Journal of Electronical & Cyber Defence
Vol. 9, No. 2, 2021, Serial No. 34

Concurrent Detection of Condense Anomalous Subgraphs in Large Social Networks

M. Shahossein, A. Mahabadi*

*Shahed University

(Received: 12/11/2020, Accepted: 11/01/2021)

ABSTRACT

Classical anomaly detection approaches suffer from lack of detection accuracy, few concurrent detected anomalous subgraphs, high computational complexity, low detection speed, and the compliance to use a combination of local and global network graph information. This paper presents a new signal processing approach to simultaneous detection of small condense anomalous subgraphs based on local information of each unknown large-scale social graph. The approach introduces a novel sampling algorithm based on compressive sensing to retrieve the sparse properties of static networks, and aims to improve the speed and accuracy of detected anomalous subgraphs by reducing the complexity of data sampling. The experimental results with generated datasets of real and artificial graphs of large social networks show that the proposed approach in comparison with the state-of-the-art methods, increases the accuracy of simultaneous detection of anomalous condense subgraphs and reduces the computational complexity from to in the n node networks. Hence the proposed approach is easily applicable to anomaly detection of large complex dynamic networks.

Keywords: Anomaly Detection, Condense Anomalous Subgraphs, Signal Processing, Compressive Sensing, Graph Theory

* Corresponding Author Email: mahabadi@shahed.ac.ir

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۲، تابستان ۱۴۰۰، ص ۱۹۴-۱۷۹

علمی – پژوهشی

تشخیص هم‌زمان زیرگراف‌های متراکم ناهنجار در شبکه‌های اجتماعی بزرگ

ملیحه شاه‌حسینی[۱]، امین‌اله مه‌آبادی[۲*]

۱- کارشناسی ارشد گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه شاهد، تهران، ایران
۲- استادیار گروه مهندسی کامپیوتر و مرکز تحقیقات آکوستیک، دانشگاه شاهد، تهران، ایران
(دریافت: ۱۳۹۹/۰۸/۲۲، پذیرش: ۱۳۹۹/۱۰/۲۲)

چکیده

روش‌های سنتی تشخیص ناهنجاری بر اساس سیگنال در شبکه‌های بزرگ از چالش‌های دقت تشخیص ناهنجاری، تعداد کم زیرگراف‌های هم‌زمان ناهنجار تعیین‌شده، حجم بالا و پیچیدگی محاسباتی، سرعت کم تشخیص، الزام به استفاده ترکیبی از اطلاعات سراسری و محلی گراف شبکه رنج می‌برند. این مقاله رویکرد جدید پردازش سیگنال برای تشخیص هم‌زمان زیرگراف‌های متراکم ناهنجار بر اساس اطلاعات‌محلی از گراف‌های بزرگ ناشناخته اجتماعی پیشنهاد داده که هدفش بهبود سرعت و دقتِ تشخیص زیرگراف‌ها است. الگوریتم نوین نمونه‌برداری مبتنی‌بر نمونه‌برداری فشرده برای تقلیل نمونه‌برداری داده‌ها ارائه داده که از طریق بازیابی ویژگی‌های تنک شبکه‌های ثابت، پیچیدگیِ محاسباتی شبکه را کاهش داده است. نتایج آزمایش‌های تجربی با داده‌های تولیدی گراف‌های مصنوعی و واقعیِ شبکه‌های اجتماعی بزرگ در مقایسه با مهم‌ترین روش‌های علمی نشان داد که رویکرد پیشنهادی علاوه بر برخورداری از دقت تشخیص‌همزمان چندین زیرگراف متراکم، پیچیدگی محاسباتی را از $O(n^4\sqrt{\log n})$به $O(n^2)$ در شبکه n گره‌ای کاهش داده است. روش پیشنهادی به آسانی قابل کاربرد در تشخیص ناهنجاری‌های شبکه‌های بزرگ پویا و پیچیده است.

کلیدواژه‌ها: تشخیص ناهنجاری، زیرگراف‌های متراکم ناهنجار، پردازش سیگنال، نمونه‌برداری فشرده، نظریه گراف

۱- مقدمه

اینترنت به‌عنوان قلب زیرساخت‌های ارتباطی مدرن اساس شبکه‌های جامعه اطلاعاتی است زیرا به عملکرد مناسب بسیاری از سرویس‌های اصلی که باید ایمن، مقاوم و تا حد امکان نزدیک به ارتباطات اجتماعی انسان باشند کمک کرده است. رفتارها و ساختارهای نامتعارف در قالب ناهنجاری در این شبکه‌ها موجب ایجاد اختلال و نا امنی است که در کاهش کارآیی عملکرد، انسجام همبندی و روابط آن اثر منفی دارد. رشد شبکه‌های اجتماعی آنلاین به‌عنوان پدیده جدید، زندگی انسان را از طرق مختلف تحت تأثیر قرار داده و نیازمند توسعه روش‌های تشخیص رفتارهای غیرمتعارف هم‌زمان در استفاده روز افزون آن است [۱]. ناهنجاری در شبکه از طریق گره‌های خاص شبکه‌ها با تغییرات ناگهانی در همبندی شبکه به وجود می‌آید یا با الگوهای تعامل آنها ایجاد می‌کند به‌گونه‌ای که به‌طور قابل توجه با همتایان خود متفاوت باشند [۲]. تأثیرات این رفتار ناهنجار را می‌توان در نظارت بر ساختار شبکه مشاهده کرد که انحراف قابل‌توجه با رفتار طبیعی آن در اشکال مختلف اینترنت اشیا[۱]، اینترنت اشیاء چندرسانه‌ای[۲]، اینترنت اشیاء پزشکی[۳]، اینترنت اشیاء شهرهای هوشمند[۴] و موارد دیگر در قالب ناهنجاری دارد. رفتار و ساختار غیرطبیعی به‌عنوان پَرت، استثناء و ناهنجاری[۵] شناخته می‌شود [۳-۴].

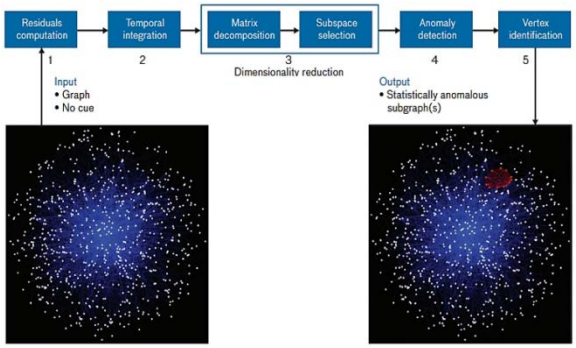
۱-۱- تشخیص ناهنجاری

تشخیص ناهنجاری برای شناسایی حملات رایانه‌ای [۵]، تهدیدهای اینترنت اشیاء [۳]، مخاطرات امنیتی[۶] [۶]، فعالیت‌های مخرب رایانه‌ای[۷] [۷]، تشخیص ازدحام جمعیت[۸] [۸] و به اشتراک‌گذاری اخبار جعلی[۹] در میان کاربران [۱] استفاده می‌شود

[1] Internet of Things (IoT)
[2] Internet of Multimedia Things (IoMT)
[3] Internet of Medical Things (IoMET)
[4] Internet of Smart city Things (IoSCT)
[5] Anomally
[6] Security Risks
[7] Computer Attacks
[8] Crowed
[9] Fake News

* رایانامه نویسنده مسئول: mahabadi@shahed.ac.ir

که مشابه پیش‌بینی لینک[1] یا پیشنهاد لینک‌ها در آینده [۹-۱۰] است.



شکل (۱): تشخیص زیرگراف ناهنجار بر اساس سیگنال[۲۰].

گرچه هدف از تشخیص ناهنجاری[2]، تعیین پیوندها[3] و گره‌های[4] ناهنجار است که به‌صورت زیرگراف‌های[5] هم‌زمان هم قابل ارائه باشند. این ناهنجاری‌ها با رفتار عادی موجودیت‌ها، ارتباطات آن‌ها یا همبندی شبکه متفاوت هستند. شبکه این موجودیت‌ها می‌تواند توسط گراف به‌صورت گره‌ها و ارتباطات آن‌ها مدل شود تا به‌عنوان بخش مهمی از ابزار تجزیه و تحلیل شبکه‌های مختلف مطرح شود [۸]. با ارائه چنین گرافی، یافتن ناهنجاری‌ها به‌صورت زیرگروه‌های ناهنجارِ هم‌زمان مورد توجه بسیاری از برنامه‌ها مانند کشف نفوذ، کشف تروریست، جنایات و جرائم سازمان‌یافته قرار دارد. این موارد شامل تحقیق در شناسایی زیرگراف‌های ناهنجار برای بسیاری از شبکه‌های بزرگ دارای روش‌های متعددی است [۱۱-۱۲]. روش‌های تشخیص ناهنجاری به دو رویکرد مبتنی‌بر پردازش سیگنال (تعیین ناهنجاری به‌صورت نویز در سیگنال) و پردازش گراف (تعیین ناهنجاری به‌صورت گره‌ها و ارتباط) تقسیم می‌شوند. روش‌های سنتی تشخیص ناهنجاری بر اساس سیگنال در شبکه‌های بزرگ از چالش‌های دقت تشخیص ناهنجاری، تعداد کم تعیین زیرگراف‌های هم‌زمان ناهنجار، حجم بالای پیچیدگی محاسباتی، سرعت کم تشخیص، الزام به استفاده از اطلاعات سراسری گراف یعنی جستجوی کل شبکه و اطلاعات محلی گراف رنج می‌برند.

توانایی شناسایی زیرگراف‌های کوچک متراکم در بسیاری از زمینه‌ها از شناسایی ترافیک مخرب در شبکه‌های رایانه‌ای تا کشف فعالیت‌های تهدیدآمیز با مخاطرات امنیتی در شبکه‌های اجتماعی بزرگ از چالش پیچیدگی پردازش رنج می‌برد. این حوزه تحقیقاتی نیازمند ارائه روش‌های جدید برای کاهش شدید حجم محاسبات در تشخیص ناهنجاری‌های شبکه‌های بزرگ در قالب زیرگراف‌های هم‌زمان است [۱۳]. تشخیص ناهنجاری در شبکه‌های پویا دارای دامنه وسیعی از برنامه‌های امنیتی مانند تشخیص نفوذ، حوادث رانندگی و حمله انکار خدمات برای کشف مناطق ناهنجار است [۱۴]. تشخیص دقیق زیرگراف‌های ناهنجار شبکه و زمان‌بندی آن در شبکه‌های بزرگ دنیای واقعی چالش بزرگی محسوب می‌شود که نیازمند احتساب زمان‌بندی و تأثیر آن در تشخیص دقیق و هم‌زمان زیرگراف‌های شبکه به‌عنوان مساله سخت مطرح است. این امر چالش‌های محاسباتی را به ماهیت ترکیباتی راه‌حل‌های احتمالی تحت قیود همبندی گراف می‌افزاید. ویژگی مهم دیگر روش تشخیص ناهنجاری، نوع ناهنجاری تشخیصی واحدهای ناهنجار یعنی گره‌ها، لینک‌ها، زیرگراف‌ها و یا سلسله رویدادهای شبکه است که تشخیص زیرگراف‌ها و تقلیل ابعاد گراف از سخت‌ترین نوع آن می‌باشد. این مساله در گام سوم از تشخیص ناهنجاری در روش سیگنال شکل (۱) مشهود است.

۱-۲- چالش‌های ناهنجاری

روش‌های جدید موجود تشخیص ناهنجاری مانند نمونه‌برداری فشرده[6] با SparcePCA با تمرکز بر سری‌های زمانی مرتبط با لبه‌های خاص شبکه یا بر روی ناهنجاری‌های سراسری، برای بیشینه‌سازی تشخیص دقیق بر روی کل شبکه با پیچیدگی[7] $O(n^4 \frac{\sqrt{\log n}}{\varepsilon})$ در شبکه n گره‌ای با پارامتر دقت ε تأثیر می‌گذارد [۱۳]. استفاده و گسترش این روش در شبکه‌های بسیار بزرگ مانند فیس‌بوک برای تشخیص دقیق و هم‌زمان زیرگراف‌های ناهنجار با چالش جدید کاهش پیچیدگی برای داده‌های عظیم روبرو است و نیازمند اندازه‌گیری غیرمستقیم یا نمونه‌برداری فشرده CS است. بهبود کارآیی تشخیص ناهنجاری مبتنی‌بر سیگنال، کاری مهم در تجزیه و تحلیل گراف‌های بسیار بزرگ برای افزایش ظرفیت گراف است که نمونه‌برداری مناسب می‌تواند با خوش‌بینی دقت تشخیص را افزایش دهد و بدبینانه وقت‌گیر باشد [۱۵]. با توجه به پیچیدگی و تنوع شبکه‌های بزرگ، چند نقطه بحرانی مقیاس‌پذیری، نمونه‌برداری، پیچیدگی، هم‌زمانی، میزان تراکم و دقت برای کاهش پیچیدگی و افزایش سرعت تصمیم‌گیری در تشخیص دقیق ناهنجاری زیرگراف‌های متراکم برای گراف‌های مقیاس بزرگ با روش پردازش سیگنال وجود دارد.

