Iran. J. Insur. Res., **(*): *-*, ****** ***

Iranian Journal of Insurance Research
(IJIR)
Homepage: https://ijir.irc.ac.ir/?lang=en

A mathematical model for identifying and validating suspicious clusters associated with organized fraud in auto insurance

A. Hamzeh[1,*], M. J. Nadjafi-Arani[2]

[2] *Associate Professor, Department of Computer Science, Mahallat institute of higher education, Markazi, Iran*

[1] *Assistant Professor of Insurance Research Center, Tehran, Iran*

ARTICLE INFO

***KEYWORDS*:**
Car insurance
Graph theory
Poisson distribution
labeling

ABSTRACT

BACKGROUND AND OBJECTIVES: Insurance fraud is a common challenge in the industry, leading to significant losses both in terms of financial interests and public trust. Financial and monetary institutions are keenly seeking to accurately identify the activities of fraudsters and fraudsters. Due to its direct effect on serving the clients of institutions, this will lead to the reduction of operating costs, gaining the trust of other insurers, and maintaining and improving the market share of insurers as reliable financial service providers. One of the most prevalent forms of fraud occurs in auto insurance, where organized and opportunistic fraudulent activities are widespread. Intentional accidents, especially those involving groups, staged injuries, and orchestrated scenes, are among the common fraudulent practices in this domain.

METHODS: One of the techniques used to detect fraud is network analysis. In the network analysis, the communication between people and different real and legal personalities are evaluated and new dimensions of these communication are identified. The objective of this paper is to introduce a mathematical model based on graph theory for identifying suspicious clusters associated with organized fraud. In our research, we first introduce a network called the “accident network” using graph theory. We demonstrate that this network exhibits characteristics of a random graph. Suspicious clusters within this network are then identified using an algorithm based on graph theory. Subsequently, we examine the occurrence probability of such clusters in a random accident network by defining a binomial distribution over its edges.

FINDINGS: This process leads to assigning a label (indicating fraudulent or non-fraudulent) to each accident and individual. Considering the structure of the algorithm and its complexity, we can conclude that the proposed algorithm is simply capable of analyzing a lot of data.

CONCLUSION: Investigating this topic enables insurers to tailor different policies based on the labels assigned to individuals or accidents, ultimately aiming to reduce financial losses and enhance public trust.

Iran. J. Insur. Res., **(*): *-*, ****** ***

Iranian Journal of Insurance Research
(IJIR)
Homepage: https://ijir.irc.ac.ir/?lang=fa

یک مدل ریاضی برای شناسایی و اعتبارسنجی خوشه‌های مشکوک به تقلب سازمان‌یافته در بیمه خودرو

اسماء حمزه [*۱]، محمدجواد نجفی آرانی [۲]

[۱] استادیار، گروه فناوری‌های نوین بیمه‌ای، پژوهشکده بیمه، تهران، ایران

[۲] دانشیار، گروه علوم کامپیوتر، مرکز آموزش عالی محلات، محلات، ایران

چکیده

پیشینه و اهداف: تقلب در صنعت بیمه یکی از مشکلات رایج در این حوزه است که موجب خسارات سنگینی چه از جهت منافع مادی و چه از جهت اعتماد عمومی در این صنعت می‌گردد. مؤسسات مالی و پولی به شدت به دنبال شناخت دقیق فعالیت‌های کلاهبرداران و متقلبان هستند. این امر به دلیل اثر مستقیم آن روی خدمت‌رسانی به مشتریان مؤسسات، منجر به کاهش هزینه‌های عملیاتی، جلب اعتماد سایر بیمه‌گذاران و حفظ و ارتقاء سهم بازار بیمه گران به عنوان ارائه‌دهندگان خدمات مالی قابل اطمینان خواهد شد. یکی از رایج‌ترین تخلفات، تقلب‌های سازمان‌یافته و فرصت‌طلبانه در بیمه خودرو است. تصادفات عمدی بالاخص در غالب گروهی، صدمه دیدن افراد توسط وسیله نقلیه و یا صحنه‌سازی از جمله تقلب‌های رایج در این حوزه هستند. هدف این مقاله معرفی مدل ریاضی مبتنی بر نظریه گراف (شبکه) برای شناسایی خوشه‌های مشکوک برای تقلب‌های سازمان‌یافته است.

روش‌شناسی: یکی از تکنیک‌هایی که برای شناسایی تقلب کاربرد دارد، تحلیل شبکه است. در تحلیل شبکه ارتباطات بین افراد و شخصیت‌های حقیقی و حقوقی مختلف مورد ارزیابی قرار گرفته و ابعاد جدیدی از این ارتباطات شناسایی می‌شوند. در این پژوهش، ابتدا با استفاده از نظریه گراف، شبکه‌ای به نام شبکه تصادفات معرفی می‌شود. سپس نشان داده می‌شود که شبکه حاصل از تصادفات خودروها یک فرآیند تصادفی است. سپس در شبکه ساخته شده از تصادفات، مجموعه خودروهای مشکوک که در این ساختار تصادفی ایجاد نظم می‌کنند، با معرفی یک الگوریتم شناسایی می‌شوند.

یافته‌ها: این فرایند باعث تخصیص یک برچسب از جهت متقلب بودن یا نبودن به هر تصادف و به هر فرد می‌گردد. با توجه به ساختار الگوریتم و پیچیدگی آن می‌توان نتیجه گرفت الگوریتم پیشنهادی به سادگی قادر به تحلیل داده‌های بسیار زیاد است.

نتیجه‌گیری: بررسی این موضوع موجب می‌شود تا بیمه‌گر بتواند وابسته به برچسب هر فرد یا تصادف، سیاست‌گذاری‌های متفاوتی را برای برخورد با متخلفان اتخاذ کند تا بتواند در جهت کاهش زیان مالی و افزایش اعتماد عمومی گام بردارد.

کلمات کلیدی:
بیمه خودرو
نظریه گراف
توزیع پواسون
برچسب‌گذاری

۱.مقدمه

تقلبات بیمه‌ای انواع گوناگونی دارد و در تمام حوزه‌های بیمه‌ای رخ می‌دهند و طیف گسترده‌ای از ادعاهای اغراق‌آمیز تا تصادف‌ها و خسارت‌های تعمدی را در بر می‌گیرند. این تقلب‌ها سبب افزایش هزینه‌ها و در پی آن، افزایش مبلغ حق‌بیمه می‌شوند. از این‌رو به ضرر سایر بیمه‌گذاران نیز خواهد بود. با وجود پیشرفت‌های فراوان در شناسایی این تقلب‌ها، هزینه‌های ایجاد شده برای شرکت‌های بیمه‌ای در اثر این کلاهبرداری‌ها در حال افزایش است. شناسایی هوشمند تقلب در ادعای خسارت مشتریان قبل از پرداخت خسارت، می‌تواند شرکت‌های بیمه را تا حد زیادی در برابر هزینه‌های تحمیلی ناشی از کلاهبرداری‌های بیمه‌ای به نوعی ایمن کند. با توجه به حجم و نوع داده‌ها، روش‌های گوناگونی برای کشف تقلبات بیمه‌ای وجود دارند. استفاده از هر یک از مدل‌ها برای شناسایی تقلب، این امکان را به متخصصین شرکت‌های بیمه می‌دهد که با صرف زمان و هزینه کمتری تشخیص دهند که ادعای خسارت اعلام شده مشکوک به تقلب است یا خیر. اگرچه این روش‌ها بسیار سودمند هستند، اما محققان به علت وجود داده‌های زیاد اغلب مجبور به استفاده از الگوریتم‌های فرااکتشافی مانند الگوریتم ژنتیک و شبکه‌های عصبی می‌شوند. استفاده از این‌گونه ابزارها بلاخص الگوریتم‌های فرااکتشافی متنوعی که در هر یک از روش‌های فوق استفاده می‌گردد دارای نقایصی جدی می‌باشند. نقص اصلی این گونه ابزارها آن است که یافتن رابطه بین تعداد زیادی داده غیرمتعادل[1] نمی‌تواند پاسخ روشنی را به مسأله ارایه دهد. به طور دقیق‌تر، در استفاده از این گونه ابزارها ابتدا به داده‌هایی که افراد یا خودروهای متقلب را از غیرمتقلب جدا کنند، احتیاج داریم تا الگوریتم بتواند یادگیری را بر روی آنها انجام دهد و سپس به مرحله تست برسیم. اما از آنجایی که داده‌ها غیرمتعادل هستند، یعنی تعداد زیادی داده داریم که غیرمتقلب می‌باشند و تعداد داده‌های متقلب بسیار اندک هستند، لذا یادگیری الگوریتم داده‌های شرکت بیمه که افراد متقلب را از غیرمتقلب جدا می‌کند دارای دقت بسیار پایین خواهد بود. محققان برای رفع این مشکل از تکنیک‌های تحدید کردن فضای نمونه استفاده می‌کنند. به عبارت دیگر محققان تعداد داده‌هایی که غیرمتقلب هستند را برای یادگیری کاهش می‌دهند و یا تعداد داده‌هایی که متقلب هستند را در یادگیری افزایش می‌دهند، به این اعمال به ترتیب کم نمونه‌گیری[2] یا زیاد نمونه‌گیری[3] گویند (Zhai et al., 2021; Tarawneh et al., 2022; Bernarda & Della valle, 2022). مشکل اصلی زیاد نمونه‌گیری آن است که اصطلاحا در الگوریتم، بیش‌برازش[4] رخ می‌دهد و کم نمونه‌گیری باعث از دست دادن اطلاعات و لذا کاهش دقت می‌گردد.

