JCWR2022
May 11-12, 2022; Tehran, Iran
8th International Conference on Web Research

A Label Propagation Based Algorithm for Community Detection in Social Networks

Azadeh Tabatabaei*[1], Mozhgan Zandi Nezhad[2]

[1] Assistant Professor of Computer Engineering Department, Science and Culture University, Tehran, Iran
a.tabatabaei@usc.ac.ir

[2] Master of Computer Engineering Department, Science and Culture University, Tehran, Iran
m.zandinezhad@usc.ac.ir

Abstract

Social network analysis with large volumes of data and complex communication structures is so difficult and time-consuming. Community detection is one of the major challenges in network analysis. A community is a set of individuals or organizations whose communication density is more than other network entities. Community detection or clustering can reveal the structure of groups in social networks, or relationships between entities. The label propagation algorithms with neighbor node influence have less complexity than traditional algorithms, such as clustering, to recognize communities. Also, the algorithms can identify overlapping communities. In our label propagation algorithm, which is based on the neighbor node influence, important nodes are more likely to publish their labels, while less important nodes have a small chance of spreading the label. The degree of similarity of nodes and the effect of nodes in a social network depends on the parameter of path length between nodes. In the proposed method, increasing this parameter leads to more accurate identification of overlapping and stable communities. The proposed algorithm detects overlapping communities with the same accuracy as the previous algorithms with fewer iterations, in less time. The algorithm is implemented on real and artificial social networks with weightless graphs and weighted graphs with weighting by Jacquard similarity criterion, in all of which the execution time is improved.

Keywords: Social Networks, Community Detection, Label Propagation Algorithm, Overlapping Communities

ICWR2022
May 11-12, 2022; Tehran, Iran
8th International Conference on Web Research

الگوریتمی مبتنی بر انتشار برچسب در تشخیص اجتماعات در شبکه‌های اجتماعی

آزاده طباطبائی*[1]، مژگان زندی نژاد[2]

[1] استادیار گروه مهندسی کامپیوتر دانشگاه علم و فرهنگ، تهران،
a.tabatabaee@usc.ac.ir

[2] کارشناسی ارشد گروه مهندسی کامپیوتر دانشگاه علم و فرهنگ، تهران،
m.zandi@usc.ac.ir

چکیده

تجزیه و تحلیل شبکه‌های اجتماعی با حجم زیادی از داده‌ها و ساختار ارتباطی پیچیده، فرایند بسیار دشوار و زمان‌بری است. تشخیص اجتماعات یکی از چالش‌های مهم در تحلیل شبکه‌های اجتماعی است. اجتماع در واقع مجموعه‌ای از افراد یا سازمان‌ها است که چگالی ارتباطی آن‌ها با هم بیشتر از سایر موجودیت‌های شبکه است. تشخیص اجتماعات و یا خوشه‌بندی می‌تواند ساختار گروه‌ها در شبکه‌ها و یا ارتباطات بین موجودیت‌ها را آشکار نماید. الگوریتم‌های انتشار برچسب بر مبنای تاثیر گره همسایه نسبت به الگوریتم‌های سنتی تشخیص اجتماعات مانند خوشه‌بندی پیچیدگی کمتری دارند که قادر به شناسایی جوامع همپوشان نیز هستند. در روش پیشنهادی ما که براساس الگوریتم انتشار برچسب بر مبنای تاثیر گره همسایه است، گره‌های مهم‌تر با احتمال بیشتری برچسبشان را منتشر می‌کنند، ضمن اینکه گره‌های کم اهمیت هم شانس انتشار برچسب را دارند. میزان مشابهت گره‌ها و تاثیر گره‌ها در یک شبکه اجتماعی وابسته به پارامتر طول مسیر بین گره‌ها است. در روش ارائه شده افزایش این پارامتر منجر به شناسایی دقیق‌تر جوامع همپوشان وپایدار می‌شود. الگوریتم پیشنهادی با تعداد تکرار کمتر، در نتیجه در زمان کمتر جوامع همپوشان را با همان دقت الگوریتم‌های قبلی شناسایی می‌کند. الگوریتم در شبکه‌های اجتماعی واقعی و مصنوعی با گراف بدون وزن و گراف وزن‌دار با وزن دهی توسط معیار شباهت جاکارد پیاده‌سازی شده است، که در همه آن‌ها زمان اجرا بهبود یافته است.

