Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 12, Number 3, Autumn 2023, pp. 85-98
Received: 01/30/2022, Revised: 11/24/2023, Accepted: 01/07/2024

Increasing the Accuracy of Sibyl Attack Detection in Social Networks using Hybrid Clustering Method based on a Graph-structure

Amirmohamad Shahparastan[1], Amineh Amini[2*] and Hadi Saboohi[3]
1- Master student, Department of Computer Engineering, Karaj Branch, Islamic Azad University Karaj, Iran.
2*- Assistant Professor, Department of Computer Engineering, Karaj Branch, Islamic Azad University Karaj, Iran.
3- Assistant Professor, Department of Computer Engineering, Karaj Branch, Islamic Azad University Karaj, Iran.
[1]Amirshahparastan90@gmail.com, [2*] aamini@kiau.ac.ir, and [3]saboohi@kiau.ac.ir

Corresponding author's address: Amineh Amini, Assistant Professor, Department of Computer Engineering, Islamic Azad University Karaj Branch, Iran.

Abstract- Sybil attacks are increasingly growing and expanding in social networks. A malicious user with a fake identity, known as a Sybil attack, can create a large number of fake accounts to generate spam, impersonate other users, commit fraud, and gain access to many legitimate users' information. For security reasons, such fake accounts should be identified and disabled. Various identification methods have been proposed to deal with fake accounts. However, most of these methods detect fake accounts using social structure graphs, which leads to poor performance, or use machine learning methods, which have low accuracy for identifying Sybil attacks. In this paper, a hybrid clustering method called CRNM is proposed. The proposed method is based on clustering, so that by combining different community detection methods; A new community detection method is presented. The combination of these methods has led to higher accuracy, more reliable results and more stability. The CRNM method has been evaluated on datasets collected from Twitter, Reddit, Instagram and Facebook. Unlike other machine learning-based approaches, the proposed method focuses on different levels of user profile features. The evaluation results have shown that the CRNM method detects Sybil nodes with an accuracy of 85.13%.

Keywords- Sibyl attack, Fake accounts, Hybrid clustering, Detection of fake accounts, Community detection, CNRM

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۲، شماره ۳، پائیز ۱۴۰۲ ، صفحه ۸۵-۹۸
دریافت: ۱۴۰۰/۱۱/۱۰، بازنگری: ۱۴۰۲/۰۹/۰۳، پذیرش: ۱۴۰۲/۱۰/۱۷

افزایش دقت تشخیص حملات سیبیل در شبکه‌های اجتماعی با استفاده از روش خوشه بندی ترکیبی بر روی گراف ساختاری

امیرمحمد شاه‌پرستان[۱]، امینه امینی*[۲]، هادی صبوحی[۳]
۱- دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران.
*۲- استادیار، گروه مهندسی کامپیوتر، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران.
۳- استادیار، گروه مهندسی کامپیوتر، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران.
[1]Amirshahparastan90@gmail.com, [2*]aamini@kiau.ac.ir, [3]saboohi@kiau.ac.ir

* نشانی نویسنده مسئول: امینه امینی، استادیار، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کرج، ایران.

چکیده- حملات سیبیل به طور فزاینده‌ای در شبکه‌های اجتماعی در حال رشد و گسترش است. یک کاربر مخرب با هویت جعلی که از آن تحت عنوان حمله سیبیل یاد می‌شود، می‌تواند تعداد زیادی حساب جعلی برای تولید هرزنامه، جعل هویت سایر کاربران، کلاهبرداری و دسترسی به بسیاری از اطلاعات کاربران قانونی ایجاد کند. به دلایل امنیتی، چنین حساب‌های جعلی باید شناسایی و غیرفعال شوند. روش‌های شناسایی مختلفی برای مقابله با حساب های جعلی پیشنهاد شده است. با این حال، بیشتر این روش‌ها حساب‌های جعلی را یا با استفاده از گراف‌های ساختاری اجتماعی شناسایی می‌کنند که منجر به عملکرد ضعیف می‌شود و یا از روش های یادگیری ماشین استفاده می‌شود که دقت پایینی برای شناسایی حملات سیبیل دارند. در این مقاله، یک روش به‌نام خوشه‌بندی ترکیبی پیشنهاد شده است که CNRM نامیده شده است. روش پیشنهادی مبتنی بر خوشه‌بندی می‌باشد، بدین صورت که با ترکیب روش‌های مختلف تشخیص اجتماعات؛ یک روش تشخیص اجتماع جدید ارائه شده است. ترکیب این روش‌ها منجر به دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری شده است. روش CNRM بر روی مجموعه داده‌های جمع آوری شده از توییتر، ردیت، اینستاگرام و فیسبوک ارزیابی شده است. برخلاف سایر رویکردهای مبتنی بر یادگیری ماشین، روش پیشنهادی بر روی سطوح مختلفی از ویژگی‌های پروفایل کاربران تمرکز می‌کند. نتایج ارزیابی نشان داده است که روش CNRM گره‌های سیبیل را با دقت ۸۵,۱۳% تشخیص می‌دهد.

واژه‌های کلیدی: حمله سیبیل، حساب‌های جعلی، خوشه‌بندی ترکیبی، تشخیص حساب‌های جعلی، تشخیص جوامع، CNRM.

۱- مقدمه

تحلیل شبکه‌های اجتماعی، روشی برای بررسی نظم و سازمان موجود روابط انسان‌ها در اجتماع است. نتیجه‌ای که تحلیل شبکه‌های اجتماعی در اختیارمان قرار می‌دهد، اجازه می‌دهد که بتوان به‌طور دقیق و کمی، الگوهایی که در میان روابط مردم در اجتماع وجود دارد را تشخیص داده، ویژگی‌های این الگوها را مورد اندازه‌گیری قرار داد و آن‌ها را به شکلی که برای تحلیل‌های خودکار مناسب است درآورد [۳]. خوشه‌بندی مشتریان و کاربران وب که علایق و سلیقه‌های مشترکی دارند و یا ازلحاظ جغرافیایی نزدیک به هم هستند، ممکن است باعث ارتقا و بهبود کیفیت خدمات ارائه‌شده به آن‌ها شود، به این دلیل که هر خوشه‌ای از مشتریان، می‌تواند توسط یک کارگزار وقف‌شده، مورد خدمت‌رسانی قرار بگیرد [۲]. تعیین محبوبیت کاربران از شبکه‌های اجتماعی، خطر بالایی را برای کاربران این شبکه‌ها ایجاد می‌کند؛ زیرا مقدار زیادی از داده‌های شخصی که کاربران در این بسترها به اشتراک می‌گذارند، آن‌ها را به

یک هدف مطلوب برای مهاجمان (هکرها) تبدیل می‌کند. یکی از شایع‌ترین این حملات، حمله سیبیل (Sybil) است. حمله سیبیل، حمله‌ای است که در آن مهاجم با ایجاد تعداد زیادی از هویت‌های جعلی، شهرت یک سرویس شبکه اجتماعی را مختل می‌کند تا بتواند بدینوسیله در آن شبکه تأثیر گذاری نامطلوب داشته باشد [۴، ۵].

در مقاله [۱] از الگوریتم تله سیبیل (SybilTrap) استفاده شده است که از یک تکنیک نیمه نظارت شده استفاده می‌کند که به طور خودکار ویژگی‌های اساسی فعالیت‌های کاربر را با ساختار اجتماعی در یک سیستم ادغام می‌کند. هدف این روش کار بر روی نمودارهای تعاملی و اجتماعی است. این در جایی که تعداد زیادی لبه حمله وجود دارد موثرتر است. در مقاله [۱] از خوشه‌بندی نیمه‌نظارتی که از داده‌های برچسب‌گذاری‌شده بسیار کمی برای شناسایی دقیق گره‌های سیبیل در مقیاس بزرگ استفاده می‌کنند، بهره گرفته شده است. بزرگترین چالشی که این روش با آن مواجه هست مقیاس پذیری است که تنها بر روی شبکه های اجتماعی بزرگ و اجتماعات عظیم مورد استفاده است. همچنین در مقاله [۲] از روش حد سیبیل(SybilLimit) استفاده شده است، که یک دفاع تقریباً بهینه در برابر حملات سیبیل از شبکه‌های اجتماعی است. اما در چندین مسئله دچار نقص است از جمله: استفاده از چندین نمونه مستقل از پروتکل مسیر تصادفی برای انجام بسیاری از مسیرهای تصادفی کوتاه و بهره برداری از تقاطع ها در لبه ها به جای گره‌ها؛ که باعث می‌شود مدت زمان اجرای طولانی را طی کند.

روش‌های موجود معمولاً به علت استفاده نکردن از تفکیک سازی جوامع دقت بالایی ندارند. با استفاده از تکنیک یادگیری ماشین نیمه نظارتی، خوشه بندی ترکیبی و تفکیک گراف کاربران به گراف‌های اجتماعی و تعاملی، دقت تشخیص و مقایسه کاربران عادی و جعلی نسبت به هم افزایش می‌یابد. با توجه به این موضوع در این روش بر روی داده‌های بدون برچسب و بر چسب دار کار می‌شود که در شناسایی حملات هدفمند موثر و دقیق است، زیرا می‌تواند سطوح مختلف ویژگیهای پروفایل کاربران را به طور دقیق و موثرتری شناسایی کند.

برای ارزیابی دقت، چندین آزمایش جهت بررسی اینکه آیا تکنیک‌های موجود تأثیر مثبتی بر استراتژی‌های دفاعی موجود در تشخیص سیبیل دارند، انجام شده و جوامع با هم مقایسه شده‌اند. روش‌های موجود دارای دقت پایینی هستند چون اولاً شبکه‌های اجتماعی دارای تعداد گره و خوشه و دسته‌ی ثابتی نیست و درضمن دارای داده‌های پرت زیادی می‌باشد.

با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های پایه‌ای، در این مقاله، یک روش خوشه‌بندی ترکیبی پیشنهاد شده است که CNRM نامیده شده است. در CNRM، با استفاده از ترکیب روش‌های مختلف تشخیص اجتماعات، یک روش جدید ارائه شده است. که در نهایت ترکیب این روش‌ها منجر به دقت تشخیص بالاتر در شناسایی گره‌های سیبل می‌شود.

این مقاله بدین شرح سازماندهی شده است. بخش ۲، پیشینه پژوهش و کارهای مرتبط توصیف و تشریح شده است. بخش ۳، روش پیشنهادی و جزئیات آن بیان شده است. بخش ۴، معیار ارزیابی و آزمایشات مورد بحث قرار گرفته است. بخش ۵، به بحث و نتیجه گیری پرداخته شده است.

۲-پیشینه پژوهش

در [۱۰] نویسندگان از نزدیکی و گام‌های تصادفی برای ارزیابی نرمال بودن گره‌ها در گراف‌های دو بخشی استفاده کردند. در این پژوهش دو مورد را در گراف در نظر گرفته‌اند: ۱) شناسایی گره‌های مشابه (تشکیل محله) و ۲) یافتن گره‌های غیرطبیعی (تشخیص ناهنجاری). همچنین الگوریتم‌هایی برای محاسبه محله برای هر گره با استفاده از پیاده‌روی تصادفی با راه‌اندازی مجدد و پارتیشن‌بندی نمودار پیشنهاد شده است.

