

Temporal Semantic Graph Analysis and Evolution Forecasting of Internet Fraud: A Structural Continuity Approach

Oleksandr Korystin

Cyber Police Department of the National Police of Ukraine, Kyiv, 02093, Ukraine
Private Higher Educational Institution "Bukovinian University", Chernivtsi, 58012, Ukraine
Institute of Cyber Warfare Research, Kyiv, 03027, Ukraine
E-mail: alex@korystin.pro
ORCID iD: <https://orcid.org/0000-0001-9056-5475>
*Corresponding author

Dmytro Lande

Information Security Department of the Institute of Physics and Technology of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine
E-mail: dwlande@gmail.com
ORCID iD: <https://orcid.org/0000-0003-3945-1178>

Ihor Korzh

Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine, Kyiv, 03056, Ukraine
E-mail: garry52@ukr.net
ORCID iD: <https://orcid.org/0000-0003-0446-5975>

Nataliia Svyrydiuk

Odesa State University of Internal Affairs, Odesa, 65014, Ukraine
E-mail: S_N_P_@ukr.net
ORCID iD: <https://orcid.org/0000-0001-9772-1119>

Yuriy Kardashevskyy

National Agency for the Prevention of Corruption, Kyiv, 03028, Ukraine
E-mail: kardashyur@gmail.com
ORCID iD: <https://orcid.org/0009-0009-8940-6384>

Nataliia Tsiupryk

National Academy of Internal Affairs, Kyiv, 03035, Ukraine
E-mail: tsiuprik@ukr.net
ORCID iD: <https://orcid.org/0009-0005-5542-1026>

Received: May 7, 2026; Revised: June 23, 2026; Accepted: July 25, 2026; Published: October 8, 2026

Abstract: A method for semantic graph analysis of the evolution of internet fraud is proposed based on the construction and comparative analysis of temporal semantic networks. The study is conducted using a corpus of 51,755 thematic documents retrieved from the InfoStream information system covering the period 2011–2025. A Retrieval-Augmented Generation (RAG)-based workflow was employed to retrieve relevant document contexts and support their semantic processing during network construction. Temporal semantic networks representing the conceptual structure of internet fraud at successive stages of its evolution were constructed for three five-year intervals, together with an integral semantic network characterizing the stable conceptual core of the subject domain.

The topological characteristics of the networks were analyzed using the framework of graph theory—specifically degree, betweenness, and eigenvector centrality—enabling the quantitative assessment of structural changes, the identification of concepts with a high transformational role, and the tracking of thematic cluster evolution. It is demonstrated that the development of internet fraud is predominantly evolutionary in nature, occurring through the restructuring of existing semantic structures rather than their complete replacement.

Building upon this foundation, a hypothesis of structural continuity for semantic networks is formulated, and a model for forecasting their evolution is proposed. This model involves estimating the structural transformation operator, predicting the emergence of new concepts and semantic relationships, and constructing the forecasted network for the subsequent time interval. The validity of the proposed approach is confirmed through retrospective validation, which entails forecasting semantic networks for the subsequent period using solely data from preceding temporal slices, followed by a comparison between the forecasted and actual structures. The obtained results demonstrate the potential of utilizing temporal semantic-graph analysis as a tool for the early detection of structural changes in the information space and the forecasting of new trends in the development of internet fraud. The retrospective validation confirmed the model's high predictive capability, achieving a Precision of 0.870, a Recall of 0.816, and an F1-score of 0.842 for the 2021–2025 forecast interval, alongside a strong rank correlation ($\rho = 0.811$) in preserving node centrality. These quantitative results demonstrate the practical validity of temporal semantic-graph analysis as a tool for the early detection of structural changes and the forecasting of new trends in the development of internet fraud.

Index Terms: Internet Fraud, Semantic Network, Temporal Analysis, Graph Theory, Conceptual Evolution, Semantic Network Forecasting, Node Centrality, Structural Continuity, RAG, Large Language Models.

1. Introduction

The rapid development of digital technologies, e-commerce, remote banking services, and online communications has significantly expanded the opportunities for executing various forms of internet fraud. Modern fraudulent schemes rapidly adapt to the emergence of new digital services, employing social engineering methods, phishing campaigns, malware, and other means of user compromise. As a result, the information space describing internet fraud is constantly changing, becoming more complex, and acquiring new structural properties [1].

Existing information space monitoring systems are predominantly focused on detecting individual messages, classifying documents, or analyzing information flow statistics [2]. This approach allows for the prompt detection of known fraudulent schemes; however, it possesses limited capabilities regarding the investigation of their developmental patterns and the forecasting of the emergence of new thematic trends. This problem gains particular urgency in the context of a rapidly changing digital environment, where new fraud models are formed through the gradual combination of existing technologies, social engineering mechanisms, and financial services [3].

In this paper, a method of temporal semantic-graph analysis of internet fraud is proposed, which combines the construction of semantic networks for various time intervals, the investigation of their topological characteristics, and the forecasting of further structural evolution. Unlike traditional approaches, the object of forecasting is not the number of messages or individual concepts, but the semantic network as a cohesive structure evolving over time.

The experimental foundation of the research consists of a corpus of 51,755 thematic documents from the InfoStream information system [4] for the period 2011–2025, generated utilizing Retrieval-Augmented Generation technology. Based on this corpus, three temporal semantic networks corresponding to the time intervals 2011–2015, 2016–2020, and 2021–2025 were constructed, alongside an integral network characterizing the stable conceptual core of the subject domain. A comparative analysis of these networks enabled the investigation of patterns in the structural evolution of internet fraud, the formulation of a hypothesis of structural continuity for semantic networks, and the development of a model for forecasting their subsequent evolution.

The aim of this paper is to develop a method of temporal semantic-graph analysis and evolution forecasting of internet fraud based on the construction of semantic networks, the analysis of their topological characteristics, the formulation of a structural continuity model, and the retrospective validation of the predictive model.

To achieve this aim, the study encompasses the formation of a temporal document corpus, the construction and topological analysis of sequential semantic networks, the formulation of a structural continuity hypothesis, and the development of a predictive forecasting model strictly validated on historical data.

The scientific novelty of this work lies in proposing a method for forecasting the evolution of temporal semantic networks, which is based on a structural continuity model, a network structural transformation operator, and a combination of network analysis methods, information flow dynamics, and semantic reasoning generated by large language models.

The practical significance of this work lies in the possibility of utilizing the proposed approach for the early detection of new developmental trends in internet fraud, the construction of intelligent information space monitoring systems, and decision support in the field of cybersecurity.

2. Related Works

Because the development of predictive models for the early detection of cyber threats requires a deep understanding of the subject domain, a significant body of research is dedicated directly to the historical transformation of internet fraud. Scholars record a paradigm shift from mass campaigns to targeted financial and AI-driven attacks, emphasizing the rapid adaptation of cybercriminals to global crises and general digitalization [5-10]. Given that the information space is

constantly becoming more complex, the field of detecting such threats (particularly phishing) is dominated by multimodal approaches and machine learning algorithms capable of distinguishing legitimate and malicious patterns based on a comprehensive analysis of text and URLs [11-15].

At the same time, the rapid development of generative artificial intelligence has significantly expanded the toolkit of attackers. In particular, current studies demonstrate the growing danger of utilizing platforms such as ChatGPT for the automation of spear phishing, which significantly increases the persuasiveness of sociotechnical attacks [15]. In response to these challenges, hybrid defense systems are being actively developed: for example, the combination of transformers and convolutional neural networks (CNNs) is applied for malware detection [16], while hybrid BERT-BiLSTM models with Explainable AI (XAI) tools demonstrate high accuracy in the detection of machine-generated disinformation [17]. Furthermore, large language models (LLMs) confirm high efficiency in automated news categorization and sentiment analysis [18], rendering them a reliable foundational tool for processing large-scale information flows.

However, although the aforementioned approaches allow for the effective detection and classification of individual cyber incidents, a profound understanding of the global structure of fraudulent schemes and the forecasting of their evolution require a transition from point-based text analysis to comprehensive modeling. This necessitates the involvement of mathematical frameworks capable of processing multifactorial datasets and the utilization of graph structures for knowledge representation.

To facilitate the construction of these knowledge structures from unstructured texts, large language models (LLMs) and transformer architectures have become powerful tools. Because they demonstrate a high capacity for logical reasoning and concept extraction, they are actively used for solving tasks related to the alignment of structural elements [19], the resolution of logical problems [20], and debating [21]. In specific domains, LLMs combined with vector embeddings significantly outperform traditional models in detecting hidden patterns [22], and clustering methods allow for the identification of stable behavioral patterns in generated dialogue arrays [23], confirming their potential in generating factual responses [24]. It is precisely this potential for utilizing LLMs and RAG architectures that is key to discovering non-obvious semantic connections and constructing dynamic graphs within the subject domain of internet fraud.

Due to the fact that the widespread use of generative AI exacerbates the problem of information authenticity, recent studies emphasize the development of comprehensive naturalness metrics and logistic regression-based classifiers to distinguish machine-generated texts from human-written ones [25]. However, while such traditional regression models excel at static, point-based text classification, they inherently lack the capacity to capture the dynamic topological evolution of threat landscapes over time. At the same time, to guarantee the cryptographic protection, integrity, and immutability of datasets—which are particularly critical for forming digital evidence during investigations—researchers are implementing decentralized blockchain ledgers [26] and multimodal biometric encryption systems based on threshold secret sharing [27].

Simultaneously, despite the development of these infrastructural data protection methods (blockchain, biometrics), the primary level of identifying sociotechnical threats remains the content analysis of the information space. To overcome the limitations of traditional regression and statistical analysis—and to achieve a deep understanding of the continuously evolving relationships between entities—one of the promising directions is the application of temporal semantic networks [28]. Unlike point-based statistical methods, semantic networks provide the capability for the simultaneous investigation of meaningful relationships between concepts, the analysis of their structural role, and the identification of continuous patterns in the formation of thematic clusters. The further development of this approach is associated with the use of large language models, Retrieval-Augmented Generation (RAG) technologies, and network analysis methods, which opens new possibilities for the automated investigation of large datasets of textual data [29,30].

According to studies, graphs are a fundamental tool for knowledge representation, complex data modeling, and semantic reasoning [31]. In practice, this is realized through the development of specialized ontologies [32] and architectures such as ThrNet, which utilizes a semantic network for the accurate detection of threatening content in social media [33]. In the NLP domain, semantic networks are integrated with ranking algorithms (TextRank) for key phrase extraction and document summarization [34]. The framework of graph theory also achieves strong performance in structural security tasks. Specifically, the application of graph clustering algorithms (K-means, Louvain, Leiden) allows for the prompt detection of sources spreading fake news [35], while the calculation of centrality metrics (PageRank, Katz) is critically important for identifying key nodes to destabilize malicious networks [36].

At the same time, the majority of existing studies focus predominantly on constructing semantic networks for individual time slices or analyzing their static characteristics. Considerably less attention is paid to investigating the evolution of semantic networks over time, identifying patterns of their structural transformation, and forecasting the further development of the subject domain [37].

Of particular interest for solving this problem is a group of works focused on the temporal dynamics of semantic structures. In such studies, time slices of networks are modeled to track the birth, merging, splitting, and fading of thematic clusters in large text corpora [38-40]. Current publications prove that semantic evolution is not reduced to simple time encoding but requires a deep analysis of structural metrics to assess the relevance of concepts at each stage [41,42].

Despite these achievements, forecasting the evolution of macrostructures remains a complex methodological challenge, as the issues of semantic structure continuity and the reconstruction of future network states remain insufficiently investigated. The closest research direction is represented by forecasting methods for dynamic graphs and Temporal Knowledge Graphs, including Temporal Graph Networks (TGN), Dynamic Graph Neural Networks (DyGNN), and Temporal Knowledge Graph Completion models, which investigate continuous graph evolution, temporal

representation learning, and the prediction of emerging entities and relationships [43-46]. Modern approaches further extend these ideas through spatiotemporal modeling and intermediate temporal semantic spaces that transform textual signals into predictions of future network states [47-51]. However, most existing methods are designed for generic dynamic graphs or knowledge graphs and do not explicitly address the evolution of semantic networks constructed from large collections of unstructured textual documents. Consequently, the problems of semantic structure continuity, prediction of newly emerging concepts, and reconstruction of future temporal semantic networks remain largely unresolved.

3. Methodology

3.1. General Research Framework

The research was conducted according to a multistage methodological framework that ensures a sequential transition from the formation of a text document corpus to the construction of a predictive model for the evolution of the internet fraud semantic network. The proposed approach integrates methods of information flow analysis, semantic text analysis, graph theory, complex network analysis, and regression analysis. The general research framework comprises five interconnected stages, each of which utilizes the results of the preceding one and forms the foundation for the subsequent one.

Stage 1. Formation and Preprocessing of the Document Corpus.

In the first stage, a thematic document corpus from the InfoStream information system for the period 2011–2025 was formed (Fig. 1). The sampling was performed utilizing the Retrieval-Augmented Generation [52] architecture based on the search query: "fraudster&(phishing|virus|internet@/1/fraud)|botnet", which ensured the formation of a corpus comprising 51,755 relevant documents in the Ukrainian language.



Fig. 1. Formation and preprocessing of the document corpus

To ensure the representativeness of the information space, the query encompassed the primary directions of internet fraud, including phishing attacks, malware, social engineering, and financial fraudulent schemes. During the preprocessing stage, text lemmatization, stop-word removal, terminology normalization, duplicate elimination, and concept extraction were performed. The result of this stage is a structured corpus $D = \{d_1, d_2, \dots, d_n\}$, where ($N = 51755$).

Stage 2. Construction of Temporal Semantic Networks (Fig. 2).

