# تشخیص اجتماعات ترکیبی در شبکه‌های اجتماعی

**حسین علیزاده[۱]  رسول حسین‌زاده[۲]  اسلام ناظمی[۳]**

۱- دانش‌آموخته دکتری- دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران – تهران – ایران
halizadeh@iust.ac.ir
۲- دانش‌آموخته کارشناسی ارشد- دانشکده مهندسی برق و کامپیوتر - دانشگاه شهید بهشتی – تهران – ایران
rasoulhossienzadeh@gmail.com
۳- استادیار- دانشکده مهندسی برق و کامپیوتر - دانشگاه شهید بهشتی – تهران – ایران
nazemi@sbu.ac.ir

**چکیده: یکی از چالش‌های مهم در تحلیل شبکه‌های اجتماعی، تشخیص اجتماعات است. اجتماع مجموعه افراد یا سازمان‌هایی هستند که چگالی ارتباط آن‌ها با هم بیشتر از سایر موجودیت‌های شبکه است. خوشه‌بندی یا تشخیص اجتماعات، ساختار گروه‌ها در شبکه‌های اجتماعی و ارتباطات پنهان بین مؤلفه‌های آنها را آشکار خواهد نمود. اکثر روش‌های رایج تشخیص اجتماعات موجود قطعی نیستند و نتایج آن‌ها به مقادیر اولیه‌ای که در اکثر مواقع به صورت تصادفی انتخاب می‌شود بستگی دارد. اما خوشه‌بندی ترکیبی، بدون توجه به مقادیر اولیه تصادفی هر کدام از الگوریتم‌های پایه‌اش، با ترکیب آنها، نتایج مستحکم و پایداری تولید می‌کند.در این مقاله یک روش ترکیبی تشخیص اجتماعات با الهام از خوشه‌بندی ترکیبی پیشنهاد شده است. از مشخصه‌های روش پیشنهادی تشخیص اجتماعات ترکیبی، توانایی ترکیب با روش‌های دیگر است به گونه‌ای که می‌توان از الگوریتم‌های دقیق‌تر نیز در چهارچوب پیشنهادی استفاده کرد. نتایج تجربی در این مقاله نشان می‌دهند که روش ترکیبی پیشنهادی نسبت به متوسط روش‌های تشخیص اجتماعات پایه‌ای مورد استفاده در آن از کارایی بسیار بالاتری برخوردار بوده و در اکثر موارد حتی از بهترین الگوریتم پایه‌ای خود نیز بهتر عمل کرده است. نتایج این مقاله می‌تواند در بسیاری از مسائل از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موتورهای جستجو مورد استفاده قرار گیرد.**

**کلمات کلیدی: خوشه‌بندی، خوشه‌بندی ترکیبی، تشخیص اجتماعات، تشخیص اجتماعات ترکیبی، تحلیل شبکه‌های اجتماعی.**

**تاریخ ارسال مقاله :** ۱۳۹۱/۱۱/۱۷
**تاریخ پذیرش مشروط مقاله :** ۱۳۹۲/۰۸/۲۷
**تاریخ پذیرش مقاله :** ۱۳۹۲/۱۰/۱۶

**نام نویسنده‌ی مسئول:** حسین علیزاده

**نشانی نویسنده‌ی مسئول:** ایران – تهران – نارمک – خ هنگام – دانشگاه علم و صنعت ایران – دانشکده مهندسی کامپیوتر – آزمایشگاه داده‌کاوی

## ۱- مقدمه

امروزه اینترنت و خدمات وب به سرعت در حال گسترش هستند و همزمان شبکه‌های اجتماعی مجازی نقش مهمی در زندگی واقعی افراد دارد. در واقع شبکه‌های اجتماعی، شبکه‌های تعاملی هستند که از اینترنت به عنوان رسانه‌ای برای ایجاد ارتباط بین افراد استفاده می‌کنند. با افزایش سریع کاربران شبکه‌های اجتماعی، کاوش در مقیاس بالای داده‌ها می‌تواند کارایی بهتر و مؤثرتری از پتانسیل پنهان این شبکه‌ها فراهم کند[1].

تشخیص اجتماعات[۱] یا خوشه‌بندی[۲] یکی از مراحل اصلی در داده‌کاوی است که وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. با توجه به پیچیدگی مسئله خوشه‌بندی و ضعف روش‌های پایه‌ای آن، امروزه برخی مطالعات در خصوص خوشه‌بندی به سمت روش‌های خوشه‌بندی ترکیبی[۳] هدایت شده است. همان‌گونه که در بالا اشاره شد، تشخیص اجتماعات و خوشه‌بندی شباهت‌های بسیار زیادی با هم دارند و منابع زیادی این دو را یکسان می‌دانند[1][2][3][4].

رویکرد خوشه‌بندی ترکیبی بدین‌صورت می‌باشد که با استفاده از روش‌های پایه‌ای و یا هر روش ممکن موجود، با ترکیب تک‌تک این روش‌ها با یکدیگر درصدد پیدا نمودن یک جواب پایدار از نتایج متفاوت می‌باشد. به‌طوری که جواب نهایی دارای دقت، پایداری، صحت و از اجماع نتایج روش‌های مختلف بدست می‌آید، که دارای اطمینان بالاتری می‌باشد. با توجه به قابلیت ادغام خوشه‌بندی ترکیبی و تشخیص اجتماعات در این مقاله برای بهبود نتایج نهایی در تشخیص اجتماعات از «تشخیص اجتماعات ترکیبی» استفاده شده است. دقت و صحت اجتماعات نهایی در تشخیص اجتماعات مسئله مهمی می‌باشد، چرا که الگوریتمی در تشخیص اجتماعات وجود ندارد که در تمام حالات و داده‌های گوناگون جواب کاملاً دقیق را داشته باشد.

در این مقاله سعی داریم که با استفاده از تشخیص اجتماعات ترکیبی در شبکه‌های اجتماعی به بررسی اجتماعات در این شبکه‌ها بپردازیم. بدین‌صورت که با اعمال روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت و در نهایت ترکیب این روش‌ها تعدادی اجتماع که دارای دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری هستند را پیدا نماییم. نتایج این بررسی‌ها می‌تواند در مسائل بسیاری از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موتورهای جستجو مورد استفاده قرار گیرد.

قسمت‌های موجود در این مقاله بدین شرح می‌باشد که بعد از شرح تعاریف و مفاهیم و نیز بررسی کارهای انجام شده موجود، یک روش پیشنهادی در خصوص پیدا نمودن اجتماعات مطرح و در نهایت این روش را مورد ارزیابی قرار می‌دهیم. استفاده از این روش دارای مزایای و معایبی نیز می‌باشد که در قسمت نتیجه‌گیری به آن اشاره شده است.

## ۲- تعاریف و مفاهیم

در این قسمت به بررسی تعاریف و مفاهیمی که در این مقاله مورد استفاده قرار گرفته است می‌پردازیم.

### ۲-۱- شبکه‌های اجتماعی

شبکه‌های اجتماعی ساختاری اجتماعی است و از افراد یا سازمان‌هایی تشکیل شده است که گره‌های شبکه را تشکیل می دهند. گره‌ها توسط یک یا چند نوع خاص از وابستگی به هم متصل هستند، برای مثال تبادلات مالی، دوستی‌ها، خویشاوندی، تجارت، لینک‌های وب، سرایت بیماری‌ها(اپیدمولوژی) یا مسیرهای هواپیمایی نمونه‌هایی از ارتباط هستند. اما ساختارهای حاصل از این شبکه‌ها اغلب بسیار گسترده و پیچیده هستند. تحلیل شبکه اجتماعی[۴] عبارت است از نگاشت و اندازه‌گیری روابط و همکاری‌ها در بین افراد، گروه‌ها، سازمان‌ها و هر موجودیتی که قابلیت پردازش اطلاعات و دانش را داشته باشد. برای نمایش و تحلیل شبکه‌های اجتماعی معمولا از تئوری گراف[۵] استفاده می‌شود. مولفه‌های موجود در تئوری گراف گره[۶] و لبه[۷] است[1].

یک شبکه $N$ را می‌توان به صورت یک گراف به شکل $G =\prec V, E \succ$ تعریف نمود، که در آن گراف $G$ شامل $V$ مجموعه گره‌های گراف و $E$ مجموعه لبه‌ها است. بعد از بدست آوردن گراف فوق می‌توان از روی گراف، ماتریس مجاورتی متناظر با آن را بدست آورد. ماتریس $A$ را ماتریس مجاورتی گراف $G$ گویند، این ماتریس یک ماتریس مربعی به ابعاد تعداد گره‌های گراف متناظر با آن بوده و در صورت وجود ارتباط بین دو گره به آن ماتریس عدد یک و درغیر این صورت عدد صفر منظور می‌گردد.