از سوی دیگر، در [۱۱] نویسندگان پیشنهاد کردند که فرد می‌تواند از محاسبه مقیاس‌های مختلف مرتبط با گره‌ها در ساختار گراف معین، گراف‌های دوتایی، سه‌تایی، جوامع و همچنین ساختار گراف جهانی بدست آورد. این پژوهش همچنین از الگوریتمی برای تشخیص گره‌های غیرعادی در گراف‌های وزن‌دار به نام OddBall استفاده کرده است.

پژوهش [۱۲] الگوریتم پیشنهادی در میزبان‌های موجود در شبکه را به کلاس‌های ترافیک، برقرارکننده اتصال و پاسخ دهنده اتصال دسته‌بندی کرده است. کلاس ترافیک شامل میزبان‌هایی است که میزان ارسال ترافیک در آن‌ها بیشتر از میزان دریافت ترافیک می‌باشد. میزبان‌هایی که درجه خروجی بسیار بالا و نامتعارف دارند، در کلاس برقرارکننده اتصال قرار می‌گیرند. همچنین میزبان‌هایی که در ارتباطات دوسویه بسیاری وجود داشته باشند در کلاس پاسخگو قرار می‌گیرند. این کلاس‌ها به گونه‌ای تعریف می‌شوند که تنها میزبان‌های مشکوک در آن‌ها قرار می‌گیرند.

در پژوهش [۱۳] رویکرد متفاوتی در شناسایی اجتماعات مخرب تحت عنوان کرم ارائه شده است. نویسندگان به بررسی ارتباط میان جریان‌های شبکه و اطلاعات ضبط شده توسط طعمه‌ها پرداختند. آن‌ها نیاز دارند تا به یک سری سیستم مطمئن جهت اعتبارسنجی ترافیک شبکه دسترسی داشته باشند. آن‌ها با جمع آوری اطلاعات جریان شبکه و اطلاعات ضبط شده طعمه‌ها یک پایگاه داده تولید

کرده و بر مبنای آن امضای مبتنی بر جریان کرم‌ها را به دقت استخراج کردند.

در [۱۴] نویسندگان ساختار اجتماعی در شبکه‌ها را براساس مسئله تشخیص و توصیف این ساختار اجتماعی را تشریح کرده‌اند؛ که یکی از موضوعات برجسته در مطالعه سیستم‌های شبکه‌ای است. روش پیشنهادی برای تشخیص این جامعه‌ها، حول ایده استفاده از شاخص‌های محوری برای یافتن مرزهای جامعه توصیف می‌گردد.

در پژوهش [۱۵] روشی جهت شناسایی کرم لیست برخورد ارائه کردند. استراتژی پویش مرحله‌ای مورد استفاده توسط کرم لیست برخورد از یک لیست از میزبان‌های آسیب‌پذیر تشکیل شده، که به طور مداوم آنلاین هستند. این لیست معمولاً توسط حمله‌گر آماده می‌شود. دلیل استفاده از لیست برخورد این است که مرحله مقداردهی اولیه جهت انتشار کرم بسیار کند است. استفاده از یک لیست برخورد باعث سرعت بخشیدن به مرحله انتشار کرم می‌شود.

بر اساس پژوهش [۱۶]، در طی سال‌های ۲۰۰۱ تا ۲۰۰۴ به طور متوسط در هر ساعت ۲۴.۵ آدرس IP متفاوت در سراسر دنیا قربانی حمله محدود کننده سرویس می‌گردند. این مسئله ضرورت توجه پژوهشگران به حملات محدود کننده سرویس را نشان می‌دهد. شایان ذکر است که روش شناسایی نفوذ مبتنی بر جریان در برابر حملات محدود کننده سرویس که ترافیک شبکه را دچار سربار مضاعف می‌کند قادر به شناسایی و انجام واکنش است. در مورد حملات کند کننده سرویس مفهومی که اختلال در ارائه سرویس به دلیل محتوای بسته مشکوک است، این روش‌ها قادر به شناسایی و انجام واکنش نمی‌باشند.

مرجع [۱۷] روش تشخیص جامعه جدید در مدل شبکه را ارائه می‌کند. در این پژوهش الگوریتم‌هایی برای کشف ساختار جامعه در شبکه‌ها تقسیمات طبیعی گره‌های شبکه به زیر گروه‌های متراکم متصل پیشنهاد و مطالعه شده است. الگوریتم‌های این پژوهش همه دارای دو ویژگی قطعی هستند: اول، آن‌ها شامل تکرار لبه‌ها از شبکه برای تقسیم آن در جوامع هستند، لبه‌های برداشته‌شده با استفاده از هر یک از تعدادی از اقدامات «بین» ممکن و در مرحله دوم، شناسایی می‌شوند. پس از هر حذف، دوباره محاسبه می‌شود.

در پژوهش [۱۸] در ابتدا با استفاده از یک ساختار با نام روش اولیه، اجتماع داده‌های جریان شبکه را ذخیره می‌کنند. روش اولیه یک جدول یک بعدی درهم است که جهت دسترسی سریع به داده‌های ذخیره شده بسیار مناسب می‌باشد. توسط این جدول می‌توان تعداد رخداد مرتبط با یک رویداد را شمارش کرد. روش اولیه این امکان را فراهم می‌کند که یک سری ویژگی‌های آماری در مورد تغییرات ترافیک در طول زمان به دست آورده شود.

جدول ۱. جزئیات تحقیقات اخیر

منبع	مشکل	هدف	روش	نتیجه
[۱۰]	پیچیدگی بسیاری از پروفایل‌های سیبیل	شناسایی و محافظت در برابر سیبیل	یادگیری عمیق	رسیدن به دقت ۸۰٪
[۱۱]	اتلاف زیاد زمان	حفاظت در برابر سیبیل	خوشه‌بندی و قدم‌زدن تصادفی در گراف	۱۴ ٪ کاهش در نرخ مثبت کاذب
[۱۲]	اخلال در امنیت و حریم خصوصی و تأخیر در شناسایی	تاثیر بهتر در برابر حمله سیبیل	یادگیری ماشین	۷۶٪ از موارد شناسایی سیبیل در جوامع است (با روش لوین تشخیص داده می‌شود)
[۱۳]	تشخیص درست جوامع	شناسایی سیبیل	مبتنی بر شباهت (گراف بیس)	عملکرد خوب هم در شناسایی جامعه و هم در دفاع از حمله سیبیل
[۱۴]	حملات سایبری و نقض حریم خصوصی	حفاظت در برابر سیبیل	خوشه بندی انتها به انتها	عملکرد بهتر در طبقه بندی جوامع
[۱۵]	اتلاف زیاد زمان	دفاع در برابر حملات سیبیل	قدم زدن تصادفی (گراف بیس)	٪۹۳.۳۳TPR= ٪۷- ٪۰FPR=
[۱۶]	حمله سیبیل	مقابله با سیبیل	قدم زدن تصادفی (گراف بیس)	متمایزسازی جوامع
[۱۷]	کارهای خصمانه توسط پروفایل‌های جعلی	حذف و شناسایی زود هنگام حملات سیبیل	گراف بیس	دستیابی به عملکرد شناسایی بیش از ۷۵٪
[۱۸]	مشکلات در شناسایی سیبیل	به حداقل رساندن نشت اعتماد	قدم زدن تصادفی (گراف بیس)	پایداری سیستم در برابر حملات
[۱۹]	فعالیت‌های مخرب کاربران جعلی (حملات سیبیل)	شناسایی پروفایل‌های جعلی	یادگیری ماشین	طبقه‌بندی جوامع با دقت بالا
[۲۰]	محدودیت‌های روش‌های موجود	شناسایی سیبیل	مبتنی بر ساختار	۷۶٪ از موارد شناسایی سیبیل در جوامع
[۲۱]	حملات سیبیل	شناسایی سیبیل	جستجوی اول عمیق (گراف بیس)	پایداری سیستم در برابر حملات

در [۲۰] از تکنیک بهینه‌سازی کلونی مورچگان جهت پیشنهاد دوست استفاده کردند. در واقع در این پژوهش از دو تکنیک هوش مصنوعی به عنوان یک روش بهینه شده ترکیبی جهت ارائه دوستان در شبکه‌های برخط با هدف شناسایی اجتماع و حلقه‌های دوستی استفاده‌شده است. تمرکز اصلی این روش بر روی ارتباطات موجود بین اعضای گراف بوده است. درنهایت یک لیست مرتب شده از پیشنهادات تولید و به کاربران ارائه می‌گردد.

در پژوهش [۲۱] ابتدا انواع حملات محدود کننده سرویس در قالب الگوهای ترافیکی بررسی شده است. در واقع یک الگوی ترافیکی یا امضای حمله عبارت است از تعداد و اندازه جریان‌های شبکه و بسته‌ها و نیز میزان پهنای باند که در خلال حمله استفاده می‌شود.

جدول (۱) جزئیات و معیارهای ارزیابی تحقیقات اخیر را نشان می‌دهد.

۳- روش پیشنهادی

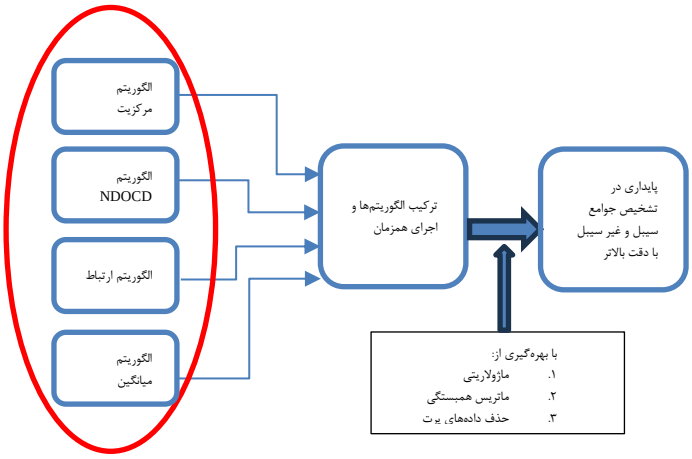
روش پیشنهادی مبتنی بر خوشه‌بندی می‌باشد. بدین صورت که با اسفاده از روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت، یک روش تشخیص اجتماع جدید پیشنهاد و ارائه شده است؛ که ترکیب این روش‌ها منجر به دقت بالاتر و نتایج مطمئن‌تر و پایداری بیشتری خواهد داشت. به طور کلی الگوریتم‌های خوشه‌بندی یکی از روش‌های اصلی در داده کاوی است که در جهت کاوش الگوهای پنهان مورد استفاده قرار می‌گیرد و با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های پایه‌ای آن، در این مقاله از روش خوشه‌بندی ترکیبی استفاده شده است [۱۶,۳۱]. در شکل (۱) فرآیند کلی روش پیشنهادی بنام خوشه بندی ترکیبی CNRM مشخص شده است.CNRM به اختصار از ترکیب الگوریتم‌های زیر انتخاب شده است: الگوریتم مرکزیت (Centrality)، الگوریتم رابطه (Relation)، الگوریتم تشخیص همپوشانی جوامع با استفاده از تجزیه شبکه (NDOCD) و الگوریتم میانگین (Mean). الگوریتم روش پیشنهادی مطابق شبه کد (۱) می‌باشد:

شبه کد ۱. شبه کد روش پیشنهادی CNRM

Algorithm: *CNRM*
Input: $G = \{V.E\}$
Output: *Stability in community recognition* (*Sibyl*)
1: Perform following four algorithms (A, B, C and D) simultaneously
2: **Begin**
3: A. Centrality algorithm (Refer to as Psuodocode 2)
4: B. NDOCD algorithm (Refer to as Psuodocode 3)
5: C. Relation algorithm (Refer to as Psuodocode 4)
6: D. K-mean algorithm (Refer to as Psuodocode 5)
7: **End**

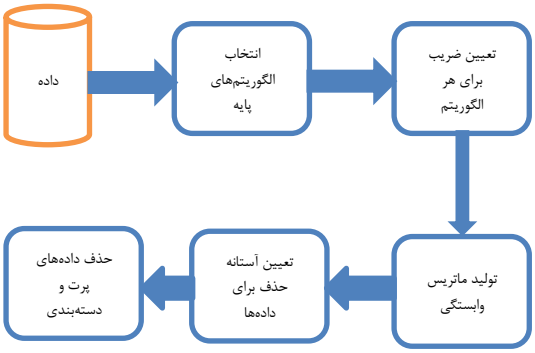
بعد از اعمال الگوریتم‌های مشخص شده در فلوچارت بر روی شبکه مورد تحلیل، یک سری نتایج متفاوت به‌دست خواهد آمد که به صورت $\Pi^* = \{\pi_1.\pi_2 \ldots .\pi_m\}$ تعریف می‌گردد. با ترکیب صحیح روش‌های مختلف با استفاده از خوشه‌بندی می‌توان به یک نتیجه بسیار بهتر، مطمئن‌تر، پایدارتر و دقیق رسید.

بنابراین تشخیص اجتماعات با ترکیب نتایج به دست آمده از الگوریتم‌های پایه جهت پیدا نمودن نتایج دقیق‌تر از نتایج به دست آمده‌ی قبلی می‌باشد. وجه تمایز این پژوهش این است که بعد از تشخیص و حذف داده‌ی پرت با استفاده از ماتریس همبستگی و ماژولاریتی با استفاده از ترکیب چند الگوریتم دقت بهبود یافته است.



شکل ۱. فرآیند کلی روش CNRM

بلوک دیاگرام کلی روش پیشنهادی در شکل (۲) نشان داده شده است. داده در شکل (۲) گراف مورد بررسی خواهد بود که از ارتباطات موجود در شبکه‌ی اجتماعی مورد بررسی ایجاد شده است. مجموعه‌ی این ارتباطات در نهایت یک گراف ساختاری می‌باشد که از آن یک ماتریس همبستگی ایجاد می‌گردد که در صورت وجود ارتباط بین گره‌ها مقدار ۱ و در غیر این صورت مقدار ۰ منظور می‌گردد، و نیز کلیه روش‌ها قابل تعمیم بر روی گراف‌های جهت‌دار و وزن‌دار می‌باشد.



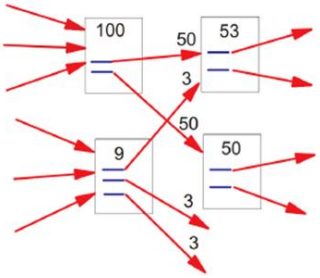
شکل ۲. فرآیند تشخیص جامعه برای هر الگوریتم

با اعمال الگوریتم‌های مختلف از قبیل ماژولاریتی و حذف داده‌های پرت، ماتریس همبستگی و گرفتن نتایج متفاوت و در بعضی موارد یکسان می‌توان از روش پیشنهادی استفاده کرد. بدین صورت که اگر گره‌ای در چند نتیجه مختلف جزئی از یک اجتماع باشد احتمال قرارگیری آن گره در آن اجتماع بیشتر خواهد بود و چنانچه گره‌ای در نتایج مختلف جزئی اجتماع‌های مختلف در آید، آن گره به عنوان داده یا گره پرت لحاظ شده و حذف می‌گردد. در نهایت با ترکیب نتایج اولیه یک ماتریس همبستگی دیگری ایجاد خواهد شد که یک گراف وزن دار از نتایج این تحقیق می‌باشد. همچنین توصیف و

تشریح الگوریتم‌ها در ادامه ارائه شده است.

۳-۱- الگوریتم مرکزیت Centrality

هرچه تعداد یال‌های ورودی یک رأس بیشتر باشد مرکزیت آن رأس بیشتر است. از طرفی هرچه رأس‌هایی که به یک رأس یال خروج دارند، مرکزیت بالاتری داشته باشند مرکزیت آن رأس هم بیشتر خواهد بود. این معیار به‌منظور بهبود نتایج موتورهای جستجو ارائه شد تا موتورهای جستجو بتوانند علاوه بر ارزیابی محتوای صفحات به‌منظور رتبه‌بندی آن‌ها، از جایگاه آن صفحه در میان صفحات وب نیز استفاده نند و برای صفحات مهم و معتبر، اولویت بالاتری قائل شوند. روش محاسبه پیشنهادی اول در شکل (۳) نمایش داده شده است. در هر مرحله هر رأس امتیاز خود را بین رأس‌های که به آن‌ها یال خروج دارد تقسیم می‌کند. درصورتی‌که یال‌ها وزن‌دار باشند این تقسیم امتیاز برحسب وزن یال‌ها صورت می‌گیرد. این عملیات آنقدر تکرار می‌شود تا امتیاز رأس‌ها همگرا شده و تقریباً ثابت شوند. در این روش به‌منظور اطمینان از همگرا شدن وزن‌ها، هر رأس بخش کوچکی از وزن خود را بین تمام رأس‌ها تقسیم می‌کند و مابقی آن را به رأس‌های که به آن‌ها یال خروج دارد می‌دهد. از طرف دیگر این نوع تقسیم وزن‌ها باعث می‌شود درصورتی‌که در گراف چاهک وزن وجود داشته باشد آن بخش وزن غیرعادی بالا نگیرد و وزن بقیه‌ی گراف را به‌طور غیرطبیعی کوچک نکند.



شکل ۳. تقسیم وزن‌ها.

به‌طورکلی روش اول یک روش بالا به پایین است که مبتنی بر افراز است. این روش ابتدا تمام گراف را به عنوان یک خوشه در نظر می‌گیرد و سپس برای اینکه مشخص شود که راس‌ها به همدیگر نزدیک هستند یا نه از گام زنی تصادفی استفاده می‌شود؛ به این صورت که از هر راس با یک احتمال یکنواخت به راس دیگری متصل به آن راس می‌رود و سپس برای مشخص کردن اینکه راس دیگری برای اضافه شدن به خوشه جاری مناسب است یا خیر از الگوریتم رتبه‌بندی بهره گرفته می‌شود (این تکنیک به این صورت عمل می‌کند که گره‌های دارای رتبه پایین‌تر به عنوان گره‌ی بالقوه مخرب رتبه‌بندی می‌شوند)، سپس هر خوشه به دو بخش افراز می‌شود تا جایی که افراز کردن بیشتر نتایج خوشه‌بندی را بدتر کند روش افراز بندی بر اساس پیمانگی بهینه‌سازی شده عمل می‌کند. این روش ابتدا کل گراف را به‌عنوان یک خوشه در نظر گرفته و سپس پیمانگی بهینه سازی شده نقطه شکست را نشان می‌دهد که برای خوشه بعد یک رأس دلخواه را انتخاب کرده و آن را به عنوان یک خوشه در نظر می‌گیرد سپس به‌صورت تکراری آن خوشه را بسط داده و این کار را تا در نظر گرفتن کل گراف به‌عنوان یک خوشه ادامه می¬دهد بهترین خوشه که طی تکرارها به وجود می‌آید به همراه خوشه‌ی شامل بقیه‌ی گراف، دو افراز این مرحله را تشکیل می‌دهد معیاری که در این روش برای تشخیص میزان خوب بودن یک افراز بندی درنظرگرفته می‌شود، پیمانگی بهینه‌سازی شده است که حتی دورترین رأس‌ها را نیز دربر می‌گیرد و محدودیت وضوح ندارد به‌طوری‌که خوشه‌های حاصل از دید کاربر بیرونی منطقی به نظر می‌رسد. شبه کد الگوریتم مرکزیت مطابق شبه کد (۲) می‌باشد.

شبه کد ۲. شبه کد الگوریتم مرکزیت

```
Algorithm: centrality
Require: ppr_u(v) = PPR score of v from u and G
                  = {V. E}
1: u ← a randomly chosen vetrex
2: S ← {u}
3: T ← V − S
4: mod = 0
5: mod_max ← 0
6: G ← WalkRandomly
5: while T ≠ 0 do
6:     GR → walkrandomly (G, V, S, T)
7:  mod ← mod + optimal modularity
8:     mod ← mod + GR
9:     if mod > mod_max then
10:        mod_max ← mod
11:        S_max ← S
12:        T_max ← T
13:    end if
14:    v ← FindBestVertex(G, S, T)
15:    S ← S ∪ {v}
16:    T ← T − {v}
17: end while
18: if mod_max > 0 then
19:    return (S_max. T_max. mod_max)
20: else
21:    return NIL
```

تغییر پیمانگی حاصل از انتقال یک رأس دلخواه v از خوشه T به خوشه S که از رابطه (۱-۳) به دست می‌آید:

$$\Delta Q = \frac{1}{m}\left(\sum \text{Avj} - \text{kv}\frac{\sum \text{j2S kj}}{2m}\right) - \frac{1}{m}\left(\sum \text{Avj} - \text{kv}\frac{\sum \text{j}\epsilon \text{T}' \text{kj}}{2m}\right) \quad (۱)$$

مقدار متوسط Q عددی بین ۰ و ۱ است که هرچقدر این عدد بزرگ تر باشد تشخیص جامعه بهتر صورت گرفته است؛ همچنین $\sum Avj$ مجموع انتقال رئوس دلخواه و kv مقدار ثابت تغییر پیمانگی حاصل از یک رأس دلخواه می‌باشد، که در رابطه (۱) $T' = T - \{v\}$ درصورتی‌که $\sum j\epsilon S\ kj$ و $\sum j\epsilon Tkj$ را ذخیره کرده و در هر مرحله با هزینه‌ی زمانی O(1) آن‌ها را به‌روزرسانی می‌کند. در گرافی با n رأس و m یال به‌طور متوسط با هزینه‌ی زمانی $O(\frac{2m}{n})$ می‌توان تغییر پیمانگی حاصل از انتقال یک رأس را محاسبه کرد.