[1] Link Prediction
[2] Anomally Detection
[3] Link
[4] Subgraph
[5] Anomally Detection
[6] Compressive Sensing (CS)
[7] Complexity

- **مقیاس‌پذیری:** مدل‌سازی تشخیص ناهنجاری جهت توسعه و کاربرد از شبکه‌های کوچک تا شبکه‌های بسیار بزرگ پیچیده؛
- **پیچیدگی:** کاهش زمان اجرای پردازش تشخیص زیرگراف‌ها در مجموعه داده‌ها یا گراف در مقیاس بزرگ به‌عنوان مساله سخت؛
- **هم‌زمانی:** افزایش هم‌زمانی زیرگراف‌ها جهت سرعت بخشیدن به تشخیص، کاهش پیچیدگی زمانی و تقلیل حافظه مصرفی با پیاده‌سازی مناسب در کنار افزایش هسته‌های پردازنده؛
- **میزان تراکم:** برای تشخیص زیرگراف متراکم به معیارهایی برای میزان تراکم در واحد سطح، محدوده‌های مجاز تراکم، و میزان شفافیت گراف برای هر مساله تعیین و تعریف شود تا در اندازه‌گیری دقت تشخیص مؤثر باشد؛
- **نمونه‌برداری:** کاهش پیچیدگی محاسبات شبکه از طریق نمونه‌برداری کم در با حفظ خواص گراف شبکه‌ها در گراف‌های با مقیاس بزرگ؛
- **دقت:** افزایش دقت تشخیص زیرمجموعه‌های ناهنجار متراکم و عناصر مربوطه نیازمند دست‌یابی به ویژگی‌های کمتر برای کاهش ابعاد و اندازه مساله است.

در پردازش سیگنال برای تشخیص ناهنجاری گراف شبکه به ماتریس تبدیل و عملیات ماتریسی برای تعیین گره‌ها و لینک‌های ناهنجار و تعریف مجموعه آن‌ها به‌عنوان زیرگراف صورت می‌گیرد. با افزایش ابعاد شبکه محاسبات ماتریسی با مشکل پیچیدگی حجم روبرو می‌شود. نمونه‌برداری از شبکه از راه‌حل‌های تقلیل ابعاد ماتریس‌های شبکه است که موجب کاهش ویژگی‌های مهم شبکه است. به‌منظور کاهش هم‌زمان اندازه گراف و حفظ ویژگی‌های آن، ارائه بسیاری از رویکردهای پیچیده امکان‌پذیر نیست و پرهزینه است. گرچه از الگوریتم‌های تقریب چندجمله‌ای استفاده می‌شود ولی این روش‌ها نیز بسیار پرهزینه و پیچیده هستند. روش‌های موجود در بیشتر مطالعات فقط نتیجه پیش‌بینی را برای شبکه مورد استفاده به‌طور قابل‌توجهی بهبود می‌بخشند ولی به اندازه کافی قابل‌اعتماد برای حل مشکل در فضای بزرگ جستجو و پیش‌بینی دقیق در شبکه‌های با مقیاس بزرگ برخط جهت تصمیم‌گیری واقعی نیستند. این مشاهدات بیشتر طراحان را به روش‌های ترکیبی سوق می‌دهد که در جنبه‌های هم‌زمانی تشخیص، مقیاس‌پذیری روش، الگوریتم انتخابی، ویژگی‌های برگزیده، پیچیدگی متعارف و معیارهای اندازه‌گیری از سهم بیشتری برخوردار باشند. برخی از کارهای ارائه‌شده در فرآیند تشخیص ناهنجاری به‌عنوان محرک روش پیشنهادی ما هستند [۱۳].

۱-۳- نمونه‌برداری فشرده

نمونه‌برداری فشرده الگوی جدیدی در پردازش سیگنال و نظریه اطلاعات است که نمونه‌برداری و بازیابی سیگنال‌های مختصر را از تعداد کمی از اندازه‌گیری‌های غیرانطباقی یا مشاهدات ناقص به چالش می‌کشد [۱۳]. هدف اصلی آن نمونه‌برداری، فشرده‌سازی سیگنال‌های پراکنده با بعد کمتر (بردار پراکنده و ماتریس کم‌درجه) در فضای مناسب و بازیابی کامل گراف با تعداد کم اندازه‌گیری است [۱۴]. نیاز اصلی نمونه‌برداری فشردهِ موفق، قابلیت فشرده‌سازی یا پراکندگی سیگنال ورودی است. سیگنال پراکنده در مقایسه با ابعاد شبکه خودش از تعداد کم عنصر غیر صفر یعنی اجزای فعال برخوردار است. از این مفهوم در کاربردهای مختلف سیگنال زیست‌پزشکی و پردازش تصویر [۱۷,۱۶] تا گفتار [۱۸] و شبکه حسگر [۱۹] استفاده شده است. گرچه نقش آن در شبکه‌های اجتماعی هنوز کامل حس نشده [۲۰] و فصل جدیدی را در ادبیات الگوریتم‌های پایش و نظارت شبکه با فقط تعداد کمی اندازه‌گیری باز کرده است [۲۱]. نمونه‌برداری توجه به تقلیل ابعاد ماتریس‌های محاسباتی شبکه است.

تحقیقات موجود در زمینه اینترنت اشیاء اساساً بر روی قابلیت‌های سنجش، کنش و روش‌های شبکه متمرکز است [۲۲]. معمول است که فقط بخش کوچکی از لینک‌های شبکه با ازدحام یا نرخ از دست‌دادن بسته‌های بزرگ روبرو هستند و به نظر می‌رسد که نمونه‌برداری فشرده ابزار مناسبی برای استنباط این مشخصات پراکنده باشد. استفاده از نظریه نمونه‌برداری فشرده در تشخیص ناهنجاریِ شبکه شامل شروط اندازه بزرگ شبکه[1]، تنکی[2] داده‌های شبکه و قابلیت تجمیع‌پذیری داده‌ها است. گسترش شبکه‌های اجتماعی منجر به تشکیل شبکه‌های پیچیده بزرگ‌شده که دست‌یابی به دانش‌کل از شبکه مستلزم صرف زمان زیاد و هزینه بالا است که این موارد برای شبکه‌های برخط بزرگ قابل تصور و دست‌یابی نیست. این امر مخصوصا در گراف شبکه‌های پنهان یا ناشناخته نسبت به شبکه‌های شناخته‌شده برخوردار از دانش محلی که آگاهی از دانش شبکه وجود ندارد دارای هزینه‌های بسیار زیاد است.

بیشتر شبکه‌های دنیای واقعی از رأس‌ها و لینک‌های بسیاری تشکیل شده و ساختارهای پیچیده‌ای دارند. تعداد کاربران شبکه اجتماعی فیس‌بوک بیش از چند میلیارد نفر با روابط پیچیده است. نمونه‌برداری مستقیم از تک‌تک اعضاء شبکه سربار

[1] Large Networks
[2] Sparse

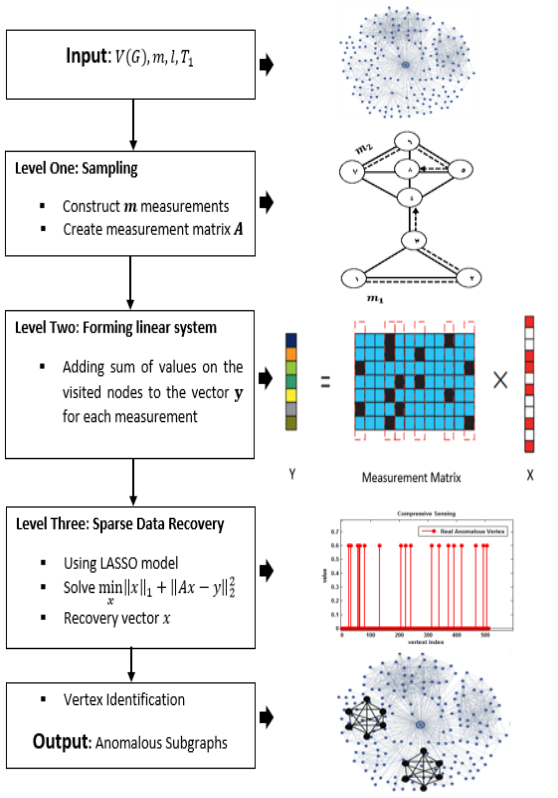
محاسباتی بالایی دارد. استفاده از نمونه‌برداری فشرده برای توموگرافی شبکه[1] به‌دلیل ارتباط نزدیک آن با گراف در نوع خود کاملاً متفاوت و جالب است. هدف از کارهای جدید تحقیقاتی، پر کردن شکاف بین نمونه‌برداری فشرده و تئوری گراف[2] است که در آن بردارهای سیگنال قابل بازیابی، بردارهای پراکنده[3] هستند که پارامترهای لینک‌های گراف را نشان می‌دهند. کارهای اخیر شروع به گسترش چارچوبی برای استنباط کارآمد ماتریس‌های درجه پایین[4] کرده‌اند [۲۴ و ۲۳]. ما نمونه‌برداری فشرده از گراف‌هایی را مطالعه می‌کنیم که مجاز به اندازه‌گیری کل گراف هستند و می‌خواهد کشف زیرگراف‌های زمانی متصل[5] متشکل از لبه‌ها را بررسی کند که به‌طور قابل‌توجهی از حالت عادی منحرف شده‌اند تا به سمت تشخیص سریع‌تر داده‌های واقعی سوق داده شوند در حالی که حل کارآمد نمونه‌های بزرگ آن حتی در زمان‌های طولانی نیز غیرقابل اجرا هستند.

تشخیص زیرگراف‌های ناهنجار از روش‌های مستقیم مانند sparsePCA[6] با فرض شناخته‌شدن گراف شبکه، محاسبات خود را روی ماتریس مدولاریتی[7] انجام می‌دهند. این روش‌ها کل گراف شبکه را جستجو می‌کنند و از پیچیدگی محاسباتی بالا رنج می‌برند. با توجه به تعداد کم رأس‌های در زیرگراف‌های ناهنجار نسبت به کل رأس‌های شبکه یعنی تنکی‌بردار داده، به‌راحتی می‌توانند از طریق نمونه‌برداری فشرده بازیابی شوند. قابلیت تجمیع‌پذیری داده‌ها شرط دیگری برای استفاده از این نظریه است. در تشخیص ناهنجاری به‌دلیل آن که نمونه‌برداری فشرده، اطلاعات را فشرده، تجمیع و از طریق رأس‌ها منتقل می‌کند این فشرده‌سازی سبب محفوظ ماندن حریم خصوصی کاربران مبنی بر ناهنجاری آن‌ها در شبکه است. به‌دلیل فراهم‌بودن این شروط، در این مقاله ایده استفاده از نظریه نمونه‌برداری فشرده برای تشخیص ناهنجاری در شبکه‌های اجتماعی در جهت تقلیل ابعاد ماتریس‌ها پیشنهاد شده است.

۱-۴- معماری روش پیشنهادی

آخرین روش‌های تشخیص ناهنجاری، متمرکز بر روش نمونه‌برداری فشرده است. مشاهده سریع و همزمان زیرگراف‌های ناهنجار دارای چالش مهم پیچیدگی محاسبات، سرعت و دقت تعیین عناصر و روابط ناهنجار است. هرچه قدرت تشخیص زیرگراف‌های ناهنجار بالاتر رود پیچیدگی محاسباتی نیز افزایش می‌یابد. الگوریتم‌های مختلفی برای حل این مساله ارائه شده است. بیشتر الگوریتم‌های پیشنهادی مقیاس‌پذیر نیستند و متکی بر کشف تعداد کم گره، لینک و زیرگراف با صرف زمان زیاد هستند که کاربردی در داده‌های عظیم گراف‌های بزرگ ندارند. در این مقاله تمرکز رویکرد پیشنهادی بر افزایش سرعت تشخیص و ردیابی همزمان دقیق زیرگراف‌های زیاد برای مقیاس‌پذیری روش جهت پشتیبانی از داده‌های عظیم در رشد گراف‌های بزرگ با فرض عدم اطلاع از وضعیت شبکه آن است (شکل۲).