کلمات کلیدی

شبکه‌های اجتماعی، تشخیص اجتماعات، الگوریتم‌های انتشار برچسب، اجتماعات همپوشان

۱- مقدمه

با توسعه فناوری اطلاعات، استفاده از رسانه‌های اجتماعی برای کاربردهای مختلف، به یک امر اجتناب ناپذیر در زندگی بشر تبدیل شده است. در اکثر رسانه‌های اجتماعی کاربران با تبادل ایده در جوامع مجازی با یکدیگر در تعامل هستند. شبکه‌های اجتماعی با مقیاس بزرگ، با ده‌ها هزار گره یا میلیون‌ها گره ساختار اجتماعی پیچیده‌ای دارند. یکی از مسائلی که در تحلیل شبکه‌های پیچیده بزرگ مطرح می‌شود تشخیص ساختار جامعه است. به این ترتیب که یک شبکه به گروه‌هایی (مجزا یا همپوشان) که هرکدام شامل مجموعه‌ای از کاربران است، تجزیه می‌شود. این کاربران معمولا وابستگی ویژه‌ای به نسبت سایر کاربران به همدیگر دارند. فرایند کشف این گروه‌ها را تشخیص جامعه می‌نامند.

تشخیص ساختارهای اجتماعی در شبکه‌های پیچیده در مقیاس بزرگ به تجزیه و تحلیل خصوصیات گروهی کمک می‌کند. تشخیص اجتماعات یکی از مهم‌ترین حوزه‌های تحقیقاتی در زمینه تحلیل شبکه‌های اجتماعی است. ارائه اطلاعات درون اجتماع یا توصیه محصولات به اجتماع، کاربردهای تجاری، اجتماعی، آموزشی و غیره دارد. همچنین اجتماعات مختلف، می‌توانند برای تبلیغات هدفمند، جرم شناسی، بهداشت عمومی و سیاست استفاده شوند[1].

از ویژگی‌های منحصر به فرد جوامع، عضویت چندگانه یک کاربر به گروه‌های متفاوت است که منجر به ایجاد جوامع همپوشان می‌شود. به این ترتیب که یک عضو شبکه می‌تواند وابستگی به چندین اجتماع داشته باشد. برای مثال، یک نویسنده در یک شبکه همکاری ممکن است با چندین گروه

تحقیقاتی همکاری کند. یکی دیگر از ویژگی‌های مهم شبکه‌های اجتماعی امکان تکامل و تغییرآنها در طول زمان است، که از اصطلاح شبکه‌های پویا برای اشاره به آنها استفاده می‌شود [2].

در سال‌های اخیر، محققان الگوریتم‌های بسیاری برای تشخیص اجتماعات، در تجزیه و تحلیل شبکه‌های اجتماعی پیشنهاد کرده‌اند[3]. در ابتدا، محققان برخی از روش‌های سنتی مانند روش خوشه‌بندی۱، الگوریتم نیومن۲ و غیره را پیشنهاد کردند. تعلق افراد در یک شبکه اجتماعی به چندین اجتماع باعث شده است که بسیاری از محققان به تشخیص اجتماعات همپوشان روی بیاورند، در شکل (۱) جوامع همپوشان دیده می‌شوند. با توسعه استفاده از شبکه‌های اجتماعی، آنها بسیار پیچیده‌تر و بسیار بزرگ‌تر می‌شوند. روش‌های سنتی متمرکز بر تقسیم کل گراف شبکه اجتماعی به تعدادی گروه هستند که در شبکه‌های پیچیده و بزرگ با ده‌ها هزار گره یا میلیون‌ها گره، به لحاظ هزینه کارایی زیادی ندارند[5,4]. بنابراین برخی محققان تشخیص اجتماعات محلی۳ را مورد مطالعه قرار می‌دهند [6] که از میان آنها الگوریتم‌های انتشار برچسب با توجه به سرعت بالا و سهولت پیاده‌سازی برای شبکه‌های اجتماعی بزرگ می‌توانند مناسب‌تر باشند[7].

الگوریتم‌های انتشار برچسب معمولا با اختصاص دادن یک برچسب منحصرد به فرد به هر عضو آغاز می‌شود و طبق فرآیندی برچسب برخی از گره‌ها به سایر گره‌ها منتشر می‌شوند و در پایان افرادی که در یک جامعه هستند دارای برچسب‌های یکسان خواهند بود. همچنین در الگوریتم‌هایی که جوامع همپوشان را شناسایی می‌کنند یک عضو شبکه در پایان می‌تواند چند برچسب داشته باشد. در شکل (۱)، گره۱ وابستگی به دو جامعه مجاور دارد.

