S گره، تعریف شده است:

$$\text{Flake}-\text{ODF}(S) = \frac{\left|\{u : u \in S, |\{(u,v) : v \in S\}| < \deg(u)/2\}\right|}{n_S} \qquad (۱-۵)$$

منظور از deg (u) درجه رأس u است. مقدار این معیار برای کل گراف نیز قابل محاسبه است (با متوسط‌گیری روی همه انجمن‌ها).

معیار کندوکتانس [۹۲] نقش یال‌های مرزی یک انجمن (حاصل از الگوریتم انجمن‌یابی) را نشان می‌دهد. مطابق رابطه زیر، تعداد یال‌های داخلی انجمن S (تعلق هر دو رأس به S) با m_S بیان می‌شود:

$$m_S = \left|\{(u,v) : u \in S, v \in S\}\right| \qquad (۲-۵)$$

تعداد یال‌های مرزی انجمن S (یالی که فقط یک رأس آن به انجمن S تعلق دارد) مطابق رابطه (۱۰) با c_S نشان داده شده است:

$$c_S = \left|\{(u,v) : u \in S, v \notin S\}\right| \qquad (۳-۵)$$

با توجه به روابط بالا کندوکتانس مطابق رابطه (۵-۴) محاسبه می‌شود:

$$cond(S) = \frac{c_S}{2m_S + c_S} \qquad (۴-۵)$$

این معیار برای کل گراف هم به‌سادگی با متوسط‌گیری روی تمام انجمن‌ها، قابل محاسبه است. کمتر بودن مقدار این معیار، معادل با ساختار انجمنی قوی‌تر و عملکرد مطلوب‌تر در روش مورد استفاده است.

معیار «نسبت همپوشانی»[۱] [۱۶] مطابق رابطه (۱۲) روی گراف تعریف می‌شود:

$$ov - ratio = \frac{1}{n}\sum_{i=1}^{n}\left|u_i \in S, S \in G\right| \qquad (۵-۵)$$

برای هر رأس u_i، تعداد انجمن‌هایی که به آنها تعلق دارد مشخص می‌شود و متوسط‌گیری روی مقادیر رئوس گراف انجام خواهد شد.

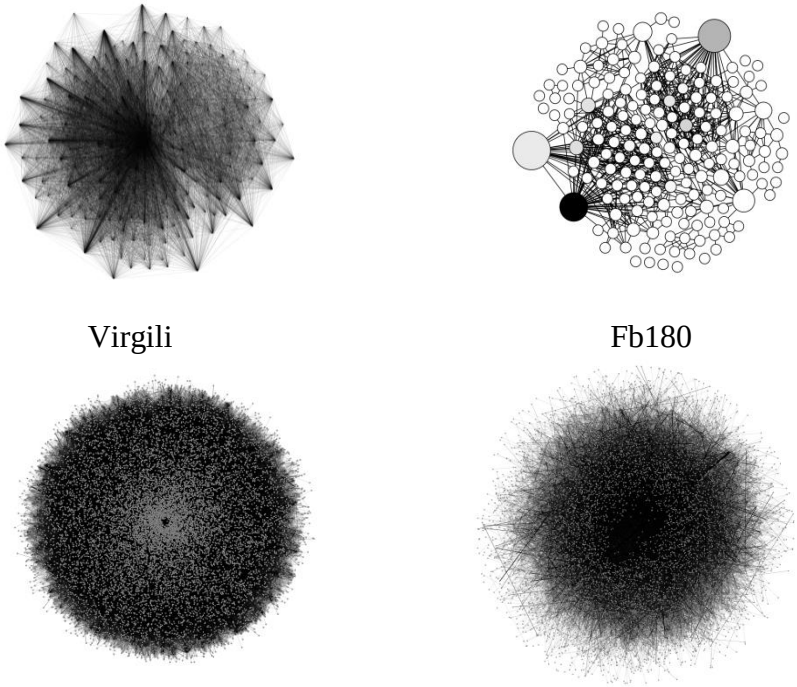
در محاسبه معیار ماژولاریتی (پیمانه‌ای) برای انجمن‌های همپوشان از رابطه۲-۱و برای انجمن‌های ناهمپوشان از رابطه۲-۱۰ استفاده شده است.

۵-۳ دادگان‌های مورد استفاده

همانطور که در فصل اول هم اشاره شد شبکه اجتماعی شکل خاصی از شبکه‌های پیچیده و مستقل از مقیاس است. عمده مقالات و تحقیق‌ها علاوه بر شبکه‌های اجتماعی روی دیگر انواع شبکه‌ها هم تست

[۱] Overlap Ratio

می‌شوند. از هر سه دسته ذکرشده دادگان انتخابی داشته‌ایم. اولین دادگان با نام Fb180، گراف دوستان یک کاربر شبکه فیس‌بوک بوده که با نرم‌افزار Netvizz از این شبکه استخراج شده است. دادگان دوم، یک شبکه ارتباطی میان گروهی از دانشجویان دانشگاه Virgili در کشور اسپانیا و دادگان سوم مربوط به پیوندهای بین کاربران در سایت hamsterster.com می‌باشد. هر دو دادگان از آرشیو پایگاه KONECT انتخاب شده‌اند. نهایتاً دادگان چهارم به نام Astro_ph، همکاری‌های علمی گروهی از اخترفیزیکدانان (ثبت‌شده در سایت arXiv) را نشان می‌دهد. این دادگان از پایگاه SNAP انتخاب شده است.



شکل ۵-۱: گراف‌های متناظر با دادگان‌ها

در انتخاب دادگان‌ها، هم از دادگان تولیدشده توسط نویسندگان مقاله و هم از دادگان استاندارد و پرکاربرد، استفاده کرده‌ایم. شکل ۵-۱ یک تجسم از این دادگان‌ها را نمایش می‌دهد.

جدول ۵-۱: مشخصات دادگان‌ها

دادگان	n	m	متوسط درجه یک گره	بیشینه درجه یک گره	تعداد مؤلفه‌های همبند	n_{GCC}	m_{GCC}	چگالی گراف	قطر گراف	APL
Fb180	۱۸۰	۷۵۴	۸,۳۷	۵۴	۱۳	۱۵۲	۷۲۷	۰,۰۴۶۸۰۳	۸	۳,۲۴
Virgili	۱۱۳۳	۵۴۵۱	۹,۶۲	۷۱	۱	۱۱۳۳	۵۴۵۱	۰,۰۰۸۵۰۰	۸	۳,۶۵
Hamsterster	۲۴۲۶	۱۶۶۳۱	۱۳,۷۱	۲۷۳	۶۹	۲۰۰۰	۱۵۷۲۳	۰,۰۰۵۶۵۳	۱۰	۳,۶۷
Astro_ph	۱۸۷۷۱	۱۹۸۰۵۰	۲۱,۱۰	۵۰۴	۵۳	۱۷۹۰۳	۱۸۹۷۵۷	۰,۰۰۱۱۲۴	۱۴	۴,۱۷

کسب اطلاعات اولیه درباره شبکه‌های کوچک، با نگاه به گراف آنها امکان‌پذیر است. مثلاً برای دادگان اول در شکل ۵-۱، اندازه هر گره متناسب با بینابینی آن و شدت رنگ هر گره متناسب با درجه آن انتخاب شده است. مشخص است که برای این دادگان، گره دارای بیشترین درجه الزاماً همان گره دارای بیشترین بینابینی نیست. جدول ۵-۱ مشخصات اولیه هر یک از دادگان‌ها و مقادیر پارامترهای مهم گراف متناظر آنها را بیان می‌کند. توپولوژی یک گراف وابسته به ماهیت و نحوه تشکیل آن بر اساس ارتباطات تعریف‌شده میان موجودیت‌هایش است. لذا ممکن است یک گراف بزرگ، یکپارچه و همبند بو ده و فقط از یک (یا چند مؤلفه معدود) تشکیل شده باشد و در مقابل، گرافی با m و n کمتر، تعداد مؤلفه‌های همبند بیشتری داشته باشد. برای چهار دادگان انتخابی، تعداد مؤلفه‌های همبند از ۱ تا ۶۹ متغیر است.

