

تجزیه و تحلیل تطبیقی معیارهای ساختاری گرافیکی برای شناسایی ناهنجاری‌ها در شبکه‌های اجتماعی آنلاین

DOR: 20.1001.1.27832570.1399.1.1.6.0

مجتبی اعجمی*؛ ناصر عسگری؛

۱- دانشکده مهندسی برق و کامپیوتر - دانشگاه آزاد اسلامی واحد زنجان - زنجان - ایران - aajami@iauz.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه آزاد اسلامی واحد زنجان - زنجان - ایران - asgari.naser@gmail.com

چکیده: شبکه‌های اجتماعی به دلیل استفاده وسیع و محبوبیت آن‌ها در معرض مشکلات امنیتی هستند. بنابراین، شناسایی فعالیت‌های غیرعادی در شبکه‌های اجتماعی، به این دلیل که کمک می‌کند تا اطلاعات مهم و قابل توجهی در مورد رفتار کاربران غیرعادی به دست آورده و آن‌ها را شناسایی کنی؛ مورد نیاز است. به منظور تشخیص ناهنجاری‌ها در شبکه‌های اجتماعی، محققان عمدتاً به رویکردهای مبتنی بر رفتار و ساختار متکی هستند. ما با استفاده از معرفی و تجزیه و تحلیل معیارهای مهم گراف برای تشخیص فعالیت‌های غیرعادی، رویکرد مبتنی بر ساختار گراف را گسترش می‌دهیم. مقایسه و اثربخشی اقدامات بر اساس سنجش‌های آماری مانند دقت، بازخوانی و F-Score و همچنین بر اساس نمرات غیرعادی محاسبه و ارائه شده است. ارزیابی نظری و تجربی روی چند مجموعه داده بزرگ نشان می‌دهد که رابطه بین گره واسط و تعداد لبه‌ها برای تشخیص و رتبه بندی حداکثری تعداد ناهنجاری‌ها به درستی کمک می‌کند.

واژه‌های کلیدی: ناهنجاری، مرکزیت بینابینی، گره واسط، دسته‌ها، شبکه‌های اجتماعی اینترنتی، شبکه‌های ستاره ای

Comparative analysis of graphical structural metrics for identifying anomalies in online social networks

Mojtaba Aajami¹, Naser Asgari²

1- Faculty of Electrical and Computer Engineering, Islamic Azad University of Zanjan, Zanjan, Iran, Email: aajami@iauz.ac.ir
2- Faculty of Electrical and Computer Engineering, Islamic Azad University of Zanjan, Zanjan Iran, Email: asgari.naser@gmail.com

Abstract: Due to their widespread use and popularity, social networks are subject to fraudulent attacks and illegal activities and security issues. Therefore, identifying abnormal activities, especially in social networks, is important because it helps to get important and significant information about the behavior of abnormal users and identify them. In order to detect abnormalities in social networks, researchers mainly rely on behavioral and structure-based approaches. We extend the graph-structure-based approach by introducing and analyzing critical graph criteria for detecting abnormal activities. The comparison and effectiveness of the measures are based on statistical measurements such as accuracy, refreshment and F-Score and also based on unusual scores. Theoretical and empirical evaluation on a Several large datasets shows that the relationship between the node node and the number of edges to help determine the maximum number of abnormalities correctly helps.

Keywords: Anomaly, Betweenness Centrality, Brokerage nodes, Clique, Social networks, Star network

تاریخ ارسال مقاله: ۱۳۹۹/۱۰/۰۲

تاریخ پذیرش مقاله: ۱۳۹۹/۱۲/۰۵

*: نویسنده مسئول

۱. مقدمه

شبکه‌های اجتماعی بر خط به عنوان یک حوزه تحقیقاتی پرطرفدار، توجه پژوهشگران فراوانی را به خود جلب کرده است. در حال حاضر، تقریباً هر حوزه‌ای مثل سرگرمی، آموزش، تجارت، کسب و کار، ارتباطات و ... متأثر از این شبکه‌ها می‌باشد. مشهورترین این شبکه‌ها، فیسبوک، در حدود ۲,۲۷ میلیارد نفر کاربر به صورت ماهانه دارد [۱]. تمایل افراد به ارتباط با یکدیگر و اشتراک گذاری اطلاعات باعث رشد سریع و روزافزون این شبکه‌ها شده است.

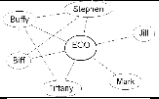
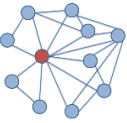
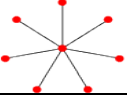
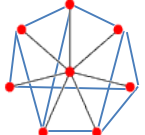
مطالعات بسیاری بر روی جنبه‌های مختلف این شبکه‌ها انجام شده است. یکی از جنبه‌های مهم تحلیل شبکه‌های اجتماعی بررسی حضور ناهنجاری‌ها است. ناهنجاری‌ها در حوزه شبکه‌های اجتماعی، دلالت بر رفتار نامنظم و اغلب غیرقانونی دارند [۲]. روش‌های بسیاری برای تشخیص انواع ناهنجاری‌ها در شبکه‌های اجتماعی پیشنهاد شده‌اند. بر اساس رویکرد مورد استفاده می‌توان این روش‌ها را در سه دسته مبتنی بر خوشه‌بندی / انجمن، مبتنی بر ساختار شبکه و مبتنی بر پردازش سیگنال دسته بندی کرد [۳]. در رویکردهای مبتنی بر خوشه بندی، هدف، پیدا کردن ناهنجاری‌ها بعد از خوشه بندی یا تشخیص انجمن‌ها می‌باشد. دومین دسته رویکردهای مبتنی بر ساختار شبکه است. از ساختار شبکه داده شده برای استخراج ویژگی‌های ساختاری گراف محور مانند، تعداد گره‌ها، تعداد یال‌ها، مرکزیت بینابینی هر گره و غیره استفاده می‌کنند.