### ۲-۲ تشخیص اجتماعات

در شبکه‌های اجتماعی برخی گره‌ها در مقایسه با کل گره‌های شبکه، ارتباط بیشتری با هم دارند که به آن‌ها اجتماع گفته می‌شود[3]. هدف از تشخیص اجتماعات، جدا کردن گروه‌ها یا اجتماعاتی است که ارتباط بیشتری با هم دارند[2]. در واقع تشخیص اجتماعات، تقسیم‌بندی‌های موجود در شبکه را نشان می‌دهد و اجتماعات یک گراف را از هم مجزا می‌کند. تشخیص اجتماعات به ما کمک می‌کند تا دید بهتری نسبت به ساختار شبکه پیدا کنیم[4].

در [3] تشخیص اجتماعات را بدین‌صورت زیر تعریف می‌نماید:

اگر $A$ را ماتریس مجاورتی یک گراف درنظر بگیریم، درجه یک گره در گراف را با $k_i$ نشان می‌دهند که برابر است با: $k_i = \sum_j A_{ij}$



Journal of Iranian Association of Electrical and Electronics Engineers - Vol.11- No.2- Fall & Winter 2014

می‌باشد. با در نظر گرفتن یک زیر گراف $S$ به‌طوری که $S \subseteq G$ باشد.

(۱)

$$S \subseteq G \Rightarrow k_i(S) = k_i^{in}(S) + k_i^{out}(S)$$

در رابطه (۱) نیز درجه هر گره را برای اجتماعات داخلی و خارجی به‌صورت $k_i^{in}(S) = \sum_{j \in S} A_{ij}$ و $k_i^{out}(S) = \sum_{j \notin S} A_{ij}$ تعریف می‌کنیم. بدین معنی که درجه یک گره به نام i که عضو اجتماع j می باشد را به عنوان مجموع اجتماعات داخلی بدین صورت تعریف می کنیم $k_i^{in}(S) = \sum_{j \in S} A_{ij}$ و درجه یک گره به نام i که عضو اجتماع j نمی باشد را به عنوان مجموع اجتماعات خارجی $k_i^{out}(S) = \sum_{j \notin S} A_{ij}$ تعریف می‌کنیم.

اجتماعاتی که مجموع گره‌های داخلی آن خیلی بیشتر از مجموع گره‌های خارجی آن باشد را به عنوان یک زیر گراف از کل گراف ورودی در نظر می‌گیریم و به انجام این کار تشخیص اجتماعات گویند و درصدد پیدا نمودن بهینه‌ترین و دقیق‌ترین اجتماعات هستیم که به صورت رابطه (۲) تعریف می‌شود.

حال به تعریف مسئله اصلی که همان تشخیص اجتماعات می‌باشد می‌پردازیم.

$$\sum_{i \in S} k_i^{in}(S) >> \sum_{i \notin S} k_i^{out}(S) \qquad (۲)$$

با توجه به رابطه (۲)، $\sum_{i \in S} k_i^{in}(S)$ مجموع درجه‌های تمامی گره‌هایی که در اجتماع یا زیرگراف $S$ وجود دارد می‌باشد و $\sum_{i \notin S} k_i^{out}(S)$ مجموع درجه‌های تمامی گره‌هایی که در اجتماع یا زیرگراف $S$ وجود ندارد می‌باشد. بنابراین مسئله اصلی در زمینه تشخیص اجتماعات این است که بدانیم چگونه به بهترین حالت شبکه را به گروه‌های اصلی آن تقسیم کنیم. در شبکه‌ها واقعی هیچ اطلاعاتی درباره تعداد اجتماعات وجود ندارد. در بعضی روش‌های تشخیص اجتماعات فرض بر آن است که تعداد اجتماعات شبکه را از قبل می‌دانیم. در حالی که در بسیاری از شبکه‌ها، هیچ دانش اولیه‌ای در مورد اجتماعات شبکه وجود ندارد و روش‌های جدید بدنبال برطرف کردن این نقیصه هستند.

روش‌های تشخیص اجتماعات معمولا با استفاده از گراف ایجاد شده از ارتباطات بین افراد در شبکه‌های اجتماعی ماتریس همسایگی[8] متناظر با آن را محاسبه می‌کنند و بعد از آن با اعمال الگوریتم‌های تشخیص اجتماعات بر روی این ماتریس می‌توان نتایج گوناگون بر روی داده‌های متفاوت را بدست آورد. هر یک از روش‌ها با توجه به روش انتخابی و نیز نوع و حجم داده ورودی، اجتماعات متفاوتی را پیدا می‌نمایند، که در بعضی شرایط یک الگوریتم جواب خوب و در شرایط دیگر جواب نامناسبی ایجاد می‌نماید، بنابراین هر روشی برای مجموعه داده‌های خاص مناسب و مورد استفاده قرار می‌گیرد.

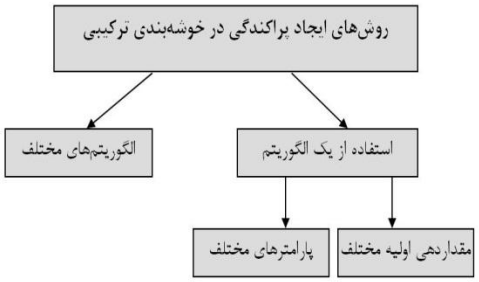
## ۲-۳- خوشه‌بندی ترکیبی

خوشه‌بندی یکی از مراحل اصلی در داده‌کاوی است که در جهت کاوش الگوهای پنهان در داده‌های بدون برچسب مورد استفاده قرار می‌گیرد. روش‌های خوشه‌بندی عادی بدین‌صورت عمل می‌نمایند که با تعریف یک الگوریتم سعی در برطرف نمودن مشکل خوشه‌بندی دارند. درصورتی که اعمال هر یک از این الگوریتم برروی داده‌های یکسان دارای نتایج گوناگونی خواهد.

در واقع هدف اصلی خوشه‌بندی ترکیبی جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است. پراکندگی در نتایج اولیه یکی از مهم‌ترین عواملی است که می‌تواند در کیفیت نتایج نهایی خوشه‌بندی ترکیبی اثر گذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب موثر است. به طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد[5]:

- تولید نتایج متفاوت از خوشه‌بندی‌ها، به عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی[9] می‌نامند.
- ترکیب نتایج به دست آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کار توسط تابع توافقی[10] (الگوریتم ترکیب کننده) انجام می‌شود.

برای ایجاد نتایج مختلف در این مقاله از دو روش که در شکل(۱) نشان داده شده است استفاده کرده‌ایم.




**شکل(۱): ایجاد نتایج اولیه مختلف**

همان‌طوری که در شکل(۱) مشاهده می‌شود، برای ایجاد نتایج مختلف می‌توان از یک الگوریتم یا از الگوریتم‌های متفاوت استفاده نمود. بدین‌صورت که با مقداردهی اولیه متفاوت و تغییر در پارامترهای آن می‌توان نتایج مختلفی را ایجاد نمود.

پس از اینکه نتایج اولیه مخلتف ایجاد شد، معمولاً با استفاده از یک تابع ترکیب کننده این نتایج ترکیب می‌شوند. یکی از متداول‌ترین

روش‌های ترکیب نتایج استفاده از ماتریس همبستگی[11] است. روش خوشه‌بندی ترکیبی انباشت مدارک (EAC[12]) که مبتنی بر ماتریس همبستگی است اولین بار توسط فرد و جین در [6] مطرح شد و خیلی زود به صورت یک روش متداول درآمد. امروزه روش‌های دیگری نیز مبتنی بر ماتریس همبستگی ارایه شده‌اند[7].

بنابراین به صورت کلی این دو مرحله در اصل یک چارچوب برای هر روش ترکیبی خواهد بود. در ابتدا نتایج مختلف را از اجرای روش‌های پایه بدست می‌آوریم و سپس با اجرای یک تابع توافقی که در بخش بعدی به آن پرداخته می‌شود نتایج مختلف را با هم ترکیب می‌کنیم تا یک نتیجه دقیق و مستحکم و از همه مهم‌تر پایدار را ایجاد نماییم.

## ۳- روش‌های موجود

در این مقاله از روش‌های مختلف خوشه‌بندی ترکیبی در جهت بهبود تشخیص اجتماعات استفاده شده است. در این بخش به بررسی این دو مورد می‌پردازیم.

### ۳-۱- تشخیص اجتماعات

روش‌های مختلفی برای بدست آوردن اجتماعات در شبکه وجود دارد. بررسی تک‌تک گره‌ها و قرار دادن آن‌ها در اجتماعات مختلف و در نهایت ارزیابی آن دارای هزینه زمانی و محاسباتی بالایی است و این رویکرد به‌طور عملی امکان‌پذیر نیست. برای غلبه بر این مشکل روش‌های مکاشفه‌ای[13] به کمک آمده‌اند. که از جمله این روش‌ها می‌توان به روش‌های طیفی[14]، تقسیم‌کننده[15]، تجمعی[16]، روش‌های مبتنی بر بهینه‌سازی ماژولاریتی[17]، تکنیک‌های حریصانه[18]، گداختگی شبیه‌سازی شده[19]، الگوریتم‌های تکاملی[20] اشاره نمود. هر کدام از این روش‌ها به نوعی در بهبود عملکرد تشخیص اجتماعات موثر است[4][3].