۳-۲- الگوریتم NDOCD

در این مبحث یک روش برای تشخیص همپوشانی جوامع با استفاده از تجزیه شبکه و از جنبه پارتیشن‌بندی گره‌های تناوبی و اتصالات یک شبکه پرداخته می‌شود. این الگوریتم Network Decomposition Overlapping Community Detection و به اختصار NDOCD نامیده می‌شود. از آنجا که تشخیص همپوشانی جوامع هنوز به عنوان یک چالش بزرگ شناخته می‌شود، لذا در این زمینه الگوریتم‌های زیادی ارائه‌شده است. NDOCD تکنیک خوشه‌بندی گره‌ها را جهت شناسایی اتصالات جوامع بکار گرفته است و به‌طور تکراری همه اتصالات تعیین‌کننده جوامع را حذف و شبکه را به یک شبکه با مولفه‌های کوچک‌تر تقسیم می‌کند. تجزیه شبکه و بهینه‌سازی تکنیک خوشه‌بندی گره به‌طور مشترک برای ساخت یک الگوریتم کارا و باصرفه از لحاظ زمانی در الگوریتم پیشنهادی بکار می‌رود. NDOCD تکنیک بهینه‌سازی خوشه‌بندی برای کشف لینک اجتماعات می‌باشد، بنابراین از مصرف انرژی اضافی همانند رویکردهای قبلی خصوصا برای شبکه‌های با چگالی بالا جلوگیری کرده است. دو مرحله اصلی دیگر در زیر شرح داده‌شده است:

فرض می‌شود I(A, B) یالی بین گره‌های A و B است و این گره‌ها نقاط انتهایی این یال نامیده می‌شوند. همچنین فرض می‌شود Γ(A,B) مجموعه همسایه‌های گره A می‌باشد. درنتیجه [۱۷]:

$$\text{Jaccard(A, B)} = \frac{\Gamma(A) \cap \Gamma(B)}{(A) \cup \Gamma(B)} \qquad (۲)$$

تنها اشتراک بین همسایه‌های بدون فاصله گره‌های A و B درنظرگرفته می‌شود؛ یعنی گره‌هایی که دارای فاصله‌ای به طول یک از گره‌های A و B هستند.

رابطه (۳) تعریف ضریب داخلی بودن است. در این رابطه اگر Jaccard(A,B) از β بزرگ‌تر باشد، آنگاه مقدار ضریب داخلی بودن یا L2I(I(A, B))=1 در نظر گرفته می‌شود. دلیل معرفی این شرط این است که اگر β درصد از همسایگان A و B مشترک باشند، آنگاه احتمال زیادی وجود دارد که هر دو گره A و B متعلق به یک جامعه باشند. از طریق انجام آزمایش‌های متعدد مشاهده شد بازه مقادیر مناسب برای β برابر [۰.۵ - ۰.۲] است. به‌صورت شهودی مشخص است که مقادیر نزدیک به ۰.۵ می‌تواند منجر به نتایج دقیق‌تر شود؛ اما دفعات متعدد آزمایش نشان داد که مقادیر پایین مانند ۰.۲ برای β کافی است و می‌تواند منجر به نتایجی با دقت قابل‌توجه در تشخیص جوامع شود.

همچنین معرفی این آستانه، باعث می‌شود که زمان لازم برای ضریب Jaccard(I(A, B)) کاهش یابد در غیر این صورت ضریب داخلی بودن یا L2I برابر با Jaccard (I(A, B)) در نظر گرفته می‌شود.

$$\text{L2I (I (A, B))} = 1; \begin{cases} Jaccard\ (A,B) > \beta \\ Jaccard\ (A,B);\ otherwise \end{cases} \qquad (۳)$$

بعد از انتخاب اولیه برچسب‌ها، برخی از گره‌های گراف با بالاترین درجه به عنوان نقاط اولیه برای جستجوی گره‌ها انتخاب می‌شود. سپس به هرکدام از این نقاط، تعدادی عامل نسبت داده می‌شود، که تعداد این عامل‌های نسبت داده‌شده با درجه این گره‌ها برابر خواهد بود. تعداد گره‌هایی که به عنوان نقاط اولیه انتخاب می‌شوند باید به‌گونه‌ای باشد که تعداد حداقلی از عامل‌های جستجو کننده را برای گراف ورودی تضمین کند. فرض کنید H مجموعه گره‌هایی باشند که به عنوان نقاط اولیه در جستجو انتخاب‌شده‌اند، آنگاه نسبت بین تعداد عامل‌های جستجو کننده و تعداد یال‌های گراف به‌صورت رابطه (۴) خواهد بود:

$$\text{p} = \frac{\sum i\epsilon H\ degree(i)}{total\_number\_of\_links} \qquad (۴)$$

نکته‌ای که اهمیت دارد این است که نقاط اولیه به‌صورت یکنواخت از تمام نواحی گراف انتخاب شوند. با انتساب عامل‌ها به گره‌هایی که دارای درجه بالایی هستند، یک تمایل اولیه به سمت این گره‌ها به وجود می‌آید. عامل‌هایی که به یک نقطه اولیه خاص نسبت داده‌شده‌اند، برچسب متناظر با آن گره را در نزدیکی آن پخش خواهند کرد.

این یک ویژگی مفید از الگوریتم خواهد بود، به این دلیل که جوامع ساختارهایی محلی هستند و احتمال بالایی وجود دارد که گره‌هایی که در نزدیکی یک نقطه اولیه قرار دارند، متعلق به جامعه‌ی همان نقطه اولیه باشند. بعد از قرار دادن عامل‌ها بر روی نقاط اولیه، بررسی‌های محلی عامل‌ها شروع خواهد شد. همگی عامل‌ها برچسبی را از حافظه نقطه‌ی اولیه‌ای که روی آن قرار دارند، انتخاب می‌کنند و سپس به‌صورت تصادفی یکی از یال‌های مربوط به این نقطه اولیه را انتخاب می‌کنند و سپس یال انتخاب‌شده را پیمایش کرده تا به گره‌ی جدیدی برسند.

در نهایت، عامل‌ها برچسب را به گره‌ی جدید می‌دهند. انتخاب یک برچسب از حافظه متناظر با یک گره با استفاده از یک قاعده احتمالی ساده صورت می‌گیرد. برای هر برچسبی در حافظه متناظر با یک گره، یک احتمال انتخاب وجود دارد. احتمال انتخاب یک برچسب برابر است با بسامد دریافت آن برچسب توسط گره تقسیم‌بر جمع کل بسامدهای دریافت تمامی برچسب‌هایی که گره موردبررسی، دریافت کرده است. اگر f(I) نشان‌دهنده‌ی بسامد دریافت برای برچسب I باشد، آنگاه احتمال انتخاب یک برچسب برابر خواهد بود با رابطه (۵):

$$S(I) = \frac{f(I)}{\sum K \in RLv\, f(K)} \qquad (۵)$$

در این رابطه RLv مجموعه‌ی تمامی برچسب‌هایی می‌باشد که گره‌ی v دریافت کرده است. عامل‌ها بر اساس احتمال متناظر با انتخاب برچسب‌ها، تصمیم می‌گیرند که چه برچسبی را به گره بعدی بدهند. هرزمانی که یک عامل به یک گره جدید رسید، برچسبی که این عامل از حافظه‌ی گره قبلی انتخاب کرده است را به گره‌ی جدید می‌دهد. گره قبلی گره‌ای است که عامل کنونی، آن را در گام قبلی از جستجوهایش ملاقات کرده است شبه کد این الگوریتم مطابق شبه کد (۳) است.

۳-۳ الگوریتم ارتباط Relation

هدف از الگوریتم ارتباط، برقراری ارتباط بین یک سری از ورودی‌ها و خروجی‌ها و یا ارتباط دادن یک سری از متغیرها با سیگنال اندازه‌گیری شده آن‌ها می‌باشد. در آموزش شبکه یک ورودی وارد شبکه شده و مقدار وزن برای هر نورون محاسبه می‌شود سپس این مقدار در تابع انتقال نورون قرار گرفته و خروجی نورون محاسبه می‌شود. خروجی هر نورون در هر لایه به عنوان ورودی برای نورون‌های موجود در لایه‌های پایین دست محسوب می‌شود.

عملیات ذکر شده در نورون‌های پایین دست به همان صورت انجام می‌گیرد و در نهایت خروجی لایه آخر که خروجی شبکه می‌باشد محاسبه می‌شود. سپس این خروجی با مقدار مطلوب که همان سیگنال اندازه‌گیری شده می‌باشد مقایسه شده و اختلاف این دو به عنوان خطا تلقی می‌شود، این مقدار خطا به لایه‌های بالادست انتشار داده شده و بر اساس آن وزن‌های نورون‌ها در لایه‌های بالادست تصحیح می‌گردد. عمل تصحیح تا اولین لایه مخفی ادامه می‌یابد. سپس با وزن‌های جدید شبکه دوباره خروجی می‌دهد و دوباره این خروجی با مقدار مطلوب مقایسه شده و خطای آن محاسبه می‌شود، اگر خطا به حدی که مدنظر ما است نرسیده باشد دوباره عمل تصحیح وزن‌ها انجام می‌گیرد (هر کدام از این چرخه‌ها یک اپوک نامیده می‌شود) و این مراحل تکرار می‌شود تا خطا به مقدار مدنظر برسد. شبه کد این الگوریتم در شبه کد (۴) آمده است.

شبه کد ۳. شبه کد الگوریتم NDOCD.

Algorithm NDOCD

$require: ppr_u^{\rightarrow}(v) = ppr\ score\ of\ v\ from\ u\ and\ G = \{V, E\}$
$u \rightarrow a\ randomly\ chosen\ vertex$
$GR \rightarrow walkrandomly\ (G, V, S, T)$
$MOD \rightarrow mod + optinal\ modularity$
$find\ the\ similarities\ of\ each\ node\ i\ and\ j\ using\ eigenvectors\ with\ Eq\ (4)$
$compute\ the\ most\ similar\ node\ i\ using\ j\ Eq\ (5)$
if $mod > mod_{max}$ **then**
  $mod \rightarrow mod_{max}$
  $S \rightarrow S_{max}$
  $T \rightarrow T_{max}$
end if
$v \rightarrow find\ best\ vertex\ (G, S, T)$
$S \rightarrow S \cup \{v\}$
$T \rightarrow T - \{v\}$
if $mod_{max} > 0$ **then**
  $return\ (S_{max}, T_{max}, mod_{max})$
else
  $return\ NIL$

شبه کد ۴. شبه کد الگوریتم Relation

Algorithm Relation

$wait\ for\ any\ message\ from\ master\ node$
$id, data(id) = decode(message)$
if $(id == node_{id})$
  $send\ message(node_{id}, data(node_{id}))$
end if
for $id == n$
  $graph\ message\ (id, data(id))$
end for
while $(poll\ 5ms\ for\ requested\ message)$ **do**
  $id, data(id) = decode(message)$
end while

۳-۴- الگوریتم میانگین Mean

در این الگوریتم از روش K-Mean استفاده شده است. الگوریتم خوشه‌بندی K-Mean یک الگوریتم خوشه‌بندی بر مبنای فاصله متداول است که از فاصله به عنوان شاخص شباهت استفاده می‌کند، بدین معنی که هر چه دو شیء به یکدیگر نزدیکتر باشند، شباهت بیشتری دارند. قواعد الگوریتم خوشه‌بندی K-Mean ساده بوده، به سهولت قابل پیاده‌سازی است و سرعت و کارآمدی خوبی دارد. مهم‌ترین عیب آن این است که مقدار K به عنوان مرکز خوشه اولیه باید ابتدا معین باشد. بنابراین می‌توان از مقادیر بیشینه، کمینه و مابین به عنوان سه مرکز خوشه اولیه در الگوریتم استفاده کند. شبه کد این الگوریتم مطابق شبه کد (۵) می‌باشد.