در سال 2007 کومارا و همکاران یک الگوریتم انتشار برچسب براساس ساختار محلی شبکه‌ها برای شناسایی جوامع که فقط قادر به شناسایی جوامع مجزا بود، ارائه کردند[6]. الگوریتم‌هایی نیز با اولویت دادن به گره‌های با درجه بیشتر، اطلاعات معنایی بیشتر و مرکزیت ارائه شدند که آنها نیز قادر به شناسایی جوامع همپوشان نبودند[8, 9].

گریگوری یک الگوریتم انتشار برچسب برای تشخیص جوامع با همپوشانی ارائه داد. به این ترتیب که، کار با دادن برچسب منحصر به فرد با ضریب تعلق ۱ به هر راس آغاز می‌شود و سپس هر راس برچسب‌های خود را با تجمیع و نرمال کردن ضرایب تعلق راس‌ها در مجموعه همسایه‌هایش بروزرسانی می‌کند. پس از انتشار، در پایان هر راس برچسب‌های متعددی دارد[10].

گسترش دیگری از تکنیک انتشار برچسب در [11]، ارائه شد که آن هم می‌تواند جوامع با همپوشانی را شناسایی کند. هر گره برچسب منحصر به فرد خود را دارد و سپس یک گره به عنوان شنونده انتخاب می‌شود. هر همسایه از گره منتخب به طور تصادفی یک برچسب انتخاب کرده و برچسب انتخاب شده را به شنونده ارسال می‌کند. شنونده یکی از برچسب‌های منتشر شده را با توجه به قانون شنونده می‌پذیرد. در نهایت، از یک آستانه۴ برای ایجاد اجتماعات همپوشان به عنوان پس پردازش۵ استفاده می‌شود، این روش دقت و پایداری به نسبت پایینی در شناسایی جوامع دارد. یک الگوریتم انتشار برچسب مبتنی بر همین روش برای شناسایی جوامع در شبکه‌های پویا ارائه شده که با تغییر ارتباطات (یال‌های گراف) جوامع را تغییر می‌دهد، ولی تغییرات در گره‌ها را لحاظ نمی‌کند[12].

در پژوهش انجام شده در [13]، از فرایند رای گیری و در نظر گرفتن برچسب غالب۶ بین برچسب‌های ارسال شده به هر گره برای تشخیص اجتماعات استفاده می‌شود، این روش هم پیچیدگی فضایی بالایی دارد. با ایده گرفتن از روش‌های اخیر در سال ۲۰۱۸، یک الگوریتم انتشار برچسب مبتنی بر تاثیر گره همسایه۷ با در نظر گرفتن برچسب غالب، برای غلبه بر ضعف‌هایی مانند دقت پایین، ناپایداری ارائه شد[4]. در این روش با اتخاذ توالی انتشار برچسب بر مبنای ترتیب صعودی اهمیت گره‌ها، برچسب‌ها محاسبه شده و بروزرسانی می‌شوند. در نهایت از روی تاریخچه برچسب‌های هر گره، ساختار جوامع همپوشان را شناسایی می‌کند.

در این پژوهش یک الگوریتم انتشار برچسب مبتنی بر تاثیر گره همسایه ارائه می‌دهیم، به طوری‌که گره‌های مهم‌تر با احتمال بیشتری برچسبشان را منتشر می‌کنند، ضمن اینکه گره‌های کم اهمیت هم شانس انتشار برچسب را دارند. الگوریتم پیشنهادی با تعداد تکرار کمتر، در نتیجه در زمان کمتر جوامع همپوشان را با همان دقت الگوریتم قبل شناسایی می‌کند. تغییر پارامتر طول مسیر در محاسبه میزان مشابهت گره‌ها و تاثیر گره‌ها در یک شبکه اجتماعی منجر به شناسایی دقیق‌تر جوامع همپوشان و پایدار در الگوریتم پیشنهادی می‌شود.

در ادامه ساختار مقاله بدین صورت است. در بخش دوم، مفاهیم پایه‌ای مرتبط با الگوریتم‌های انتشار بیان می‌شود. در بخش سوم، الگوریتم اصلی تشریح می‌شود. در بخش چهارم، نتایج به دست آمده از آزمایش‌ها ارائه می‌گردد. در پنجم نتیجه‌گیری و کارهای آتی آمده است.

