



Views & Comments

Data Inference: Data Security Threats in the AI Era

Zijun Wang^{a,b}, Ting Liu^{a,*}, Yang Liu^a, Enrico Zio^{b,c}, Xiaohong Guan^a^a Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China^b Department of Energy, Politecnico di Milano, Milano 20156, Italy^c CRC, Mines Paris–PSL University, Paris 06904, France

1. Introduction

Data inference (DInf) is a data security threat in which critical information is inferred from low-sensitivity data. Once regarded as an advanced professional threat limited to intelligence analysts, DInf has become a widespread risk in the artificial intelligence (AI) era.

In 1966, Japanese intelligence analysts inferred the location, scale, and production capacity of China's Daqing Oilfield using only five publicly available reports from Chinese journals that subsequently supported the development of specialized oil-extraction equipment for Daqing [1]. In 2014, Cambridge Analytica analyzed psychological test data from approximately 50 million Facebook users to infer their personality traits and political preferences, enabling the delivery of targeted political advertisements and news during the 2016 US presidential election [2]. In 2022, the Cyberspace Administration of China imposed a fine of over 1.2 billion USD on DiDi Global for violating cybersecurity and data-protection laws, citing the illegal collection of user information and inference of travel patterns [3] (Fig. 1).

These incidents demonstrate how DInf has transitioned from an advanced professional threat to a common risk. Historically, DInf originated in intelligence fields, where it required specialized expertise and advanced analytical skills to extract sensitive information from limited data [4]. Today, advancements in AI methods, combined with improvements in software and hardware technologies, have significantly expanded attackers' access to data and computational resources. These developments have lowered the barriers to complex data analysis, increasing the likelihood and accuracy of DInf while reducing the need for specialized expertise. Consequently, DInf attacks, once restricted to professional intelligence agencies, have evolved into a pervasive security challenge.

DInf threats pose significant risks to national security, trade secrets, and personal privacy, with attackers employing covert and diverse methods. The integration of AI technologies with public data has introduced unforeseen threats, which heighten data holders' reluctance and concerns about data sharing [5], thereby impeding the circulation and effective utilization of valuable information.

2. Factors constraining DInf

DInf can be categorized into three types: state inference, parameter inference, and joint inference. State inference utilizes dynamic correlations and real-time data to infer the dynamic states of targets, such as vehicle trajectories [6] and social behaviors [7]. Parameter inference relies on long-term observations and statistical analysis to infer the intrinsic static parameters of targets, such as identity attributes [8] and line admittance [9]. Joint inference integrates both approaches, enabling the simultaneous inference of dynamic states and static parameters. Most studies on DInf focus on extensive theoretical investigation and targeted modeling to demonstrate the feasibility of inference processes. However, this specialized, problem-centric approach obscures the broader prevalence and complexity of inference threats in real-world scenarios, leading to significant underestimation of their risks. Furthermore, due to the difficulty of DInf, it is ignored in data security classification and grading, which primarily consider the value of the data itself. In practice, the widespread occurrence of inference threats is constrained by the following three critical factors (Fig. 2(a)):

(1) **Heterogeneity of multi-domain data.** Data from different domains differ significantly in format, precision, and scope, complicating the creation of unified representations and obscuring potential interrelationships among datasets. Most DInf techniques, limited by computational capacities, are designed to handle a single data type or domain. Consequently, they encounter substantial challenges in processing multi-domain heterogeneous data, resulting in an underestimation of the risks associated with cross-domain inference.

(2) **Ambiguity of data relationships.** Traditional inference methods heavily depend on expert knowledge to identify relationships between data through explicit associations. However, uncovering potential associations and inference pathways using common sense remains challenging, blurring the line between non-sensitive and sensitive data. Limited in their ability to process complex non-linear relationships and implicit patterns, these methods often overlook numerous potential pathways. As a result, public data is perceived as weakly related to sensitive data, leading to a widespread underestimation of DInf risks.