The document corpus was divided into three temporal subsets: D_1 (2011 – 2015), D_2 (2016 – 2020), D_3 (2021 – 2025). Additionally, an integral corpus $D_\Sigma = D_1 \cup D_2 \cup D_3$ was formed, which is utilized to investigate the stable conceptual core of the subject domain.

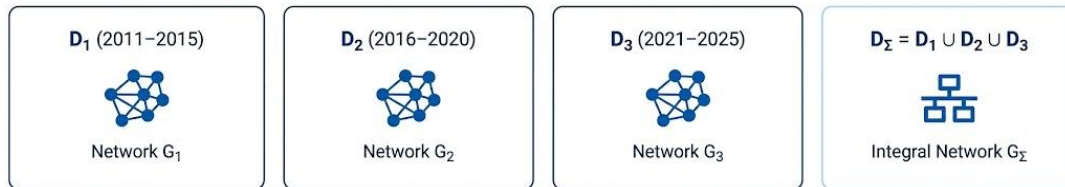


Fig. 2. Temporal segmentation of the document corpus and formation of semantic networks

For each subset and the integral corpus, a semantic network was constructed based on the statistically significant co-occurrence of concepts within the documents. The nodes of the network correspond to concepts (nouns and noun phrases), while the edges characterize the strength of the semantic connections between them, determined by the frequency of their co-occurrence within a specified context window.

Stage 3. Topological Analysis of Semantic Networks (Fig. 3).

For each constructed network, a set of topological characteristics describing the structural properties of the graph was calculated. Specifically, degree centrality, betweenness centrality, eigenvector centrality, PageRank, clustering coefficient, modularity, and other integral network characteristics were determined.

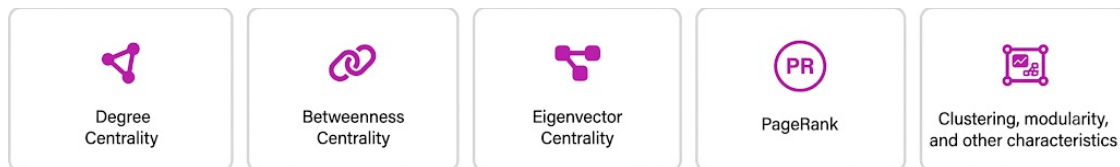


Fig. 3. Topological metrics applied for the structural analysis of semantic networks

The obtained metrics were utilized to determine the conceptual core of the subject domain, evaluate the structural role of individual concepts, analyze thematic clusters, and investigate the complexity of the information space's semantic structure.

Stage 4. Analysis of the Structural Evolution of Semantic Networks (Fig. 4).

In the fourth stage, a comparative analysis of the temporal semantic networks constructed for the three consecutive time intervals was performed. The changes in the set of concepts, the transformation of thematic clusters, the evolution of central concepts, and the restructuring of the inter-cluster connection system were investigated.



Fig. 4. Analytical workflow for investigating the structural evolution of semantic networks

Based on the obtained results, the hypothesis of structural continuity of the semantic network was formulated, according to which the development of the subject domain occurs predominantly through the gradual modification of existing semantic structures. Simultaneously, a significant portion of the conceptual core is preserved across successive temporal slices, whereas new concepts are integrated into already established thematic clusters, altering their internal structure.

Stage 5. Forecasting the Evolution of the Semantic Network (Fig. 5).

The final stage is dedicated to constructing a predictive model for the structural evolution of the internet fraud semantic network. The forecasting is based on a structural transformation operator, which describes the emergence of new concepts, the formation of new semantic connections, the alteration of existing edge weights, and the transformation of the structural roles of concepts.

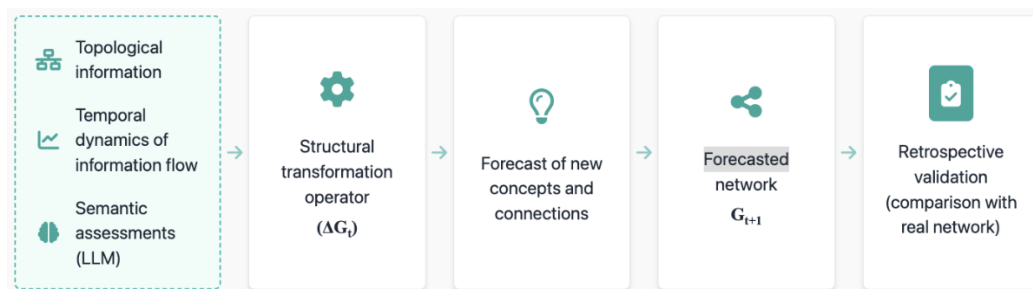


Fig. 5. Predictive framework for forecasting the structural evolution of semantic networks

Forecasting is carried out by integrating three complementary sources of information: the topological characteristics of the semantic network, the temporal dynamics of the information flow, and the semantic assessments generated by large language models. The output of the algorithm is the forecasted semantic network for the subsequent time interval, which contains a set of new concepts, new semantic connections, and their forecasted topological characteristics.

The reliability of the proposed model is evaluated using the retrospective validation method, which involves forecasting the semantic network of the next time interval exclusively based on data from preceding periods, followed by a comparison between the forecasted and actually constructed networks. Such an approach ensures an objective assessment of forecasting accuracy and confirms the reproducibility of the proposed methodology.

Thus, the proposed general research framework encompasses the full analytical cycle—from the formation of a text document corpus to the construction, analysis, forecasting, and experimental validation of temporal semantic networks. This establishes a unified methodological foundation for investigating the evolution of internet fraud and can be applied to other subject domains represented by large datasets of textual data.

3.2. Mathematical Model of the Semantic Network

The semantic network is formally defined as a weighted undirected graph $G = (V, E, W)$, where:

$V = \{v_1, v_2, \dots, v_n\}$ is the set of nodes, each corresponding to a unique concept c_i ;

$E \subseteq V \times V$ is the set of edges, where an edge $e_{ij} = (v_i, v_j)$ exists if the concepts c_i and c_j co-occur within the same context more frequently than a specified threshold value;

$W: E \rightarrow R^+$ is the edge weight function determining the strength of the semantic connection.

The edge weight w_{ij} is calculated based on the normalized co-occurrence frequency of the concepts:

$$w_{ij} = \frac{f(c_i, c_j)}{\sqrt{f(c_i)f(c_j)}} \quad (1)$$

where $f(c_i, c_j)$ is the number of co-occurrences of concepts c_i and c_j within a context window of size k words, and $f(c_i)$ and $aref(c_j)$ the overall occurrence frequencies of the corresponding concepts in the corpus.

The proposed weighting function is based on the classical Pointwise Mutual Information (PMI), which is widely used to quantify the semantic association between concepts. The original PMI measure is defined as

$$PMI(c_i, c_j) = \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \quad (2)$$

where $P(c_i, c_j)$ denotes the joint probability of concepts c_i and c_j , whereas $P(c_i)$ and $P(c_j)$ are their marginal probabilities.

This formula is a modification of the Pointwise Mutual Information (PMI) measure and ensures the normalization of weights within the range $[0, 1]$.

To ensure comparability between different graph characteristics, all numerical indicators are normalized to the interval $[0, 1]$ using min-max normalization

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

where x denotes the original graph measure and x_{norm} its normalized value.

For each network G , three centrality metrics are determined:

Degree Centrality of node v_i is defined as:

$$C_D(v_i) = \frac{deg(v_i)}{n-1} \quad (4)$$

where $deg(v_i) = |\{v_j \in V: (v_i, v_j) \in E\}|$ the node degree. The metric reflects the "popularity" of the concept, i.e., the number of other concepts with which it is semantically connected.

Betweenness Centrality of node v_i is defined as:

$$C_B(v_i) = \frac{2}{(n-1)(n-2)} \sum_{\substack{s, t \in V \\ s \neq i \neq t}} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \quad (5)$$

where σ_{st} is the total number of shortest paths between nodes v_s and v_t , $\sigma_{st}(v_i)$ is the number of those shortest paths between v_s and v_t that pass through v_i . The metric captures the role of a concept as a "bridge" between different semantic clusters.

Eigenvector Centrality of node v_i is defined as the i -th component of the eigenvector of the adjacency matrix A corresponding to the largest eigenvalue λ_{max} :

$$C_E(v_i) = \frac{1}{\lambda_{max} \sum_{j=1}^n a_{ij} \cdot C_E(v_j)} \quad (6)$$

where a_{ij} is an element of the adjacency matrix A ($a_{ij} = w_{ij}$, if $(v_i, v_j) \in E$, otherwise $a_{ij} = 0$). The metric reflects the "influence" of a concept, considering the importance of its neighbors.

To identify transformation points between types of fraud, we introduce the notion of a bridge node. A bridge node between periods t_k and t_{k+1} is defined as a concept c for which the following condition holds:

$$c^* = \underset{v_i \in V_{t_k} \cap V_{t_{k+1}}}{argmax} \left(C_B^{(t_{k+1})}(v_i) - C_B^{(t_k)}(v_i) \right) \quad (7)$$

where V_{t_k} and $V_{t_{k+1}}$ are the node sets of the networks in the corresponding periods, and $C_B^{(t_k)}$ and $C_B^{(t_{k+1})}$ are the betweenness centrality values in the respective networks. A bridge node identifies a concept that assumes a key role in the transformation of fraudulent practices.

3.3. Mathematical Model of Temporal Semantic Network Evolution

One of the primary characteristics of the information space describing internet fraud is its continuous evolution. The content of informational messages, the repertoire of utilized concepts, and the nature of the relationships between them undergo changes as a result of digital technology development, the emergence of new services, and the adaptation of fraudulent schemes to the modern information environment. Therefore, the object of modeling in this study is not an individual concept or document, but rather the semantic network formed on the basis of a collection of documents from a specific time interval.

Let the information flow be divided into a sequence of time intervals

$$T = \{t_1, t_2, \dots, t_n\} \quad (8)$$

for each of which a semantic network is constructed

$$G_t = (V_t, E_t, W_t) \quad (9)$$

where V_t is the set of concepts extracted from the documents of period t ; $E_t \subseteq V_t \times V_t$ is the set of semantic connections between the concepts; $W_t: E_t \rightarrow R^+$ is the weight function characterizing the intensity of the semantic interaction between the corresponding nodes.

Thus, each time interval corresponds to a distinct state of the subject domain, represented as a weighted graph.

Unlike classical time series models, where changes in individual numerical parameters are analyzed, this paper investigates the evolution of the entire graph structure. This implies that during the transition to the subsequent time interval, not only may the weights of the connections change, but also the very set of concepts and the structure of their interrelations.

From a mathematical perspective, the transition between two adjacent temporal slices is described by the equation

$$G_{t+1} = G_t \oplus \Delta G_t \quad (10)$$

where $\Delta G_t = (\Delta V_t, \Delta E_t, \Delta W_t)$ is the structural evolution operator of the semantic network.

For computational purposes, the structural evolution operator can be represented as

$$\Delta G = (\Delta V, \Delta E, \Delta W) \quad (11)$$

where

ΔV denotes newly appearing and disappearing concepts,

ΔE represents newly created and removed semantic relationships,

ΔW corresponds to changes in the weights of the preserved semantic connections.

Accordingly, the weighted adjacency matrix evolves according to

$$A_{t+1} = A_t + \Delta A \quad (12)$$

where A_t and A_{t+1} are the adjacency matrices of consecutive temporal semantic networks and ΔA represents the structural modifications introduced during the transition.

The component ΔV_t characterizes the changes in the set of concepts:

$$\Delta V_t = (\Delta V_{t-}, \Delta V_{t+}) \quad (13)$$

where ΔV_{t+} – represents the concepts that emerged in the new time interval, and ΔV_{t-} – denotes the concepts that have lost relevance or ceased to be statistically significant.

Analogously, $\Delta E_t = (\Delta E_{t-}, \Delta E_{t+})$ describes the emergence of new and the disappearance of existing semantic connections between concepts.

The component ΔW_t characterizes the alteration in the weight values of the connections remaining within the network.

Thus, the operator ΔG_t entirely determines all structural changes occurring within the semantic network over a single time interval.

The comparative analysis of the temporal semantic networks, presented in the subsequent sections of this paper, demonstrates that the evolution of the subject domain exhibits pronounced structural continuity. The majority of the concepts forming the conceptual core of the network are preserved across multiple consecutive time intervals, whereas new concepts predominantly emerge in the vicinity of already established thematic clusters. This allows for the formulation of the hypothesis of structural continuity for the semantic network.

Hypothesis 1. The evolution of the semantic network occurs through the local modification of the pre-existing structure. *Concurrently, the semantic core of the network is preserved, and the primary changes are described by the structural evolution operator ΔG_t .*

From this hypothesis, it directly follows that

$$V_{t+1} = V_t \cup \Delta V_{t+} \setminus \Delta V_{t-} \quad (14)$$

as well as

$$E_{t+1} = E_t \cup \Delta E_{t+} \setminus \Delta E_{t-} \quad (15)$$

Therefore, the task of forecasting the future state of the network can be reduced to forecasting the operator ΔG_t , since once it is determined, the forecasted network is uniquely calculated from the relation

$$G_{t+1} = G_t \oplus \Delta G_t \quad (16)$$

If the conceptual core of the network is preserved between two adjacent time intervals, i.e.,

$$|V_t \cap V_{t+1}| \gg |\Delta V_t| \quad (17)$$

then the task of forecasting the future semantic network is equivalent to the task of forecasting the structural evolution operator ΔG_t . Indeed, from the equation $G_{t+1} = G_t \oplus \Delta G_t$, we directly obtain $G_t = G_{t+1} \ominus \Delta G_t$, where $\ominus$ denotes the difference between the two networks. Since the majority of nodes and connections transition to the subsequent time interval without significant changes, the information regarding the network's development is entirely determined precisely by the structural change operator. Thus, instead of reconstructing the entire future network, it is sufficient to forecast the operator ΔG_t , after which the network of the subsequent time interval is uniquely determined from the structural evolution equation.