مراحل الگوریتم به قرار زیر است:

۱. در ورودی داده، مقادیر بیشینه، کمینه و مابین به عنوان سه مرکز خوشه اولیه c_1، c_2، c_3 انتخاب می‌شوند.

۲. برای سایر اشیای داده در مجموعه داده، رابطه (۶) برای محاسبه فاصله اقلیدسی تا مراکز خوشه‌ها استفاده می‌شود. بر طبق قاعده نزدیکترین همسایه، داده در نزدیکترین خوشه قرار داده می‌شود، در اینجا x_j بیانگر j امین نقاط داده، c_i i امین مرکز خوشه هستند.

$$d_{ij} = |x_j - c_i|, (i = 1,2,3) \qquad (۶)$$

۳. روابط (۸) و (۷) برای محاسبه دوباره مراکز خوشه‌ها استفاده می‌شوند، که در آن d_i بیانگر مقدار Mean برای تمام نقاط داده در خوشه i، n بیانگر تعداد نقاط داده در خوشه i و x_j نقاط داده در خوشه i هستند. مرکز جدید خوشه x_j است به نحوی که x_j عبارت $\min\{f(x_j)\}$ را برآورده سازد.

$$d_i = \frac{\sum x_j}{n}, (x_j \in c_i,\ i = 1,2,3) \qquad (۷)$$

$$f(x_j) = |x_j - d_i|, \quad (i = 1,2,3) \qquad (۸)$$

۴. گام‌های(۲) و (۳) تکرار می‌شوند، تا زمانیکه E کمتر از یک مقدار آستانه داده شده‌شود و الگوریتم پایان می‌یابد. به صورت رابطه (۹).

$$E = \sum_{i=1}^{3} \sum_{x_j \in c_i} (x_j - c_i)^2, \quad (i = 2,3) \qquad (۹)$$

روش پیشنهادی مبتنی بر خوشه‌بندی می‌باشد. بدین صورت که با اعمال روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت یک روش تشخیص اجتماع جدید پیشنهاد و ارائه شده که در نهایت ترکیب این روش‌ها منجر به دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری خواهد شد. خوشه‌بندی یکی از مراحل اصلی داده کاوی است که در جهت کاوش الگوهای پنهان مورد استفاده قرار می‌گیرد [۴۰].

روش‌های خوشه‌بندی عادی بدین صورت عمل می‌کند که با تعریف یک الگوریتم سعی در برطرف نمودن مت خوشه‌بندی دارند. درصورتی که اعمال هر یک از این الگوریتم‌ها بر روی داده‌های یکسان دارای نتایج گوناگونی خواهد بود. با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های پایه‌ای آن، در این تحقیق از روش خوشه‌بندی ترکیبی به اختصار با نام CNRM استفاده شده است. در نتیجه این روش پیشنهادی هر الگوریتم بطور جداگانه به تشخیص جامعه و حذف داده های پرت با استفاده از ماکزیمم کردن ماژولاریتی و ماتریس همبستگی می‌پردازد. در نهایت با یک اجماع به یک جامعه با دقت بالاتر خواهد رسید

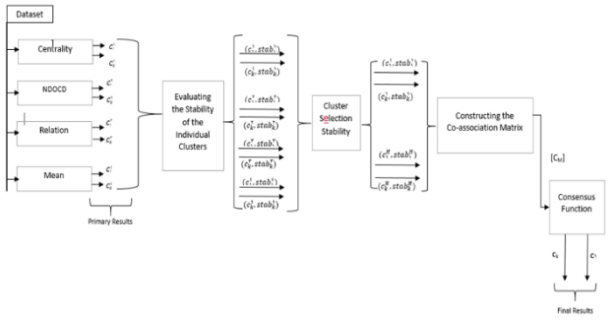
شبه کد ۵. شبه کد الگوریتم Mean.

Algorithm K-means

$require: d_i , f(x_j)$ & E
Output: Hierarchically clustered communities
1: Find the eigenvectors of points
2:Find the similarities of each node i and j using eigenvectors with Eq (3-6)
3: Compute the most similar node i using Eq. 7 and 8
4: Agglomerate the node i and j if the two nodes chose each other as the most similar node
5: Re-initialize the graph with the agglomerated nodes and start the next iteration
6: Agglomerate the nodes into hierarchical clusters when the iteration is finished

.

همچنین فرآیند ترکیب چهار الگوریتم به شرح زیر می‌باشد:در این روش خوشه‌های پایدارتر از خوشه‌های اولیه شناسایی میشوند و ماتریس همبستگی نهایی تنها از این خوشه‌های پایدار تشکیل میشود. یک خوشه پایدار، خوشه‌ای است که اگر آن روش خوشه‌بندی چندبار دیگر هم، رو آن مجموعه داده‌ها با همان الگوریتم اجرا شود باز هم همان خوشه دیده خواهد شد. در این روش ابتدا با نمونه برداری مجموعه داده‌های بدون نویز و داده پرت ایجاد میشود، حال برای بررسی شباهت میان خوشه‌ها از طریق نمونه برداری بررسی میشود. سپس میانگین این معیارهای شباهت بعنوان میزان پایداری این خوشه برگردانده می‌شود.



شکل ۴. دیاگرام روش ترکیب الگوریتم‌های خوشه‌بندی

در مرحله ترکیب نتایج اولیه الگوریتم‌ها، ابتدا خوشه‌های انتخاب شده، ماتریس همبستگی را تشکیل میدهند. نتایج m خوشه‌بندی روی داده‌های نمونه برداری شده در ماتریس همبستگی n*n ذخیره میشوند. برای ترکیب با استفاده از ماتریس همبستگی باید معیاری را برای نشان دادن همبستگی نمونه‌ها تعریف کرد که بتواند شباهت بین نمونه‌ها را با وجود تنها زیر مجموعه‌ای از خوشه‌های اولیه به درستی استخراج و محاسبه کرد. پس از ساخت ماتریس همبستگی، میتوان خوشه‌های نهایی را استخراج نمود. همچنین موارد زیر باید برای ترکیب الگوریتم‌ها و خوشه‌بندی رعایت میشود:

- باید تابع f_j که توسط الگوریتم‌های خوشه‌بندی خاص A_j بهینه میشود، ارتباط منطقی داشته باشد. به عبارت دیگر، مقدار بیشتر برای $g(C_j,D)$ به این معنی باشد که خوشه C_j نسبت به تابع f_j و به تناظر نسبت به الگوریتم خوشه‌بندی خاص A_j بهینه‌تر است.
- باید نسبت به توابع خوشه‌بندی مختلف قابل مقایسه باشد. به عبارت دیگر اگر $g_j(C_j,D) > g_i(C_i,D)$ آنگاه باید نتیجه گرفت که کیفیت خوشه C_i باتوجه به تابع f_j از تابع f_i بهتر است. یعنی کیفیت خوشه C_i در خوشه‌بندی j-ام ازکیفیت خوشه C_j در خوشه‌بندی i-ام بهتر است.
- مقدار تابع f خوشه باید نسبت به خوشه‌های مختلف قابل مقایسه باشد. به عبارت دیگر، $g_j(C_j,D) = g_i(C_i,D)$ باید خروجی بدهد که خوشه‌های C_i C_j نسبت به تابع f برابری دارند. یعنی این دو خوشه در خوشه‌بندی j-ام از کیفیت، شباهت و بهینه بودن وضعیت برابری دارند.

یک معیار شباهت بین خوشه C_i و خوشه‌بندی اولیه (P(D)) با پارامتر $sim(C_i,P(D))$ نشان داده میشود. با استفاده از این معیار، شباهت آن خوشه را با خوشه‌بندی‌های مختلف حاصل از نمونه برداری محاسبه می‌شود. سپس میانگین این معیارهای شباهت، بعنوان میزان پایداری این خوشه $g_j(C_j,D)$ برگردانده میشود. $sim(Ci,P(D))$ میزان اعتبار خوشه C_i را در خوشه‌بندی P روی مجموعه داده D مشخص می‌کند.

برای محاسبه $sim(C_i,P(D))$ بصورت زیر عمل می‌شود:

ابتدا تمام نمونه‌های دیگر متعلق به مجموعه داده D که در خوشه Ci قرار ندارند، بصورت یک خوشه مستقل D/C_i نمایش داده میشود. حال یک خوشه‌بندی شامل دو خوشه C_i و D/C_i ایجاد شده است که آن را P1 می‌نامیم $P1=\{C_i,D/C_i\}$. اکنون خوشه‌بندی P(D) که روی داده اعمال شده است. همه خوشه‌ها در P(D) به دو خوشه C* و D/C* تقسیم می‌شوند. خوشه C* از اجتماع همه خوشه‌هایی که بیش از ۶۰٪ از نمونه‌هایشان در خوشه C_i وجود دارند، تشکیل میشود و مابقی خوشه‌ها نیز در خوشه D/C* قرار می‌گیرند. این خوشه را P_n مینامیم. حال از اطلاعات هنجارسازی شده که معیار متداول برای ارزیابی شباهت بین چندین افراز است، برای اندازه‌گیری شباهت بین چندین خوشه‌بندی استفاده میشود. هر داده ورودی از ماتریس همبستگی در این روش بصورت $C(i,j) = \{n(i,j) / max(n(i,j))\}$ تعریف میشود. n_i تعداد دفعاتی است که نمونه i در خوشه‌های انتخاب شده ظاهر شده است. بطور مشابه n_j نیز تعداد دفعاتی است که نمونه j در خوشه‌های انتخاب شده ظاهر شده است. همچنین n(i,j) تعداد دفعاتی است که جفت نمونه‌های i و j باهم در یک خوشه از خوشه‌های انتخاب شده ظاهر شده‌اند. بدیهی است با درنظر گرفتن تعداد خوشه‌های ثابت در خوشه‌بندی‌های اولیه همواره n(i,j) کمتر از تعداد کل افرازهای اولیه و همچنین، تعداد کل خوشه‌های ممکن میباشد. همچنین پس از این که ماتریس همبستگی با روش ذکر شده ساخته شد، از اجرا و ترکیب همزمان چهار الگوریتم برای استخراج خوشه‌های نهایی از این ماتریس استفاده میشود.

از روی این ماتریس یک خوشه ترسیم میشود که در واقع این خوشه بعنوان خوشه نهایی بعنوان خروجی از خوشه‌های چهار الگوریتم ترکیب میشود. این خوشه نهایی را با استفاده از الگوریتم ولگشت [۷] (random walk) جهت تشخیص جوامع سیبیل و غیر سیبیل مورد استفاده قرار میگیرد. این پیاده روی‌ها-مسیرها با توجه به ارزش اعتماد هر گره اصلاح میشوند. براساس این اطلاعات، الگوریتم تشخیص سیبیل کار میکند و در مورد اعتبار گره مشکوک نسبت به گره تأییدکننده تصمیم گیری میکند. در این روش مطالب منتشر شده توسط گره توسط طبقه‌بندی یادگیری ماشین در ماژول نرم افزار نصب شده در گره جمع آوری و تحلیل میشود. سپس خروجی (مقدار اطمینان) خود را میدهد که به صورت محلی ذخیره میشود و توسط الگوریتم تشخیص سیبیل قابل دسترسی است. برای بررسی گره مشکوک با توجه به گره تأیید کننده، الگوریتم تشخیص سیبیل از اطلاعات تصادفی مسیر اصلاح شده گره تأییدکننده و مشکوک استفاده می‌کند.