(3) **Complexity of combinatorial space.** The inference process involves multiple data categories, where one category can be inferred from one or more others. As the number of data categories

* Corresponding author.

E-mail address: tingliu@xjtu.edu.cn (T. Liu).

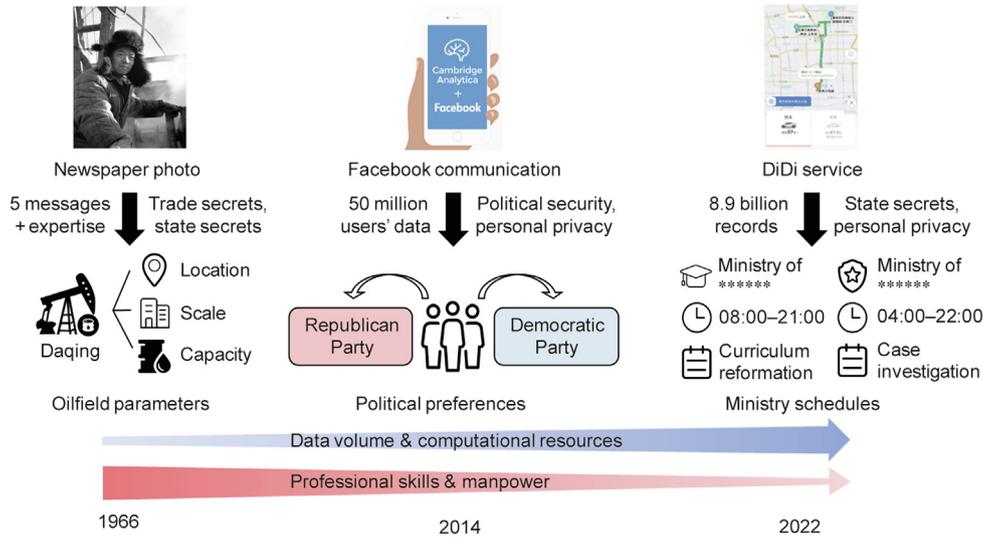


Fig. 1. Data inference incidents and trends.

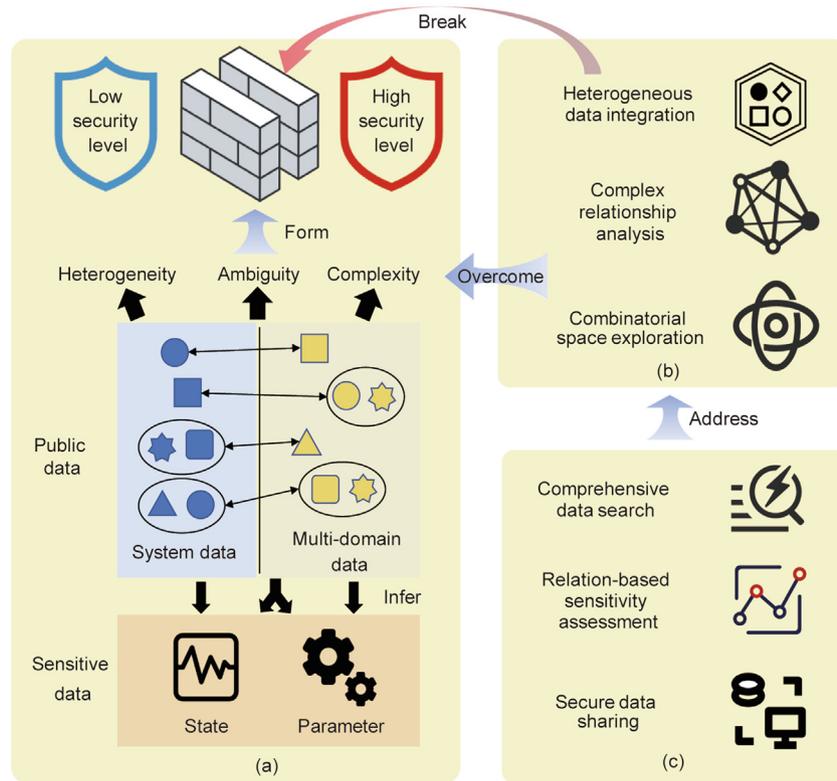


Fig. 2. DInf threats and protection frameworks. (a) The process and challenges of DInf, forming the foundation of traditional data-protection frameworks; (b) AI-driven capabilities in DInf; (c) AI-empowered data-protection frameworks.