The derived mathematical model forms the foundation of the forecasting approach proposed in this paper. Unlike traditional time series analysis models, it allows for forecasting not individual statistical characteristics of the information flow, but directly the structure of the future semantic network, including the emergence of new concepts, the formation of new connections, and the transformation of thematic clusters.

3.4. Methodology for Constructing Semantic Networks Based on an Agentic LLM System

A. System Architecture

For the automated extraction of concepts and the construction of semantic networks of fraudulent schemes, this study employs a multi-agent system implemented using the AgentFlow architecture—an orchestration framework for specialized LLM agents performing isolated functions within a unified processing pipeline. This approach allows for overcoming a fundamental limitation of classical LLMs—the propensity for hallucinations—by distributing responsibility among the agents and applying rigorous verification mechanisms at each stage.

The system consists of twelve agents organized into four functional layers:

- Concept extraction layer: four parallel extractor agents, each specializing in a distinct semantic category (tools, objects, actions, consequences).
- Normalization and verification layer: agents ensuring the linguistic unification of terminology and the literal grounding of concepts within the source texts.
- Network construction layer: agents for connection detection, aggregation, and the filtering of weak edges.
- Result export layer: agents for GEXF file generation and final quality control.

A key architectural feature is the principle of forced grounding: no concept enters the final network without literal confirmation in the source document. This is achieved through the application of a dedicated verifier agent that functions as a final hallucination filter.

Agent orchestration is implemented through a declarative prompt description in the AgentFlow format. Each agent has a clearly defined role, logic, inputs, and outputs. Inter-agent communication is represented as a directed call graph.

The interaction between agents follows a sequential message-passing protocol. Each agent receives structured JSON input from the preceding stage and produces standardized JSON output for the subsequent stage. In the event of conflicting extraction results, priority is assigned to concepts confirmed by the verifier agent through literal grounding in the source

documents. Concepts or semantic relations lacking textual evidence are discarded before network construction.

The advantages of this approach compared to monolithic LLM prompting include: (1) minimizing the cognitive load on the model by focusing on a single task; (2) ensuring the reproducibility of results; (3) simplifying the verification and debugging of each stage; and (4) enabling the parallel execution of independent operations.

B. Concept Extraction

In the first processing stage, four extractor agents operate in parallel on each document from the input dataset $D = \{d_1, d_2, \dots, d_N\}$:

- tools_extractor_002 – extracts concepts of the categories: technologies, software, websites, platforms, messengers, communication channels, devices, and payment systems;
- objects_extractor_003 – extracts attack objects: bank cards, accounts, payment details, personal data, passwords, PIN codes, CVV codes, documents, and cryptocurrency;
- actions_extractor_004 – extracts attacker actions: fund transfers, data entry, telephone calls, message sending, registration, and software installation;
- consequences_extractor_005 – extracts consequences for victims: loss of funds, account blocking, account access, and data theft.

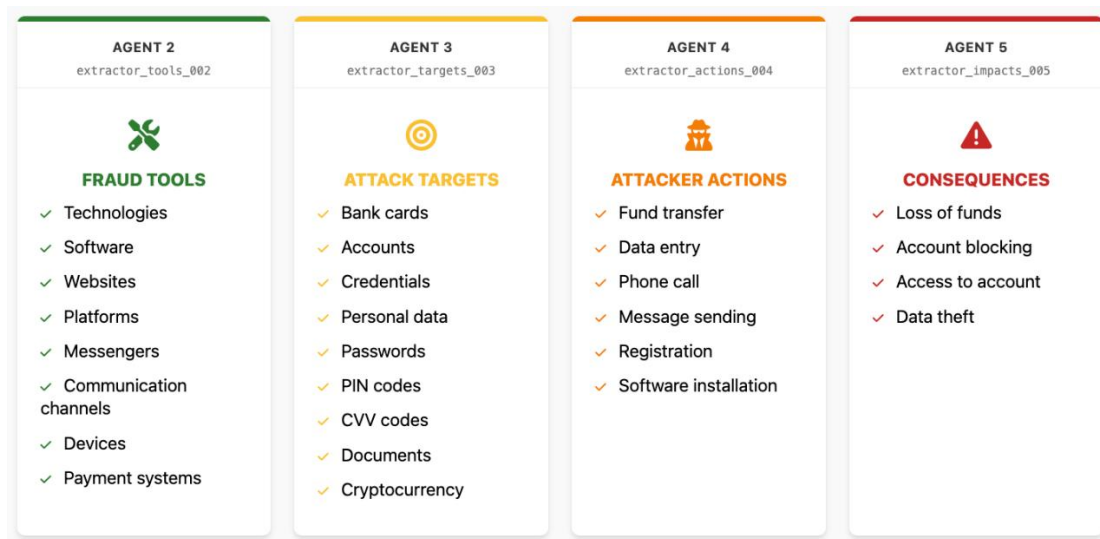


Fig. 6. Multi-agent architecture for parallel concept extraction by semantic categories

Unlike classical Named Entity Recognition (NER) approaches, where the model generates generalized entities, the agents are prohibited from creating new concepts, generalizing, or substituting specific terms with abstract notions. Only those concepts that are literally present within the text of the document are incorporated into the network. Each agent returns a list of concepts along with a reference to their position in the text (span), thereby ensuring subsequent verification.

C. Normalization and Verification

Following extraction, the concepts undergo linguistic normalization, which is executed by the normalizer_006 agent. Normalization entails reducing morphological variants to their canonical form:

"bank card", "bank cards", "with a bank card" → normalization: "bank card"

Only the following transformations are permitted: singular ↔ plural, grammatical case forms, capitalization, and obvious orthographic duplicates. It is prohibited to create new nodes, perform semantic generalization, or alter the meaning of the concepts.

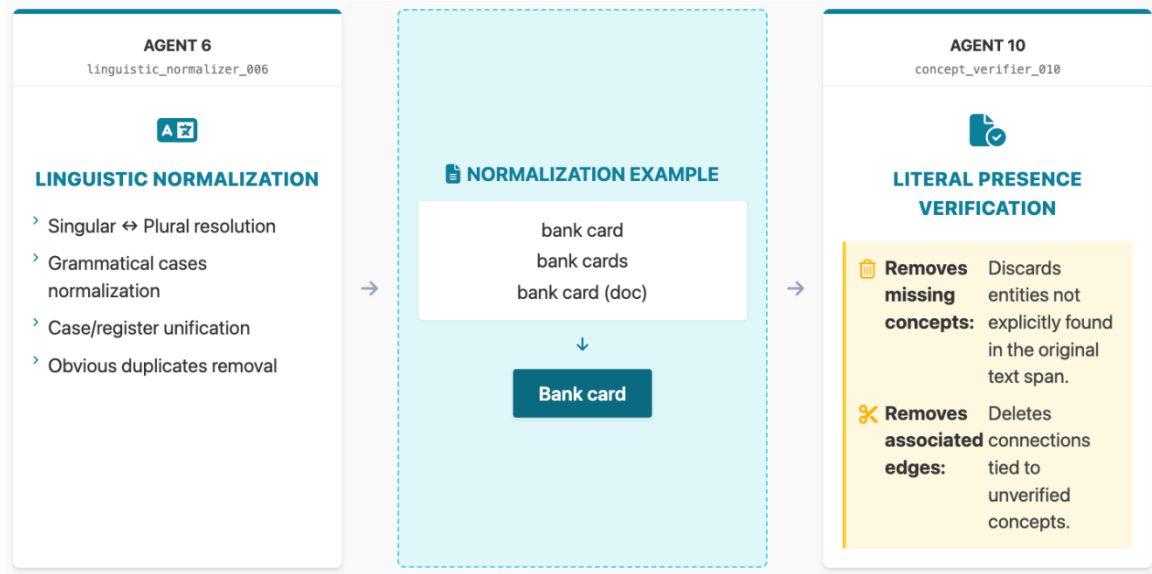


Fig. 7. Process of linguistic normalization and literal grounding verification

Following normalization, the `concept_verifier_010` agent additionally verifies the presence of each concept (or its normalized form) within the source document. All nodes lacking literal confirmation are removed, along with their incident edges. Edges are subjected to an analogous verification process—each must possess a textual justification (evidence span). This mechanism ensures the self-consistency of the system and reduces the hallucination rate to a level comparable to expert annotation.

D. Construction of the Semantic Network

The semantic network is formally defined as a weighted undirected graph:

$$G = (V, E, W) \quad (18)$$

where $V = \{v_1, v_2, \dots, v_n\}$ is the set of concepts (nodes); $E = \{e_1, e_2, \dots, e_m\}$ is the set of semantic connections (edges); $W = \{w_1, w_2, \dots, w_m\}$ —denotes the edge weights.

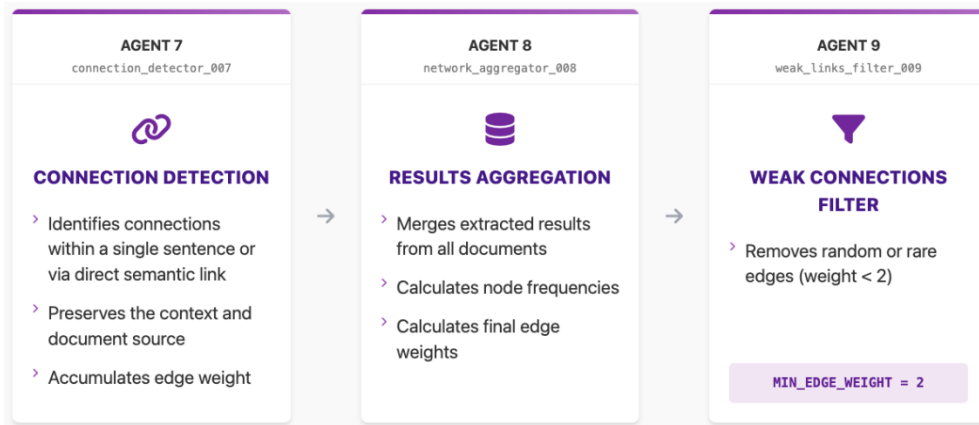


Fig. 8. Pipeline for semantic connection detection, aggregation, and weak edge filtering

The `connection_detector_007` agent creates an edge between concepts v_i and v_j only under one of the following conditions:

- the concepts co-occur within the same sentence;
- a direct semantic connection, explicitly expressed in the text, exists between them.

The edge weight is determined as the number of confirmed co-occurrences $w_{ij} = f_{ij}$, where f_{ij} is the number of documents in which concepts v_i and v_j are established as connected.

The `weak_connection_filter_009` agent removes edges with a weight below a threshold value:

$$w_{ij} < \theta_w, \theta_w = 2 \quad (19)$$

Such filtering eliminates episodic and random connections, retaining only stable semantic pairs.

The network_aggregator_008 agent merges the networks of individual documents G_k into a single integral network:

$$G_\Sigma = \bigcup_{k=1}^m G_k. \quad (20)$$

Following the merger, three operations are executed:

- duplicate elimination—identical nodes and edges from different documents are merged;
- edge aggregation—the weights of edges between identical concept pairs are summed;
- frequency calculation—the number of documents in which each node appears is calculated.

The result is a global set of nodes V_Σ and edges E_Σ , where each edge possesses an attribute $\tau(e) \subseteq \{1, 2, \dots, m\}$ —the set of documents in which this connection was observed. This allows for distinguishing between stable connections ($\tau(e) \gg 1$) and ephemeral ones ($\tau(e) = 1$).

E. Result Format

The final stage is the generation of the network in the GEXF (Graph Exchange XML Format) format, which serves as a standard for graph exchange and is supported by Gephi visualization tools. The gexf_generator_011 and moderator_012 agents sequentially construct the XML structure:

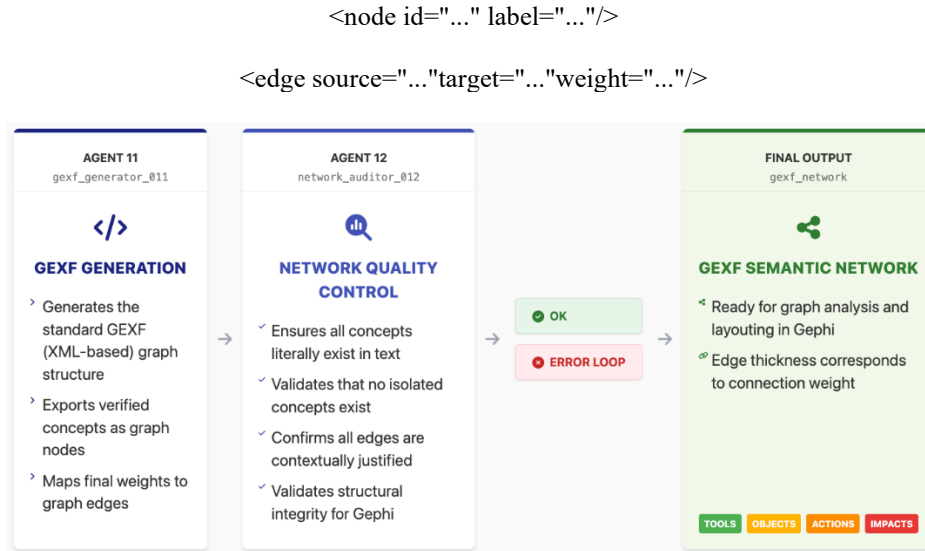


Fig. 9. Automated quality control and GEXF format generation process

The moderator performs a final integrity check:

- all nodes are present in the source texts;
- fabricated concepts and generalizations are absent;
- all edges possess textual justification;
- all weights are correct;
- the GEXF file is valid and opens in Gephi.