چارچوب چند سطحی پیشنهادی دارای چهار سطح نمونه‌برداری، خطی‌سازی سیستم، بازیابی داده‌های‌تنک و شناسایی رأس‌ها است. در سطح اول، نمونه‌برداری با فرض ناشناخته‌بودن گراف شبکه صورت می‌گیرد و ماتریس اندازه‌گیری تولید می‌شود. در سطح دوم، مقادیر رأس‌های ملاقات‌شده در هر اندازه‌گیری با هم جمع و تشکیل بردار مشاهدات را می‌دهند تا سیستم خطی را تشکیل دهند. در سطح سوم، تنک‌ترین راه‌حل سیستم خطی از طریق حل مساله بهینه‌سازی بازیابی می‌شود. در نهایت در سطح چهارم رأس‌های مربوط به زیرگراف‌های ناهنجار به‌طور دقیق شناسایی می‌شوند. نتایج آزمایش‌های تجربی با داده‌های مصنوعی و واقعی نشان می‌دهد که رویکرد پیشنهادی دارای عملکرد مناسب است.



شکل (۲): معماری روش چندسطحی پیشنهادی تشخیص ناهنجاری.

[1] Network Tomography
[2] Graph Theory
[3] Sparse Vectors
[4] Low-rank Matrices
[5] Connected Temporal Subgraphs
[6] Sparse principal component analysis (sparse PCA)
[7] Modularity

۱-۵- محرک و نوآوری

محرک ما در این مقاله مستخرج از مدل تشخیص ناهنجاری در گراف‌های استاتیک است [۱۳]. هدف اصلی غلبه بر پیچیدگی محاسباتی تشخیص هم‌زمان زیرگراف‌هایِ کوچکِ متراکمِ ناهنجار[1] در شبکه‌های بزرگ است که منجر به کاهش محاسبات شناسایی، رتبه‌بندی و اعتبارسنجی شود. این امر در حالی است که گره‌ها و لینک‌های ناهنجار در یک گراف بزرگ پیچیده قرار دارند و مدل شناسایی باید صحیح و دقیق زیرگراف‌های ناهنجار شبکه‌های اجتماعی را استخراج کند [۱۱]. این مقاله به گنجاندن مدیریت ناهنجاری در الگوی محاسبات اجتماعی در مقیاس بزرگ کمک می‌کند تا بتوان ناهنجاری‌ها را در مکان‌ها و گره‌های مختلف تشخیص داد. این روشِ مدل‌سازی با رشد شبکه، مقیاس‌پذیر است و به‌عنوان نمایشی برای شناسایی هم‌زمان، مدل‌سازی و نظارت بر زیرگراف‌ها قابل ارائه است [۱۳]. نوآوری ما در رویکرد جدید پیشنهادی، ارائه روش چندسطحیِ جریانی و مقیاس‌پذیر، با قدرت تشخیص دقیق و هم‌زمان زیرگراف‌های متراکم ناهنجار از طریق الگوریتم پیشنهادی نمونه‌برداری بر اساس نمونه‌برداری فشرده در جهت بازیابی خصوصیات پراکنده گراف‌ها است تا بتواند کاهش قابل‌توجه پیچیدگی محاسباتی را از $O(n^4\sqrt{\log n})$ به $O(n^2)$ سوق دهد. در ضمن فرض ناشناخته‌بودن شبکه مورد بررسی بر قدرت و دامنه کاربرد روش پیشنهادی در گراف‌های بزرگ و شبکه‌های پیچیده می‌افزاید.

در ادامه بخش دوم به معرفی ناهنجاری، نظریه گراف و بررسی آخرین کارهای علمی و مهم تشخیص ناهنجاری در شبکه‌های اجتماعی و نظریه نمونه‌برداری فشرده می‌پردازد. بخش سوم روش‌های مبتنی بر پردازش‌سیگنال و چارچوب پیشنهادی را معرفی می‌کند. بخش چهارم مبین نتایج آزمایش‌های تجربی بر روی داده‌های مختلف است. در نهایت بخش پنجم به نتیجه‌گیری و بیان کارهای آینده می‌پردازد.

۲- کارهای مرتبط

کارهای انجام‌شده در تشخیص ناهنجاری گراف‌ها برای تعیین گره‌های ناهنجار، یال‌ها ناهنجار و زیرگراف‌های ناهنجار صورت گرفته است. که در این بخش به‌دلیل ماهیت رویکرد حل مساله این مقاله فقط تمرکز بر کارهای مرتبط با رویکرد زیرگراف‌ها است. الگوریتم‌های توسعه‌یافته تشخیص زیرگراف‌های ناهنجار متراکم در شبکهِ ایستایِ بدونِ مشخصه بر اساس رویکردهای مبتنی بر ساختار شبکه و پردازش سیگنال می‌باشد.

[1] Condense Small Anomolus Subgraphs

رویکرد مبتنی‌بر ساختار شبکه: در اسپم ایمیل[2] و بازاریابی ویروسی در شبکه‌های اجتماعی، کاربران کلاه‌بردار مجموعه‌ای از هویت‌های جعلی ایجاد کرده و از این هویت‌ها برای برقراری ارتباط با مجموعه بزرگ تصادفی از کاربران بی‌گناه استفاده می‌کنند. در مقاله [۲۳] نویسندگان برای مدل‌سازی چنین فعالیت‌های مخرب، حملات تصادفی لینک RLA[3] را تعریف کردند. الگوریتم‌هایی را برای استخراج زیرگراف‌ها با خصوصیت RLA پیشنهاد دادند که در دو مرحله تشخیص داده می‌شود. **مرحله اول**: شناسایی گره‌های مشکوک که احتمالاً بخشی از خوشه هکر[4] هستند از طریق انجام دو آزمون به نام‌های آزمون خوشه‌بندی و آزمون همسایگی مستقل[5] روی هر گره در شبکه انجام می‌شود. ازآنجاکه کاربران بی‌گناه به‌طور تصادفی انتخاب می‌شوند احتمال کمی دارد تا آن‌ها به هم متصل باشند لذا الگوی ستاره مانندی در شبکه تشکیل می‌دهند. به‌منظور شناسایی گره‌های مشکوک در شبکه، سه‌تایی‌ها[6] در ابزارهای جمع‌آوری اطلاعات[7] شمارش می‌شوند. تعداد سه‌تایی‌های کمتر، نشان‌دهنده هکر است.

مرحله دوم: مجموعه حمله، توسط رشد همسایگان گره‌های مشکوک شناسایی می‌شود.

در مقاله [۲۴] نویسندگان با استفاده از سیستم SUBDUE، دو روش برای شناسایی الگوهای غیرمعمول[8] در شبکه‌های ایستا معرفی کردند. روش اول، زیرساختارهای[9] خاص و غیرمعمول درون یک شبکه را شناسایی می‌کند. روش دوم، زیرگراف‌های ناهنجار توسط بخش‌بندی شبکه در زیرگراف‌های متمایز و جداگانه تشخیص می‌دهد. سپس هرکدام از آن‌ها را در برابر زیرگراف‌های دیگر برای کشف رویدادهای غیرمعمول مقایسه می‌کند. ایده اصلی این دو روش آن است که به‌طورکلی زیرگراف‌های مکرر از زیرگراف‌های کم تکرار، ناهنجاری کمتری دارند. سیستم SUBDUE زیرساختارهای مکرر را شناسایی و سپس شبکه را فشرده می‌کند. به‌دلیل این که زیرگراف‌های غیرعادی فقط شامل زیرساختارها با تکرار کم هستند متراکم‌سازی کمتری را در مقایسه با دیگر زیرگراف‌ها تجربه می‌کنند.

[2] Email Spam
[3] Random Link Attacks (RLA)
[4] Hacker
[5] Neighborhood Independence Test
[6] Triangles
[7] Egonet
[8] Unusual Patterns
[9] Substructures

همچنین روش دیگری در [۲۵] بر اساس سیستم SUBDUE پیشنهاد شد. سه الگوریتم تشخیص ناهنجاری مبتنی بر گراف GBAD[1]، برای کشف ناهنجاری‌ها در سه تغییر ممکن در شبکه شامل درج گره/ یال، دست‌کاری برچسب گره/یال و حذف گره/یال تطبیق داده شده با فعالیت‌های ناهنجار معرفی شد. هدف هر سه الگوریتم تشخیص یکی از این ناهنجاری‌ها و به‌کارگیری اصل MDL، برای شناسایی زیرساخت‌هایی است که گره‌ها و یال‌های ناهنجار را تشکیل می‌دهند. الگوریتم GBAD-P، احتمال درج گره/یال ناهنجار را تشخیص می‌دهد. الگوریتم GBAD-MDL، دست‌کاری برچسب گره/یال ناهنجار را تعیین می‌کند و الگوریتم GBAD-MPS حذف گره/یال ناهنجار را کشف می‌کند. ایده اصلی این الگوریتم‌ها یافتن ناهنجاری‌هایی است که زیرساختِ ناهنجارِ شبکه بخشی از الگوی هنجار[2] است. الگوریتم‌های مذکور، فرموله شده و می‌توانند برای هر کاربردی شامل داده‌های شبکه استفاده شوند. ولی به علت آن که همه اتفاقات نادر نمی‌توانند به‌عنوان ناهنجاری اعلام شوند نرخ مثبت کاذب[3] این الگوریتم‌ها بسیار بالا است.

رویکردهای مبتنی‌بر پردازش سیگنال: در مقاله [۲۶] چارچوبی مبتنی بر پردازش سیگنال پیشنهاد شد که از ویژگی‌های نُرمِ یکِ بردارهای ویژه ماتریس مدولاریتی گراف شبکه برای تعیین حضور زیرگراف ناهنجار استفاده می‌کند. مدولاریتی به‌عنوان معیار اندازه‌گیری کیفیت گروه‌بندی به‌عنوان گروه‌های چگال با ارتباطات تنک بین گروهی است که میزان تفکیک‌پذیری گراف به گروه‌ها را دربر دارد. بررسی می‌کند که آیا شبکه از فعالیت‌های معمولی تشکیل شده یا زیرگراف کوچک با همبندی ناهنجار درون شبکه جاسازی شده است. هدف مساله تشخیص زیرگراف ناهنجار، حل آزمون فرضیه دودویی زیر و ایجاد تمایز بین دو فرض صفر و یک است [۱۱]. فرض صفر یعنی گراف پس‌زمینه توسط فرآیند نویز تصادفی (تنها نویز) ایجاد شده و فرض یک یعنی‌گراف پس‌زمینه توسط همان فرآیند تصادفی ایجاد و با زیرگراف ناهنجار (سیگنال و نویز) ترکیب شده است. در پژوهشی دیگر [۲۷] چارچوب پیشنهادی برای مساله تشخیص زیرگراف تهدید جاسازی‌شده در شبکه‌های اجتماعی ارائه شد.

در [۱۳ و ۲۸] چارچوبی بر اساس فضای ویژه اصلی[4] از ماتریس مانده‌های[5] گراف شبکه پیشنهاد شد که در آن شبکه با مقدار امید[6] آن برای یافتن زیرگراف‌های ناهنجار مقایسه می‌شود.

[1] Graph-Based Anomaly Detection (GBAD)
[2] Normative Pattern
[3] False Positive
[4] Principal Eigenspace
[5] Residual Matrix
[6] Expected Value

مانده‌های بزرگ‌تر نشان‌دهنده وجود زیرگراف‌های ناهنجار هستند. در این پژوهش سه نوع الگوریتم مختلف به نام‌های مربع خی[7]، نُرم یک بردار ویژه[8] و تحلیل مؤلفه‌های اصلی تنک sparsePCA[9] برای یافتن زیرگراف‌های ناهنجار با باقی‌مانده‌های بزرگ‌تر پیشنهاد شد. شبکه‌های اجتماعی پویا تحت تغییرات ساختار و یا مشخصه‌هایشان هستند. وظیفه اصلی در شناسایی ناهنجاری‌ها، شناسایی نقاط تغییر یا رویدادها در زمانی است که اکثریت گره‌ها یا یال‌ها از رفتار عادی‌شان انحراف دارند و بخش‌های خاصی از شبکه که مسئول این نقاط تغییر هستند شناسایی شوند. با توجه به آن که شبکه‌های اجتماعی پویا از حوزه بررسی این پژوهش خارج است برای مطالعه انواع روش‌های تشخیص ناهنجاری در این نوع شبکه‌ها به منابع [۲۹] و [۳۰] مراجعه شود. در جدول (۱) ویژگی‌ها و چالش‌های روش‌های موجود و هدف تشخیص ناهنجاری‌های شبکه‌های ایستا بدون مشخصه را دارند ارایه شده است. قابل مشاهده است که رویکردهای مبتنی بر پردازش سیگنال از جدیدترین روش‌ها در حوزه تشخیص زیرگراف‌های ناهنجار هستند.

جدول (۱): روش‌های تشخیص ناهنجاری زیرگراف در شبکه‌های ایستا.