ماژولاریتی از پرکاربردترین معیارهای مورد استفاده در روش‌های مختلف است. این معیار کمیتی از گروه‌بندی که از کل گراف بدست آمده است، ارائه می‌کند و نقش بسزایی در تعیین صحت گروه‌بندی دارد. معمولا برای ارزیابی هر روش تشخیص اجتماعات در شبکه، مقدار ماژولاریتی گروه‌بندی پیشنهادی آن روش را برای شبکه‌ها و گراف‌های مختلف محاسبه می‌کنند. هرچه ماژولاریتی به‌دست آمده بیشتر باشد، دقت روش موردنظر بهتر بوده است. معیار ماژولاریتی مهمترین محک در ارزیابی روش‌های تشخیص اجتماعات در شبکه‌های اجتماعی است. ماژولاریتی به‌صورت رابطه (۳) زیر تعریف می‌شود[9][8][3]:

(۳)

$$Q = \sum_i (e_{ii} - a_i^2)$$

$$a_i = \sum_j e_{ij}$$

در رابطه فوق $i$ و $j$ اندیس‌های اجتماع است و $e_{ii}$ نسبت تعداد لبه‌هایی که گره‌های داخل اجتماع $i$ را به هم متصل می‌کند به کل لبه‌های گراف است. به عبارت دیگر کسری از تعداد لبه‌ها که دو سر آن در اجتماع $i$ قرار دارد به کل لبه‌های گراف و $a_i$ نسبت تعداد لبه‌هایی که حداقل یک گره آن در اجتماع $i$ باشد به کل لبه‌های گراف است. در واقع رابطه فوق سهم هر اجتماع $i$ را در کل شبکه بیان می‌کند. هرچه این عدد بزرگتر باشد، این اجتماع سهم بیشتری در شبکه دارد و در واقع اجتماع قوی‌تر است. بدین معنی که ارتباطات داخل اجتماع در آن بیشتر از ارتباط آن با سایر اجتماعات شبکه است. اگر کل گراف شامل تنها یک اجتماع باشد یا گره‌ها بصورت تصادفی در بین اجتماعات قرار گرفته باشند $Q=0$ خواهد بود. هرچه مقدار $Q$ به یک نزدیک‌تر باشد اجتماعات بهتر جدا شده‌اند ولی هیچ‌گاه یک نمی‌شود. در عمل $Q > 0.3$ ساختار گروهی مناسبی را نشان می‌دهد. ماژولاریتی می‌تواند مقادیر منفی را نیز بپذیرد[9]. مقدار ماژولاریتی ممکن است برای یک شبکه در بهترین حالت زه‌بندی از عدد خاص بیشتر نشود. این بدان معنا نیست تشخیص اجتماع به درستی انجام نشده است، بلکه ممکن است بیشترین مقداری باشد که ماژولاریتی برای آن شبکه می‌تواند داشته باشد.

یکی از روش‌های تشخیص اجتماعات روش تقسیم‌کننده می‌باشد. در شبکه‌ها، هر چه تعداد مسیرهایی که از گره یا لبه خاصی عبور می کنند بیشتر باشد آن گره یا لبه مهم‌تر است. بنابراین با فرض اینکه کوتاه‌ترین مسیر بین دو گره محاسبه‌پذیر باشد، اهمیت یک گره یا لبه را می‌توان اندازه‌گیری نمود، که به آن مرکزیت مابینی گفته می‌شود. یکی از روش‌های بدست آوردن اجتماعات که [10] ارائه شده است مرکزیت مابینی می‌باشد. رابطه مرکزیت مابینی به صورت رابطه (۴) تعریف می‌شود:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \qquad (۴)$$

که در آن $\sigma(i,u,j)$ تعداد کوتاه‌ترین مسیرها بین دو اجتماع $i$ و $j$ است که از گره یا لبه $u$ عبور می‌کند. $\sigma(i,j)$ تعداد کل مسیرها بین دو گره $i$ و $j$ است. در این روش با پیدا نمودن لبه‌ای که بیشترین مرکزیت یا ارتباط با دیگر گره‌ها را دارد و حذف بازگشتی آن لبه در نهایت اجتماعات جدا شده بدست می‌آید. روش‌های دیگری نیز مانند ضریب خوشه‌بندی لبه نیز وجود دارد که از این ایده استفاده نموده‌اند ولی در رابطه مربوطه لبه‌هایی که اجتماعات را به یکدیگر مرتبط می‌کنند ضریب کوچکی دارند که در هر مرحله حذف می‌شوند[9].

روش‌های دیگر که از جمله روش‌های طیفی است با ایجاد یک برش نرمال بر روی گراف به تقسیم نمودن گراف می‌پردازد. این معیار بر این اساس است که یک گروه‌بندی خوب تعداد لبه‌های بین اجتماعات را کمینه می‌کند، در عین حال لبه‌های داخل اجتماع را نگه می‌دارد[11]. این معیار به صورت زیر تعریف می‌شود:

گراف $G =\prec V, E \succ$ را در نظر بگیرید که در آن $V$ مجموعه گره‌های گراف و $E$ مجموعه تمام لبه‌های گراف است. گره-



Journal of Iranian Association of Electrical and Electronics Engineers - Vol.11- No.2- Fall & Winter 2014

های گراف را به دو مجموعه مجزای $B, A$ تقسیم می‌کنیم به طوریکه داشته باشیم، $B = V - A$ باشد. مقدار این رابطه کسری از اتصالات بین $B, A$ با توجه به اتصالات جداگانه $B, A$ می‌باشد.

مقدار cut بین $B, A$ به صورت رابطه (۵) است:

$$cut(A,B) = \sum_{i \in A, j \in B} W(i,j) \qquad (۵)$$

و مقدار association بدین صورت است:

$$assoc(A,V) = \sum_{i \in A, v \in V} W(i,v) \qquad (۶)$$

$W(i,v)$ وزن بین گره $i$ و $v$ است. رابطه (۶) برای نرمال کردن اندازه خوشه‌ها(اجتماعات) بکار می‌رود. این روش را معمولا برای گراف‌های وزن‌دار استفاده می‌کنند که البته قابل تعمیم به گراف‌های بدون‌وزن نیز هست ( با در نظر گرفتن وزن یک در صورت وجود لبه و صفر در صورت عدم وجود لبه).

حال از دو رابطه اخیر برای تعریف برش نرمال شده استفاده می‌کنیم.

$$N_{cut}(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \qquad (۷)$$

همان‌طور که در رابطه (۷) می‌بینیم این رابطه فقط شامل دو اجتماع می‌باشد، که نیاز به تعمیم برای پیدا نمودن اجتماعات دیگر دارد. بدین صورت که همین روال را بر روی اجتماع کوچکتر ایجاد شده اجرا نماییم تا زمانی که به بهترین مقدار ماژولاریتی دست پیدا نماییم.

یکی از مهم‌ترین روش‌ها، روش تجمعی می‌باشد. این روش بر این حقیقت بنا شده که گره‌های یک اجتماع ویژگی‌های مشترکی دارند و می‌توان از این ویژگی‌های مشترک برای گروه‌بندی استفاده کرد. در برابر روش‌های تقسیم‌کننده، روش تجمعی در ابتدا همه گره‌ها را جدا از هم و غیر متصل در نظر می‌گیرد و آن‌ها را بر اساس ویژگی‌های مشترک به هم متصل می‌کند تا به اجتماعات برسد. در واقع نحوه عملکرد این رویکرد بدین صورت است که لبه‌های بین اجتماعات خود به خود حذف شده و تنها لبه‌های داخل اجتماع باقی می‌مانند[12].

از جمله روش‌های دیگر روش بیشینه‌سازی ماژولاریتی می‌باشد، جستجو برای یافتن ماژولاریتی بهینه(یا بیشینه) از نوع مسائل بسیار سخت می‌باشد. چرا که فضای تقسیم‌بندی‌هایی که از گراف امکان‌پذیر است با بزرگ شدن اندازه گراف به سرعت در حال افزایش است. به همین دلیل رویکردهای جستجوی مکاشفه‌ای برای محدود کردن فضای جستجو الزامی است. در[16][15][14][13] روش‌هایی بر پایه ادغام کردن اجتماعات ارائه شده است که بیشتر توسط نیومن بوده است، بطوری که ماژولاریتی بیشینه شود.

از جمله روش‌های دیگر در این زمینه روش‌های مبتنی بر الگوریتم‌های تکاملی هستند. روش‌های مختلفی در حوزه الگوریتم‌های تکاملی بر روی تشخیص اجتماعات انجام شده است. تارگین و بینگل[۲۱] [17] روشی ارائه کردند که از ماژولاریتی برای سنجش استفاده می‌کند. در این روش هر کروموزوم کل گره‌های موجود در شبکه را شامل می‌شود. در ابتدا به هر گره به صورت تصادفی یک اجتماع نسبت داده می‌شود و سپس crossover که تغییراتی روی آن صورت گرفته، انجام می شود. گاهی نیز mutation گره‌های دو اجتماع را جابجا می‌کند اعمال می‌شود و در نهایت هر کروموزوم توسط معیار ماژولاریتی سنجیده می‌شود.