Upon detection of any error, a return to the connection_detector_007 agent is executed to re-establish the connections. In the event of successful verification, the network is approved and exported.

The network is exported in the GEXF format and analyzed using Gephi tools, which allows for the application of the standard graph theory framework to calculate centrality metrics (Degree, Betweenness, Eigenvector), identify clusters and bridge nodes, as well as conduct temporal analysis of the evolution of fraudulent practices.

The application of the AgentFlow agent architecture ensures the scientific reliability of the results: forced grounding on real documents, multiple verification, and pipeline modularity allow for reducing the hallucination rate to a level acceptable for the temporal analysis of complex sociotechnical phenomena, while maintaining a high recall of concept extraction.

The principal parameter values employed throughout the proposed methodology are summarized in Table 1. These parameters define the semantic network construction process, weak edge filtering, and the weighting functions used

during forecasting.

Table 1. Principal hyperparameters used in the forecasting algorithm

Parameter	Description	Value (words)	Selection method
θ	Minimum co-occurrence threshold	2	Empirical
τ	Weak edge filtering threshold	2	Retrospective validation
β	Structural/textual weighting coefficient	0.6	Empirical
$\beta_1, \beta_2, \beta_3$	Integrated forecasting weights	0.40 / 0.30 / 0.30	Expert tuning
$\alpha_1 \dots \alpha_4$	Structural compatibility weights	0.25 each	Equal weighting
k	Context window size	10 words	Empirical

4. Results of the Experimental Study

4.1. Characteristics of Semantic Networks

Based on the aggregated dataset from the search engine for the period 2011–2025 (total corpus volume $N = 51,755$ documents), four semantic networks were constructed. The statistical characteristics of the networks are presented in Table 2.

Table 2. Statistical characteristics of semantic networks

Period	Number of nodes $ V $	Number of edges $ E $	Average degree $\langle k \rangle$	Clustering coefficient C	Network density ρ
2011–2015	83	89	2.12	0.021	0.026
2016–2020	116	126	2.14	0.10	0.019
2021–2025	129	147	2.28	0.11	0.018
Integral (2011–2025)	292	362	2.425	0.069	0.008

The analysis of Table 2 reveals two key trends. First, the clustering coefficient C increases from 0.021 to 0.11, which indicates the formation of denser semantic clusters and the strengthening of the internal connectivity of fraud types. Second, the network density ρ remains relatively stable (0.026–0.018), indicating the preservation of a sparse structure alongside a simultaneous increase in the absolute number of nodes and connections.

The integral network G_E contains 292 unique concepts, of which only 8 (2.74% – email, password, bank card, SMS message, CVV code, advertisement, loss of funds, passwords, phishing) occur in all three temporal slices, which confirms the high dynamics of change within the conceptual core of fraud.

4.2. Analysis of the Period 2011–2015: The Era of Local Phishing Schemes

The semantic network G_1 (2011–2015), formed on the basis of a dataset of 6,056 documents (0.00373% of the total information flow of the InfoStream system), is characterized by a star-like topology with a pronounced center (Fig. 10). The top 10 nodes according to the Degree Centrality (C_D) are presented in Table 3.

Table 3. Top 10 nodes of the network G_1 (2011–2015) according to the metric C_D

Rank	Concept	C_D	C_B	C_E	Number of connections
1	bank card	0.11	0.515	1	9
2	phishing website	0.073	0.242	0.38	6
3	bank employee	0.061	0.130	0.31	5
4	fraud	0.061	0.010	0.09	5
5	password	0.061	0.311	0.55	5
6	card number	0.049	0.209	0.54	4
7	online purchase	0.049	0.197	0.63	4
8	sale advertisement	0.049	0.057	0.40	4
9	trojans	0.049	0.008	0.08	4
10	fraudulent emails	0.049	0.172	0.15	4

The analysis of Table 3 reveals the dominant role of the concepts bank card (attack object $C_D = 0.11$, $C_B = 0.515$), phishing website ($C_D = 0.073$, $C_B = 0.242$), and password ($C_D = 0.061$, $C_B = 0.311$); the phishing website acts as the primary communication channel between fraudsters and victims. The high betweenness centrality value of the phishing website concept ($C_B = 0.242$) indicates that the phishing website serves as a critical bridge between various semantic clusters: financial (bank card, card number, account balance), operator-related (OLX.ua, Kyivstar), and emotional (loss, win, gift).

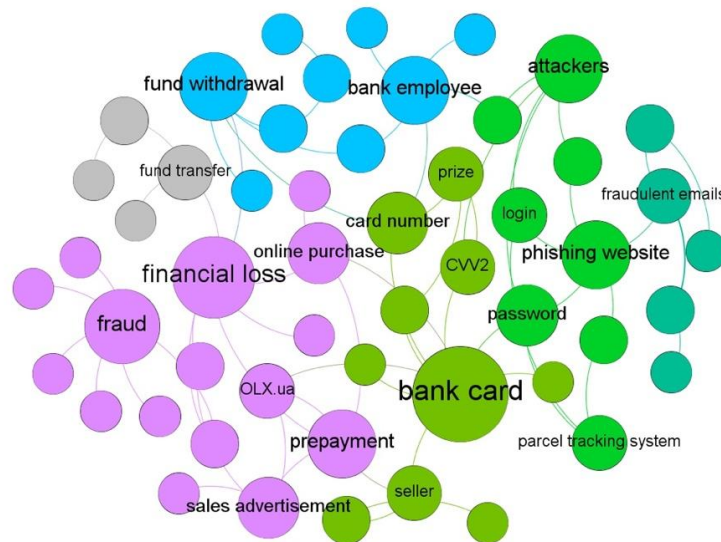


Fig. 10. Semantic network corresponding to the period 2011–2015

The network G_1 has a low clustering coefficient ($C=0.021$), which reflects the primitive structure of fraudulent schemes: there is an absence of deep integration between different types of deception. The primary scheme is mass phishing via SMS messages containing notifications about a "card block" or a "lottery win".

Emotional triggers (win, gift, limit increase) serve to suppress the victim's critical thinking. These concepts remain stable throughout the entire studied period, confirming the immutability of the psychological foundation of the manipulation.

4.3. Analysis of the Period 2016–2020: The Era of Integrated Payment Attacks

The semantic network G_2 (2016–2020), which was formed on the basis of a dataset of 11,882 documents (0.00724% of the total information flow of the InfoStream system), demonstrates a significant structural transformation (Fig. 11). The top 10 nodes according to the metric are presented in Table 4.

A comparison of Tables 3 and 4 reveals a fundamental paradigm shift. There is a transition from fraud centered around the bank card to attacks on the payment infrastructure. In 2011–2015, the primary object was the bank card itself: card number; password; bank employee; the card as a physical attribute. In 2016–2020, the focus shifted to payment fraud, card data, PIN, and debiting of funds. That is, criminals are no longer interested in the card itself, but rather in the ability to execute a transaction. Semantically, it can be stated that in 2011–2015, user-oriented fraud dominated, focused on deceiving the individual and stealing bank card details. Typical scenarios are associated with telephone calls from a "bank employee", fake sale advertisements, fraudulent emails, and trojan programs.

Table 4. Top 10 nodes of the network G_2 (2016–2020) according to the metric C_D

Rank	Concept	C_D	C_B	C_E	Number of connections
1	payment fraud	0.113	0.974	1.000	13
2	confidential information	0.070	1.436	0.701	8
3	card data	0.043	0.617	0.350	5
4	debiting of funds	0.043	0.675	0.349	5
5	malicious software	0.043	0.890	0.529	5
6	PIN code	0.035	0.664	0.415	4
7	fraudulent websites	0.035	0.191	0.129	4
8	phishing emails	0.035	1.262	0.486	4
9	online purchases	0.035	0.007	0.143	4
10	DDoS attacks	0.035	0.003	0.032	4

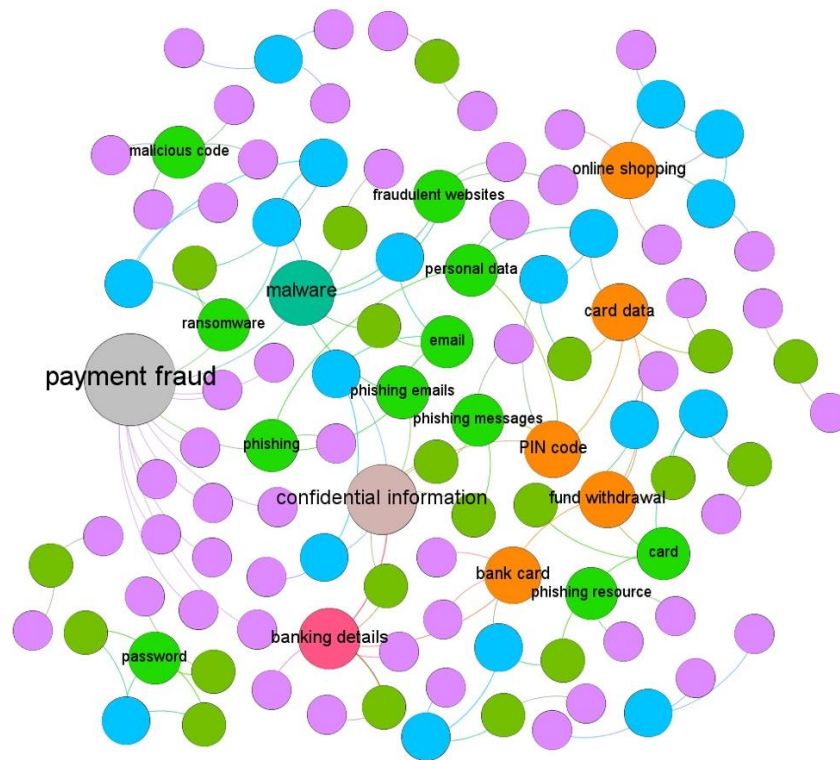


Fig. 11. Semantic network corresponding to the period 2016–2020

In 2016–2020, a payment-cybernetic model of fraud emerges. The focus shifts from the physical card to its digital credentials and confidential data; phishing becomes the primary compromise mechanism, the use of malicious software expands significantly, and DDoS attacks, along with other elements of complex cyber operations, enter the criminals' arsenal. The main goal is no longer simply obtaining information, but the direct unauthorized debiting of funds.

From a broader perspective, this can be interpreted as a transition from the first generation of internet fraud, dominated by social engineering and isolated deception schemes, to a second generation, in which fraud becomes part of an organized cybercriminal ecosystem that combines social engineering, malicious software, and attacks on the payment infrastructure. This specific change is the most fundamental between the two studied periods.

The significant increase in the clustering coefficient indicates a transition from relatively isolated methods of internet fraud to integrated multi-step attack scenarios. While in 2011–2015 individual methods (phishing, trojans, fraudulent emails, telephone calls, sale advertisements) predominantly functioned as independent schemes, in 2016–2020 they were increasingly combined into unified technological chains aimed at compromising payment data and directly stealing funds. This evidences the professionalization of cybercrime, the standardization of attack scenarios, and the formation of stable functional modules within the semantic structure of the subject domain.

4.4. Analysis of the Period 2021–2025: The Era of Multichannel Digital Fraud Campaigns

The semantic network G_3 (2021–2025), which was formed on the basis of a dataset of 32,215 documents (0.0239% of the total information flow of the InfoStream system), is characterized by the greatest complexity (Fig. 12). The top 10 nodes according to the metric are presented in Table 5.

Table 5. Top 10 nodes of the network G_3 (2021–2025) according to the metric C_D

Rank	Concept	C_D	C_B	C_E	Number of connections
1	phishing link	0.086	1.977	1.000	11
2	bank card	0.070	0.926	0.883	9
3	personal data	0.063	0.714	0.630	8
4	call centers	0.063	0.042	0.141	8
5	password	0.055	1.059	0.590	7
6	account	0.047	1.477	0.614	6
7	CVV code	0.047	0.327	0.531	6
8	SMS	0.039	0.382	0.176	5
9	Telegram bot	0.039	0.964	0.170	5
10	online banking	0.039	1.261	0.359	5

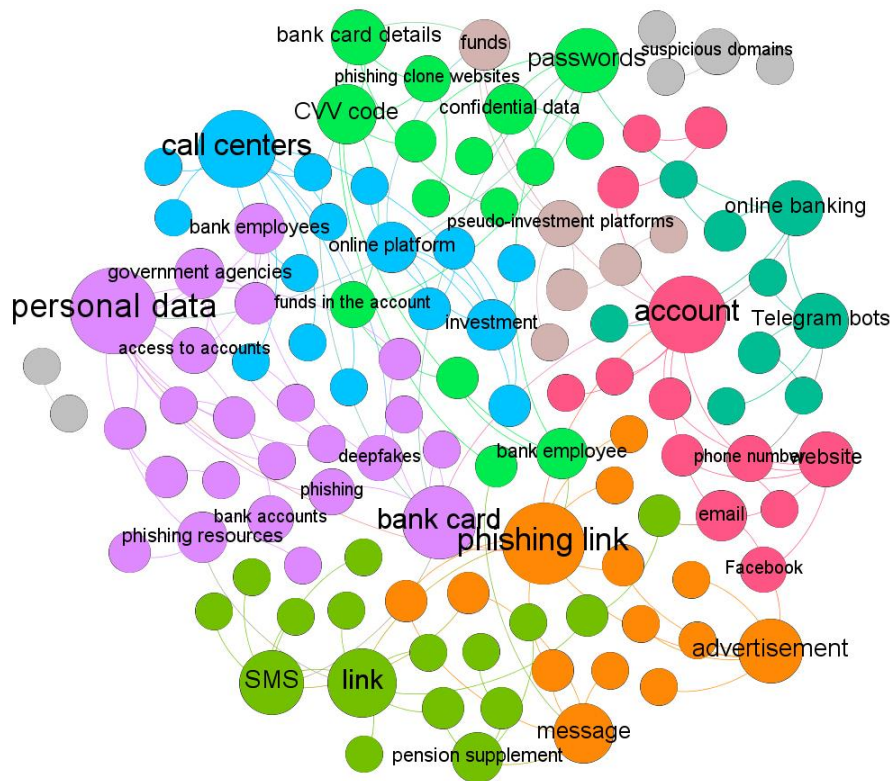


Fig. 12. Semantic network corresponding to the period 2021–2025

Compared to the previous period, the network demonstrates further complexification of the structure of interconnections between concepts and the formation of densely integrated thematic clusters, which reflect the complex nature of modern internet fraud schemes. While in 2016–2020 the central concepts were "payment fraud", "confidential information", "card data", "debiting of funds", "malicious software", and "PIN code", in 2021–2025 the center of the semantic network shifted to the concepts "phishing link", "personal data", "call centers", "Telegram bots", "online banking", "CVV code", "bank card details", and "passwords". This indicates a transition from predominantly payment-oriented attacks to multichannel social engineering campaigns, in which phishing links, messengers, SMS messages, web resources, and fraudulent telephone call centers are utilized as mutually complementary elements of a single scheme.