۲- مفاهیم مورد نیاز

با توسعه اینترنت و علوم رایانه، افراد بیشتری به شبکه های اجتماعی می‌پیوندند. افراد با یکدیگر ارتباط برقرار می‌کنند و نظرات خود را در رسانه‌های اجتماعی بیان می‌کنند، که این رابطه‌ها شبکه‌های پیچیده‌ای را تشکیل می‌دهند. در شبکه‌های اجتماعی ساختار رابطه‌ای۸، از طریق ارتباطات مختلفی تشکیل می‌شود که باعث انتشار اطلاعات زیادی در شبکه می‌شود. تشخیص اجتماع برای آشکار ساختن ساختار شبکه‌ها، بررسی نظرات گروه‌های مردمی، تجزیه و تحلیل اطلاعات و درک و کنترل احساسات عمومی بسیار مهم است[7].

در تحلیل شبکه‌های اجتماعی، شبکه‌ها به صورت گراف‌هایی طراحی می‌شوند که در آن هر گره (راس) نشان دهنده یک کاربر شبکه است و لبه‌های گراف (خطوط بین این گره‌ها) نشان دهنده روابط آنها، مانند دوستی یا تعامل متقابل است شکل (۱) را ببینید. اکنون سه معیار اساسی، اهمیت گره (NI)، شباهت بین گره‌ها (Sim)، تاثیر گره همسایه (NNI) را توضیح می‌دهیم. این معیارها در توالی انتشار برچسب در الگوریتم‌های انتشار برچسب و همچنین راهکار پیشنهادی در این پژوهش اهمیت دارند [13,4] .

۲-۱- اهمیت گره (NI)

برای یک شبکه مشخص با گراف G=(V,E)، اهمیت گره، احتمال اینکه یک گره به طور بالقوه مرکز یک اجتماع باشد را بیان می‌کند. چنین گره‌ای ویژگی‌های زیر را دارد: این گره با گره‌های همسایه بیشتری ارتباط دارد و

همسایه‌های این گره نیز با یکدیگر ارتباط دارند. هر چه گره‌های همسایه یک گره بیشتر به هم متصل باشند، و درجه گره بزرگ‌تر باشد، مثلث‌های بیشتری توسط گره و گره‌های مجاور آن شکل می‌گیرد.

همانطور که در شکل (۱) نشان‌داده‌شده، گره ۵ دارای لبه‌های بیشتری متصل به گره‌های همسایه خود نسبت به گره ۳ است، و گره ۵ با همسایه‌هایش بیشتر از گره ۳ مرتبط است، بنابراین گره ۵ به عنوان مرکز یک جامعه بالقوه تعیین می‌شود. براساس فرضیات بالا، ما از درجه گره $u \in V$ و تعداد مثلث‌های تشکیل‌شده توسط گره و گره‌های همسایه آن برای اندازه‌گیری اهمیت گره استفاده می‌کنیم. بنابراین اهمیت گره به این صورت در فرمول (۱) تعریف می‌شود:

$$NI(u) = \frac{1}{2} + \frac{1}{2} * \frac{(e_u + k_u) - \min_{i \in V}(e_i + k_i)}{\max_{j \in V}(e_j + k_j) - \min_{i \in V}(e_i - k_i)} \quad (۱)$$

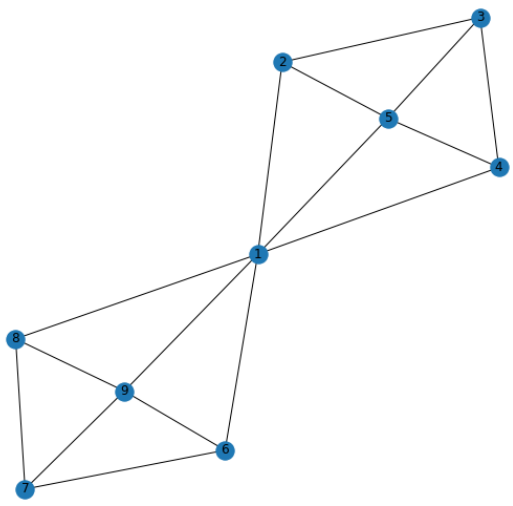
در فرمول (۲) e_u نشان‌دهنده تعداد مثلث حاوی گره u است، و k_u در فرمول (۳) درجه گره u، برای شبکه‌های وزن دار با گراف G=(V,E,W) است.

$$e^u = \sum_{x,y \in Ng(u), e_{xy} \in E} \frac{W(u,x)+W(u,y)+W(x,y)}{3} \quad (۲)$$

$$K_u = \sum_{v \in Ng(u)} W(u,y) \quad (۳)$$

تعریف می‌شوند. $Ng(u)$ نشان دهنده مجموعه گره‌های مجاور گره u و $W(x,y)$ نشان دهنده وزن یال e_{xy} بین گره‌های $x, y \in V$ است.