در گراف‌های ناهمبند، بسیاری از پردازش‌های گراف روی «بزرگ‌ترین مؤلفه همبند»[۱] انجام می‌شود. نظر به اهمیت این مؤلفه، تعداد رئوس (n_{GCC}) و تعداد یال‌های آن (m_{GCC}) برای دادگان‌ها در جدول ۵-۱ ذکر شده است. برای یک گراف با n رأس و m یال، چگالی یا میزان تراکم گراف که بیانگر نسبت یال‌های موجود به یال‌های ممکن می‌باشد، با $2m/(n(n-1))$ بیان می‌شود.

طول مسیر مابین دو گره، با حداقل تعداد یال‌های ممکن میان آن دو گره مشخص می‌شود. مفهوم «قطر گراف» بیانگر مقدار بیشینه در طول مسیر بین دو گره از گراف است و مفهوم APL[۲] به

[۱] Greatest (Giant) Connected Component (GCC)

[۲] Average Path Length (APL)

متوسط‌گیری روی تمام این طول مسیرها اشاره دارد. در گراف‌های با چند مؤلفه همبند، مقادیر قطر و APL روی بزرگ‌ترین مؤلفه همبند گراف تعیین می‌شوند. با توجه به مقادیر APL در جدول (۲)، دادگان‌های انتخابی جزو شبکه‌های جهان کوچک هستند.

۵-۴ بررسی نتایج روی دادگان دنیای واقعی و استخراجی

این بخش به توضیحات مربوط به پیاده‌سازی چند روش انجمن‌یابی روی دادگان‌های معرفی‌شده و مقایسه عملکرد روش‌ها با یکدیگر اختصاص دارد. عملکرد روشها روی دادگان دنیای واقعی و گراف استخراجی در جدول ۵-۲ بیان شده است.

برای یکسان‌سازی در مقایسه روش‌ها، پیاده‌سازی‌ها روی حالت بدون‌وزن و غیرجهت‌دار گراف دادگان‌های انتخابی انجام شده است. مکانیزم پیاده‌سازی COPRA[۱] و پیاده‌سازی CPM[۲] با زبان جاوا و پیاده‌سازی LFM[۳] با زبان C++ (بهره‌گیری از کدهای منتشرشده توسط نویسندگان مقالات) انجام شده است.

پیاده‌سازی الگوریتم دیگر روش‌ها، با زبان پایتون (استفاده از ماژول Networkx آن) انجام شد. البته با توجه به غیرقطعی‌بودن روش‌های LPA و COPRA، نتیجه نهایی آنها از متوسط‌گیری روی ده بار اجرا محاسبه شده است. مقدار k در روش‌های ساختارگرای CPM و GCE برابر با ۵ در نظر گرفته شده و پارامتر γ رابطه (۵) روش LFM برابر با ۰٫۸ انتخاب شده است. ضمناً به دلیل ناهمگونی در پلتفرم‌ها و زبان‌های پیاده‌سازی روش‌ها، زمان اجرا را در مقایسه عملکرد روش‌ها لحاظ نکرده‌ایم.