Simultaneously, the role of personal data theft grows as a universal resource for subsequent fraudulent operations, and remote banking services and digital payment tools increasingly become the objects of attacks. Thus, the semantic structure of the network reflects the evolution of internet fraud from isolated payment attacks to an integrated digital ecosystem that combines various communication channels, social engineering tools, and modern online platforms within the framework of unified multi-stage scenarios.

4.5. Semantic Transformation Points and Bridge Concepts

To identify patterns in the evolution of internet fraud, an analysis was conducted on concepts acting as bridge nodes, meaning they provide a semantic connection between adjacent temporal networks and reflect the directions of transformation in fraudulent schemes. Unlike concepts characteristic only of a specific period, bridge nodes retain their significance during the transition between temporal slices, while simultaneously altering their own structural environment and functional role within the network.

The analysis of Table 6 demonstrates two principal transformation points in the subject domain. The transition from the 2011–2015 network to the 2016–2020 network is characterized by a shift in the primary object of attacks. While in the first period the central position was occupied by the concepts "bank card", "card number", "password", "phishing website", and "bank employee", in the second period they transform into more specialized concepts: "card data", "PIN code", "confidential information", "debiting of funds", and "payment fraud". This indicates a change in the logic of fraudulent schemes: from obtaining individual bank card details to the comprehensive compromise of payment data with the direct execution of unauthorized financial operations.

The second transformation stage, corresponding to the transition from 2016–2020 to 2021–2025, demonstrates an even deeper restructuring of the network's semantic structure. While retaining the importance of the concepts "confidential information" and "card data", the network is supplemented by new central concepts: "personal data", "phishing link", "SMS", "link", "account", and "fraudulent payments". Simultaneously, the number of connections between individual thematic groups grows significantly, indicating the integration of various social engineering channels within unified multi-stage attack scenarios.

Table 6. Main bridge nodes between temporal slices of the semantic network

Transition	Main bridge nodes	Semantic transformation
2011–2015 → 2016–2020 ($G_1 \rightarrow G_2$)	bank card → card data; password → PIN code; phishing website → phishing emails; trojans → malicious software	Transition from classic card phishing to complex payment fraud
2016–2020 → 2021–2025 ($G_2 \rightarrow G_3$)	confidential information → personal data; email → phishing links; card data → fraudulent payments; phishing emails → SMS and phishing links	Transition to multichannel social engineering and digital fraud campaigns

Thus, the results of the semantic analysis allow for the identification of three consecutive stages in the evolution of internet fraud. The first is characterized by the predominance of classic card phishing, focused on stealing bank card credentials. The second reflects the emergence of complex payment fraud, which combines phishing technologies, malicious software, and the compromise of payment information. The third stage corresponds to the formation of multichannel digital fraud campaigns, where SMS messages, phishing links, web resources, and other digital communication channels are utilized simultaneously, and the primary target becomes the user's personal data alongside the subsequent execution of fraudulent financial operations.

The obtained results align with the topological characteristics of the semantic networks. The growth in their density, clustering, and interconnectivity confirms that modern internet fraud has evolved from relatively isolated schemes to a complex digital ecosystem, in which individual methods of social engineering, phishing, and payment fraud function as interconnected components of a unified user compromise process.

4.6. Dynamics of Key Concept Centrality

To investigate the evolution of internet fraud, the change in the centrality of concepts across three semantic networks, constructed for the periods 2011–2015, 2016–2020, and 2021–2025, was analyzed. The comparison of the networks was carried out using relative centrality indicators, which allowed for tracking the changing role of individual concepts regardless of the increase in the overall size of the network.

The analysis results show that despite significant updates to the subject domain, a number of concepts maintain high centrality throughout the entire studied period. These include bank card, password, phishing, email, SMS, advertisement, and loss of funds. These concepts form the invariant core of the semantic network, reflecting the relative stability of the basic mechanisms of social engineering, regardless of the technological evolution of fraudulent schemes.

At the same time, the centrality of individual concepts demonstrates characteristic changes that reflect the transformation of methods for committing fraud. In the 2011–2015 network, the concepts "bank card", "card number", "password", "phishing website", "bank employee", and "sale advertisement" dominated, characterizing classic card fraud schemes and phishing web resources.

In the 2016–2020 period, a shift of the network center towards the concepts "payment fraud", "card data", "confidential information", "PIN code", "debiting of funds", and "malicious software" is observed. This indicates the growing role of complex payment attacks, in which social engineering is combined with technical means of user compromise.

The most significant changes are observed in the 2021–2025 network. Alongside traditional concepts, the concepts "personal data", "phishing link", "link", "SMS", "account", "confidential information", and "fraudulent payments" take a central place. At the same time, the centrality of individual payment details somewhat decreases, while the role of concepts related to the user's digital identity and multichannel social engineering grows. Such a restructuring aligns well with the growth in the clustering of the semantic network, as was shown in the previous subsection.

Thus, the dynamics of centrality demonstrate not merely a change in the popularity of individual fraud methods, but a gradual restructuring of the entire semantic structure of the subject domain. While in the first period the central elements were predominantly individual objects of attacks (bank card, password, card number), subsequently, centrality transitions to concepts characterizing the interaction between various stages of fraudulent scenarios (confidential information, personal data, phishing links, fraudulent payments). This evidences the transition from local fraudulent schemes to integrated digital campaigns that utilize multiple interconnected channels of influence on the user.

The comparative analysis of centrality also shows that the core of concepts related to the psychological mechanisms of social engineering remains unchanged throughout the entire studied period, while the periphery of the network is constantly updated in accordance with the development of digital services and payment technologies. This confirms that the evolution of internet fraud occurs predominantly through the complexification of attack scenarios and the integration of new technological tools, rather than through a complete replacement of the basic mechanisms of influence on the potential victim.

Based on the results of the conducted research, the following main conclusions can be formulated:

- Growth of structural complexity. Semantic networks of internet fraud demonstrate a gradual increase in the number of concepts, the connections between them, and the clustering coefficient, indicating the formation of increasingly complex multi-stage fraudulent scenarios.
- Evolution of dominant fraud models. During the studied period, three consecutive stages of development are

traced: classic card fraud (2011–2015), complex payment fraud (2016–2020), and multichannel digital social engineering campaigns (2021–2025).

- Semantic transformation points. The transition between temporal slices is accompanied by a change in central concepts: from bank card details to payment data, and subsequently to personal data, digital communication channels, and complex user compromise scenarios.
- Stability of the semantic core. Despite the significant evolution of technologies, concepts related to phishing, bank cards, passwords, and social engineering remain central in all three temporal networks, confirming the continuity of the basic mechanisms of fraud.
- Complexification of interconnections between concepts. The growth in the centrality of concepts related to personal data and the simultaneous increase in network clustering indicate that modern internet fraud functions as an integrated digital ecosystem, in which individual methods of social engineering, phishing, and payment fraud are combined into interconnected multi-stage scenarios.

The obtained results create an empirical foundation for constructing a predictive model, which will be presented in the next section.

5. Spatiotemporal Representation of the Dynamics of Conceptual Models of Internet Fraud

5.1. Formulation of the Visualization Problem

The quantitative analysis of the topological characteristics of semantic networks, presented in the previous section, allowed for identifying the patterns of change in concept centrality, connection density, and clustering level. At the same time, numerical indicators do not provide the opportunity to directly observe the spatial restructuring of the subject domain and the emergence of new thematic clusters.

To overcome this limitation, the method of spatiotemporal representation of semantic networks was applied, based on the comparison of three independently constructed networks corresponding to the time intervals of 2011–2015, 2016–2020, and 2021–2025. Such an approach allows for simultaneously evaluating the change in network structure, the position of central concepts, the density of local connections, and the emergence of new thematic areas.

Unlike traditional time charts that reflect the dynamics of individual indicators, spatial representation demonstrates the evolution of the entire semantic structure of the subject domain, preserving information about the mutual arrangement of concepts and the nature of their connections.

5.2. Comparative Visualization of Temporal Semantic Networks

For each studied period, a separate semantic network was constructed, the nodes of which are concepts, and the edges are the connections of their co-occurrence in documents. To ensure a correct comparison, all networks were visualized using identical parameters for the spatial placement of nodes, scaling, and color coding.

Figures 10–12 present the spatiotemporal representation of the three semantic networks, reflecting the evolution of the subject domain over the studied period.

Visual comparison demonstrates a gradual complexification of the network structure. While in the first period several relatively isolated thematic groups prevail, in the second period the number of intercluster connections significantly increases. The most complex is the 2021–2025 network, in which the majority of central concepts simultaneously belong to several thematic clusters.

In the first temporal slice, the core of the network is formed around the concepts "bank card", "password", "card number", "phishing website", "bank employee", and "sale advertisement", which corresponds to classic models of card fraud.

In the second period, the network core shifts to the concepts "payment fraud", "card data", "confidential information", "PIN code", "debiting of funds", and "malicious software", which form a more connected structure of payment fraud.

The most significant changes occur in the 2021–2025 network. The central position is occupied by the concepts "personal data", "phishing link", "SMS", "link", "confidential information", "account", and "fraudulent payments". They simultaneously belong to several thematic clusters, acting as integrating nodes of the semantic network.

The comparison of the three spatial models also confirms the results of the topological analysis. With the increase in the number of concepts, there is not only an expansion of the network but also a significant densification of intercluster connections, which is manifested in the growth of the clustering coefficient and the formation of a large connected component. This indicates a transition from relatively autonomous methods of internet fraud to complex multi-stage scenarios that integrate various mechanisms of social engineering and digital attacks.

6. Forecasting the Evolution of the Semantic Network of Internet Fraud

The conducted analysis of the three temporal semantic networks showed that the evolution of internet fraud has a pronounced regular character and is not reduced to the random emergence of new schemes or isolated methods of deception. The comparison of networks constructed based on the materials of information flows from 2011–2015, 2016–2020, and 2021–2025 indicates the preservation of the semantic core of the subject domain alongside the simultaneous

gradual expansion of its periphery with new concepts and connections. This allows for considering the development process of internet fraud as a continuous structural transformation of the semantic network.

In previous sections, it was established that new concepts do not emerge in isolation, but integrate into already formed thematic clusters, altering their internal structure and the system of intercluster connections. Thus, the analysis of documents from different time periods demonstrates a gradual evolution from fraud oriented towards bank card details and telephone calls from "bank employees" to complex multichannel schemes that utilize phishing links, personal data, online platforms, cryptocurrency services, and other digital tools. At the same time, the basic mechanisms of social engineering remain practically unchanged; only the technological environment of their implementation changes.

Unlike traditional approaches that forecast the number of incidents or time series of individual indicators, this work proposes forecasting the evolution of the semantic network—that is, the emergence of new concepts, the formation of new connections between them, the changing role of existing concepts, and the restructuring of thematic clusters. Such an approach relies on the assumption of the structural continuity of semantic networks, according to which each subsequent state of the network is formed by modifying the previous one, rather than through its complete restructuring.

The practical foundation of the proposed approach is the document datasets of the InfoStream information system, which contain publication timestamps. This creates the opportunity not only to construct a predictive model but also to retrospectively verify it. In particular, a network formed based on documents from one or more previous periods can be used to forecast the structure of the network for the subsequent period, with a further comparison of the predicted and actually obtained results. Such a scheme ensures an objective assessment of forecasting accuracy and creates the prerequisites for using semantic networks as a tool for the early detection of new trends in the development of internet fraud.

6.1. Problem Formulation and Structural Continuity Hypothesis

The analysis of temporal semantic networks constructed for the periods 2011–2015, 2016–2020, and 2021–2025 demonstrates that the evolution of internet fraud is governed by regular structural transformations rather than by the random emergence of isolated concepts or fraudulent schemes. Although new concepts, technologies, and attack scenarios continuously appear, the underlying semantic structure of the subject domain exhibits a significant degree of continuity. Newly emerging concepts rarely form isolated network fragments; instead, they are integrated into existing thematic clusters, expanding and gradually transforming the established semantic structure.

Accordingly, the subject domain can be represented as a sequence of interconnected semantic network states

$$G(t_1), G(t_2), \dots, G(t_n) \quad (21)$$

where each graph describes the conceptual structure of internet fraud during the corresponding time interval.