تشخیص اجتماعات، تقسیم بندی‌های موجود در شبکه را نشان می‌دهد و گروه‌های آن را از هم مجزا می‌کند. تشخیص اجتماعات کمک می‌نماید تا دید بهتری نسبت به ساختار شبکه پیدا کرد. در روش پیشنهادی که برای تشخیص سیبیل در شبکه‌های اجتماعی با استفاده روش از یادگیری ماشین بر روی گراف ساختاری انجام می‌شود، بر خلاف سایر رویکردهای مبتنی بر یادگیری ماشین، بر روی ترکیبی از داده‌های بدون برچسب و برچسب گذاری شده کار می‌شود تا در شناسایی حملات هدفمندتر عمل کند؛ زیرا سطوح مختلف ویژگی‌های پروفایل‌های کاربر را شناسایی می‌کند، که همین امر منجر به افزایش دقت بالا در شناسایی سیبیل می‌شود. بنابراین شناسایی این نوع حمله در شبکه‌های اجتماعی می‌تواند بخشی از امنیت و اعتماد این بستر را میان کاربران افزایش دهد. بدین صورت که، ویژگی استخراج شده کاربران به صورت گراف در شبکه‌های اجتماعی تبدیل می‌شود، سپس با استفاده از تکنولوژی یادگیری ماشین نیمه نظارتی که هم داده‌های بدون برچسب و هم داده‌های برچسب دار را مورد پردازش قرار می‌دهد، گراف ساختاری را پردازش

می‌کند. آنگاه با استفاده از الگوی بدست آمده از کاربران، می‌توان جوامع سیبیل و غیر سیبیل را مورد بررسی قرار داد تا رفتاران کاربران دسته بندی شود، سپس جوامع تفکیک می‌شوند و کاربران سیبیل تشخیص داده می‌شوند. از این روش برای از هرزنامه ها، جلوگیری از سرقت اطلاعات افراد، کاهش آسیب‌های اجتماعی و روانی کاربران، بالابردن محبوبیت شبکه اجتماعی، افزایش اعتماد و اطمینان کاربران به شبکه اجتماعی و روابط سالم میان کاربران برای کاربرد روش پیشنهادی یاد کرد. زیرا با بالابردن دقت تشیخص حملات سیبیل در شبکه‌های اجتماعی موارد ذکر شده محقق خواهد شد. روش CNRM بدین صورت است که با اعمال روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت یک روش تشخیص اجتماع جدید پیشنهاد و ارائه شده که در نهایت ترکیب این روش‌ها منجر به دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری خواهد شد. درصورتی که اعمال هر یک از این الگوریتم‌ها بر روی داده‌های یکسان دارای نتایج گوناگونی خواهد بود. با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های پایه‌ای آن، در این تحقیق از روش خوشه‌بندی ترکیبی CNRM استفاده شده است.

۴- معیار ارزیابی و نتایج

در این بخش با استفاده از مجموعه داده‌هایی که در ادامه معرفی شده‌اند و انجام آزمایشات و نتایج شبیه‌سازی روش پیشنهادی با استفاده از نرم افزار متلب ارائه شده است؛ همچنین در ادامه این بخش نتایج این آزمایشات تشریح و تعریف شده‌اند.

۴-۱- معرفی داده‌های استفاده شده

در این مقاله از چهار مجموعه داده توییتر، رِدیت، فیسبوک و اینستاگرام استفاده شده است که مشخصات آن‌ها شامل تعداد گره و تعداد یال در جدول (۲) ارائه شده است.

جدول ۲. مشخصه های داده‌های شبیه‌سازی.

نام	تعداد گره	تعداد یال	مجموعه داده
توییتر	۳۲	۴۶۰	۱
ردیت	۸۹۹	۳۳۷۲۰	۲
فیسبوک	۱۸۹۹	۵۹۸۳۵	۳
اینستاگرام	۱۴۸	۱۲۳۰	۴

استفاده از روش‌های مبتنی بر ساختار برای شناسایی گره‌های سیبیل کافی نیست. لازم است معیاری تعریف شود که بتواند به طور کمی مشروعیت یک کاربر را توصیف کند. این مسئله مطرح است که حمله سیبیل یک مشکل خوشه‌بندی و برچسب‌گذاری شده است. تکنیک‌های خوشه‌بندی سنتی به داده‌های برچسب‌گذاری شده تکیه می‌کنند. داده‌های نیمه‌نظارت‌شده از شباهت‌های بین داده‌های بدون برچسب استفاده می‌کنند که وقتی با داده‌های برچسب‌گذاری شده ترکیب می‌شوند، خوشه‌بندی بهتری را شکل می‌دهند.

با بهره‌گیری از الگوریتم جنگل تصادفی، دادگان به ۷۵٪ و ۲۵٪ مجموعه آموزشی و آزمون تقسیم شده است. جنگل تصادفی، به عنوان یک مدل یادگیری ماشین نظارت شده، یاد می‌گیرد که در فاز آموزش، داده‌ها را به خروجی‌ها نگاشت کند. در طول آموزش، داده‌های تاریخی به مدل داده می‌شوند.که طبق موارد زیر سطوح مختلف آن ارزش گذاری شده است.

ویژگی‌های استخراج شده در سه سطح پروفایل کاربر، شبکه تعاملی و سطح محتوا دسته بندی می‌شوند. هر سطح دارای ویژگی های مشخصی است که می‌تواند در زمان واقعی محاسبه شود و همچنین ویژگی‌های ضمنی که نیاز به محاسبه آفلاین دارند. سطح محتوا شامل ویژگی‌های پُست (مانند طول، تعداد پاسخ‌ها یا تعداد بازنوشته‌ها) است که ممکن است اهمیت آن پست را منعکس کند. سطح کاربر شامل ویژگی‌های صریح (مانند تعداد دنبال‌کنندگان، تعداد دوستان یا تعداد پست‌ها) و ویژگی‌های ضمنی (مانند تعداد پست‌های بازنوشته شده، نرخ پست هفتگی، یا پاسخ‌ها به پست‌ها) است. سطح سوم، سطح شبکه است که در آن ویژگی‌های استخراج‌شده، تعداد رویدادهای تعامل بین دو یا چند کاربر در گراف را در یک زمان خاص با توجه به یک رویداد تعریف می‌کنند. فرآیند استخراج ویژگی‌ها در این سطح شبیه به روش تجمیع بر اساس هر دو سطح محتوا و کاربر است.

۴-۲- معرفی معیارهای مورد استفاده

برای ارزیابی و مقایسه مدل‌های روش پیشنهاد شده، پارامترهای آماری مختلفی بکار گرفته شده‌اند، که هر یک ماهیت مستقلی از خطا را می‌دهند. ضریب تعیین (R2) معیاری برای تعیین همبستگی نسبی بین دو مجموعه متغیر می‌باشد. این پارامتر در مرحله آموزش مدل، نقش مهم‌تری نسبت به مرحله آزمون ایفا می‌کند. RMSE مشهورترین معیار خطا می‌باشد، که خطاهای بزرگ را بیشتر از خطاهای کوچک جذب می‌کند. ضریب پراکندگی (SI) معرف پراکندگی متغیر به صورت مطلق می‌باشد، در واقع میزان پراکندگی داده‌ها از خط بهینه می‌باشد، اگر تمام داده‌ها درست پیش بینی شوند میزان پراکندگی صفر می‌باشد. میانگین خطاهای انحرافی (Mean Bias Error) یا به اختصار MBE از دیگر پارامترهای آماری مهم می‌باشد، که هرچه به عدد صفر نزدیک باشد بهتر است، مقادیر مثبت برای MBE نشان از پیش‌بینی دست بالا و مقادیر منفی نشان

از مقادیر دست پایین می‌باشد. که در جدول (۳) پارامترای‌های ذکر شده لیست شده‌اند.

همچنین معیار F یک نوع میانگین بین پارامتر دقت و یادآوری است. دقت (p)، دقت سیستم در میان دادهای پیش‌بینی شده‌است.

جدول ۳. تعریف پارامترهای آماری.

پارامتر	رابطه
R2	$R^2=(1-\frac{\sum_{i=1}^{n}(M_i-P_i)^2}{\sum_{i=1}^{n}(P_i-\bar{P}_i)^2})$
RMSE	$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(P_i-M_i)^2}{n}}$
SI	$SI=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pi-Mi)^2}}{\bar{M}}$
MBE	$MBE=\frac{1}{n}\sum_{i=1}^{n}(P_i-M_i)$
F	$F_\beta=(1+\beta^2).\frac{precision.recall}{(\beta^2.precision)+recall}$

یادآوری (r) نسبت تعداد داده‌های پیش‌بینی شده، به تعداد کل داده‌های مورد انتظار برای پیش‌بینی است. ارزیابی عملکرد مدل‌های پیشنهاد شده توسط پارامترهای آماری برای هر مجموعه داده در جدول (۴) مشخص شده است.

جدول ۴. ارزیابی عملکرد مدل‌های پیشنهاد شده توسط پارامترهای آماری.

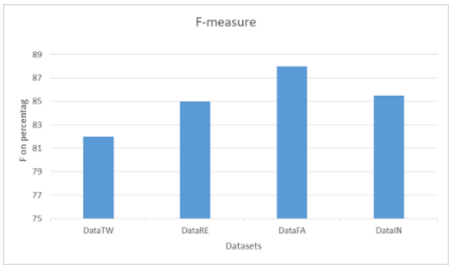
داده	R2	RMSE	Bias	SI%
DataTW	۰.۷۵۵	۱۱.۴۷۳	۰.۰۴۵	۹.۹۶۷
DataRE	۰.۹۰۵	۱۲.۱۹۰	۰.۰۳۳	۹.۳۵۱
DataFA	۰.۹۲۶	۱۲.۷۶۳	۰.۰۲۲	۹.۳۶۹
DataIN	۰.۹۲۵	۱۳.۴۸۱	۰.۰۱۶	۸.۸۷۱

۴-۳- معرفی سخت افزار و نرم افزار در آزمایشات

برای شبیه‌سازی روش پیشنهادی از نرم افزار متلب و از سیستمی با مشخصات CPU CORE i7 Intel، GPU AMD 2GB با سیستم عامل ویندوز نسخه ۱۰.۱ استفاده شده است.

