Iranian Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Iranian Insurance Research Center (IIRC)

ORIGINAL RESEARCH PAPER

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

Asma Hamzeh[1,*], Mohammad Javad Nadjafi-Arani[2]

[1] *Assistant professor, Department of New Insurance Technologies, Insurance Research Institute, Tehran, Iran*

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

ARTICLE INFO

Article History:
Received: 26 October 2024
Revised: 25 December 2024
Accepted: 11 March 2025
Early online access: 15 March 2025
Published: 1 July 2025

Keywords:
Car insurance
Graph theory
labeling
Poisson distribution

*Corresponding Author:
Email: *Hamzeh@irc.ac.ir*
Phone: +9821 22084084
ORCID: 0000-0001-7736-7626

DOI: 10.22056/ijir.2025.03.03

ABSTRACT

BACKGROUND AND OBJECTIVES: Insurance fraud presents a persistent challenge within the insurance industry, leading to substantial financial losses and eroding public trust. Financial institutions are actively seeking accurate methods to identify the activities of fraudsters and scammers. 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 rampant. Fabricated accidents, especially those involving groups, staged injuries, and orchestrated scenes, are among the common fraudulent practices in this realm. Opportunistic fraud is typically committed by an individual who simply seizes an opportunity to inflate a claim or receive an exaggerated estimate for damages or repairs from their insurance companies. In contrast, professional fraud is often carried out by organized groups. These rings typically target multiple fake identities, organizations, or even brands. These criminal networks frequently rely on insiders to help them defraud companies, simultaneously using various schemes. Although the amounts involved in professional fraud cases are much larger, they occur less frequently than opportunistic insurance fraud. Combating insurance fraud is a challenging issue. Most traditional systems can detect opportunistic fraud; however, due to the significant financial losses involved, insurance companies are particularly focused on identifying organized fraud rings. Consequently, insurers need to adopt advanced technologies and sophisticated systems to effectively address this problem.

METHODS: Network analysis is a valuable technique for fraud detection, enabling the evaluation of communications among individuals and entities (both real and legal) to uncover new dimensions of these interactions. This paper introduces a mathematical model, based on graph theory, to identify suspicious clusters associated with organized fraud. A network, termed the "accident network," was first introduced using graph theory in this research. This network demonstrates characteristics of a random graph. Subsequently, suspicious clusters within this network are identified using a graph theory-based algorithm. The occurrence probability of such clusters in a random accident network is then examined by defining a binomial distribution over its edges.

FINDINGS: This process assigns a label (indicating fraudulent or non-fraudulent) to each accident and individual involved. Given the algorithm's structure and complexity, the proposed method is capable of efficiently analyzing large datasets.

CONCLUSION: Insurance fraud is an act committed to defraud insurers for financial gain. Insurance fraud has existed since the formation of commercial enterprises and has so far imposed billions of dollars in costs on insurance companies annually. Insurance fraud comes in various forms and occurs in all insurance domains, covering a wide range of claims from exaggerated ones to fabricated accidents and damages.Auto insurance fraud, particularly the organized fraud studied in this research, is often carried out through group structures. This structure leads to significant cost increases for insurers and consequently higher insurance premiums. Today, given the necessity of fraud detection in various fields, data mining and machine learning techniques such as artificial neural networks, fuzzy logic, and genetic algorithms have become common tools for fraud detection due to their high capabilities in modeling and navigating complex problems.Another tool used for detecting organized fraud is graph theory. In this approach, the problem is first mathematically modeled. This means the accident network is first modeled as a graph, and then organized fraud is detected using available tools. Then, computer science concepts are utilized to more precisely identify networks suspected of fraud. More accurately, in structures like a country's accident data, the amount of available data is very large finding relationships among them is quite difficult.While using tools like data mining, machine learning, neural networks, fuzzy logic, genetic algorithms, etc., whose main purpose is to find relationships among data, is very useful, they have some shortcomings. These tools, if meta-heuristic algorithms are used, will have inaccuracies or overfitting in imbalanced data. In heuristic algorithms, finding relationships among large amounts of data has very high computational complexity, which in some cases may take weeks or more to execute.In this research, the researchers have tried to address this shortcoming using mathematical models while accurately examining the probability of suspicious events. Therefore, in this research, first, the accident network was modeled using graph theory, and then it was shown that this model is a random process, and the presence of regular elements in the model indicates sets of vehicles suspected of fraud. Subsequently, based on an algorithm for finding suspicious subgraphs written as an m-file script in MATLAB, suspicious vehicles were extracted from all vehicles.Finally, it was proven that the accident network is a Poisson process, and its occurrence probability can be determined. This reasoning, based on graph modeling structure, helps assign a credibility degree to each accident and each vehicle regarding suspicion of fraud. For future research, it is suggested that a more extensive network, including all stakeholders in organized fraud, should be created and examined. More precisely, the network should examine the label assignment of main beneficiaries who profit from an accident based on their profit shares. Specifically, future work should investigate labeling main beneficiaries based on their profit shares from an accident, enabling insurers to adopt tailored policies for various stakeholders (e.g., policyholders, vehicle occupants, repair shops) to reduce financial losses and restore public trust.

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله پژوهشی

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

اطلاعات مقاله

تاریخ های مقاله:

تاریخ دریافت: ۵ آبان ۱۴۰۳

تاریخ داوری: ۵ دی ۱۴۰۳

تاریخ پذیرش: ۲۵ اسفند ۱۴۰۳

تاریخ زودآیند: ۲۵ اسفند ۱۴۰۳

تاریخ انتشار: ۱۰ تیر ۱۴۰۴

کلمات کلیدی:

بیمه خودرو

برچسب‌گذاری

توزیع پواسون

نظریهٔ گراف

*نویسنده مسئول:

ایمیل: Hamzeh@irc.ac.ir

تلفن: ۲۲۰۸۴۰۸۴ ۹۸۲۱+

ORCID: 0000-0001-7736-7626

DOI: 10.22056/ijir.2025.03.03

چکیده:

پیشینه و اهداف: تقلب در صنعت بیمه یکی از مشکلات رایج در این حوزه است که موجب خسارات سنگینی، چه به‌لحاظ منافع مادی و چه به‌لحاظ اعتماد عمومی در این صنعت می‌شود. مؤسسات مالی و پولی به‌شدت در پی شناخت دقیق فعالیت‌های کلاهبرداران و متقلبان هستند. این امر به‌دلیل اثر مستقیم آن بر خدمت‌رسانی به مشتریان مؤسسات، به کاهش هزینه‌های عملیاتی، جلب اعتماد سایر بیمه‌گذاران و حفظ و ارتقای سهم بازار بیمه‌گران به‌عنوان ارائه‌دهندگان خدمات مالی قابل اطمینان منجر خواهد شد. یکی از رایج‌ترین تخلفات، تقلب‌های سازمان‌یافته و فرصت‌طلبانه در بیمهٔ خودرو است. تصادفات عمدی به‌ویژه در قالب گروهی، صدمه دیدن افراد توسط وسیلهٔ نقلیه یا صحنه‌سازی از جمله تقلب‌های رایج در این حوزه هستند. هدف این مقاله معرفی مدل ریاضی مبتنی بر نظریهٔ گراف (شبکه) برای شناسایی خوشه‌های مشکوک برای تقلب‌های سازمان‌یافته است.