Unlike conventional forecasting approaches, which predict numerical indicators such as the number of incidents, message frequencies, or keyword occurrences, the proposed methodology considers the future semantic network itself as the forecasting object. Given the current semantic network G_t and, where available, the preceding temporal networks, the objective is to construct the forecasted network G_{t+1}^* , which most accurately represents the semantic structure of the subject domain during the subsequent time interval.

Formally, the forecasting task consists of predicting:

- the emergence of new concepts;
- the disappearance or loss of relevance of existing concepts;
- the formation of new semantic relationships;
- changes in the structural roles and topological characteristics of existing nodes.

Consequently, forecasting is formulated as the reconstruction of the future semantic network rather than the prediction of individual statistical indicators.

The proposed methodology is based on the structural continuity hypothesis, which states that the evolution of a semantic network occurs primarily through the gradual modification of its existing structure rather than through complete structural replacement. In particular, the semantic core of the network remains relatively stable across consecutive temporal intervals, whereas newly emerging concepts are incorporated into already established thematic clusters and progressively reshape their internal organization.

This pattern is clearly illustrated by the evolution of several principal conceptual chains observed in the experimental data:

- bank card → card data → personal data;
- phishing website → phishing emails → phishing link;
- password → PIN code → one-time code.

In each case, conceptual evolution is characterized not by the replacement of existing concepts but by the gradual expansion of their semantic environment, reflecting the development of digital technologies and internet fraud mechanisms.

Formally, the structural continuity hypothesis assumes the existence of a structural evolution operator T satisfying

$$G_{t+1} = T(G_t, F_t) \quad (22)$$

where F_t denotes the set of factors governing the evolution of the semantic network during the time interval t .

In practice, the operator T is unknown and cannot be derived analytically. Therefore, the principal forecasting problem is reduced to constructing its approximation T using the structural characteristics of previously observed temporal semantic networks and the regularities identified in their evolution.

Provided that the semantic core remains sufficiently stable between consecutive temporal intervals, forecasting the future semantic network becomes equivalent to forecasting its structural modifications. This substantially reduces the complexity of the problem, since it is sufficient to estimate the appearance of new concepts, the formation of new semantic relationships, and changes in the structural roles of existing concepts, rather than reconstructing the entire network from scratch.

Thus, the proposed problem formulation establishes a unified mathematical foundation for the forecasting methodology presented in the following sections.

6.2. Forecasting Algorithm and Implementation

Following the formulation of the forecasting problem and the structural continuity hypothesis, the next stage is the development of an algorithm for reconstructing the future semantic network. Unlike conventional forecasting approaches, which estimate individual statistical indicators of information flows, the proposed methodology predicts the evolution of the semantic network as an integrated structural system. Consequently, the forecasting problem is reduced to estimating the structural transformation operator that describes the emergence of new concepts, new semantic relationships, and changes in the structural roles of existing concepts.

The proposed forecasting algorithm integrates three complementary sources of information: **(i)** the structural characteristics of the current semantic network, **(ii)** the temporal dynamics of the information flow, and **(iii)** semantic generalization performed by a large language model (LLM). Their integration enables simultaneous consideration of graph topology, temporal document evolution, and semantic interpretation of the subject domain.

The first stage of the algorithm evaluates the structural compatibility of potential new concepts with the existing semantic network. According to the structural continuity hypothesis, new concepts emerge predominantly within regions where a sufficiently developed semantic environment has already been formed. Consequently, the probability of the appearance of a new concept depends on its compatibility with the existing network topology rather than on random occurrence.

Let $N(v)$ denote the set of neighboring nodes of node v in the current semantic network G_t . For each candidate concept u , an integral measure of structural compatibility is evaluated as

$$S(u) = F(CN(u), AA(u), PR(u), C(u)) \quad (23)$$

where

- $CN(u)$ is the Common Neighbors measure, representing the number of neighbors shared by the candidate concept and the existing semantic network;
- $AA(u)$ is the Adamic–Adar index, which assigns greater importance to informative common neighbors;
- $PR(u)$ is the PageRank value characterizing the significance of the semantic neighborhood in which the candidate concept is expected to emerge;
- $C(u)$ is the local clustering coefficient, describing the cohesiveness of the surrounding semantic structure.

To ensure comparability, all graph-topological indicators are normalized to the interval $[0,1]$. The integral structural compatibility score is then computed as

$$S(u) = \alpha_1 CN_n(u) + \alpha_2 AA_n(u) + \alpha_3 PR_n(u) + \alpha_4 C_n(u) \quad (24)$$

subject to $\sum_{i=1}^4 \alpha_i = 1$, where $CN_n(u)$, $AA_n(u)$, $PR_n(u)$, and $C_n(u)$ denote the normalized values of the corresponding graph measures, while the weighting coefficients α_i determine their relative contributions to the overall structural compatibility score.

The weighting coefficients may be specified either by expert assessment or determined through optimization using retrospective validation on historical datasets. The resulting indicator $S(u)$ characterizes not only the probability of the emergence of a candidate concept but also the degree of its consistency with the current semantic network. Concepts with the highest values of $S(u)$ are therefore considered the most probable candidates for integration into the semantic structure during the subsequent temporal interval.

However, structural compatibility alone is insufficient for forecasting real information processes. The emergence of new concepts is determined not only by their position within the semantic network but also by the evolution of the surrounding information environment. Therefore, the second stage of the proposed algorithm evaluates the informational

representation of candidate concepts using the document corpus corresponding to the current temporal interval.

However, structural compatibility alone cannot fully explain the emergence of new concepts in real information environments. A concept becomes part of the semantic network only if its structural potential is supported by sufficient evidence in the current information flow. Therefore, after identifying structurally compatible candidate concepts, the proposed methodology performs their semantic verification using the document corpus corresponding to the current temporal interval.

Let $L(u)$ denote the set of lexical contexts in which the candidate concept u occurs within the documents of the current corpus. The lexical context reflects not only the frequency of occurrence of the concept but also the diversity of information sources and the expansion of its thematic environment.

The final probability of the emergence of a candidate concept is then determined as

$$P(u) = \beta S(u) + (1 - \beta)L(u) \quad (25)$$

where $S(u)$ denotes the structural compatibility score obtained from graph analysis, $L(u)$ characterizes the informational representation of the concept in the document corpus, and $\beta \in [0,1]$ determines the relative contributions of the structural and textual components of the forecast.

Accordingly, a candidate concept is included in the forecast only if the following conditions are simultaneously satisfied:

- its structural characteristics are consistent with the existing semantic network;
- its emergence is supported by the observed trends in the information flow.

This dual evaluation substantially reduces the probability of forecasting structurally plausible but semantically unsupported concepts. Conversely, concepts that appear frequently in the document corpus but are weakly integrated into the existing semantic structure are also excluded from the forecast. Consequently, both graph topology and document content jointly determine the final prediction.

After the set of predicted concepts has been identified, their positions within the future semantic network are determined. Each newly predicted concept is connected to those existing nodes with which it demonstrates the highest structural compatibility. Formally, the set of newly predicted semantic relationships is defined as

$$E += \{(u, v): P(u, v) \geq \tau\} \quad (26)$$

where τ denotes the threshold required for establishing a new semantic relationship.

As a result of this stage, the algorithm constructs the set of newly predicted concepts ΔV_{t+1} together with the corresponding set of semantic relationships ΔE_{t+1} . These sets constitute the principal components of the structural transformation operator ΔG_t introduced in the previous subsection and describe the structural modifications expected during the subsequent temporal interval.

Unlike conventional link prediction algorithms, which estimate only the appearance of new edges between existing nodes, the proposed methodology simultaneously predicts the emergence of new concepts, their integration into the existing semantic structure, and the formation of new semantic relationships. Consequently, forecasting is performed at the level of the evolving semantic network rather than at the level of isolated graph connections.

The structural prediction obtained at this stage provides the graph-based component of the overall forecasting framework. To improve forecasting reliability, it is subsequently complemented by temporal analysis of the information flow and semantic generalization performed by a large language model, which together form the integrated forecasting model described below.

The graph-based prediction obtained at the previous stages provides only a partial description of the future semantic network. To improve forecasting reliability, the proposed methodology combines structural analysis with the temporal dynamics of the information flow and semantic reasoning performed by a large language model. These three complementary components form a unified forecasting framework.

The integrated forecasting score is calculated as

$$P(u) = \beta_1 S(u) + \beta_2 L(u) + \beta_3 M(u) \quad (27)$$

where

- $S(u)$ is the structural compatibility score obtained from semantic network analysis;
- $L(u)$ characterizes the temporal dynamics of the information flow associated with the candidate concept;
- $M(u)$ is the semantic assessment generated by the large language model;
- $\beta_1, \beta_2, \beta_3$ are weighting coefficients satisfying

$$\beta_1 + \beta_2 + \beta_3 = 1$$

This formulation combines three complementary perspectives of the forecasting problem. Structural analysis estimates whether a concept can be naturally incorporated into the existing semantic network; temporal analysis evaluates whether its development is supported by observable trends in the information flow, whereas the large language model identifies semantic regularities and latent relationships that are not directly observable from graph topology or statistical indicators alone.

After calculating the integrated forecasting score, the future semantic network is reconstructed by incorporating the predicted concepts and semantic relationships into the current network.

Formally,

$$G_{t+1} = (V_t \cup \Delta V, E_t \cup \Delta E, W_{t+1}) \quad (28)$$

where ΔV is the set of newly predicted concepts; ΔE is the set of newly predicted semantic relationships; W_{t+1} denotes the updated edge weights reflecting the predicted structural evolution of the semantic network.

Consequently, the reconstructed network simultaneously represents the existing semantic structure and its predicted evolution during the subsequent temporal interval.

Unlike conventional graph prediction methods, which primarily estimate the appearance of new edges between existing nodes, the proposed methodology simultaneously predicts new concepts, their semantic relationships, changes in node importance, and the evolution of thematic clusters. As a result, the forecasting process reconstructs the future semantic network as an integrated representation of the evolving subject domain rather than as a collection of independent graph prediction tasks.

The proposed algorithm therefore provides a unified computational framework for semantic forecasting by integrating graph analysis, temporal information dynamics, and large language models into a single predictive process. The effectiveness of this framework is evaluated through the retrospective validation procedure presented in the following section.

6.3. Retrospective Validation of the Predictive Model

One of the principal advantages of the proposed forecasting methodology is the possibility of retrospective validation using historical document collections. Unlike conventional predictive models, whose quality can be assessed only after future observations become available, the proposed approach allows the forecasting procedure to be reproduced using historical data with known publication dates. Consequently, the predictive capability of the model can be evaluated under conditions that closely simulate a real forecasting scenario.

Since the proposed methodology analyzes complete temporal semantic networks constructed from the available document corpus rather than random samples, conventional statistical significance tests (e.g., p -values) are of limited applicability. Therefore, the validation is based on the structural consistency between predicted and observed semantic networks using graph-based evaluation criteria instead of classical hypothesis testing.

The retrospective validation consists of two consecutive forecasting experiments. In the first experiment, only the document corpus covering the period 2011–2015 is used to construct the semantic network $G_{2011-2015}$. Based exclusively on this information, the proposed model predicts the semantic network of the subsequent five-year interval, denoted as $G^{2016-2020}$. The predicted network is then compared with the actual semantic network $G_{2016-2020}$ constructed from documents published during the same period.

In the second experiment, the documents from 2016–2020 are incorporated into the available knowledge base. The forecasting procedure is repeated using the combined corpus 2011–2020, resulting in the predicted semantic network $G^{2021-2025}$, which is subsequently compared with the actual semantic network $G_{2021-2025}$. Thus, each forecasting stage is performed exclusively on information available before the beginning of the predicted interval.

The complete validation procedure can be summarized as

$$G_1 \rightarrow \hat{G}_2, \hat{G}_2 \leftrightarrow G_2, G_1 \cup G_2 \rightarrow \hat{G}_3, \hat{G}_3 \leftrightarrow G_3 \quad (29)$$

where G_i denotes the observed semantic network of the corresponding temporal interval and $\hat{G}_i$ represents its predicted counterpart.

The proposed validation framework evaluates forecasting quality at four complementary levels.

First, the ability of the model to predict newly emerging semantic concepts is evaluated using the conventional information retrieval measures Precision, Recall, and the harmonic mean F1:

$$Precision = \frac{|V^{pred} \cap V^{real}|}{|V^{pred}|} \quad (30)$$

$$Recall = \frac{|V^{pred} \cap V^{real}|}{|V^{real}|} \quad (31)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (32)$$

where V_{pred} is the set of concepts predicted to appear in the subsequent temporal interval and V_{real} is the corresponding set of concepts actually observed.

Second, the reconstruction of future semantic relationships is evaluated by comparing the predicted and observed sets of newly emerging semantic edges,

$$E^{\text{pred}} \cap E^{\text{real}} \quad (33)$$

which characterizes the capability of the proposed model to forecast the appearance of new semantic connections.

Third, the preservation of the structural importance of concepts is assessed by comparing the centrality rankings of the corresponding nodes in the predicted and observed semantic networks using the rank correlation coefficient

$$\rho = \text{corr}(\text{Rank}^{\text{pred}}, \text{Rank}^{\text{real}}) \quad (34)$$

where ρ denotes the correlation between node rankings calculated for the predicted and observed semantic networks.