روش‌های دیگری نیز در خصوص تشخیص اجتماعات وجود دارد که در اینجا به این چند روش اشاره کردیم که در این مقاله نیز تا حدودی استفاده شده است. در [18] به الگوریتم‌های دیگری برای تشخیص اجتماعات اشاره شده است.

## ۳-۲- خوشه‌بندی ترکیبی

در این بخش روش‌های مختلف خوشه‌بندی ترکیبی با استفاده از ماتریس همبستگی را مورد بررسی قرار می‌دهیم.

در [6] یک روش برای ترکیب خوشه‌های ایجاد شده از مرحله یکم برای بدست آوردن خوشه‌های نهایی ارائه شده است. ایده اصلی در این روش که انباشت مدارک نام دارد، بدین صورت می‌باشد که شباهت بین نتایج مختلف بدست آمده در مرحله اول از خوشه‌بندی ترکیبی را مورد بررسی قرار می‌دهد. روش کار بدین صورت می‌باشد که نتایج $m$ الگوریتم خوشه‌بندی توسط تابع توافقی بررسی و در یک ماتریس همبستگی $n \times n$ ذخیره می‌شود. تابع توافقی بدین صورت رابطه(۸) تعریف می‌گردد:

$$C(i,j) = \frac{n_{i,j}}{m} \qquad (۸)$$

که $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های $i$ و $j$ با هم در یک خوشه گروه‌بندی شده‌اند و $m$ تعداد الگوریتم‌های استفاده شده یا نتایج $m$ الگوریتم خوشه‌بندی می‌باشد. $C$ نیز ماتریس همبستگی حاصله از ترکیب نتایج می‌باشد، پس از ساخت ماتریس همبستگی می‌توان با استفاده از یکی از الگوریتم‌های سلسله مراتبی نظیر اتصال منفرد یا اتصال میانگین[۲۲]، خوشه‌های نهایی را استخراج کرد. این نکته قابل یادآوری است که در اینجا تنها یک ماتریس استخراج می‌شود و خوشه‌های نهایی نیز به سادگی با اعمال یکی از الگوریتم‌های سلسله مراتبی از این ماتریس به دست می‌آیند.

در [7] نیز سه روش دیگر بر مبنای ماتریس همبستگی ارائه شده است که عبارتند از CTS[۲۳]، SRS[۲۴]، ASRS[۲۵] می‌باشد. در هر سه الگوریتم ارائه شده یک ماتریس همبستگی متفاوت ایجاد می‌شود که هر یک از این ماتریس‌ها شباهت‌های میان روش‌های خوشه‌بندی اولیه را نشان می‌دهد.

**CTS**: ایده این روش بر پایه تخمین تعداد مثلث‌های مرتبط با هم می‌باشد. در این روش با محاسبه تعداد مثلث‌های ایجاد شده بین

لبه‌های دو خوشه که با هم در ارتباط هستند به ایجاد ماتریس همبستگی می‌پردازد. بدین صورت که تعداد مثلث‌های مرتبط با هم در بین دو خوشه به مجموع تعداد مثلث‌های ممکن، ماتریس مشابهت را تشکیل می‌دهد. مثلث بدین معنی می باشد که دارای سه راس می باشد، اگر تمام این سه راس درون دو خوشه $i$ و $j$ باشد آن را به عنوان $CT_{ij}$ معرفی می کنیم.بدین صورت که تعداد کل مثلث‌های ممکن بدین شکل را محاسبه می‌کنیم و تعداد کل مثلث‌های ممکن بدین صورت که حداقل یک راس آن در خوشه $i$ و $j$ باشد را $CT_{\max}$ می گوییم. با محاسبه $CT_{ij}$ و $CT_{\max}$ می‌توان شباهت بین دو خوشه $i$ و $j$ را طبق رابطه (۹) محاسبه نمود.

$$SimCT(i,j)=\frac{CT_{ij}}{CT_{\max}} \qquad (۹)$$

همان‌طوری که در رابطه (۹) مشاهده می‌شود $CT_{ij}$ تعداد مثلث‌های مرتبط با هم که همه رئوس آن بین دو خوشه $i$ و $j$ می‌باشد و $CT_{\max}$ تعداد کل مثلث‌های مرتبط با هم که حداقل یک راس آن بین دو خوشه $i$ و $j$ می‌باشد.

**SRS**: ایده این روش بدین صورت است که همسایه‌ها، در یک ماتریس شبیه به هم خواهند بود اگر به اندازه کافی درجه شباهت به هم داشته باشند. به عبارت دیگر شباهت بین دو گره با استفاده از لبه‌هایی که به طور مستقیم به این دو گره متصل هستند بدست می‌آید. بدین معنی که هرچه گره دارای درجه بالاتری باشد و به گره مدنظر نیز متصل باشد، شباهت آن گره با استفاده از گره‌های همسایه بدست می‌آید. با محاسبه تعداد لبه‌های متصل به دو گره مختلف می‌توان شباهت بین آن‌ها در خوشه‌بندی‌های مختلف در یک ماتریس همبستگی بدست آورد.

$$SRS(i,j)=\frac{SR(i,j)}{SR_{\max}}\times DC \qquad (۱۰)$$

در رابطه (۱۰) صورت کسر $SR(i,j)$ درجه دو گره $i$ و $j$ خواهد بود، به شرطی که این دو گره در یک خوشه‌بندی نباشند و مخرج آن $SR_{\max}$ بیشترین مقداری می‌باشد که $SR(i,j)$ بدون توجه به هم خوشه بودن می‌تواند داشته باشد است. مقدار $DC=[0.1)$ می‌باشد که نشان دهنده درجه اعتمادی می‌باشد که دو گره‌ای که در یک خوشه‌بندی نباشند شبه هستند یا خیر. معمولا مقدار $DC$ را با توجه به گراف مورد بررسی مقدار ثابتی در نظر می‌گیریم.

**ASRS**: این روش نیز همانند روش SRS می‌باشد با این تفاوت که در SRS شباهت بین دو گره در تمام روش‌های خوشه‌بندی مورد بررسی قرار می‌گیرد ولی در این روش تقریبی از تمامی روش‌های خوشه‌بندی برای بدست آوردن ماتریس همبستگی نهایی استفاده می‌شود.

در این مقاله از این روش‌ها به عنوان تابع توافقی برای اجرای بر روی نتایج روش‌های تشخیص اجتماعات استفاده می‌نماییم.

## ۴- روش پیشنهادی

با توجه به اینکه اکثر روش‌های تشخیص اجتماعات پایه روی جنبه‌های خاصی از داده‌ها تمرکز می کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید نماید[5].

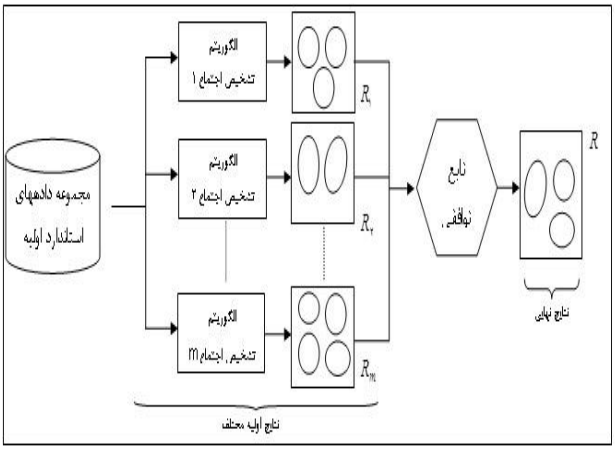
بنابراین تشخیص اجتماعات ترکیبی با ترکیب نتایج بدست آمده از الگوریتم‌های پایه درصدد پیدا نمودن نتایج دقیق‌تر، مطمئن‌تر و پایدارتر از نتایج بدست آمده قبلی می‌باشد. همان‌طوری که در شکل(۳) مشاهده می‌شود، انباره داده نشان داده شده در این شکل در اصل همان گراف ما خواهد بود، که از ارتباطات موجود در شبکه اجتماعی مورد بررسی ایجاد شده است. مجموعه این ارتباطات در نهایت یک گراف و از آن یک ماتریس همسایگی ایجاد می‌گردد که در صورت وجود ارتباط بین گره‌ها یک و در غیر این‌صورت صفر منظور می‌گردد. با اعمال الگوریتم‌های متفاوت بر روی این ماتریس همسایگی و گرفتن نتایج متفاوت و در بعضی موارد یکسان می‌توان از روش ترکیبی استفاده نمود.