انواع ناهنجاری‌هایی که توسط روش‌های تشخیص ناهنجاری در شبکه‌های اجتماعی مبتنی بر ساختار تشخیص داده می‌شوند، شامل رأس ناهنجار، یال ناهنجار و زیرگراف ناهنجار است [۳]. اگر هدف

شناسایی کاربرانی در شبکه باشد که رفتارشان از رفتار عادی مورد انتظار، متفاوت است؛ زیرمجموعه‌ای از رأس‌ها به عنوان ناهنجاری‌ها گزارش می‌شوند. اگر هدف، شناسایی فعل و انفعالات غیرقانونی میان کاربران باشد، زیر مجموعه‌ای از یال‌ها به عنوان ناهنجاری‌ها شناسایی می‌شوند و در نهایت در تشخیص زیر گراف ناهنجار، هدف پیدا کردن گروه‌هایی از افراد در شبکه است که الگوهای تعامل بین آن‌ها در مقایسه با رأس‌های دیگر در شبکه غیرعادی است که به این نوع ناهنجاری‌ها، ناهنجاری به لحاظ اتصال^۱ نیز گویند. شناسایی گره‌های هدف در گذشته مورد توجه بسیار کمی قرار گرفته است [۴] [۵]. در حالی که شناسایی عوامل دیگر در صدر مطالعات بسیاری از محققان [۶] [۷] بوده است.

برخی از معیارهای ساختاری گراف (تعداد گره، یال‌ها، مقادیر ویژه، مرکزیت بینابینی و...) برای شناسایی ناهنجاری‌ها در شبکه‌های اجتماعی استفاده شده است، اما در راستای تلاش برای بهبود دقت و کاهش تعداد مثبت و منفی‌های کاذب، دیگر معیارهای گراف (مانند واسط یا Brokerage) نیز می‌تواند برای تشخیص ناهنجاری‌ها در شبکه‌های اجتماعی واقعی بکار رود. با توجه به نوع ناهنجاری‌هایی که در این پژوهش مورد توجه قرار گرفته است، اجرای کار پیشنهادی بر روی گراف‌های ساده ایستا انجام شده است. بر این اساس، ویژگی‌های میانی مختلف گراف مبتنی بر ساختار همسایگان هر گره (Egonet) مورد مطالعه و بررسی قرار گرفته‌اند. برازش منحنی، تحلیل رگرسیون و تخصیص نمره غیر عادی به گره، به شناسایی و رتبه بندی ناهنجاری‌ها مطابق درجه ناهنجاری آن‌ها کمک می‌کند. برخی از اصطلاحات دیگر مرتبط با کار ما که اغلب در این زمینه استفاده می‌شود در جدول ۱ ارائه شده است.

جدول ۱: شرح مختصری از اصطلاحات مهم.

| واژه | تعریف | ملاحظات |
|---|---|---|
| Ego | یک Ego یک شخص در گراف است. |  |
| Egonet | گراف دوستان سطح یک یک گره (دوستان یک کاربر) یک ساختار Egonet را تعریف می‌کند. |  |
| Near Star | یک شبکه با مجموعه‌ای گره‌های تقریباً متصل، یک شبکه ستاره‌ای نزدیک را نمایش می‌دهد. |  |
| Near Clique | یک شبکه با مجموعه‌ای تقریباً تمام متصلی از گره‌ها، کلاس نزدیک را نمایش می‌دهد. |  |
| Betweenness Centrality (BC) مرکزیت بینابینی | مرکزیت بینابینی به عنوان تعدادی از کوتاه‌ترین مسیر بین هر جفت از گره‌ها در شبکه تعریف شده است که از طریق «ا» بوسیله تعداد کل کوتاهترین مسیر بین این رأس‌ها عبور می‌کند [۴]. | این در پیش بینی مرکزی ترین گره در شبکه و بنابراین جریان اطلاعات کمک می‌کند. |
| Average Betweenness Centrality (ABC) | ABC یک گره در داخل Egonet است که به عنوان میانگین مرکزیت بینابینی آن گره و مرکزیت بینابینی همه گره‌ها در آن Egonet تعریف شده است. [۵] | این نشان دهنده پیشرفت مرکزیت بینابینی یک گره با شناسایی با نفوذ ترین گره در |

۲- کارهای مرتبط با شناسایی زیرگراف‌های غیرمعمول

شناسایی گره‌های هدف در گذشته مورد توجه بسیار کمی قرار گرفته [۴] [۵] در حالی که شناسایی عوامل دیگر در صدر مطالعات بسیاری از محققان [۶] [۷] بوده است. به منظور شناسایی اجزای غیرعادی فوق، تکنیک‌های تشخیص غیرعادی مبتنی بر رفتار، مبتنی بر ساختار و مبتنی بر طیف طبقه بندی شده‌اند. روش‌های مبتنی بر رفتار [۹] [۱۰] با ویژگی‌های رفتاری افراد که شامل تعداد و محتوای پیام‌ها، تعداد لایک‌ها یا کامنت‌های یک پست (متن)، مدت زمان مکالمه، محتوای موارد به اشتراک گذاشته شده، سرو کار دارند. روش‌های مبتنی بر ساختار [۱۳] [۱۲] [۱۱] با ویژگی‌های ساختاری شبکه سر و کار دارند. چنین روش‌هایی ممکن است شامل استفاده از تعدادی از معیارهای گراف برای تعیین فعالیت‌های غیر عادی [۱۱] [۱۴] باشد. روش‌های مبتنی بر طیف [۱۵] [۱۶] با اندازه گیری‌های گرافیکی پیچیده مانند مقادیر ویژه و بردار ویژه مرکزیت‌ها^۲ یا با الگوریتم‌های ابرگراف پیچیده سرو کار دارند. همانطور که در [۱۳] توضیح داده شده است، روش‌های مبتنی بر ساختار، سودمندتر هستند، زیرا تقلید و ساختن ویژگی‌های ساختاری و فراداده‌هایی که در ساختار شبکه‌ها وجود دارد، برای کاربر مشکل است. در رویکردهای دیگر، یک کاربر می‌تواند برخی از روش‌های فرار، از قبیل پنهان کردن هویت و دادن اطلاعات نادرست در مورد نمایه، سن، شغل و سایر ویژگی‌های نمایه شان را به کار ببندد تا از شناسایی جلوگیری کرده و در نتیجه عموم را فریب دهد، اما مخفی کردن (ساختاری) فراداده همچون اتصالاتی که یک کاربر با گره‌های دیگر دارد، بسیار دشوار است.