increases, the total number of possible inference combinations grows exponentially. This vast combinatorial space significantly complicates attackers' efforts to comprehensively explore all possible combinations within limited computational resources, making it more challenging to identify feasible inferences.

The challenges of heterogeneity, ambiguity, and complexity significantly increase the technical difficulties encountered by traditional DInf methods, posing substantial barriers to effective inference. These challenges also underpin the theoretical foundations of current data-protection frameworks (Fig. 2), which assume that traditional DInf techniques cannot overcome these obstacles.

3. AI-enabled DInf can overcome previous constraints

The advanced capabilities of AI technologies in integrating heterogeneous data, analyzing complex relationships, and exploring combinatorial spaces can overcome traditional data barriers. Thus, DInf is challenging the foundational principles of existing data-protection frameworks. Data security grading, which assigns security levels to data based on sensitivity and importance, serves as a critical safeguard [10,11]. However, the rapid advancement of AI technologies is progressively eroding the security boundaries of traditional frameworks (Fig. 2(b)).

Advancements in multimodal modeling, particularly transformer-based technologies such as ChatGPT [12], have significantly increased the integration of heterogeneous data. These models process text, images, and structured data simultaneously, facilitating cross-domain analysis. In the case of the Daqing Oilfield, intelligence personnel integrated heterogeneous data sources manually. In multimodal systems, various data types are preprocessed and projected into a unified embedding space. Textual data, such as oilfield operation reports, are encoded into contextual embeddings using pretrained language models. Visual data, such as satellite images or photographs, are processed through vision transformers or convolutional neural networks to extract spatial features. Structured data, such as geographic coordinates, are normalized and embedded into the same vector space. Attention mechanisms within these models align embeddings across modalities, enabling relationships between disparate data sources to be discovered (Fig. 3). For instance, these models can correlate textual descriptions of “oilfield depth” from public reports with visual features of drilling infrastructure captured in satellite imagery. By incorporating geospatial data, they can also infer critical parameters, such as the precise location and production capacity of oilfields. Furthermore, pretrained models streamline the unification of diverse data formats and structures, effectively lowering barriers to processing heterogeneous datasets.

To address the challenge of ambiguous data relationships, physics-informed neural networks (PINNs) integrate domain knowledge into deep learning models, enhancing the analysis of complex nonlinear relationships and uncovering hidden inference pathways for sensitive data. In power systems, PINNs embed operational mechanisms, such as power balance constraints and power flow equations, directly into the learning process, enabling the inference of unobserved system parameters. For example, the incorporation of physics guidance into a deep learning model reduced inference errors in key system parameters, such as voltage phase angles and active powers, by 1–3 orders of magnitude, significantly improving power system inference accuracy [13]. Adversaries could exploit the inference pathways uncovered by PINNs to infer sensitive operational information, posing significant threats to power system security. Graph neural networks (GNNs) are highly effective in modeling graph-structured relationships, enabling AI to identify deep latent associations. In social networks, GNNs can identify hidden interaction paths between individuals without direct connections. For example,

the model proposed by Ref. [14] achieves up to 73.9% higher precision in temporal link prediction tasks compared with traditional methods, demonstrating its strength in inferring latent social relationships. In transportation systems, spatiotemporal GNNs outperform traditional methods such as auto-regressive integrated moving average model (ARIMA), achieving up to 63.5% higher accuracy [15]. These models reveal critical propagation paths of congestion, exposing correlations between vehicle trajectories and geographic regions, which malicious actors could exploit to identify and target vulnerable components of transportation infrastructure.

