فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Spring 2021, 2 (1), 1-9
DOR: 20.1001.1.27832570.1399.1.1.6.0

Received: 3 Jan 2021
Accepted: 20 Feb 2021

Comparative Analysis of Graphical Structural Metrics for Identifying Anomalies in Online Social Networks

Mojtaba Aajami[1*], Naser Asgari[2]

1.Assistant professor, Faculty of Electrical and Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran. *(Corresponding Author)* aajami@iauz.ac.ir
2. MSc Student, Faculty of Electrical and Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran. asgari.naser@gmail.com

Abstract

Introduction: Social networks are exposed to a variety of security problems due to their wide use and popularity. Therefore, identifying unusual activities in social networks is of paramount importance as it helps to obtain significant information about the behavior of unusual users and identify them. One of the important aspects of social network analysis is to check the presence of anomalies. Anomalies in the field of social networks imply irregular and often illegal behavior. A host of methods have been proposed to detect different kinds of anomalies in social networks. According to the employed approach, these methods can be classified into three categories, namely, clustering-based, based on network structure-based, and signal processing-based. In this paper, we extend the graph structure-based approach by introducing and analyzing important graph metrics to detect abnormal activities. Theoretical and experimental evaluation using several large data sets demonstrate that the relationship between the interface node and the number of edges helps to correctly detect and rank the maximum number of anomalies.

Method: The proposed method is a combination of graphical and statistical theory. First, various metrics and graph structures are calculated, and then statistical methods are used to identify and analyze unusual structures (stars and clusters).

Results: Statistical and visual analysis shows that the area covered by the curve is maximum for the interface (B) compared to the number of edges (E). The results show that the proxy is a scale that can correctly detect many abnormalities. It can also be said that the relationship between the (B) and the (E) helps to predict most anomalies.

Discussion: In this research, a structure-based method was presented by using graph criteria to predict abnormalities. The curve fitting method based on the graph structure was extended to detect anomalies using the combination of new graph criteria. It was observed that the relationship between the interface and the edges helped to predict a large number of anomalies that were either misclassified or missed by the Oddball method and the ABC relationship to E. The abnormality scores assigned to the nodes help predict the degree of anomalies and rank the nodes according to their irrational behavior.

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

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران

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

تاریخ دریافت: ۱۳۹۹/۱۰/۱۳
تاریخ پذیرش: ۱۳۹۹/۱۲/۰۲

سال دوم، بهار ۱۴۰۰
شمارهٔ اول ، صص: ۱ - ۹

مجتبی اعجمی*[۱]، ناصر عسگری[۲]

۱. استادیار، دانشکده مهندسی برق و کامپیوتر، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران. (نویسنده مسئول) aajami@iauz.ac.ir
۲. کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران. asgari.naser@gmail.com

چکیده: شبکه‌های اجتماعی به دلیل استفاده وسیع و محبوبیت میان کاربران، با مشکلات امنیتی بسیاری مواجهند. بنابراین، شناسایی فعالیت‌های غیرعادی در شبکه‌های اجتماعی، به این دلیل که کمک‌می‌کند تا اطلاعات مهم و قابل‌توجهی در مورد رفتار کاربران غیرعادی به‌دست‌آورده و آن‌ها را شناسایی‌کنیم؛ 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 گسترش‌می‌دهیم. مقایسه و اثربخشی اقدامات بر اساس سنجش‌های آماری مانند دقت، بازخوانی و F-Score و همچنین بر اساس نمرات غیرعادی محاسبه و ارائه‌شده‌است. ارزیابی نظری و تجربی روی چند مجموعه داده بزرگ نشان‌می‌دهد که رابطه بین گره واسط و تعداد لبه‌ها برای تشخیص و رتبه‌بندی حداکثری تعداد ناهنجاری‌ها به‌درستی کمک‌می‌کند.

واژه‌های کلیدی: ناهنجاری، مرکزیت بینابینی، گره واسط، دسته‌ها، شبکه‌های اجتماعی اینترنتی، شبکه‌های ستاره‌ای.