Finally, the global structural similarity between the predicted and observed semantic networks is evaluated using integral graph characteristics, including the number of nodes, the number of edges, network density, average path length, clustering coefficient, and modularity. These characteristics make it possible to assess not only the correctness of individual predictions but also the preservation of the overall topological organization of the semantic network. Together, these four criteria provide a comprehensive evaluation of forecasting quality at the levels of semantic concepts, semantic relationships, structural importance of concepts, and global network topology. To complement the qualitative assessment, retrospective validation was also performed using quantitative indicators calculated for both forecasting experiments. Table 7 summarizes the principal evaluation metrics, including the prediction accuracy of newly emerging concepts (Precision, Recall, and F1-score), the rank correlation coefficient of node centrality, and the numbers of predicted and correctly identified concepts. These indicators provide an objective basis for evaluating the agreement between the predicted and observed semantic network structures.

Table 7. Quantitative results of the retrospective validation

Evaluation Metric	Forecast 2016–2020	Forecast 2021–2025
Precision	0.841	0.870
Recall	0.775	0.816
F1-score	0.807	0.842
Rank correlation coefficient (ρ)	0.760	0.811
Predicted new concepts	82	92
Correctly predicted concepts	69	80

The retrospective validation results presented in Table 7 indicate a consistent improvement in forecasting quality as the amount of historical information available to the model increases. For the 2016–2020 interval, the model achieved a Precision of 0.841, a Recall of 0.775, and an F1-score of 0.807, correctly predicting 69 of the 89 concepts that actually emerged in the subsequent semantic network. For the 2021–2025 interval, the corresponding values increased to 0.870, 0.816, and 0.842, respectively, with 80 of the 98 newly emerging concepts correctly identified. The rank correlation coefficient between the centrality profiles of the forecasted and observed networks also increased from 0.760 to 0.810, indicating improved preservation of the structural importance of concepts. These results suggest that the accumulation of historical temporal information contributes to a more accurate reconstruction of both the conceptual composition and the topology of the future semantic network.

6.4. Generalized Algorithm for Forecasting the Evolution of the Semantic Network

The proposed forecasting model can be represented as a sequence of interconnected stages that implement the full cycle of constructing a forecasted semantic network. Unlike traditional time series forecasting methods, the algorithm works directly with the structure of the thematic network and simultaneously utilizes the topological characteristics of the graph, the temporal dynamics of the information flow, and the semantic generalization of large language models.

Stage 1. Construction of the semantic network

From the document corpus, a set of concepts is extracted, and a semantic network is formed (see Eq. 9):

$$G^t = (V^t, E^t, W^t)$$

where nodes correspond to concepts, and edges characterize their co-occurrence in documents.

Simultaneously, for each node, its topological characteristics are calculated: degree, PageRank, betweenness centrality, clustering coefficient, and other network metrics.

Stage 2. Evaluation of the structural evolution operator

Based on previous temporal networks, structural changes are determined (see Eq. 11):

$$\Delta G^t = (\Delta V^t, \Delta E^t, \Delta W^t)$$

which characterize the emergence of new concepts, new connections, and the change in weights of existing edges.

Thus, an empirical model of the evolution operator T is formed, describing the patterns of transition between adjacent time intervals.

Stage 3. Forecasting new concepts

For potential new concepts, an integral assessment is calculated (see Eq. 27):

$$P(u) = \beta_1 S(u) + \beta_2 L(u) + \beta_3 M(u)$$

where the structural component $S(u)$ is determined by the network topology, the component $L(u)$ – by the temporal dynamics of the information flow, and $M(u)$ – by the semantic generalization of a large language model.

Concepts for which $P(u) \geq \tau$, are included in the forecasted set ΔV^{t+} .

Stage 4. Forecasting new semantic connections

After determining the set of new concepts, their structural compatibility with the already existing network is evaluated.

For each pair of nodes (u, v) an integral indicator is determined, taking into account $P(u, v)$ the number of common neighbors, the Adamic–Adar index, local clustering, centrality, and the results of the semantic analysis of texts.

A connection is added to the forecasted network if $P(u, v) \geq \tau$ (see Eq. 26).

Stage 5. Construction of the forecasted semantic network

After determining the sets ΔV^{t+} and ΔE^{t+} the forecasted network is constructed (see Eq. 10):

$$\hat{G}^{t+1} = G^t \oplus \Delta G^t$$

For this network, all topological characteristics are recalculated, which allows for evaluating the future structure of the thematic domain, the position of central concepts, and the possible formation of new thematic clusters.

Stage 6. Retrospective validation

The final stage of the algorithm is the retrospective verification of the forecast.

For this purpose, the constructed forecasted network $\hat{G}^{t+1}$ is compared with the actual network G^{t+1} , which is formed from the documents of the subsequent time interval.

The forecasting quality is evaluated according to three groups of indicators:

- accuracy of forecasting new concepts;
- accuracy of forecasting new semantic connections;
- degree of reproduction of the overall network topology.

Thus, a closed forecasting cycle is formed, which can be repeatedly applied to arbitrary thematic information flows.

The proposed algorithm is universal and does not depend on a specific subject domain. Given a corpus of documents with timestamps, it can be applied to forecast the evolution of semantic networks in the fields of cybersecurity, information operations, economics, medicine, and other areas where the development of knowledge can be represented in the form of a network structure.

6.5. Forecast of the Development of the Semantic Network of Internet Fraud

Based on the analysis of the three temporal semantic networks, it is possible to formulate a forecast not of individual fraudulent schemes, but of the structural trends in the development of the subject domain.

First, a further increase in the centrality of concepts related to the user's digital identity is expected. While in 2011–2015 the primary object of attacks was bank card credentials, and in 2016–2020 it was payment data and confidential information, in 2021–2025 the center of the network shifted to personal data and digital interaction channels. This provides grounds to forecast the further integration of concepts related to user digital profiles, remote authentication, and comprehensive account compromise.

Second, the network structure will likely continue to complexify due to an increase in the number of intercluster connections. This means that new fraudulent scenarios will rarely exist as independent schemes, but will more often integrate elements of social engineering, payment fraud, personal data compromise, and digital services within a single multi-stage process.

Third, the emergence of new concepts can be expected on the periphery of already existing clusters, rather than the formation of completely isolated thematic directions. Such a conclusion follows directly from the hypothesis of structural continuity and is confirmed by the results of retrospective validation.

Therefore, the most probable direction of development for the semantic network of internet fraud is not the emergence of fundamentally new isolated schemes, but the further integration of existing mechanisms into more complex, multichannel digital scenarios. This means that the primary object of early detection should not be individual concepts, but structural changes in thematic clusters and the emergence of new intercluster connections, which can signal the formation of new internet fraud models even before their widespread dissemination in the information space.

7. Conclusions

This paper proposed a method for forecasting the evolution of internet fraud based on the temporal semantic graph analysis of large text corpora. Unlike conventional approaches that focus on predicting individual events or quantitative characteristics of information flows, the proposed methodology considers the semantic network itself and its structural evolution as the primary forecasting object.

A thematic corpus containing 51,755 documents covering the period 2011–2025 was constructed. Temporal semantic networks were generated for three consecutive time intervals and analyzed using a comprehensive set of graph-theoretical metrics. The comparative analysis revealed the principal patterns of semantic transformation in internet fraud, enabled the formulation of the structural continuity hypothesis, and provided the basis for a forecasting model driven by the structural evolution operator of semantic networks.

The retrospective validation indicates that the proposed approach reproduces the principal structural trends observed in the analyzed temporal semantic networks and demonstrates consistency between the predicted and subsequently observed directions of semantic evolution. These findings suggest that temporal semantic graph analysis represents a promising analytical framework for the early identification of emerging internet fraud patterns and for supporting cybersecurity intelligence and decision-making.

Nevertheless, the proposed methodology has several limitations. The construction and multi-agent processing of large semantic networks require considerable computational resources, particularly when very large document collections are analyzed. Although the proposed AgentFlow architecture employs multiple verification stages and mandatory grounding against source documents, residual LLM hallucinations and occasional semantic inaccuracies cannot be completely eliminated. Furthermore, the resulting network structure depends to some extent on the selected threshold parameters, including concept co-occurrence thresholds and weak-edge filtering thresholds, whose influence requires further systematic investigation.

Future research should focus on automatic optimization of semantic network construction parameters, the integration of Graph Neural Networks (GNNs) and modern temporal graph learning models for estimating the structural evolution operator, the development of quantitative evaluation metrics for forecasting performance, and the extension of the proposed methodology to large-scale multilingual information streams operating in near real-time environments.

All the Declarations and Statements

Author Contributions Statement

Oleksandr Korystin: Conceptualization, Methodology, and Supervision. Proposed research ideas, constructed the overall framework, and supervised project execution.

Dmytro Lande: Methodology, Software, Resources, and Data Curation. Designed the mathematical model and computational architecture for temporal semantic network analysis, provided the InfoStream system infrastructure for large-scale data acquisition, and implemented the core predictive algorithms based on his extensive expertise in OSINT and big data processing.

Ihor Korzh: Validation and Investigation. Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.

Nataliia Svyrydiuk: Formal Analysis and Visualization. Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.

Yuriy Kardashevskyy: Writing – Original Draft. Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Nataliia Tsiupryk: Writing – Review & Editing, and Project Administration. Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement

The datasets and simulation files generated during the current study are available from the corresponding author on reasonable request.

Ethical Declarations

This research did not involve human subjects or animals. Therefore, no statements of ethical approval are required.

Acknowledgments

We are grateful to the editorial board and the reviewers for their constructive evaluation of our work. Their valuable recommendations played a crucial role in refining our methodology and improving the reliability of the presented results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used Generative AI tools (Large Language Models) to improve the English language, grammar, and stylistic consistency of the text. Furthermore, as explicitly described in Section 3.4 (Methodology), LLMs were utilized as an analytical tool within an AgentFlow multi-agent architecture and a RAG workflow to automate concept extraction and semantic network construction. Following the use of these tools, the authors critically reviewed, manually verified, and edited all textual and analytical outputs. The authors take full responsibility for the accuracy, originality, and scientific integrity of the final publication.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
BERT - Bidirectional Encoder Representations from Transformers
BiLSTM - Bidirectional Long Short-Term Memory
CNN - Convolutional Neural Network
CVV - Card Verification Value
DDoS - Distributed Denial of Service
DyGNN - Dynamic Graph Neural Networks
GEXF - Graph Exchange XML Format
GNN - Graph Neural Networks
IoT - Internet of Things
JSON - JavaScript Object Notation
LLM - Large Language Model
NER - Named Entity Recognition
NLP - Natural Language Processing
OSINT - Open-Source Intelligence
PIN - Personal Identification Number
PMI - Pointwise Mutual Information
RAG - Retrieval-Augmented Generation
SMS - Short Message Service
TGN - Temporal Graph Networks
XAI - Explainable AI
XML - eXtensible Markup Language

Appendix

None.

References

- [1] I. Lypkan, "The concept of virtual space as an environment for committing cyber fraud", *Law and Public Administration*, (3), pp. 399–403, 2022. doi:10.32782/pdu.2022.3.61
- [2] D. Lande, I. Subach, A. Puchkov, "System of analysis of big data from social media", *Information & Security*, Vol. 47, No. 1, pp.