AI addresses the challenges of combinatorial explosions by employing heuristic search and metaheuristic algorithms. For instance, genetic algorithms mimic natural selection to optimize the search for high-risk inference pathways iteratively. The process begins with the generation of an initial population of random data subsets. Each subset is evaluated using a fitness function designed to quantify inference risk, defined as follows:

$$f(S) = \alpha \sum_{d_i \in S} \text{Rel}(d_i) + \beta \sum_{d_i \in S} \text{Leak}(d_i) + \gamma \sum_{d_i \in S} \text{Freq}(d_i) + \delta \sum_{(d_i, d_j) \in S} \text{CmbRisk}(d_i, d_j) \quad (1)$$

where S denotes a data subset, d_i and d_j represent individual data items within the subset, $\text{Rel}(\cdot)$ measures the relevance of a data item to sensitive parameters, $\text{Leak}(\cdot)$ quantifies its historical leakage rate, $\text{Freq}(\cdot)$ indicates its occurrence frequency in the dataset, and $\text{CmbRisk}(\cdot)$ captures the combined inference risk between pairs of data items. The parameters α , β , γ , and δ are weighting factors that balance the importance of each metric. High-risk subsets are refined further through crossover and mutation operations to generate new candidate subsets. This evolutionary process continues until the algorithm converges on the most critical inference pathways.

Heuristic methods can become computationally prohibitive under real-time constraints. To address this, clustering analysis reduces the search space by partitioning data items into highly correlated clusters. These clusters are formed based on co-occurrence in historical inference pathways, attribute similarity, or statistical features, effectively eliminating many irrelevant combinations. Additionally, parallel computing significantly increases scalability. Clustered data subsets can be distributed across multiple processing units based on geographic region, data type, or temporal

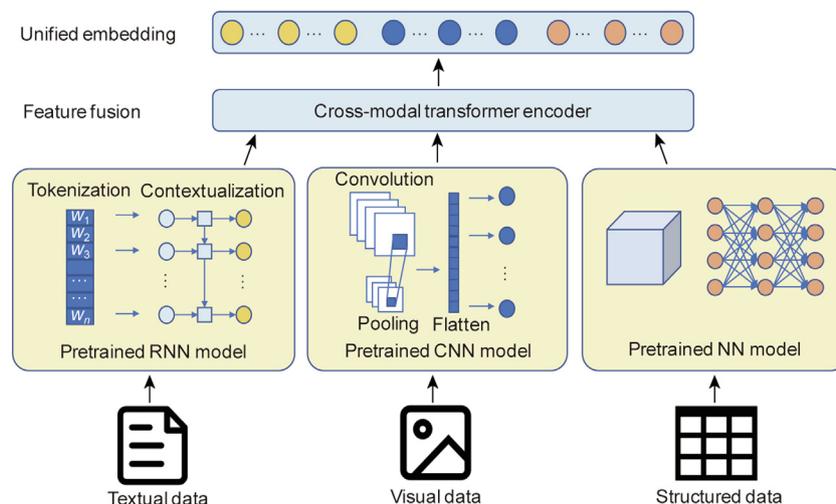


Fig. 3. Schematic workflow of a multimodal model. RNN: recurrent neural network; CNN: convolutional neural network; NN: neural network.

window, enabling parallel execution and accelerating the discovery of critical inference pathways.

4. Mitigating DInf risks

To mitigate the risks of DInf, it is necessary to address the limitations of traditional protection frameworks in data discovery, sensitivity assessment, and data sharing, thereby establishing an AI-empowered data-protection framework. Such a framework encompasses the following aspects (Fig. 2(c)).

4.1. Comprehensive data searching

Identifying potential DInf risks requires both domain-specific and cross-domain exploration. Data providers and operators must thoroughly analyze their data and its relationships with external datasets. This process is inherently complex, involving diverse data modalities and extensive analytical scopes. To address these challenges, a modular framework for AI-assisted data searching is proposed, comprising the following components.