۱. مقدمه

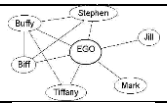
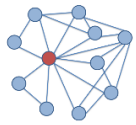
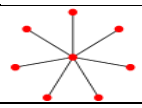
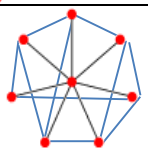
شبکه‌های اجتماعی بر خط به عنوان یک حوزه تحقیقاتی پرطرفدار، توجه پژوهشگران فراوانی را به خود جلب‌کرده‌است. در حال حاضر، تقریباً هر حوزه‌ای مثل سرگرمی، آموزش، تجارت، کسب و کار، ارتباطات و ... متأثر از این شبکه‌ها می‌باشد. مشهورترین این شبکه‌ها، فیسبوک، در حدود ۲,۲۷ میلیارد نفر کاربر به صورت ماهانه دارد[۱].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

مطالعات بسیاری بر روی جنبه‌های مختلف این شبکه‌ها انجام‌شده‌است. 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 پیشنهادشده‌اند. بر اساس رویکرد مورداستفاده می‌توان این روش‌ها را در سه دسته مبتنی بر خوشه‌بندی /انجمن ، مبتنی بر ساختار شبکه و مبتنی بر پردازش سیگنال دسته‌بندی‌کرد [۳]. در رویکردهای مبتنی بر خوشه‌بندی، هدف، پیداکردن ناهنجاری‌ها بعد از خوشه‌بندی یا تشخیص انجمن‌ها می‌باشد. دومین دسته رویکردهای مبتنی بر ساختار شبکه است. از ساختار شبکه داده‌شده برای استخراج ویژگی‌های ساختاری گراف‌محور مانند، تعداد گره‌ها، تعداد یال‌ها، مرکزیت بینابینی هر گره و غیره استفاده‌می‌کنند .

انواع ناهنجاری‌هایی که توسط روش‌های تشخیص ناهنجاری در شبکه‌های اجتماعی مبتنی بر ساختار تشخیص داده‌می‌شوند، 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؛ زیرمجموعه ای از رأس‌ها به عنوان ناهنجاری‌ها گزارش می‌شوند. اگر هدف، شناسایی فعل و انفعالات غیرقانونی میان کاربران باشد، زیر مجموعه‌ای از یال‌ها به عنوان ناهنجاری‌ها شناسایی‌می‌شوند و در نهایت در تشخیص زیر گراف ناهنجار، هدف پیداکردن 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 [1] نیز گویند. شناسایی گره‌های هدف در گذشته کمتر مورد بوده‌است [۴][۵]. در حالی که شناسایی عوامل دیگر در صدر مطالعات بسیاری از محققان [۶] [۷] بوده‌است.

برخی از معیارهای ساختاری گراف (تعداد گره، یال‌ها، مقادیر ویژه، مرکزیت بینابینی و...) برای شناسایی ناهنجاری‌ها در شبکه‌های اجتماعی استفاده‌شده‌است، 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 به‌کاررود. با توجه به نوع ناهنجاری‌هایی که در این پژوهش مورد توجه قرارگرفته‌است، اجرای کار پیشنهادی بر روی گراف‌های ساده ایستا انجام‌شده‌است. بر این اساس، ویژگی‌های میانی مختلف گراف مبتنی بر ساختار همسایگان هر گره(Egonet) مورد مطالعه و بررسی قرارگرفته‌اند. برازش منحنی، تحلیل رگرسیون و تخصیص نمره غیرعادی به گره، به شناسایی و رتبه‌بندی ناهنجاری‌ها مطابق درجه ناهنجاری آن‌ها کمک‌می‌کند. برخی از اصطلاحات دیگر مرتبط با کار ما که اغلب در این زمینه استفاده‌می‌شود در جدول ۱ ارائه شده‌است.

