**Issue 15, 2024**

https://doi.org/10.23939/sisn2024.15.341

**UDK 004.6**

# DISTRIBUTED DATA ANALYSIS IN CLOUD SERVICES FOR INSURANCE COMPANIES

**Oleksandr Lutsenko[1], Serhii Shcherbak[2]**

Lviv Polytechnic National University,
Information Systems and Networks Department, Lviv, Ukraine
[1] E-mail: Oleksandr.V.Lutsenko@lpnu.ua, ORCID: 0009-0008-7644-6056
[2] E-mail: serhii.s.shcherbak@lpnu.ua, ORCID: 0000-0003-2914-2101



**This article embarks on an insightful journey through the realm of advanced data analysis techniques which can be used in the insurance area, with a keen focus on the applications and capabilities of Graph Neural Networks (GNN) in the following sector. The article is structured into several chapters, which include the overview of existing and commonly used approaches of the data representation, the possible ways of data analysis of the data in such a representation, deep dive into the concept of GNN for the graph data analysis and the applicability of each approach in the insurance industry.**

**The initial chapter introduces the two main concepts of the data representation, which are the commonly used relational database and the more modern approach of dimensional data design. Then the focus is moved to the graph data representation, which also can be used for data analysis in the cloud environment. To achieve the best applicability in the insurance industry, particularly in underwriting and claims management, the article analyzes the advantages of each approach to the data representation as well as its drawbacks. To conclude the chapter, the comparison table of the three approaches is presented. Based on the comparison table, the decision to use the graph representation is made as it enables the industry to unravel complex relationships and dependencies amid various data points—such as policyholder history, incident particulars, and third-party information—resulting in more accurate risk assessments and efficient claim resolutions.**

**Then the article presents the concept of Graph Neural Networks, a rather new concept which can be used to analyze the data, represented in a graph form using machine learning algorithms. The potential of using this approach for the data analysis in the insurance area and some possible use cases are described. The advantages of using this approach include ability to effectively capture and leverage the complex relationships inherent in graph-structured data and a powerful framework for analyzing and processing graph-structured data. However, the potential drawbacks of the approach such as complexity to design and difficulties in scaling are also considered.**

**Further along, the article probes the strategic integration of Graph Neural Networks with real-time and dynamic data environments, examining their adaptability to evolving network patterns and temporal dependencies. We discuss how this adaptability is paramount**

**in contexts like real-time decision-making and predictive analysis, which are crucial for staying agile in a rapidly changing market landscape.**

**Then the exact use cases of the GNN applicability in the insurance area are provided, including the claim assignment and underwriting process are described in detail. Furthermore, the simplified mathematical formulation of the underwriting process is provided, which elaborates the role GNNs play in propelling actuarial science with their capability to incorporate node attributes, edge information, and graph structure into a composite risk assessment algorithm.**

**The article concludes by describing that with the new technologies, the graph representation may become the new standard for the data analysis in the cloud environment, especially for the insurance area, stressing the pivotal role of GNNs in navigating the complexities of interconnected, dynamic data and advocating for their continued research and development to unlock even greater potentials across various sectors.**

**Key words: Relational database, Dimensional data design, Graph, Graph Neural Network (GNN), Underwriting, Insurance.**

## Introduction

In the modern era of Big Data, traditional methods of data analysis have often fallen short when it comes to capturing the intricacies and interconnections present in vast and complex datasets. The aim of the article is to choose the best data representation that will suit for the purpose of data analysis in the cloud environment in the insurance industry for several specific use cases such as underwriting and claim assignment. The paper seeks to illuminate the potential of graph databases for the data representation and Graph Neural Networks as a transformative approach to data analysis in dynamic and quickly changing environment of the claim and underwriting process which can be very different for each client.

We begin with an exploration of different data representation techniques in the insurance industry such as relational databases and dimensional data design.

Building on the established techniques in the insurance sector, the field of data analysis has historically leveraged structured databases to understand and predict customer behavior, risk, and financial viability. Traditional relational databases have been a staple, organizing data into tabular formats that insurance companies use to track and manage countless records, from policyholder information to claims data. Although effective for structured data management and complex querying, these models lack flexibility when dealing with the intricate web of relationships intrinsic to insurance data, such as networks of policyholders, their claims history, and interactions with service providers.

Dimensional data modeling, with its focus on multi-faceted analysis, has also found its place in the insurance domain. Data warehouses built on dimensional models enable insurers to perform sophisticated analyses, such as identifying trends in claims or understanding the impact of risk factors across different dimensions of the data, such as geography, time, and customer demographics. However, the static nature of these models means they struggle to adapt to the real-time, interconnected nature of today's data streams.

The advent of Graph Neural Networks marks a pivotal development, particularly for the insurance industry, where the ability to grasp and evaluate complex, dynamic relationships can redefine risk assessment and fraud detection. GNNs provide a powerful tool to visualize and analyze the interdependencies and intricate network of relationships, such as those between policyholders, claims, and incidents, in a way that traditional data representation methods simply cannot. This advanced approach paves the way for more nuanced and real-time decision-making, crucial for adapting to the evolving landscape of risks in the insurance sector.

## Formulation of the problem

The problem addressed in this article is the challenge of distributed data analysis for insurance companies in cloud services. Specifically, the focus will be on application of Graph Neural Networks in

handling and analyzing large-scale graph data relevant to the insurance industry. This problem is of critical importance as insurance companies handle vast amounts of interconnected data, and there is a need for efficient, scalable, and accurate analysis methods to extract valuable insights from this data. This article aims to review the current approaches and techniques in the utilization of GNNs for distributed data analysis in cloud services for insurance companies.

## Analysis of recent research and publications

### Common Approaches of Data Representation in the Insurance Industry

When considering the representation of data for distributed data analysis, it is essential to consider standard approaches such as relational databases and dimensional data design. Relational databases are widely used for structuring data with a focus on how data elements relate to one another. They provide a flexible and efficient way to store and retrieve data, making them suitable for a wide range of applications.

On the other hand, dimensional data design is specifically tailored for data analysis and reporting. It involves organizing data into a structure that is optimized for querying and summarizing large datasets. This approach is commonly used in data warehouses and business intelligence systems to facilitate complex analytical queries.

Relational databases are commonly used in the insurance industry for storing and managing large volumes of data. These databases are structured to store information in tables with rows and columns, allowing for efficient data retrieval and manipulation. Insurance companies can use relational databases to store customer information, policy details, claims data, and other relevant information in a structured and organized manner.