روش‌شناسی: یکی از روش‌هایی که برای شناسایی تقلب کاربرد دارد، تحلیل شبکه است. در تحلیل شبکه ارتباطات بین افراد و شخصیت‌های حقیقی و حقوقی مختلف ارزیابی و ابعاد جدیدی از این ارتباطات شناسایی می‌شود. در این پژوهش، ابتدا با استفاده از نظریهٔ گراف، شبکه‌ای به نام شبکهٔ تصادفات معرفی می‌شود. سپس نشان داده می‌شود که شبکهٔ حاصل از تصادفات خودروها یک فرایند تصادفی است. سپس در شبکهٔ ساخته‌شده 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 نظر متقلب بودن یا نبودن به هر تصادف و به هر فرد می‌شود. با توجه به ساختار الگوریتم و پیچیدگی آن می‌توان نتیجه گرفت الگوریتم پیشنهادی به‌سادگی قادر به تحلیل داده‌های بسیار زیاد است.

نتیجه‌گیری: بررسی این موضوع موجب می‌شود که 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.

توجه: مدت‌زمان بحث و انتقاد برای این مقاله تا ۱ اکتبر ۲۰۲۵ در وب‌سایت IJIR در «نمایش مقاله» باز است.

مقدمه

تقلب‌های بیمه‌ای انواع گوناگونی دارد و در تمام حوزه‌های بیمه‌ای رخ می‌دهد و طیف گسترده‌ای از ادعاهای اغراق‌آمیز تا تصادف‌ها و خسارت‌های تعمدی را دربرمی‌گیرد. این تقلب‌ها سبب افزایش هزینه‌ها و در پی آن، افزایش مبلغ حق‌بیمه می‌شوند. ازاین‌رو به ضرر سایر بیمه‌گذاران نیز خواهد بود. با وجود پیشرفت‌های فراوان در شناسایی این تقلب‌ها، هزینه‌های ایجادشده 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 به‌نوعی ایمن کند. با توجه به حجم و نوع داده‌ها، روش‌های گوناگونی برای کشف تقلب‌های بیمه‌ای وجود دارند. استفاده از هریک از مدل‌ها برای شناسایی تقلب، این امکان را به متخصصان شرکت‌های بیمه می‌دهد که با صرف زمان و هزینهٔ کمتری تشخیص دهند که ادعای خسارت اعلام‌شده مشکوک به تقلب است یا خیر. اگرچه این روش‌ها بسیار سودمندند، اما محققان به‌علت وجود داده‌های زیاد غالباً مجبور به استفاده از الگوریتم‌های فرااکتشافی مانند الگوریتم ژنتیک و شبکه‌های عصبی می‌شوند. استفاده از این‌گونه ابزارها، به‌ویژه الگوریتم‌های فرااکتشافی متنوعی که در هریک از روش‌های فوق استفاده می‌شود، نقایصی جدی دارند. نقص اصلی این‌گونه ابزارها آن است که یافتن رابطه بین تعداد زیادی دادهٔ نامتعادل[1] نمی‌تواند پاسخ روشنی را به مسئله ارائه دهد. به‌طور دقیق‌تر، در استفاده از این‌گونه ابزارها ابتدا به داده‌هایی که افراد یا خودروهای متقلب را از غیرمتقلب جدا کنند، لازم است که الگوریتم بتواند یادگیری را بر روی آن‌ها انجام دهد و سپس به مرحلهٔ آزمایش برسیم. اما ازآنجاکه داده‌ها نامتعادل‌اند، یعنی تعداد زیادی داده داریم که غیرمتقلب‌اند و تعداد داده‌های متقلب بسیار اندک‌اند، لذا یادگیری الگوریتم داده‌های شرکت بیمه که افراد متقلب را از غیرمتقلب جدا می‌کند دقت بسیار پایینی خواهد داشت. محققان برای رفع این مشکل از روش‌های تحدید فضای نمونه استفاده می‌کنند. به‌عبارت‌دیگر محققان تعداد داده‌هایی را که غیرمتقلب‌اند برای یادگیری کاهش می‌دهند یا تعداد داده‌هایی را که متقلب‌اند در یادگیری افزایش می‌دهند، به این اعمال به‌ترتیب کم‌نمونه‌گیری[2] یا بیش‌نمونه‌گیری[3] گویند (Zhai et al., 2022; Tarawneh et al., 2022; Bernardo & Della valle, 2022). مشکل اصلی بیش‌نمونه‌گیری آن است که اصطلاحاً در الگوریتم، بیش‌برازش[4] رخ می‌دهد و کم‌نمونه‌گیری باعث از دست دادن اطلاعات و لذا کاهش دقت می‌شود.

نکتهٔ دوم آنکه در اغلب مقالات، به‌ویژه مقالاتی که از هوش مصنوعی استفاده می‌شود، به داده‌هایی نیاز داریم که بیمه‌گر برچسب تقلب به آن‌ها تخصیص دهد. اما غالب داده‌های موجود شرکت‌های بیمه این نوع برچسب‌گذاری را ندارند، حتی اگر هم تعداد محدودی این برچسب‌گذاری را انجام دهند در موارد کاملاً اثبات‌شده خواهند بود که بسیار محدود است. این امر باعث دقت پایین این‌گونه الگوریتم‌ها در قسمت کاربرد می‌شود.

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

در این پژوهش سعی شده است این نقیصه‌ها با استفاده از مدل‌های ریاضی مبتنی بر شبکه و نظریهٔ گراف رفع شود. لذا در این پژوهش با استفاده از نظریهٔ گراف به مدل‌سازی شبکهٔ تصادفات پرداخته می‌شود. سپس الگوریتمی اکتشافی پیشنهاد می‌شود که مشکل نقیصهٔ 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 به پیچیدگی آن نیز پرداخته شده است، می‌توان نتیجه گرفت الگوریتم پیشنهادی به‌سادگی قادر به تحلیل داده‌های بسیار زیاد است.

مبانی نظری پژوهش

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 به از بین رفتن باور و اعتماد عملکردی در بین ارکان صنعت بیمه می‌انجامد، تقلب در صنعت بیمه محسوب می‌شود. تقلب و کلاهبرداری در صنعت بیمه بسیار متنوع است و به‌صورت 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 را به‌طور معمول شخصی انجام می‌دهد که به‌سادگی فرصتی برای افزایش قیمت یک خسارت یا دریافت تخمینی مبالغه‌شده‌ای برای خسارت یا تعمیرات از شرکت بیمه خود دارد، درحالی‌که کلاهبرداری حرفه‌ای را غالباً گروه‌های سازمان‌یافته انجام می‌دهند. 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 با استفاده از برخی راه‌ها به‌طور 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 (SAS Institute Inc., 2011). مبارزه با تقلب بیمه‌ای مشکلی چالش‌برانگیز است. بیشتر سیستم‌های سنتی قادر به یافتن کلاهبرداری‌های فرصت‌طلبانه هستند، اگرچه شرکت‌های بیمه به‌دلیل یادشده (بیشترین ضرر مالی) علاقهٔ فراوانی به شناسایی گروه‌های سازمان‌یافته دارند. در نتیجه شرکت‌های بیمه برای مقابله با این مشکل باید از فناوری‌های مدرن و سیستم‌های هوشمند استفاده