[1] Imbalance
[2] Under-sampling
[3] Over-sampling
[4] Overfitting

نکته دوم آنکه در اغلب مقالات بالاخص مقالاتی که از هوش مصنوعی استفاده می‌گردد احتیاج به داده‌هایی داریم که برچسب تقلب توسط بیمه گر به آنها تخصیص داده شود. اما غالب داده‌های موجود شرکت‌های بیمه این نوع برچسب‌گذاری را ندارند، حتی اگر هم تعداد محدودی این برچسب گذاری را انجام دهند در موارد کاملا اثبات شده خواهند بود که بسیار محدود است. این امر باعث دقت پایین این‌گونه الگوریتم‌ها در قسمت کاربرد می‌گردد.

سومین نقیصه برمی‌گردد به تعداد داده‌های تصادفات که بسیار زیاد است و پیچیدگی محاسباتی بسیار بالایی را ایجاد می‌کنند. در برخی موارد ممکن است زمان اجرای یک الگوریتم هفته‌ها به طول بیانجامد.

در این پژوهش سعی شده است، این نقیصه‌ها با استفاده از مدل‌های ریاضی مبتنی بر شبکه و نظریه گراف رفع شود. لذا در این پژوهش، با استفاده از نظریه گراف به مدل‌سازی شبکه تصادفات پرداخته می‌شود. سپس الگوریتمی اکتشافی پیشنهاد می‌شود که مشکل نقیصه اول را حل می‌کند. سپس الگوریتم پیشنهادی با توجه به برچسبی که هم به تصادفات و هم به اشخاص می‌دهد در همان زمان تصادف، به مشکوک بودن آن می‌تواند پی ببرد. با توجه به ساختار الگوریتم و همان‌طور که در بخش‌های آتی پیچیدگی آن نیز مورد بحث قرار گرفته است، می‌توان نتیجه گرفت الگوریتم پیشنهادی به سادگی قادر به تحلیل داده‌های بسیار زیاد است.

۲. مبانی نظری

تقلب در صنعت بیمه، عملی ارادی برای به دست آوردن مزایا و بهره‌مندی غیرقانونی از سازمان‌ها و شرکت‌های بیمه است. به بیان دیگر، تمامی اموری که منجر به از بین رفتن باور و اعتماد عملکردی در بین ارکان صنعت بیمه می‌شود، تقلب در صنعت بیمه محسوب می‌گردد. تقلب و کلاهبرداری در صنعت بیمه بسیار متنوع است و به صورت مکرر و روزانه در اطراف ما اتفاق می‌افتد که از بارزترین نمونه‌های تقلب‌های بیمه‌ای، می‌توان به ایجاد تصادفات ساختگی برای دریافت خسارت‌های مالی و بدنی اشاره کرد. در کل دو نوع کلاهبرداری بیمه‌ای شامل فرصت‌طلبانه و سازمان‌یافته (حرفه‌ای) وجود دارد. کلاهبرداری فرصت‌طلبانه به طور معمول توسط شخصی انجام می‌شود که به سادگی فرصتی برای افزایش قیمت یک خسارت یا دریافت تخمینی مبالغه‌شده‌ای برای خسارت یا تعمیرات از شرکت بیمه خود دارد، در حالی که کلاهبرداری حرفه‌ای اغلب توسط گروه‌های سازمان‌یافته انجام می‌شود. آنها در واقع هویت‌های دروغین، سازمان‌ها یا برندهای متعدد را هدف قرار می‌دهند. این حلقه‌های متخلف اغلب از طریق خودی‌ها انجام می‌شود تا به آن‌ها کمک کنند تا با استفاده از برخی راه‌ها به طور همزمان از شرکت کلاهبرداری کنند. اگرچه مبالغ در حادثه‌هایی با کلاهبرداری حرفه‌ای بسیار بیشتر هستند، اما وقوع آنها کمتر از کلاهبرداری فرصت‌طلبانه از بیمه است (White, 2011). مبارزه با تقلب بیمه‌ای یک مشکل چالش‌برانگیز است. اکثر سیستم‌های سنتی قادر به یافتن کلاهبرداری‌های فرصت‌طلبانه هستند، اگرچه شرکت‌های بیمه به دلیل ذکر شده (بیشترین ضرر مالی) علاقه زیادی به شناسایی گروه‌های سازمان‌یافته دارند. در نتیجه شرکت‌های بیمه برای مقابله با این مشکل نیاز به استفاده از فناوری‌های مدرن و سیستم‌های هوشمند دارند. هدف این پژوهش استفاده از مدل‌های ریاضی مبتنی بر شبکه و نظریه گراف برای کشف تقلب است. لذا در ادامه مفاهیم موردنیاز بیان می‌شود.

یک شبکه، که در ریاضیات، گراف نیز نامیده می‌شود، از مجموعه‌ای ناتهی از اشیاء به نام رأس تشکیل شده (که با V نمایش داده می‌شوند) و همچنین مجموعه‌ای شامل یال‌ها، که رأس‌ها را به هم وصل می‌کنند و با E نشان داده می‌شوند. چنین گرافی را با $G = (V, E)$ نشان داده و به آن گراف ساده گویند. در گراف ساده بین هر دو رأس حداکثر یک یال وجود دارد. اگر یال e دو رأس v_1 و v_2 را به هم وصل کند، آنگاه آن را با $e = v_1v_2$ نشان می‌دهند. گراف وزن‌دار، گرافی است که به هر یک از یال‌ها یا به هر یک از رأس‌های آن عددی نسبت داده شده باشد. این اعداد وزن یال یا رأس نامیده می‌شوند. وزن یال می‌تواند نشان‌دهنده هزینه، مسافت، زمان یا هر مشخصه دیگری از یال باشد. تعداد رأس‌های گراف G یعنی $|V(G)|$ را مرتبه گراف گویند که با $n(G)$ نمایش داده می‌شود و تعداد یال‌های گراف یعنی $|E(G)|$ را اندازه گراف G گویند که با $m(G)$ بیان می‌شود. به طور معمول برای سهولت در کار و در صورت مشخص بودن گراف G به جای $n(G)$ از n و به جای $m(G)$ از m استفاده می‌گردد. درجه رأس v در گراف G برابر با تعداد یال‌هایی از گراف G است که به رأس v متصل هستند و آن را با $deg_G(v)$ یا به‌طور ساده‌تر با $deg(v)$ یا $d(v)$ نمایش می‌دهند. دو رأس u و v را دو رأس همسایه یا مجاور گویند، هرگاه توسط یالی به هم متصل شده باشند، به عبارت دیگر $uv \in E(G)$ برقرار باشد. یک زیرگراف از گراف G گرافی است که مجموعه رأس‌های آن زیرمجموعه‌ای از مجموعه رأس‌های گراف G و مجموعه یال‌های

آن زیرمجموعه‌ای از مجموعه یال‌های G باشد. زیرگراف H از G را القائی گویند، هرگاه $V(H) \subseteq V(G)$ و میان رأس‌های H تمام یال‌های موجود بین همین رأس‌ها در گراف G وجود داشته باشد. گرافی را که درجه تمام رأس‌های آن با هم مساوی است منظم گویند و اگر به ازای هر رأس v داشته باشیم $deg(v) = k$، آنگاه گراف G را k-منظم می‌نامند. گرافی را که هر رأس آن با تمام رأس‌های دیگر، مجاور باشد گراف کامل نامیده می‌شود. گراف کامل n رأسی با K_n نمایش داده می‌شود. اگر u و v دو رأس از گراف G باشند، یک مسیر از u به v را یک $u - v$ مسیر گویند هرگاه در G دنباله‌ای از رأس‌های دو به دو متمایز وجود داشته باشد که از u شروع و به v ختم می‌شود، به طوری که هر دو رأس متوالی این دنباله در G مجاور هم باشند. طول یک مسیر برابر با تعداد یال‌های موجود در آن مسیر (یکی کمتر از تعداد رأس‌های موجود در آن مسیر) است. یک مسیرِ n رأسی را با P_n نمایش می‌دهند. یک دور به طول n، C_n، مسیر بسته‌ای به طول n است. به عبارت دیگر، ابتدا و انتهای این مسیر رأس‌های یکسانی می‌باشند. وتر یا قطر، یالی است که دو رأس غیرمجاور از یک دور را به هم وصل می‌کند. گراف G را همبند می‌نامند، هرگاه بین هر دو رأس آن حداقل یک مسیر وجود داشته باشد، در غیر این صورت آن را ناهمبند گویند. گراف n رأسی بدون دور که درجه یک رأس $n - 1$ و درجه مابقی رأس‌های آن یک باشد را گراف ستاره گویند. یک برش یا مجموعه جداکننده گراف همبند G، مجموعه‌ای از رأس‌ها است که با حذف آنها، گراف G ناهمبند می‌شود. عدد همبندی یا عدد همبندی رأسی که با $\kappa(G)$ نشان داده می‌شود تعداد کمینه رأس‌های مجموعه برشی است. یک گراف، k-همبند یا k-همبند رأسی نامیده می‌شود اگر تعداد رأس‌های همبندی آن، حداقل k باشد. این بدین معنی است که گراف G را زمانی k -همبند می‌نامند که در آن، یک مجموعه $k - 1$ عضوی از رأس‌ها که با حذف آنها گراف ناهمبند شود، وجود نداشته باشد. مجموعه برش رأس‌های u و v مجموعه‌ای از رأس‌ها است که با حذف آنها، ارتباط بین u و v قطع می‌شود. عدد همبندی محلی $\kappa(u, v)$ برابر با کمترین تعداد رأس‌هایی است که با حذف آنها، ارتباط u و v قطع می‌شود. واضح است که همبندی محلی برای گراف‌های بدون جهت، متقارن است (یعنی $\kappa(u, v) = \kappa(v, u)$) و در ضمن به استثنای گراف‌های کامل، $\kappa(G)$ به ازای هر دو انتخاب دلخواه از رأس‌های u و v برابر با کمینه $\kappa(u, v)$ است. مفاهیم مشابهی را می‌توان برای یال‌ها نیز تعریف کرد. به عنوان نمونه، اگر حذف یک یال خاص، آن گراف را ناهمبند کند، در این صورت این یال، پل نامیده می‌شود. به‌طور کلی، مجموعه یال برشی گراف G مجموعه‌ای از یال‌ها است که حذف آن‌ها، گراف را ناهمبند می‌کند. عدد همبند یالی $\kappa'(G)$ برابر با اندازه کوچکترین مجموعه یال‌های برشی است و عدد همبند یالی محلی $\kappa'(u, v)$ رأس‌های u و v برابر با اندازه کوچکترین مجموعه یال برشی است که u و v را از هم جدا می‌کند. همبندی یالی محلی نیز متقارن است. یک گراف k-همبند یالی نامیده می‌شود، اگر عدد همبندی یالی آن حداقل $\kappa'(G)$ باشد. مجموعه‌ای از مسیرها بین u و v را مجزای درونی (یالی) می‌نامند، اگر هیچ دوتایی از آن‌ها، رأس (یال) مشترک نداشته باشند (به جز رأس‌های u و v). قضیه مشهور منگر بیان می‌کند که عدد همبندی (عدد همبند یالی) یک گراف برابر تعداد مسیرهای مجزای درونی (یالی) بین رأس‌های مشخص می‌باشد (West, 2001).