ویژگی‌های ساختاری در یک شبکه مانند تعداد و نوع اتصالات بین گره‌ها، تراکم (چگالی) گراف، مواردی هستند که انکار آن‌ها برای کاربر دشوار است و از این رو کار کردن روی آن مهم است. به عنوان مثال، به منظور گسترش هرزنامه‌ها، هرزنامه‌نویسان معمولاً برای ایجاد پروفایل‌های جعلی یا سوء استفاده از حساب‌های قانونی انتخاب شده، هدف اصلی خود را بزرگ و متصل به هم قرار داده و به دنبال این، آن‌ها گسترش دایره ای خود را به‌طور تصادفی ادامه می‌دهند. تشخیص چنین افرادی با استفاده از ویژگی‌های رفتاری و نمایه ای کمک زیادی نخواهد کرد زیرا اگر پروفایل‌ها جعلی باشند، این اطلاعات رفتاری در ظاهر مشروع خواهد بود. علاوه بر این، تجزیه و تحلیل محتوای پست‌ها و پیام‌ها یک روش بسیار پیچیده و خسته کننده است. اما مشخصات ساختاری مانند چگالی گراف، مرکزیت گره به طور خودکار توسعه یافته و جعل کردن آن‌ها در دست کاربر نیست.

ویژگی‌های ساختاری گره‌ها توسط بسیاری از محققان برای تحلیل الگوهای مختلف در میان گره‌ها استفاده شده است. شناسایی روابط غیر

عادی با استفاده از کشف لینک غیر عادی (ALD) [۱۷]، تشخیص لینک‌های از دست رفته و خراب با استفاده از استخراج پیوند ساختاری [۱۸]، تشخیص کاربران و پیام‌های غیر عادی با استفاده از همکاری‌های خوشه بندی [۱۹]، تشخیص رهبران تفکرات خاص با استفاده از معیارهای اعتماد [۲۰]. محققان همچنین استفاده از معیارهای گراف را برای تشخیص گره‌ها و الگوهای غیرعادی به کار بسته‌اند. به عنوان مثال، Akoglu و همکاران [۱۱] الگوریتمی به نام Oddball را که تعدادی از قوانین قدرت را برای تشخیص سه نوع مختلف ناهنجاری‌ها، یعنی نزدیک ستاره‌ها / دسته‌ها، لبه‌های غالب و همسایگی سنگین پیشنهاد کرده است.

رضا و همکاران [۱۳] استفاده از برخی از معیارهای جدید مانند، ABC (میانگین مرکزیت بینابینی) و انجمن انسجام را برای پیش‌بینی ناهنجاری‌های نزدیک ستاره‌ای / دسته‌ای پیشنهاد کرده‌اند. به‌طور مشابه، رضایی و همکاران [۱۲] روش مشابهی را با پیروی از رابطه بین تعداد گره‌ها و لبه‌ها در مجموعه داده‌های توپیتر برای تشخیص چنین ناهنجاری‌هایی مورد استفاده قرار داده‌اند. هندرسون و همکاران [۱۴] بر روی ویژگی‌های مبتنی بر گره و مبتنی بر egonet کار کرده‌اند. آن‌ها ویژگی‌های موجود را با استفاده از محاسبه مقادیر جمع شده خاص طبق ویژگی‌های موجود به صورت بازگشتی تجزیه و تحلیل کرده‌اند. Trifunovic و همکاران [۲۱] نزدیکی اتصالات برای یافتن اعتماد^۳ اجتماعی صریح و ضمنی در شبکه‌های فرصت طلبانه^۴ را بکار برده‌اند. سینگ و همکاران [۲۲] از معیارهای گراف از قبیل درجه خروجی^۵ و درجه ورودی^۶ همراه با رتبه داده شده به ویژگی‌های مختلف به عنوان یک عامل قابل اعتماد بودن افراد در شبکه‌های اجتماعی مبتنی بر وب، استفاده کردند. اقدام و نویمپور [۲۰] از عواملی مانند روابط اجتماعی، محتوای اطلاعات متنی و روابط اعتماد برای یافتن سرگروه‌های یک تفکر در شبکه‌های اجتماعی استفاده کردند. بر اساس شباهت ساختاری و همچنین اجتماعی در میان گره‌ها، ضریب تشابه جاکرد محاسبه شد. سرگروه‌های نظرات بر اساس ارزش اعتماد محاسبه شده از نظرات کاربران انتخاب شدند. برخی از برنامه‌های کاربردی، با تخصیص یک عدد مبنی بر درجه پرت بودن هر شیء، بر میزان رفتار غیر عادی تمرکز کرده‌اند. این درجه به عنوان عامل پرت محلی (LOF^۷) توسط Breunig و همکاران اشاره شده است [۲۴]. در این پژوهش اثربخشی واسط برای تشخیص ناهنجاری ستاره ای نزدیک / دسته ای و مقایسه آن با دیگر معیارهای گراف مانند N نسبت به E و ABC نسبت به E را مورد بررسی قرار گرفته است.