[۱] http://www.cs.bris.ac.uk/~steve/networks/

[۲] http://www.cfinder.org

[۳] http://sites.google.com/site/andrealancichinetti/software

جدول ۵-۲: عملکرد الگوریتم‌های انجمن‌یابی روی دادگان دنیای واقعی و استخراجی

	Astro_ph				Hamsterster				Virgili				Fb180			
	Ov-ratio	Conductance	Flake-ODF	Q	Ov-ratio	Conductance	Flake-ODF	Q	Ov-ratio	Conductance	Flake-ODF	Q	Ov-ratio	Conductance	Flake-ODF	Q
GN	۱	**۰,۰۵**	۰,۱۵	۰,۱۳	۱	**۰,۰۷**	۰,۱۴	۰,۲۷	۱	**۰,۰۴**	۰,۱۱	۰,۳۸	۱	**۰,۰۳**	۰,۱۲	۰,۴۴
CONGA	۱,۵۳	۰,۲۰	۰,۲۳	۰,۲۰	۱,۸۵	۰,۲۴	۰,۲۳	۰,۳۰	۲,۰۴	۰,۱۹	۰,۲۰	۰,۳۹	۲,۴۱	۰,۱۷	۰,۲۵	۰,۴۷
LFM	۱,۵۸	۰,۱۴	۰,۱۳	**۰,۲۲**	۱,۴۵	۰,۱۳	۰,۱۶	**۰,۳۲**	۱,۳۰	۰,۱۰	۰,۱۱	۰,۴۰	۱,۲۳	۰,۰۸	۰,۱۰	۰,۴۸
LPA	۱	۰,۰۷	۰,۱۴	۰,۱۵	۱	۰,۱۰	۰,۱۶	۰,۲۹	۱	۰,۰۶	۰,۰۷	۰,۳۵	۱	۰,۰۴	۰,۰۵	۰,۴۴
COPRA	۱,۷۴	۰,۱۷	۰,۲۳	**۰,۲۲**	۲,۳۱	۰,۲۰	۰,۲۰	۰,۳۱	۲,۲۸	۰,۱۶	۰,۲۱	**۰,۴۱**	**۳,۱۴**	۰,۱۴	۰,۲۷	**۰,۴۹**
CPM	۰,۶۹	۰,۰۹	**۰,۰۷**	۰,۱۱	۰,۶۳	۰,۱۲	**۰,۰۹**	۰,۱۸	۰,۷۸	۰,۰۸	**۰,۰۵**	۰,۳۰	۰,۷۱	۰,۰۶	**۰,۰۴**	۰,۴۰
GCE	۰,۷۷	۰,۰۹	۰,۰۹	۰,۱۳	۰,۷۱	۰,۱۳	۰,۱۰	۰,۲۱	۰,۸۲	۰,۰۸	۰,۰۶	۰,۳۲	۰,۷۳	۰,۰۶	**۰,۰۴**	۰,۴۱
Infomap	۱	۰,۰۷	۰,۱۳	۰,۱۴	۱	۰,۰۹	۰,۱۱	۰,۲۵	۱	۰,۰۷	۰,۰۸	۰,۳۴	۱	۰,۰۴	۰,۱۱	۰,۴۲
LC	**۲,۰۹**	۰,۱۱	۰,۱۵	۰,۱۶	**۲,۳۵**	۰,۱۵	۰,۱۴	۰,۲۶	**۲,۳۳**	۰,۱۱	۰,۱۱	۰,۳۷	**۱۳۰**	۰,۰۹	۰,۱۰	۰,۴۴
|c|	۱۲۹۱	۰,۱۳	۰,۱۵	۰,۱۳	۱	۰,۱۴	۰,۱۴	۰,۲۷	۱۰۴۷	۰,۱۱	۰,۱۱	۰,۳۸	۴۹	۰,۱۱	۰,۱۲	۰,۴۴
Q	۰,۱۳	۰,۲۰	۰,۲۳	۰,۲۰	۰,۲۷	۰,۲۳	۰,۲۳	۰,۳۰	۰,۳۹	۰,۲۰	۰,۱۱	۰,۴۰	۰,۳۹	۰,۱۹	۰,۲۵	۰,۴۷
Flake-ODF	۰,۱۵	۰,۲۳	۰,۱۳	**۰,۲۲**	۰,۱۴	۰,۲۳	۰,۱۶	**۰,۳۲**	۰,۱۱	۰,۱۰	۰,۱۱	۰,۴۰	۰,۱۱	۰,۰۸	۰,۱۰	۰,۴۸
Conductance	**۰,۰۵**	۰,۲۰	۰,۱۴	۰,۱۵	**۰,۰۷**	۰,۲۴	۰,۱۳	۰,۱۳	۰,۱۰	۰,۱۹	**۰,۰۴**	۰,۳۸	۰,۱۱	۰,۰۸	**۰,۰۳**	۰,۱۷
Ov-ratio	۱	۱,۵۳	۱,۵۸	۱	۱	۱,۸۵	۱,۴۵	۱	۱,۳۰	۲,۰۴	۱	۱۰۴۷	۹۲۱	۵۸۲	۷۷۰	۵۱۱
|c|	۱۲۹۱	۴۰۱۷	۱۵۷۷	۹۵۱	۱۶۵۴	۷۹۲	۸۳۹	۱۰۸۵	۴۱۸۲							
Q	۰,۱۳	۰,۲۰	**۰,۲۲**	۰,۱۵	**۰,۲۲**	۰,۱۱	۰,۱۳	۰,۱۴	۰,۱۶							
Flake-ODF	۰,۱۵	۰,۲۳	۰,۱۳	۰,۱۴	۰,۲۳	**۰,۰۷**	۰,۰۹	۰,۱۳	۰,۱۵							
Conductance	**۰,۰۵**	۰,۲۰	۰,۱۴	۰,۰۷	۰,۱۷	۰,۰۹	۰,۰۹	۰,۰۷	۰,۱۱							
Ov-ratio	۱	۱,۵۳	۱,۵۸	۱	۱,۷۴	۰,۶۹	۰,۷۷	۱	**۲,۰۹**							

بهترین نتیجه به‌دست‌آمده برای هر معیار ارزیابی کارایی، به‌صورت برجسته و سایه‌دار مشخص شده است. در رابطه با |c| باید گفت که افزایش یا کاهش تعداد انجمن‌های یافته‌شده در مقایسه بین روش‌ها، الزاماً بیانگر مزیت خاصی نیست و به سطح تفکیک[1] تعریف‌شده در مطالعه گراف یک شبکه مربوط می‌شود.

در حالت کلی، انتخاب تعداد انجمن‌ها بین ۱ (انتخاب خود گراف به‌عنوان انجمن) وn (انتخاب هر رأس به‌عنوان انجمن) امکان‌پذیر است. مشخص است که هیچ‌یک از این دو حالت، اطلاعات خاصی را بیان نمی‌کند. بنا بر ماهیت عملکرد خود، روش‌های LC و CONGA عمدتاً بیشترین تعداد انجمن‌ها را یافته‌اند. کمترین مقادیر توسط روش‌های مبتنی بر ساختار نظیر CPM و GCE حاصل شده است. روش‌های ساختاری به دنبال هسته‌های چگال (کلیک‌ها) هستند تا بر پایه آنها، انجمن‌ها را بیابند. این روش‌ها، رئوس با درجه اندک را در انجمن‌یابی شرکت نمی‌دهند. لذا می‌توان برای این روش‌ها انتظار داشت که هرچه گراف مورد بررسی، چگال‌تر (درهم‌تنیده‌تر) شود تعداد انجمن‌های آن کاهش یابد.

همپوشانی عمدتاً باعث افزایش در مقدار تابع پیمانه‌ای می‌شود زیرا ترم‌های بیشتری را در محاسبه آن دخالت می‌دهد. بیشترین مقادیر این معیار برای روش‌های LFM و COPRA و کمترین مقادیر برای CPM حاصل شده است.

مقادیر مندرج برای معیارFlake-ODF ، با متوسط‌گیری از مقادیر همه انجمن‌ها محاسبه شده است. بهترین نتیجه توسط روش‌های CPM و GCE به دست آمده است. استفاده از ساختار کلیک که یک گراف کامل است سبب شده تا گره‌های انجمن‌های حاصل از این روش همپوشان، دارای پیوندهای داخلی قوی باشند. این گره‌ها عمدتاً از تعداد بیشتری یال داخلی نسبت به یال مرزی برخوردار هستند. در مقابل، دو روش همپوشانی COPRA و CONGA نتیجه خوبی برای این معیار ندارند؛ زیرا به دلیل نحوه عملکردشان، انجمن‌های زیادی تولید می‌کنند که این امر سبب افزایش تعداد یال‌های مرزی می‌شود.

[1] Granularity

بهترین نتایج برای معیار کندوکتانس توسط روش ناهمپوشان GN حاصل شده است چون بنای کار این روش بر حذف یال‌های فی‌مابین می‌باشد. می‌توان گفت روش‌های مبتنی بر مرکزگرایی روی دادگان‌هایی که به‌خوبی قابل تفکیک باشند نتایج مطلوبی می‌گیرند، چون بنای کار را بر حذف یال‌ها و نواحی کم‌تراکم می‌گذارند. البته این روش‌ها روی گراف‌های چگال و درهم‌تنیده به‌خوبی عمل نمی‌کنند.

در انجمن‌یابی، رأس با درجه صفر (رأس ایزوله) به‌عنوان انجمن مستقل در نظر گرفته شده و رأس با درجه یک، به انجمن گره همسایه‌اش اختصاص می‌یابد. بنابراین، در روش‌های ناهمپوشان، مقدار معیار نسبت همپوشانی برابر با ۱ خواهد بود زیرا هر گره فقط و فقط به یک انجمن تعلق دارد.