۲-۱. فرآیندهای تصادفی

در این بخش به مفاهیم فرآیندهای تصادفی می‌پردازیم که نقش عمده‌ای را در نتایج اصلی این پژوهش بازی می‌کنند. مرجع اصلی مطالب این بخش، کتاب (Ghahramani, 2005) می‌باشد. برای مطالعه پدیده‌های دنیای واقعی که در آن سیستم‌ها به صورت تصادفی عمل می‌کنند به جای مدل‌های قطعی به مدل‌های احتمالی نیاز است. پایه و اساس مدل‌های احتمالی را فرآیندهای تصادفی می‌گویند.

در آمار و احتمال، متغیر تصادفی یا ورتنده کاتوره‌ای متغیری است که مقدار آن از اندازه‌گیری برخی از انواع فرآیندهای کاتوره‌ای بدست می‌آید. به‌طور رسمی‌تر، متغیر تصادفی تابعی از فضای نمونه به اعداد حقیقی است. تابع توزیع احتمال بیانگر احتمال وقوع هر یک از مقادیر متغیر تصادفی می‌باشد. تابع جرم احتمال یک متغیر تصادفی به صورت رابطه (۲) تعریف می‌شود:

$$f_X(x) = p(X = x) = p(x) \qquad (۲)$$

متغیرهای تصادفی نقشی اساسی در این پژوهش بازی می‌کنند. در ادامه این بخش، متغیرهای تصادفی و توابع توزیع احتمال که در این پژوهش به آنها اشاره می‌شود، به طور مختصر شرح داده خواهند شد.

متغیرهای تصادفی برنولی از جمله ساده‌ترین نوع متغیرهای تصادفی هستند که تنها شامل دو برآمد موفقیت و شکست می‌باشند که به طور معمول به‌ترتیب آنها را با s و f نمایش می‌دهند. به عبارت دقیق‌تر، متغیر تصادفی X که به صورت $X(s) = 1$ و $X(f) = 0$ تعریف می‌شود، یک متغیر تصادفی برنولی است. اگر p احتمال موفقیت باشد، آنگاه $1 - p$ احتمال شکست متغیر تصادفی برنولی است. لذا تابع جرم احتمال X به صورت رابطه (۳) است.

$$p(x) = \begin{cases} p & if\ x = s \\ 1 - p & if\ x = f \\ 0 & o.w \end{cases} \qquad (۳)$$

(Ghahramani, 2005).

اگر n امتحان برنولی همه با امتحان موفقیت p به طور مستقل انجام شوند، آنگاه X، تعداد موفقیت‌های این امتحان، تشکیل یک متغیر تصادفی جدید به نام متغیر تصادفی دوجمله‌ای می‌دهد. این متغیر تصادفی، یک متغیر تصادفی دوجمله‌ای با دو پارامتر n و p نامیده می‌شود. در متغیر تصادفی دوجمله‌ای تابع جرم احتمال X به صورت رابطه (۴) تعریف می‌شود.

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \dots, n \\ 0 & o.w \end{cases} \qquad (۴)$$

امید ریاضی (که با $E(X)$ نمایش داده می‌شود) و واریانس (که با $Var(X)$ بیان می‌شود) برای متغیر تصادفی دوجمله‌ای با پارامترهای n و p در رابطه (۵) تعریف شده‌اند.

$$E(X) = np, \qquad Var(X) = np(1-p) \qquad (۵)$$

دقت کنید امید ریاضی (یا مقدار چشم‌داشتی یا مقدار انتظاری) در نظریه احتمالات، مقدار قابل انتظار از یک متغیر تصادفی گسسته است که برابر با مجموع حاصل‌ضرب احتمال وقوع هر یک از حالات ممکن در مقدار آن حالت می‌باشد. در نتیجه میانگین برابر است با مقداری که به‌طور متوسط از یک فرآیند تصادفی با بی‌نهایت تکرار انتظار می‌رود. همچنین وردایی (یا واریانس)، در نظریه احتمالات و آمار، نوعی سنجش پراکندگی است. ریشه دوم وردایی که انحراف معیار نامیده می‌شود دارای واحدی یکسان با متغیر اولیه است (Ghahramani, 2005).

فرآیند پواسون

در بخش قبل به معرفی توزیع برنولی و دوجمله‌ای پرداخته شد. در بسیاری از موارد در این نوع توزیع، یافتن $p(x)$ از فرمول توزیع دوجمله‌ای غیرممکن است، زیرا برای مقادیر نه چندان بزرگ n، مقدار $n!$ از بزرگترین عددی که یک کامپیوتر می‌تواند در خود ذخیره کند، بزرگ‌تر است. لذا برای غلبه بر این مشکل روش‌های گوناگونی به کار گرفته شده است که ازجمله این روش‌ها تقریب پواسون است. به عبارت دقیق‌تر، فرمول تقریبی برای تابع توزیع احتمال دوجمله‌ای وقتی تعداد امتحان‌ها بزرگ ($n \to \infty$)، احتمال موفقیت کوچک ($p \to 0$) و متوسط تعداد موفقیت‌ها ثابت باقی بماند ($np = \lambda$ که در آن λ مقداری ثابت است) نیز معرفی می‌شود. به عبارت دیگر، توزیع پواسون همواره به عنوان تقریبی برای توزیع دوجمله‌ای می‌باشد. این مفهوم را می‌توان به صورت زیر تعریف کرد.

تعریف ۱. فرض کنید X یک متغیر تصادفی گسسته با مقادیر ممکن $0,1,2, \dots$ باشد. آنگاه X را یک متغیر پواسون با پارامتر λ گویند هرگاه:

$$P(\mathrm{X} = \mathrm{n}) = \frac{(\lambda)^n e^{-\lambda}}{n!}. \qquad n = 0,1,2, \dots \qquad (۶)$$

تحت شرایطی که بیان شد، احتمال‌های دوجمله‌ای را می‌توان به وسیله احتمال‌های پواسون تقریب زد. در حالت کلی می‌توان گفت برای متغیر تصادفی پواسون X با پارامتر λ داریم:

$$E(X) = Var(X) = \lambda \quad (۷)$$

که در آن $E(X)$ نشان‌دهنده امید ریاضی X و $Var(X)$ نمایش‌دهنده واریانس آن می‌باشد.

توزیع پواسون به خودی خود و جدا از تقریبی برای توزیع دوجمله‌ای، اغلب در رابطه با مطالعه دنباله پیش‌آمدهای تصادفی که در طول زمان رخ می‌دهند، تحقق پیدا می‌کند، مانند تعداد تصادفاتی که در یک تقاطع رخ می‌دهد. فرض کنید با شروع از زمان $t = 0$، تعداد پیشامدها شمارش شده‌اند. بدین ترتیب برای هر مقدار t عددی مانند $N(t)$ به دست می‌آید که برابر با تعداد پیشامدهایی است که در بازه $[0, t]$ رخ می‌دهند. برای مطالعه توزیع $N(t)$ سه فرض ساده و طبیعی مانایی، استقلال خطی و تناسب خطی در نظر گرفته می‌شود (Ghahramani, 2005).

به عبارتی، فرآیند $\{N(t): t \geq 0\}$ را فرآیند پواسون با پارامتر $\lambda(> 0)$ گوییم اگر:

۱- $N(0) \equiv 0$.

۲- $\{N(t): t \geq 0\}$ با نموهای مستقل باشد.

۳- $\{N(t): t \geq 0\}$ با نموهای مانا باشد.

۴- به ازای هر $t > s \geq 0$ نمو $N(t) - N(s)$ به عنوان یک متغیر تصادفی دارای توزیع پواسون با پارامتر $\lambda(t - s)$ باشد.

۳. روش‌شناسی پژوهش

یکی از تکنیک‌هایی که برای شناسایی تقلب کاربرد دارد، تحلیل شبکه است. در تحلیل شبکه ارتباطات بین افراد و شخصیت‌های حقیقی و حقوقی مختلف مورد ارزیابی قرار گرفته و ابعاد جدیدی از این ارتباطات شناسایی می‌شوند. در عمل افراد حقیقی و حقوقی درون و بیرون از سازمان با یکدیگر ارتباطات مختلفی را برقرار می‌سازند. در هر شبکه چندین گره وجود داشته که بواسطه لینک‌های ارتباطی با یکدیگر متصل می‌شوند. در عمل شبکه ها از تعداد بسیار زیادی گره با ارتباطات بسیار زیاد شکل گرفته است. شبکه‌ها طبیعی‌ترین نمایش چنین حوزه رابطه‌ای هستند که امکان فرمول بندی و تجزیه‌وتحلیل روابط پیچیده بین موجودیت‌ها را فراهم می‌کنند.