جدول ۱ : 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 تعریف‌می‌کند.	
Near Star	یک شبکه با مجموعه‌ای گره‌های تقریباً منفصل، یک شبکه ستاره‌ای نزدیک را نمایش‌می‌دهد.	
Near Clique	یک شبکه با مجموعه‌ای تقریبا تمام متصالی از گره‌ها، کلاس نزدیک را نمایش می‌دهد.	
Betweenness Centrality (BC) مرکزیت بِینابِینی	مرکزیت بِینابِینی به عنوان تعدادی از کوتاه ترین مسیر بین هر جفت از گره‌ها در شبکه تعریف شده‌است که از طریق «i» به‌وسیله تعداد کل کوتاهترین مسیر بین این رأس‌ها عبورمی‌کند [۴].	این در پیش‌بینی مرکزی‌ترین گره در شبکه و بنابرین جریان اطلاعات کمک‌می‌کند.
Average Betweenness Centrality (ABC)	ABC یک گره در داخل egonet است که به عنوان مبانگین مرکزیت بینابینی آن گره و مرکزیت بینابینی همه گره‌ها در آن egonet تعریف شده‌است. [۵]	این نشان‌دهنده پیشرفت مرکزیت بیتابینیِ یک گره با شناسایی با نفوذترین گره در

میانگین مرکزیت بینابینی	زمانی است که بسیاری از گره‌ها در کوتاه-ترین مسیر دخیل هستند.
Brokerage واسط(کارگزاری)	واسط(کارگزاری) حاکی از تعداد دفعاتی است که یک بازیگر به عنوان یک پل بین دو بازیگر دیگر که خودشان ارتباط مستقیمی با یکدیگر ندارند عمل‌می‌کند [۶].

۲. کارهای مرتبط با شناسایی زیرگراف‌های غیرمعمول

شناسایی گره‌های هدف در گذشته کمتر موردتوجه قرارگرفته [۴] [۵] در حالی که شناسایی عوامل دیگر در صدر مطالعات بسیاری از محققان [۶] [۷] بوده‌است. به منظور شناسایی اجزای غیرعادی فوق، تکنیک‌های تشخیص غیرعادی مبتنی بر رفتار، مبتنی بر ساختار و مبتنی بر طیف طبقه‌بندی‌شده‌اند. 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 گذاشته‌شده، سروکار دارند. روش‌های مبتنی بر ساختار [۱۳] [۱۲] [۱۱] با ویژگی‌های ساختاری شبکه سروکار دارند. چنین روش‌هایی ممکن است شامل استفاده از تعدادی از معیارهای گراف برای تعیین فعالیت‌های غیرعادی[۱۱] [۱۴] باشد. روش‌های مبتنی بر طیف [۱۵] [۱۶] با اندازه‌گیری‌های گرافیکی پیچیده مانند مقادیر ویژه و بردار ویژه مرکزیت‌ها[2] یا با الگوریتم‌های اَبَرگراف پیچیده سروکار دارند. همانطور که در [۱۳] توضیح داده‌شده‌است، روش‌های مبتنی بر ساختار سودمندترند، زیرا تقلید و ساختن ویژگی‌های ساختاری و فراداده‌هایی که در ساختار شبکه‌ها وجوددارد، برای کاربر مشکل است. در رویکردهای دیگر، یک کاربر می‌تواند برخی از روش‌های فرار، از قبیل پنهان‌کردن هویت و دادن اطلاعات نادرست در مورد نمایه، سن، شغل و سایر ویژگی‌های نمایه‌شان را به کار ببندد تا از شناسایی جلوگیری‌کرده و در نتیجه عموم را فریب‌دهد، اما مخفی‌کردن (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 کارکردن روی آن مهم است. به عنوان مثال، به منظور گسترش هرزنامه‌ها، هرزنامه‌نویسان معمولاً برای ایجاد پروفایل‌های جعلی یا سوء استفاده از حساب‌های قانونی انتخاب‌شده، هدف اصلی خود را بزرگ و متصل به هم قرارداده و به‌دنبال این، آن‌ها گسترش دایره ای خود را به‌طور تصادفی ادامه‌می‌دهند. تشخیص چنین افرادی با استفاده از ویژگی‌های رفتاری و نمایه‌ای کمک زیادی نخواهدکرد زیرا اگر پروفایل‌ها جعلی باشند، این اطلاعات رفتاری در ظاهر مشروع خواهدبود. علاوه بر این، تجزیه و تحلیل محتوای پست‌ها و پیام‌ها یک روش بسیار پیچیده و خسته‌کننده است. اما مشخصات ساختاری مانند چگالی گراف، مرکزیت گره به طور خودکار توسعه‌یافته و جعل آن‌ها در دست کاربر نیست.