۴-۴- معرفی انواع آزمایشات و هدف

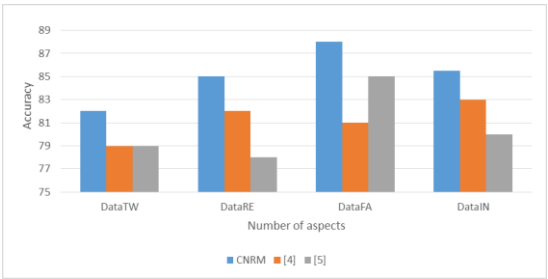
شکل (۵) نمودار معیار F روش مطرح شده را به تفکیک مجموعه داده‌ها نشان می‌دهد. معیار F یک نوع میانگین بین پارامتر p(صحت) و پارامتر r (بازخوانی) است. همانطور که در نمودار نشان داده شده است، برای هر چهار مجموعه داده‌ی مورد استفاده معیار F بالاتر از ۸۰٪ می‌باشد.



شکل ۵. معیار F روش پیشنهادی به تفکیک داده‌ها.

۴-۵- ارائه‌ی نتایج آزمایشات

در این بخش نتایج بدست آمده از مقایسه‌های صورت گرفته با روش مقالات [۴] و [۵] تجزیه و تحلیل شده‌اند. به‌طور کلی در این دو روش [۴] و [۵]؛ هر گره در شبکه به عنوان یک تأیید کننده و یک فرد مشکوک در نظر گرفته می‌شود، زیرا یک کاربر مخرب می‌تواند پیام‌های ارسالی توسط پروتکل را شنود کرده و بدست آورد. در چنین سیستم توزیع شده‌ای، به منظور شناسایی مدل حمله، از یادگیری ماشین به وسیله راه رفتن تصادفی بر روی گراف ساختاری شبکه اجتماعی استفاده می‌شود. پیمایش‌های تصادفی از جایگشت‌های تصادفی محاسبه شده ویژه ای حاصل می‌شود که در یافتن روابط خارجی بین مسیرها مهم هستند. اما دو روش ذکر شده در مقایسه با روش پیشنهادی این مقاله به‌نام خوشه‌بندی ترکیبی CNRM نقطه ضعف‌هایی را دارا می‌باشند که در ادامه مقاله به صورت ارائه نتایج آزمایشات مختلف به آن پرداخته خواهد شد. شکل (۶) مقایسه نتایج شبیه‌سازی این تحقیق با کارهای قبلی را نشان می‌دهد. در اینجا دقت همان دقت تشخیص حملات سیبل در شبکه‌های اجتماعی می‌باشد.

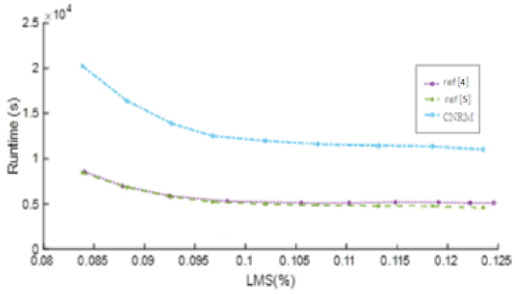


شکل ۶. مقایسه دقت این تحقیق با کارهای قبلی.

در شکل (۶) دقت روش CNRM با دو روش موجود بر روی چهار مجموعه داده مقایسه شده است.

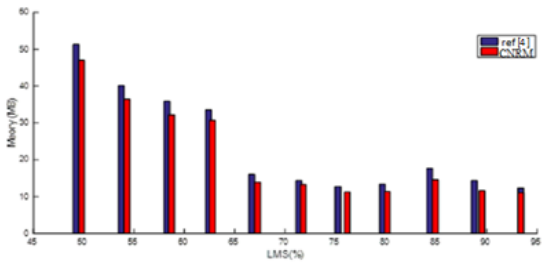
هر چهار الگوریتم پیشنهادی دارای زمان شبیه‌سازی در رنج دو کار قبلی ذکر شده هستند، ولی زمان شبیه‌سازی هر چهار الگوریتم و

رسیدن به اجماع تقریبا دو برابر یک الگوریتم واحد می‌باشد. دلیل آن این است که هر چهار الگوریتم بطور همزمان اجرا می‌شوند و با یک زمان یکسان به پایان می‌رسند و در نهایت زمانی هم صرف اجماع نتایج چهار الگوریتم می‌گردد. همچنین زمان اجرا برای دو روش [۴] و [۵] با الگوریتم پیشنهادی مقایسه می‌شود. شکل (۷) نمودار زمان اجرا را برای الگوریتم‌ها نشان می‌دهد. می‌توان دید که ابتدا امر زمان استفاده از روش پیشنهادی دو برابر سایر الگوریتم‌های ذکر شده می‌باشد اما به مرور زمان اجرا به شدت کاهش یافته است. در واقع با اینکه در روش پیشنهادی از چهار الگوریتم استفاده شده است اما زمان شبیه‌سازی تقریبا دو برابر بوده و چهار برابر نشده است. علت این امر این است که هر چهار الگوریتم بطور همزمان کار می‌کنند و اندکی همزمانی برای اجماع آن‌ها مورد نیاز می‌باشد. با وجود زمان شبیه‌سازی کمی بیشتر، اما دقت روش پیشنهادی بالاتر می‌باشد. منظور از Runtime زمان اجرای الگوریتم و LMS (Line Measure System) میزان پیشبرد الگوریتم در سیستم می‌باشد.



شکل ۷. زمان اجرای الگوریتم‌ها.

در شکل (۸) مصرف حافظه‌ی روش‌ها باهم مقایسه شده است. نتایج مصرف حافظه نشان می‌دهد که الگوریتم پیشنهادی مصرف حافظه-ی تقریبا مشابه الگوریتم مرجع [۴] دارد. در واقع مصرف حافظه به معنی میزان حافظه اصلی در حین اجرای الگوریتم می‌باشد و چون هر چهار الگوریتم موازی هم کار می‌کنند و همچنین قسمت اجماع بعد از اجرای هر چهار الگوریتم می‌باشد بنابراین مصرف حافظه با افزایش چهار الگوریتم تغییر زیادی نداشته است. همچنین با توجه به مرکزیت مابینی، ماکزیمم کردن ماژولاریتی، ماتریس همبستگی و حذف داده‌های پرت؛ در نتیجه روتین بودن الگوریتم‌ها به‌ویژه بعد از فرآیند یادگیری، میزان مصرف حافظه‌ی روش پیشنهادی کمی کمتر می‌باشد.



شکل ۸. مصرف حافظه.

در مقاله [۴] از الگوریتم تله سیبیل (SybilTrap) استفاده شده است که از یک تکنیک نیمه نظارت شده استفاده می‌کند که به طور خودکار ویژگی‌های اساسی فعالیت‌های کاربر را با ساختار اجتماعی در یک سیستم ادغام می‌کند. هدف این روش کار بر روی نمودارهای تعاملی و اجتماعی است. این در جایی که تعداد زیادی لبه حمله وجود دارد موثرتر است. در مقاله [۴] از خوشه‌بندی نیمه‌نظارتی که از داده‌های برچسب‌گذاری‌شده بسیار کمی برای شناسایی دقیق گره‌های سیبیل در مقیاس بزرگ استفاده می‌کنند، بهره گرفته شده است. بزرگترین چالشی که این روش با آن مواجه هست مقیاس پذیری است که تنها بر روی شبکه های اجتماعی بزرگ و اجتماعات عظیم مورد استفاده است.

همچنین در پژوهش [۵] از روش حد سیبیل (SybilLimit) استفاده شده است، که یک دفاع تقریباً بهینه در برابر حملات سیبیل از شبکه‌های اجتماعی است. اما در چندین مسئله دچار نقص است از جمله: استفاده از چندین نمونه مستقل از پروتکل مسیر تصادفی برای انجام بسیاری از مسیرهای تصادفی کوتاه و بهره برداری از تقاطع ها در لبه ها به جای گره‌ها؛ که باعث می‌شود مدت زمان اجرای طولانی را طی کند.

روش پیشنهادی CNRM در چهار مجموعه داده استفاده شده به ترتیب به دقت‌های ٪۸۲، ٪۸۵، ٪۸۸ و ٪۸۵.۵ رسیده است که به دقت کلی ٪۸۵.۱۳ رسیده است؛ درصورتی که دو روش مطرح شده تله سیبیل [۴] به دقت کلی ٪۸۱.۲۵ و حد سیبیل [۵] به دقت ٪۸۱ کلی رسیده بوده‌اند.

در روش پیشنهادی زمان اجرا افزایش می‌یابد. در واقع با اینکه در روش پیشنهادی از چهار الگوریتم استفاده شده است اما زمان شبیه‌سازی تقریبا دو برابر بوده است و چهار برابر نشده است. علت این امر این است که هر چهار الگوریتم بطور همزمان کار می‌کنند و اندکی همزمانی برای ترکیب آن‌ها مورد نیاز می‌باشد. با وجود زمان شبیه‌سازی کمی بیشتر، دقت روش پیشنهادی بسیار بالاتر می‌باشد.

جدول ۵. مقادیر ورودی و خروجی پارامترهای پژوهش.

الگوریتم	مجموع یال و گره گراف ورودی	دقت	زمان اجرا	حجم حافظه
مرکزیت	۲۴۵۵	۸۲٪	۰.۰۹۴۳	۴.۵
NDOCD	۲۴۵۵	۸۵%	۰.۰۱۱۰	۲.۶
ارتباط	۲۴۵۵	۸۸%	۰.۱۱۸۰	۹.۴۵
میانگین	۲۴۵۵	۸۵.۵٪	۰.۱۰۵۶	۷.۵۴
مجموع (CNRM)	۹۸۲۲۳	۸۵.۱۳٪	۱.۳۱ ثانیه	۲۴.۰۹ مگابایت

جدول ۶. پایداری و افراز خوشه‌ها.

	پایداری هر خوشه $Stability(C_i)$	افراز خوشه‌های مشابه $C(i,j)=\frac{n_{i,j}}{m_{i,j}}$
الگوریتم مرکزیت	۰.۸۰۱	۰.۸۷۷۵
الگوریتم NDOCD	۰.۵۶۹	۰.۸۵
الگوریتم ارتباط	۰.۷۴	۰.۸۹
الگوریتم میانگین	۰.۸۶۹	۰.۸۷۹

جدول (۵) مقایسه مقادیر دقت، زمان اجرا و حجم حافظه چهار الگوریتم را با CNRM نشان می دهد. در جدول (۶) پایداری و افراز خوشه‌ها در چهار الگوریتم مقایسه شده است.