همان‌طوری که در روش‌های خوشه‌بندی ترکیبی ارائه شد برای ترکیب نتایج از توابع توافقی استفاده می‌شود. توابع توافقی استفاده شده در این مقاله همان‌توابع EAC، CTS، SRS، ASRS می‌باشد. بدین‌صورت که اگر گره‌ای در چند نتیجه مختلف جزء یک اجتماع باشد احتمال قرارگیری آن گره در آن اجتماع بیشتر خواهد بود، در نهایت با ترکیب نتایج اولیه یک ماتریس همبستگی ایجاد خواهد شد که یک گراف وزن‌دار از نتایج ما می‌باشد، از مجموع نتایج کمک می‌گیریم تا به یک نتیجه بهتر، دقیق‌تر، پایدارتری برسیم.

```
CD: Community Detection
Input: A(Adjacency matrix)
  for Number of CD method do
                while (Can change initialization parameter) do
                        ECD+:=run(CD method)
                end while
                if(Can change final state CD method) then
                        ECD+:=run(CD method)
                end if
end for
ECD: Ensemble members of Community Detection methods
/*Each column show one result of community detection method*/
        Co.assosiation matrix:=run(Consensus function)
        C:=run(Hierarchical clustering method)
Output:R(Final CD result)
```

**شکل (۲): شبه کد روش تشخیص اجتماعات ترکیبی**

در این مقاله یک روش جدید برای افزایش دقت نتـایج تشـخیص اجتماعات ارائه شده است. این روش که در شکل(۲) نشـان داده شـده است، روش تشخیص اجتماعات ترکیبی می‌باشد بـا اسـتفاده از روش-های پایه‌ای برای بدست آوردن نتایج پایدارتر و دقیق‌تر مـورد اسـتفاده قرار می‌گیرد. شبه کد نشان داده شـده در شـکل(۲) بـا دریافـت یـک مـاتریس همسـایگی و اجـرای روش ترکیبـی بـا اسـتفاده از مـاتریس همبستگی در تشخیص اجتماعات نتیجـه پایـدارتر و دقیـق‌تـری را بـه عنوان نتیجه نهایی ارائه می‌کند.




**شکل(۳): الگوریتم پیشنهادی برای تشخیص اجتماعات ترکیبی**

ماتریس همبستگی ایجاد شده برآیندی از تمامی روش‌های اولیـه خواهد بود، در این ماتریس برای هر لبـه در گـراف مـورد نظـر عـددی منظور می‌شود که این عدد مقدار وابستگی آن لبه به گره‌ای کـه بـا آن در تماس می‌باشد را نشان می‌دهد. به عبارت دیگـر اجمـاع روش‌هـای مختلف در مورد اینکه این دو گره به چه میزان با هم در ارتباط هستند یا خیر را نشان خواهد داد. لبه‌های که ارتباط ضعیف دارند، در الگوریتم تشخیص اجتماعات ترکیبی سیاست حذف آن‌ها اتخاذ می‌گردد که بـه دقیق‌ترین نتایج نزدیک باشد. در نهایت با اعمال یکی از الگوریتم‌هـای سلسـله مراتبـی بـرای اسـتخراج نتـایج نهـایی بـر روی آن مـاتریس همبستگی، تشخیص اجتماعات کامل‌تر و دقیق‌تری را خواهیم داشت.

بعد از اعمال الگوریتم‌های مختلف بـر روی شـبکه مـورد تحلیـل همان‌طوری که در شکل(۳) نیز نشان داده می‌شود، یک سـری نتـایج متفاوت بدست خواهد آمد که به‌صـورت $R = \{R_1, R_2, ..., R_m\}$ تعریف می‌گردد. این نتایج تنها با اجرای روش‌های تشخیص اجتماعات متفاوت بدست آمده‌اند. بدین معنی که الگوریتم تشـخیص اجتماعـات مختلف را اجرا می‌نماییم و نتایج مختلف را به عنـوان ورودی بـه تـابع توافقی ارسال می کنیم، که بـا ترکیـب صـحیح روش‌هـای مختلـف بـا استفاده از توابع توافقی که در این مقاله به آن اشاره شده اسـت، مـی-توان به یک نتیجه بسیار بهتر، مطمئن‌تر، پایدارتر و دقیق رسید. چـون این نتیجه از اجماع یا هم‌فکری روش های مختلف ایجاد می‌شـود ایـن قابلیت اعتماد را نیز برای ما به همراه می‌آورد، نتایج ارزیـابی‌هـا دلیـل براین ادعا می‌باشد.

از نظر هزینه اجرایی، هزینه اجرایی این روش بالا و مقدار مصـرف حافظه آن نیز بالا می‌باشد و علت آن نیز بدلیل اسـتفاده از روش‌هـای مختلف و به تعداد زیاد می‌باشد. ولی هزینه اجرایی ایـن روش برابـر بـا بدترین حالت اجرای یک الگوریتم از بین تمام الگوریتم‌هـای تشـخیص اجتماعات خواهد بود، چرا که تعداد الگوریتم‌های مورد بررسی در ایـن فرایند ترکیبی در نهایت محدود خواهند بود.

به طور کلی روش پیشنهادی تشخیص اجتماعات ترکیبی درصدد ترکیب نتایج مختلف تشـخیص اجتماعـات بـرای بدسـت اوردن نتـایج دقیق‌تر و پایدارتر با بررسی نتایج اولیه می‌باشد.

## ۴-۱- ارزیابی روش پیشنهادی

در این مقاله چهار روش مختلف تشخیص اجتماعات با استفاده از MATLAB (ver R2009a) پیاده‌سازی شده اسـت، کـه درصـدد آن هستیم که با ترکیب این چهار روش با استفاده از خوشه‌بندی ترکیبـی به یک خوشه‌بندی یا اجتماعات دقیق‌تری نسـبت بـه تـک‌تـک نتـایج برسیم. از مزایای این روش می‌توان به مقیاس‌پـذیری بـالای آن اشـاره نمود بدین معنا که هر چه تعداد روش‌های ما برای تشخیص اجتماعات بیشتر باشد نتایج بهتری را در خوشه‌بندی ترکیبی به ارمغان می‌آورد و نیز هر روش را بدون محدودیت و با تغییـر شـرط اتمـام آن مـی‌تـوان استفاده نمود، چرا که ما در خوشه‌بنـدی ترکیبـی بـه نـوع روش نگـاه نمی‌کنیم بلکه نتیجه روش را مورد بررسی قرار می‌دهـیم و بـا بررسـی نتایج بدست آمده از تک‌تک روش‌ها یک خوشه‌بندی دقیق‌تری را ارائه می‌دهیم.

در اینجا به بررسی این چهـار الگـوریتم پیـاده‌سـازی شـده بـرای تشخیص اجتماعات می‌پردازیم که تک‌تک به چه صورتی اجتماعـات را بدست می‌آورند. کلیه این روش‌ها از روی ماتریس مجاورت ایجاد شـده از روی گراف به طوری که اگر ارتباطی بین گره‌ها وجود داشـته باشـد یک و در غیر این‌صورت صفر خواهد بود استفاده می‌کنند و نیـز کلیـه این روش‌ها نیز قابل تعمیم برروی گراف‌های جهـت‌دار و وزن‌دار مـی-باشد.

الگوریتم اول که با نام Newman یا N در این مقاله معرفی شـده است، یک روش برمبنای مرکزیت مابینی می‌باشـد، کـه در روش‌هـای موجود به آن اشاره شد. در این روش لبه‌هایی که بیشترین مرکزیـت را داشته باشند، به‌صورت بازگشتی حذف نموده و به این ترتیب یکسـری اجتماعات جدا از هم را خواهیم داشت[9].

الگوریتم دوم که با نام Newman Greedy یا NG در این مقالـه معرفی شده است، یک روش بر مبنای روش تجمعی ارائه شـده اسـت. ایده اصلی در این روش بهینه‌سازی ماژولاریتی و حریصانه می‌باشد، در این روش با استفاده از رابطه ماژولاریتی درصدد پیدا نمودن اجتماعاتی با بیشنه‌ترین حالت ماژولاریتی هستند. این روش چون از انـواع روش-

های تجمعی می‌باشد به یک معیار مشابهت نیاز داریم که در این روش ماژولاریتی معیار مشابهت خواهد بود[14].

الگوریتم سوم که با نام Newman & Girvan یا GN در این مقاله معرفی شده است، یک روش مشابه مرکزیت مابینی می‌باشد با این تفاوت که به جای بررسی تک‌تک گره‌ها در هر مرحله و بدست آوردند یک لبه برای حذف شدن در هر مرحله لبه‌ها با ضریب بالا را حذف نموده و در نهایت یک دندروگرام از کل گراف ایجاد می‌شود که یک نمایش سلسله مراتبی می‌باشد، با برش این دندروگرام می‌توان اجتماعات را تفکیک نمود ولی در برش می‌بایست دقت نمود که بهینه-ترین حالت را داشته باشد، برای پیدا نمودن بهینه‌ترین حالت نیز از ماژولاریتی استفاده شده است[15].