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

کنند. هدف این پژوهش استفاده از مدل‌های ریاضی مبتنی بر شبکه و نظریهٔ گراف برای کشف تقلب است. لذا در ادامه مفاهیم مورد نیاز 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 می‌دهند و به آن گراف ساده می‌گویند. 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 هریک از یال‌ها یا به هریک از رأس‌های آن عددی نسبت داده شده باشد. این اعداد وزن یال یا رأس نامیده می‌شوند. وزن یال می‌تواند نشان‌دهندهٔ هزینه، مسافت، زمان یا هر مشخصهٔ دیگری از یال باشد. تعداد رأس‌های گراف G یعنی $|V(G)|$ را مرتبهٔ گراف گویند که با $n(G)$ نمایش داده می‌شود و تعداد یال‌های گراف یعنی $|E(G)|$ را اندازهٔ گراف G گویند که با $m(G)$ بیان می‌شود. به‌طور معمول برای سهولت در کار و در صورت مشخص بودن گراف G به‌جای $n(G)$ از n و به‌جای $m(G)$ از m استفاده می‌شود. درجهٔ رأس v در گراف G برابر با تعداد یال‌هایی از گراف G است که به رأس v متصل‌اند و آن را با $deg_G(v)$ یا به‌طور ساده‌تر با $deg(v)$ یا $d(v)$ نمایش می‌دهند. دو رأس u و vرا دو رأس همسایه یا مجاور گویند، هرگاه توسط یالی به هم متصل شده باشند، به‌عبارت‌دیگر $uv\in E(G)$ برقرار باشد. یک زیرگراف از گراف G گرافی است که مجموعه رأس‌های آن زیرمجموعه‌ای از مجموعه رأس‌های گراف Gو مجموعه یال‌های آن زیرمجموعه‌ای از مجموعه یال‌های G باشد. زیرگراف H از G را القایی 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 درجهٔ تمام رأس‌های آن با هم مساوی است، منظم گویند و اگر به‌ازای 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 گویند. گراف کامل n رأسی با K_n نمایش داده می‌شود. اگر u و vدو 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، به‌طوری 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. به‌عبارت‌دیگر، ابتدا و انتهای این مسیر رأس‌های یکسانی هستند. 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 را که درجه یک رأس $n-1$ و درجه مابقی رأس‌های آن یک باشد، گراف ستاره گویند. یک برش یا مجموعه جداکننده گراف همبند G، مجموعه‌ای از رأس‌هاست 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 رأس‌هاست 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 به‌استثنای گراف‌های کامل، $\kappa(G)$ به‌ازای 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. برای 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 یال‌هاست که حذف آن‌ها، گراف را ناهمبند می‌کند. عدد همبند یالی $\kappa'(G)$ برابر با اندازهٔ کوچک‌ترین مجموعه یال‌های برشی است و عدد همبند یالی محلی $\kappa'(u,v)$ رأس‌های u و v برابر با اندازهٔ 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 (به‌جز رأس‌های u و v). قضیهٔ مشهور منگر بیان می‌کند که عدد همبندی (عدد همبند یالی) یک گراف برابر تعداد مسیرهای مجزای درونی (یالی) بین رأس‌های مشخص است (West, 2001).

فرایندهای تصادفی

در این بخش به مفاهیم فرایندهای تصادفی می‌پردازیم که نقش عمده‌ای در نتایج اصلی این پژوهش دارند. مرجع اصلی مطالب این بخش، کتاب (Ghahramani, 2005) است. برای مطالعهٔ پدیده‌های دنیای واقعی که در آن سیستم‌ها به‌صورت تصادفی عمل می‌کنند به‌جای مدل‌های قطعی به مدل‌های احتمالی نیاز است. پایه و اساس مدل‌های احتمالی را فرایندهای تصادفی می‌گویند.

در آمار و احتمال، متغیر تصادفی یا ورتنده کاتوره‌ای متغیری است که مقدار آن از اندازه‌گیری برخی از انواع فرایندهای کاتوره‌ای به دست 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 است. تابع جرم احتمال یک متغیر تصادفی به‌صورت رابطهٔ (۱) تعریف می‌شود:

$$f_X(x)=p(X=x)=p(x) \qquad (۱)$$

متغیرهای تصادفی نقشی اساسی در این پژوهش دارند. در ادامهٔ این بخش، متغیرهای تصادفی و توابع توزیع احتمال که در این پژوهش به آن‌ها اشاره می‌شود، به‌طور مختصر شرح داده خواهد شد.

متغیرهای تصادفی برنولی از جمله ساده‌ترین نوع متغیرهای تصادفی‌اند که تنها شامل دو برآمد موفقیت و شکست هستند که به‌طور معمول به‌ترتیب آن‌ها را با s و f نمایش می‌دهند. به‌عبارت دقیق‌تر، متغیر تصادفی X که به‌صورت $X(s)=1$ 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 به‌صورت رابطهٔ (۲) است (Ghahramani, 2005).

$$p(x)=\begin{cases} p & if\ x=s \\ 1-p & if\ x=f \\ 0 & o.w \end{cases} \qquad (۲)$$

اگر n امتحان برنولی همه با امتحان موفقیت p به‌طور مستقل انجام شوند، آنگاه X، تعداد موفقیت‌های این امتحان، متغیر تصادفی جدیدی به نام متغیر تصادفی دوجمله‌ای را تشکیل 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 رابطهٔ (۳) تعریف می‌شود.

$$p(x)=\begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x=0,1,\ldots,n \\ 0 & o.w \end{cases} \qquad (۳)$$

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 رابطهٔ (۴) تعریف شده‌اند.

$$E(X)=np, \qquad Var(X)=np(1-p) \qquad (۴)$$

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

فرایند 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 ناممکن است، زیرا برای مقادیر نه‌چندان 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 از جمله این روش‌ها تقریب پواسون است. به‌عبارت 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. به‌عبارت‌دیگر، توزیع پواسون همواره به‌عنوان تقریبی برای توزیع دوجمله‌ای است. این مفهوم را می‌توان به‌صورت زیر تعریف کرد.

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

$$P(X=n)=\frac{(\lambda)^n e^{-\lambda}}{n!}. \qquad n=0,1,2,\ldots \qquad (۵)$$

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

$$E(X)=Var(X)=\lambda \qquad (۶)$$

که در آن $E(X)$ نشان‌دهندهٔ امید ریاضی X و $Var(X)$ نمایش‌دهندهٔ واریانس آن است.

توزیع پواسون به‌خودی‌خود و جدا از تقریبی برای توزیع دوجمله‌ای، اغلب در رابطه با مطالعهٔ دنبالهٔ 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. بدین‌ترتیب برای هر مقدار t عددی مانند $N(t)$ به دست می‌آید که برابر با تعداد پیشامدهایی است که در بازهٔ $[0,t]$ رخ می‌دهند. برای مطالعهٔ توزیع $N(t)$ سه فرض ساده و طبیعی مانایی، استقلال خطی و تناسب خطی در نظر گرفته می‌شود (Ghahramani, 2005).

به‌عبارتی، فرایند $\{N(t):t\ge 0\}$ را فرایند پواسون با پارامتر $\lambda(>0)$ گوییم، اگر

۱- $N(0)\equiv 0$.

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

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

۴- به‌ازای هر $t>s\ge 0$ نمو $N(t)-N(s)$ به‌عنوان یک متغیر تصادفی دارای توزیع پواسون با پارامتر $\lambda(t-s)$ باشد.

مروری بر پیشینهٔ پژوهش

روش‌های مختلفی برای کشف تقلب وجود دارد که تمام این روش‌ها در مطالعات مختلف استفاده شده که در این بخش از پژوهش به‌طور مختصر به بررسی آن‌ها پرداخته می‌شود.