افزایش مقدار این معیار در روش‌های همپوشان بیانگر انعطاف روش در تخصیص رئوس به انجمن‌های متعدد است. عمدتاً بیشترین مقدار نسبت همپوشانی برای روش LC حاصل شده است. شایان ذکر است که این معیار برای CPM و GCE، مقداری کمتر از ۱ را نتیجه می‌دهد زیرا این روش‌ها، گره‌های با درجه اندک (کمتر از k-1) را جزو هیچ‌یک از انجمن‌ها دسته‌بندی نمی‌کنند. مقادیر جدول، برای k=5 به دست آمده‌اند یعنی رئوس با درجه کمتر از ۴ بررسی نمی‌شوند. اگر مقدار بزرگ‌تری انتخاب بشود، تعداد انجمن‌ها کاهش خواهد یافت.

انتخاب مقادیر کوچک برای k، افزایش تعداد انجمن‌ها را سبب شده و انتخاب مقادیر خیلی بزرگ برای k، فقط انجمن‌های بسیارچگال را نشان می‌دهد. روش CPM خصوصاً وقتی که گراف یک شبکه دارای تعداد زیادی مؤلفه کوچک باشد عملکرد خوبی ندارد.

روش‌های infomap و LPA میانگین عملکرد خوبی دارند هرچند در هیچ‌یک از معیارها اول نیستند. ضمناً برخلاف روش‌های ساختاری که در گراف‌های چگال عملکرد خوبی دارند، روش‌های مبتنی بر یال نواحی همپوشان متراکم را به‌خوبی تشخیص نداده و ممکن است ناحیه همپوشان متراکم میان دو انجمن متمایز را عملاً یک انجمن مستقل در نظر بگیرند.

۵-۵ بررسی نتایج روی محکها

محکهای متعددی با توجه به مکانیزم تولید آن که در بخش ۳-۴-۳ مطرح شده است، مورد بررسی قرار گرفته‌اند. مشخصات این محک‌ها با توجه به پارامترهای قابل تعیین به همراه پیکربندی مناسب برای شبکه یادگیری عمیق متناظر با گره‌های هر محک، در جدول ۵-۳ نشان داده شده است.

در شکل ۵-۲ نیز منحنی درجات گره‌ها برای محک‌های تولیدشده به نمایش درآمده است. مطابق این شکل، مشخص است که دادگان تولیدی بطور کامل در حیطه شبکه‌های مستقل از مقیاس قرار می‌گیرد چرا که منحنی توزیع درجات آن از نوع قانون توزیع توانی بوده و تعداد زیادی از گره‌ها از درجه اندک و تعداد کمی از گره‌های از درجه زیاد برخوردار هستند.

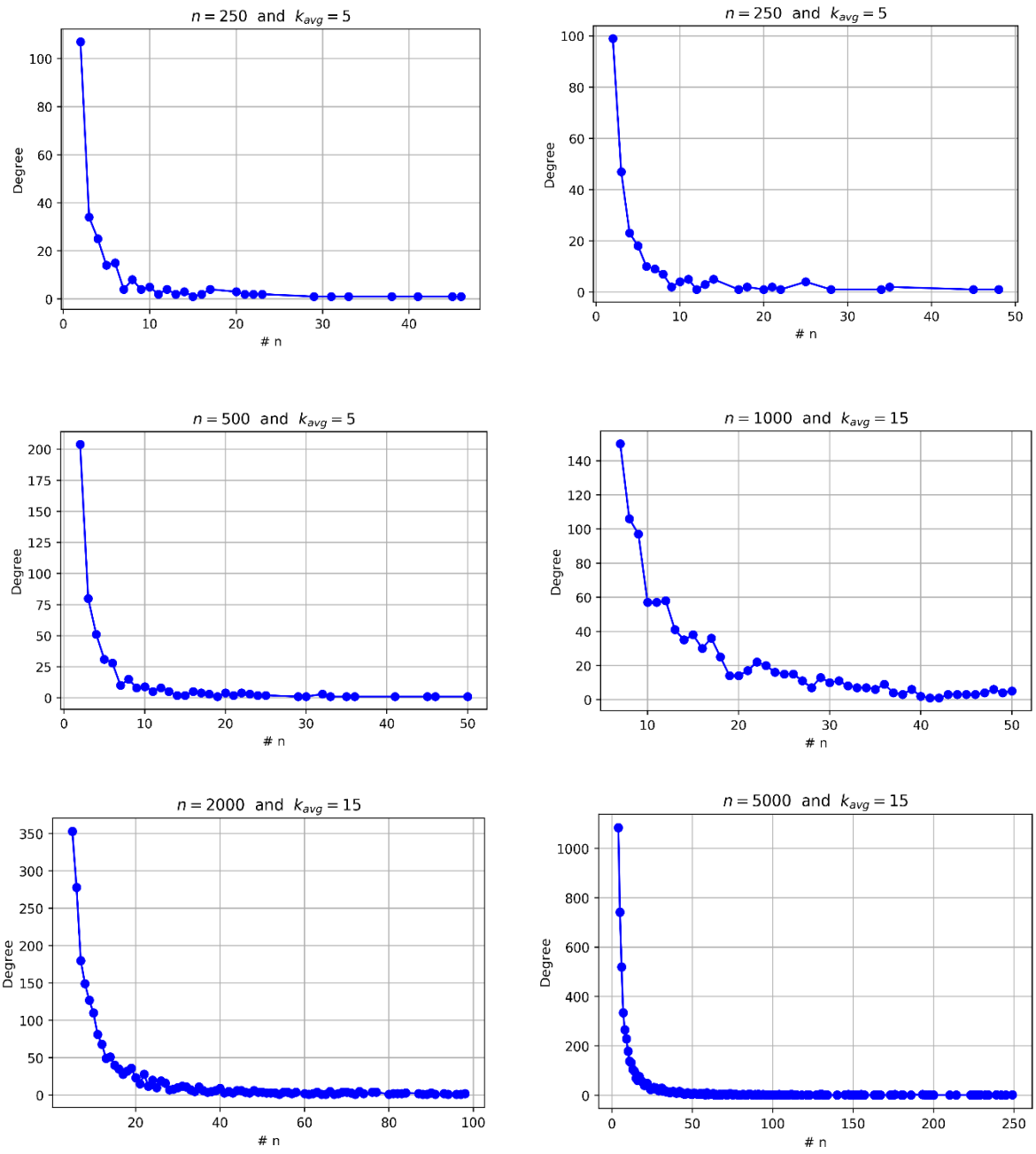
جدول ۵-۳: مشخصات و پیکربندی شبکه عمیق برای محکها

Dataset	n	m	k_{avg}	k_{max}	μ	min_C	max_C	O_m	O_n	Layer Configuration
Bench1	250	731	5	50	0.1	10	25	2	0.1	250-128-64-32
Bench2	500	1413	5	50	0.1	20	100	2	0.1	500-128-64-32
Bench3	1000	7685	10	50	0.1	20	100	2	0.1	1000-512-256-128-64
Bench4	2000	15488	10	100	0.1	25	250	3	0.2	2000-512-256-128-64
Bench5	5000	37769	15	250	0.2	50	500	4	0.3	5000-1024-512-256-128-64

نتایج حاصله با توجه به چهار معیار ارزیابی کارایی رسانایی، ماژولاریتی، NMIو F-score در جداول ۵-۴ تا ۵-۷ قید شده‌اند.