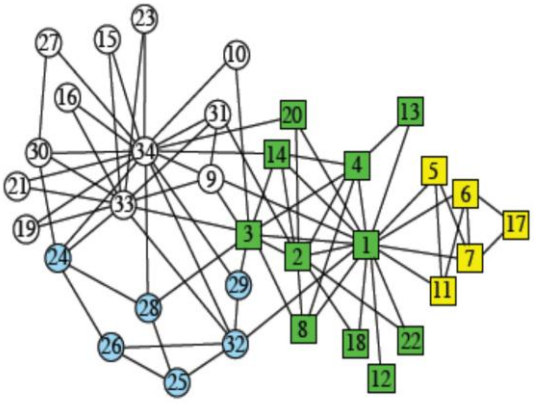
الگوریتم چهارم که با نام Radicchi & et al یا R در این مقاله معرفی شده است، نیز یک روش بر مبنای بهینه‌سازی ماژولاریتی می-باشد که با استفاده از ماژولاریتی و تعریف متغیرهای دیگری که در آن مقادیر ماژولاریتی و نیز حالات گره‌ها در آن وضعیت ذخیره می‌شود. این روش باعث افزایش سرعت در پیدا نمودن اجتماعات و نیز با داشتن یک پیش‌زمینه از مراحل قبلی با دقت بالاتری به ترکیب گره‌ها می‌پردازد[19].

حال با ترکیب این چهار روش با استفاده از توابع توافقی که مبتنی بر ماتریس همبستگی می‌باشد، به ترکیب نتایج می‌پردازیم، تا بتوانیم در تشخیص اجتماعات نتایج دقیق‌تری داشته باشیم. برای ارزیابی روش پیشنهادی از تعدادی داده استاندارد استفاده شده است که به بررسی آن‌ها می‌پردازیم.

## ۴-۲- داده‌های مورد استفاده

داده‌های مورد استفاده داده‌های استانداردی می‌باشند که تقریبا در تمامی روش‌های تشخیص اجتماعات از آن‌ها استفاده می‌شود. از جمله این داده‌ها می‌توان به باشگاه کاراته زاکاری[۲۶][20]، مجموعه داده‌های موسیقی دانان جاز[۲۷][21]، لیگ فوتبال دانشگاه امریکا[۲۸][9]، شبکه متابولیک، مجموعه داده‌های ایمیل Ernon توسط FERC[۲۹] [22] و مجموعه داده NetSciecne اشاره نمود.

**باشگاه کاراته زاکاری:** همان‌طوری که درشکل (۴) مشاهده می‌شود، این مجموعه یک گراف استاندارد برای تشخیص اجتماع است و در بسیاری از مقالات از آن استفاده شده است. این گراف شامل ۳۴ گره و ۷۸ لبه است، که اعضای یک باشگاه کاراته در ایالات متحده هستند و در طول ۳ سال مشاهده شده‌اند. لبه‌ها، افرادی را که خارج از فعالیت‌های باشگاه با هم ارتباط دارند به هم متصل می‌کند. گره ۳۴ مسئول باشگاه و گره ۱ مربی می‌باشد، در بعضی نقاط درگیری بین مسئول باشگاه با مربی موجب ایجاد شکاف بین اعضای باشگاه شده است و آن‌ها را به دو گروه طرفدار این دو فرد تقسیم کرده است(با دایره و مربع نشان داده شده است). با نگاه به شکل (۴) می‌توان دو تجمع را تشخیص داد. یکی اطراف گره ۳۳ و ۳۴ و دیگری اطراف گره ۱ قرار دارد[20]. رنگ‌های نشان داده شده در این شکل نیز تعداد گره-های این باشگاه را به ۴ اجتماع مختلف تفکیک نموده است که هر رنگ نشان دهنده یک اجتماع می‌باشد.




**شکل(۴): گراف باشگاه کاراته[3]**

**مجموعه داده‌های موسیقی دانان جاز:** در این پایگاه‌داده ۱۹۶ باند موسیقی وجود دارد که بین سال‌های ۱۹۱۲ و ۱۹۴۰ فعالیت می-کردند که بیشتر آن‌ها در سال‌های اطراف ۱۹۲۰ مشغول بوده‌اند. این پایگاه‌داده موسیقی‌دان‌هایی را لیست می‌کند که در این باندها می-نواختند. این گراف شامل ۱۹۶ گره و ۲۷۴۲ لبه است، اگر دو موسیقی دان در یک باند می‌نواختند لبه‌ها بین آن‌ها ترسیم می‌شود[21].

**لیگ فوتبال دانشگاه امریکا:** این شبکه که لیگ فوتبال دانشگاه امریکا را نشان می‌دهد، از ۱۱۵ گره تشکیل شده است. هر گره نشان دهنده یک تیم می‌باشد و در صورت بازی بین دو تیم در یک فصل یک لبه بین آن دو ایجاد می‌گردد. تعداد لبه‌ها در این شبکه ۶۱۶ لبه می‌باشد که نشان دهنده تعداد بازی‌های انجام شده بین دو تیم بوده است. این مشاهدات برای سال‌های ۲۰۰۰ تا ۲۰۰۱ بوده است[9].

**شبکه متابولیک:** در یک سلول یا میکرو ارگانیزم فرآیندهایی که انتقال انرژی و اطلاعات را تولید می‌کنند، به طور یکپارچه از طریق شبکه پیچیده‌ای از مولکول‌های سلولی و واکنش‌هایشان است. با وجود این که نقش کلیدی این شبکه‌ها حفظ کارکردهای سلولی است. اساساً ساختار بزرگ آن‌ها ناشناخته است. این شبکه، یک شبکه متابولیکی کرم‌ها به نام C.Elegans است که ۴۵۳ گره دارد. گره‌ها نمایانگر متابولیک‌ها و لبه‌ها نشان‌دهنده واکنش‌های متابولیکی است[16].

**مجموعه داده‌های ایمیل:** مجموعه داده‌های ایمیل Enron توسط FERC منتشر شد و البته اشکالاتی هم داشت. پس از آن توسط افراد دیگری بررسی شد و بسیاری از اشکالات آن بر طرف شد. این داده‌ها همه نوع ایمیل‌های شخصی و اداری است. این نسخه از داده‌ها شامل ۴۳۱،۵۱۷ ایمیل از ۱۵۱ کاربر است که در ۳۵۰۰ فولدر توزیع شده است. این گراف شامل ۱۱۳۳ گره و ۵۴۵۱ لبه است این



Journal of Iranian Association of Electrical and Electronics Engineers - Vol.11- No.2- Fall & Winter 2014

ایمیل‌ها بخش ضمیمه شده(مثل عکس، فیلم یا هر فایل دیگری) ندارد[22].

**مجموعه داده NetSciecne**: در این شبکه تعداد ۱۵۸۹ نویسنده وجود دارد که با هم در مورد تئوری شبکه پژوهش می‌کنند. این مجموعه داده با بررسی انجام شده بر روی مقالات مختلف در تئوری شبکه توسط نیومن در سال ۲۰۰۶ گردآوری شد. بدین‌صورت که ارتباطات این نویسندگان در مقالات مختلف را به صورت یک شبکه مدل می‌نماید. تعداد لبه‌ها در این شبکه ۲۷۴۲ لبه می‌باشد که نشان دهنده ارتباطات بین نویسندگان است.

حال که پیش‌زمینه لازم برای روش پیشنهادی فراهم شد و داده‌های مورد استفاده در ارزیابی نیز مورد بررسی قرار گرفت، حال به بررسی نتایج گوناگون می‌رسیم.

## ۴-۳- نتایج ارزیابی

جدول(۱) نتایج اعمال الگوریتم‌های مختلف را نشان می‌دهد. این جدول شش نوع داده رایج که برای ارزیابی روش‌های تشخیص اجتماعات مورد استفاده قرار می‌گیرند را نشان می‌دهد. تعداد گره‌ها و لبه‌های این داده‌های استاندارد در جلوی آن‌ها مشخص گردیده است. یک الگوریتم خوب می‌بایست در تعداد کم و زیاد گره‌ها خوب و بهینه جواب دهد چرا که در شبکه‌های اجتماعی تعداد گره‌ها بسیار متنوع می‌باشند و در بعضی موارد تعداد آن به کل کاربران در شبکه جهانی تامین پیدا می‌کند. روش ارزیابی این الگوریتم‌ها نیز ماژولاریتی می‌باشد که در روش‌های موجود به آن اشاره گردیده است، و این روش (که مقدار آن بین منفی یک و یک می‌باشد) مهمترین محک برای ارزیابی تشخیص اجتماعات شناخته شده است.

همان‌طوری که در جدول(۱) مشاهده می‌شود معیار ارزیابی برای روش پیشنهادی ماژولاریتی می‌باشد که مقدار آن بین (1+,1-) می‌باشد. اگر مقدار ماژولارتی صفر شود یعنی تمامی گره‌ها را دریک اجتماع قرار داده‌ایم و اگر منفی شود به معنی خطا زیاد در پیدا نمودن اجتماعات می‌باشد. در تمامی نتایج بالا مقدار آن مثبت بوده است، روش ما در تمامی حالت دارای دقت، پایداری و صحت بالایی می‌باشد. بدین معنی که همواره جواب بهینه بدست آمده توسط روش‌های شخیص اجتماعات معمولی را دارد ودرصدد پیدا نمودن جواب بهینه‌تری از ترکیب نتایج می‌باشد. با این حال که نوع داده در حال تغییر است ولی به جواب بهینه نزدیک می‌باشد. اگر این روش را بروی داده‌های بسیار بزرگ تست گردد این تفاوت مشهودتر خواهد بود چراکه این روش‌های تشخیص اجتماعات معمولی بر روی داده‌های بزرگ تقریبا جواب خوبی را نخواهند داشت، ولی برآیند نتایج جواب بهینه‌تری را خواهد داشت.