در این پژوهش، ابتدا با استفاده از نظریه گراف، شبکه‌ای به نام شبکه تصادفات معرفی می‌شود. سپس نشان داده می‌شود که شبکه حاصل از تصادفات خودروها یک فرآیند تصادفی است. سپس در شبکه ساخته شده از تصادفات، مجموعه خودروهای مشکوک که در این ساختار تصادفی ایجاد نظم می‌کنند، با معرفی یک الگوریتم شناسایی می‌شوند.

۴. مروری بر پیشینه پژوهش

تکنیک‌های مختلفی برای کشف تقلب وجود دارد که تمام این تکنیک‌ها در مطالعات مختلف مورد استفاده قرار گرفته‌اند که در این بخش از پژوهش به طور مختصر به بررسی آنها پرداخته می‌شود.

اکثر سیستم‌های موجود تشخیص ناهنجاری تقلب، برای شناسایی گروه‌های مشکوک به کار برده می‌شوند (به عنوان نمونه، مرجع Nian et al., 2016) را ببینید). ادبیات مربوط به کشف تقلب سازمان‌یافته بسیار پراکنده است. به عنوان مثال، آنها می‌توانند شامل تجزیه و تحلیل PRIDIT باشند که توسط (Brockett & Levine, 1977) پیشنهاد شده است. در واقع این روش‌ها براساس امتیازات RIDIT و تجزیه‌وتحلیل مؤلفه اصلی بنا نهاده شده‌اند که توسط (Šubelj et al., 2011) برای تشخیص مؤلفه‌های مشکوک به کار گرفته شده است. نویسندگان در این مقاله یک سیستم خبره را برای شناسایی و بررسی گروه‌های کلاهبرداری بیمه خودرو پیشنهاد کرده‌اند. این سیستم با جزئیات زیاد توصیف و بررسی

شده است. در ضمن چندین مشکل فنی در کشف تقلب نیز در نظر گرفته شده است تا در عمل قابل اجرا باشد. نهادهای متقلب با استفاده از یک الگوریتم ارزیابی جدید، با نام الگوریتم ارزیابی تکراری (IAA)، پیدا می‌شوند. علاوه بر ویژگی‌های ذاتی موجودیت‌ها، الگوریتم روابط بین آن‌ها را نیز بررسی می‌کند. در (Bodaghi & Teimourpour, 2018)، یک سیستم خودکار برای شناسایی گروه‌های مجرم در بیمه خودرو پیشنهاد می‌کنند. این سیستم از تجزیه‌وتحلیل شبکه برای شناسایی رفتارهای مشکوک در تصادفات خودرو استفاده می‌کند. نویسندگان بدون در نظر گرفتن میزان اثرگذاری هر گره در تقلب مشکوک (عدم انتصاب برچسب رتبه‌بندی به گره‌ها)، روش جدیدی را بر مبنای شبکه تصادفات جهت شناسایی خوشه‌های متقلب سازمان‌یافته معرفی می‌کنند. سپس، در بخش ارزیابی با سیستم نمونه اولیه، روش پیشنهادی براساس داده‌های دنیای واقعی ارزیابی و نتیجه‌گیری می‌شود. در (Nian et al. 2016) یک روش رتبه‌بندی برای تشخیص ناهنجاری تقلب با در نظر گرفتن رتبه‌بندی طبقاتی نادر ارائه شده است که در آن ناهنجاری با توجه به یک طبقه اکثریت واحد و رتبه‌بندی ناهنجاری با توجه به بیش از یک الگوی اصلی ارزیابی می‌شود. در (Zhou et al., 2015) نویسندگان روشی برای مدل‌سازی تصادفات به صورت شبکه‌ای وزن‌دار بر مبنای علت تصادف ارائه کرده‌اند. با توجه به مقاله این نویسندگان، نظریه تئوری تحلیل شبکه‌های پیچیده برای درک و تحلیل علت حوادث در سیستم‌های پیچیده مورد توجه قرار گرفت. آنها روش جدیدی را برای ایجاد شبکه وزن‌دار جهت‌دار (DWACN) برحسب علت تصادف برای شعبه‌های مورد مطالعه معرفی کردند و علت تصادف را براساس داده‌های کشور انگلستان بررسی کردند. در (Noble & Cook, 2003)، محققان ناهنجاری‌های مشکوک به تقلب سازمان‌یافته در شبکه‌های بزرگ با انواع مختلف گره‌ها مشخص کرده‌اند. اگرچه در روش پیشنهادی آنها ساختارهای مشکوک کشف شده‌اند، ولی گره‌های مشکوک نادیده گرفته شده‌اند. محققان همچنین معیارهای مرکزیت و فرآیندهای تصادفی را برای شناسایی ناهنجاری‌ها ارائه کرده‌اند، این در حالی است که این‌گونه رویکردها به طور عمده فقط بر ویژگی‌های رابطه‌ای گره‌ها تمرکز دارند.

(Óskarsdóttir et al., 2022) در مقاله خود، با پیوند دادن ادعاها با همه طرف‌های درگیر، از جمله بیمه‌شدگان، کارگزاران، کارشناسان و تعمیرکاران، یک شبکه تشکیل داده‌اند. سپس آنها کلاهبرداری را به عنوان یک پدیده اجتماعی در شبکه ایجاد کرده و از الگوریتم Bi Rank و بردار جستجوی خاص کلاهبرداری برای محاسبه امتیاز تقلب برای هر ادعا استفاده کرده‌اند. درنهایت، با استفاده از ویژگی‌های شبکه، مدل نظارت‌شده‌ای را برای کشف تقلب در بیمه اتومبیل پیاده‌سازی نموده‌اند. نتایج آنها نشان می‌دهد که مدل‌هایی با ویژگی‌های مشتق‌شده از شبکه هنگام تشخیص تقلب عملکرد خوبی دارند. ترکیب شبکه و ویژگی‌های خاص ادعا، عملکرد مدل‌های یادگیری تحت نظارت برای کشف تقلب را بیشتر ثابت می‌کند. مدل به دست آمده، ادعاهای خسارت مشکوک را که نیاز به بررسی بیشتر دارند، نشان می‌دهد. همچنین مقاله مروری (Pourhabibi et al., 2020) و مقاله (Rajan et al., 2019) از جمله مقالاتی هستند که با دیدگاه نظریه گراف و شبکه‌ها به موضوع تقلب در صنعت بیمه پرداخته‌اند که این بخش از ریاضیات تمرکز اصلی محققان در این پژوهش می‌باشد. همان‌طور که پیش از این ذکر گردید، هدف این پژوهش یافتن خوشه‌های ناهنجاری شبکه بوده تا با استفاده از این موضوع خوشه‌های مشکوک مشخص شوند. این روش می‌تواند علاوه بر در بر گرفتن کمبودهای برخی از مقالات بیان شده در بالا، راهکار جدیدی را نیز در جهت یافتن ناهنجاری‌های مشکوک به تقلب‌های سازمان‌یافته ارائه دهد.

۵. نتایج اصلی

در این بخش ابتدا نشان داده می‌شود که شبکه حاصل از تصادفات خودروها شرایط مانایی و مستقل افزایشی و تناسب خطی را دارا است و لذا نمایش یک فرایند پواسون مبتنی بر زمان می‌باشد. سپس در شبکه ساخته شده از تصادفات، مجموعه خودروهای مشکوک که در این ساختار تصادفی ایجاد نظم می‌کنند، شناسایی می‌شوند. مجموعه‌ای از خودروها را در یک محدوده جغرافیایی در نظر بگیرید. این خودروها در کل بازه زمانی که مورد بررسی قرار گرفته است، در این محدوده جغرافیایی در حال تردد هستند. واضح است که احتمال وقوع n تصادف در بازه‌های زمانی با اندازه‌های یکسان Δ_1 و Δ_2 مستقل از زمان و با هم برابر است (مانایی). تعداد تصادفات رخ داده در بازه زمانی $(t, t+s)$ مستقل از تصادفات رخ داده در قبل یا بعد از این زمان می‌باشد (استقلال افزایشی). تعداد رخداد چند تصادف دقیقا در یک زمان احتمال نزدیک به صفر دارد، به عبارت دیگر دو تصادف همزمان رخ نمی‌دهند (تناسب خطی). لذا بنابر تعریف ۱ تعداد تصادفات جاده‌ای در یک مکان مشخص یک فرایند پواسون است.

آنچه که در یک فرآیند تصادفی دارای اهمیت است، عدم وجود یک نظم مشخص در آن است. همان‌طور که از تعریف حرکت تصادفی برخورد بین خودروها مشخص است، امکان برخورد دو یا چند خودرو یکسان به یکدیگر در حوادث متفاوت امری بعید است که نمایش یک عدم تناسب در این فرآیند تصادفی خواهد بود. در ادامه، ساختارهایی گروهی یا انفرادی منظم در این فرآیند تصادفی جستجو می‌شوند که می‌توانند به عنوان موارد مشکوک شناسایی شوند. در شبکه‌ای که در این پژوهش تعریف و مورد بررسی قرار خواهد گرفت، ارتباطات بین افراد یا خودروهایی که با یکدیگر برخورد داشته‌اند، مورد ارزیابی قرار گرفته و ابعاد جدیدی از این ارتباطات شناسایی می‌شوند. در عمل افراد یا خودروها به عنوان گره در نظر گرفته شده و اگر دو خودرو به یکدیگر برخورد کرده باشند یک پیوند آنها را به یکدیگر متصل می‌کند. این نوع نمایش تشکیل شبکه‌ای از تصادفات می‌دهد که در ادامه برای تحلیل و نمایش داده‌های موجود از این ساختار استفاده خواهد شد. در این پژوهش از شبکه وزن‌دار جهت نمایش داده‌ها برای شناسایی و بررسی گروه‌های کلاهبرداران سازمان‌یافته بیمه خودرو استفاده می‌شود. در ادامه به صورت دقیق‌تر شبکه تصادفات تعریف شده و نظمی را که از نظر نویسندگان فرآیند تصادفی را در شبکه مشکوک تبدیل می‌کند، توصیف می‌شود.