ویژگی‌های ساختاری گره‌ها توسط بسیاری از محققان برای تحلیل الگوهای مختلف در میان گره‌ها استفاده‌شده‌است. شناسایی روابط غیر عادی با استفاده از کشف لینک غیرعادی(ALD) [۱۷]، تشخیص لینک‌های ازدست‌رفته و خراب با استفاده از استخراج پیوند ساختاری [۱۸]، تشخیص کاربران و پیام‌های غیرعادی با استفاده از همکاری‌های خوشه‌بندی [۱۹]، تشخیص رهبران تفکرات خاص با استفاده از معیارهای اعتماد [۲۰]. محققان همچنین استفاده از معیارهای گراف را برای تشخیص گره‌ها و الگوهای غیرعادی به‌کاربسته‌اند. به عنوان مثال، Akoglu و همکاران [۱۱] الگوریتمی به‌نام Oddball را که تعدادی از قوانین قدرت را برای تشخیص سه نوع مختلف ناهنجاری‌ها، یعنی نزدیک ستاره‌ها / دسته‌ها، لبه‌های غالب و همسایگی سنگین پیشنهاد کرده‌است.

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 پیشنهادکرده‌اند. رضایی و همکاران [۱۲] نیز روش مشابهی را با پیروی از رابطه بین تعداد گره‌ها و لبه‌ها در مجموعه داده‌های توییتر برای تشخیص چنین ناهنجاری‌هایی مورد استفاده قرارداده‌اند. هندرسون و همکاران [۱۴] بر روی ویژگی‌های مبتنی بر گره و مبتنی بر egonet کارکرده‌اند. آن‌ها ویژگی‌های موجود را با استفاده از محاسبه مقادیر جمع‌شده خاص طبق ویژگی‌های موجود به صورت بازگشتی تجزیه و تحلیل کرده‌اند. Trifunovic و همکاران [۲۱] نزدیکی اتصالات برای یافتن اعتماد[3] اجتماعی صریح و ضمنی در شبکه‌های فرصت‌طلبانه[4] را به‌کاربرده‌اند. سینگ و همکاران [۲۲] از معیارهای گراف از قبیل درجه خروجی[5] و درجه ورودی[6] همراه با رتبه داده‌شده به ویژگی‌های مختلف به عنوان یک عامل قابل‌اعتماد بودن افراد در شبکه‌های اجتماعی مبتنی بر وب، استفاده کردند. اقدم و نویمیپور [۲۰] از عواملی مانند روابط اجتماعی، محتوای اطلاعات متنی و روابط اعتماد برای یافتن سرگروه‌های یک تفکر در شبکه‌های اجتماعی استفاده‌کردند. بر اساس شباهت ساختاری و همچنین اجتماعی در میان گره‌ها، ضریب تشابه جاکارد محاسبه‌شد. سرگروه‌های نظرات بر اساس ارزش اعتماد محاسبه‌شده از نظرات کاربران انتخاب‌شدند. برخی از برنامه‌های کاربردی، با تخصیص یک عدد مبنی بر درجه پرت‌بودن هر شیء، بر میزان رفتار غیرعادی تمرکزکرده‌اند. این درجه به عنوان عامل پرت محلی(LOF[7]) توسط Breunig و همکاران اشاره‌شده‌است [۲۴]. در این پژوهش اثربخشی واسط برای تشخیص ناهنجاری ستاره‌ای نزدیک / دسته‌ای و مقایسه آن با دیگر معیارهای گراف مانند N نسبت به E و ABC نسبت به E را مورد بررسی قرارگرفته‌است.

۳. شناسایی ناهنجاری‌ها

کار حاضر ترکیبی از نظریه گرافیکی و آماری است. اولاً، معیارها و ساختارهای مختلف گراف محاسبه‌می‌شوند و بعد از آن برای شناسایی و تجزیه و تحلیل ساختارهای غیرمعمول (ستاره‌ها و دسته‌ها) از روش‌های آماری استفاده‌می‌شود.