۳- شناسایی ناهنجاری‌ها

کار ما ترکیبی از نظریه گرافیکی و آماری است. اولاً، معیارها و ساختارهای مختلف گراف محاسبه می‌شوند و بعد از آن برای شناسایی و تجزیه و تحلیل ساختارهای غیرمعمول (ستاره‌ها و دسته‌ها) از روش‌های آماری استفاده می‌شود.

۳-۱- انتخاب معیارهای گراف تشخیص ناهنجاری

تعدادی از معیارهای گراف در یک شبکه وجود دارد که می‌تواند مورد بررسی و تحلیل قرار گیرد. معیارهای گراف مورد استفاده در این کار شامل تعداد گره، لبه‌ها، میانگین مرکزیت بینابینی و واسط هستند، زیرا به نظر می‌رسد که تعریف آن‌ها مناسب ترین نوع تشخیص نوع ناهنجاری‌هایی است که در پژوهش مورد توجه قرار گرفت. برای انجام کلیه محاسبات مربوط به ساختار egonet و معیارهای مربوط به گراف از نرم افزار Matlab نسخه 2018a استفاده شد.

۳-۲- محاسبه منحنی برازش

به منظور درک ارتباط بین یک کاربر و دوستانش، رابطه بین معیارهای گراف مشخص شده با استفاده از قوانین توزیع مانند قانون خطی و قدرت مدل سازی می‌شود. برای محاسبه نمره غیر عادی برای هر کاربر، یک پیش نیاز برای تعریف یک مدل، که می‌تواند یک مدل آماری، مدل رگرسیون، یا سایر مدل‌ها باشد، وجود دارد. انتخاب مدل یک کار مهم است و در این کار بر استفاده از مدل رگرسیون با استفاده از منحنی برازش برای شناسایی ناهنجاری‌ها تمرکز می‌شود. همانطور که تحلیل رگرسیون یک روش آماری برای تحلیل رابطه بین متغیرهای مختلف است، از این رو برای کار پیشنهادی مناسب تر است. وابستگی متقابل بین گره‌های مختلف مورد مطالعه قرار گرفته است و تغییر رفتار یک متغیر به عنوان یک نتیجه از تغییر در دیگر، برای بررسی ارتباط بین معیارهای مختلف تجزیه و تحلیل می‌شود.

۳-۳- محاسبه امتیاز ناهنجاری

امتیاز ناهنجاری با استفاده از فرمول مورد استفاده توسط Akoglu و همکاران [۱۱] در الگوریتم Oddball و رضا و همکاران [۱۳] در کارشان، محاسبه شده است. فاصله گره از منحنی برازش به شناسایی رفتار غیرعادی کمک می‌کند.

$$ascore(i) = \quad (1-3)$$

$$\frac{\max(y_i, Cx_i^\theta)}{\min(y_i, Cx_i^\theta)} * \log(|y_i - Cx_i^\theta| + 1)$$

فرمول بالا، فاصله هر گره i را از منحنی برازش محاسبه می‌کند. روند مشترک و آشکار بیان می‌کند که گره‌های ناهنجار به دور از منحنی قرار داشته و فاصله آن‌ها از منحنی بیشتر خواهد بود در حالی که گره‌های طبیعی بر روی منحنی یا نزدیک به آن قرار دارند. در حین محاسبه فاصله به عنوان میزان امتیاز، فرمول هر گره را بوسیله مقدار انحراف بین مقدار محاسبه شده (y_i) و مقدار مورد انتظار (Cx_i^θ) همراه با لگاریتم

میزان انحراف جریمه می‌کند. آن y و x نشانگر معیارهای گراف هستند و θ مربوط به شاخص قدرت قانون است.

۳-۴- برچسب گذاری گره‌ها

گره‌های ناهنجار با استفاده از مجموعه ای خاص از محاسبات ریاضی و کدهای برنامه نویسی در MATLAB برچسب گذاری می‌شوند. فرمول‌های مورد استفاده برای برچسب گذاری ساختار ستاره و دسته در گره‌ها به شرح زیر است:

برای یک ساختار دسته، تعداد لبه‌های اطراف یک گره از الگوی $n(n-1)/2$ پیروی می‌کند که نشان دهنده حداکثر اتصالات است. به طور مشابه، در مورد ساختار ستاره یک گره تنها حدود $(n-1)$ گره در مجاورت خود (و یا egonet) دارد. همانطور که قبلاً گفته شد، درصد اتصالات میان گره‌ها، cliqueness یا starness در شبکه را تعیین می‌کند. بنابراین، با نگاه کردن به درصد $cliqueness / starness$ و با استفاده از اندازه گیری درجه در فرمول فوق، گره‌ها به عنوان داشتن یا نداشتن ساختار دسته/ ستاره ای برچسب گذاری می‌شوند. هدف اصلی این کار، بررسی اثربخشی معیارهای مختلف برای تشخیص ناهنجاری‌های تعریف شده است. با استفاده از چنین نمونه‌های برچسب گذاری شده، اثربخشی اقدامات مختلف مورد بررسی قرار می‌گیرد.

در پایان، مقدار آستانه روی امتیازهای ناهنجاری که برای هر معیار محاسبه می‌شود، تنظیم می‌شود تا اندر F-score را به حداکثر برساند. هدف نهایی این است که تعداد مثبت‌ها و منفی‌های کاذب که ممکن است با آن‌ها مواجه شود، به حداقل برسد. مقادیر آستانه با استفاده از حساسیت / اختصاصیت نسبت به نمره طرح، به دنبال هر مقدار ممکن که بتواند به همان اندازه عمل کند، تعیین می‌شود. ناحیه (ROC) منحنی (AUC) برای مقادیر مختلف اندازه گیری کمک می‌کند تا بهترین اندازه شناسایی شود.