رویکرد	روش	ناهنجاری	چالش‌ها
ساختار شبکه	[۲۳]	زیرمجموعه‌ای از رأس	تطبیق رفتار دنیای واقعی با ویژگی‌های گراف محور. حدودیت ویژگی‌ها و انواع ناهنجاری‌های تشخیصی. همه اتفاقات نمی‌توانند ناهنجاری اعلام شوند.
	[۲۴]	امتیازهای ناهنجاری	
	[۲۵]	امتیازهای ناهنجاری	
پردازش سیگنال	[۱۱]	زیرمجموعه‌ای از رأس	افزایش دقت تشخیص ناهنجاری و افزایش پیچیدگی محاسباتی. تشخیص تنها یک زیرگراف ناهنجار.
	[۱۳]	زیرمجموعه‌ای از رأس	
	[۲۶]	زیرمجموعه‌ای از رأس	
	[۲۸]	زیرمجموعه‌ای از رأس	
نمونه-برداری فشرده	پیشنهادی	زیرمجموعه‌ای از رأس	وابستگی احتمال تشخیص درست به مقداردهی تنکی بردار رأس‌ها.

روش‌های نمونه‌برداری فشرده: حجم و اندازه بردار اطلاعات ذخیره‌شده در شبکه، داده شبکه‌ای[10]، با توجه به کاربرد با تعداد گره‌ها و تعداد یال‌های آن تعیین می‌شود. برای مثال در شبکه‌های ارتباطی[11]، داده شبکه یک بردار به‌اندازه تعداد لینک‌های موجود در شبکه است. درحالی‌که در شبکه‌های حسگر

[7] Chi-Squared
[8] Eigenvector L_1 Norm
[9] Sparse Principle Component Analysis (sparsePCA)
[10] Networked Data
[11] Communications Networks

بی‌سیم[1]، بردار شبکه هم‌اندازه تعداد گره‌های فعال در شبکه خواهد بود. در اولی هر مؤلفه داده شبکه بیان‌کننده تأخیر[2] لینک مربوطه و در دومی بیان‌کننده دما و یا هر داده‌ای است که توسط حسگر دریافت می‌شود [۳۱]. چالش عمده آنجایی است که در بسیاری از کاربردها امکان دسترسی مستقیم برای اندازه‌گیری داده‌های هر یک از مؤلفه‌های تشکیل‌دهنده شبکه امکان‌پذیر نیست. حتی درصورتی‌که امکان دسترسی مستقیم وجود داشته باشد تعداد بسیار بالای آن‌ها در بعضی از محیط‌ها، اندازه‌گیری مستقیم را توسط یک نقطه مرکزی عملاً غیرممکن می‌کند و به‌نوعی حمله منع سرویس[3] برای ایستگاه پایه به‌صورت ناخواسته رخ می‌دهد. این مشکل باعث پیشنهاد استفاده از نظریه نمونه‌برداری فشرده بر روی داده‌های شبکه‌ای شد [۳۱].

فرض می‌شود که هر شبکه دلخواه از n مؤلفه و هریک دارنده مقدار داده X_i به‌ازای $i = 1, \dots, n$ تشکیل شده است. این داده می‌تواند مربوط به ویژگی‌های رأس‌های شبکه یا آمارهای امنیتی باشد. مثلا دمای اندازه‌گیری‌شده توسط حسگرها یا نرخ اتلاف بسته[4] هر لینک در شبکه اینترنت باشد. همچنین فرض می‌شود بیشتر مؤلفه‌های شبکه مقدار داده ناچیزی دارند و تنها تعداد کمی از آن‌ها مقدار متفاوتی دارند. در این‌صورت شبکه حاوی بردار داده $x = [x_1, \dots, x_n]$ است که در اکثر موارد به‌جز برای تعداد کمی موقعیت ثابت است [۳۲] که در بسیاری از کاربردهای امنیتی مورد علاقه است. از آنجا که فرض بر kتنک بودن داده موجود در شبکه است پس نمونه‌برداری از تمامی مؤلفه‌ها برای دست‌یابی به آن بهینه نیست. ارائه راه‌کاری که با m نمونه‌برداری $m \ll n$ بتواند به‌صورت فشرده اطلاعات موردنیاز را از شبکه استخراج کند بسیار مطلوب است. مساله بازیابی اطلاعات شبکه درصورتی‌که بردار داده آن تنک باشد به‌راحتی با استفاده از نظریه نمونه‌برداری فشرده مدل می‌شود [۳۲].

در چند دهه اخیر نمونه‌برداری فشرده به خاطر توانایی‌اش برای ضبط و استخراج کارآمد اطلاعات تنک توجه زیادی را به خود جلب کرد. به‌کارگیری نمونه‌برداری فشرده در حوزه شبکه‌ها هنوز در مراحل اولیه است و چالش‌های بسیاری در اتصال این آن‌ها وجود دارد که مهم‌ترین این چالش‌ها، ساخت ماتریس اندازه‌گیری به‌دلیل برخی محدودیت‌های ساختاری شبکه‌ها است. ساخت ماتریس اندازه‌گیری باید مطابق دو محدودیت امکان‌پذیر باشد [۱۴]: (۱) گرچه در بیشتر مسائل موجود در CS هر زیرمجموعه‌ای از ورودی‌های بردار می‌توانند با یکدیگر جمع‌آوری و تشکیل ماتریس اندازه‌گیری را بدهند ولی این فرض در مورد مسائل نظارت شبکه که فقط گره‌های واقع‌شده روی یک مسیر یا زیرگراف همبند[5] در همان اندازه می‌توانند با یکدیگر جمع‌آوری شوند لزوماً درست نیست و اندازه‌گیری‌ها به‌واسطه محدودیت‌های همبندی شبکه محدود هستند. (۲) در شبکه‌ها، ماتریس اندازه‌گیری محدود به استفاده از فقط اعداد صحیح غیرمنفی است درحالی‌که معمولاً در CS از ماتریس‌های تصادفی گوسی[6] استفاده می‌شود.

درنتیجه نمونه‌برداری فشرده در شبکه‌ها در مقایسه با دیگر مسائل CS به‌طورکلی متفاوت است. از نظریه نمونه‌برداری فشرده برای بازیابی داده‌های تنک در شبکه‌ها استفاده شده است. برای اولین بار برای یافتن لینک‌هایی دارای ازدحام و نرخ خرابی بالا در شبکه‌های ارتباطی و از اندازه‌گیری‌های غیرمستقیم استفاده شد [۱۴]. لینک‌های با ازدحام بالا را یک ویژگی تنکی در شبکه‌های ارتباطی معرفی می‌کنند به‌طوری‌که فقط بخش کوچکی از لینک‌ها در شبکه ازدحام بالا دارند یعنی تأخیر این لینک‌ها بسیار بزرگ‌تر از تأخیر دیگر لینک‌ها است. از قدم‌زنی تصادفی[7] روی گراف شبکه برای تولید اندازه‌گیری‌ها استفاده شده است. در کارهای دیگر از نمونه‌برداری فشرده برای نظارت بر شبکه‌ها و بازیابی داده‌های تنک در شبکه‌های نظیر به نظیر[8] استفاده‌شد [۳۱ و ۳۳].

در شبکه‌های اجتماعی با مقیاس بزرگ روش‌هایی توسعه داده شد با اندازه‌گیری‌های بسیار کمتر از ابعادشان شبکه با دانش قبلی از ویژگی‌های رأس‌ها و یا یال‌ها بازیابی شوند. مثلا ویژگی تنکی در شبکه وجود داشته باشد. با این مفهوم و در کارهای دیگری برای اولین بار از نمونه‌برداری فشرده در شبکه‌های اجتماعی برای شناسایی مرکزی‌ترین رأس‌ها و یال‌های بین انجمن‌ها درروشی غیرمستقیم استفاده شد [۳۴-۳۵]. مرکزی‌ترین رأس‌ها و یال‌های بین انجمن‌ها به‌عنوان ویژگی‌های تنکی در شبکه‌های اجتماعی معرفی شد. برخلاف کار قبلی برای تولید اندازه‌گیری‌ها، آن که در هر اندازه‌گیری مقادیر کدام گره‌ها دخالت داده شوند و مقادیرشان باهم جمع شود اهمیت پیدا کردتا روی دقت داده بازیابی‌شده و شباهت آن به داده اصلی تأثیر بگذارد. دو الگوریتم به نام‌های CS-ComDet و CS-TopCent طراحی شد. جدول (۲) به‌صورت خلاصه کارهای مرتبط با نمونه‌برداری فشرده را به لحاظ نوع شبکه، روش نمونه‌برداری و نوع کاربرد با یکدیگر مقایسه می‌کند.

[1] Wireless Sensor Networks (WSN)
[2] Delay
[3] Denial Of Service Attack (DOSA)
[4] Packet Loss Rate (PLR)
[5] Connected
[6] Gaussian Random Matrix (GRM)
[7] Random Walk (RW)
[8] Peer To Peer Networks

جدول (۲): مقایسه کارهای مرتبط با نمونه‌برداری فشرده.

روش	نـوع شبکه	روش نمونه‌برداری	تشخـیص
[۱۴]	شبکه ارتباطی	RW	لینک‌های با ازدحام بالا
[۳۱] [۳۳]	نظیر به نظیر	RW	لینک‌های با ازدحام بالا
[۳۴]	شبکه‌اجتماعی	CS-TopCent	مرکـزی‌ترین k رأس
[۳۵]	شبکه اجتماعی	CS-ComDet	لینک‌های بین انجمن‌ها
پیشنهادی	شبکه اجتماعی	CS-AnomalyDet	ناهنجاری‌ها

۳- روش‌پیشنهادی

در روش پیشنهادی هدف کاهش پیچیدگی با نمونه‌برداری غیرمستقیم شبکه است. فرض می‌شود گراف شبکه ناشناخته و دانش نسبت به آن محلی است. این روش از چهار مرحله نمونه‌برداری، سیستم خطی‌سازی، بازیابی داده‌های تنک به نام بردار رأس‌ها و شناسایی رأس‌ها تشکیل شده است که در ادامه بیان می‌شود (شکل(۱)).

۳-۱- مدل‌سازی ناهنجاری

مساله یافتن زیرگراف‌های ناهنجار در شبکه با استفاده از چارچوب مبتنی بر نمونـه‌بـرداری فشـرده فرمـول‌بنـدی مـی‌شـود. شـبکه اجتماعی توسط گراف $G = (V, E)$، مجموعه V شـامل رأس‌هـای گراف مدل می‌شـود بـرخلاف V مجموعـه E شـامل لینـک‌هـای ناشناخته است. برای هر رأس $v \in V$ تنها همسـایگان سـطح اول $N^1(v)$، تعداد رأس‌ها $|V| = n$ و تعـداد لینـک‌هـا $|E|$ مشـخص اسـت. فـرض مـی‌شـود هـر رأس i مقـدار واقعـی x_i دارد و بردار $x = (x_i, i = 1,2, \dots, |V|)$ مـرتبط بـا مجموعـه V و $|V| = n$ است. نُرم‌پی از بردار x از رابطه (۱) محاسبه می‌شود [۹].

$$\|x\|_p = (\sum_{i=1}^{n} |x|^p)^{1/p} \qquad (۱)$$

بـرای $p = 0$ ، $\|x\|_0$ تعـداد مؤلفـه‌هـای غیرصـفر در x و بـرای $p = 1$، $\|x\|_1$ جمع مقادیر مطلق مؤلفه‌های x و بـرای $p = 2$، $\|x\|_2$ نُـرم اقلیدسـی و بـرای $p = \infty$ ، $\|x\|_\infty$ بیشـینه مقـادیر مطلق xاست. اگر x فقط k مؤلفه غیرصفر داشته باشد، x یـک بردار رأس k تُنک است. مثلا ناهنجاری‌ها در شبکه‌های اجتماعی ویژگی تنکی دارند به‌طوری‌که ناشـی از k رأس نـامزد هسـتند و تعداد این رأس‌های نامزد بسیار کمتر از تعداد کل رأس‌های شبکه است. اگر $x \in R^n$ بردار غیرمنفـی باشـد کـه p امـین ورودی آن مقدار روی رأس p است و $y \in R^m$ بردار مشاهدات، با m مؤلفـه به‌تعداد m اندازه‌گیری که qامین ورودی آن نشان‌دهنده مجمـوع مقادیر رأس‌ها در زیرگراف همبند از گراف شبکه باشد. همچنین A ماتریس اندازه‌گیری با ابعاد $m \times n$ اسـت کـه iامـین ردیـف مربوط به $i = 1, \dots, m$ امین اندازه‌گیری از شبکه است. بـرای $j = 1, \dots, n$ و $A_{ij} = 1$، j است اگر و فقط اگر i امین اندازه‌گیـری شامل رأس j باشد و در غیر این صورت صفر است. برای مثال در شـبکه دارای $|V| = 8$ و $|E| = 12$ بـا $m = 2$ انـدازه‌گیـری، ماتریس اندازه‌گیری A طبق رابطه (۲) تعریف شده است. سیستم خطی با رابطه (۳) بیان می‌گردد که در آن $m \ll n$ است.