بنابراین e_u میانگین وزن سه یال در تمام مثلث‌هایی است که حاوی گره u هستند. همچنین، k_u مجموع وزن همه یال‌هایی است که با گره u در ارتباط هستند. بنابراین، تعریف اهمیت گره‌ها می‌تواند، برای شبکه‌های وزن‌دار هم اعمال شود.



شکل (۱): یک شبکه اجتماعی و وابستگی گره ۱ به دو جامعه مجاور

۲-۲- تشابه بین گره‌ها (Sim)

با استفاده از معیار تشابه بین گره‌ها[۹]، قدرت اتصال بین دو گره مجاور اندازه‌گیری می‌شود و میزان تاثیر پذیری آن‌ها از یکدیگر مشخص می‌شود. در الگوریتم پیشنهادی از یکی از رایج‌ترین معیارهای شباهت که براساس α-path[۱۰]، (مسیرهای به طول α) است، استفاده می‌کنیم [4]. این معیار به دلیل استفاده از تمام مسیرهای بین دو گره، پیچیدگی محاسباتی نسبتا بالایی همانند سایر معیارها دارد.

همانطور که در فرمول (۴) نشان داده شده است، نوعی از شباهت جاکارد، که از سطح آستانه طول مسیر α برای کنترل پیچیدگی محاسبات استفاده می‌کند، به طور کارایی، تاثیر اختلاف درجه بین دو گره بر تشابه را تشخیص می‌دهد.

در فرمول (۵)، p نشان‌دهنده یکی از مسیرهای بین گره‌های i, j است و $|p|$ نشان دهنده طول مسیر p است که طول آن بین ۱ و α است.

برای شبکه های بدون وزن $A^{|p|} = 1$ است و مقدار آن برای شبکه های وزن دار در فرمول (۶) مجموع وزن یال‌هایی است که مسیر p را می‌سازند.

$$sim(u,v) = \frac{s(u,v)}{\sqrt{\sum_{x \in Ng(u)} s(u,x) * \sum_{y \in Ng(v)} s(v,y)}} \quad (۴)$$

$$s(i,j) = \sum_{|p|=1}^{\alpha} \frac{(A^{|p|})_{ij}}{|p|} \quad (۵)$$

$$A^{|P|} = \frac{\sum_{x,y \in P \cap e_{xy} \in E} W_{xy}}{|P|} \quad (۶)$$

۲-۳- تاثیر گره همسایه (NNI)

تاثیر گره همسایه v بر گره u به صورت (۷) تعریف می‌شود:

$$NNI_v(u) = \sqrt{NI(v) * \frac{sim(v,u)}{\max_{h \in Ng(u)} sim(h,u)}} \quad (۷)$$

در فرمول (۷)، $Ng(u)$ نشان دهنده مجموعه گره‌های همسایه گره u می‌باشد[۴]. NI و Sim بین گره‌ها برای محاسبه NNI ترکیب می‌شود و از NNI برای اندازه‌گیری تاثیر گره‌های همسایه بر روی گره‌هایی که نیاز بروزرسانی برچسب‌ها دارند، استفاده می‌شود، به طوری که برچسب‌های گره می‌توانند به طور معقول و منطقی بروز شوند. گره‌های همسایه با NI و Sim بزرگتر تأثیر بیشتری برای بروزرسانی برچسب‌ها دارند. برای مثال در گراف شکل (1)، گره ۲ مطابق با فرمول (7)، تاثیرات $NNI_3(2) = 0.66$ $NNI_1(2) = 0.80$ $NNI_5(2) = 0.89$ را از سه گره همسایه خود به طور جداگانه می‌گیرد. بدیهی است با توجه به مقادیر بدست آمده گره ۵ تاثیر بیشتری نسبت به گره ۱ و ۳ دارد، همچنین از روی خود گراف با توجه به کوچک بودن آن هم می‌توان متوجه شد که تاثیر گره ۵ نسبت به گره ۱ و بعد از آن گره ۳ بیشتر می‌باشد.