اکثر سیستم‌های موجود تشخیص ناهنجاری تقلب، برای شناسایی گروه‌های مشکوک به کار برده می‌شوند [برای نمونه، مرجع (Nian et al., 2016) را ببینید]. ادبیات مربوط به کشف تقلب سازمان‌یافته بسیار پراکنده است. مثلاً، آن‌ها می‌توانند شامل تجزیه‌وتحلیل PRIDIT باشند که براکت و لوین پیشنهاد کرده‌اند (Brockett & Levine, 1977). در واقع این روش‌ها براساس امتیازات RIDIT و تجزیه‌وتحلیل مؤلفهٔ اصلی بنا نهاده شده‌اند که سوبلج و همکاران برای تشخیص مؤلفه‌های مشکوک از آن‌ها استفاده کرده‌اند. نویسندگان در این مقاله یک سیستم خبره را برای شناسایی و بررسی گروه‌های کلاهبرداری بیمهٔ 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. علاوه‌بر ویژگی‌های ذاتی موجودیت‌ها، الگوریتم روابط بین آن‌ها را نیز بررسی می‌کند (Šubelj et al., 2011). در پژوهش بداغی و تیمورپور، یک سیستم خودکار برای شناسایی گروه‌های مجرم در بیمهٔ 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 شبکهٔ تصادفات به‌منظور شناسایی خوشه‌های متقلب سازمان‌یافته معرفی می‌کنند. سپس، در بخش ارزیابی با سیستم نمونهٔ اولیه، روش پیشنهادی براساس داده‌های دنیای واقعی ارزیابی و نتیجه‌گیری می‌شود (Bodaghi & Teimourpour, 2018). در پژوهش نیان و همکاران یک روش رتبه‌بندی برای تشخیص ناهنجاری تقلب با در نظر گرفتن رتبه‌بندی طبقاتی نادر ارائه شده است که در آن ناهنجاری با توجه به طبقهٔ اکثریت واحد و رتبه‌بندی ناهنجاری با توجه به بیش از یک الگوی اصلی ارزیابی می‌شود (Nian et al., 2016). در پژوهش ژاو و همکاران نویسندگان روشی برای مدل‌سازی تصادفات به‌صورت شبکه‌ای وزن‌دار بر مبنای علت تصادف ارائه کرده‌اند. با توجه به مقالهٔ این نویسندگان، نظریهٔ تحلیل شبکه‌های پیچیده برای درک و تحلیل علت حوادث در سیستم‌های پیچیده مورد توجه قرار گرفت. آن‌ها روش جدیدی برای ایجاد شبکهٔ وزن‌دار جهت‌دار (DWACN) برحسب علت تصادف برای شعبه‌های مورد مطالعه معرفی کردند و علت تصادف را براساس داده‌های کشور انگلستان بررسی کردند (Zhou et al., 2015). در مطالعهٔ نوبل و کوک، 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 فرایندهای تصادفی را برای شناسایی ناهنجاری‌ها ارائه کرده‌اند، این در حالی است که این‌گونه رویکردها به‌طور عمده فقط بر ویژگی‌های رابطه‌ای گره‌ها تمرکز دارند (Noble & Cook, 2003).

اوشکارسدوتیر و همکاران در مقالهٔ خود، با پیوند دادن ادعاها با همهٔ طرف‌های درگیر، از جمله بیمه‌شدگان، کارگزاران، کارشناسان و تعمیرکاران، یک شبکه تشکیل داده‌اند. سپس آن‌ها کلاهبرداری را به‌عنوان پدیده‌ای اجتماعی در شبکه ایجاد کرده و از الگوریتم Bi Rank و بردار جست‌وجوی خاص کلاهبرداری برای محاسبهٔ امتیاز تقلب برای هر ادعا استفاده کرده‌اند. در نهایت، با استفاده از ویژگی‌های شبکه، مدل نظارت‌شده‌ای را برای کشف تقلب در بیمهٔ خودرو 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 به‌دست‌آمده، ادعاهای خسارت مشکوک را که نیاز به بررسی بیشتر دارند، نشان می‌دهد (Óskarsdóttir et al., 2022).

همچنین مقالهٔ مروری پورحبیبی و همکاران و مقالهٔ راجان و همکاران از جمله مقالاتی هستند که با دیدگاه نظریهٔ گراف و شبکه‌ها به موضوع تقلب در صنعت بیمه پرداخته‌اند که این بخش از ریاضیات تمرکز اصلی محققان در این پژوهش است. همان‌طور که پیش از این ذکر شد، هدف این پژوهش یافتن خوشه‌های ناهنجاری شبکه بوده تا با استفاده از این موضوع خوشه‌های مشکوک مشخص شوند. این روش می‌تواند علاوه‌بر دربرگرفتن کمبودهای برخی مقالات بیان‌شده در بالا، راه‌کار جدیدی را نیز در جهت یافتن ناهنجاری‌های مشکوک به تقلب‌های سازمان‌یافته ارائه دهد (Rajan et al., 2021; Pourhabibi et al., 2020).

روش‌شناسی پژوهش

یکی از روش‌هایی که برای شناسایی تقلب کاربرد دارد، تحلیل شبکه است. در تحلیل شبکهٔ ارتباطات بین افراد و شخصیت‌های حقیقی و حقوقی مختلف ارزیابی 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 به‌واسطهٔ لینک‌های ارتباطی با یکدیگر متصل می‌شوند. در عمل شبکه‌ها از تعداد بسیار زیادی گره با ارتباطات بسیار زیاد شکل گرفته است. شبکه‌ها طبیعی‌ترین نمایش چنین حوزهٔ رابطه‌ای هستند که امکان فرمول‌بندی و تجزیه‌وتحلیل روابط پیچیده بین موجودیت‌ها را فراهم می‌کنند.

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

نتایج و بحث

در این بخش ابتدا نشان داده می‌شود که شبکهٔ حاصل از تصادفات خودروها شرایط مانایی و مستقل افزایشی و تناسب خطی را داراست و لذا نمایش یک فرایند پواسون مبتنی بر زمان است. سپس در شبکهٔ ساخته‌شده از تصادفات، مجموعه خودروهای مشکوک که در این ساختار تصادفی ایجاد نظم می‌کنند، شناسایی می‌شوند. مجموعه‌ای از خودروها را در یک محدودهٔ جغرافیایی در نظر بگیرید. این خودروها در کل بازهٔ زمانی که بررسی شده است، در این محدودهٔ جغرافیایی در حال ترددند. 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 تصادف‌های رخداده در بازهٔ زمانی $(t,t+s)$ مستقل از تصادفات رخداده در قبل یا بعد از این زمان است (استقلال افزایشی). تعداد رخداد چند تصادف دقیقاً در یک زمان احتمال نزدیک به صفر دارد، به‌عبارت‌دیگر دو تصادف هم‌زمان رخ نمی‌دهند (تناسب خطی). لذا بنا بر تعریف ۱ تعداد تصادفات جاده‌ای در یک مکان مشخص یک فرایند پواسون است.

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

یک شبکهٔ تصادفات را به این صورت در نظر بگیرید، به‌طوری که مجموعهٔ رئوس شامل تمامی خودروهای دارای بیمه (ثالث/بدنه) است و مجموعهٔ یال‌ها شامل تصادفات بین خودروهاست. به‌عبارت دقیق‌تر، دو رأس (خودرو) a و b با هم مجاورند اگر و تنها اگر این دو با هم تصادف کرده باشند. ابتدا قسمت‌های محدودی از این شبکه بررسی می‌شود که ساختار منظم داشته باشد. توجه داشته باشید که شبکهٔ تصادفات احتمالاً شامل دورهایی با تنها سه گره است که برخورد سه خودرو به یکدیگر را نشان می‌دهد. ممکن است برخی یا اغلب این‌گونه تصادفات به‌صورت مصنوعی رخ دهند که برای بررسی گروه‌های سازمان‌یافته مناسب نیستند. بنابراین، این پژوهش فقط بر 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، قابل توجه باشد.