در این پژوهش الگوهای محلی شبکه به جای سناریوی سراسری تحلیل می‌شود زیرا الگوهای محلی می‌توانند اطلاعات خوب و قابل اعتماد در مورد مفهوم سراسری را نیز ارائه دهند. البته هر دو مفهوم مرتبط هستند. ویژگی‌های محلی در استخراج الگوها در یک egonet کمک می‌کنند در حالی که الگوهای سراسری، قوانین استخراج شده را برای تعریف رفتار کاربران مختلف تعمیم می‌دهند. بنابراین، برای تشخیص ناهنجاری، اگر هر کاربر مشخص در فضای ویژگی‌های محلی خود (egonet) از الگوی کلی تعریف شده توسط فاکتورهای سراسری پیروی کرد آن را به عنوان عادی در نظر گرفته، در غیر این صورت آن بعنوان ناهنجاری در نظر گرفته می‌شود. علاوه بر این، ساختار پیوند که نشان دهنده ستاره‌ها یا دسته‌ها است، بهتر است از دید محلی تجزیه و تحلیل شود، زیرا تمرکز اصلی تجزیه و تحلیل ارتباطات، در egonet می‌باشد.

۴- نتایج ارزیابی‌ها

کار پیشنهادی در سه مجموعه داده جمع آوری شده از وب سایت <http://networkrepository.com/> استفاده شده است که شامل

دسته بندی از مجموعه داده‌های مربوط به زمینه‌های مختلف است که شبکه‌های اجتماعی یکی از آن‌ها است. در جدول ۲، جزئیات هر یک از مجموعه داده‌ها ارائه شده است. ما چهار مجموعه داده را پوشش دادیم که در هر یک از این شبکه‌ها، وجود ساختارهای نزدیک به حالت ستاره ای / دسته‌ای در egonet کاربر، رفتار ناهنجار را نشان می‌دهد.

جدول ۲: شرح مجموعه داده‌ها.

| SOCFB-MICHIGAN23 | SOCFB-A-ANON | SOC-FLICKR | SOC-DIGG | مجموعه داده |
|----------------------|----------------------|---------------------------|----------------|-----------------------|
| شبکه دوستی در فیسبوک | شبکه دوستی در فیسبوک | شبکه اجتماعی اشتراک تصویر | شبکه خبری | نوع شبکه |
| ۳۰,۱۴۷ | ۳,۰۹۷,۱۶۵ | ۵۱۳,۹۶۹ | ۷۷۰,۷۹۹ | تعداد گره‌ها |
| ۱,۱۷۶,۵۱۴ | ۲۳,۶۶۷,۳۹۲ | ۳,۱۹۰,۴۵۳ | ۵,۹۰۷,۱۳۳ | تعداد یال‌ها |
| رابطه اجتماعی/دوستی | رابطه اجتماعی/دوستی | روابط دوستی | روابط کاربران | نوع از تبادلات |
| شبه ستاره/دسته | شبه ستاره/دسته | شبه ستاره/دسته | شبه ستاره/دسته | نوع ناهنجاری یافت شده |

عادی روی منحنی یا در نزدیکی آن قرار می‌گیرند در حالی که گره‌های ناهنجار دورتر از منحنی قرار می‌گیرند. همچنین معیاری مثل مقدار آستانه از لحاظ فاصله (نمره ناهنجاری) به گونه‌ای تنظیم می‌شود تا حد مجاز (فاصله مجاز جهت عادی بودن) را تعیین کند. مقدار آستانه باید به مقداری تنظیم شود که تعداد مثبت و منفی کاذب به حداقل برسد تا دقت بالاتر را بدست بیاورد. F-score به عنوان یک واحد اندازه گیری برای تعیین دقت معیارهای مختلف استفاده می‌شود. محاسبه نمره غیرعادی با استفاده از فرمول بیان شده در معادله (۳-۱) به شناسایی ناهنجاری‌ها در مجموعه داده‌ها کمک می‌کند.

برای شناسایی ناهنجاری‌ها، برخی از معیارهای مهم گراف مانند تعداد گره‌ها، لبه‌ها، مرکزیت بینابینی و واسط ابتدا از لحاظ نظری مورد بررسی و سپس عملاً برای بررسی اثر بخشی آن‌ها مورد استفاده قرار گرفت. به منظور تشخیص ناهنجاری، فرمول مبتنی بر فاصله ذکر شده در معادله (۳-۱) استفاده شده است که فاصله یک گره را از منحنی برازش محاسبه می‌کند. به دلیل مقدار بالای واسط و تعداد کم لبه‌ها در ساختار egonet، چنین گره‌ای در جایی در بخش پایین سمت راست قرار خواهد گرفت. بدیهی است که این نوع از گره‌ها دورتر از منحنی قرار خواهند گرفت. پس از محاسبه منحنی‌های برازش، فاصله گره از منحنی برازش برای تعیین گره‌های ناهنجار بکار می‌رود. گره‌های

جدول ۳: مقایسه اثرگذاری معیارهای مختلف گراف.