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \qquad (۲)$$

$$y_{m\times 1} = A_{m\times n}\, x_{n\times 1} \qquad (۳)$$

$$\min_x \|x\|_0 \quad subject \quad to \quad y = Ax \qquad (۴)$$

$$\min_x \|x\|_1 \quad subject \quad to \quad y = Ax \qquad (۵)$$

$$\min_x \|x\|_1 + \|Ax - y\|_2^2 \qquad (۶)$$

سیستم فوق فرومعین است و بی‌نهایت جواب دارد. لازم است تا با اضافه کردن قیدی فضای راه‌حل محدود شود. محدودیت موردنظر در این مساله، قید تنکی بردار رأس x که تعداد رأس‌های زیرگراف‌های ناهنجار نسبت به تعداد کل رأس‌های گراف بسیار کمتر است می‌باشد و اغلب فرضی قابل قبول خواهد بود $(k \ll n)$.

در نظریه نمونه‌برداری فشرده بیان شده که تنک‌ترین راه‌حل می‌تواند از طریق حل مساله بهینه‌سازی رابطه (۴) به‌دست آید [۲ و ۳۶]. حل آن به‌دلیل وجود نُرم صفر مساله سخت است لذا با نُرم یک جایگزین و به رابطه (۵) تبدیل می‌شود. با ترکیب حداقل مربعات[1] این رابطه، تابع هدف به‌شکل رابطه (۶) تغییر می‌کند و به‌عنوان مدل LASSO شناخته می‌شود [۳۷] که مساله را حتی در حضور نویز یا مقادیر ناقص در ماتریس A و بردار y حل می‌کند. LASSO ازجمله بهینه‌سازی‌هایی است که برای دست‌یابی به مجموعه جواب‌های تنک مساله کارایی زیادی دارد و در حل مساله CS به‌کار می‌رود. با درنظرگرفتن m اندازه‌گیریِ شبکه و با وجود ویژگی تنک ناهنجاری‌ها در شبکه‌های اجتماعی، هدف از این اندازه‌گیری‌ها شناسایی رأس‌های زیرگراف‌های ناهنجار است لذا در ابتدا باید m اندازه‌گیری از گراف ورودی مساله تولید شود.

با فرض ناشناخته‌بودن گراف شبکه، دانش محلی و آن‌که از رأس‌ها و یال‌های بین آن‌ها آگاهی وجود ندارد باید از شبکه نمونه‌برداری شود و ماتریس اندازه‌گیری A تولید گردد. الگوریتم پیشنهادی ما CS-AnomalyDet برای تولید ماتریس اندازه‌گیری مبتنی‌بر الگوریتم CS-TopCent [۳۴] جهت شناسایی مرکزی‌ترین رأس‌ها است. در الگوریتم پیشنهادی برخلاف

[1] Least Squares

روش‌های نمونه‌برداری تصادفی، سعی‌شده تا در طی مراحل ساخت ماتریس اندازه‌گیری، اطلاعات برجسته در بردار داده x (مرکزی‌ترین رأس‌ها) حفظ شود و از این رأس‌ها با احتمال بالاتری گذر کند. یافتن رأس‌های زیرگراف‌های ناهنجار در این الگوریتم پیشنهادی تغییر یافته است. مراحل تولید ماتریس A در الگوریتم (۱) ارائه شده است. هر ردیف ماتریس اندازه‌گیری در ۶ مرحله ساخته می‌شود. **اول** برای هر رأس $v \in V$ به‌طور محلی وزن $W(v)$ محاسبه می‌شود (خطوط۹-۷). **دوم** برای شروع، یک رأس ابتدایی نسبت به احتمالات شروع $P(v)$ که قبلاً برای همه رأس‌های $v \in V$ در گراف G محاسبه شده (خطوط ۱۴-۱۱) انتخاب می‌شود. **سوم** ماتریس انتقال[1] $P_{trans}(v_{current}, u)$ بر اساس احتمال‌گذر از رأس v به رأس u ساخته می‌شود (خطوط ۱۹-۱۶). **چهارم** رأس بعدی از میان همسایگان رأس کنونی و بر اساس احتمالات $p_{trans}(v_{current}, u)$ انتخاب می‌شود (خط ۲۰).

الگوریتم (۱): الگوریتم CS-AnomalyDet جهت تولید ماتریس اندازه‌گیری.

Algorithm 1 Construction of measurement Matrix Method

Input : V(G), m, l, T_1
1: $V(G)$: Set of network vertex
2: m: Number of measurements
3: l : Number of measurements lengths
4: T_1 : Average degree of graph
5: A = NULL /*Initializing Measurements Matrix*/
6: P_{trans} = NULL /*Initializing Transition Matrix*/
7: **Foreach** $v \in V$ **do** /*Local Computation at each node*/
8: $\forall u \in N^1(v)$, $W(v) = \frac{2|\{e_{uw}: u,w \in N^1(v),\ e_{uw} \in E\}|}{|N^1(v)|\,(|N^1(v)|-1)} \times \frac{|nodes\ (u)\ with\ Degree < T_1|}{|N^1(v)|}$
9: **end for**
10: **for** $i = 1 \rightarrow m$ **do**
11: **Foreach** $v \in V$ **do** /* First Node Selection*/
12: $P(v) = \frac{1}{|V|-1}\left(\frac{W(v)}{\sum_{u \in V} W(u)}\right)$
13: **end for**
14: $v_{current}$ = Select first node relative to $P(v)$
15: **for** $j = 1 \rightarrow l$ **do**
16: **if** $\exists u \in N^1(v_{current})$ **then** /* Next Node Selection*/
17: **For each** $u \in N^1(v_{current})$ **do**
18: $P_{trans}(v_{current}, u) = \frac{W(u)}{\sum_{u \in N^1(v_{current})} W(u)}$
19: **end for**
20: v_{next} = Select next node relative to $P_{trans}(v_{current}, u)$
21: $N^1(v_{current}) = N^1(v_{current}) - \{v_{next}\}$
22: $N^1(v_{next}) = N^1(v_{next}) - \{v_{current}\}$
23: **else**
24: v_{next} = Back Track to the previous node
25: **end if**
26: $v_{current} = v_{next}$
27: **end for**
28: Add the visited nodes to the matrix A as a new row
29: **end for**
Output: measurements matrix of A

[1] Transition Matrix

۳-۲- نمونه‌برداری

از آنجا که هدف ما بازیابی رأس‌های زیرگراف‌های ناهنجار به‌عنوان ویژگی تنک در شبکه‌های اجتماعی است لذا الگوریتم CS-AnomalyDet طراحی شده تا برای ساخت ماتریس اندازه‌گیری از رأس‌های زیرگراف‌های ناهنجار با احتمال بالاتری گذر کند. به این منظور برای هر رأس در شبکه، وزنی در نظر گرفته می‌شود که مقدار این وزن نشان‌دهنده وضعیت[2] هر رأس $v \in V$ در زیرگراف حاصل از همسایگان سطح اول آن رأس است. این وزن توسط هر رأس به‌طور محلی طبق رابطه (۷) و بر اساس دو معیار ضریب خوشه‌بندی محلی[3] هر رأس و تعداد همسایگان هر رأس محاسبه و با درجه کمتر از آستانه T_1 گزارش می‌شود. با فرض این‌که هر رأس، همسایگان خود را می‌شناسد محاسبه وزن هر رأس می‌تواند توسط خود رأس به‌طور توزیع‌شده[4] و با استفاده از معیارهای محلی نظیر درجه خود رأس و همسایگانش انجام شود. ایده اصلی این معیار آن است که رأس‌های زیرگراف‌های کوچک متراکم، وزن بالایی را دریافت می‌کنند.

$$W(v) = \frac{2|\{e_{uw}: u,w \in N^1(v),\ e_{uw} \in E\}|}{|N^1(v)|\,(|N^1(v)|-1)} \times \frac{|nodes\ (u)\ with\ Degree < T_1|}{|N^1(v)|} \qquad (۷)$$

طی ساخت ماتریس اندازه‌گیری از این رأس‌ها با احتمال بالاتری گذر می‌شود. همچنین دیگر رأس‌های عادی وزن تقریباً ناچیزی را محاسبه می‌کنند. دلیل ارائه این وزن به شرح زیر است. اولین عبارت این رابطه برابر ضریب خوشه‌بندی محلی هر رأس است و این ضریب بررسی می‌کند که به چه مقدار همسایگان هر رأس به یکدیگر نزدیک هستند و تشکیل یک دسته گراف کامل را داده‌اند. نسبت یال‌های بین همسایگان یک رأس تقسیم‌بر تعداد یال‌هایی است که می‌تواند بین آن‌ها وجود داشته باشد. معمولاً اتصالات بین همسایگان رأس‌های زیرگراف‌های ناهنجار از یک زیرگراف عادی با همان اندازه در شبکه بسیار زیاد می‌باشد لذا مقدار این ضریب برای رأس‌های زیرگراف‌های ناهنجار بزرگ است. دومین عبارت آن تعداد همسایگان یک رأس با درجه کمتر از T_1 را می‌شمارد. طبق رابطه (۷) محاسبه ماتریس مانده‌های گراف $b_{ij} = a_{ij} - \frac{k_i k_j}{2|E|}$ صورت می‌گیرد. زمانی که مانده بین دو جفت رأس v_i و v_j مقدار بالایی می‌شود که ضرب درجه این دو رأس k_i و k_j کوچک شود. زمانی مقدار مانده بین دو رأس بزرگ می‌شود که درجه هر دو رأس یا یکی از این دو رأس کوچک باشد. ولی ازآنجاکه هدف یافتن زیرگراف‌های کوچک ناهنجار معمولاً

[2] Status
[3] Local Clustering Coefficiant
[4] Distributed

همسایگان یک رأس در یک زیرگراف ناهنجار درجه کمی دارند. لذا مقدار عددی عبارت دوم نیز برای رأس‌های ناهنجار دارای مقدار بالایی است. بهاین دو دلیل، رأس‌های یک زیرگراف ناهنجار وزنشان بالا است.

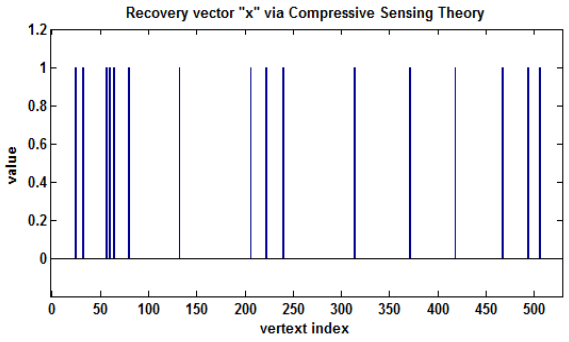
دلیل استفاده از عبارت دوم رابطه (۷) تأثیر خوشه‌بندی بالا در کاهش وزن رأس است. از آنجا که ممکن است رأس‌هایی در گراف باوجود ضریب خوشه‌بندی بالا دارای چندین همسایه با درجه بالا باشند لذا عبارت دوم از مقدار کمی محاسبه برخوردار می‌شود و خوشه‌بندی بالا به کاهش وزن رأس منجر می‌شود. در مرحله اول الگوریتم هر رأس وزن خود را به‌صورت محلی محاسبه می‌کند. به‌منظور بازیابی کارآمد رأس‌های زیرگراف ناهنجار در بردار داده‌ها، مطابق مراحل دوم و سوم یک رأس خوب برای شروع در هر m اندازه‌گیری نسبت به احتمالات شروع $P(v)$ انتخاب می‌شود. همچنین به هریک از همسایگان این رأس، احتمالات مناسب P_{trans} برای انتخاب بهترین رأس بعدی اختصاص می‌دهد. رأس بعدی l بار انتخاب می‌شود و l طول هر اندازه‌گیری است. به دلیل حذف یال بین رأس کنونی و رأس بعدی ممکن است که رأسی هیچ همسایه‌ای برای انتخاب به‌عنوان رأس بعدی نداشته باشد لذا به رأس قبلی ملاقات‌شده (خط ۲۴) برمی‌گردد. مجموعه رأس‌های ملاقات‌شده یک ردیف جدید در ماتریس اندازه‌گیری A را تولید می‌کنند.

۳-۳- سیستم خطی‌سازی

بعد از تولید ماتریس اندازه‌گیری A و افزودن وزن به رأس‌های بازدید شده به بردار مشاهدات y در هر اندازه‌گیری در زیرگراف همبند، سیستم خطی $y_{m\times1} = A_{m\times n}\, x_{n\times1}$ تشکیل می‌شود. در این سیستم بردار رأس‌ها x مجهول است و دستگاه مربوطه فرومعین و بی‌نهایت جواب دارد. لذا باید با اعمال محدودیت‌هایی، اندازه مجموعه جواب‌های ممکن را به یک کاهش داد. این قید همان شرط kتنکی بردار x است که در این مساله شرط قابل قبولی می‌باشد. یعنی تعداد رأس‌های زیرگراف‌های ناهنجار نسبت کمی از تعداد کل رأس‌های شبکه را تشکیل می‌دهند.