۳- الگوریتم پیشنهادی

در ابتدا برای همه گره‌های متعلق به شبکه کمیت‌های $NI(u)$، $Sim(u)$ و $Ng(u)$ محاسبه می‌شود. سپس گره‌ها بر اساس معیار $NI(u)$ به صورت صعودی مرتب می‌شوند.

در فاز دوم، در ابتدا به هر گره مجموعه برچسب $L_u = \{1, u\}$ نسبت داده می‌شود (در ابتدا هر گره متعلق به جامعه مختص خودش با ضریب تعلق ۱ است.) سپس در T تکرار برچسب گره‌های مجاورش در آن منتشر می‌شود. به طوریکه در هر تکرار به مجموعه‌ای از جوامع با ضرایب تعلق خاصی وابسته می‌شود. از نماد b(c,u) برای تعریف ضریب وابستگی گره u به جامعه c استفاده می‌شود، یعنی، احتمال اینکه گره u متعلق به جامعه c باشد. در زمان انتشار برچسب هر گره، برچسب جامعه‌ای که بیشترین ضریب تعلق به آن را دارد، منتشر می‌کند.

به نظر می‌آید که هرچه اهمیت گره (کمیت $NI(u)$) بیشتر باشد، احتمال اینکه این گره مرکز اجتماع باشد، بیشتر است. علی رغم الگوریتم‌های قبلی [13,4] که راس‌های شبکه به صورت تصادفی یا به ترتیب بر اساس $NI(u)$، برای دریافت برچسب از گره‌های مجاور، انتخاب می‌شدند، ما در این پژوهش گره‌ها را به صورت تصادفی ولی متناسب با مقدار $NI(u)$ انتخاب می‌کنیم بنابراین هر گره که $NI(u)$ بیشتری/کمتری دارد، با احتمال بیشتری برای انتشار/دریافت برچسب انتخاب می‌شود. بعد از اینکه یک گره u انتخاب می‌شود مجموعه برچسب L_{Ng} را از گره‌های مجاورش دریافت می‌کند:(فرمول (۸))

$$L_{Ng} = \{l(c_1, b_1), l(c_2, b_2), \dots l(c_v, b_v)\}, v \in Ng(u) \quad (8)$$

که در فرمول (۸) $l(c_v, b_v)$ بیانگر برچسب غالب همسایه گره v، b_v نشان‌دهنده ضریب تعلق گره v به اجتماع c_v می‌باشد.

حال، براساس NNI، از روی L_{Ng}، ضریب تعلق جدید $b'(c,u)$ گره u، به اجتماع c طبق فرمول فرمول (۹) محاسبه می‌شود:

$$b'(c,u) = \frac{\sum_{l(c_v,b_v)\in L_{Ng}, v\in Ng(u), c_v=c} b(c_v,v)*NNI_V(u)}{\sum_{l(c_v,b_v)\in L_{Ng}, v\in Ng(u)} b(c_v,v)*NNI_V(u)} \quad (9)$$

با استفاده از برچسب‌هایی که از گره‌های همسایه گرفته است ضریب تعلقش به اجتماعاتی که همسایه‌هایش به آن‌ها تعلق دارند را پیدا می‌کند.سپس مجموعه برچسب L' گره u شامل ضرایب تعلق گره u به اجتماع‌های مختلف که همسایه‌هایش به آن‌ها تعلق دارند ایجاد می‌شود.

$$L' = \{l(c_1, b'_1), l(c_2, b'_2), \dots l(c_{|L'|}, b'_{L'|})\}, \quad \sum_{l(c,b)\in L'} b'(c,u) = 1 \quad (10)$$

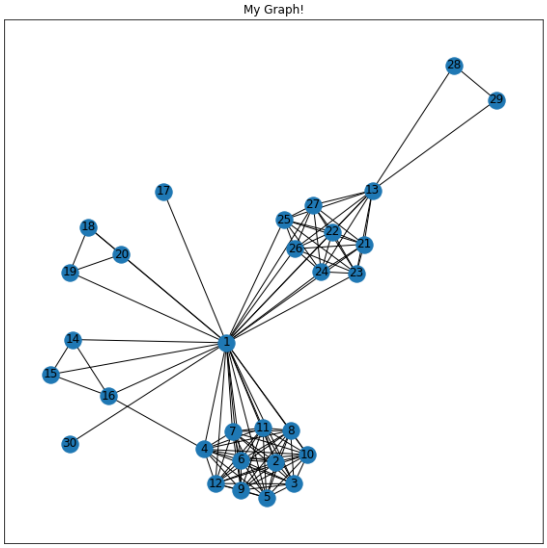
که در فرمول (۱۰) $|L'|$ تعداد برچسب ها در L' می‌باشد. سپس برچسب‌های بی‌مصرف را که $b'(c,u) < 1 < 1/|L'|$ است، حذف کرده و برچسب‌های باقی مانده، مجموعه برچسب L^* را تشکیل می‌دهند.