۱،۳. انتخاب معیارهای گراف تشخیص ناهنجاری

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

۲،۳. محاسبه منحنی برازش

به منظور درک ارتباط بین یک کاربر و دوستانش، رابطه بین معیارهای گراف مشخص‌شده با استفاده از قوانین توزیع مانند قانون خطی و قدرت مدل‌سازی‌می‌شود. برای محاسبه نمره غیرعادی برای هر کاربر، یک پیش‌نیاز برای تعریف یک مدل، که می‌تواند یک مدل آماری، مدل رگرسیون، یا سایر مدل‌ها باشد، وجوددارد. انتخاب مدل یک کار مهم است و در این کار، بر استفاده از مدل رگرسیون با استفاده از منحنی برازش برای شناسایی ناهنجاری‌ها تمرکزمی‌شود. همانطور که تحلیل رگرسیون یک روش آماری برای تحلیل رابطه بین متغیرهای مختلف است، از این رو برای کار پیشنهادی مناسب‌تر 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، محاسبه‌شده‌است. فاصله گره از منحنی برازش به شناسایی رفتار غیرعادی کمک‌می‌کند.

$$ascore(i) = \frac{\max(y_i.Cx_i^{\theta})}{\min(y_i.Cx_i^{\theta})} * \log(|y_i - Cx_i^{\theta}| + 1) \quad (۳-۱)$$

فرمول بالا، فاصله هر گره i را از منحنی برازش محاسبه‌می‌کند. روند مشترک و آشکار بیان‌می‌کند که گره‌های ناهنجار به‌دور از منحنی قرار داشته و فاصله آن‌ها از منحنی بیشتر خواهد بود در حالی که گره‌های طبیعی بر روی منحنی یا نزدیک به آن قراردارند. در حین محاسبه فاصله به عنوان میزان امتیاز، فرمول هر گره را به‌وسیله مقدار انحراف بین مقدار محاسبه‌شده (y_i) و مقدار مورد انتظار (Cx_i^{θ}) همراه با لگاریتم میزان انحراف جریمه‌می‌کند. آن y و x نشانگر معیارهای گراف هستند و θ مربوط به شاخص قدرت قانون است.

۴،۳. برچسب گذاری گره‌ها

گره‌های ناهنجار با استفاده از مجموعه ای خاص از محاسبات ریاضی و کدهای برنامه‌نویسی در MATLAB برچسب‌گذاری‌می‌شوند. فرمول‌های مورد استفاده برای برچسب‌گذاری 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 پیروی‌می‌کند که نشان‌دهنده 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 گفته‌شد، درصد اتصالات میان گره‌ها، cliqueness یا starness در شبکه را تعیین می‌کند. بنابراین، با نگاه‌کردن به درصد starness / cliqueness و با استفاده از اندازه‌گیری درجه در فرمول فوق، گره‌ها به عنوان داشتن یا نداشتن ساختار دسته/ ستاره‌ای برچسب‌گذاری‌می‌شوند. هدف اصلی این کار، بررسی اثربخشی معیارهای مختلف برای تشخیص ناهنجاری‌های تعریف‌شده‌است. با استفاده از چنین نمونه‌های برچسب‌گذاری‌شده، اثربخشی اقدامات مختلف مورد بررسی قرارمی‌گیرد.

در پایان، مقدار آستانه روی امتیازهای ناهنجاری که برای هر معیار محاسبه می‌شود، تنظیم می‌شود تا اندره F-score را به حداکثر برساند. هدف نهایی این است که تعداد مثبت‌ها و منفی‌های کاذب که ممکن است با آن‌ها مواجه‌شود، به حداقل برسد. مقادیر آستانه با استفاده از حساسیت / اختصاصیت نسبت به نمره طرح، به‌دنبال هر مقدار ممکن که بتواند به همان اندازه عمل‌کند، تعیین‌می‌شود. ناحیه(ROC) منحنی(AUC) برای مقادیر مختلف اندازه‌گیری کمک‌می‌کند تا بهترین اندازه شناسایی‌شود.