با توجه به آنچه دربارهٔ تصادفی بودن برخورد بین خودروها ذکر شد، از نظر نویسندگان ساختار منظم در شبکهٔ 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 بررسی می‌کنند آن است که به‌طور مثال در یک دور یا یک دور وتردار تعدادی تصادف پی‌درپی رخ داده است که خودروی اول و آخر تصادف یکسان هستند، 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 یا کامل باشد، عدد همبندی زیرگراف افزایش می‌یابد و لذا درصد همبستگی بین رئوس آن بیشتر خواهد شد؛ بنابراین مشکوکیت بیشتری خواهد داشت. این‌گونه زیرگراف‌های مشتق‌شده 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 واقع گراف ستاره نمایش خودرویی است که با چندین خودرو تصادف کرده است و از جهت انفرادی بررسی خواهد شد. برای شناسایی این ساختارهای مشکوک به الگوریتم‌هایی نیاز است که در بخش بعدی توضیح داده خواهند شد.

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 شبکهٔ 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 تعریف‌شده در این بخش، از روش زیر استفاده می‌شود.

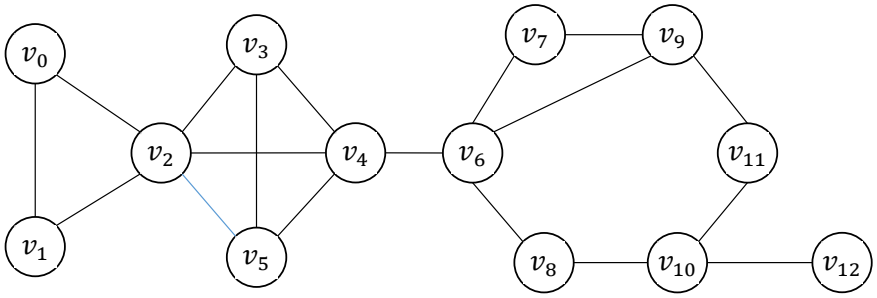
فرض کنید $e=uv$ یال دلخواهی باشد، به‌طوری که درجهٔ رئوس هر دو سر یال حداقل برابر با سه باشد و با حذف یال e گراف همچنان همبند باقی بماند. انتخاب یال با این ویژگی‌ها نشان می‌دهد که یال الزاماً درون حداقل یک دور قرار می‌گیرد، زیرا با حذف یال e گراف همبند باقی می‌ماند، یعنی مسیر دیگری از u به v در گراف به 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 یادشده شامل دو دور کوچک‌تر است که یال f را در خود دارند. حال ازآنجاکه

```
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
```

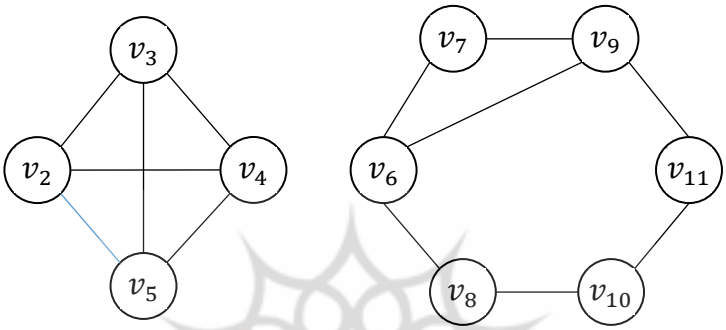
شکل ۱. الگوریتم شبه کد روش پیشنهادی
Fig 1. Pseudo-code algorithm of the proposed method

درجهٔ رأس u و v در یال انتخابی e هر دو بالاتر از سه هستند، لذا اگر یال e وتر یک دور باشد خودبه‌خود دور C_2 موجود خواهد بود که $C_1 \cup C_2 - e$ یک دور بزرگ‌تر مانند C را نمایش می‌دهد. با یافتن تمامی مسیرهایی که u را به v متصل می‌کند و با تکرار این عمل روی یال‌هایی که در الگوریتم بررسی می‌شوند، می‌توان تمام دورها، دورهای دارای وتر، زیرگراف‌های منظم و زیرگراف‌های کامل حداقل چهار رأسی را به دست آورد. برای فهرست 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 تاکنون ملاقات نشده‌اند، در نظر گرفته می‌شوند. سپس برای ایجاد مسیرها، از رأس u به یکی از رأس‌های مجاور آن مراجعه می‌شود، به‌طوری که این رأس تاکنون 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 مجاور دیده‌نشده نداشته باشد با بازگشت به عقب (backtrack) به اولین رأسی بازگردید که شرایط ادامهٔ مسئله را داشته باشد (یعنی دارای رأس مجاوری باشد که تاکنون 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 فرایند یافتن مسیرهای بین u و v (یعنی زمانی که مسیر دیگری بین آن‌ها وجود نداشته باشد)، تمام رأس‌های موجود در مجموعهٔ $Con(u,v)$ به‌عنوان ملاقات‌شده علامت زده می‌شود. این امر سبب می‌شود مسیرهای تکراری در فهرست 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 به‌صورت یک اسکریپت در محیط محاسباتی متلب نوشته شده و شبه کد آن در شکل ۱ بیان شده است. حال به‌منظور بررسی اجرای الگوریتم یک مثال ساده بررسی می‌شود.

مثال ۱. گراف سادهٔ زیر (شکل ۲) را به‌عنوان یک شبکهٔ تصادف



شکل ۲. یک شبکهٔ ساده تصادفات
Fig 2. A simple network of accidents



شکل ۳. زیرگراف‌های مشکوک به تقلب
Fig 3. Fraud-suspected subgraphs

بسیار کوچک در نظر بگیرید.

ابتدا الگوریتم یال V_2V_3 را به‌عنوان یالی در نظر می‌گیرد که رئوس دو طرف آن از درجه بزرگ‌تر یا مساوی سه است. سپس الگوریتم بررسی می‌کند که یال مورد نظر برشی نیست. آنگاه مسیرهای $V_2V_4V_3$ و $V_2V_5V_3$ و $V_2V_5V_4V_3$ به‌عنوان مسیرهای دیگری که دو رأس V_2 و V_3 را به یکدیگر متصل می‌کنند توسط الگوریتم مشخص می‌شوند. تمامی رئوسی که در این مسیرها قرار دارند مشکوک‌اند. به‌عبارت‌دیگر، زیرگراف القایی حاصل از مجموعه رئوس $\{V_2,V_3,V_4,V_5\}$ یک زیرگراف مشکوک در شبکهٔ تصادفات است که احتمالاً درصد مشکوکیت هر رأس آن متفاوت با رأس دیگری است. سپس الگوریتم یال V_4V_6 را انتخاب می‌کند که درجهٔ دو سر آن بزرگ‌تر یا مساوی سه است و ازآنجاکه مسیر دیگری این دو یال را به یکدیگر متصل نمی‌کند از آن گذر کرده و یال V_6V_7 را به‌عنوان یال پایه در نظر می‌گیرد و مسیرهای $V_6V_9V_7$ و $V_6V_8V_{10}V_{11}V_9V_7$ به‌عنوان مسیرهای مشکوک توسط الگوریتم شناسایی می‌شوند. همان‌طور که در مثال مشخص شد، زیرگراف‌های شکل ۳ به‌عنوان زیرگراف‌های مشکوک در این گراف توسط الگوریتم شناسایی می‌شوند.