Dimensional data design is particularly useful for insurance companies when it comes to analyzing and reporting on their data. This design approach involves organizing data into dimensions and facts, making it easier to perform complex queries and analytics. For example, an insurance company can use dimensional data design to analyze policy sales across different regions, time periods, and customer segments, enabling them to gain valuable insights into their business performance.

In the context of insurance companies, the use of relational databases and dimensional data design is crucial for managing and analyzing complex and interconnected data sets. These approaches enable insurance companies to make informed decisions, improve operational efficiency, and enhance customer service through the analysis of their data.

### How Relational Models of Data Representation Work

Relational models of data representation work by structuring data into tables, with each table representing an entity or a relationship between entities. Relational models of data representation work by structuring data into tables that are interlinked through common attributes or keys. These tables represent entities or objects, and the relationships between these entities are established through these common attributes. The relational model allows for the creation of complex queries that can retrieve data from multiple tables based on the defined relationships.

In a relational database, the relationships between tables are established using foreign keys, which define the links between the data in different tables. This enables the database to efficiently retrieve related data and maintain referential integrity. Additionally, the relational model supports normalization, which reduces data redundancy and ensures data consistency by eliminating update anomalies [1].

The insurance industry can greatly benefit from utilizing a relational representation for data in cloud systems. Relational databases provide a standardized and efficient way to organize and manage structured data, making it easier to maintain large datasets. They also offer the foundation needed to enforce data integrity constraints and security measures, which are crucial in the insurance sector.

However, there are limitations when using relational representation for big data in the insurance industry. As the volume of data increases, managing complex relationships between tables can become challenging and

may lead to performance issues. Additionally, the rigid structure of relational databases may not be well-suited for unstructured or semi-structured data commonly found in claims information or customer profiles.

While relational models of data representation provide a solid framework for managing structured insurance-related information like policies and premiums; they might not be as effective when handling big volumes with complex interdependencies such as claim histories across various product lines or unstructured elements like digital documentation through alternative approaches could prove more beneficial – graph databases or NoSQL solutions might better cater to these needs within distributed environments prevalent throughout different departments within an insurer's organization.

**Dimensional Model specifics, benefits, and drawbacks**

On the other hand, one of the most commonly used layouts for the data representations in modern Data Lakes and Data warehouses is a dimensional model. It is a data structure technique optimized for data warehousing tools, designed to improve data readability and write efficiency for more straightforward querying and analysis. Dimensional models allow data to be stored in two types of tables: facts and dimensions. The fact tables store quantitative data or measurements of the business process, and the dimension tables store the context (dimension) of the facts which includes descriptive attributes related to the fact data [2].

A multidimensional model extends the basic concept of the dimensional model by organizing data into a cube-like structure that allows for multiple dimensions to be analyzed simultaneously. This cube structure enables complex calculations, trend analysis, and data mining in an efficient manner. Each axis of the cube represents a different dimension, and the cells within the cube represent the data values at the intersections of these dimensions.

There are several benefits of using a multidimensional data model for representation of the huge volumes of the data in insurance industry [3]:

- Performance: Optimized for query speed, allowing the fast retrieval of complex and aggregated data.
- Intuitive: Closely aligns with the way business users think and talk about information, facilitating easier data understanding and interpretation. Although it may become a lot harder with the growth of connections between data
- Analysis: Simplifies analysis by enabling the slicing and dicing of data across various dimensions.
- Scalability: Handles large amounts of data efficiently, which is essential for big data environments.

However, there are also some drawbacks of using this approach for the data representation:

- Complexity: Can be complex to design and implement, requiring specialist knowledge to build and maintain.
- Inflexibility: Once a cube is constructed, making changes can be difficult and time-consuming.
- Cost: Typically involves higher storage and computational costs due to the pre-aggregation of data.

**Exploring Graph Data Representation**

While relational databases and dimensional models have their strengths, graph data representation provides an alternative approach that is particularly well-suited for handling interconnected and complex data dependencies. Graph databases store data in the form of nodes, edges, and properties, allowing for efficient representation and analysis of relationships between various entities. This makes graph data representation ideal for scenarios where the relationships between data elements are just as important, if not more so, than the data itself. Graph data representation has gained significant attention in recent years, especially for distributed analysis tasks.

Graph databases excel in representing highly connected and dynamic data, such as social networks, recommendation systems, and network topologies. The flexibility and scalability of graph data representation make it an attractive option for distributed data analysis in cloud services. By leveraging graph data

representation, organizations can effectively model and analyze complex relationships and dependencies within their data, providing valuable insights that might be challenging to uncover using traditional relational or dimensional approaches.

A graph database is a database designed to treat the relationships between data as equally important to the data itself. The key concept is that the relationships allow data in the store to be linked together directly and, in most cases, retrieved with one operation.

Multilayer graphs extend directed labeled graphs by introducing edge identifiers, which allows for a more nuanced representation of relationships and helps reduce the need for complex reification constructs. In a multilayer graph, edges can themselves be nodes, and layers of additional information can be added, creating a richer and more flexible model that supports various graph formats like RDF and property graphs [4].

Multilayer graphs offer various benefits when applied in the insurance industry, specifically in underwriting and claim assignment:

- Flexibility: Multilayer graphs naturally support complex data models without needing intricate workarounds like reification. This can be valuable for representing intricate relationships between policyholders, assets, and risk factors.
- Powerful Representation: These graphs provide an efficient way to model and store complex knowledge such as policies, claims history, and risk assessments associated with different layers of information relevant to the insurance domain.
- Versatility: Widely adopted in diverse domains, multilayer graphs show wide applicability in managing interrelated data about insured entities including relationships between policyholders or properties within a portfolio.

However, the drawbacks include:

- Complexity: While reducing the need for reification, multilayer graph structures can become significantly more complex than simpler graph models. This complexity may pose difficulties for understanding without specific tools but offers potential value through advanced analysis techniques involving interconnected insurance-related data points.
- Querying: Utilizing multilayer graphs may require more sophisticated querying mechanisms specific to identifying patterns on grouped risks across multiple layers while gauging their impact on underwriting decisions or forecasting claim distributions appropriately due to amalgamated historical trends at each layer of representation Interoperability has been recently improved enabling tighter integrations within legacy systems mostly built around traditional relational databases. However, ensuring seamless communication particularly with those designed strictly for other types of graph databases remains a challenge that demands further exploration.