در این پژوهش الگوهای محلی شبکه به‌جای سناریوی سراسری تحلیل‌می‌شود زیرا الگوهای محلی می‌توانند اطلاعات خوب و قابل اعتماد در مورد مفهوم سراسری را نیز ارائه‌دهند. البته هر دو مفهوم مرتبط هستند. ویژگی‌های محلی در استخراج الگوها در یک egonet کمک می‌کنند در حالی که الگوهای سراسری، قوانین استخراج‌شده را برای تعریف رفتار کاربران مختلف تعمیم‌می‌دهند. بنابراین، برای تشخیص ناهنجاری، اگر هر کاربر مشخص در فضای ویژگی‌های محلی خود(egonet) از الگوی کلی تعریف‌شده توسط فاکتورهای سراسری پیروی کرد آن را عادی و در غیر این صورت ناهنجاری در نظرگرفته می‌شود. علاوه بر این، ساختار پیوند که نشان‌دهنده 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 جمع‌آوری‌شده از وب سایت http://networkrepository.com/ استفاده‌شده‌است 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 ارائه‌شده‌است. ما چهار مجموعه داده را پوشش‌دادیم که در هریک از این شبکه‌ها، وجود ساختارهای نزدیک به حالت ستاره‌ای / دسته‌ای در egonet کاربر، رفتار ناهنجار را نشان‌می‌دهد.

جدول ۲ : شرح مجموعه داده‌ها.

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

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 اثربخشی آن‌ها مورد استفاده قرارگرفت. به منظور تشخیص ناهنجاری، فرمول مبتنی بر فاصله ذکر شده در معادله (۳-۱) استفاده شده‌است که فاصله یک گره را از منحنی برازش محاسبه‌می‌کند. به دلیل مقدار بالای واسط و تعداد کم لبه‌ها در ساختار egonet ، چنین گره‌ای در جایی در بخش پایین سمت راست قرارخواهدگرفت. بدیهی است که این نوع از گره‌ها دورتر از منحنی قرارخواهندگرفت. پس از محاسبه منحنی‌های برازش، فاصله گره از منحنی برازش برای تعیین گره‌های ناهنجار به‌کارمی‌رود. گره‌های عادی روی منحنی یا در نزدیکی آن قرارمی‌گیرند در حالی که گره‌های ناهنجار دورتر از منحنی قرارمی‌گیرند. همچنین معیاری مثل مقدار آستانه از لحاظ فاصله (نمره ناهنجاری) به گونه‌ای تنظیم‌می‌شود تا حد مجاز (فاصله مجاز جهت عادی‌بودن) را تعیین‌کند. مقدار آستانه باید به مقداری تنظیم‌شود که تعداد مثبت و منفی کاذب به حداقل برسد تا دقت بالاتر را به‌دست بیاورد. F-score به عنوان یک واحد اندازه‌گیری برای تعیین دقت معیارهای مختلف استفاده‌می‌شود. محاسبه نمره غیرعادی با استفاده از فرمول بیان‌شده در معادله(۳-۱) به شناسایی ناهنجاری‌ها در مجموعه داده‌ها کمک‌می‌کند.