دقت کنید در این گراف در حال حاضر دور سه‌تایی $V_0V_1V_2$ به‌عنوان زیرشبکهٔ مشکوک قرار نمی‌گیرد، مگر آنکه یک یال دیگری بین رئوس همین دور سه‌تایی یا زیرگراف کامل چهاررأسی مجاور آن به وجود بیاید.

در ادامه به مقایسهٔ ویژگی‌های الگوریتم معرفی‌شده در این مقاله با مقالات دیگر و وجه تمایز آن می‌پردازیم.

آنالیز بخش اول الگوریتم

شایان ذکر است که در فرایند این الگوریتم تمامی دورها پیدا نمی‌شوند، مثلاً دوری به طول چهار از نظر نویسندگان مشکوک نیست، اما اگر دور به طول چهار دارای وتر باشد یا حداقل دو خودرویی که پیش از این با هم تصادف کرده‌اند با خودروهای دیگری به‌جز خودروهای مورد بررسی در دور نیز برخورد داشته باشند مشکوک تلقی شده و بررسی خواهند شد. در واقع نویسندگان بنا بر این اصل چنین موضوع را در نظر گرفته‌اند که احتمال وجود دور تنها در تصادفات زنجیری زیاد است و اگر بخواهد این دور نشان‌دهندهٔ یک تخلف باشد، آنگاه الزاماً متخلفان برای ادامهٔ کار خود این دور را وتردار خواهند کرد یا متخلفان پیش از این، تصادفات متخلفانهٔ دیگری داشته‌اند. از نظر نویسندگان این نگاه می‌تواند معیار دقیق‌تری را نسبت به انتخاب دور تنها که در مقالهٔ بداغی و تیمورپور (Bodaghi & Teimourpour, 2018) در نظر گرفته شده است، برای بررسی مشکوکیت ایجاد کند. همچنین در مقالهٔ اوشکارسدوتیر و همکاران (Óskarsdóttir et al., 2022) نشان داده شده است که بیش از ۷۰ درصد خودروها که دارای

فقط دو برخورد هستند، مشکوک به تقلب نیستند. لذا داده‌های این مقاله نویسندگان را بر آن داشت تا دو خودرو را زمانی به‌عنوان پایه یک تخلف سازمان‌یافته در نظر بگیرند که دارای حداقل سه تصادف باشند. به‌عبارت‌دیگر، استفاده از رئوس با درجه سه یا بیشتر برای شروع الگوریتم امری مطابق بر آمار است.

نکتهٔ دوم آنکه هنگامی که یک یال درون دورهای مختلفی قرار می‌گیرد، در واقع همبندی محلی یال افزایش می‌یابد و این بدان معناست که رئوس (خودروهای) منتسب به دو سر این یال دارای مشکوکیت بالاتری نسبت به مابقی رئوسی (خودروهایی) هستند که در دورهای شامل این یال قرار دارند. در بخش اعتبارسنجی میزان همبستگی بین این رئوس که از عدد همبندی محلی آن‌ها در گراف مشخص می‌شود در اعتبار هر خودرو یا تصادف تأثیر بسزایی دارد. دقت کنید در فرایند اجرای الگوریتم عدد همبندی کل گراف اصلاً اهمیت ندارد. نویسندگان در واقع مانند مقالهٔ پینهایرو (Pinheiro, 2011) عدد همبندی گراف را بررسی نکرده‌اند. 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، اصلاً به‌عنوان زیرگراف مشکوک شناسایی نشود. مثلاً، k مسیر به طول بزرگ‌تر از سه را در نظر بگیرید که ابتدا و انتهای همهٔ آن‌ها یکسان بوده و گره مشترک دیگری بین این k مسیر وجود نداشته باشد (k مسیر مجزای درونی). 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، درحالی‌که بنا بر قضیه منگر این گراف 2-همبند است و عدد همبندی محلی دوسر این k مسیر برابر با k است. اما اگر این گراف شامل فقط یک وتر شود کل گراف در فرایند الگوریتم مشکوک به حساب می‌آید.

تجزیه‌وتحلیل الگوریتم بخش مهمی از نظریهٔ پیچیدگی محاسباتی است که تخمین نظری مناسبی برای منابع مورد نیاز یک الگوریتم، مانند زمان و حافظهٔ مورد نیاز 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 رایانهٔ خاص، پیش‌بینی کند. یکی از مهم‌ترین تحلیل‌ها، محاسبهٔ بزرگی زمانی الگوریتم است. حال برای این منظور نیاز است تعداد یال‌ها مشخص شود. بیشینهٔ 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 فهرست 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 باقی‌مانده را به‌صورت زیر انتخاب کرد:

$$\binom{n-2}{k}=\frac{(n-2)!}{k!(n-2-k)!} \qquad (۷)$$

و آن‌ها را می‌توان به هر ترتیب دلخواهی در مسیر قرار داد. لذا تعداد کل مسیرها با جمع کردن این مقادیر برای همهٔ kهای ممکن به شکل زیر به دست می‌آید:

$$\sum_{k=0}^{n-2}\frac{(n-2)!}{k!(n-2-k)!}=(n-2)!\times\sum_{i=0}^{n-2}\frac{1}{i!} \qquad (۸)$$

ازآنجاکه

$$e=\sum_{i=0}^{\infty}\frac{1}{i!} \qquad (۹)$$

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. بااین‌حال، مهم است که توجه داشته باشید که این بدترین سناریوی حالتی است که شبکهٔ تصادفات را یک گراف کامل در نظر بگیریم که در آن همهٔ اتومبیل‌ها با یکدیگر برخورد می‌کنند. چنین وضعیتی با توجه به ساختار شبکهٔ تصادفات واقعی نیست. برای تراز کردن مسئله با شرایط دنیای واقعی، فرض کنید که بزرگ‌ترین زیرگراف کامل در شبکهٔ تصادف از m گره تشکیل شده که $m \ll n$ یک عدد ثابت است.

در این مورد، الگوریتم ما با در نظر گرفتن همهٔ زیرگراف‌های کامل متمایز با m رأس شروع می‌شود. این زیرگراف‌های کامل گراف‌های سطح ۱ نامیده می‌شوند. حداکثر تعداد نمودارهای سطح ۱ برابر با n/m خواهد بود. در سطح ۲، هر گراف سطح ۱ را به‌عنوان یک گره در نظر می‌گیریم. اگر یک یال بین دو گراف سطح ۱ وجود داشته باشد، 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 فرایند 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. ازآنجاکه 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 فرایند پواسون تعیین می‌شود. ازآنجاکه 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 به‌تفصیل بیان می‌شوند. سپس سطح اعتباری برای هر فرد بیمه‌شده براساس میزان ظن مطالبات آن مشخص می‌شود. این سطح اعتبار به‌عنوان یک

برچسب به هرتصادف و هر خودرو در شبکه اختصاص داده می‌شود.