Overall, multilayered graph representations present an intriguing approach towards analyzing vast amounts inherent complexities found amidst longitudinally stored multitude restrictions tagged against numerous entity schema's all neatly intertwined likely having immediate attention whilst making handling dynamically evolving Insurance Use Cases.

**Comparison of the approaches to the data representation**

Having defined the commonly used standard approaches to the data representation in the cloud environment and the graph database, it is required to address all the differences between them to come up with the best solution for the insurance area use case. To achieve this goal, the comparison table may be used (Table 1).

*Table 1.*

**Comparison table of different data storage approaches for analytical purposes**

| Features | Relational databases | Dimensional design | Graph databases |
|---|---|---|---|
| Highly structured data storage | + Best fit | +- Dimensions can change often, facts no | - Suit better for semi-structured data with changing features |
| Transactional storage | + Suits best for usually updated transactions | - Does not support transactions | - Does not support transactions |
| Performance | +- Great for quick transactional updates, but not for analytical queries | + Optimized for query speed, allowing the fast retrieval of complex and aggregated data | + Suits best for running analytical queries or passing the graph through GNN |
| Analysis | - Not a great fit for analytical purposes, querying history data takes a lot of time | + Huge volumes of data can be queried and processed quickly | + Good fit running analytical queries based on a few nodes |
| Business understanding | + Used for a long time, so is pretty easy to understand, but may become frustrating with a lot of connections between tables | +- Connections between dimension and fact tables can become hard to understand quickly | + As it translates directly to the graphical representation, it is a great fit for business understanding |
| Structure Complexity | +- Easy to define the structure, but hard to manage a lot of dependencies | + Can be complex to design and implement | +- Structure is intuitive for simple layer graphs, but can become hard for multi-layer |
| Dependency management | +- Gets progressively harder to query and manage with more dependencies | +- Gets progressively harder to manage the structure with more dependencies | + Easy support of complex data models with a lot of dependencies |
| Scalability | - Hard to scale, requires vertical scaling | + Easy to scale to petabytes of data for dimensions | + Easy to scale as nodes do not hold a lot of information |
| Fit for insurance data analysis (underwriting process, claim assignment) | - Suits best for transactional operations | +- Can be a good choice for the data analysis, but gets hard to manage, considering the nature of the data with a lot of connections between different tables | + The best fit for the nature of the data, considering a lot of interconnections and is easily understandable for the business |

Considering the comparison of the approaches, it is safe to assume that graph data representation is the best suit for the purpose of the data analysis in the cloud environment for the insurance area.

## Goal formulation and task setting

Having analyzed the existing approaches to the data representation in the insurance industry, the goal of the article can be defined as exploration of the potential of using neural networks in the analysis of

the graph data representation in the insurance industry. To achieve that, a few use cases of the graph data representation in the insurance industry can be presented such as underwriting process and claim assignment after the insurance incident. Having defined the graphs for the aforementioned processes, the analysis of the graphs using GNN should be described. Additionally, the article aims to highlight the importance of efficient, scalable, and accurate analysis methods for extracting valuable insights from interconnected data within the insurance sector with the help of GNN.

## Presenting main material

### The Graph Neural Network Models in the Insurance Area

Graph neural networks have emerged as a powerful and versatile class of models for addressing the challenges in analyzing graph-structured data.

The concept of neural networks that operate on graph-structured data can be traced back to the work of Scarselli et al. in 2008, where they introduced the concept of graph neural networks as a way to extend neural network models to handle graph data. However, it wasn't until the advancements in deep learning and the need to effectively model complex relational data that GNNs began to gain widespread attention.

One of the primary reasons for the current surge in interest in GNNs is the growing availability of graph-structured data in diverse fields, from social networks and knowledge graphs to protein-protein interaction networks [5]. Traditional neural networks and other machine learning models are ill-equipped to effectively capture the rich structural information and dependencies present in such data. This is where GNNs shine, as they are designed to operate directly on graph structures, effectively capturing the relational information and local dependencies inherent in these data types.

The ability of GNNs to learn and propagate information across the graph structure, taking into account the local neighborhood of each node, has made them a compelling choice for a wide range of tasks, including node classification, link prediction, and graph classification.

Furthermore, the evolution of hardware and software technologies has also contributed to the resurgence of interest in GNNs. The development of efficient algorithms for training GNNs, coupled with the availability of specialized hardware for graph-based computations, has made it more practical to apply GNNs to large-scale graph data.

In conclusion, the emergence of graph neural network models represents a significant advancement in the field of machine learning and data analysis. With their ability to effectively capture and leverage the complex relationships inherent in graph-structured data, GNNs are poised to play a crucial role in addressing the challenges of distributed analysis in cloud services, making them more relevant than ever in today's data-driven landscape [6].

However, while graph neural networks offer numerous advantages in analyzing graph-structured data, they also come with certain drawbacks that need to be considered, which include the following:

1. Complexity of Design and Implementation: The design and implementation of graph neural networks can be more complex compared to traditional neural networks, especially when dealing with graph-structured data. GNNs require specialized architectures and algorithms to effectively capture and process the structural information present in the graph, which can lead to increased complexity in model development and deployment.

2. Limited Interpretability: Due to the complex nature of graph neural network models, interpreting the learned representations and decision-making processes can be challenging. Understanding the inner workings of GNNs and the reasoning behind their predictions and classifications may require specialized expertise and tools, limiting their interpretability compared to simpler machine learning models.

3. Scalability and Efficiency: While GNNs have shown promise in processing large-scale graph data, achieving optimal scalability and efficiency in distributed computing environments can still be a challenge. As the size and complexity of graph data increase, the computational and memory requirements of GNN algorithms may also escalate, posing scalability issues in certain scenarios.

4. Overfitting and Generalization: GNNs may be susceptible to overfitting when training on sparse or noisy graph data, leading to suboptimal generalization performance. Balancing model complexity and generalizability in the context of graph-structured data presents a unique set of challenges for GNNs, especially when dealing with diverse and heterogeneous graph structures.

In summary, while graph neural networks offer a powerful framework for analyzing and processing graph-structured data, it is important to consider and address the associated complexities and limitations. As the field of graph neural networks continues to evolve, ongoing research and advancements aim to mitigate these drawbacks and further enhance the applicability and effectiveness of GNNs in diverse domains.