یک شبکه تصادفات را به این صورت در نظر بگیرید، به طوری که مجموعه رئوس شامل تمامی خودروهای دارای بیمه (ثالث/بدنه) می‌باشد و مجموعه یال‌ها شامل تصادفات بین خودروها است. به عبارت دقیق‌تر، دو رأس (خودرو) a و b با هم مجاور هستند اگر و تنها اگر این دو با هم تصادف کرده باشند. ابتدا قسمت‌های محدودی از این شبکه مورد بررسی قرار می‌گیرد که ساختار منظم داشته باشد. توجه داشته باشید که شبکه تصادفات احتمالا شامل دورهایی با تنها سه گره است که برخورد سه خودرو به یکدیگر را نشان می‌دهد. ممکن است برخی یا اغلب این‌گونه تصادفات به صورت مصنوعی رخ دهند که برای بررسی گروه‌های سازمان‌یافته مناسب نیستند. بنابراین این پژوهش فقط بر روی شبکه‌هایی متمرکز می‌شود که تعداد گره‌های آنها بیش از سه گره باشند و در صورتی سه گره نیز بررسی می‌شود که بین آنها درصد هزینه‌ای که به بیمه‌گذار تحمیل شده است، قابل‌توجه باشد.

با توجه به آنچه در مورد تصادفی بودن برخورد بین خودروها ذکر گردید، از نظر نویسندگان ساختار منظم در شبکه فوق زیرگراف‌هایی به شکل دور، دورهای شامل وتر، زیرگراف‌های منظم، زیرگراف‌های کامل و زیرگراف‌هایی به شکل ستاره هستند که تعاریف آنها در بخش مبانی نظری ذکر شد. علت آنکه نویسندگان این ساختارها را مورد بررسی قرار می‌دهند آن است که به عنوان مثال در یک دور و یا یک دور وتردار تعدادی تصادف پی‌درپی رخ داده است که خودرو اول و آخر تصادف یکسان می‌باشند، حال داشتن وتر در این دور یعنی خودروهای مورد بررسی در این دور دارای همبندی محلی بیشتری هستند؛ لذا وابستگی بیشتری به یکدیگر خواهند داشت. حال اگر این دور مشمول در یک گراف منظم و یا کامل باشد، عدد همبندی زیرگراف افزایش می‌یابد و لذا درصد همبستگی بین رئوس آن بیشتر شده و بنابراین مشکوکیت بیشتری را خواهند داشت. این گونه زیرگراف‌های مشتق شده از دورها زیرگراف‌هایی دوهمبند هستند. دقت کنید که هر زیرگراف دوهمبند بیشینه را یک بلوک می‌نامند. زیرگراف دیگری که ممکن است مشکوک باشد، زیرگراف ستاره است. در حقیقت گراف ستاره نمایش خودرویی است که با چندین خودرو تصادف کرده است و از جهت انفرادی مورد بررسی قرار خواهد گرفت. برای شناسایی این ساختارهای مشکوک به الگوریتم‌هایی نیاز است که در بخش بعدی توضیح داده خواهند شد.

۵-۱. الگوریتم یافتن ساختارهای منظم سازمان‌یافته بر مبنای همبندی و فرایند پواسون

الگوریتم شامل دو بخش اصلی است. بخش اول الگوریتم به یافتن مجموعه‌های مشکوک به تقلب در شبکه تصادفات می‌پردازد و بخش دوم الگوریتم میزان این مشکوکیت را اعتبارسنجی و برچسب‌گذاری می‌کند.

بخش اول الگوریتم یافتن زیرگراف‌های متناظر به مشکوکیت در تقلب‌های سازمان‌یافته

برای یافتن ساختارهای منظم مبتنی بر دور تعریف شده در این بخش، از روش زیر استفاده می‌شود.

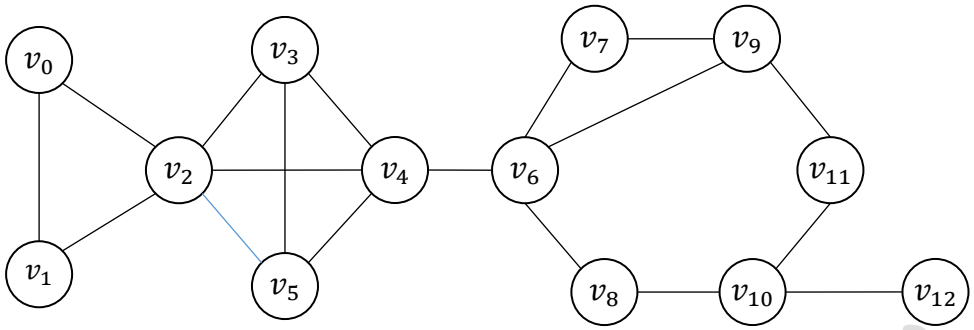
فرض کنید $e = uv$ یال دلخواهی باشد به طوری که درجه رئوس هر دو سر یال حداقل برابر با سه باشد و با حذف یال e گراف همچنان همبند باقی بماند. انتخاب یال با این ویژگی‌ها نشان می‌دهد که یال الزاما درون حداقل یک دور قرار می‌گیرد، زیرا با حذف یال e گراف همبند باقی می‌ماند، یعنی مسیر دیگری از u به v در گراف به غیر از یال e موجود است. این دور C_1 نامگذاری می‌شود. برای یک یال دلخواه مانند f، اگر این یال وتر یک دور باشد، دور مذکور شامل دو دور کوچک‌تر است که یال f را در خود دارند. حال از آنجایی که درجه رأس u و v در یال انتخابی e هر دو بالاتر از سه می‌باشند، لذا اگر یال e وتر یک دور باشد خود به خود دور C_2 موجود خواهد بود که $C_1 \cup C_2 - e$ یک دور بزرگ‌تر مانند C را نمایش

می‌دهد. با یافتن کلیه مسیرهایی که u را به v متصل می‌کند و با تکرار این عمل روی یال‌هایی که در طی الگوریتم مورد بررسی قرار می‌گیرند، می‌توان کلیه دورها، دورهای دارای وتر، زیرگراف‌های منظم و زیرگراف‌های کامل حداقل چهار رأسی را بدست آورد. برای لیست کردن مسیرها، ابتدا یالی مانند $e = uv$ که دو سر آن رأس‌های u و v با درجه بزرگ‌تر یا مساوی ۳ هستند (یعنی $deg(u) \geq 3, deg(v) \geq 3$) و تا به حال ملاقات نشده‌اند، در نظر گرفته می‌شوند. سپس برای ایجاد مسیرها، از رأس u به یکی از رأس‌های مجاور آن مراجعه می‌شود به طوری که این رأس تا به حال در مسیر دیده نشده باشد و این روند ادامه می‌یابد تا رأس v دیده شود. دقت داشته باشید که اگر به رأسی وارد شدید که رأس مجاوری دیده نشده، نداشته باشد با بازگشت به عقب (backtrack) به اولین رأسی باز گردید که شرایط ادامه مساله را داشته باشد (یعنی دارای رأس مجاوری باشد که تا به حال در مسیر دیده نشده است). بیان این نکته حائز اهمیت است که حداقل یک مسیر بین این دو رأس پیدا خواهد شد، زیرا بین دو رأس u و v یک دور وجود دارد. پس از یافتن اولین مسیر، تمام رأس‌های این مسیر به یک مجموعه با نام $Con(u, v)$ که شامل تمام رأس‌های بین u و v است اضافه می‌شود. در انتهای فرآیند یافتن مسیرهای بین u و v (یعنی زمانی که مسیر دیگری بین آنها وجود نداشته باشد)، تمام رأس‌های موجود در مجموعه $Con(u, v)$ به عنوان ملاقات‌شده علامت زده می‌شود. این امر سبب می‌شود مسیرهای تکراری در لیست نهایی حضور نداشته باشند، زیرا مسیرهای بین دو رأس مانند $x, y \in Con(u, v)$ با مسیرهای بین u و v یکسان هستند. این الگوریتم به صورت یک اسکریپت در محیط محاسباتی متلب نوشته شده و شبه کد آن در الگوریتم ۱ بیان شده است. حال جهت بررسی اجرای الگوریتم یک مثال ساده بررسی می‌شود.

الگوریتم ۱: شبه کد روش پیشنهادی

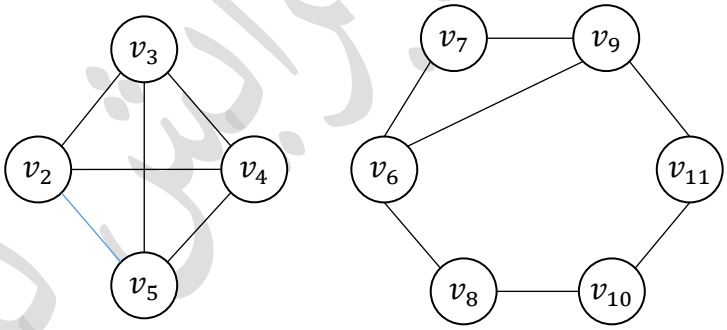
```
procedure AllPaths(graph G)
    foreach e = (u, v) in E
        if (deg(u) ≥ 3 and deg(v) ≥ 3 and Visited(u)==false and Visited(v)==false)
            FindAllPaths(G, u, v)
endprocedure
procedure FinaAllPaths(graph G, node u, node v)
    Con = u, v
    P =
    BacktrackingPaths(G, u, v, , P)
    foreach x in Con(u, v)
        Visited(x)=true
endprocedure
procedure BacktrackingPaths(graph G, node u, node v, Path p, PathList P)
    if u==v then
        add p to path list P
        foreach x in p
            add x to Con
    if there is no adjacent node for u then
        return false
    add u to p
    BacktrackingPaths(G,adjacent(u),v,p,P)
Endprocedure
```

مثال ۱. گراف ساده زیر (شکل ۲) را به عنوان یک شبکه تصادف بسیار کوچک در نظر بگیرید.