جدول ۵-۴: نتایج انجمن‌یابی با توجه به معیار F-score

	CONGA	COPRA	SLPA	GCE	LC	Deepwalk	Proposed (α=0.95)	Proposed (α=0.975)
Bench1	0.3213	0.3622	0.3757	0.2841	0.3092	0.4219	0.4710	**0.5118**
Bench2	0.3185	0.3370	0.3510	0.2380	0.2738	0.3772	0.4043	**0.4297**
Bench3	0.2957	0.3149	0.3278	0.2147	0.2409	0.3408	0.3721	**0.3818**
Bench4	0.2753	0.2792	0.2792	0.1975	0.2140	0.3087	**0.3514**	0.3379
Bench5	0.2314	0.2514	0.2541	0.1712	0.1946	0.2749	**0.3109**	0.3049



شکل ۵-۲: منحنی توزیع درجات گره‌ها در محک‌های تولیدشده

جدول ۵-۵: نتایج انجمن‌یابی با توجه به معیار NMI

	CONGA	COPRA	SLPA	GCE	LC	Deepwalk	Proposed (α=0.95)	Proposed (α=0.975)
Bench1	0.3818	0.4417	0.4713	0.3619	0.3118	0.4629	0.5012	**0.5369**
Bench2	0.2823	0.3742	0.3921	0.3654	0.2673	0.3376	0.4090	**0.4229**
Bench3	0.2492	0.2944	0.2783	0.2504	0.2138	0.2887	0.3239	**0.3498**
Bench4	0.1793	0.2286	0.2438	0.2141	0.1782	0.2381	0.2612	**0.2779**
Bench5	0.1627	0.1940	0.2073	0.1807	0.1437	0.2028	0.2340	**0.2395**

جدول ۵-۶: نتایج انجمن‌یابی با توجه به معیار ماژولاریتی

	CONGA	COPRA	SLPA	GCE	LC	Deepwalk	Proposed (α=0.95)	Proposed (α=0.975)
Advogato	0.1302	0.1275	0.1548	0.1184	0.1033	0.1670	**0.1594**	0.1562
Hamsterster	0.1608	0.1543	0.1705	0.1449	0.1501	0.1836	0.1973	**0.2028**
Virgili	0.1758	0.1743	0.1810	0.1538	0.1594	0.1877	0.2190	**0.2274**
Bench1	0.2789	0.2865	0.3105	0.2419	0.2377	0.3019	0.3217	**0.3245**
Bench2	0.2566	0.2597	0.2873	0.2187	0.2041	0.2791	0.3025	**0.3092**
Bench3	0.2271	0.2289	0.2547	0.1940	0.1897	0.2492	0.2691	**0.2751**
Bench4	0.1625	0.1586	0.1749	0.1307	0.1248	0.1704	**0.2036**	0.2004
Bench5	0.1294	0.1160	0.1672	0.1037	0.0973	0.1662	**0.1893**	0.1854

از مقادیر حاصله برای سه پایگاه داده دنیای واقعی هم بعنوان شاهد استفاده شده است تا حیطه مقایسه روشن‌تر و دقیقتر باشد.

جدول ۵-۷: نتایج انجمن‌یابی با توجه به معیار رسانایی

	CONGA	COPRA	SLPA	GCE	LC	Deepwalk	Proposed (α=0.95)	Proposed (α=0.975)
Advogato	0.0834	0.1408	0.0702	0.2148	0.2931	0.0997	0.0541	**0.0510**
Hamsterster	0.0507	0.1003	0.0615	0.1229	0.1309	0.0626	0.0692	**0.0630**
Virgili	0.0459	0.0992	0.0603	0.1107	0.1194	0.0578	0.0254	**0.0237**
Bench1	0.0094	0.0251	0.0142	0.0297	0.0309	0.0114	0.0069	**0.0051**
Bench2	0.0186	0.0493	0.0280	0.0587	0.0601	0.0338	0.0120	**0.0093**
Bench3	0.0457	0.0784	0.0457	0.0945	0.0938	0.0502	0.0314	**0.0278**
Bench4	0.0676	0.1055	0.0530	0.1395	0.1487	0.0760	**0.0480**	0.0489
Bench5	0.0910	0.1602	0.0886	0.2309	0.2411	0.1033	**0.0527**	0.0554

عمده بهتری نتایج متعلق به روش پیشنهادی با $\alpha=0.975$ است که مشخصا از تمامی روشهای کلاسیک و نیز روشی که Deepwalk برای بازنمایی گره‌ها استفاده می‌کند، بهتر عمل کرده است.

می‌توان گفت روش‌های مبتنی بر مرکزگرایی روی دادگان‌هایی که به‌خوبی قابل تفکیک باشند نتایج مطلوبی می‌گیرند، چون بنای کار را بر حذف یال‌ها و نواحی کم‌تراکم می‌گذارند. البته این روش‌ها روی گراف‌های چگال و درهم‌تنیده به‌خوبی عمل نمی‌کنند. در انجمن‌یابی، رأس با درجه صفر (رأس ایزوله) به‌عنوان انجمن مستقل در نظر گرفته شده و رأس با درجه یک، به انجمن گره همسایه‌اش اختصاص می‌یابد. بنابراین، در روش‌های ناهمپوشان، مقدار معیار نسبت همپوشانی برابر با ۱ خواهد بود زیرا هر گره فقط و فقط به یک انجمن تعلق دارد.

عمدتاً بیشترین مقدار نسبت همپوشانی برای روش LC حاصل شده است. شایان ذکر است که این معیار برای CPM و GCE، مقداری کمتر از ۱ را نتیجه می‌دهد زیرا این روش‌ها، گره‌های با درجه اندک (کمتر از k-1) را جزو هیچ‌یک از انجمن‌ها دسته‌بندی نمی‌کنند. مقادیر جدول، برای k=5 به دست آمده‌اند یعنی رئوس با درجه کمتر از ۴ بررسی نمی‌شوند.

روش‌های انجمن‌یابی مبتنی بر ساختار گراف و استفاده از مفهوم کلیک، روش‌هایی ساده و دقیق هستند و مبانی ریاضی قدرتمندی دارند. این روش‌های همپوشان هیچ عنصر غیرقطعی ندارند و در هر بار اجرا جواب یکسانی تولید می‌کنند. انتخاب مقادیر کوچک برای k، باعث افزایش تعداد انجمن‌ها شده و انتخاب مقادیر خیلی بزرگ برای k، فقط انجمن‌های بسیارچگال و بزرگ یک گراف را نشان می‌دهد. از نقاط ضعف روش‌های مبتنی بر کلیک می‌توان به پیچیدگی زمانی نامناسب و غیرمقاوم‌بودن به تغییرات اندک در تغییرات شبکه اشاره کرد: با حذف یا اضافه شدن حتی یک گره (یال) در شبکه، باید تمام محاسبات را از ابتدا انجام داد. همچنین، گره‌های با درجه پایین (کمتر از مقدار k-1) جزو هیچ‌یک از انجمن‌ها دسته‌بندی نمی‌شوند.