اعتبارسنجی هر تصادف

با توجه به مدل پیشنهادی، هر یال نشان‌دهندهٔ یک تصادف است. A را به‌عنوان 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 همهٔ مسیرها باشد. اگر وجود یک مسیر مجزای درونی بین u و v را به‌عنوان موفقیت با احتمال p تعریف کنیم، آنگاه X یک متغیر تصادفی دوجمله‌ای است. ازآنجاکه 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. براساس قضیه منگر به‌ازای 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 براساس فرمول (۵) داریم:

$$Pr\big(X=\kappa(u,v)\big)=\frac{e^{-\lambda}\lambda^{\kappa(u,v)}}{\kappa(u,v)!} \qquad (۱۰)$$

مقدار $\kappa(u,v)$ در طول بخش اول فرایند الگوریتم پیشنهادی به دست می‌آید. مقدار λ به‌صورت تقریبی از فرمول mp به دست می‌آید. بااین‌حال، مقدار به‌دست‌آمده برای یال $e=uv$ از معادلهٔ بالا مستقل از زیرگراف حاصل از بخش اول الگوریتم است و آن را به‌خوبی تعریف می‌کند. یک مقدار احتمال کمتر برای یک یال نشان‌دهندهٔ 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 نشان‌دهندهٔ سطح بیشتری از مشکوکیت مرتبط با تصادف است. بنابراین، برچسب یال $e=uv$ به‌صورت زیر تعریف می‌شود:

$$l_{uv}=1-Pr\big(X=\kappa(u,v)\big) \qquad (۱۱)$$

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

تاکنون، 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 به‌طور مشابه از تعداد مسیرهای مجزای درونی یالی $\kappa'(u,v)$ و تعداد همهٔ مسیرها $\kappa''(u,v)$ به‌ازای هر یال $e=uv$ برچسب‌های دیگری را معرفی کنیم. در تمام این موارد، اتصال بین گره‌ها به روش‌های مختلف در نظر گرفته می‌شود. به‌راحتی می‌توان مشاهده کرد که $\hat{e}(u,v)\le\hat{e}'(u,v)\le\hat{e}''(u,v)$ تمرکز اصلی $\kappa(u,v)$ روی همبندی محلی بین رئوس است، درحالی‌که $\kappa'(u,v)$ نشان‌دهندهٔ همبندی محلی بین یال‌هاست و $\kappa''(u,v)$ همبستگی[5] در گراف را نشانه می‌گیرد. ازآنجاکه هریک از متریک‌ها ویژگی متفاوتی را در نظر می‌گیرند، بنابراین دو برچسب جدید l'_{uv} و l''_{uv} به‌صورت زیر تعریف می‌شود:

$$l'_{uv}=1-Pr\big(X=\kappa'(u,v)\big),$$
$$l''_{uv}=1-Pr\big(X=\kappa''(u,v)\big). \qquad (۱۲)$$

مثلاً، 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

$$l_{v_6v_9}=l'_{v_6v_9}=l''_{v_6v_9}=1-Pr(X=3)=1-\frac{e^{-\lambda}\lambda^3}{3!} \qquad (۱۳)$$

برای یال v_6v_7 داریم:

$$l_{v_6v_7}=l'_{v_6v_7}=1-Pr\big(X=\kappa(v_6,v_7)\big)=1-\frac{e^{-\lambda}\lambda^2}{2!},$$
$$l''_{v_6v_7}=1-Pr\big(X=\kappa(v_6,v_7)\big)=1-\frac{e^{-\lambda}\lambda^3}{3!}. \qquad (۱۴)$$

اگرچه در این حالت سه مسیر متفاوت مابین v_6 و v_7 وجود دارد، اما تنها دو تا از آن‌ها رأس مجزا و یال مجزا هستند، بنابراین می‌توان نتیجه گرفت:

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

درحالی‌که

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

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

5 betweennes

از این زیرگراف‌ها نیستند، عدد همبندی محلی ۱ اختصاص داده می‌شود.

برچسب‌گذاری و اعتبارسنجی خودروها

هدف بعدی این مقاله، اختصاص یک برچسب مشکوک به هر بیمه‌گذار است. تاکنون، برچسب‌های روی یال‌ها براساس احتمال پواسون تعیین شدند. مرحلهٔ بعدی شامل انتقال این برچسب‌ها به رئوس است. ازآنجاکه چندین یال به هر رأس متصل است، لازم است در هنگام انتقال برچسب یال به رأس، تأثیر هر یال (که نشان‌دهندهٔ میزان سوءظن در مورد هر تصادف است) در نظر گرفته شود. بااین‌حال، به‌دلیل وجود چندین یال متصل به یک رأس، نرمال‌سازی برچسب‌های یال برای به دست آوردن تأثیر عادلانهٔ هر یال بر روی رأس بسیار مهم است. برای این منظور، برچسب‌های یال به‌صورت زیر نرمال می‌شوند.

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

$$L_u=\left\{\sum_{v\in N_\bullet(u)} l_{uv} | v\in V(G)\right\}. \quad (۱۵)$$

فرض کنید l_{max} و l_{min} به‌ترتیب نمایش بزرگ‌ترین و کوچک‌ترین مقدار L_u به‌ازای uهای مختلف باشد. در این صورت برچسب رأس u که با l_u نمایش داده می‌شود به‌صورت زیر تعریف می‌شود:

$$l_u=\frac{\sum_{v\in N_G(u)} l_{uv}-l_{min}}{l_{max}-l_{min}}. \quad (۱۶)$$

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 نشان‌دهندهٔ سطح سوء ظن بالاتر در شبکه است، تأثیر بیشتری بر برچسب رأس u خواهد داشت.

واضح است که متناسب با برچسب‌های l'_{uv} و l''_{uv} که در بخش قبل معرفی شدند، می‌توان به‌ترتیب برچسب‌های l'_u و l''_u را نیز برای هر خودرو تعریف کرد که متناسب با شرایط مورد نظر بیمه‌گذار هریک می‌توانند در موارد مختلف کارایی داشته باشند و مقایسهٔ آن‌ها نیز باعث ایجاد اطمینان بیشتر به اعتبار تخصیص‌یافته می‌شود.

در این پژوهش یافتن تقلب‌های سازمان‌یافته در بیمهٔ خودرو با پیشنهاد مدلی ریاضی بر مبنای نظریهٔ گراف هدف قرار گرفت تا نواقصی را برطرف کند که دیگر الگوریتم‌های پیشنهادشده دارد. لذا نشان داده شد که تصادفات بین خودروها و صدمه دیدن افراد یا خودروها فرایندی تصادفی بوده و هرگونه دخالت سازمان‌یافته توسط افراد در این فرایند تصادفی، سبب ایجاد نوعی نظم شده که می‌تواند سرنخی برای شناسایی تخلف در تصادفات باشد. به‌این‌ترتیب زیرگراف‌های مشکوک براساس شبکهٔ تصادفات و یال‌های مجازی که نشان‌دهندهٔ هزینه‌های زیاد یک خسارت نسبت به میانگین خسارت‌های پرداخت‌شده توسط بیمه‌گر است، کشف شدند.