شکل ۲. یک شبکه ساده تصادفات

ابتدا الگوریتم یال v_2v_3 را به عنوان یالی که رئوس دو طرف آن از درجه بزرگ‌تر یا مساوی سه است را در نظر می‌گیرد. سپس الگوریتم بررسی می‌کند که یال مورد نظر برشی نمی‌باشد. آنگاه مسیرهای $v_2v_4v_3$ و $v_2v_5v_3$ و $v_2v_5v_4v_3$ به عنوان مسیرهای دیگری که دو رأس v_2 و v_3 را به یکدیگر متصل می‌کنند توسط الگوریتم مشخص می‌گردند. تمامی رئوسی که در این مسیرها قرار دارند مشکوک هستند. به عبارت دیگر، زیرگراف القایی حاصل از مجموعه رئوس $\{v_2, v_3, v_4, v_5\}$ یک زیرگراف مشکوک در شبکه تصادفات است که احتمالا درصد مشکوکیت هر رأس آن متفاوت از رأس دیگری است. سپس الگوریتم یال v_4v_6 که درجه دو سر آن بزرگ‌تر یا مساوی سه است را انتخاب می‌کند و از آنجایی که مسیر دیگری این دو یال را به یکدیگر متصل نمی‌کند از آن گذر کرده و یال v_6v_7 را به عنوان یال پایه در نظر می‌گیرد و مسیرهای $v_6v_9v_7$ و $v_6v_8v_{10}v_{11}v_9v_7$ به عنوان مسیرهای مشکوک توسط الگوریتم شناسایی می‌گردند. همان‌طور که در مثال مشخص گردید، زیرگراف‌های شکل ۳ به عنوان زیرگراف‌های مشکوک در این گراف توسط الگوریتم شناسایی می‌شوند.



شکل ۳. زیرگراف‌های مشکوک به تقلب

دقت کنید در این گراف در حال حاضر دور سه‌تایی $v_0v_1v_2$ به عنوان زیرشبکه مشکوک قرار نمی‌گیرد مگر آنکه یک یال دیگری بین رئوس همین دور سه‌تایی یا زیرگراف کامل چهار رأسی مجاور آن بوجود بیاید.

در ادامه به مقایسه ویژگی‌های الگوریتم معرفی شده در این مقاله با مقالات دیگر و وجه تمایز آن می‌پردازیم.

آنالیز بخش اول الگوریتم

لازم به ذکر است که در طی فرآیند این الگوریتم کلیه دورها پیدا نمی‌شوند، به عنوان مثال دوری به طول چهار از نظر نویسندگان مشکوک نمی‌باشد، اما اگر دور به طول چهار دارای وتر باشد یا حداقل دو خودروی که پیش از این با هم تصادف کرده‌اند با خودروهای دیگری به جز خودروهای مورد بررسی در دور نیز برخورد داشته باشند مشکوک تلقی شده و مورد بررسی قرار خواهند گرفت. در واقع نویسندگان بنابر این اصل چنین موضوع را در نظر گرفته‌اند که احتمال وجود دور تنها در تصادفات زنجیری زیاد است و اگر بخواهد این دور نشان‌دهنده یک تخلف باشد، آنگاه الزاما متخلفان برای

ادامه کار خود این دور را وتردار خواهند کرد و یا متخلفان پیش از این، تصادفات متخلفانه دیگری را داشته‌اند. از نظر نویسندگان این نگاه می‌تواند معیار دقیق‌تری را نسبت به انتخاب دور تنها که در مقاله (Bodaghi & Teimourpour, 2018) در نظر گرفته شده است، برای بررسی مشکوکیت ایجاد کند. همچنین در مقاله (Óskarsdóttir, 2022) نشان داده شده است که بیش از ۷۰ درصد خودروها که دارای تنها دو برخورد هستند، مشکوک به تقلب نمی‌باشند. لذا داده‌های این مقاله نویسندگان را بر آن داشت تا دو خودرو را زمانی به عنوان پایه یک تخلف سازمان‌یافته در نظر بگیرند که دارای حداقل سه تصادف باشند. به عبارت دیگر، استفاده از رئوس با درجه سه یا بیشتر برای شروع الگوریتم امری مطابق بر آمار است.

نکته دوم آن‌که هنگامی که یک یال درون دورهای مختلفی قرار می‌گیرد، در واقع همبندی محلی یال افزایش پیدا می‌کند و این بدان معنا است که رئوس (خودروهای) منتصب به دو سر این یال دارای مشکوکیت بالاتری نسبت به مابقی رئوسی (خودروهایی) که در دورهای شامل این یال قرار دارند، هستند. در بخش اعتبارسنجی میزان همبستگی بین این رئوس که از عدد همبندی محلی آن‌ها در گراف مشخص می‌گردد در اعتبار هر خودرو یا تصادف تأثیر به سزایی دارد. دقت کنید در فرآیند اجرای الگوریتم عدد همبندی کل گراف اصلا دارای اهمیت نمی‌باشد. نویسندگان درواقع مانند مقاله (Pinheiro, 2011) عدد همبندی گراف را مورد بررسی قرار نداده‌اند. آنچه برای نویسندگان اهمیت دارد، همبندی محلی رئوس دو سر یک یال است. در واقع ممکن است گراف k-همبندی وجود داشته باشد که پس از بررسی الگوریتم، اصلا به عنوان زیرگراف مشکوک شناسایی نشود. به عنوان مثال، k مسیر به طول بزرگ‌تر از سه که ابتدا و انتهای همه آنها یکسان بوده و گره مشترک دیگری بین این k مسیر وجود نداشته باشد (k مسیر مجزای درونی) را در نظر بگیرید. در این حالت هیچ یالی که دو سر آن از درجه بزرگ‌تر یا مساوی سه باشد، وجود ندارد و لذا الگوریتم هیچ زیرگراف مشکوکی را شناسایی نمی‌کند، در حالی که بنابر قضیه منگر این گراف 2-همبند است و عدد همبندی محلی دوسر این k مسیر برابر با k است. اما اگر این گراف شامل تنها یک وتر شود کل گراف در فرایند الگوریتم مشکوک به حساب می‌آید.

تجزیه‌وتحلیل الگوریتم بخش مهمی از نظریه پیچیدگی محاسباتی است که تخمین نظری مناسبی برای منابع مورد نیاز یک الگوریتم، مانند زمان و حافظه موردنیاز برای حل یک مشکل محاسباتی خاص، ارائه می‌دهد. در واقع، این تجزیه‌وتحلیل به طراحان کمک می‌کند تا رفتار یک الگوریتم را بدون پیاده‌سازی آن بر روی یک رایانه خاص، پیش‌بینی کند. یکی از مهم‌ترین تحلیل‌ها، محاسبه بزرگی زمانی الگوریتم است. حال برای این منظور نیاز است تعداد یال‌ها مشخص شود. بیشینه تعداد یال‌ها در یک گراف کامل رخ می‌دهد که برای گرافی با n رأس این تعداد برابر با $n(n-1)/2$ و دارای بزرگی زمانی $O(n^2)$ است. در ادامه، برای هر دو رأس یک یال، مسیرها لیست می‌شوند. دقت داشته باشید که بیشترین تعداد مسیرها بین دو گره نیز در گراف کامل ایجاد می‌شود. حال فرض کنید n تعداد رأس‌های گراف G است. برای دو رأس u و v، می‌توان k رأس از بین $n-2$ رأس باقی مانده را به صورت زیر انتخاب کرد:

$$\binom{n-2}{k} = \frac{(n-2)!}{k!\,(n-2-k)!} \qquad (۸)$$

و آنها را می‌توان به هر ترتیب دلخواهی در مسیر قرار داد. لذا تعداد کل مسیرها با جمع کردن این مقادیر برای همه k های ممکن به شکل زیر حاصل می‌شود:

$$\sum_{k=0}^{n-2} \frac{(n-2)!}{k!\,(n-2-k)!} = (n-2)! \times \sum_{i=0}^{n-2} \frac{1}{i!} \qquad (۹)$$

از آنجایی که

$$e = \sum_{i=0}^{\infty} \frac{1}{i!} \qquad (۱۰)$$

لذا تعداد مسیرها برابر با $e \times (n-2)!$ و دارای بزرگی زمانی $O(n!)$ است. در نتیجه بزرگی زمانی روش پیشنهادی برابر با $O(n^2 \times n!)$ است که مقداری بسیار بزرگ است. با این حال، مهم است که توجه داشته باشید که این بدترین سناریو حالتی است که شبکه تصادفات را یک گراف کامل

در نظر بگیریم که در آن همه اتومبیل‌ها با یکدیگر برخورد می‌کنند. چنین وضعیتی با توجه به ساختار شبکه تصادفات واقعی نیست. برای تراز کردن مسئله با شرایط دنیای واقعی، فرض کنید که بزرگترین زیرگراف کامل در شبکه تصادف متشکل از m گره است که $m \ll n$ یک عدد ثابت است.