(1) **Heterogeneous data source integration.** A unified data-acquisition module supports diverse data interfaces, enabling seamless access to relational databases, not only structured query language (NoSQL) stores, and unstructured file systems. By leveraging data type recognition algorithms, the system automatically parses and normalizes various formats into a unified semantic representation. Automated tools extract metadata from connected sources and construct multidimensional indices to enable efficient querying, retrieval, and management.

(2) **Data identification and annotation.** Natural language processing techniques perform semantic analysis and classify data content. Pretrained named-entity recognition models identify sensitive entities, such as personally identifiable information and financial data. These models are integrated with rule-based engines to assign sensitivity levels.

(3) **Dynamic update mechanism.** An event-driven, change-detection-based monitoring module automatically tracks insertions, modifications, or deletions within data sources. This ensures real-time updates to data assets while maintaining system performance and scalability.

Although AI has significantly improved the efficiency and scope of data discovery, challenges persist when processing unstructured or low-quality data. Issues such as noise interference, incomplete fields, and contextual ambiguity can lead to misclassification or inaccurate annotation. These challenges are particularly evident when models encounter semantically ambiguous expressions or domain-specific terminology. Furthermore, biases in training data exacerbate error rates in automated classification and annotation. To mitigate these risks, safeguards such as confidence threshold filtering and human-in-the-loop reviews can be employed. These measures increase the accuracy and reliability of data discovery outcomes while reducing the risks associated with AI-driven automation.

4.2. Relation-based sensitivity assessment

Traditional data-grading methods evaluate the sensitivity of individual data points in isolation, overlooking the relational complexities among them. To address this limitation, structured modeling with knowledge graphs offers a robust approach to uncover latent relationships ignored by conventional methods. By constructing knowledge graphs using triples extracted from open-source intelligence, the performance of entity-relation recognition can be significantly improved, with the F_1 -score rising from 0.62 to 0.81 [16]. Relationships embedded within knowledge graphs are

critical for sensitivity assessment. Experimental results show that incorporating joint entity-relation embeddings increases classification accuracy for sensitive data, achieving a 4.2% increase in the F_1 -score [17]. This relation-based sensitivity assessment framework enables more precise sensitivity evaluations and significantly strengthens the ability to defend against DInf.

Using urban mobility as a representative scenario, knowledge graphs can model causal and semantic relationships among key attributes, such as ride timestamps, geographic location labels, user identifiers, and payment metadata. In this graph structure, nodes represent data entities, while edges capture their semantic or causal dependencies. For instance, an edge from “timestamp” to “geographic location” represents travel patterns during specific time intervals, whereas an edge from “geographic location” to “payment amount” reflects the influence of spatial context on ride costs. To increase the graph’s expressiveness for complex behavioral patterns, domain knowledge can be incorporated through structural labels and edge weights. For example, peak periods are assigned as categorical attributes of timestamp nodes, location nodes are weighted based on access frequency, and edge strengths are adjusted according to statistical co-occurrence. On this foundation, graph embedding methods can be applied to learn vectorized representations of nodes. In the embedding space, structurally similar or semantically related nodes exhibit higher vector similarity. Leveraging this similarity, sensitive information can propagate to adjacent nodes in the embedding space, enabling risk perception even for data points not labeled as sensitive.

To ensure scalability and interpretability, an evolutionary update mechanism combining automated learning with human supervision is employed. Node embeddings are periodically updated to autonomously capture new relationships arising from changes in user behavior or system policies. Simultaneously, expert reviews validate high-sensitivity data and newly discovered relationships, ensuring the accuracy and reliability of sensitivity recognition. Additionally, an incremental update strategy is implemented to revise newly added entities and edges locally, avoiding the overhead of structural reconstruction. This approach improves maintenance efficiency and supports the long-term evolution of the graph structure.

4.3. Secure data sharing

The value of data comes from sharing and circulation, but it is challenging to ensure that shared data does not expose sensitive information. Addressing this issue requires robust methods to balance data utility and security.