۴-۳- بازیابی بردار راس‌ها

تنک‌ترین راه‌حل خطی $y_{m\times1} = A_{m\times n}\, x_{n\times1}$ از طریق رابطه (۶) یافت می‌شود. شکل (۳) بردار بازیابی‌شده راس x با تنکی ۴درصد را در یک گراف ۵۱۲رأسی از طریق نمونه‌برداری فشرده با دو زیرگراف ناهنجار نشان می‌دهد. مؤلفه‌های غیرصفر این بردار مربوط به رأس‌های نامزد برای زیرگراف‌های ناهنجار هستند. در میان رأس‌های بازیابی‌شده، برخی از آن‌ها عضو زیرگرافی کوچک و متراکم هستند در حالی که همسایگانشان در بردار x بازیابی نشده‌اند در نتیجه زیرگراف‌های حاصل از این رأس‌ها ناهنجار نیستند. در مرحله بعد، رأس‌های زیرگراف‌های ناهنجار به‌طور دقیق شناسایی می‌شوند.



شکل (۳): بردار بازیابی‌شده رأس x با نمونه‌برداری فشرده در گراف ۵۱۲رأسی.

۵-۳- شناسایی راس‌ها

رأس‌های زیرگراف ناهنجار تنها رأس‌هایی هستند که به‌طور کامل بازیابی می‌شوند لذا زیرگراف‌های ناهنجار بالاترین تعداد رأس بازیابی شده را دارند. الگوریتم (۲) شناسایی زیرگراف‌های ناهنجار را نشان می‌دهد. در این الگوریتم پیشنهادی برای هر یک از مؤلفه‌های غیرصفر بردار x، بررسی می‌شود که هر مؤلفه با چند مؤلفه‌های بازیابی‌شده همسایه است (خطوط ۴-۱۵). برای جلوگیری از بررسی تکراری، بعد از شمارش مقدار هر یک از همسایه‌ها در بردار x صفر قرار داده می‌شود (خطوط ۹و ۱۴). درنهایت t رأس که دارای بالاترین مقدار شمارنده هستند به همراه همسایگانشان به‌عنوان زیرگراف‌های ناهنجار اعلام می‌شوند (خطوط ۱۶-۱۷).

الگوریتم (۲): الگوریتم پیشنهادی شناسایی زیرگراف‌های ناهنجار.

Algorithm2 Anomalous Subgraph Identification

Input: vector 'x'
1: Create $neighborsList$ HashMap <Integer, List<Integer>>
2: Create $counterList$ HashMap<Integer, Integer>
3: $Counter = 0$; /*Initializing counter*/
4: **Foreach** nonzero elements of x, x_i do
5: **if** $(x(i) == 0)$ continuous;
6: **Foreach** nonzero elements of x, x_j do
7: **if** (x_i contains neighbor (x_j))
8: $counter++$;
9: $x(j) = 0$;
10: $neighborsList.put(x_i, x_j)$;
/*set key to x_i & values to its neighbor*/
11: **end if**
12: **end for**
13: $counterList.put(x_i, counter)$;
/*set key to x_i & values to relevent counter*/
14: Reset $counter$ and $x(i) = 0$.
15: **end for**
16: Sort descending the $\boldsymbol{counterList}$ by value.
17: Get vertices and their neighbors' equivalent to t-top keys of the $\boldsymbol{neighborsList}$.
Output: Anomalous Subgraphs

۴- نتایج آزمایش‌های تجربی

نتایج آزمایش‌های تجربی در چهار بخش ارائه می‌شود. ابتدا تنظیمات آزمایش‌های، ابزار و مجموعه‌های داده‌ای ارائه می‌گردد. سپس به توضیح موارد مطالعه می‌پردازد و معیارهای ارزیابی آزمایش‌ها و الگوریتم معرفی و در نهایت نتایج تجربی بیان و تحلیل می‌شوند.

۴-۱- تنظیمات و مجموعه داده‌ها

از زبان برنامه‌نویسی جاوا و ابزار تحلیل گراف JGraphT [۵۴] برای نمونه‌برداری و پیاده‌سازی الگوریتم CS-AnomalyDet و از نرم‌افزار Matlab برای بخش بهینه‌سازی مساله به‌ویژه کتابخانه SPAMS[۵۵] و توابع بهینه‌ساز LASSO استفاده‌شده است. برای پیاده‌سازی روش‌های پردازش سیگنال نیز از زبان R و ابزار تحلیل گراف igraph [۵۶] و کتابخانه DSPCA [۵۷] در نرم‌افزار Matlab استفاده‌شده است. نتایج در ماشین Corei5، CPU 2.3GHz و حافظه اصلی RAM 4G بر روی سیستم‌عامل Win8 استخراج شده است.

از چالش‌های اساسی تشخیص ناهنجاری در شبکه‌های اجتماعی، عدم دسترسی به مجموعه داده‌های واقعی با حقیقت‌پایه[1] برای ارزیابی و مقایسه روش‌های پیشنهادی مختلف است. به این دلیل از روش تولید مجموعه داده‌های مصنوعی[2] و تزریق ناهنجاری به آن استفاده کرده‌ایم [۳۸]. مدل مولد گراف تصادفی اردوش-رینی[3] است که در آن احتمال رخداد همه یال‌ها در شبکه ثابت و برابر p در نظر گرفته می‌شود تا مدل مولد گراف‌تصادفی R-MAT تولید آن گراف را مشابه گراف‌های دنیای واقعی بسازد. همچنین ساختار انجمنی‌شبکه آن حفظ و قطر شبکه کوچک شود [۳۹]. برای ارزیابی از مولد R-MAT تحت لینوکس جهت تولید مجموعه داده‌های استفاده‌شده است [۴۰].

گراف‌های R-MAT با استفاده از ماتریس احتمالات پایه با مقادیر $a = 0.5$، $b = c = 0.125$ و $d = 0.25$ و ۱۲ تکرار تولید شده است. بعد از تولید گراف با استفاده از مولد R-MAT، زیرگراف ناهنجار درون گراف شبکه جاسازی شده است. مثلاً برای جاسازی یک زیرگراف ناهنجار شش رأسی درون گراف شبکه، ابتدا به‌صورت تصادفی شش رأس از میان رأس‌های گراف با درجه حداکثر شش کنترل‌کننده اتصالات خارجی زیرگراف انتخاب می‌شوند و تشکیل یک دسته می‌دهند. جاسازی زیرگراف ناهنجار، روی مجموعه یال‌ها تأثیر می‌گذارد. جدول (۳) مشخصات مرتبط با مجموعه داده‌های تولیدی توسط مولد R-MAT را بعد از تزریق ناهنجاری‌ها را نشان می‌دهد. همچنین در آن مجموعه داده‌های واقعی شبکه اجتماعی فیس‌بوک جهت یافتن زیرگراف‌های کوچک ناهنجار درج شده است [۶۰].

۴-۲- روش‌های مورد مقایسه

در این بخش به مقایسه روش‌های مبتنی‌بر پردازش سیگنال شامل دو روش Eigenvector L1 norm [۲۶] و sparsePCA [۲۸] با چارچوب پیشنهادی پرداخته می‌شود. در جدول کارایی این روش‌ها از نظر پیچیدگی محاسباتی و تعداد زیرگراف‌های ناهنجار تشخیصی مورد مقایسه قرار گرفته‌اند. بعد از پیاده‌سازی دو روش Eigenvector L1 norm و spaesePCA و استخراج نتایج آن‌ها، قدرت تشخیص این روش‌ها در مقایسه با چارچوب پیشنهادی با رسم منحنی ROC در گراف ۱۰۲۴ رأسی با تعبیه یک زیرگراف ناهنجار ۸رأسی مورد بررسی قرار گرفت. سطح زیر منحنی[4]AUC، به‌عنوان معیار مقایسه قدرت تشخیص هرکدام از روش‌ها اندازه‌گیری است. در چارچوب پیشنهادی تنکی‌بردار رأس‌ها، تعداد اندازه‌گیری‌ها و طول هر اندازه‌گیری به‌ترتیب با $\frac{k}{n} = 4\%$، $m = \frac{n}{5}$ و $l = \frac{n}{2}$ تنظیم شده است.

۴-۳- معیارهای ارزیابی

معیارهای اصلی ارزیابی را می‌توان به‌چهار دسته TP[5]، FP[6]، TN[7] و FN[8] تقسیم‌بندی نمود. TP تعداد رأس‌های ناهنجار که به‌درستی پیش‌بینی شده‌اند. FP تعداد رأس‌هایی که به‌اشتباه ناهنجار پیش‌بینی شده‌اند. TN تعداد رأس‌هایی که به‌درستی عادی پیش‌بینی شده‌اند. FN تعداد رأس‌های ناهنجار که به‌اشتباه عادی پیش‌بینی شده‌اند. معیارهای ارزیابی شامل نرخ فراخوانی R[9] یا نرخ تشخیص‌درست[10] رابطه(۸)، نرخ تشخیص‌خطا[11] رابطه (۹) و دقت[12] رابطه (۱۰) هستند. معمولاً برای ایجاد مصالحه[13] بین دو معیار R و P از شاخص F به‌عنوان ترکیبی از هر دو معیار مطابق رابطه (۱۱) استفاده می‌شود.

$$Recall, Probability\ of\ detection, TPR = \frac{TP}{TP+FN} \quad (۸)$$

[1] Ground Truth
[2] Synthetic Data Set
[3] Erdos-Renyi (ER)

[4] Area Under Curve (AUC)
[5] True Positive (TP)
[6] False positive (FP)
[7] True Negative (TN)
[8] False Negative (FN)
[9] Recall
[10] True Positive Rate (TPR)
[11] False Positive Rate (FPR)
[12] Precision
[13] Trade-off

$$Probability\ of\ false\ alarm, FPR = \frac{FP}{FP+TN} \quad (۹)$$

$$Precision = {TP}/{(TP + FP)} \quad (۱۰)$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (۱۱)$$

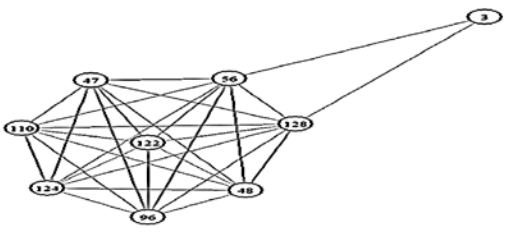
۴-۴- اعتبار سنجی و موارد مطالعه

چارچوب پیشنهادی ازنظر کاربرد و بررسی مورد خاص در گراف‌های تصادفی و شبکه‌های اجتماعی واقعی مورد مطالعه قرار می‌گیرد که نتایج حاصل از نظر ویژگی‌های مختلف زیرگراف‌های ناهنجار تشخیصی تحلیل می‌شوند. اعتبار سنجی روش پیشنهادی بر روی داده‌های دارای حقیقت‌پایه صورت گرفت و سپس قدرت تشخیص آن بر روی داده‌های گراف‌های تولیدی تصادفی تایید گردید.

گراف‌های تصادفی: چارچوب پیشنهادی روی گراف‌های تصادفی ۵۱۲ رأسی و ۱۰۲۴رأسی برای تشخیص زیرگراف‌های ناهنجار مورد بررسی قرار گرفته است. شکل (۴) زیرگراف ناهنجار ۶ رأسی با ۱۵یال، تعبیه‌شده در گراف ۵۱۲ رأسی را نشان می‌دهد. حداقل و حداکثر اتصال خارجی رأس‌های این زیرگراف به ترتیب سه (رأس ۵۰۶) و شش (رأس ۳۱۴ و ۴۱۸) است. نتایج نشان‌دهنده آن است که این زیرگراف کوچک روی رأس‌هایی تشکیل شده که اتصالات خارجی آن‌ها کم اما تعاملات بین خود رأس‌ها بالا است.

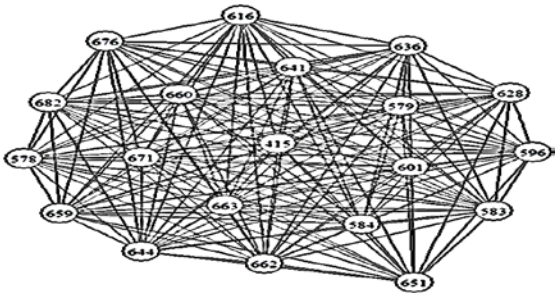
شکل (۴): زیرگراف ناهنجار ۶رأسی با ۱۵یال، تعبیه شده در گراف ۵۱۲رأسی.

شکل (۵) زیرگراف ناهنجار دیگری را با ۸رأس و ۲۸یال، تعبیه شده در گراف ۱۰۲۴رأسی را نشان می‌دهد. حداقل و حداکثر اتصال خارجی رأس‌های این زیرگراف به ترتیب یک (رأس ۵۶) و سه (رأس۱۲۲) است. لذا با وجود تعاملات بسیار کم رأس‌های این زیرگراف با دیگر رأس‌های شبکه، تراکم داخلی آن‌ها بالا است. پس این زیرگراف‌ها یک ناهنجاری ساختاری را در شبکه نشان می‌دهند.