الگوریتم OKM که در مرحله سوم مدل پیشنهادی استفاده می‌شود، گره‌های گراف را بصورت همپوشان به خوشه‌ها (انجمن‌ها) تخصیص می‌دهد. این امر معادل این است که در ماتریس W که نحوه تخصیص گره‌ها (سطرهای ماتریس) به انجمن‌ها (ستون‌های ماتریس) را نشان می‌دهد، سطرهای متناظر

با گره‌های همپوشان دارای دو یا چند مدخل برابر با ۱ باشند. این امر در شکل ۵-۳ نشان داده شده است.

```
Clustering matrix:
      [,1] [,2] [,3] [,4]
 [1,]    0    0    1    0
 [2,]    0    0    1    1
 [3,]    0    0    1    1
 [4,]    0    0    1    1
 [5,]    0    0    1    0
 [6,]    0    0    1    0
 [7,]    0    0    1    1
 [8,]    0    0    1    0
 [9,]    0    0    1    1
[10,]    0    0    1    1
[11,]    0    0    1    0
[12,]    0    0    1    1
[13,]    0    0    1    1
[14,]    0    0    1    1
[15,]    0    0    1    0
[16,]    0    0    1    0
```

شکل ۵-۳: قسمتی از خروجی الگوریتم خوشه‌بندی همپوشان

فصل ۶ : نتیجه گیری و کارهای آتی

در این رساله، تعدادی از روش‌های شاخص انجمن‌یابی با توجه به رویکرد روش به تعریف و نحوه شکل‌گیری انجمن و نیز توجه به مفهوم همپوشانی بررسی شدند. مدل‌سازی مفهوم همپوشانی (امکان تعلق هم‌زمان اعضا به دو یا چند انجمن) در چند عنوان شامل روش‌های طیفی و مرکزگرایی، تابع کیفیت، انتشار برچسب، ساختار گراف، نزدیکی رئوس به یکدیگر، و دسته‌بندی یال‌ها، طبقه‌بندی و بررسی شده‌اند. استفاده از ایده‌های خلاقانه و متنوع این روش‌ها نظیر کار با معیارهای مرکزگرایی در تفکیک اجزای گراف، عملکرد تابع چگالی، نحوه انتشار اطلاعات، سهولت دسترسی اعضا به یکدیگر، و توجه به الگوهای خاص در ساختار گراف، مورد مطالعه و بررسی قرار گرفت.

انجمن‌یابی یک مسئله بهینه‌سازی است و رویکرد به آن با توجه به مجهولات خواسته‌شده و مورد مطالعه در مسئله انجام می‌شود. دو موضوع مهم در انجمن‌یابی هر شبکه‌ای عبارتند از: چگونه انجمن‌ها یافته می‌شوند؟ و ما به دنبال یافتن چه نوع انجمن‌هایی هستیم؟ لذا فرآیند انجمن‌یابی به شدت به خود صورت مسئله و فرض‌های آن وابسته است.

مقادیر به‌دست‌آمده برای معیارهای ارزیابی کارایی روش‌ها روی مجموعه دادگان انتخابی، نشان می‌دهد که هیچ روشی بهترین مقادیر را برای تمام معیارهای ذکرشده نتیجه نمی‌دهد. هر روش با توجه به مکانیزم عملکرد خود، به تعیین انجمن‌ها می‌پردازد و لذا ممکن است روی شبکه خاصی (با توجه به ویژگی‌ها و توپولوژی آن) نتیجه بهتری بدهد. بنابراین هیچ روشی نمی‌تواند خود را به‌عنوان بهترین و جامع‌ترین مکانیزم انجمن‌یابی (خصوصاً برای شبکه‌های دنیای واقعی) مطرح کند.

پیشنهاد می‌شود برای هر شبکه خاص، انجمن‌یابی توسط دو یا چند روش انجام شود و نتایج حاصل با هم مقایسه گردند تا بهترین روش با توجه به ویژگی‌های آن شبکه انتخاب شود.

کاری که در رساله حاضر انجام پذیرفته است، در زمره یادگیری ماشین و داده‌کاوی قرار می‌گیرد. نتایج اجرای این کار و تلاشهای مرتبط با آن را می‌توان در رئوس زیر خلاصه کرد:

- برآورده کردن همپوشانی به عنوان یک فرض مهم در انجمن‌یابی شبکه‌های اجتماعی

- استفاده از نقاط قوت روشهای بازنمایی و توکاری گراف برای نگاشت گره‌های آن به یک فضای با پیچیدگی کمتر بعنوان قسمتی از کار انجمن‌یابی (با ایده از آنچه در پردازش متن و مدلسازی زبان استفاده می‌شود)
- استفاده از قدرت بالای مولفه‌های یادگیری عمیق در استخراج ویژگی‌ها برای پردازش گراف و انجام فرایند یادگیری در سطوح مختلف انتزاع (غیرخطی بودن مکانیزم)
- استفاده از خوشه‌بندی همپوشان روی داده‌های پردازش‌شده (خروجی مرحله یادگیری عمیق که با توجه به خطای بازسازی در کدکننده خودکار پشته‌شده، نمایش کاملا واضح و متمایزی از بازنمایی گره‌ها را ارائه می‌دهد)
- ارائه یک روش جامع و همپوشان که دارای پیچیدگی زمانی مطلوبی برای اکثر حالات است: پیچیدگی زمانی برای بدترین حالت از مرتبه $O(n^2)$ یا $O(|V|^2)$ است که از تجمیع پیچیدگی زمانی سه مرحله شاخص روش یعنی توکاری گره، یادگیری عمیق، و خوشه‌بندی همپوشان حاصل شده است. حتی در این حالت هم مثلا در مقایسه با روش SDNE با پیچیدگی $O(|V||E|)$ یا روش LINE با پیچیدگی $O(|E|d)$ عملکرد خوبی دارد.
- روش پیشنهادی، معیارهای مجاورت و ساختار گراف را بخوبی حفظ می‌کند، قابلیت کار با گرافهای وزن‌دار را دارد، و امکان اجرای سریعتر با توجه به قابلیتهای توازی موجود در ساختارهای یادگیری عمیق روی پایانه‌های مختلف را داراست.

برخی از موارد مهم در توسعه رساله حاضر به شرح زیر هستند:

- انجام کار روی دادگان‌های متنوع‌تر
- ساخت دادگان استانداردشده برای انجمن‌یابی همپوشان
- تعمیم روش به گراف‌های جهت‌دار

- تعمیم روش به گراف‌های دو یا چند قسمتی
- استفاده از دیگر روش‌های خوشه‌بندی همپوشان: روش OKM بکاررفته، به‌صورت Exhaustive عمل می‌کند. استفاده از روش‌های non-exhaustive برای این امر می‌تواند محمل خوبی برای مطالعات بعدی در نگاشت گراف‌ها (به‌عنوان داده‌های بزرگ ساخت‌نیافته) به فضاهای کوچک ساخت‌یافته باشد.
- استفاده از ساختارهای پیچیده‌تر در پیکربندی شبکه عمیق و نیز دیگر ساختارهای کدکننده خودکار نظیر کدکننده ضدنویز یا variational
- انجام انجمن‌یابی برای یک هدف خاص و مسئله خاص (task-based)

مراجع

[۱] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: threats and solutions", ***IEEE Communications Surveys & Tutorials***, vol. 16, pp. 2019-2036, 2014.