**Handling Dynamic and Real-time Data with Graph Neural Networks**

As the landscape of the data continues to evolve, the need to handle dynamic and real-time data has become increasingly important. Traditional machine learning models often struggle to adapt to the changing nature of data, especially in scenarios where data is constantly updated or when temporal dependencies play a critical role. Graph neural networks have shown promise in addressing these challenges, particularly in the context of dynamic graph structures and real-time data analysis.

Dynamic graph structures arise in various domains such as social networks, transportation networks, and communication networks, where the relationships and connections between nodes are subject to change over time. Adapting GNNs to handle dynamic graph structures involves extending the model's capability to capture and process evolving relationships and temporal dependencies within the graph [7].

However, existing approaches that utilize node embeddings and recurrent neural networks may have limitations in scenarios where the node set frequently changes or completely differs at different time steps. To overcome these limitations, recent research has explored novel approaches for handling dynamic graph structures with GNNs. One such approach is the use of graph evolution networks, which aim to model the temporal dynamics of a graph by explicitly encoding the changes in node set and

Furthermore, the concept of message passing in GNNs can be extended to account for temporal information, where messages exchanged between nodes carry not only structural and attribute information but also temporal cues. This enables GNNs to effectively propagate and update information based on the dynamic nature of the graph, offering a powerful framework for real-time analysis of evolving networks.

**Utilizing Graph Data Representation in Insurance**

The insurance industry relies heavily on data to assess risk, make underwriting decisions, and detect fraudulent activities. With the increasing volume and complexity of data, the role of data analysis has become paramount in the insurance sector. Data analysis allows insurance companies to gain valuable insights into customer behavior, market trends, and risk factors, ultimately leading to more accurate pricing and personalized insurance. In particular, the use of advanced analytics, including machine learning and data mining techniques, has revolutionized the insurance industry by allowing for more accurate pricing, improved customer segmentation, and enhanced fraud detection.

The underwriting process, in particular, benefits significantly from data analysis. The underwriting process involves the evaluation of insurance applications to determine the level of risk associated with insuring a particular individual or entity. This assessment typically includes analyzing an applicant's personal information, medical history, and any other relevant data to calculate the likelihood of a claim being made. By leveraging historical data and using predictive analytics, insurance underwriters can make more informed decisions, leading to more accurate risk assessment and appropriate premium calculations [8].

By leveraging historical data and predictive analytics, insurance underwriters can more accurately assess risk and determine appropriate premiums. Data analysis allows for the identification of key risk factors and patterns, leading to better risk classification and pricing strategies. Moreover, the use of data analysis in underwriting helps insurance companies adapt to dynamic market conditions and evolving customer needs, ensuring competitiveness and profitability in a rapidly changing landscape.

For the actual representation of the underwriting process in the form of a directed graph, the main idea is to represent every process or role as a vertex (or node) and the relationships between these vertices as edges (or arcs).

Set of Vertices (V):

- V1: The Applicant
- V2: Underwriting System (or Underwriter)
- V3: Policy Issuance
- V4: Reinsurance Company (if applicable)
- V5: Claims Department

These vertices are stages or actors involved in the underwriting process.

Set of Edges (E):

- E1: (V1, V2) The applicant submits an application to the underwriting system.
- E2: (V2, V3) Once the application is assessed and if it's accepted, the underwriting system forwards the application to policy issuance.
- E3: (V2, V4) If the risk is too high for the insurance company to fully undertake, the underwriting system may involve a reinsurance company.
- E4: (V3, V1) Once the policy is issued, it is given to the applicant.
- E5: (V1, V5) If there's a claim by the policyholder, the process involves the claims department.
- E6: (V5, V2) In some cases, the claims department may need to liaise with the underwriting department for clarification or assessment.

Please note that this is a simplistic representation of the underwriting process (Fig.1). It may vary depending upon the company's size, geographical location, number of intermediaries, regulations, and the complexity and type of insurance product.

In a directed graph or a digraph, each edge has an orientation (i.e., direction from one vertex to another), which here indicates the direction of the underwriting process flow. The layout of the directed graph would depend upon the exact sequence of processes within the specific company for which the underwriting process is being graphed.

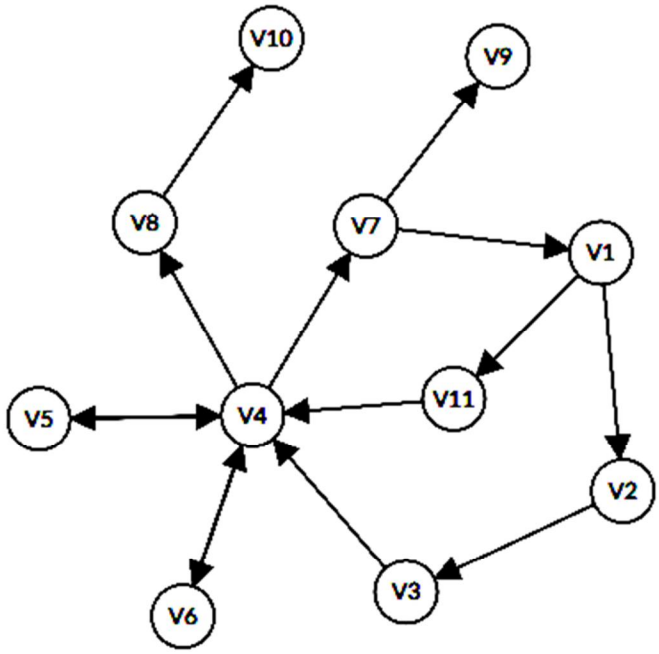


*Fig. 1. The graph representation of the underwriting process*

Incorporating graph neural networks into the underwriting process offers additional advantages. GNNs are well-suited for analyzing complex relationships and interconnected data, which are inherent in insurance portfolios. By processing graph-structured data, GNNs can uncover hidden correlations and dependencies, providing underwriters with a more comprehensive understanding of risk factors and customer profiles. This, in turn, leads to more accurate risk assessment, enhanced portfolio management, and improved decision-making in underwriting.

The use of GNNs in insurance underwriting also addresses the need for real-time data analysis, especially in response to dynamic market shifts and changing customer behaviors. GNNs' ability to adapt to evolving graph structures and process real-time streams of information aligns well with the demands of the insurance industry, where timely insights and decision-making are critical for sustainable business operations.