۵- بحث و نتیجه‌گیری

روش پیشنهادی CNRM در چهار مجموعه داده استفاده شده به ترتیب به دقت‌های ٪۸۲، ٪۸۵، ٪۸۸ و ٪۸۵.۵ رسیده است که به دقت کلی ٪۸۵.۱۳ رسیده است؛ درصورتی که دو روش مطرح شده تله سیبیل به دقت کلی ٪۸۱.۲۵ و حد سیبیل به دقت٪۸۱ کلی رسیده بوده‌اند. روش این مقاله در میزان مصرف حافظه نسبت به روش مقاله [۴] مورد بررسی قرار گرفت، که چون هر چهار الگوریتم موازی با هم کار می‌کنند و همچنین قسمت اجماع بعد از اجرای هر چهار الگوریتم می‌باشد لازم به ذکر است که مصرف حافظه با افزایش چهار الگوریتم تغییر زیادی نداشته است و از این جهت یک جنبه خوب و مثبت محسوب می‌شود. همچنین در روش پیشنهادی زمان اجرا به شدت کاهش می‌یابد. در واقع با اینکه در روش پیشنهادی از چهار الگوریتم استفاده شده است اما زمان شبیه‌سازی تقریبا دو برابر بوده و چهار برابر نشده است. علت این امر این است که هر چهار الگوریتم بطور همزمان کار می‌کنند و اندکی همزمانی برای اجماع آن‌ها مورد نیاز می‌باشد. با وجود زمان شبیه‌سازی کمی بیشتر، دقت روش پیشنهادی بسیار بالاتر می‌باشد. اما از نظر مدت زمان اجرا نسبت به سایر روش‌های مطرح شده طولانی‌تر خواهد بود. تشخیص اجتماعات، تقسیم بندی‌های موجود در شبکه را نشان می‌دهد و گروه‌های آن را از هم مجزا می‌کند. تشخیص اجتماعات کمک می‌کند تا دید بهتری نسبت به ساختار شبکه پیدا کرد. در جدول (۷) مقدار دقت، مصرف حافظه و زمان اجرا با روش‌های [۴] و [۵] مقایسه شده است.

جدول ۷. خلاصه مقادیر نتیجه گیری شده.

معیار ارزیابی	روش CNRM	روش [۴]	روش [۵]
دقت	٪۸۵.۱۳	٪۸۱.۲۵	٪۸۱
مصرف حافظه	در زمان ۹۵ میلی ثانیه: ۲۴.۰۹ مگابایت	در زمان ۹۵ میلی ثانیه: ۲۰ مگابایت	در زمان ۹۵ میلی ثانیه: ۱۹.۸ مگابایت
زمان اجرا	در ٪۱۰۰ روندکار: ۱.۳۱ ثانیه	در ٪۱۰۰ روندکار: ۰.۶۵ ثانیه	در ٪۱۰۰ روندکار: ۰.۴۵ ثانیه

روش CNRM بدین صورت است که با اعمال روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت یک روش تشخیص اجتماع جدید پیشنهاد و ارائه شده که در نهایت ترکیب این روش‌ها منجر به دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری خواهد شد. روش‌های خوشه‌بندی عادی بدین صورت عمل می‌کند که با تعریف یک الگوریتم سعی در برطرف نمودن مشکل خوشه‌بندی دارند. درصورتی که اعمال هر یک از این الگوریتم‌ها بر روی داده‌های یکسان دارای نتایج گوناگونی خواهد بود. با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش‌های پایه‌ای آن، در این تحقیق از روش خوشه‌بندی ترکیبی CNRM استفاده شده است.

در ادامه‌ی روند این مقاله پیشنهادات زیر ارائه می‌گردد:

۱. ارائه مدلی برای شبیه‌سازی شبکه‌های اجتماعی، که در آن یال ها (افراد) در یک فضای اجتماعی دریک نقطه خاص جای گرفته‌اند. در این فضا، مکان یک فرد بازتاب دهنده ویژگی‌های اوست که بر اساس هرکدام از این ویژگی‌ها یک نقطه از یک بعد خاص را اشغال می‌کنند. در نهایت جایگاه فرد در فضا بازتاب دهنده ویژگی‌های او و البته تفاوت و شباهت او با دیگر افراد جامعه است. افراد با احتمالی که با بیشتر شدن فاصله کمتر می‌شود، یال‌ها (ارتباطات) را بینشان به اشتراک می‌گذارند. به طور کلی این مدل گرافی با ضریب خوشه بندی نسبتاً بزرگ، به اضافه همبستگی درجه به درجه با شرکت پذیری مثبت را از خود بروز می‌دهد. به‌علاوه برای یک بازه خاص از احتمال ارتباط، مدل گرافی دارای خاصیت انجمنی با انجمن‌های (تشابه به لحاظ ساختاری) خود متشابه را تولید می‌کند.

۲. استفاده از دو نوع جستجوی سراسری و محلی، جستجوی محلی از دنبال کردن راس‌هایی که در جستجوی سراسری انتخاب شدند، همچنین راس‌هایی را بر می‌گزینند که ماهیت

محلی جستجوی نوع دوم باعث می‌شود خاصیت‌های خوشه بندی، شرکت پذیری مثبت و انجمنی درگراف به وجود بیاید.

۳. اضافه کردن عنصر همانندی مبنایی بخصوص فاصله جغرافیایی را به صورتی به مدل پیشنهادی. با در نظر گرفتن فضای دو بعدی و توزیع پواسون راس‌ها در این فضا، سعی نمود مدلی را ارائه داد که علاوه بر دخیل کردن همانندی مبنایی، ویژگی های اصلی شبکه های اجتماعی را در گراف تولید شده توسط مدل، نهفته داشته باشد.

مراجع

[۱] سیدمحمد طباطبائی پارسا، حسن شاکری، رویکرد جدید برای تشخیص حملات سیبیل غیرمستقیم در شبکه‌های حس‌گر بی‌سیم مبتنی بر اعتماد آگاه از اطمینان با درنظرگرفتن عامل زمان، مجله افتا، ۱۳۹۶.

[۲] روح اله شاه مرادیان، حملات SYBIL و انواع روش های مقابله با آن، اولین کنفرانس ملی مهندسی کامپیوتر و فناوری اطلاعات، ۱۳۹۵.

[۳] محمد حجاریان، مروری بر انواع حملات سیبیل در شبکه های اجتماعی، کنفرانس ملی تحقیقات بین رشته ای در مهندسی کامپیوتر، برق، مکانیک و مکاترونیک، آذر ۱۳۹۸.

[4] Muhammad Al-Qurishi. and Sk Md Mizanur Rahman, SybilTrap: A graph-based semi-supervised Sybil defense scheme for online social networks, EERI Research Paper Series, 2019.

[5] Yu H, Gibbons PB, Kaminsky M, Xiao F. SybilLimit: A near-optimal social network defense against Sybil attacks. IEEE/ACM TRANSACTIONS ON NETWORKING. 2020.

[6] Xiong, F., Y. Liu, and J. Cheng, Modeling and predicting opinion formation with trust propagation in online social networks. Communications in Nonlinear Science and Numerical Simulation, 2017.

[7] Meshram, S.A. and D. Sable, A SURVEY ON SYBIL ATTACKS IN SOCIAL NETWORKS۲۰۱۸.

[8] Bansal, H. and M. Misra. Sybil detection in online social networks (osns). in 2016 IEEE 6th International Conference on Advanced Computing (IACC). 2016.

[9] Al-Qurishi, M., et al., Sybil defense techniques in online social networks: a survey. IEEE Access, ۲۰۱۷. ۵: p.۱۲۰۰-۱۲۱۹ .

[10] Liu, S., L. Zhang, and Z. Yan, Predict pairwise trust based on machine learning in online social networks: A survey. IEEE Access, 2018.

[11] Zhang, H., et al. Improving Sybil detection via graph pruning and regularization techniques. in Asian Conference on Machine Learning. 2016.

[12] Jiang, W., et al., Understanding graph-based trust evaluation in online social networks: Methodologies and challenges. ACM Computing Surveys (CSUR), 2016.

[13] Yuan, D., et al. Detecting fake accounts in online social networks at the time of registrations. in Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security2019.

[14] Al-Qurishi, M., et al., A prediction system of Sybil attack in social network using deep-regression model. Future Generation Computer Systems, 2018.

[15] Gao, T., et al., A Content-Based Method for Sybil Detection in Online Social Networks via Deep Learning. IEEE Access, 2020.

[16] Misra, S., A.S.M. Tayeen, and W. Xu. SybilExposer: An effective scheme to detect Sybil communities in online social networks. in 2016 IEEE International Conference on Communications (ICC).

[17] Zhou, Q. and G. Chen, An Efficient Victim Prediction for Sybil Detection in Online Social Network. IEEE Access, 2020.

[18] Zheng, H., et al., Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks. arXiv preprint arXiv:1709.06916, 2017.

[19] Höner, J., et al. Minimizing trust leaks for robust sybil detection. in International Conference on Machine Learning. 2017.

[20] Kumar, S., P. Kumar, and B. Bhasker, Interplay between trust, information privacy concerns and behavioural intention of users on online social networks. Behaviour & Information Technology, 2018.

[21] Wang, B., L. Zhang, and N.Z. Gong. SybilSCAR: Sybil detection in online social networks via local rule based propagation. in IEEE INFOCOM 2017-IEEE Conference on Computer Communications.

[22] Asadian, H. and H.H.S. Javadi, Identification of Sybil attacks on social networks using a framework based on user interactions. Security and Privacy, 2018. 1(2): p. e19.

[23] Boshmaf, Y., et al., Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs. Computers & Security, 2016.

[24] Boshmaf, Y., et al. Integro: leveraging victim prediction for robust fake account detection in OSNs. in Ndss. 2015.

[25] Ullah, F. and S. Lee, Community clustering based on trust modeling weighted by user interests in online social networks. Chaos, Solitons & Fractals, 2017.

[26] Wu, Y., et al., A novel framework for detecting social bots with deep neural networks and active learning. Knowledge-Based Systems, 2021.

[27] Wanda, P. and H.J. Jie, DeepProfile: Finding fake profile in online social network using dynamic CNN. Journal of Information Security and Applications, 2020.

[28] Roy, P. and M. Sood. Implementation of Ensemble-Based Prediction Model for Detecting Sybil Accounts in an OSN. in International Conference on Innovative Computing and Communications. 2021.

[29] RAM, A. and R.K. GALAV, Detection and Identification of Bogus Profiles in online Social Network using Machine Learning Methods. European Journal of Molecular & Clinical Medicine, 2020.

[30] Orabi, M., et al., Detection of bots in social media: a systematic review. Information Processing & Management, 2020.

[31] Kumari, A. and M. Sood. Performance Analysis of the ML Prediction Models for the Detection of Sybil Accounts in an OSN. in International Conference on Innovative Computing and Communications. 2021.

[32] Kadam, N. and H. Patidar, Social Media Fake Profile Detection Technique Based on Attribute Estimation and Content Analysis Method. in International Conference on Innovative Computing and Communications.2017.

[33] Jiang, Z., et al., Similarity-Based and Sybil Attack Defended Community Detection for Social Networks. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020.

[34] Jabardi, M.H. and A.S. Hadi, Ontology Meter for Twitter Fake Accounts Detection2019.

[35] Breuer, A., R. Eilat, and U. Weinsberg. Friend or faux: Graph-based early detection of fake accounts on social networks. in Proceedings of The Web Conference 2020.

[36] 35. Bhavani, D.Y., et al., Fake profiles detection on social media using machine learning. IEEE Access, ۲۰۱۷.

[37] Chavoshi, N., H. Hamooni, and A. Mueen. Identifying correlated bots in twitter. in International conference on social informatics. 2016.

[38] Liu, S., L. Zhang, and Z. Yan, Predict pairwise trust based on machine learning in online social networks: A survey. IEEE Access, 2018.

[39] Tiwari, V. Analysis and detection of fake profile over social network. in 2017 International Conference on Computing, Communication and Automation (ICCCA). 2017.

[40] Amini A, Wah TY, Saboohi H (2014) On density-based data streams clustering algorithms: a survey. *J Comput Sci Technol* 29:116–141.