[۲] A.-L. Barabási, "Scale-free networks: a decade and beyond", ***science***, vol. 325, pp. 412-431, 2009.

[۳] M. R. Lee and C. Tsung Teng, "Understanding Social Computing Research", ***IT Professional***, vol. 15, pp. 56-62, 2013.

[۴] A.-L. Barabási, "Emergence of scaling in random networks", ***Science***, vol 286, pp. 509-512, 1999.

[۵] B. A.-L. Barabási and E. Bonabeau" ,Scale-free Networks", ***Scientific American,*** vol. 288, pp. 50-59, 2003.

[۶] D. J. Watts, *Six degrees: The science of a connected age*: WW Norton & Company, 2004.

[۷] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook", in *Proceedings of the 2nd ACM workshop on Online social networks*, 2009, pp. 37-42.

[۸] S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On Deep Learning for Trust-Aware Recommendations in Social Networks", ***IEEE Transactions on Neural Networks and Learning Systems,*** vol. 28, pp. 1164-1177, 2017.

[۹] N. Gulbahce and S. Lehmann, "The art of community detection", ***BioEssays***, vol. 30, pp. 934-938, 2008.

[۱۰] S. Fortunato, "Community detection in graphs", ***Physics reports***, vol. 486, pp. 75-174, 2010.

[۱۱] E. Esen, S. Ozkan, I. Atil, M. A. Arabaci, and S. Tankiz, "Image community detection", in *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, 2014, pp. 738-741.

[۱۳] Y. Shunsen, Y. Xinyu, Z. Chao ,and E. Spyrou, "Using social network theory for modeling human mobility", ***Network, IEEE***, vol. 24, pp. 6-13, 2010.

[۱۴] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach", in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 587-596.

[۱۵] S. Fortunato and D. Hric, "Community detection in networks: A user guide", ***Physics reports,*** vol. 659, pp. 1-44, 2016.

[۱۶] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks", ***Statistical Analysis and Data Mining,*** vol. 4, pp. 512-546, 2011.

[۱۷] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, *et al.*, "Community detection in large-scale networks: a survey and empirical evaluation", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, pp. 426-439, 2014.

[۱۸] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study", *ACM Computing Surveys (csur)*, vol. 45, p. 43, 2013.

[۱۹] S. E. Schaeffer, "Graph clustering", ***Computer science review,*** vol. 1, pp. 27-64, 2007.

[۲۰] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using bayesian non-negative matrix factorization", ***Physical review E***, vol. 83, p. 066114, 2011.

[۲۱] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection", in *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 1170-1175.

[۲۲] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks", ***Physical review E,*** vol. 69, p. 026113, 2004.

[۲۳] M. Girvan and M. E. Newman, "Community structure in social and biological networks ,"*Proceedings of the national academy of sciences*, vol. 99, pp. 7821-7826, 2002.

[۲۴] S. Gregory, "An algorithm to find overlapping community structure in networks", in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2007, pp. 91-102.

[۲۵] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks", in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 29-42.

[۲۶] D. Rhouma and L. B. Romdhane, "An efficient algorithm for community mining with overlap in social networks", ***Expert Systems with Applications***, vol. 41, pp. 4309-4321, 2014.

[۲۷] W. Zhi-Xiao, L. Ze-chao, D. Xiao-fang, and T. Jin-hui, "Overlapping community detection based on node location analysis", ***Knowledge-Based Systems***, vol. 105, pp. 225-235, 2016.

[۲۸] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks", ***Physical review E,*** vol. 70, p. 066111, 2004.

[۲۹] M. E. Newman, "Analysis of weighted networks", ***Physical review E***, vol. 70, p. 056131, 2004.

[۳۰] M. E. Newman, "Modularity and community structure in networks", ***Proceedings of the national academy of sciences***, vol. 103, pp. 8577-8582, 2006.

[۳۱] V. D. Blondel, J.-L. Guillaume ,R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", ***Journal of Statistical Mechanics: Theory and Experiment,*** p. P10008, 2008.

[۳۲] M. Chen, K. Kuzmin, and B. K. Szymanski, "Extension of Modularity Density for overlapping community structure", in *2014 IEEE/ACM International Conference*

on Advances in Social Networks Analysis and Mining (ASONAM 2014), 2014, pp. 856-863.

[۳۳] J. Baumes, M. Goldberg, and M. Magdon-Ismail, "Efficient identification of overlapping communities", in *IEEE International Conference on Intelligence and Security Informatics*, 2005, pp. 27-36.

[۳۴] S. Kelley, "The existence and discovery of overlapping communities in large-scale networks", PhD Thesis, Rensselaer Polytechnic Institute, NY, 2009.

[۳۵] A. Lancichinetti and S. Fortunato, "Community detection algorithms: a comparative analysis", ***Physical review E,*** vol. 80, p. 056117, 2009.

[۳۶] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks", ***PloS one***, vol. 6, p. e18961, 2011.

[۳۷] S. Bandyopadhyay, G. Chowdhary, and D. Sengupta, "FOCS: Fast Overlapped Community Search", ***IEEE Transactions on Knowledge and Data Engineering***, vol. 27, pp. 2974-2985, 2015.

[۳۸] C. Pizzuti, "GA-Net: A Genetic Algorithm for Community Detection in Social Networks", in *PPSN*, 2008, pp. 1081-1090.

[۳۹] C. Pizzuti, "A Multiobjective Genetic Algorithm to Find Communities in Complex Networks", ***IEEE Transactions on Evolutionary Computation,*** vol. 16, pp. 418-430, 2012.

[۴۰] W. Chen, Z. Liu, X. Sun, and Y. Wang, "A game-theoretic framework to identify overlapping communities in social networks", ***Data Mining and Knowledge Discovery***, vol. 21, pp. 224-240, 2010.

[۴۱] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", ***Physical review E***, vol. 76, p. 036106, 2007.

[۴۲] S. Gregory, "Finding overlapping communities in networks by label propagation", ***New Journal of Physics***, vol. 12, p. 103018, 2010.

[۴۳] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks", in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2012, pp. 25-36.

[۴۴] J. Xie, B. K. Szymanski, and X. Liu, "Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process", in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 344-349.

[۴۵] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon :a local-first discovery method for overlapping communities", in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 615-623.

[۴۶] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels", ***Journal of Machine Learning Research,*** vol. 9, pp. 1981-2014, 2008.

[۴۷] A. McDaid and N. Hurley, "Detecting highly overlapping communities with model-based overlapping seed expansion", in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, 2010, pp. 112-119.

[۴۸] P. J. McSweeney, K. Mehrotra, and J. C. Oh, "A Game Theoretic Framework for Community Detection", in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, 2012, pp. 227-234.

[۴۹] L. Zhou, K. Lü, P. Yang, L. Wang, and B. Kong, "An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory", ***Expert Systems with Applications***, vol. 42, pp. 9634-9646, 2015.