We can use the GNN in the underwriting process to solve the following tasks:

- Node Classification: If you're using a GNN for a node classification task, for each node in the graph, the GNN would return a probability distribution over the potential classes for each node. In the context of an underwriting process, classes could be categories like approved, declined, or pending further checks.
- Link Prediction: In a link prediction task, for each potential or existing edge in the graph, the GNN would return a score reflecting the likelihood of that edge existing or being created. This could be useful for predicting a possible future stage in the underwriting process.
- Node Regression: If you're doing a node regression task, the GNN produces continuous values for each node. This could be useful for scoring or ranking applications in an underwriting process.

It is worth noting that whatever the GNN returns, the actual result would depend on how the machine learning problem is framed. There is necessity to label and annotate graph-based data accordingly to train the model. It is important to consider the problem at hand, the type of data available and the business needs choosing the appropriate GNN technique to apply.

The potential usage of the graph described above via GNN may look as follows:

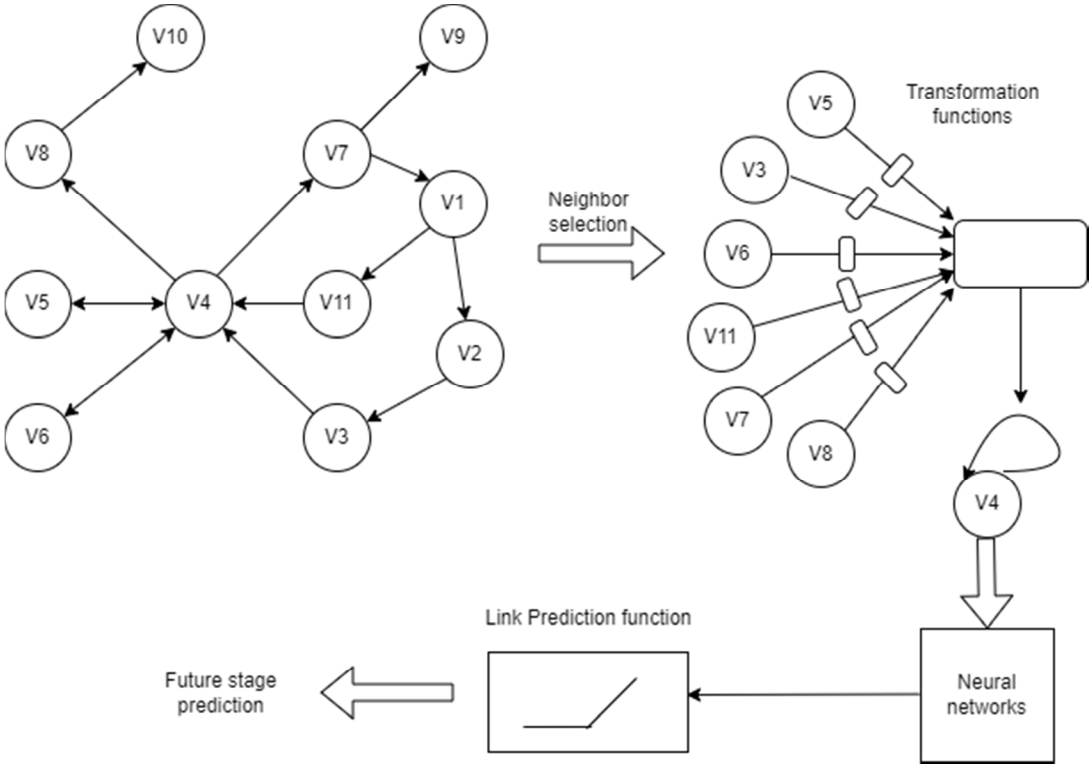


*Fig. 2. The underwriting process in GNN*

Another example of graph visual representation in the insurance industry is the resolution of a claim payment after a car incident. Let's consider a scenario where an insurance company receives a claim for a car accident. By utilizing graph data representation, the company can map out the interconnected variables associated with the claim, such as the policyholder's driving history, the location and time of the incident, the severity of the accident, and any third-party involvement (fig. 3).

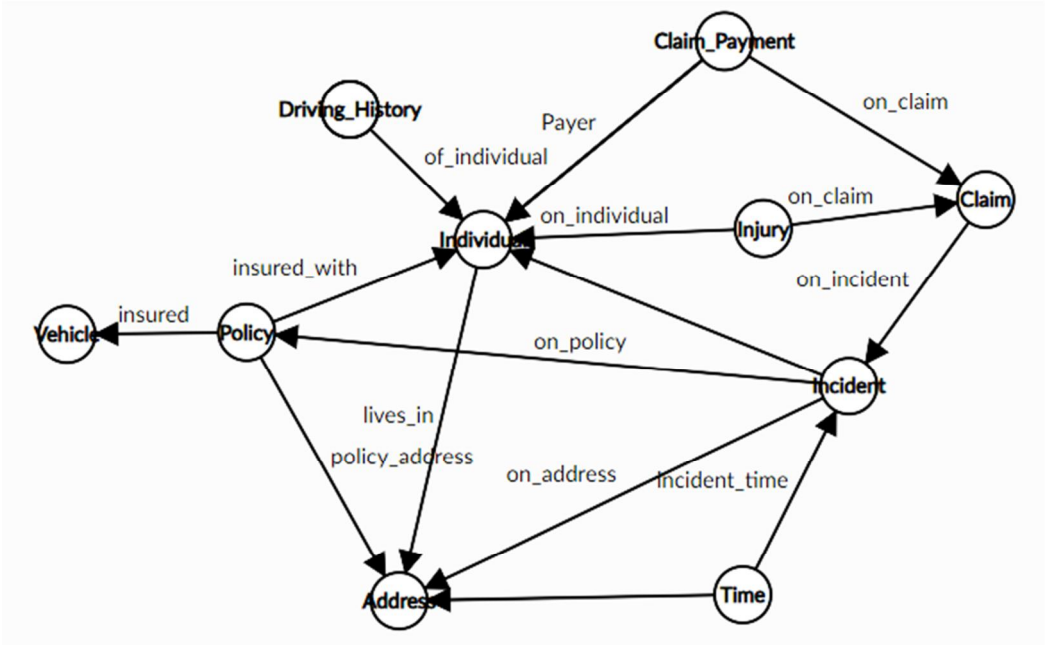


*Fig. 3. The graph representation of claim assignment for a car accident*

With graph neural networks, the insurance company can effectively analyze the relationships among these variables to determine the legitimacy of the claim and assess the appropriate coverage and payout. The GNN can process the structured data and identify patterns and dependencies that may indicate fraudulent activities or help in making accurate decisions.