جدول۳ : مقایسه اثرگذاری معیارهای مختلف گراف

مجموعه داده	روش	منحنی برازش	دقت (%)	Recall (%)	F-Score(%)
SOC-DIGG	B نسبت به E (قانون خطی)	$E = 0.03492B + 516.7$	۴۲,۸۱	۶۷,۶۷	۵۲,۴۴
	ABC نسبت به E (قانون خطی)	$E = 29.41ABC - 634.2$	۲۰,۴۲	۸۴,۱۲	۳۲,۸۶
	N نسبت به E (قانون خطی)	$E = 26.63N - 1294$	۷۴,۰۰	۳۳,۶۰	۴۶,۲۲
	B نسبت به E (قانون قدرت)	$E = 2.569 \times B^{0.6773}$	۷۹,۲۱	۹۲,۶۳	۸۵,۳۹
	ABC نسبت به E (قانون قدرت)	$E = 25.1 \times ABC^{1.013}$	۸۷,۱۴	۸۰,۷۱	۸۳,۸۰
	N نسبت به E (قانون قدرت)	$E = 1.38 \times N^{1.445}$	۷۸,۰۶	۸۸,۳۰	۸۲,۸۶
SOC-FLICKR	B نسبت به E (قانون خطی)	$E = 0.03357B + 1555$	۵۱,۶۵	۴۷,۵۹	۴۸,۵۵
	ABC نسبت به E (قانون خطی)	$E = 79.94ABC - 1917$	۴۳,۴۵	۲۷,۰۷	۳۳,۳۶
	N نسبت به E (قانون خطی)	$E = 71.9\ N - 3717$	۱۲,۲۷	۱۰۰,۰۰	۲۱,۸۶
	B نسبت به E (قانون قدرت)	$E = 13.45 \times B^{0.6179}$	۸۷,۵۴	۷۹,۱۶	۸۳,۱۴
	ABC نسبت به E (قانون قدرت)	$E = 63.61 \times ABC^{1.022}$	۱۰۰,۰۰	۶۷,۴۳	۸۰,۵۵
	N نسبت به E (قانون قدرت)	$E = 8.124 \times N^{1.302}$	۹۱,۸۰	۷۴,۳۹	۸۲,۱۹
SOCFB-A-ANON	B نسبت به E (قانون خطی)	$E = 0.0061B + 320.5$	۳۱,۶۷	۱۷,۵۶	۲۲,۵۹
	ABC نسبت به E (قانون خطی)	$E = 8.349ABC - 144.4$	۱۰,۹۳	۱۰۰,۰۰	۱۹,۷۰
	N نسبت به E (قانون خطی)	$E = 8.011\ N - 240.5$	۱۰۰,۰۰	۱۲,۴۸	۲۲,۱۹
	B نسبت به E (قانون قدرت)	$E\ 2.343 \times B^{0.5943}$	۷۲,۷۹	۸۲,۴۰	۷۷,۳۰
	ABC نسبت به E (قانون قدرت)	$E = 3.778 \times ABC^{1.121}$	۶۱,۰۴	۷۸,۹۲	۶۸,۸۴
	N نسبت به E (قانون قدرت)	$E = 1.998 \times N^{1.211}$	۶۲,۵۷	۸۵,۸۹	۷۲,۴۰
	B نسبت به E (قانون خطی)	$E = 0.0502B + 543.2$	۴۴,۸۷	۹۷,۲۲	۶۱,۴۰

SOCFB-MICHIGAN23	ABC نسبت به E (قانون خطی)	$E = 98.44ABC + 1205$	۸۴,۰۹	۴۷,۴۴	۶۰,۶۶
	N نسبت به E (قانون خطی)	$E = 21.41\ N - 1043$	۵۸,۱۴	۶۴,۱۰	۶۰,۹۸
	B نسبت به E (قانون قدرت)	$E = 1.789 \times B^{0.7072}$	۹۳,۸۵	۷۸,۲۱	۸۵,۳۱
	ABC نسبت به E (قانون قدرت)	$E = 1603 \times ABC^{0.309}$	۸۰,۰۰	۷۶,۹۲	۷۸,۴۳
	N نسبت به E (قانون قدرت)	$E = 0.9947 \times N^{1.505}$	۱۰۰	۷۳,۰۸	۸۴,۴۴

تنظیم مقدار آستانه با نسبت حساسیت/خصوصیت نسبت به نمره Score انجام‌می‌شود که این به انتخاب بهترین آستانه از تمام مقادیر ممکن که می‌توان درنظرگرفت، کمک‌می‌کند. همانطور که در مرحله ۹ شبه کد گفته‌شد، برای هر مقدار ممکن آستانه، پارامترهای مختلف مانند دقت، جامعیت، صحت و اندازه‌گیری F-Score برای هر مجموعه از معیارها محاسبه‌می‌شود و یکی از موارد با بالاترین دقت و صحت به عنوان بهترین مقدار انتخاب‌می‌شود.