در این مورد، الگوریتم ما با در نظر گرفتن همه زیرگراف‌های کامل متمایز با m رأس شروع می‌شود. این زیرگراف‌های کامل به عنوان گراف‌های سطح ۱ نامیده می‌شوند. حداکثر تعداد نمودارهای سطح ۱ برابر با n/m خواهد بود. در سطح ۲، هر گراف سطح ۱ را به عنوان یک گره در نظر می‌گیریم. اگر یک یال بین دو گراف سطح ۱ وجود داشته باشد، یک یال مربوطه را بین گره‌های مربوطه در سطح ۲ ایجاد می‌کنیم. طبق فرض، تعداد گراف‌های کامل در سطح ۲ برابر با n/m^2 خواهد بود. این فرآیند برای سطوح بعدی ادامه می‌یابد تا زمانی که یک گراف کامل در سطح k تشکیل شود، جایی که $k = \log_m n$. واضح است که در سطح i، حداکثر $m! \times m/n^i$ مسیر وجود خواهد داشت. از آنجایی که m یک عدد ثابت است، پیچیدگی زمانی تعیین مسیرها در کل شبکه $O(n \log n)$ است.

بخش دوم الگوریتم اعتبارسنجی و برچسب‌گذاری

هدف این بخش تحلیل دو موضوع مهم است. ابتدا سطح مشکوکیت هر ادعا بررسی می‌شود و به هر حادثه یک برچسب اختصاص داده می‌شود. این برچسب با ترکیب عدد همبندی محلی و فرآیند پواسون تعیین می‌گردد. از آنجایی که همبندی محلی را می‌توان با استفاده از یال‌ها یا رئوس بررسی کرد، انواع مختلفی از برچسب‌ها را می‌توان تولید کرد. این برچسب‌ها در بخش بعدی به تفصیل بیان می‌شوند. سپس سطح اعتباری برای هر فرد بیمه‌شده بر اساس میزان ظن مطالبات آن مشخص می‌گردد. این سطح اعتبار به عنوان یک برچسب به هرتصادف و هر خودرو در شبکه اختصاص داده می‌شود.

اعتبارسنجی هر تصادف

با توجه به مدل پیشنهادی، هر یال نشان‌دهنده یک تصادف است. A را به عنوان یک شبکه تصادف از مرتبه n و اندازه m در نظر بگیرید. فرض کنید $e = uv$ یک یال دلخواه باشد و X تعداد مسیرهای مجزای درونی بین u و v در بین همه مسیرها باشد. اگر وجود یک مسیر مجزای درونی بین u و v را به عنوان موفقیت با احتمال p تعریف کنیم، آنگاه X یک متغیر تصادفی دوجمله‌ای است. از آنجایی که تعداد مسیرها زیاد است ($m \to \infty$)، احتمال موفقیت کوچک است ($p \to 0$)، بنابراین امید ریاضی λ ثابت باقی می‌ماند. با توجه به توضیح بخش قبل، می‌توان نتیجه گرفت که این توزیع دوجمله‌ای یک متغیر تصادفی پواسون با پارامتر λ است. بر اساس قضیه منگر به ازای عدد همبندی محلی $\kappa(u,v)$ به همین تعداد مسیرهای مجزای درونی بین u و v موجود است، لذا می‌توانیم استنباط کنیم $\kappa(u,v)$ یک متغیر تصادفی پواسون است. اکنون بر اساس فرمول (۶) داریم:

$$Pr\big(X = \kappa(u,v)\big) = \frac{e^{-\lambda}\lambda^{\kappa(u,v)}}{\kappa(u,v)!} \qquad (۱۱)$$

مقدار $\kappa(u,v)$ در طول بخش اول فرآیند الگوریتم پیشنهادی به دست می‌آید. مقدار λ به صورت تقریبی از فرمول mp بدست می‌آید. با این حال، مقدار بدست امده برای یال $e = uv$ از معادله فوق مستقل از زیرگراف حاصل از بخش اول الگوریتم است و آن را به خوبی تعریف می‌کند. یک مقدار احتمال کمتر برای یک یال نشان‌دهنده سطح بالاتری از همبستگی بین گره‌های آن در شبکه است، زیرا آنها از طریق مسیرهای مختلف رخ می‌دهند. در واقع، یک عدد برچسب کمتر در یال نشان‌دهنده سطح بیشتری از مشکوکیت مرتبط با تصادف است. بنابراین، برچسب یال $e = uv$ به صورت زیر تعریف می‌شود:

$$l_{uv} = 1 - Pr\big(X = \kappa(u,v)\big) \qquad (۱۲)$$

توجه داشته باشید که وقتی احتمال $X = \kappa(u,v)$ روی یک یال افزایش می‌یابد، برچسب تخصیص یافته به صفر نزدیک می‌شود، که نشان می‌دهد تصادف مربوطه کمتر مشکوک است. به همین ترتیب، اگر احتمال یک یال نزدیک به صفر باشد، نشان‌دهنده سطح شک بالاتر برای تصادف در شبکه است.

تا به حال، تعداد مسیرهای مجزای درونی رأس بین u و v یعنی $\kappa(u,v)$ را در فرمول پواسون اعمال کرده‌ایم تا یک برچسب به هر ادعا اختصاص دهیم. با همان آرگومان و بدون از دست دادن کلیت، می‌توانیم به طور مشابه از تعداد مسیرهای مجزای درونی یالی $\kappa'(u,v)$ و تعداد همه مسیرها $\kappa''(u,v)$ به ازای هر یال $e = uv$ برچسب‌های دیگری را معرفی کنیم. در تمام این موارد، اتصال بین گره‌ها به روش‌های مختلف در نظر گرفته می‌شود. به راحتی می‌توان مشاهده کرد که $\kappa(u,v) \le \kappa'(u,v) \le \kappa''(u,v)$ تمرکز اصلی $\kappa(u,v)$ روی همبندی محلی بین رئوس است، در حالی که $\kappa'(u,v)$ نشان‌دهنده همبندی محلی بین یال‌ها است و $\kappa''(u,v)$ همبستگی[1] در گراف را نشانه می‌گیرد. از آنجایی که هر یک از متریک‌ها ویژگی متفاوتی را در نظر می‌گیرند، بنابراین دو برچسب جدید l'_{uv} و l''_{uv} به صورت زیر تعریف می‌شود:

$$l'_{uv} = 1 - Pr\big(X = \kappa'(u,v)\big), \qquad (۱۳)$$

$$l''_{uv} = 1 - Pr\big(X = \kappa''(u,v)\big).$$

به عنوان مثال، زیرگراف مشکوک $v_6v_8v_{10}v_{11}v_9v_7$ را در شکل ۳ در نظر بگیرید. برای یال v_6v_9 ، سه مسیر داخلی رأس مجزا (و همچنین سه مسیر یال مجزا و همچنین سه مسیر مختلف) بین دو رأس v_6 و v_9 وجود دارد. بنابراین، برچسب‌های این یال‌ها برابر است با:

$$l_{v_6v_9} = l'_{v_6v_9} = l''_{v_6v_9} = 1 - Pr(X = 3) = 1 - \frac{e^{-\lambda}\lambda^3}{۳!} \qquad (۱۴)$$

برای یال v_6v_7 داریم:

$$l_{v_6v_7} = l'_{v_6v_7} = 1 - Pr\big(X = \kappa(v_6,v_7)\big) = 1 - \frac{e^{-\lambda}\lambda^2}{۲!}, \qquad (۱۵)$$

$$l''_{v_6v_7} = 1 - Pr\big(X = \kappa(v_6,v_7)\big) = 1 - \frac{e^{-\lambda}\lambda^3}{۳!}.$$

اگرچه در این حالت سه مسیر متفاوت مابین v_6وv_7 وجود دارد، اما تنها دوتا از آنها رأس مجزا و یال مجزا هستند، بنابراین می‌توان نتیجه گرفت:

$$\kappa(v_6,v_7) = \kappa'(v_6,v_7) = 2$$

در حالی که

$$\kappa''(v_6,v_7) = 3.$$

توجه داشته باشید که الگوریتم با تمرکز بر زیرگراف‌های مشکوک منتج از بخش اول سعی می‌کند با بررسی برچسب‌های یال‌ها، سطح شک و تردید آنها را مشخص کند. یافتن برچسب‌ها روی زیرگراف‌های مشکوک، پیچیدگی محاسباتی را در مقایسه با کاوش کل شبکه کاهش می‌دهد. بنابراین، برای یال‌هایی که بخشی از این زیرگراف‌ها نیستند، عدد همبندی محلی ۱ اختصاص داده می‌شود.

برچسب‌گذاری و اعتبارسنجی خودروها

[1] betweennes

هدف بعدی این مقاله، اختصاص یک برچسب مشکوک به هر بیمه‌گذار است. تاکنون، برچسب‌های روی یال‌ها بر اساس احتمال پواسون تعیین شدند. مرحله بعدی شامل انتقال این برچسب‌ها به رئوس است. از آنجایی که چندین یال به هر رأس متصل است، لازم است در هنگام انتقال برچسب یال به رأس، تأثیر هر یال (که نشان‌دهنده میزان سوءظن در مورد هر تصادف است) در نظر گرفته شود. با این حال، به دلیل وجود چندین یال متصل به یک رأس، نرمال‌سازی برچسب‌های یال برای به دست آوردن تأثیر عادلانه هر یال بر روی راس بسیار مهم است. برای این منظور، برچسب‌های یال به صورت زیر نرمال می‌شوند.