(1) **Generative adversarial networks (GANs)** simulate the statistical properties of real data to generate synthetic datasets, effectively mitigating DInf risks [18]. However, GANs may unintentionally preserve latent correlations among sensitive attributes, leaving them vulnerable to membership or attribute inference attacks. To address this issue, differential privacy-enhanced training methods have been developed, primarily leveraging noisy stochastic gradient descent (NoisySGD) and private aggregation of teacher ensembles (PATE) frameworks [19].

For example, in the Cambridge Analytica incident, psychological testing data were exploited to infer individuals’ political preferences. NoisySGD-based approaches mitigate such risks by injecting noise into the discriminator’s gradient updates, preventing it from memorizing associations between personal traits and political preferences. In contrast, PATE-based approaches utilize multiple teacher models trained on distinct data subsets. The differentially private aggregated outputs of these teachers are then used to train the generator. Since the generator never directly accesses sensitive information and is trained under differential privacy guarantees, the exposure of sensitive relationships is effectively prevented.

Compared with other synthetic data-generation methods, rule-based data augmentation relies on manual transformation strategies and fails to capture deep statistical properties. Variational autoencoders excel at synthesizing continuous data but struggle with categorical or unstructured data. By integrating differential privacy mechanisms, GANs generate high-quality synthetic datasets with strong privacy guarantees, making them particularly suitable for mitigating DInf risks.

(2) **Federated learning (FL)** is particularly effective for distributed datasets, such as user mobility data collected by ride-hailing platforms. FL enables collaborative model training without sharing raw location or trip data [20]. A global model is sent to user devices and trained locally on private data, and only gradient updates are transmitted back to a central server. However, real-world datasets often exhibit highly non-independent and identically distributed (Non-IID) characteristics. For instance, urban users generate frequent and dense mobility data, while rural users produce sparse and irregular data. This data imbalance reduces global model performance and increases the risk of sensitive information leakage through shared gradient updates.

To address these challenges, personalized FL techniques customize the global model for diverse user groups by adding local adaptation layers or fine-tuning the model with user-specific data. These techniques improve performance across heterogeneous populations and mitigate the negative effects of Non-IID data distributions. Additionally, gradient masking can be applied to shared gradients to reduce the risk of sensitive information leakage.

Unlike other privacy-preserving methods, homomorphic encryption enables computation directly on encrypted data, but it incurs extremely high computational overhead, limiting practical deployment. Secure multi-party computation protects the confidentiality of participant inputs but requires extensive communication, posing scalability challenges. In contrast, FL achieves a balance between usability and security, making it suitable for large-scale collaborative training in resource-constrained environments. However, FL remains vulnerable to risks such as model update leakage and inference attacks. To increase security, additional techniques such as differential privacy and cryptographic methods must be integrated.

5. Conclusions

DInf illustrates the transformative impact of AI on cybersecurity. Traditional security methods, such as data grading based on inference barriers or password systems relying on random character combinations, are rooted in the intractability of complex problems. However, these approaches are increasingly disrupted by advancements in AI technologies [21]. Defensive strategies against DInf leverage AI to empower defenders, enabling proactive detection and mitigation of potential risks. Additionally, these strategies offer valuable insights into using AI to counter AI-driven threats, providing a broader framework for addressing emerging cybersecurity challenges.

CRedit authorship contribution statement

Zijun Wang: Writing – original draft, Visualization, Investigation. **Ting Liu:** Conceptualization, Writing – review & editing, Funding acquisition. **Yang Liu:** Writing – review & editing, Funding acquisition. **Enrico Zio:** Writing – review & editing. **Xiaohong Guan:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2022YFB2703503), the National Natural Science Foundation of China (62293501, 62525210, and 62293502), and the China Scholarship Council (202306280318).