مقدار آستانه نسبت به نوع مجموعه داده‌هایی که بر روی آن‌ها کار می‌کنیم، عدد متفاوتی خواهدبود. مقادیر آستانه برای هریک از معیارهای مورد بحث به صورت جدول زیر به‌دست آمده‌است:

جدول۳ : محاسبه مقدار آستانه امتیازهای به‌دست آمده

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

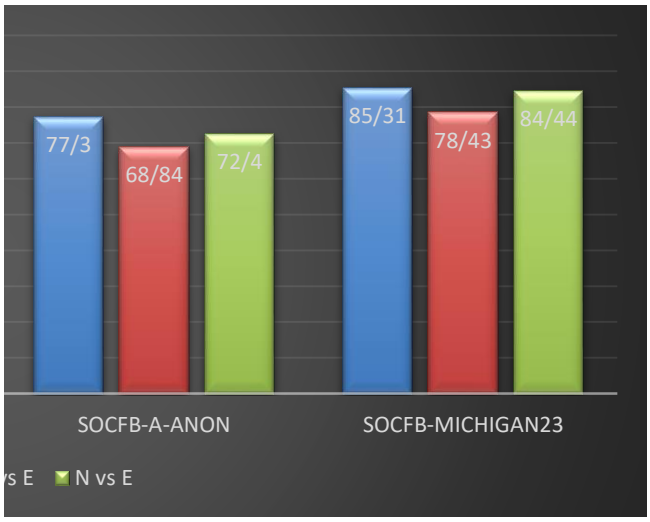
منحنی ROC و سطح زیر منحنی(AUC) در شکل‌های صفحات بعدی برای کلیه مجموعه داده‌های این پژوهش نشان داده شده‌است. تجزیه و تحلیل آماری و بصری نشان‌می‌دهد که منطقه تحت پوشش منحنی برای رابطه B نسبت به E حداکثر است. مقدار AUC برای باقی‌مانده هر یک از مجموعه داده‌های پژوهش برای هر یک از رابطه‌های B نسبت به E، ABC نسبت به E و N نسبت به E به شکل جدول زیر محاسبه شده‌است:

جدول۴ : 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

از این رو به نظر می‌رسد کارگزار(واسط) مقیاسی است که به‌درستی می‌تواند بسیاری از ناهنجاری‌ها را تشخیص‌دهد. با توجه به جدول۳ می‌توان‌گفت که رابطه بین واسط(B) نسبت به تعداد لبه‌ها(E) به پیش بینی اکثر ناهنجاری‌ها کمک‌می‌کند. در جدول ۴، رابطه بین معیارهای مختلف گراف نشان‌داده‌شده‌است. برای هریک از مجموعه داده‌ها، مقادیری برای پارامترهای مختلف مانند دقت، Recall و F-Score محاسبه‌شده و مشاهده‌می‌شود که استفاده از مقدار واسط در مدل عملکرد بهتری از مقیاس‌های موجود ارائه‌شده در روش "Oddball" [۱۱] و مرکزیت بینابینی در [۲۱] دارد. نتایج از نظر آماری با استفاده از محاسبه نمره F-Score و همچنین بررسی گرافیکی در شکل ۱ نمرات ارزیابی‌شده‌اند.



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

۵. نتیجه‌گیری

در این پژوهش، استفاده از معیارهای گراف در پیش‌بینی ناهنجاری‌ها و تأکید بر اهمیت روش‌های مبتنی بر ساختار ارائه‌شد. روش برازش منحنی مبتنی بر ساختار گراف، برای تشخیص ناهنجاری‌ها با استفاده از ترکیب معیارهای جدید گراف و بیان اهداف نظری و عملی آن‌ها، گسترش‌داده‌شد و مشاهده‌شد که رابطه میان واسط و لبه‌ها به پیش‌بینی تعداد زیادی از ناهنجاری‌ها که یا به اشتباه طبقه‌بندی شده‌بودند و یا با روش Oddball و رابطه ABC نسبت به E نادیده گرفته‌شده‌بود، کمک‌کرد. نمرات غیرعادی تخصیصی به گره‌ها کمک‌می‌کند تا میزان ناهنجاری‌ها را پیش‌بینی و گره‌ها را با توجه به رفتار غیرمنطقی آن‌ها رتبه‌بندی‌کنند.

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 .

[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 Proce. 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 .

پی‌نوشت

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