فرض کنید l_{uv} نمایش برچسب یال با uv شد، L_u به صورت زیر تعریف می‌شود:

$$L_u = \left\{\sum_{v\in N_{\dot{}}(u)} l_{uv} \,\middle|\, v \in V(G)\right\}. \qquad (۱۶)$$

فرض کنید l_{max} و l_{min} به ترتیب نمایش بزرگ‌ترین و کوچک‌ترین مقدار L_u به ازای u های مختلف باشد. در این صورت برچسب رأس u که با l_u نمایش داده می‌شود به صورت زیر تعریف می‌شود:

$$l_u = \frac{\sum_{v\in N_G(u)} l_{uv} - l_{min}}{l_{max} - l_{min}}. \qquad (۱۷)$$

در واقع، زمانی که یال‌های متصل به رأس u مشکوکیت کمتری داشته باشند، تأثیر کمتری بر این رأس خواهند داشت. برعکس، اگر برچسب روی یک یال نزدیک به یک باشد، که نشان‌دهنده سطح سوء ظن بالاتر در شبکه است، تأثیر بیشتری بر روی برچسب رأس u خواهد داشت.
واضح است که متناسب با برچسب‌های l'_{uv} و l''_{uv} که در بخش قبل معرفی شدند، می‌توان به ترتیب برچسب‌های l'_u و l''_u را نیز برای هر خودرو تعریف کرد که متناسب با شرایط موردنظر بیمه‌گذار هر یک می‌توانند در موارد مختلف کارایی داشته باشند و مقایسه آن‌ها نیز باعث ایجاد اطمینان بیشتر به اعتبار تخصیص‌یافته می‌گردد.

در این پژوهش یافتن تقلب‌های سازمان‌یافته در بیمه خودرو با پیشنهاد مدلی ریاضی برمبنای نظریه گراف هدف قرار گرفت تا نواقصی که سایر الگوریتم‌های پیشنهادی قبلا ارائه شده دارد را برطرف نماید. لذا نشان داده شد که تصادفات بین خودروها و صدمه دیدن افراد یا خودروها فرآیندی تصادفی بوده و هرگونه دخالت سازمان‌یافته توسط افراد در این فرآیند تصادفی، سبب ایجاد نوعی نظم شده که می‌تواند سرنخی برای شناسایی تخلف در تصادفات باشد. به این ترتیب زیرگراف‌های مشکوک براساس شبکه تصادفات و یال‌های مجازی که نشان‌دهنده هزینه‌های زیاد یک خسارت نسبت به میانگین خسارت‌های پرداخت شده توسط بیمه‌گر است، کشف شدند.

۶. جمع‌بندی و نتیجه‌گیری

تقلب بیمه‌ای عملی است که با هدف کلاهبرداری از بیمه‌گر، برای کسب منافع مالی انجام می‌گیرد. تقلب بیمه‌ای از زمان شکل‌گیری بنگاه‌های تجاری وجود داشته و سالانه میلیاردها دلار، هزینه را به شرکت‌های بیمه تحمیل نموده است. تقلبات بیمه‌ای انواع گوناگونی دارد و در تمام حوزه‌های بیمه‌ای رخ می‌دهد و طیف گسترده‌ای از ادعاهای اغراق‌آمیز تا تصادف‌ها و خسارت‌های تعمدی را در بر می‌گیرد. تقلب‌های بیمه خودرو بالاخص تقلبات سازمان‌یافته که در این پژوهش مورد مطالعه قرار گرفته است، اغلب به صورت ساختار گروهی انجام می‌گیرد. این ساختار سبب افزایش کلان هزینه‌های بیمه‌گر و در پی آن، افزایش مبلغ حق‌بیمه می‌شود. امروزه با توجه به ضرورت کشف تقلب در حوزه‌های مختلف، استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین، مانند شبکه‌های عصبی مصنوعی، منطق فازی و الگوریتم‌های ژنتیک، به دلیل توانمندی بالایی که در مدل کردن مسائل پیچیده دارند، به ابزار رایج در کشف تقلب تبدیل شده‌اند. ابزار دیگری که به هدف کشف تقلب‌های سازمان‌یافته مورد استفاده قرار می‌گیرد، نظریه گراف است. در این نگاه ابتدا به مدل‌سازی ریاضی مسأله پرداخته می‌شود. بدین مفهوم که ابتدا شبکه تصادفات به صورت یک گراف مدل شده و با ابزارهای در دسترس به کشف تقلبات سازمان‌یافته پرداخته می‌شود. سپس از مفاهیم علوم کامپیوتری استفاده می‌گردد تا شبکه مشکوک به تقلب به صورت دقیق‌تر مشخص شود. به عبارت دقیق‌تر، در ساختارهایی مانند تصادفات یک کشور تعداد داده‌های موجود بسیار زیاد هستند و یافتن

رابطه بین آنها عملی دشوار است. استفاده از ابزارهایی مانند داده‌کاوی و یادگیری ماشین، شبکه‌های عصبی، منطق فازی، الگوریتم‌های ژنتیک و غیره که هدف اصلی آنها یافتن رابطه بین داده‌ها است اگرچه بسیار سودمند است، اما دارای نقایصی می‌باشند. این گونه ابزارها اگر از الگوریتم های فرا اکتشافی استفاده کنند عدم دقت یا بیش برازش در داده های نامتعادل را خواهند داشت، در الگوریتم های اکتشافی یافتن رابطه بین تعداد زیادی داده دارای پیچیدگی محاسباتی بسیار بالایی است که در برخی موارد ممکن است زمان اجرای یک الگوریتم هفته‌ها به طول بیانجامد. در این پژوهش نویسندگان سعی کرده‌اند این نقیصه را با استفاده از مدل‌های ریاضی رفع نمایند و در ضمن احتمال رخداد رویدادهای مشکوک را نیز بررسی کنند. لذا در این پژوهش، ابتدا با استفاده از نظریه گراف به مدل‌سازی شبکه تصادفات پرداخته شده و سپس نشان داده شده است که این مدل یک فرآیند تصادفی است و وجود عناصر منظم در مدل بیان‌کننده مجموعه خودروهای مشکوک به تقلب خواهد بود. در ادامه براساس الگوریتم یافتن زیرگراف مشکوک که به صورت یک فایل اسکریپت (فایل m) در محیط محاسباتی متلب نوشته شده است، خودروهای مشکوک از کلیه خودروها استخراج شده است. در پایان ثابت شده است که شبکه تصادف یک فرآیند پواسون است و احتمال رخداد آن می‌تواند مشخص شود. این استدلال که بر پایه ساختار مدل‌سازی گرافی بنا شده است، کمک می‌کند تا به هر تصادف و به هر خودرو درجه اعتباری به جهت مشکوک بودن به تقلب تخصیص یابد. پیشنهاد برای تحقیقات آتی، ایجاد و بررسی شبکه‌ای گسترده‌تر شامل همه ذینفعان یک تقلب سازمان‌یافته است. به عبارت دقیق‌تر، شبکه تخصیص برچسب ذینفعان اصلی که در یک تصادف سود می‌برند وابسته به مقدار سود آنها، مورد بررسی قرار گیرند. بررسی این موضوع موجب می‌شود تا بیمه‌گر بتواند سیاست‌گذاری‌های متفاوتی را برای برخورد با ذینفعان مختلف در یک تصادف، مانند بیمه‌گذار، سرنشینان خودرو، تعمیرکاران و... اتخاذ کند تا بتواند در جهت کاهش زیان مالی و اعتماد عمومی گام بردارد.

منابع:

Brockett, P. L.; Levine, A., (1977). On a Characterization of RIDITs. Annals of Statistics, 5(6), 1245–1248 (6 Pages).

Bernardo, A.; Della Valle, E., (2022). An extensive study of C-SMOTE, a Continuous Synthetic Minority Oversampling Technique for Evolving Data Streams. Expert Systems with Applications. 196. 116630. 10.1016/j.eswa.2022.116630.

Bodaghi, A.,; Teimourpour, B., (2018). The detection of professional fraud in automobile insurance using social network analysis, arXiv preprint arXiv:1805.09741 (37 Pages).

Ghahramani, S., (2005). Fundamentals of probability with stochastic processes.—3rd edition. Pearson/Prentice Hall p. cm. Includes Index. ISBN: 0-13-145340-8 (624 Pages).

Nian, K.; Zhang, H.; Tayal, A.; Coleman, T.; Li, Y., (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58–75. 16 A (18 Pages).

Noble, C. C.; Cook, D. J., (2003). Graph-based anomaly detection. In Proceedings of the ACM SIGKDD international conference on knowledge discovery and datamining, 631– 636 (6 Pages).

Oskarsdottir, M.; Ahmed, W.; Antonio, K.; Baesens, B.; Dendievel, R.; Donas, T.; Reynkens, T., (2022). Social network analytics for supervised fraud detection in insurance, https://doi.org/10.48550/arXiv.2009.08313 (19 Pages).

Pinheiro, C., (2011). Highlighting unusual behavior in insurance based on social network analysis, AR SAS Institute Inc, Oi, Rio de Janeiro, Brazil.

Pourhabibi, T.; Ong, K.L.; Kam, B.; Boo, Y.L., (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems. 133(4):113303 (15 Pages).

Rajan, R. S.; Shantrinal, A. A.; Kumar, K. J.; Rajalaxmi, T.; Fan, J.; Fan, W., (2019). Embedding complete multi-partite graphs into cartesian product of paths and cycles, arXiv preprint arXiv:1901.07717 (19 Pages).

Šubelj, L.; Furlan, Š.; Bajec, M., (2011). An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems with Applications, 38(1), 1039–1052 (14 Pages).

Tarawneh, A.; Hassanat, A.; Altarawneh, G.; Almuhaimeed, A., (2022). Stop Oversampling for Class Imbalance Learning: A Review. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3169512 (18 Pages).

West, D.B., (2001). Introduction to Graph Theory. 2nd Edition, Prentice-Hall, Inc., Upper Saddle River.

White Paper., (2011). The insurance fraud race: Using information and analytics to stay ahead of criminals. Featuring as an example: SAS Fraud Framework for Insurance.

Zhou, J.; Xu, W.; Guo, X.; Ding, J., (2015). A method for modeling and analysis of directed weighted accident causation network (DWACN). Physica A: Statistical Mechanics and Its Applications, 437, 263–277 (15 Pages).

Zhai, J.; Qi, J.; Shen, C., (2021). Binary Imbalanced Data Classification Based on Diversity Oversampling by Generative Models. Information Sciences. 585. 10.1016/j.ins.2021.11.058 (31 Pages).