[۵۰] H. Alvari, S. Hashemi, and A. Hamzeh, "Discovering overlapping communities in social networks: A novel game-theoretic approach", ***AI Communications,*** vol. 26, pp. 161-177, 2013.

[۵۱] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion", ***IEEE Transactions on Knowledge and Data Engineering,*** vol. 28, pp. 1272-1284, 2016.

[۵۲] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", ***Nature***, vol. 435, pp. 814-818, 2005.

[۵۳] F. Hao, G. Min, Z. Pei, D. S. Park, and L. T. Yang, "K-Clique Community Detection in Social Networks Based on Formal Concept Analysis", ***IEEE Systems Journal,*** vol. PP, pp. 1-10, 2015.

[۵۴] I. Farkas, D. Ábel, G. Palla, and T. Vicsek, "Weighted network modules", ***New Journal of Physics,*** vol. 9, p. 180, 2007.

[۵۵] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, and T. Vicsek, "Directed network modules", ***New Journal of Physics,*** vol. 9, p. 186, 2007.

[۵۶] S. Lehmann, M. Schwartz, and L. K. Hansen, "Biclique communities", ***Physical review E***, vol. 78, p. 016108, 2008.

[۵۷] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki, "Sequential algorithm for fast clique percolation", ***Physical review E,*** vol. 78, p. 026109, 2008.

[۵۸] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks", ***Physica A: Statistical Mechanics and its Applications,*** vol. 388, pp. 1706-1712, 2009.

[۵۹] C. Lee, F. Reid, A. McDaid, and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion", *arXiv preprint arXiv:1002.1827*, 2010.

[۶۰] J. Yang and J. Leskovec, "Overlapping Communities Explain Core - Periphery Organization of Networks", ***Proceedings of the IEEE,*** vol. 102, pp. 1892-1902, 2014.

[۶۱] P. Pons and M. Latapy, "Computing communities in large networks using random walks", in *International Symposium on Computer and Information Sciences*, 2005, pp. 284-293.

[۶۲] F. Wei, W. Qian, C. Wang, and A. Zhou, "Detecting overlapping community structures in networks", ***World Wide Web***, vol. 12, pp. 235-261, 2009.

[۶۳] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure", *Proceedings of the national academy of sciences*, vol. 105, pp. 1118-1123, 2008.

[۶۴] D. Jin, B. Yang, C. Baquero, D. Liu, D. He, and J. Liu, "A Markov random walk under constraint for discovering overlapping communities in complex networks", ***Journal of Statistical Mechanics: Theory and Experiment,*** vol. 2011, p. P05031, 2011.

[۶۵] C. Piccardi, "Finding and testing network communities by lumped Markov chains", ***PloS one,*** vol. 6, p. e27028, 2011.

[۶۶] T. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities", ***Physical review E***, vol. 80, p. 016105, 2009.

[۶۷] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks", ***Nature***, vol. 466, pp. 761-764, 2010.

[۶۸] C. L. Staudt and H .Meyerhenke, "Engineering parallel algorithms for community detection in massive networks", ***IEEE Transactions on Parallel and Distributed Systems***, vol. 27, pp. 171-184, 2016.

[۶۹] H. Cai, V. W. Zheng, and K. Chang, "A comprehensive survey of graph embedding: problems, techniques and applications", ***IEEE Transactions on Knowledge and Data Engineering,*** 2018.

[۷۰] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey", ***Knowledge-Based Systems,*** vol. 151, pp. 78-94, 2018.

[۷۱] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization", in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 37-48.

[۷۲] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks", in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855-864.

[۷۳] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding", in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1105-1114.

[۷۴] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[۷۵] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, 2013 ,pp. 3111-3119.

[۷۶] J. Schmidhuber, "Deep learning in neural networks: An overview", ***Neural networks***, vol. 61, pp. 85-117, 2015.

[۷۷] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", ***Nature,*** vol. 521, pp. 436-444, 2015.

[۷۸] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding", in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225-1234.

[۷۹] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures", in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 37-49.

[۸۰] M. Bastian, S. Heymann, and M. Jacomy, *Gephi: An Open Source Software for Exploring and Manipulating Networks*, 2009.

[۸۱] F. Chollet, "Keras: The python deep learning library ,"*Astrophysics Source Code Library*, 2018.

[۸۲] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library", ***ACM Transactions on Intelligent Systems and Technology (TIST),*** vol. 8, pp. 1-20, 2016.

[۸۳] J. Kunegis, "Konect :the koblenz network collection", in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 1343-1350.

[۸۴] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms", ***Physical review E,*** vol. 78, p. 046110, 2008.

[۸۵] J. Long, C. Yang, W. Tianyi, H. Pan, and A. V. Vasilakos, "Understanding user behavior in online social networks: a survey", ***Communications Magazine, IEEE***, vol. 51, pp. 144-150, 2013.

[۸۶] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[۸۷] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study", *Behavior research methods*, vol. 39, pp. 510-526, 2007.

[۸۸] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization", in *Advances in neural information processing systems*, 2014, pp. 2177-2185.

[۸۹] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information", in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 891-900.

[۹۰] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[۹۱] G. Cleuziou, "An extended version of the k-means method for overlapping clustering", in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[۹۲] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection", in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 631-640

Abstract

The increasing attractiveness and capabilities of cyberspace and the widespread use of personal electronic devices have made the Internet environment the mainstay backbone of many interactions between people and businesses. A great deal of interest in online social networks is a clear sample in this regard. The equivalent graph of these networks has a high volume of data due to a large number of nodes (humans) and edges (friendships and relationships between individuals or groups), so the processing of information on these large and complex networks requires efficient methods. Community detection, as one of the most essential branches of social network analysis, extracts and categorizes information of a network by identifying its constituent components. The applications of Community detection in the various sciences have led to the development of numerous methods with different aims and approaches. In the simplest case, the community detection - that is strongly related to the type of network and how to study it - can be considered as clustering or partitioning problem. However, with the increasing size and volume of information available in networks, membership of nodes in two or more communities seems inevitable. A set of algorithms that assign each node just to a single community ignores the overlap of nodes in them and eliminates an amount of information.

In this thesis, we present a novel method for overlapping community detection. We have used an improved version of node representation and graph embedding, improved node Embedding (modeling input information to deep learning network), as well as using computational components of deep learning such as stacked Autoencoders. According to performance evaluation criteria in overlapping social networks, the proposed method works better than the present ones. Apart from this, the proposed method has a moderate time complexity compared to most existing methods, due to the use of deep learning technics, it can use powerful processors to accelerate its computations, maintains a large number of proximity measures, and ultimately reflects the local and global structure of graph nodes in detected communities.

Keywords: Social network analysis, Community Detection, Overlapping, Deep Learning, Graph Representation, Graph Embedding

Shahrood University of Technology

Faculty of Computer Engineering

Ph.D. Thesis in Computer Engineering – Artificial Intelligence

Online Social Network Analysis For Detecting Overlapped Communities

By: Seyed Mohammad Mahdi Salehi

Supervisor:
Dr. Ali Akbar Pouyan

Advisor:
Dr. Hamid Hassanpour

February 2020