| F-Score(%) | Recall (%) | دقت (%) | منحنی برازش | روش | مجموعه داده |
|------------|------------|---------|--------------------------------|----------------------------|--------------|
| ۵۲,۴۴ | ۶۷,۶۷ | ۴۲,۸۱ | $E = 0.03492B + 516.7$ | B نسبت به E (قانون خطی) | SOC-DIGG |
| ۳۲,۸۶ | ۸۴,۱۲ | ۲۰,۴۲ | $E = 29.41ABC - 634.2$ | ABC نسبت به E (قانون خطی) | |
| ۴۶,۲۲ | ۳۳,۶۰ | ۷۴,۰۰ | $E = 26.63N - 1294$ | N نسبت به E (قانون خطی) | |
| ۸۵,۳۹ | ۹۲,۶۳ | ۷۹,۲۱ | $E = 2.569 \times B^{0.6773}$ | B نسبت به E (قانون قدرت) | |
| ۸۳,۸۰ | ۸۰,۷۱ | ۸۷,۱۴ | $E = 25.1 \times ABC^{1.013}$ | ABC نسبت به E (قانون قدرت) | |
| ۸۲,۸۶ | ۸۸,۳۰ | ۷۸,۰۶ | $E = 1.38 \times N^{1.445}$ | N نسبت به E (قانون قدرت) | |
| ۴۸,۵۵ | ۴۷,۵۹ | ۵۱,۶۵ | $E = 0.03357B + 1555$ | B نسبت به E (قانون خطی) | SOC-FLICKR |
| ۳۳,۳۶ | ۲۷,۰۷ | ۴۳,۴۵ | $E = 79.94ABC - 1917$ | ABC نسبت به E (قانون خطی) | |
| ۲۱,۸۶ | ۱۰۰,۰۰ | ۱۲,۲۷ | $E = 71.9N - 3717$ | N نسبت به E (قانون خطی) | |
| ۸۳,۱۴ | ۷۹,۱۶ | ۸۷,۵۴ | $E = 13.45 \times B^{0.6179}$ | B نسبت به E (قانون قدرت) | |
| ۸۰,۵۵ | ۶۷,۴۳ | ۱۰۰,۰۰ | $E = 63.61 \times ABC^{1.022}$ | ABC نسبت به E (قانون قدرت) | |
| ۸۲,۱۹ | ۷۴,۳۹ | ۹۱,۸۰ | $E = 8.124 \times N^{1.302}$ | N نسبت به E (قانون قدرت) | |
| ۲۲,۵۹ | ۱۷,۵۶ | ۳۱,۶۷ | $E = 0.0061B + 320.5$ | B نسبت به E (قانون خطی) | SOCFB-A-ANON |
| ۱۹,۷۰ | ۱۰۰,۰۰ | ۱۰,۹۳ | $E = 8.349ABC - 144.4$ | ABC نسبت به E (قانون خطی) | |
| ۲۲,۱۹ | ۱۲,۴۸ | ۱۰۰,۰۰ | $E = 8.011N - 240.5$ | N نسبت به E (قانون خطی) | |
| ۷۷,۳۰ | ۸۲,۴۰ | ۷۲,۷۹ | $E = 2.343 \times B^{0.5943}$ | B نسبت به E (قانون قدرت) | |

| | | | | | |
|-------|-------|-------|--------------------------------|----------------------------|------------------|
| ۶۸,۸۴ | ۷۸,۹۲ | ۶۱,۰۴ | $E = 3.778 \times ABC^{1.121}$ | ABC نسبت به E (قانون قدرت) | SOCFB-MICHIGAN23 |
| ۷۲,۴۰ | ۸۵,۸۹ | ۶۲,۵۷ | $E = 1.998 \times N^{1.211}$ | N نسبت به E (قانون قدرت) | |
| ۶۱,۴۰ | ۹۷,۲۲ | ۴۴,۸۷ | $E = 0.0502B + 543.2$ | B نسبت به E (قانون خطی) | |
| ۶۰,۶۶ | ۴۷,۴۴ | ۸۴,۰۹ | $E = 98.44ABC + 1205$ | ABC نسبت به E (قانون خطی) | |
| ۶۰,۹۸ | ۶۴,۱۰ | ۵۸,۱۴ | $E = 21.41 N - 1043$ | N نسبت به E (قانون خطی) | |
| ۸۵,۳۱ | ۷۸,۲۱ | ۹۳,۸۵ | $E = 1.789 \times B^{0.7072}$ | B نسبت به E (قانون قدرت) | |
| ۷۸,۴۳ | ۷۶,۹۲ | ۸۰,۰۰ | $E = 1603 \times ABC^{0.309}$ | ABC نسبت به E (قانون قدرت) | |
| ۸۴,۴۴ | ۷۳,۰۸ | ۱۰۰ | $E = 0.9947 \times N^{1.505}$ | N نسبت به E (قانون قدرت) | |

از معیارها محاسبه می‌شود و یکی از موارد با بالاترین دقت و صحت به عنوان بهترین مقدار انتخاب می‌شود. مقدار آستانه نسبت به نوع مجموعه داده‌هایی که بر روی آن‌ها کار می‌کنیم، عدد متفاوتی خواهد بود. مقادیر آستانه برای هر یک از معیارهای مورد بحث به صورت جدول زیر بدست آمده است:

تنظیم مقدار آستانه با نسبت حساسیت/خصوصیت نسبت به نمره Score انجام می‌شود که این به انتخاب بهترین آستانه از تمام مقادیر ممکن که می‌توان در نظر گرفت، کمک می‌کند. همانطور که در مرحله ۹ شبهه کد گفته شد، برای هر مقدار ممکن آستانه، پارامترهای مختلف مانند دقت، جامعیت، صحت و اندازه گیری F-Score برای هر مجموعه