در نهایت برای هر گره u برچسبی که با بیشترین ضریب تعلق به آن وابسته است، به عنوان برچسب غالب گره u در نظر گرفته می‌شود. اگر برچسب‌های متعددی با حداکثر ضرایب تعلق وجود داشته باشد و یکی از آن‌ها برچسب غالب در تکرار قبلی هم باشد، آنگاه این برچسب غالب به عنوان برچسب غالب برتر فعلی انتخاب می‌شود، در غیر این صورت به طور تصادفی یکی را به عنوان برچسب برتر انتخاب می‌کند.

٤- پیاده سازی و بررسی نتایج

برای بررسی عملکرد الگوریتم، از مجموعه داده شبکه‌های مصنوعی و همچنین از سه مجموعه داده واقعی شامل دلفین [14]، فوتبال [15] و مجموعه داده ساختگی خودمان استفاده کرده‌ایم. علت استفاده از شبکه‌های مصنوعی، این است که انجام تست کدهای مدنظر، بر روی آن‌ها آسان‌تر و دقیق تر انجام می‌شود. در واقع برای بسیاری از شبکه‌های واقعی، ساختارهای اجتماعی واقعی معمولا ناشناخته هستند، که منجر به ناتوانی در ارزیابی واقعی الگوریتم می‌شود. پیاده سازی الگوریتم پیشنهادی در محیط گوگل کولب انجام شد. کل ارزیابی‌ها در گیت‌هاب [16] قرار داده شده است. به عنوان نمونه برای شکل (۲)، جوامع شناسایی شده توسط الگوریتم پیشنهادی روی مجموعه داده خانوادگی که تهیه کرده ایم، در حالت بدون وزن، در جدول (۱) آمده است.

برای ارزیابی کارایی الگوریتم مطرح شده، نتایج به‌دست آمده از لحاظ زمان اجرا، دقت و پایداری، معیار F [11] و صحت[12] و پوشش[13] و NMI و NMI_max و omega مورد بررسی قرار گرفته‌اند.



شکل (۲): گراف شبکه خانوادگی

جدول (۱): نتیجه تشخیص اجتماعات، گره ۱ به دو جامعه تعلق دارد.

Community
1 2 3 4 5 6 7 8 9 10 11 12 17 30
1 13 21 22 23 24 25 26 27
16 14 15
18 19 20
28 29

در تمامی شاخص‌های فوق الگوریتم پیشنهادی ما با الگوریتم قبلی برابری می‌کند [16]. نکته قابل توجه دیگر این است که با توجه به اینکه ما، علی رغم پژوهش‌های قبلی که گره‌ها برای انتشار برچسب‌شان با توجه به اهمیتشان در صف انتشار قرار می‌گیرند یا کاملا تصادفی انتخاب می‌شوند، آن‌ها را به صورت احتمالاتی متناسب با اهمیتشان انتخاب می‌کنیم، یعنی گره‌های کم اهمیت هم شانس انتشار برچسب دارند. این باعث شد که بتوانیم در زمان کمتری، ۶۰ تکرار به جای ۱۰۰ تکرار و در زمان تقریبا نصف زمان پژوهش قبلی به همان نتایج برسیم.

پارامتر α (طول مسیر بین دو گره) در محاسبه میزان مشابهت و میزان تاثیرپذیری گره‌ها دخیل است. با قرار دادن مقادیر مختلف برای α به این نتیجه رسیدیم که زمانی که طول مسیر ۱ یا ۲ باشد مقادیر گره‌های همپوشان و اجتماعات مانند هم هستند ولی زمانی که طول مسیر به ۳ یا ۴ تغییر پیدا می‌کند، گره‌های همپوشان و اجتماعات همپوشان تغییر کرده و گره‌ها و اجتماعات دقیق‌تر با همپوشانی بیشتر شناسایی می‌شود. علاوه براین باید به این نکته توجه شود که هزینه اجرای الگوریتم، با افزایش این پارامتر بالا می‌برد و نمی‌توان آن را خیلی افزایش داد.