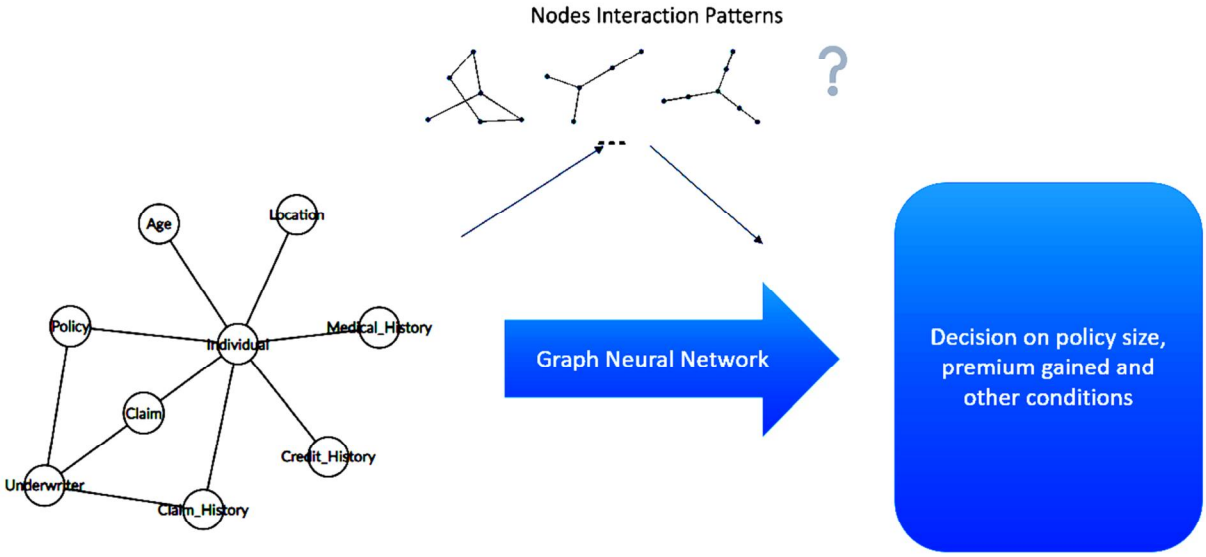


*Fig. 4: GNN usage for underwriting process*

Furthermore, the utilization of GNNs allows for real-time analysis of the evolving factors related to the claim, such as the progression of medical treatments for injuries, the repair status of the vehicles involved, and any legal proceedings. This real-time analysis facilitates prompt and informed decision-making, enabling the insurance company to respond swiftly to the evolving nature of the claim and provide fair and accurate resolutions.

By visualizing the dynamic graph representation using GNNs, insurance companies can gain a comprehensive understanding of the interconnected data points, leading to improved accuracy in claim resolution, fraud detection, and risk assessment [9]. This application of graph visualization demonstrates the power of GNNs in enhancing the efficiency and effectiveness of the claim resolution process in the insurance industry.

**Mathematical Formulation of Underwriting Process**

GNNs enable the development of algorithms that can effectively assess risk by considering the interconnected nature of insurance data. The formula for risk assessment using GNNs involves the propagation of risk-related information throughout the graph, taking into account not only the attributes of individual entities but also the complex dependencies and correlations within the insurance network.

The risk assessment formula can be represented as:

$$f(G) = G\ (V, E) \rightarrow R \quad (1)$$

Where:

- G = (V, E) where V is a set whose elements are vertices which represent different factors that must be considered during underwriting and E is a set of edges which represent connections between these factors.
- -> represents the undergoing of the graph G through GNN process.
- R represents the actual value of the risk of insuring the individual (or a company).

As the insurance industry continues to navigate increasing data complexity and market dynamics, the integration of GNNs and advanced data analysis techniques in the underwriting process holds great potential for driving innovation and enhancing risk management practices. By harnessing the power of GNNs, insurance companies can gain a competitive edge, improve underwriting accuracy, and ultimately deliver greater value to their customers.

Graph data representation has gained significance in the insurance industry, particularly for underwriting purposes. By leveraging graph neural networks for analyzing and processing complex relationships within insurance data, insurers can gain valuable insights into risk assessment and decision-making.

**The practical significance and ways of applying the results**

The actual graph data representation and its further analysis can be applied on top of the existing storage for the insurance company. For example, in case of the data, which is stored on Amazon Web Services cloud (AWS) S3 in the form of Data Lake on S3 (Simple Storage Service), the GNN can be applied for the risk assumption in the underwriting process with the following simplified architecture (Fig. 5). In this architecture we consider that the process of risk assumption for the new policyholder during the underwriting is triggered via the API by a service called Lambda function. After that the necessary data can be read with serverless ETL tool “AWS Glue” to preprocess it and prepare for further usage in the format of the graph. The next step would be to train the model in the special AWS service for ML called “Sagemaker”. After that we can transfer the data in the graph format to NoSQL graph database “Neptune”. The data from Neptune can then be used in the Sagemaker to create the model and use it to predict risks in the underwriting process.