References

- [1] Liu T, Li Q, Shi J. Data leakage risks of public data in the big data era. *Baomi Gongzuo* 2018;4:51–2. Chinese.
- [2] Confessore N. Cambridge Analytica and Facebook: the scandal and the fallout so far [Internet]. New York City: The New York Times; 2018 Apr 4 [cited 2025 Jan 2]. Available from: <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html#>.
- [3] Xiong Y, Register L, He L. China fines Didi \$1.2 billion for violating cybersecurity and data laws [Internet]. Atlanta City: Cable News Network; 2022 Jul 21 [cited 2025 Jan 2]. Available from: <https://edition.cnn.com/2022/07/21/economy/china-fines-didi-data-law-violation-intl-hnk/index.html>.
- [4] Liu T, Wang Z, Liu Y, Zhou Y, Wu J, Bao Y, et al. Data inference: data leakage paradigms and defense methods in cyber-physical systems. *Sci Sin Inf* 2023;53(11):2152–79. Chinese.
- [5] Nonnecke B, Carlton C. EU and US legislation seek to open up digital platform data. *Science* 2022;375(6581):610–2.
- [6] Liu J, Luo Y, Zhong Z, Li K, Huang H, Xiong H. A probabilistic architecture of long-term vehicle trajectory prediction for autonomous driving. *Engineering* 2022;19:228–39.
- [7] Garcia D. Leaking privacy and shadow profiles in online social networks. *Sci Adv* 2017;3(8):e1701172.
- [8] Acquisti A, Gross R. Predicting social security numbers from public data. *Proc Natl Acad Sci USA* 2009;106(27):10975–80.
- [9] Wang Z, Liu Y, Yu N, Wu Q, Wu J, Zhou Y, et al. Data inference from publicly available data: threats and defense methods in power systems. *IEEE Trans Power Syst* 2025;40(1):1049–59.
- [10] European Parliament and Council of the European Union. General data protection regulation. 2016.
- [11] Force JT. Security and privacy controls for information systems and organizations. Gaithersburg: National Institute of Standards and Technology; 2020.
- [12] Mackenzie D. Surprising advances in generative artificial intelligence prompt amazement—and worries. *Engineering* 2023;25:9–11.
- [13] Bento MEC. Physics-guided neural network for load margin assessment of power systems. *IEEE Trans Power Syst* 2024;39(1):564–75.
- [14] Qiu Z, Wu J, Hu W, Du B, Yuan G, Yu PS. Temporal link prediction with motifs for social networks. *IEEE Trans Knowl Data Eng* 2023;35(3):3145–58.
- [15] Zhao L, Song Y, Zhang C, Liu Y, Wang P, Lin T, et al. T-GCN: a temporal graph convolutional network for traffic prediction. *IEEE Trans Intell Transp Syst* 2020;21(9):3848–58.
- [16] Guo Y, Liu Z, Huang C, Wang N, Min H, Guo W, et al. A framework for threat intelligence extraction and fusion. *Comput Secur* 2023;132:103371.
- [17] Narvala H, McDonald G, Ounis I. RelDiff: enriching knowledge graph relation representations for sensitivity classification. In: Moens MF, Huang X, Specia L, Yih SW, editors. Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: Association for Computational Linguistics; 2021. p. 3671–81.
- [18] Gadotti A, Rocher L, Houssiau F, Creţu AM, de Montjoye YA. Anonymization: the imperfect science of using data while preserving privacy. *Sci Adv* 2024;10(29):eadn7053.
- [19] Ganev G, Xu K, De Cristofaro E. Graphical vs. deep generative models: measuring the impact of differentially private mechanisms and budgets on utility. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security; 2024 Oct 14–18; Salt Lake City, UT, USA. New York City: Association for Computing Machinery; 2024. p. 1596–610.
- [20] Liu Q, Yan Y, Jin Y, Wang X, Ligeti P, Yu G, et al. Secure federated evolutionary optimization—a survey. *Engineering* 2024;34:23–42.
- [21] Wang D, Zou Y, Zhang Z, Xiu K. Password guessing using random forest. In: Calandrino J, Troncoso C, editors. Proceedings of the 32nd USENIX Security Symposium; 2023 Aug 9–11; Anaheim, CA, USA. Berkeley: USENIX Association; 2023. p. 965–82.