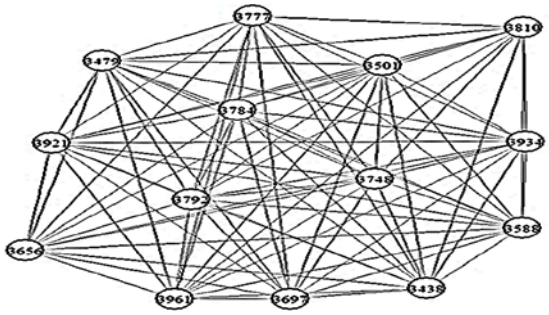
شکل (۵): زیرگراف ناهنجار ۸ رأسی با ۲۸ یال در گراف ۱۰۲۴رأسی.

شبکه‌های اجتماعی: برای بررسی کاربرد چارچوب پیشنهادی روی مجموعه داده‌های واقعی، داده‌های شبکه اجتماعی فیس‌بوک مورد استفاده قرار گرفته است. هدف یافتن زیرگراف‌های کوچک متراکم در این شبکه اجتماعی است که تراکم داخلی این زیرگراف‌ها بالاتر از حد مورد انتظار است و اتصال خارجی رأس‌های این زیرگراف پایین است. به‌طور تجربی تنکی بردار X به ۲۴درصد تنظیم شده است. شکل (۶) زیرگراف۲۰رأسی با ۱۸۸یال تشخیص داده شده توسط چارچوب پیشنهادی را نشان می‌دهد. درجه رأس ۵۷۸ کمترین مقدار یعنی ۱۹ است. که نشان‌دهنده آن است که این زیرگراف روی این رأس قرار گرفته است.



شکل (۶): زیرگراف کوچک متراکم با ۲۰رأس و ۱۸۸ یال.

همچنین بررسی‌ها نشان‌دهنده آن است که حداقل و حداکثر اتصال خارجی رأس‌های این زیرگراف به ترتیب صفر (رأس ۵۹۶) و هفت (رأس ۵۷۹) است. از این رأس‌ها، رأس ۴۱۵ با درجه ۱۵۹ که تعداد اتصالات داخلی آن در این زیرگراف ۱۹ است، در بردار x بازیابی نشده است. همچنین شکل (۷) زیرگراف ۱۴رأسی با ۸۹یال را نشان می‌دهد. در این زیرگراف، درجه رأس ۳۷۸۴ کمترین درجه یعنی ۱۳ است. بررسی‌ها نشان می‌دهد که حداقل و حداکثر اتصال خارجی رأس‌های این زیرگراف به ترتیب صفر (رأس ۳۷۸۴) و هفت (رأس ۳۷۴۸) است. علاوه، رأس ۳۴۳۸ با درجه ۵۴۷، در این زیرگراف درجه‌ای برابر ۱۳ دارد. این رأس در بردار x مقدار ناچیز و بازیابی نشده است. لذا طبق ویژگی‌های بررسی شده این زیرگراف‌ها می‌توان این زیرگراف‌های کوچک متراکم را به‌عنوان ناهنجاری‌های ساختاری شبکه تعیین کرد. وجود این نوع ناهنجاری‌ها به تحلیلگر این هشدار را می‌دهد که ممکن است فعالیت تهدیدآمیزی در شبکه در حال انجام باشد.



شکل (۷): زیرگراف کوچک متراکم با ۱۴رأس و ۸۹ یال.

۴-۵- نتایج آزمایش‌ها

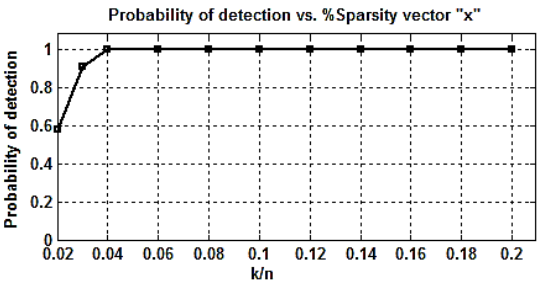
در این بخش نتایج حاصل از آزمون‌های مختلف برای ارزیابی عملکرد چارچوب پیشنهادی ارائه می‌شود. هر آزمایش‌های از پنج تکرار برخوردار است. نتایج حاصل از سه آزمون تأثیر میزان تنکی‌بردار رأس بر احتمال تشخیص، تعداد اندازه‌گیری‌های مورد نیاز جهت تشخیص، تأثیر طول هر اندازه‌گیری بر تشخیص بررسی می‌شوند. هر سه آزمایش بر روی گراف۵۱۲ رأسی R-MAT با تعبیه دو زیرگراف ناهنجار انجام‌شده است.

تأثیر میزان‌تنکی بردار رأس: آزمون تأثیر درصد تنکی بردار رأس‌ها را بر احتمال تشخیص زیرگراف‌های ناهنجار بررسی می‌کند. در این آزمایش تعداد اندازه‌گیری‌ها $m=\frac{|V|}{5}$ و طول هر اندازه‌گیری $l=\frac{|V|}{2}$ تنظیم شده است. نتایج مطابق شکل (۸) نشان می‌دهد که در گراف ۵۱۲رأسی با افزایش درصد تنکی بردار رأس‌ها، احتمال تشخیص بالاتر می‌رود. زمانی که تعداد مؤلفه‌های غیرصفر بردار x برابر ۴درصد تعداد رأس‌های گراف $\frac{k}{n}=4\%$ باشد احتمال تشخیص ۱۰۰درصد است. زمانی که تنکی در کمترین مقدار خودش $\frac{k}{n}=2\%$ باشد احتمال تشخیص بزرگ‌تر از صفر و ۵۸درصد $TPR=58\%$ است. نتایج نشان می‌دهد درحالی‌که تنکی بردار رأس‌ها k بسیار کمتر از تعداد رأس‌ها $k\ll n$ باشد احتمال تشخیص ناهنجاری‌ها به بالاترین حد ممکن می‌رسد.

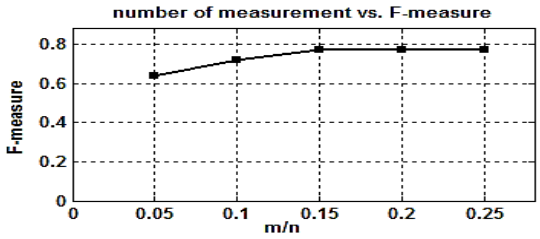
تأثیر تعداد اندازه‌گیری: این آزمون روی شاخص F در گراف ۵۱۲رأسی با تعبیه دو زیرگراف ناهنجار و با تنظیم طول اندازه‌گیری به $l=\frac{n}{2}$ مورد بررسی قرار می‌دهد. در این آزمایش تنکی بردار x به ۴درصد تنظیم شده است ($k/n=4\%$). مطابق شکل (۹)، با افزایش تعداد اندازه‌گیری‌ها، شاخص F افزایش می‌یابد و زمانی که تعداد اندازه‌گیری‌ها تنها ۱۵ درصد تعداد کل رأس‌های گراف است ($\frac{m}{n}=15\%$)، شاخصF در بالاترین مقدار خود یعنی ۷۷درصد قرار دارد. همچنین زمانی که درصد اندازه‌گیری‌ها در کمترین مقدار یعنی ۵ درصد است ($\frac{m}{n}=5\%$)، شاخص F برابر ۶۴ درصد است. نتایج نشان می‌دهد، درحالی‌که تعداد نمونه‌برداری‌ها از شبکه m بسیار کمتر از تعداد رأس‌ها است ($m\ll n$) بالاترین میزان کارایی حاصل می‌شود که هدف نمونه‌برداری فشرده است.

تأثیر طول اندازه‌گیری: این آزمون روی شاخص F در گراف ۵۱۲رأسی با تعبیه دو زیرگراف ناهنجار و با تنظیم تعداد اندازه‌گیری به $m=\frac{n}{5}$ و تنکی بردار x به ۴درصد ($k/n=4\%$) مورد بررسی قرار می‌دهد. مطابق شکل (۱۰) با افزایش طول اندازه‌گیری‌ها شاخص F نیز افزایش می‌یابد و زمانی که طول هر اندازه‌گیری تنها ۲۵درصد تعداد کل رأس‌های گراف است ($L/n=25\%$)، شاخص F در بالاترین مقدار خود یعنی ۷۳درصد قرار دارد. همچنین زمانی که نسبت اندازه‌گیری‌ها به تعداد کل رأس‌های گراف در کمترین مقدار یعنی ۵ درصد باشد شاخص F برابر ۶۶ درصد است.

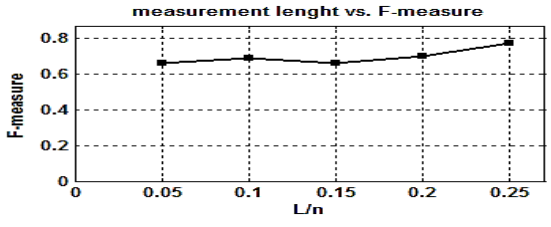
در این بخش به مقایسه روش‌های مبتنی بر پردازش سیگنال شامل دو روش Eigenvector L1 norm [۳۵] و sparsePCA [۳۷] با چارچوب پیشنهادی می‌پردازیم.



شکل (۸): تأثیر درصد تنکی بردار x بر احتمال تشخیص ناهنجاری‌ها.



شکل (۹): تأثیر تعداد اندازه‌گیری‌ها بر روی شاخصF.



شکل (۱۰): تأثیر طول اندازه‌گیری l بر شاخصF.

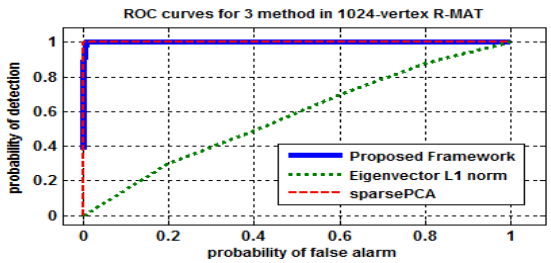
در جدول (۴) کارایی این روش‌ها ازنظر پیچیدگی محاسباتی و تعداد زیرگراف‌های ناهنجار تشخیصی مورد مقایسه قرار

گرفته‌اند. بعد از پیاده‌سازی دو روش Eigenvector L1 norm و spaesePCA و استخراج نتایج آن‌ها، قدرت تشخیص این روش‌ها در مقایسه با چارچوب پیشنهادی با رسم منحنی ROC در گراف ۱۰۲۴رأسی با تعبیه زیرگراف ناهنجار ۸رأسی مورد بررسی قرارگرفت. سطح زیر منحنی AUC[1] به‌عنوان معیار مقایسه قدرت تشخیص هرکدام از روش‌ها اندازه‌گیری است. در چارچوب پیشنهادی تنکی بردار رأس‌ها، تعداد اندازه‌گیری‌ها و طول هر اندازه‌گیری به ترتیب به $\frac{k}{n} = 4\%$ ،$m = \frac{n}{5}$ و $l = \frac{n}{2}$ تنظیم شده است.

جدول (۴): میزان پیچیدگی و تشخیص روش‌های مختلف و چارچوب پیشنهادی.

روش	پیچیدگی محاسباتی	زیرگراف ناهنجار تشخیصی
Eigenvector L_1 norm	$O(\|E\|d + \|V\|d^2)$	زیرگراف غیرهمزمان کم
sparsePCA	$O(n^4\sqrt{\log n}/\epsilon)$	فقط یک زیرگراف
Proposed Framework	$O(n^2 + m \times l \times n + k^2)$	زیرگراف هم‌زمان زیاد

شکل (۱۱) قدرت تشخیص این سه روش را با رسم منحنی‌های ROC مقایسه می‌کند. منطبق‌شدن منحنی ROC روش Eigenvector L1 norm بر قطر مربع و معیار AUC برابر ۰/۵۷ نشان‌دهنده عدم توانایی تشخیص این روش است. چارچوب پیشنهادی وروش sparsePCA تقریباً قدرت تشخیص مشابه و بالایی رادارند. معیار AUC این دو روش به ترتیب برابر ۰/۹۹و ۱ است. نتایج نشان می‌دهد که چارچوب پیشنهادی با پیچیدگی محاسباتی $o(n^2)$ کمتر از روش sparsePCA از قدرت تشخیص مشابه برخوردار است. حتی چارچوب پیشنهادی برخلاف روش sparsePCA قادر به تشخیص چندین زیرگراف ناهنجار هم‌زمان است.



شکل (۱۱): منحنی ROC روش‌های مختلف و چارچوب پیشنهادی.