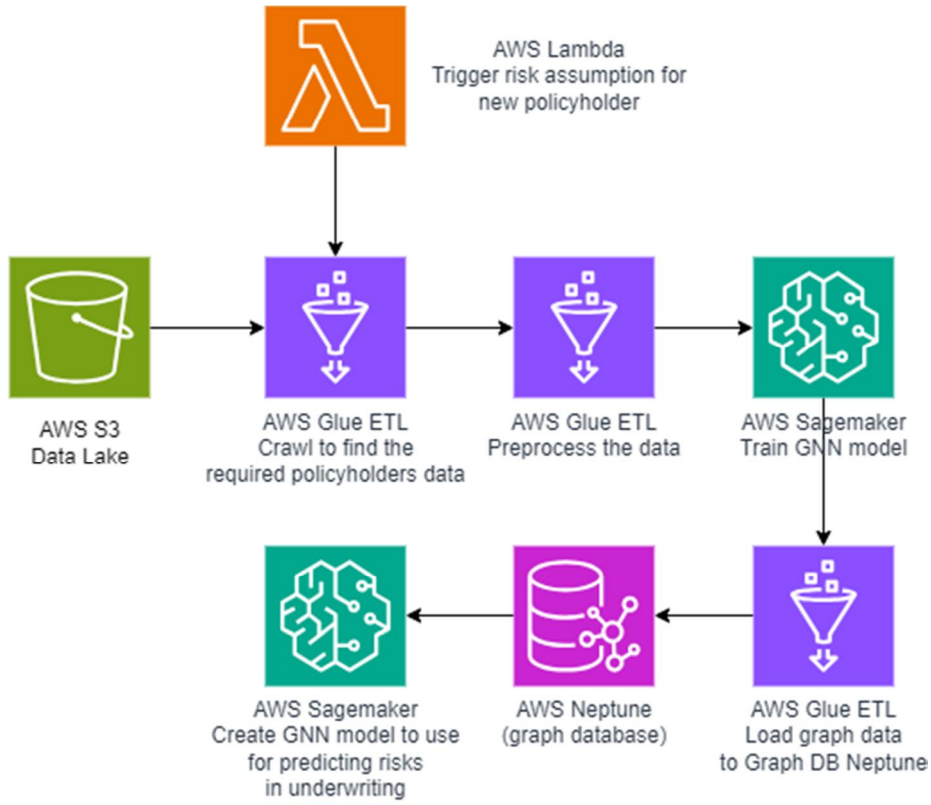


*Fig. 5. Architecture diagram for the underwriting process on AWS cloud service*

The Sagemaker runs the code on Python programming language to analyze the risks using the provided pretrained model. The code is dependent on the open-source machine learning library "Torch". The code is not provided fully, only some parts model is created in the following way:

```
class Net(torch.nn.Module):
   def __init__(self):
      super(Net, self).__init__()
      self.conv1 = GCNConv(data.num_node_features, 16)
      self.conv2 = GCNConv(16, num_classes)

…
# Initialize the model and optimizer
model = Net()
optimizer = torch.optim.Adam(model.parameters(), lr=0.01)
```

Then the model is trained in the training loop to use it in the next run of Sagemaker service for the risk assessment:

```
# Switch model to evaluation mode
model.eval()

# Pass the graph for which the risk assessment is happening
new_out = model(insurance_graph)

# The output is a probability distribution over the classes for each node
# To get the predicted class for each node, take the class with the highest probability
policy_holders_risks = new_out.argmax(dim=1)
```

The results of the Sagemaker second run then represent the estimated risks for the new policyholders and canbe used to choose the proper policy for new customers.

## Conclusions

In this paper, we have thoroughly analyzed different approaches to data analysis in a cloud environment such as dimensional model, relational and graph databases. The specific focus was provided to the applications of the provided approaches in the insurance area. The special attention has been given to the capabilities and applications of Graph Neural Networks in real-time and dynamic environments for the different possible use cases in the insurance industry. Through our investigation, it is evident that GNNs offer unparalleled advantages in processing and interpreting continuously evolving graph structures and temporal data.

The adaptability of GNNs to changing connectivity patterns, coupled with their ability to capture temporal dependencies and process real-time streams of information, makes them the ideal fit for addressing the challenges posed by dynamic and interconnected data in the insurance industry. This adaptability not only facilitates quick decision-making and response to changes in the underlying graph structure but also enhances the model's ability to make informed predictions and recommendations in real time.

Besides the analysis of the approaches for the data analysis, the specific examples of the graph data model in the insurance industry have been provided for the two use cases: one for the underwriting process and one for the claim assignment right after the car incident. The graphs represent all of the aspects of the following processes in a simplified way. The basic architecture of the possible way to analyze the provided graphs using GNN has been presented naming the advantages of such approaches, which include nearly real-time and dynamic response for both the underwriting process and the claim assignment.

In conclusion, the analysis presented in this paper underscores the significance of GNNs in advancing the state-of-the-art in the insurance industry area. The unique ability of GNNs to effectively capture temporal dependencies and process nearly real-time streams of information positions them as the frontrunners in addressing the complexities of the interconnected data. Graph data representation is perfectly suited for the insurance industry's underwriting and claims assignment processes due to their exceptional ability to model the complex, interrelated patterns of risk and behavior. Running the graphs through GNNs can provide insurers with deeper insights and a more dynamic analytical capacity for precise decision-making and enhanced fraud detection.

## References

1. Krasowski, M. D. (2015). The Power of the Relational Model. *Journal of Information Systems Management, 4*(3), 73–78. DOI: https://doi.org/10.1080/07399018808962931

2. Setyawan, R. A., & Prasetyo, E., & Girsang, A. S. (2019). Design and Implementation Data Warehouse in Insurance Company. *Journal of Physics: Conference Series, 1175*(1), 72–88. DOI: https://doi.org/10.1088/1742-6596/1175/1/012072

3. Hänel, T., & Schulz, M. (2014). Is there still a need for multidimensional data models. *Proceedings of the European Conference on Information Systems (ECIS) 2014.*

4. Angles, R., & Hogan, A., & Lassila, O., & Rojas, C., & Schwabe, D., & Szekely, P., & Vrgoc, D. (2022). Multilayer graphs: a unified data model for graph databases. *GRADES-NDA '22: Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), 5,* 1–6. DOI: https://doi.org/10.1145/3534540.3534696.

5. Scarselli, F., & Gori, M., & Tsoi, A. C., & Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks, 20*(1), 61–80. DOI: https://doi.org/10.1109/TNN.2008.2005605

6. Zhang, X., & Zhang, L., & Liu, L., & Tang, M. (2021). Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics, 12.* DOI: https://doi.org/10.3389/fgene.2021.690049

7. Zhou, J., & Cui, Z., & Hu, S., & Zhang, Z, & Yang, C., & Liu, Z., & Wang, L., & Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open, 1,* 57–81. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001

8. Ma, Y., & Guo, Z., & Ren, Z., & Tang, J., & Yin, D. (2018). Streaming Graph Neural Networks. *arXiv, 1810*. DOI: https://doi.org/10.48550/arXiv.1810.10627

9. Chen, C., & Liang, C., & Lin, J., & Wang, L., & Liu, Z., & Yang, X., & Wang, X., & Zhou, J., & Shuang, Y., & Qi, Y. (2020). InfDetect: a Large Scale Graph-based Fraud Detection System for E-Commerce Insurance. *arXiv, 2003*. DOI: https://doi.org/10.48550/arXiv.2003.02833

**Список літератури**

1. Красовські, М. Д. (2015). "Потужність реляційної моделі". *Журнал управління інформаційних систем, 4*(3), 73–78. DOI: https://doi.org/10.1080/07399018808962931

2. Сетяван, Р. А., & Прасетіо, Е., & Гірсанг, С. (2019). "Дизайн та втілення сховища даних для страхової компанії". *Журнал фізики: Серія конференцій, 1175*(1), 72–88. DOI: https://doi.org/10.1088/1742-6596/1175/1/012072

3. Ганел, Т., & Шульц, М. (2014). Чи існує потреба в мультивимірних моделях даних?. *Провадження Європейської конференції з Інформаційний Систем (ЄКІС) 2014.*

4. Енгельс, Р., & Хоган, А., & Лассіла, О., & Рохас, К., & Швабе, Д., & Секели, П., & Вргоч, Д. (2022). "Багатошарові графи: уніфікована модель даних для графових БД". *ГРАКДС-МАД '22: Проведення 5-го АСМ SIGMOD Міжнародного майстер класу про керування даними графів за допомогою досвіду та систем (ГРАКДМС) та Мережевої Аналітики Даних (МАД), 5*, 1–6. DOI: https://doi.org/10.1145/3534540.3534696.