جدول ۳: محاسبه مقدار آستانه امتیازهای بدست آمده

| مقدار آستانه | F-Score(%) | روش | مجموعه داده |
|--------------|------------|----------------------------|------------------|
| ۰,۱۴ | ۵۲,۴۴ | B نسبت به E (قانون خطی) | SOC-DIGG |
| ۰,۸۸ | ۳۲,۸۶ | ABC نسبت به E (قانون خطی) | |
| ۰,۱۲ | ۴۶,۲۲ | N نسبت به E (قانون خطی) | |
| ۰,۱۰ | ۸۵,۳۹ | B نسبت به E (قانون قدرت) | |
| ۰,۳۶ | ۸۳,۸۰ | ABC نسبت به E (قانون قدرت) | |
| ۰,۰۸ | ۸۲,۸۶ | N نسبت به E (قانون قدرت) | |
| ۰,۵۲ | ۴۸,۵۵ | B نسبت به E (قانون خطی) | SOC-FLICKR |
| ۰,۶۲ | ۳۳,۳۶ | ABC نسبت به E (قانون خطی) | |
| ۰,۰ | ۲۱,۸۶ | N نسبت به E (قانون خطی) | |
| ۰,۳۰ | ۸۳,۱۴ | B نسبت به E (قانون قدرت) | |
| ۰,۵۴ | ۸۰,۵۵ | ABC نسبت به E (قانون قدرت) | |
| ۰,۲۸ | ۸۲,۱۹ | N نسبت به E (قانون قدرت) | |
| ۰,۴۶ | ۲۲,۵۹ | B نسبت به E (قانون خطی) | SOCFB-A-ANON |
| ۰,۸۲ | ۱۹,۷۰ | ABC نسبت به E (قانون خطی) | |
| ۰,۱۸ | ۲۲,۱۹ | N نسبت به E (قانون خطی) | |
| ۰,۱۸ | ۷۷,۳۰ | B نسبت به E (قانون قدرت) | |
| ۰,۰۸ | ۶۸,۸۴ | ABC نسبت به E (قانون قدرت) | |
| ۰,۱۶ | ۷۲,۴۰ | N نسبت به E (قانون قدرت) | |
| ۰,۶۲ | ۶۱,۴۰ | B نسبت به E (قانون خطی) | SOCFB-MICHIGAN23 |
| ۰,۴۴ | ۶۰,۶۶ | ABC نسبت به E (قانون خطی) | |
| ۰,۳۲ | ۶۰,۹۸ | N نسبت به E (قانون خطی) | |
| ۰,۴۶ | ۸۵,۳۱ | B نسبت به E (قانون قدرت) | |
| ۰,۴۸ | ۷۸,۴۳ | ABC نسبت به E (قانون قدرت) | |
| ۰,۵۰ | ۸۴,۴۴ | N نسبت به E (قانون قدرت) | |

تحلیل آماری و بصری نشان می‌دهد که منطقه تحت پوشش منحنی برای رابطه B نسبت به E حداکثر است. مقدار AUC برای باقی مانده

منحنی ROC و سطح زیر منحنی (AUC) در شکل‌های صفحات بعدی برای کلیه مجموعه داده‌های این پژوهش نشان داده شده است. تجزیه و

هر یک از مجموعه داده‌های پژوهش برای هر یک از رابطه‌های B نسبت به E، ABC نسبت به E و N نسبت به E به شکل جدول زیر محاسبه شده است:

جدول ۴: محاسبه مقدار سطح زیر نمودار (AUC)

| مجموعه داده | قانون قدرت / خطی | B نسبت به E | ABC نسبت به E | N نسبت به E |
|------------------|------------------|-------------|---------------|-------------|
| SOC-DIGG | قانون قدرت | ۹۳,۵۷ | ۸۹,۰۱ | ۹۱,۳۵ |
| | قانون خطی | ۷۳,۶۳ | ۵۵,۰۸ | ۶۵,۴۷ |
| SOC-FLICKR | قانون قدرت | ۸۸,۷۹ | ۸۳,۷۱ | ۸۶,۷۳ |
| | قانون خطی | ۶۹,۸۹ | ۶۱,۰۷ | ۵۰,۰۰ |
| SOCFB-A-ANON | قانون قدرت | ۸۹,۳۱ | ۸۶,۳۷ | ۸۹,۷۲ |
| | قانون خطی | ۵۶,۴۵ | ۵۰,۰۱ | ۸۹,۷۹ |
| SOCFB-MICHIGAN23 | قانون قدرت | ۸۸,۴۱ | ۵۱,۲۲ | ۷۷,۵۶ |
| | قانون خطی | ۶۰,۱۲ | ۵۳,۰۳ | ۸۳,۵۶ |

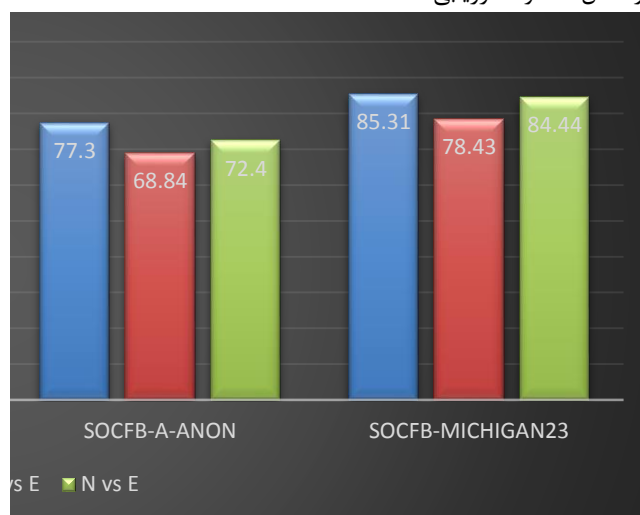
۵- نتیجه گیری

در این پژوهش، استفاده از معیارهای گراف در پیش بینی ناهنجاری‌ها و تاکید بر اهمیت روش‌های مبتنی بر ساختار ارائه شد. روش برآزش منحنی مبتنی بر ساختار گراف، برای تشخیص ناهنجاری‌ها با استفاده از ترکیب معیارهای جدید گراف و بیان اهداف نظری و عملی آنها، گسترش داده شد و مشاهده شد که رابطه میان واسط و لبه‌ها به پیش بینی تعداد زیادی از ناهنجاری‌ها که یا به اشتباه طبقه بندی شده بودند و یا با روش Oddball و رابطه ABC نسبت به E نادیده گرفته شده بود، کمک کرد. نمرات غیرعادی تخصیص داده شده به گره‌ها کمک می‌کند تا میزان ناهنجاری‌ها را پیش بینی و گره‌ها را با توجه به رفتار غیرمنطقی آنها رتبه بندی کند.