تحلیل پیچیدگی: پیچیدگی محاسباتی چارچوب پیشنهادی ناشی از الگوریتم نمونه‌برداری CS-AnomalyDet و الگوریتم شناسایی زیرگراف‌های ناهنجار است که در این پژوهش پیشنهاد شده است. در ادامه هزینه محاسباتی هرکدام از این الگوریتم‌ها محاسبه می‌شود. با توجه به گراف شبکه $G = (V, E)$، مطابق فرض الگوریتم CS-AnomalyDet که هر رأس یک جدول درهم[2] را برای همسایگانش نگه می‌دارد، بررسی اینکه آیا یک رأس، همسایه رأس دیگر است می‌تواند تقریباً در زمان ثابتی انجام شود. سپس، یک الگوریتم پیمایش گراف می‌تواند بررسی کند، چه تعداد از همسایگان رأس v_i در لیست همسایگان v_i ذکر شده‌اند که برای هر رأس v_i به‌طور محلی و در زمان $|N^1(v_i)||N^1(v_i) - 1|$ انجام می‌شود. همچنین بررسی اینکه چه نسبت از همسایگان رأس v_i درجه کمتر از میانگین درجه گراف دارند، در زمان ثابتی انجام خواهد شد.

بررسی‌های ذکرشده به‌طور محلی انجام می‌شود و هر رأس حداکثر $n-1$ همسایه دارد. هزینه محاسباتی خطوط ۷-۹ الگوریتم (۱) در بدترین حالت $O(n^2)$ با $|V| = n$ است. خطوط ۱۱-۱۴ الگوریتم (۱) می‌تواند خارج از حلقه و فقط یک‌بار انجام شود. لذا هزینه محاسباتی عبارت $\sum_{u\in V} W(u)$ برابر $O(n)$ است و محاسبه احتمالات شروع $P(v)$ برای انتخاب بهترین رأس شروع به‌صورت محلی و برای هر رأس v در زمان ثابتی انجام می‌شود. در این الگوریتم، انتخاب رأس بعدی از طریق بررسی مقادیر محاسبه‌شده در هر رأس می‌تواند در زمان $O(n)$ انجام شود (خطوط۱۶-۲۵). اختصاص رأس بعدی در خط ۲۰ در زمان‌ثابت انجام می‌شود لذا زمان محاسبات کل الگوریتم نمونه‌برداری CS-AnomalyDet در بدترین حالت برابر $O(n^2 + m \times l \times n)$ است. در الگوریتم (۲) شناسایی زیرگراف‌های ناهنجار با بالاترین هزینه محاسباتی مربوط به خطوط ۴-۱۴ است که در آن، حلقه‌های Foreach موجود در خطوط ۴و۵ با ترتیب زمانی $O(k^2)$ انجام می‌شود که در آن k میزان‌تنکی بردار x است. با توجه به اجرای پشت سر هم دو الگوریتم، هزینه‌کل محاسبات تشخیص زیرگراف‌های ناهنجار در بدترین حالت $O(n^2 + m \times l \times n + k^2)$ می‌باشد که در آن، n تعداد رأس‌های گراف، m تعداد اندازه‌گیری‌های شبکه، l طول هر اندازه‌گیری و k میزان‌تنکی بردار رأس‌ها است.

۵- نتیجه‌گیری

این مقاله رویکرد تشخیص ناهنجاری بدون علامت و الگوریتم جدید نمونه‌برداری بر اساس پردازش‌سیگنال‌های مرتبط با اطلاعات محلی ارائه داد که قادر به‌تعیین هم‌زمان زیرگراف‌های متراکم ناهنجار در گراف ناشناخته نویزی شبکه‌های اجتماعی بزرگ است. الگوریتم نمونه‌برداری پیشنهادی مبتنی‌بر نمونه‌برداری فشرده جهت بازیابی ویژگی‌های تنک شبکه‌های ثابت با هدف بهبود دقت‌تشخیص ناهنجاری در کنار کاهش

[1] Area Under Curve (AUC)

[2] Hash Table

پیچیدگیِ نمونه‌برداری داده‌ها است. نتایج آزمایش‌های تجربی نشان داد که رویکرد پیشنهادی علاوه بر برخورداری از دقت تشخیص هم‌زمان چندین زیرگراف متراکم، پیچیدگی محاسباتی را به‌طرز قابل توجهی کاهش داد. از نظریه نمونه‌برداری فشرده برای بازیابی داده‌های تنک استفاده شده که به کاهش زیاد پیچیدگی محاسباتی کمک کرده است تا هم‌زمان چندین زیرگراف ناهنجار را تشخیص دهد. معرفی ناهنجاری به‌عنوان ویژگی تنک در شبکه‌های اجتماعی و استفاده از نظریه نمونه‌برداری فشرده برای بازیابی آن‌ها معرفی شد. با توجه به موانع موجود و چالش‌های متعدد تشخیص ناهنجاری، نیاز فزاینده به مطالعه سازمان‌یافته و کاربردی در حوزه شبکه‌های اجتماعی احساس شد. کارهای آینده ما با هدف مدل‌سازی موازی چارچوب پیشنهادی جهت افزایش دقت و کاهش زمان اجرا در شبکه‌های بزرگ پیچیده با هدف تصمیم‌سازی‌های بی‌درنگ به‌کمک نمونه‌برداری فشرده و در راستای تشخیص ناهنجاری در شبکه‌های اجتماعی جهت حفظ حریم خصوصی کاربران خواهد بود. همچنین افزایش سرعت تشخیص و دقت تعیین زیرگراف‌های متراکم نیازمند تعریف تراکم، مدل تراکم، معیارهای اندازه‌گیری تراکم و تعیین محدوده آن است که بررسی خواهیم داد.

۶- مراجع

[1] Yu, Rose, et al., “A Survey on Social Media Anomaly Detection.” ACM SIGKDD Explorations Newsletter, vol. 18(1), pp. 1-14, 2016.

[2] Y. Yasami and F. Safaei, “A statistical in nite feature cascade-based approach to anomaly detection for dynamic social networks,” Computer Communications, vol. 100, pp. 52-64, 2017.

[3] Xiong, Fei, Yun Liu, and Junjun Cheng, “Modeling and predicting opinion formation with trust propagation in online social networks,” Communications in Nonlinear Science and Numerical Simulation, vol. 44, pp. 513-524, 2017.

[4] N. Rastogi and J. Hendler, “Graph Analytics for anomaly detection in homogeneous wireless networks A Simulation Approach,” arXiv preprint arXiv:1701.06823, 2017.

[5] M. Ahmed, A. N. Mahmood, and J. Hu, “ A survey of network anomaly detection techniques,” J. Netw. Comput. Appl., vol. 60, p. 1931, 2016.

[6] A. Zargar, A. Nowroozi, and R. Jalili, “XABA: A zero-knowledge anomaly-based behavioral analysis method to detect insider threats,” Information Security and Cryptology (ISCISC), 2016 13th International Iranian Society of Cryptology Conference on. IEEE, 2016.

[7] D. Mutz, F. Valeur, G. Vigna, and C. Kruegel, “Anomalous system call detection, ACM Trans,” Inf. Syst. Secur., vol. 9(1), p. 6193, 2006.

[8] R. Chaker, Z. Al Aghbari, and I. N. Junejo, “Social network model for crowd anomaly detection and localization.” Pattern Recognition, vol. 61, pp. 266-281, 2017.

[9] V. Krebs, “Mapping networks of terrorist cells,” Connections, vol. 14(3), p. 4352, 2002.

[10] E. Bastami, A. Mahabadi, and E. Taghizadeh, “A gravitation-based link prediction approach in social networks,” Swarm Evol Comput., vol. 44, pp. 176–186, 2019.

[11] B. A. Miller, et al., “Detection theory for graphs,” Lincoln Laboratory Journal, vol. 20(1), pp. 10-30, 2013.

[12] A. Mohiuddin, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques.” Journal of Network and Computer Applications. 60 (2016): 19-31.

[13] Miller, B. A., Beard, M. S., Wolfe, P. J., & Bliss, N. T. “A spectral framework for anomalous subgraph detection,” IEEE Transactions on Signal Processing, vol. 63(16), pp. 4191-4206, 2015.

[14] W. Xu, E. Mallada, and A. Tang, “Compressive sensing over graphs,” INFOCOM, Proceedings IEEE, pp. 2087-2095, 2011.

[15] S. Wang, J. Cao, and P. Yu, “Deep learning for spatio-temporal data mining: A survey,” IEEE Transactions on Knowledge and Data Engineering, 2020.

[16] Y. Wang, et al., “Data-Driven Sampling Matrix Boolean Optimization for Energy-E cient Biomedical Signal Acquisition by Compressive Sensing,” IEEE Transactions on Biomedical Circuits and Systems, 2016.

[17] V. J. Barranca, et al., “Ecient image processing via compressive sensing of integrate-and-fire neuronal network dynamics,” Neurocomputing, vol. 171, pp. 1313-1322, 2016.

[18] J. Xiaobo, et al., “An improved sparse reconstruction algorithm for speech compressive sensing using structured priors,” Multimedia and Expo (ICME), 2016 IEEE International Conference, pp. 1-6, 2016.

[19] Z. Liu, et al., “Path reconstruction in dynamic wireless sensor networks using compressive sensing,” IEEE/ACM Transactions on Networking, vol. 24(4), pp. 1948-1960, 2016.

[20] H. T. Wai, A. Scaglione, and A. Leshem, “Active sensing of social networks,” IEEE Transactions on Signal and Information Processing over Networks, vol. 2(3), pp. 406-419, 2016.

[21] J. Madhuka, et al., “Compressive sensing for efficient health monitoring and effective damage detection of structures,” Mechanical Systems and Signal Processing, vol. 84, pp. 414-430, 2017.

[22] W. Xue, et al., “Kryptein: a compressive-sensing-based encryption scheme for the internet of things,” Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 169-180, 2017.

[23] N. Shrivastava, A. Majumder, and R. Rastogi, “Mining (social) network graphs to detect random link attacks,” Proc. IEEE Int. Conf. on Data Engineering. (2008): 486-495.

[24] C. C. Noble and D. J. Cook, “Graph-based anomaly detection,” Proce. ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, pp. 631-636, 2003.

[25] W. Eberle and L. Holder, “Anomaly detection in data represented as graphs,” Intelligent Data Analysis, vol.11(6), pp. 663-689, 2007.

[26] B. Miller, N. Bliss, and P. J. Wolfe, “Subgraph detection using eigenvector L1 norms,” Proce. Int. Conf. on Neural Information Processing Systems, pp. 1633-1641, 2010.

[27] B. A. Miller, M. S. Beard, and N. T. Bliss, “Eigenspace analysis for threat detection in social networks,” in Proce. IEEE Int. Conf. on Information Fusion. (2011): 1-7.

[28] Singh N., Miller B. A., Bliss N. T., Wolfe P. J. “Anomalous subgraph detection via sparse principal component analysis,” Proce. IEEE Statistical Signal Process, pp. 485–488, 2011.

[29] P. Bindu and P. S. Thilagam, "Mining social networks for anomalies: Methods and challenges," Netw. Comput. Appl. vol. 68, pp. 213-229, 2016.

[30] Y. Yasami and F. Safaei, "A statistical infinite feature cascade-based approach to anomaly detection for dynamic social networks," Comput. Communi., 2016.

[31] A. Fattaholmanan, "Sparse Recovery in Peer-to-Peer Networks via Compressive Sensing," M.Sc. Thesis Dept. Computer Engineering (CE), Sharif University of Technology, 2012.

[32] J. Haupt, W. U. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," IEEE Signal Processing Magazine, vol. 25(2), pp. 92-101, 2008.

[33] A. Fattaholmanan, H. R. Rabiee, P. Siyari, A. Soltani-Farani, and A. Khodadadi, "Peer-to-peer Compressive Sensing for Network Monitoring," IEEE Commu. Letters, vol. 19(1), pp. 38-41, 2015.

[34] M. Mahyar, "Detection of top-k central nodes in social networks: A compressive sensing approach," IEEE Inte. Conf. on Advances in Social Networks Analysis and Mining, pp. 902-909, 2015.

[35] H. Mahyar, et al., "CS-ComDet: A compressive sensing approach for inter-community detection in social networks," Proce. IEEE Int. Conf. on Advances in Social Networks Analysis and Mining, pp. 89-96, 2015.

[36] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Commun. on pure and applied mathematics, vol. 59(8), pp. 1207-1223, 2006.

[37] R. Tibshirani, "Regression shrinkage and selection via the lasso." Royal Statistical Society. Series B (Methodological), pp. 267-288, 1996.

[38] D. Savage, et al., "Anomaly detection in online social networks," J. Soc. Netw., vol. 39, pp. 62-70, 2014.

[39] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-MAT: A Recursive Model for Graph Mining," SDM, pp. 442-446, 2004.

[40] R-MAT source code, NetMine package. [Online] Available: http://faculty.mccombs.utexas.edu/deepayan.chakrabarti/my www/software/NetMine-Basic-03-30-2004.tgz.

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۲، تابستان ۱۴۰۰، ص ۱۹۴-۱۷۹