۵- نتیجه‌گیری

در این پژوهش الگوریتمی مبتنی بر انتشار برچسب بر اساس سه شاخص اهمیت گره، شباهت بین گره‌ها و تاثیر گره همسایه برای تشخیص اجتماعات همپوشان در شبکه‌های اجتماعی ارائه دادیم. ویژگی اصلی الگوریتم ما نسبت به الگوریتم‌های مشابه قبلی تاثیر اهمیت گره‌ها در انتشار برچسب است، بدین ترتیب که گره‌های مهم‌تر با احتمال بیشتری برای انتشار برچسب انتخاب می‌شوند و این باعث می‌شود که گره‌های کم اهمیت هم شانس انتشار برچسب را داشته باشند. این تکنیک سبب شد که الگوریتم به همان نتیجه قبلی ولی با تعداد تکرار کمتر و در نتیجه در زمان کمتری جوامع همپوشان و غیر همپوشان را شناسایی کند. همچنین با تغییر پارامتر α (طول مسیر بین دو گره) الگوریتم ما قادر است تا همپوشانی جوامع را با دقت بالاتری تشخیص دهد. در ادامه کار می‌توان معیارهای شباهت، اهمیت و تاثیرگذاری را، در راستای بهبود زمان اجرای الگوریتم تغییر داد و به گونه مناسب‌تری تعریف کرد.

مراجع

[1] A. Karataş and S. Şahin, "Application areas of community detection: A review," in *2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT)*, Turkey, 2018.

[2] E. M. Win and M. A. Khine, "A Study on Community Overlapping Detection Algorithms in Social Networks," MERAL Portal-University of Computer Studies, Yangon, 2020.

[3] L. Tang and H. Liu, "Community detection and mining in social media," *Synthesis lectures on data mining and knowledge discovery*, vol. 2, no. 1, pp. 1-137, 2010.

[4] M. Lu, Z. Zhang, Z. Qu and Y. Kang, "LPANNI: Overlapping community detection using label propagation in large-scale complex networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 9, pp. 1736-1749, 2018.

[5] S. E. Garza and S. E. Schaeffer, "Community detection with the Label Propagation Algorithm: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122058, 2019.

[6] U. N. Raghavan, R. Albert and S. and Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[7] C. Wang, W. Tang, B. Sun, J. Fang and Y. Wang, "Review on community detection algorithms in social networks," in *2015 IEEE international conference on progress in informatics and computing (PIC)*, China, 2015.

[8] Y. Xing, F. Meng, Y. Zhou, M. Zhu, M. Shi and G. Sun, "A node influence based label propagation algorithm for community detection in networks," *The Scientific World Journal*, 2014.

[9] Z. Lin, X. Zheng, N. Xin and D. Chen, "CK-LPA Efficient community detection algorithm based on label propagation with community kernel," *Physica A: Statistical Mechanics and its Applications*, vol. 416, pp. 386-399, 2014.

[10] S. Gregory, "Finding overlapping communities in networks by label propagation," *New journal of Physics*, vol. 12, no. 10, p. 103018, 2010.

[11] J. Xie, B. K. Szymanski and X. Liu, "Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *2011 ieee 11th international conference on data mining workshops*, Vancouver, 2011.

[12] N. Aston, J. Hertzler and W. Hu, "Overlapping community detection in dynamic networks," *Journal of Software Engineering and Applications*, vol. 7, no. 10, p. 872, 2014.

[13] H.-L. Sun, J.-B. Huang, Y.-Q. Tian, Q.-B. Song and H.-L. and Liu, "Detecting overlapping communities in networks via dominant label propagation," *Chinese Physics B*, vol. 24, no. 1, p. 018703, 2015.

[14] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl_2, pp. S186-S188, 2003.

[15] J. Park and M. E. Newman, "A network-based ranking system for US college football," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 10, p. P10014, 2005.

[16] "HYPERLINK," LPANNI, [Online]. Available: https://github.com/mojganzandi/LPANNI.

زیرنویس‌ها

[1] Clustering
[2] Newman's algorithm
[3] Local community detection
[4] Threshold
[5] Post processing
[6] Dominant label propagation algorithm
[7] LPANNI
[8] Relation structure

[9] Similarity between nodes
[10] α–path similarity
[11] FScore
[12] Precision
[13] Recall