حال به بررسی روش‌های تشخیص اجتماعات ترکیبی و مقایسه آن با روش‌های پایه‌ای در تشخیص اجتماعات می‌پردازیم. روش‌های پایه همان چهار روش ارایه شده در روش پیشنهادی می‌باشند و روش‌های ترکیبی برای روش پیشنهادی نیز از توابع ترکیبی مبتنی بر ماتریس همبستگی استفاده می‌نمایند.

**جدول (۱): مقایسه روش‌های مختلف تشخیص اجتماعات بر اساس معیار ماژولاریتی**

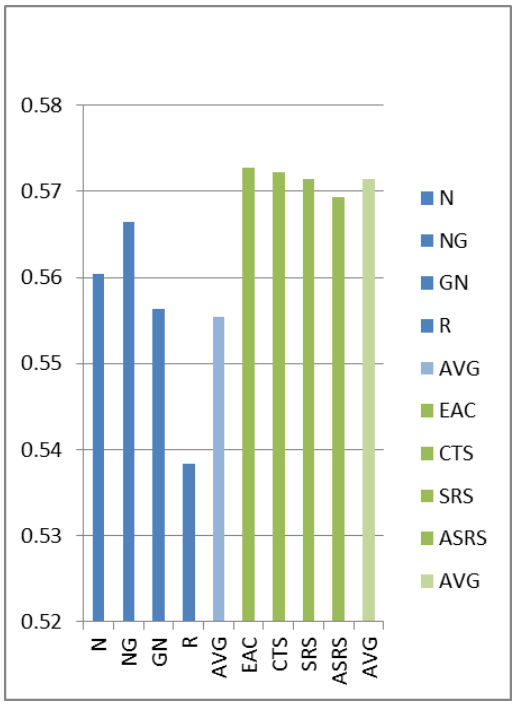
| Data Set | تعداد گره‌ها | تعداد لبه‌ها | Newman [9] | | Newman Greedy [14] | | Newman & Girvan [15] | | Radicchi & et al [19] | |
|---|---|---|---|---|---|---|---|---|---|---|
| کاراته | ۳۴ | ۷۸ | ۴ اجتماع | 0.4086 | ۴ اجتماع | 0.3980 | ۴ اجتماع | 0.4009 | ۳ اجتماع | 0.3853 |
| فوتبال | ۱۱۵ | ۶۱۶ | ۷ اجتماع | 0.5788 | ۹ اجتماع | 0.6057 | ۶ اجتماع | 0.5714 | ۶ اجتماع | 0.5581 |
| جاز | ۱۶۸ | ۲۷۴۲ | ۳ اجتماع | 0.4389 | ۴ اجتماع | 0.4442 | ۵ اجتماع | 0.4397 | ۴ اجتماع | 0.4394 |
| متابولیکی | ۴۵۳ | ۲۰۳۲ | ۱۰ اجتماع | 0.4332 | ۱۳ اجتماع | 0.4269 | ۱۱ اجتماع | 0.4334 | ۱۰ اجتماع | 0.4019 |
| ایمیل | ۱۱۳۳ | ۵۴۵۱ | ۱۱ اجتماع | 0.5485 | ۱۰ اجتماع | 0.5766 | ۱۰ اجتماع | 0.5463 | ۱۱ اجتماع | 0.5036 |
| **NetScience** | ۱۵۸۹ | ۲۷۴۲ | ۴۰۷ اجتماع | 0.9547 | ۴۰۹ اجتماع | 0.9511 | ۴۰۵ اجتماع | 0.9472 | ۴۰۳ اجتماع | 0.9419 |

همان‌طوری که در جدول(۲) ملاحضه می‌شود روش‌های N، NG، GN و R روش‌های پایه برای تشخیص اجتماعات و روش‌های EAC، CTS، SRS و ASRS روش‌های ترکیبی تشخیص اجتماعات می‌باشند. در دو ستون AVG به ترتیب مقادیر میانگین نتایج روش پایه و ترکیبی آورده شده است . در سطر آخر از جدول(۲) میانگین اعمال نتایج مختلف یک الگوریتم یا روش خاص بر روی تمامی مجموعه داده‌ها آورده شده است که نشان دهنده نحوه اجرای و یا چگونگی نتایج آن الگوریتم یا روش مورد استفاده بر روی مجموعه داده‌های متفاوت می‌باشند.

همان‌طوری که در جدول(۲) مشاهده می‌شود اعمال روش ترکیبی نتایج دقیق‌تر و پایدارتری را به نسبت روش‌های پایه‌ای به ارمغان می‌آورند. در شکل(۵) به صورت نموداری این موضوع مشهودتر می‌باشد.

همانند جدول(۲) در زیر نمودار حاصل از شکل(۵) نیز روش-های N، NG، GN و R روش‌های پایه برای تشخیص اجتماعات و روش‌های EAC، CTS، SRS و ASRS روش‌های ترکیبی تشخیص اجتماعات می‌باشند. مقادیر میانگین نتایج روش پایه و ترکیبی به ترتیب با رنگ آبی کم رنگ و سبز کم رنگ آورده شده است که میانگین اعمال نتایج مختلف یک الگوریتم یا روش خاص تشخیص اجتماعات ترکیبی بر روی تمامی مجموعه داده‌ها آورده شده است.

همان‌طوری که مشاهده می‌شود نتایج حاصل دارای دقت بالاتر و نیز پایداری بیشتر نسبت به روش‌های پایه‌ای دارد.




**شکل(۵): نمودار حاصل از نتایج مختلف پایه و پیشنهادی**

**جدول (۲): ماژولاریتی محاسبه شده توسط روش‌های اولیه و مقایسه با روش‌های ECD**

| AVG | ECD | | | | AVG | Basic Community Detection Method | | | | Data Set |
|---|---|---|---|---|---|---|---|---|---|---|
| | ASRS | SRS | CTS | EAC | | R | GN | NG | N | |
| 0.4190 | 0.4179 | 0.4197 | 0.4188 | 0.4197 | 0.3982 | 0.3853 | 0.4009 | 0.3980 | 0.4086 | کاراته |
| 0.6002 | 0.5985 | 0.5985 | 0.6019 | 0.6019 | 0.5777 | 0.5581 | 0.5714 | 0.6027 | 0.5788 | فوتبال |
| 0.4445 | 0.4445 | 0.4438 | 0.4447 | 0.4450 | 0.4405 | 0.4394 | 0.4397 | 0.4442 | 0.4389 | جاز |
| 0.4363 | 0.4334 | 0.4374 | 0.4368 | 0.4377 | 0.4238 | 0.4019 | 0.4334 | 0.4269 | 0.4332 | متابولیکی |
| 0.5745 | 0.5685 | 0.5763 | 0.5765 | 0.5769 | 0.5437 | 0.5036 | 0.5463 | 0.5766 | 0.5485 | ایمیل |
| 0.9539 | 0.9534 | 0.9528 | 0.9546 | 0.9551 | 0.9487 | 0.9419 | 0.9472 | 0.9511 | 0.9547 | NetScience |
| 0.5714 | 0.5693 | 0.5714 | 0.5722 | 0.5727 | 0.5554 | 0.5383 | 0.5564 | 0.5665 | 0.5604 | AVG |

در شکل(۵) نتایج مختلف را بر روی نمودار نشان می‌دهد، با استفاده از این روش پیشنهادی می‌توان با اطمینان بیشتری به نتایج حاصل از روش‌های تشخیص اجتماعات اتکا نمود. چرا که نتایج روش‌های تشخیص اجتماعات می‌تواند در تصمیمات بلند مدت و یا کوتاه مدت و سیاست‌های سازمانی و یا مواردی از این قبیل تاثیر بسزای داشته باشد.

## ۴-۴- چالش‌ها

با توجه به اینکه روش پیشنهادی شامل دو مرحله پشت سر هم تولید نتایج اولیه و سپس ترکیب آنهاست، برای تحلیل پیچیدگی زمانی این روش باید پیچیدگی هر دو مرحله را تحلیل کرد. برای مرحله اول، پیچیدگی برابر با پیچیدگی بدترین الگوریتم استفاده شده در روش ترکیبی می‌باشد. بدین معنا که اگر از ۱۰ روش مختلف در روش‌های ترکیبی استفاده نماییم، به دلیل آنکه تک‌تک روش‌ها می‌بایست اجرا گردند تا نتایج حاصل شود و بعد از اجرا بتوان از ترکیب آن‌ها استفاده نمود، بنابراین پیچیدگی آن برابر با بدترین حالت در هر کدام از ۱۰روش مورد استفاده می‌باشد. در مرحله ترکیب نیز پیچیدگی زمانی برابر است با پیچیدگی زمانی توابع توافقی مورد استفاده که از متداول‌ترین روش‌های ترکیبی می‌باشند. برای بررسی بیشتر تحلیل پیچیدگی روش‌های ترکیبی مورد استفاده می‌توان به منابع [۶] و [۷] رجوع کرد.