- 44–61, 2020, doi:10.11610/isij.4703
- [3] A. Bilz, L. Shepherd, G. Johnson, “Tainted love: A systematic literature review of online romance scam research”, *Interacting with Computers*, 35(6), pp. 773–788, 2023. doi:10.1093/iwc/iwad048
 - [4] InfoStream. Intelligent Monitoring of Internet News: A Scientific and Methodological Guide / Edited by D.V. Lande. — Kyiv: “Elektronni Visti” Information Center, 86 p., 2026. doi:10.13140/RG.2.2.11034.40643
 - [5] S. Demediuk, O. Korystin, “Trends and Characteristics of Online Fraud in Ukraine”, *Science and Law Enforcement*. 2024. Vol. 3–4. No. 65–66. pp. 142–154, 2025.
 - [6] O. Korystin, S. Demediuk, Y. Likhovitskyy, Y. Kardashevskyy, O. Mitina, “Priorities for the Strategic Development of Ukraine's Cybersecurity Based on the Analysis of Expert Sampling Patterns”, *International Journal of Information Technology and Computer Science*, Vol. 17, No. 2, pp. 24–35, 2025.
 - [7] S. Kemp, D. Buil-Gil, A. Moneva, F. Miró-Llinares, N. Díaz-Castaño, “Empty Streets, Busy Internet: A Time-Series Analysis of Cybercrime and Fraud Trends During COVID-19”, *Journal of Contemporary Criminal Justice*, 37, pp. 480–501, 2021. doi:10.1177/10439862211027986
 - [8] M. Abdullah, M. Nawaz, B. Saleem, M. Zahra, E. Ashfaq, Z. Muhammad, “Evolution Cybercrime—Key Trends, Cybersecurity Threats, and Mitigation Strategies from Historical Data”, *Analytics*, 2025. doi:10.3390/analytics4030025
 - [9] N. Shete, M. Maddel, Z. Shaikh, “A Comparative Analysis of Cybersecurity Scams: Unveiling the Evolution from Past to Present”, *IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–8, 2024. doi:10.1109/i2ct61223.2024.10543498
 - [10] A. Papasavva, S. Johnson, E. Lowther, S. Lundrigan, E. Mariconti, A. Markovska, N. Tuptuk, “Applications of AI-Based Models for Online Fraud Detection and Analysis”, *Crime Science*, 14, 2024. doi:10.1186/s40163-025-00248-8
 - [11] A. Karim, M. Shahroz, K. Mustofa, S. Belhaouari, S. Ramana, K. Joga, “Phishing Detection System Through Hybrid Machine Learning Based on URL”. *IEEE Access*, 11, 36805–36822, 2023. doi:10.1109/access.2023.3252366
 - [12] A. Alsufyani, S. Alzahrani, “Social Engineering Attack Detection Using Machine Learning: Text Phishing Attack”, *Indian Journal of Computer Science and Engineering*, 2021. doi:10.21817/indjce/2021/v12i3/211203298
 - [13] M. Shaukat, R. Amin, M. Muslam, A. Alshehri, J. Xie, “A Hybrid Approach for Alluring Ads Phishing Attack Detection Using Machine Learning”, *Sensors (Basel, Switzerland)*, 23, 2023. doi:10.3390/s23198070
 - [14] K. Zhang, H. Wang, M. Chen, X. Chen, L. Liu, Q. Geng, Y. Zhou, “Leveraging machine learning to proactively identify phishing campaigns before they strike”, *Journal of Big Data*, 12, 2025. doi:10.1186/s40537-025-01174-x
 - [15] A. Fahad, J. Rizwan, “Spear Phishing in Social Engineering: Leveraging ChatGPT and Numberbook”, *International Journal of Wireless and Microwave Technologies*, Vol. 16, No. 2, pp. 49–59, 2026.
 - [16] M. Gaurav, K. Pradepta, J. Shaily, T. Vikas, “Transformer Framework Enhanced by Large Language Models for Image-based Multi-class Malware Detection”, *International Journal of Information Technology and Computer Science*, Vol. 18, No. 3, pp. 170–188, 2026.
 - [17] N. Alok, “Beyond Accuracy: A Hybrid BERT-BiLSTM Framework with Explainable AI (XAI) for Detecting Machine-Generated Disinformation”, *International Journal of Wireless and Microwave Technologies*, Vol. 16, No. 3, pp. 350–358, 2026.
 - [18] D. Holubinka, V. Vysotska, S. Vladov, Y. Ushenko, M. Talakh, Y. Tomka, “Intelligent System for Recognizing Tone and Categorizing Text in Media News at an Electronic Business Based on Sentiment and Sarcasm Analysis”, *International Journal of Information Engineering and Electronic Business*, Vol. 17, No. 1, pp. 90–139, 2025.
 - [19] Suwut Tumthong, Nichanun Samakthai, Pinyaphat Tasatanattakool, “Evaluating LLM-Assisted CLO–PLO Alignment Decisions for AUN-QA–Aligned Curriculum Mapping”, *International Journal of Modern Education and Computer Science*, Vol. 18, No. 3, pp. 121–134, 2026.
 - [20] Azeddine Benelrhali, Khalid Berrada, “Exploring AI Tools and Large Language Models for Students' Performance Enhancement in Riddle Based Logical Reasoning”, *International Journal of Modern Education and Computer Science*, Vol. 17, No. 5, pp. 1–28, 2025.
 - [21] Konstantinos Mastrothanasis, Maria Kladaki, Panagiotis Alexopoulos, “Artificial Intelligence and Rhetorical Art: Argumentative Debate with ChatGPT”, *International Journal of Modern Education and Computer Science*, Vol. 18, No. 1, pp. 48–60, 2026.
 - [22] Shobhit Shrotriya, Nizar Banu P. K., Avi Kulkarni, Vinod G. Kumar, “Application of Large Language Models for Data-Driven Analytics in Oncology: Insights and Evidence Generation from Real-World Imaging Data”, *International Journal of Image, Graphics and Signal Processing*, Vol. 17, No. 6, pp. 38–59, 2025.
 - [23] January F. Naga, Julieta E. Perez, “Learner Engagement State Typologies in AI Tutoring: A Clustering Analysis of Dialogue Behaviors in Introductory Programming Sessions”, *International Journal of Modern Education and Computer Science*, Vol. 18, No. 3, pp. 152–166, 2026.
 - [24] Xuan-Quy Dao, Ngoc-Bich Le, “LLMs Performance on Vietnamese High School Biology Examination”, *International Journal of Modern Education and Computer Science*, Vol. 15, No. 6, pp. 14–30, 2023.
 - [25] Vijay H. Kalmani, Amol C. Adamuthe, Arati Premnath Gondil, Vaishnavi Prashant Patil, Riya Amar Kore, Vaishnavi Mahadev Metkari, “AI vs. Human Writing: Developing a Novel Method for Text Authenticity Detection in Education”, *International Journal of Modern Education and Computer Science*, Vol. 17, No. 3, pp. 45–58, 2025.
 - [26] Ragu Gnanaprakasam, Ramamoorthy Sriramulu, Poorvadevi Ramamoorthy, Mervin Retnadhas, “Innovative Forensics in IoT Clouds by Leveraging Blockchain for Data Integrity and Security”, *International Journal of Computer Network and Information Security*, Vol. 18, No. 1, pp. 116–130, 2026.
 - [27] Sahil Rampal, Manmohan Sharma, Parul Khurana, Mukesh Kumar, “Enhancing Cloud Storage Security Using Multi-User Biometric Encryption with Threshold Access Control”, *International Journal of Wireless and Microwave Technologies*, Vol. 16, No. 3, pp. 142–159, 2026.
 - [28] A. Paziuk, D. Lande, E. Shnurko-Tabakova, P. Kingston, “Decoding manipulative narratives in cognitive warfare: a case study of the Russia-Ukraine conflict”, *Front. Artif. Intell.* 8:1566022, 2025. doi: 10.3389/frai.2025.1566022
 - [29] P. Lewis, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks”, *Advances in neural information processing systems*, 33, pp. 9459–9474, 2020.
 - [30] N. Ibrahim, S. Aboulela, A. Ibrahim, et al., “A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges”, *Discov Artif Intell* 4, 76, 2024. doi:10.1007/s44163-024-00175-8
 - [31] Engels Rajangam, Chitra Annamalai, “Graph Models for Knowledge Representation and Reasoning for Contemporary and

- Emerging Needs – A Survey", *International Journal of Information Technology and Computer Science*, Vol. 8, No. 2, pp. 14–22, 2016.
- [32] Mahmoud Moshref, Rizik Al-Sayyad, "Developing Ontology Approach Using Software Tool to Improve Data Visualization (Case Study: Computer Network)", *International Journal of Modern Education and Computer Science*, Vol. 11, No. 4, pp. 32–39, 2019.
- [33] Fethi Fkih, Ghadeer Al-Turaif, "Threat Modelling and Detection Using Semantic Network for Improving Social Media Safety", *International Journal of Computer Network and Information Security*, Vol. 15, No. 1, pp. 39–53, 2023.
- [34] Ahmad Ashari, Mardhani Riasetiawan, "Document Summarization using TextRank and Semantic Network", *International Journal of Intelligent Systems and Applications*, Vol. 9, No. 11, pp. 26–33, 2017.
- [35] Mariia Nazarkevych, Victoria Vysotska, Vasyl Lytvyn, Dmytro Uhryn, Zhengbing Hu, "An Innovative Method for Detecting Fake News Distribution Sources based on Machine Learning Technology and Graph Theory", *International Journal of Computer Network and Information Security*, Vol. 18, No. 2, pp. 149–180, 2026.
- [36] Nisha Chaurasia, Akhilesh Tiwari, "Efficient Algorithm for Destabilization of Terrorist Networks", *International Journal of Information Technology and Computer Science*, Vol. 5, No. 12, pp. 21–30, 2013.
- [37] Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, Jianxin Li. "Temporal Knowledge Graph Completion: A Survey", *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Survey Track*. Pages 6545–6553, 2023. doi:10.24963/ijcai.2023/734
- [38] E. Momeni, S. Karunasekera, P. Goyal, K. Lerman, "Modeling Evolution of Topics in Large-Scale Temporal Text Corpora", 656–659, 2018. <https://doi.org/10.1609/icwsm.v12i1.15068>
- [39] J.T. Aparicio, E. Arsenio, F.C. Santos, R. Henriques, "Using dynamic knowledge graphs to detect emerging communities of knowledge", *Knowl. Based Syst.*, 294, 111671, 2024. doi:10.1016/j.knosys.2024.111671
- [40] D. Wu, S. Pan, "Dynamic Topic Evolution with Temporal Decay and Attention in Large Language Models", *5th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 1440–1444, 2025. doi:10.1109/eiecs67708.2025.11283454
- [41] A. Rossanez, J. Reis, R. Torres, "Representing Scientific Literature Evolution via Temporal Knowledge Graphs", 33–42, 2020.
- [42] S. Zhang, et al., "Unifying Text Semantics and Graph Structures for Temporal Text-attributed Graphs with Large Language Models", *ArXiv*, abs/2503.14411, 2025. doi:10.48550/arxiv.2503.14411
- [43] H. Huang, L.-Y. Xie, M. Liu, J. Lin, H. Shen, "An embedding model for temporal knowledge graphs with long and irregular intervals", *Knowl. Based Syst.*, 296, 111893, 2024. doi:10.1016/j.knosys.2024.111893
- [44] F. Heeg, I. Scholtes, "Using time-aware graph neural networks to predict temporal centralities in dynamic graphs", *Advances in Neural Information Processing Systems*, 37, pp. 30149–30178, 2024.
- [45] Y. Zheng, L. Yi, Z. Wei, "A survey of dynamic graph neural networks", *Frontiers of Computer Science*, 19(6), p.196323.59 — Temporal Knowledge Graph Completion, 2025.
- [46] N. Jia, C. Yao, "A brief survey on deep learning-based temporal knowledge graph completion", *Applied Sciences*, 14(19), p.8871, 2024.
- [47] Li, L., et al., "From Text to Forecasts: Bridging Modality Gap with Temporal Evolution Semantic Space", 2026.
- [48] S. Berkani, B. Guermah, M. Zakroum, M. Ghogho, "Spatio-Temporal Forecasting: A Survey of Data-Driven Models Using Exogenous Data", *IEEE Access*, 11, 75191–75214, 2023. doi:10.1109/access.2023.3282545
- [49] C. Yang, W. Zhang, Y. Zhou, "An Overview of Spatiotemporal Network Forecasting: Current Research Status and Methodological Evolution", *Mathematics*, 2025. doi:10.3390/math14010018
- [50] X. Chen, et al., UniMamba: A Unified Spatial-Temporal Modeling Framework with State-Space and Attention Integration, 2026.
- [51] L. Yang, C. Chatelain, S. Adam, "Dynamic graph representation learning with neural networks: A survey", *IEEE Access*, 12, pp. 43460–43484, 2024.
- [52] P. Zhao, et al., "Retrieval-augmented generation for ai-generated content: A survey", *Data Science and Engineering*, pp. 1–29, 2026. doi:10.1007/s41019-025-00335-5

Authors' Profiles



Oleksandr Korystin DSc (Law), Professor. In 2009 he received a DSc degree in information law from NAIA. In 2010 he received the academic title of Professor. Honored Science and Technology Figure of Ukraine.

In 2014–2016 – Rector of the Odesa State University of Internal Affairs. Head of the Risk Management & Resilience Committee NGO "Institute of Cyber Warfare Research".

An expert on implementing a risk-based approach to security. Together with the Cyber Police Department, he is implementing EUROPOL's IOCTA methodology, supplementing it with risk assessment and forecasting, as well as the identification of cyber police capabilities and cybersecurity system vulnerabilities. Research interests: cybersecurity, critical infrastructure resilience; criminology; economic security; intelligence; methodology of strategic (risk assessment, forecasting, foresight); counteraction to the hybrid threat.



Dmytro Lande DSc (Computer science), Professor. In 2007 he received a DSc degree in information technology. In 2019 he received the academic title of Professor. Honored Science and Technology Figure of Ukraine. Laureate of the Cabinet of Ministers of Ukraine Prize for the Development and Implementation of Information Technologies (2023). Head of the Department of Information Security at the Educational and Research Institute of Physics and Technology, Igor Sikorsky Kyiv Polytechnic Institute; Head of the Scientific Center for Digital Transformation and Law at the Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine.

Research interests: information and cybersecurity, information technologies, systems analysis, mathematical modeling, artificial intelligence, OSINT, and information law.



Ihor Korzh DSc (Law), Senior Research Fellow. In 2014 he received a DSc degree in Administrative Law and Procedure; Financial Law; Information Law. In 2010 he received Senior Research Fellow degree.

Head of the Center for Legal Issues of Digital Transformation at the Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine. Research interests: Information Law, Administrative Law, Information Security, Cybersecurity, Digital Transformation.



Nataliia Svyrydiuk DSc, PhD, Professor. In 2016 she received a DSc degree in information law from OSUIA. In 2021 she received the academic title of Professor.

Senior Researcher at the Odesa State University of Internal Affairs. Research interests: Administrative law, legal deontology, economic security, criminology, strategic analysis, security risk management, cybersecurity risk assessment, cybercrime, open-source intelligence (OSINT); countering hybrid threats in the field of civil security; legal research and applied research methodologies.



Yuriy Kardashevskyy PhD in Law, Head of Internal Control Department, National Agency for the Prevention of Corruption.

From 2015 to 2023 - worked at the National Anti-Corruption Bureau of Ukraine. Research interests: anti-corruption strategies, cybersecurity, economic security, cybercrime, cyber resilience, intelligence, counteraction to the hybrid threat.



Nataliia Tsiupryk PhD in Law, Associate Professor in the Department of Administrative Law at the National Academy of Internal Affairs.

Research interests: Administrative Law, Information Security, Cybersecurity, Digital Transformation, Cybercrime, Counteraction to the Hybrid Threat.

How to cite this paper:

Oleksandr Korystin, Dmytro Lande, Ihor Korzh, Nataliia Svyrydiuk, Yuriy Kardashevskyy, Nataliia Tsiupryk, "Temporal Semantic Graph Analysis and Evolution Forecasting of Internet Fraud: A Structural Continuity Approach", International Journal of Intelligent Systems and Applications (IJISA), Vol. 18, No. 5, pp. 109–136, 2026. DOI:10.5815/ijisa.2026.05.06