از این رو به نظر می‌رسد کارگزار (واسط) مقیاسی است که به درستی می‌تواند بسیاری از ناهنجاری‌ها را تشخیص دهد. با توجه به جدول ۳ می‌توان گفت که رابطه بین واسط (B) نسبت به تعداد لبه‌ها (E) به پیش بینی اکثر ناهنجاری‌ها کمک می‌کند. در جدول ۴، رابطه بین معیارهای مختلف گراف نشان داده شده است. برای هر یک از مجموعه داده‌ها، مقادیری برای پارامترهای مختلف مانند دقت، Recall و F-Score محاسبه شده و مشاهده می‌شود که استفاده از مقدار واسط در مدل عملکرد بهتری از مقیاس‌های موجود ارائه شده در روش "Oddball" [۱۱] و مرکزیت بینایی در [۲۱] دارد. نتایج از نظر آماری با استفاده از محاسبه نمره F-Score و همچنین بررسی گرافیکی در شکل ۱ نمرات ارزیابی شده‌اند.

مراجع

- [1] <https://zephoria.com/top-15-valuable-facebook-statistics/>
 [2] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," J. Soc. Netw, vol. 39, pp. 62-70, Octobr 2014.
 [3] P. Bindu and P. S. Thilagam, "Mining social networks for anomalies: Methods and challenges," J. Netw. Comput. Appl, vol. 68, pp. 213-229, June 2016.
 [9] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in Proce. Int. Conf. on Knowledge Discovery and Data Mining, pp. 410-421, June 2010.
 [4] Li H, Bhowmick SS, Sun A. Blog cascade affinity: analysis and prediction. In: Proceedings of the 18th ACM conference on information and knowledge management; 2009. p. 1117-26 .
 [۵] Li H, Bhowmick SS, Sun A, Cui J. Affinity-driven blog cascade analysis and prediction. Data Min Knowl Discov 2014;28(2):442-74 .
 [۶] Agarwal N, Liu H, Tang L, Yu PS. Identifying the influential bloggers in a community. In: Proceedings of the 2008 international conference on web search and data mining; 2008. p. 207-18 .



شکل ۱. مقایسه مقادیر F-Score اندازه گیری‌های B نسبت به E، ABC نسبت به E و N نسبت به E برای مجموعه داده‌های مختلف.

زیرنویس

1. Connectivity
2. Eigenvector Centralities
3. Trust
4. Opportunistic Networks
5. Out-Degree
6. In-Degree
7. Local Outlier Factor

- [۷] Cai K , Bao S , Yang Z , Tang J , Ma R , Zhang L , et al. OOLAM: an opinion oriented link analysis model for influence persona discovery. In: Proceedings of the fourth ACM international conference on web search and data mining; 2011. p. 645–54 .
- [۸] M. Newman, Networks: an introduction: Oxford university press, 2010.
- [۱۱] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in Proc. Int. Conf. on Knowledge Discovery and Data Mining, pp. 410-421, June 2010.
- [13] R. Hassanzadeh, R. Nayak, and D. Stebila, "Analyzing the effectiveness of graph metrics for anomaly detection in online social networks," in Int. Conf. on Web Inf. Syst. Eng., 2012, pp. 624-630.
- [۹] Vanetti M , Binaghi E , Carminati B , Carullo M , Ferrari E . Content-based filtering in on-line social networks. In: Privacy and security issues data mining and machine learning; 2011. p. 127–40 .
- [10] Viswanath B , Bashir MA , Crovella M , Guha S , Gummadi KP , Krishnamurthy B . Towards detecting anomalous user behavior in online social networks. In: Proceedings of the 23rd USENIX security symposium (USENIX security); 2014. p. 223–38 .
- [12] Rezaei A , Kasirun ZM , Rohani VA , Khodadadi T . Anomaly detection in online social networks using structure-based technique. In: 8th international conference for internet technology and secured transactions (ICITST); 2013. p. 619–22 .
- [14] Henderson K , Gallagher B , Li L , Akoglu L , Eliassi-Rad T , Tong H . It's who you know: graph mining using recursive structural features. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining; 2011. p. 663–71 .
- [15] Ying X , Wu X , BarbaráD . Spectrum based fraud detection in social networks. In: Proceedings - international conference on data engineering; 2011. p. 912–23 .
- [16] Miller B , Beard M , Wolfe P , Bliss N . A spectral framework for anomalous subgraph detection. IEEE Trans Signal Process 2014;63(16):4191–206 .
- [17] Rattigan MJ , Jensen D . The case for anomalous link discovery. ACM SIGKDD Explorations Newsletter 2004;7(2):41–7 .
- [18] Zheleva E , Getoor L , Golbeck J , Kuter U . Using friendship ties and family circles for link prediction. In: Advances in social network mining and analysis; 2010. p. 97–113 .
- [19] Yang W , Shen G-W , Wang W , Gong L-Y , Yu M , Dong G-Z . Anomaly detection in microblogging via co-clustering. J Comput Sci Technol 2015;30(5):1097–108 .
- [20] Aghdam SM , Navimipour NJ . Opinion leaders selection in the social networks based on trust relationships propagation. Karbala Int J Mod Sci 2016;2(2):88–97 .
- [21] Trifunovic S , Legendre F , Anastasiades C . Social trustin opportunistic networks. In: INFOCOM IEEE conference on computercommunications work- shops; 2010. p. 1–6 .
- [22] Singh S , Sidhu J . An approach for determining trustworthiness of individuals in a web-based social network. Arab J Sci Eng 2016;41(2):461–77 .
- [23] Breunig MM , Kriegel H-P , Ng RT , Sander J . LOF: identifying density-based local outliers. ACM Sigmod Rec 20 0 0;29(2):93–104 .
- [24] Eberle W , Holder L . Anomaly detection in data represented as graphs. Intell Data Anal 2007;11(6):663–89 .