روش پیشنهادی در نرم‌افزار مطلب شبیه‌سازی شده است و از ماتریس دو بعدی برای نمایش شبکه استفاده شده است. با توجه به این که انجام عملیات برروی ماتریس‌ها در نرم‌افزار مطلب نیاز به حجم حافظه زیادی دارد، بنابراین بررسی شبکه‌های بزرگتر از ۵۰۰۰ نود در این نرم‌افزار با مشکلات حافظه‌ای روبروست. برای رفع این مشکل و بررسی شبکه‌های با نودهای بیشتر می‌بایست اولاً عملیات برروی ماتریس‌ها را به صورتی اصلاح نمود که به حجم حافظه و عملیات کمتری نیاز داشته باشد و ثانیاً می‌توان ساختمان داده‌ای جدید تعریف کرد و آن را جایگزین ماتریس نماییم و یا اینکه از ماتریس اسپارس استفاده کنیم.

## ۵. نتیجه‌گیری

هدف این مقاله بهبود نتایج روش‌های تشخیص اجتماعات می‌باشد. روش مورد استفاده در این مقاله رویکرد ترکیبی می‌باشد، با توجه به قدرت روش‌های خوشه‌بندی ترکیبی و نتایج دقیق روش مذکور و امکان استفاده آن در روش‌های تشخیص اجتماعات در این مقاله از تشخیص اجتماعات ترکیبی برای بدست آوردن نتایج دقیق‌تر، مطمئن‌تر، پایدارتر و مستحکم‌تر استفاده شده است.

بنابراین در این مقاله یک روش جدید برای تشخیص اجتماعات ترکیبی پیشنهاد شده است. در ابتدای مقاله به بررسی شبکه‌های اجتماعی و تشخیص اجتماعات در آن پرداختیم و با معرفی خوشه‌بندی ترکیبی و ترکیب آن با روش‌های تشخیص اجتماعات روشی را با نام تشخیص اجتماعات ترکیبی را معرفی نمودیم. با پیاده‌سازی چهار روش متفاوت از روش‌های تشخیص اجتماعات و بدست آوردن نتایج آن‌ها و نیز در نهایت با استفاده از توابع توافقی در تشخیص اجتماعات ترکیبی که رویکرد اصلی این مقاله بوده است، به ترکیب نتایج بدست آمده از آن‌ها پرداختیم. برای ترکیب نتایج، یکی از روش‌های خوشه‌بندی ترکیبی با نام روش مبتنی بر ماتریس همبستگی، مورد استفاده قرار کرفته است.

همان‌گونه که مشاهده گردید از مزایای این روش استفاده از روش‌های قبلی به تعداد زیاد که نشان از مقیاس‌پذیری بالای این روش و نیز دقت بالای آن به نسبت روش‌های دیگر و در مجموع نتایج دقیق‌تر و پایدارتری را برای ما خواهد داشت. از معایب این روش به هزینه اجرایی آن که با بدترین حالت هزینه اجرایی در الگوریتم‌های استفاده شده برابر می‌باشد و نیز حافظه زیادی را نیز مصرف می‌نماید، همچنین ممکن است در همه جا بهینه‌ترین جواب را نداشته باشیم و الگوریتم هایی باشند که برای داده‌های خاصی جواب دقیق‌تری را داشته باشند و نیز در اعمال این روش می‌بایست تعداد خوشه‌ها نیز برای ما مشخص باشد ولی می‌توان از روش‌های دیگر خوشه‌بندی ترکیبی استفاده نمود که تعداد خوشه‌های در آن نامشخص باشد.

امکان استفاده از روش‌های دیگر توابع توافقی نیز وجود دارد، برای استفاده از آن روش‌ها دیگر به تعداد بیشتری نتایج گوناگون نیاز می‌باشد که با پیاده‌سازی‌های بیشتری از الگوریتم‌های تشخیص اجتماعات می‌توان از آن‌ها نیز استفاده نمود که می‌توان در کارهای بعدی به آن پرداخت. تشخیص اجتماعات ترکیبی می‌تواند در مسائلی از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه، بهبود موتورهای جستجو، ارتباطات(از قبیل تلفنی، سیار و...) و تعاملات افراد و فعالیت‌های امنیتی و جاسوسی مورد استفاده قرار گیرد.

## مراجع


[1] Easley, D. and Kleinberg, J. "*Networks, Crowds, and Markets: Reasoning about a Highly Connected World*", Cambridge University Press, 2011.

[2] Liu, B. "*Web DataMining Exploring Hyperlinks, Contents, and Usage Data*", Springer, 2007.

[3] Fortunato, S. "*Community detection in graphs*", Physics Reports, vol. 486, no. 3-5, pp. 75–174, 2010.

[4] Porter, M.A., Onnela, J.P. and Mucha, P.J. "*Communities in networks*", Notices of the American Mathematical Society, vol. 56, no. 9, pp. 1082–1097, 2009.

[5] Strehl, A. and Ghosh, J. "*Cluster Ensembles -A Knowledge Reuse Framework for Combining Multiple Partitions*", Journal of Machine Learning Research, pp. 583-617, 2003.

[6] Fred, A.L.N. and Jain, A.K. "*Data Clustering Using Evidence Accumulation*", Proc. of the 16th Intl. Conf.

## زیرنویس‌ها

[1] Community Detection
[2] Clustering
[3] Ensemble Clustering
[4] Social Network Analysis(SNA)
[5] Graph theory
[6] Node
[7] Edge
[8] Adjacency Matrix
[9] Diversity
[10] Consensus Function
[11] Co-assosiation Matrix
[12] Evidence Accumulation Clustering
[13] Heuristic
[14] Spectral
[15] Divisive
[16] Agglomerative
[17] Modularity optimization
[18] Greedy techniques
[19] Simulated annealing
[20] Evolutionary
[21] Targin and Bingol
[22] Single Linkage(SL) or Average Linkage(AL)
[23] Connected Triple based Similarity
[24] SimRank based Similarity
[25] Approximate SimRank based Similarity
[26] Zachary karate club
[27] Jazz bands
[28] American College Football League
[29] Federal Energy Regulatory Commission


on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.

[7] Iam-on, N. and Garrett, S. "*LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles*", In Journal of Statistical Software, Issue 9, Volume 36, 2011.

[8] Lancichinetti, A. "*Community detection algorithms: a comparative analysis*", Physical Review E, vol. 80, no. 5, p. 056117, 2009.

[9] Girvan, M. and Newman, M.E.J. "*Community structure in social and biological networks*", Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 12, p. 7821, 2002.

[10] Newman, M.E.J. "*A measure of betweenness centrality based on random walks*", Social networks, vol. 27, no. 1, pp. 39–54, 2005.

[11] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. "*Defining and identifying communities in networks*", Proceedings of the National Academy of Sciences of the United States of America, pp. 2658-2663, 2004.

[12] Lipczak, M. and Milios, E. "*Agglomerative genetic algorithm for clustering in social networks*", In Proceedings of the 11th Annual conference on Genetic and evolutionary computation, pp. 1243–1250, 2009.

[13] Zhang, X.S. and et al. "*Modularity optimization in community detection of complex networks*", EPL (Euro Physics Letters), vol. 87, p. 38002, 2009.

[14] Newman, M.E.J. "*Fast algorithm for detecting community structure in very large networks*", Physical review E, vol. 69, 2004.

[15] Newman, M.E.J. and Girvan, M. "*Finding and evaluating community structure in networks*", Physical review E, vol. 69, no. 2, p. 26113, 2004.

[16] Clauset, A., Newman, M.E.J. and Moore, C. "*Finding community structure in very large networks*", Physical Review E, vol. 70, no. 6, p. 66111, 2004.

[17] Tasgin, M. and Bingol, H. "*Community detection in complex networks using genetic algorithm*", Arxiv preprint cond-mat/0604419, 2006.

[18] Leskovec, J., Lang, K.J. and Mahoney, M. "*Empirical comparison of algorithms for network community detection*", In Proceedings of the 19th International conference on World Wide Web, pp. 631–640, 2010.

[19] Reichardt, J. and Bornholdt, S. "*Statistical mechanics of community detection*", Physical Review E, 2006.

[20] Zachary, W.W. "*An information flow model for conflict and fission in small groups*", Journal of Anthropological Research, vol. 33, no. 4, pp. 452–473, 1977.

[21] Gleiser, P. and Danon, L. "*Community structure in jazz*", Arxiv preprint condmat/0307434, 2003.

[22] Shetty, J. and Adibi, J. "*The Enron email dataset database schema and brief statistical report*", Information Sciences Institute Technical Report, University of Southern California, 2004.