5. Скарселлі, Ф., & Ґорі, М., & Цой, А. Ц., & Гагенбухнер, М., & Монфардіні, Ґ. (2009). Модель графової нейронної мережі. *IEEE Транзакції над нейронними мережами, 20*(1), 61–80. DOI: https://doi.org/10.1109/TNN.2008.2005605

6. Жень, К., & Жень, Л., & Лю, Л., & Танґ, М. (2021). Графові нейронні мережі та їх використання в біоінформатиці. *Фронтири в генетиці, 12.* DOI: https://doi.org/10.3389/fgene.2021.690049

7. Жау, Ж., & Суі, Ґ., & Ху, С., & Жан, Ж., & Ян, Ч., & Лю, Ч., & Ван, Л., & Лі, С., & Сан, М. (2020). Графові нейронні мережі: огляд методів та застосування. *ШІ Відкритий,1,* 57–81. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001

8. Ма, І., & Ґуо, З., & Рен, З., & Танґ, І., & Їн, Д. (2018). Потокові графові нейронні мережі. *arXiv, 1810.* DOI: https://doi.org/10.48550/arXiv.1810.10627

9. Чен, Ц., & Лян, С., & Лін, Д., & Ван, Л., & Лю, З., & Ян, К., & Ван, К., & Жау, Д., & Шуан, Й., & Кі, Й. (2020). Інфдетект: великомасштабна система визначення шахрайства на базі графів для Е-комерції страхування. *arXiv, 2003.* DOI: https://doi.org/10.48550/arXiv.2003.02833

## РОЗПОДІЛЕНИЙ АНАЛІЗ ДАНИХ В ХМАРНИХ СЕРВІСАХ ДЛЯ СТРАХОВИХ КОМПАНІЙ

**Олександр Луценко[1], Сергій Щербак[2]**

Національний Університет "Львівська політехніка", кафедра ІСМ, Львів, Україна
Національний Університет "Львівська політехніка", кафедра ІСМ, Львів, Україна
[1] E-mail: Oleksandr.V.Lutsenko@lpnu.ua, ORCID: 0009-0008-7644-6056
[2] E-mail: serhii.s.shcherbak@lpnu.ua, ORCID: 0000-0003-2914-2101



**Ця стаття розпочинає проникливу подорож сферою передових методів аналізу даних, які можна використовувати в сфері страхування, з акцентом на застосування та можливості графових нейронних мереж (ГНМ) у цій сфері. Стаття поділена на кілька розділів, які включають огляд існуючих і широко використовуваних підходів до представлення даних, можливі способи аналізу даних у такому представленні, глибоке занурення в концепцію ГНМ для аналізу даних, представлених у вигляді графа і застосовуваність кожного підходу в страховій галузі.**

**У першому розділі представлені дві основні концепції представлення даних, якими є широко використовувана реляційна база даних і більш сучасний підхід проектування даних у розмірностях. Потім фокус переміщується до графічного представлення даних, яке також можна використовувати**

**для аналізу даних у хмарному середовищі. Для досягнення найкращої застосовності в страховій галузі, зокрема в андеррайтингу та управлінні вимогами клієнтів, у статті аналізуються переваги кожного підходу до представлення даних, а також його недоліки. На завершення розділу подано порівняльну таблицю трьох підходів. На основі порівняльної таблиці прийнято рішення про використання графового представлення, оскільки воно дає змогу враховувати складні взаємозв'язки та залежності у даних, такі як історія страхувальників, відомості про інциденти та інформація третіх сторін, що призводить до більш точного визначення ризику. оцінки та ефективного вирішення претензій.**

**Потім у статті представлено концепцію графових нейронних мереж, доволі новий підхід, що може бути використаний для аналізу даних, представлених у формі графа за допомогою алгоритмів машинного навчання. Описано потенціал використання цього підходу для аналізу даних у сфері страхування та деякі можливі випадки використання. Переваги використання цього підходу включають можливість ефективного охоплення та використання складних зв'язків, притаманних графоструктурованим даним, а також потужну структуру для аналізу та обробки графоструктурованих даних. Однак також розглядаються потенційні недоліки підходу, такі як складність проектування та труднощі масштабування.**

**Далі в статті досліджується стратегічна інтеграція графових нейронних мереж із середовищами даних у реальному часі та динамічними даними, досліджується їх адаптивність до мінливих мережевих шаблонів і часових залежностей. Ми обговорюємо, наскільки ця адаптивність має першочергове значення в таких контекстах, як прийняття рішень у режимі реального часу та прогнозний аналіз, які мають вирішальне значення для збереження гнучкості в ринковому ландшафті, що швидко змінюється.**

**Пізніше у статті надано конкретні приклади застосування ГНМ у сфері страхування, включно з процесом виписування компенсації і детальним описом процесу андеррайтингу. Крім того, надається спрощене математичне формулювання процесу андеррайтингу, яке детально пояснює роль ГНМ у розвитку науки з їхньою здатністю включати атрибути вузлів, інформацію про межі та структуру графа в комплексний алгоритм оцінки ризику.**

**Стаття закінчується висновком, що з новими технологіями представлення графів може стати новим стандартом для аналізу даних у хмарному середовищі, особливо для сфери страхування, наголошуючи на ключовій ролі ГНМ у навігації між складними взаємопов'язаними динамічними даними та виступаючи за продовження досліджень і розробок, щоб розкрити ще більший потенціал у різних секторах.**

**Ключові слова: Реляційна база даних, Розмірність відношення, Граф, Графові Нейронні Мережі (ГНМ), Андеррайтинг, Страхування.**