جمع‌بندی و پیشنهادها

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 کرده است. تقلب‌های بیمه‌ای انواع گوناگونی دارد و در تمام حوزه‌های بیمه‌ای رخ می‌دهد و طیف گسترده‌ای از ادعاهای اغراق‌آمیز تا تصادف‌ها و خسارت‌های تعمدی را دربرمی‌گیرد. تقلب‌های بیمهٔ خودرو، به‌ویژه تقلب‌های سازمان‌یافته که در این پژوهش مطالعه شده است، اغلب به‌صورت 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 روش‌های داده‌کاوی و یادگیری ماشین، مانند شبکه‌های عصبی مصنوعی، منطق فازی و الگوریتم‌های ژنتیک، به‌دلیل 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 استفاده می‌شود، نظریهٔ گراف است. در این نگاه ابتدا به مدل‌سازی ریاضی مسئله پرداخته می‌شود. بدین مفهوم که ابتدا شبکهٔ تصادفات به‌صورت یک گراف مدل شده و با ابزارهای در دسترس به کشف تقلب‌های سازمان‌یافته پرداخته می‌شود. سپس از مفاهیم علوم کامپیوتری استفاده می‌گردد تا شبکهٔ مشکوک به تقلب به‌صورت دقیق‌تر مشخص شود. به‌عبارت دقیق‌تر، در ساختارهایی مانند تصادفات یک کشور تعداد داده‌های موجود بسیار زیادند و یافتن رابطه بین آن‌ها دشوار است. استفاده از ابزارهایی مانند داده‌کاوی و یادگیری ماشین، شبکه‌های عصبی، منطق فازی، الگوریتم‌های ژنتیک و امثال آن‌ها که هدف اصلی‌شان یافتن رابطه بین داده‌هاست، اگرچه بسیار سودمند است، اما نقایصی هم دارد. این‌گونه ابزارها اگر از الگوریتم‌های فرا اکتشافی استفاده کنند عدم دقت یا بیش‌برازش در داده‌های نامتعادل را خواهند داشت، در الگوریتم‌های 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 بینجامد. در این پژوهش نویسندگان سعی کرده‌اند این نقیصه را با استفاده از مدل‌های ریاضی رفع و در ضمن احتمال رخداد رویدادهای مشکوک را نیز بررسی کنند. لذا در این پژوهش، ابتدا با استفاده از نظریهٔ گراف به مدل‌سازی شبکهٔ تصادفات پرداخته شده و سپس نشان داده شده است که این مدل فرایندی تصادفی است و وجود عناصر منظم در مدل بیان‌کنندهٔ مجموعه خودروهای مشکوک به تقلب خواهد بود. در ادامه براساس الگوریتم یافتن زیرگراف

مشکوک که به‌صورت یک فایل اسکریپت (فایل m) در محیط محاسباتی متلب نوشته شده است، خودروهای مشکوک از تمامی خودروها استخراج شده است. در پایان ثابت شده است که شبکهٔ تصادف یک فرایند پواسون است و احتمال رخداد آن را می‌توان مشخص کرد. این استدلال که بر پایهٔ ساختار مدل‌سازی گرافی بنا شده است، کمک می‌کند تا به هر تصادف و به هر خودرو درجهٔ اعتباری به جهت مشکوک بودن به تقلب تخصیص یابد. پیشنهاد برای تحقیقات آتی، ایجاد و بررسی شبکه‌ای گسترده‌تر شامل همهٔ ذی‌نفعان یک تقلب سازمان‌یافته است. به‌عبارت دقیق‌تر، شبکهٔ تخصیص برچسب ذی‌نفعان اصلی که در یک تصادف سود می‌برند وابسته به مقدار سود آن‌ها، بررسی شوند. بررسی این موضوع موجب می‌شود 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.

مشارکت نویسندگان

اسماء حمزه: مفهوم و طرح مقاله، بازنگری مقاله و محتوای کیفی، تحلیل و تفسیر داده‌ها، بازخوانی و رفع ایرادها، مدل‌سازی

محمدجواد نجفی: بازنگری مقاله و محتوای کیفی، تحلیل و تفسیر داده‌ها، مدل‌سازی

تشکر و قدردانی

نویسندگان از داوران محترم مقاله که با پیشنهادهای ارزنده موجب بهبود مقاله شدند، قدردانی و سپاسگزاری می‌نمایند.

تعارض منافع

نویسندگان اعلام می‌کنند که هیچ تضاد منافعی در خصوص انتشار مقالهٔ ثبت‌شده وجود ندارد. علاوه‌براین، موارد اخلاقی از جمله سرقت ادبی، رضایت آگاهانه، سوءرفتار، جعل داده‌ها، انتشار و ارسال مجدد و مکرر ازسوی نویسندگان رعایت شده است.

دسترسی آزاد

یادداشت ناشر

منابع

Brockett, P. L., & Levine, A. (1977). On a characterization of RIDITs. *The Annals of Statistics*,1245-1248. https://doi.org/10.1214/aos/1176344010

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. https://doi.org/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. https://doi.org/10.48550/arXiv.1805.09741

Ghahramani, S. (2005). *Fundamentals of probability with stochastic processes* (3rd ed.). Pearson/Prentice Hall. https://www.amazon.com/Fundamentals-Probability-Stochastic-Processes-Third/dp/1498755011

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. https://doi.org/10.1016/j.jfds.2016.03.001

Noble, C. C., & Cook, D. J. (2003). Graph-based anomaly detection [Conference presentation]. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 631-636. https://doi.org/10.1145/956750.956831

Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., & Reynkens, T. (2022). Social network analytics for supervised fraud detection in insurance. *Risk Analysis*, *42*(8), 1872-1890. https://doi.org/10.48550/arXiv.2009.08313

Pinheiro, C. A. (2011). Highlighting unusual behavior in insurance based on social network analysis. In *SAS Global Forum* (pp. 4-7). https://support.sas.com/resources/papers/proceedings11/130-2011.pdf

Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303. https://doi.org/10.1016/j.dss.2020.113303

Rajan, R. S., Shantrinal, A. A., Kumar, K. J., Rajalaxmi, T., Fan, J., & Fan, W. (2021). Embedding complete multi-partite graphs into Cartesian product of paths and cycles. *Electronic Journal of Graph Theory and Applications*, *9*(2),

507-521. https://doi.org/10.5614/ejgta.2021.9.2.21

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. https://doi.org/10.1016/j.eswa.2010.07.143

Tarawneh, A., Hassanat, A., Altarawneh, G., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, *10*, 47643-47660. https://doi.org/10.1109/ACCESS.2022.3169512

West, D. B. (2001). *Introduction to graph theory* (Vol. 2). Upper Saddle River: Prentice hall. https://www.amazon.com/Introduction-Graph-Theory-Douglas-West/dp/0130144002

SAS Institute Inc. (2011). *The insurance fraud race: Using information and analytics to stay ahead of criminals* [White paper]. https://www.nextdeal.gr/sites/default/files/attachments/SAS.pdf

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. https://doi.org/10.1016/j.physa.2015.05.112

Zhai, J., Qi, J., & Shen, C. (2022). Binary imbalanced data classification based on diversity oversampling by generative models. *Information Sciences*, *586*, 313-343. https://doi.org/10.1016/j.ins.2021.11.058

AUTHOR(S) BIOSKETCHES — معرفی نویسندگان

اسماء حمزه (Asma Hamzeh)، استادیار گروه فناوری‌های نوین بیمه‌ای، پژوهشکده بیمه، تهران، ایران

- Email: *Hamzeh@irc.ac.ir*
- ORCID: 0000-0001-7736-7626

محمدجواد نجفی آرانی (Mohammad Javad Nadjafi-Arani)، دانشیار، گروه علوم کامپیوتر، مرکز آموزش عالی محلات، محلات، ایران

- Email: *Mjnajafiarani@gmail.com*
- ORCID: 0000-0003-1754-6694

HOW TO CITE THIS ARTICLE

Hamzeh, A., & Nadjafi-Arani, M. J. (2025). A mathematical model for identifying and validating suspicious clusters associated with organized fraud in auto insurance. *Iranian Journal of Insurance Research, 14*(3), 209-222.

DOI: 10.22056/ijir.2025.03.03

URL: https://ijir.irc.ac.ir/article_160346.html