Fundamental Research 1 (2021) 595–606

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Review

Combating emerging financial risks in the big data era: A perspective review

Xueqi Cheng[a,*], Shenghua Liu[a,*], Xiaoqian Sun[a], Zidong Wang[a], Houquan Zhou[a], Yu Shao[b], Huawei Shen[a]

[a] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190 China
[b] China Institute of Finance and Capital Markets, Beijing 100033, China

ARTICLE INFO

Article history:
Received 22 June 2021
Received in revised form 17 August 2021
Accepted 22 August 2021
Available online 30 September 2021

Keywords:
Financial risk
Big data
Risk management
Deep learning
Graphs and networks

ABSTRACT

Big data technology has had a significant impact on new business and financial services: for example, GPS and Bluetooth inspire location-based services, and search and web technologies motivate online shopping, reviews, and payments. These business services have become more connected than ever, and as a result, financial frauds have become a significant challenge. Therefore, combating financial risks in the big data era requires breaking the borders of traditional data, algorithms, and systems. An increasing number of studies have addressed these challenges and proposed new methods for risk detection, assessment, and forecasting. As a key contribution, we categorize these works in a rational framework: first, we identify the data that can be used to identify risks. We then discuss how big data can be combined with the emerging tools to effectively learn or analyze financial risk. Finally, we highlight the effectiveness of these methods in real-world applications. Furthermore, we stress on the importance of utilizing multi-channel information, graphs, and networks of long-range dependence for the effective identification of financial risks. We conclude our survey with a discussion on the new challenges faced by the financial sector, namely, deep fake technology, adversaries, causal and interpretable inference, privacy protection, and microsimulations.

1. Introduction

With the development of big data, people's everyday life has been comprehensively recorded, including their financial activities [1]. For example, smartphones, embedded with GPS or Bluetooth technology, keep track of people's visits to banks/ATMs, shopping malls, or office buildings, which consequently generates logs of activities and even reveal people's possible physical contacts [2]; time spent on social media, search queries, and clicks are collected for users' credit and preference profiling; and satellite images of the earth are periodically taken by many sites for analyzing changes in buildings and even the number of cars parked in a parking lot at a certain time [3]. Sufficiently using and screening these data can help us better understand and combat hidden financial risks [4,5].

Meanwhile, a range of algorithms for emerging technologies are widely used in finance-related systems. Recommendations generated by artificial intelligence (AI) algorithms decide which ads, shopping items, and news should be prioritized. Natural language processing (NLP) models and technologies have been adopted to understand users' sentiments and opinions. Deep neural networks and intelligent algorithms are embedded into devices for real-time monitoring, smart Q&A, and assistant robots, among others.

Nowadays, finance has become a highly complex field; it is no more limited to traditional banking and trading systems, stocks, and futures markets, but it also includes emerging digital currencies, online shopping, payment systems, and even overlaps with social media and networks, political campaigns, and a range of smart systems, including the Internet of things (IoT). For example, a breaking news can shake the foundations of a financial system, or a smart system on route planning can decide how commodities can be efficiently delivered. Fintech, a combination of Internet, big data, and finance technologies, grows new businesses in various fields (e.g., payments, deposits, loans, investments, and market facilities), having a significant impact on traditional financial business models and laying hidden dangers. Moreover, circulation of multiple digital currencies, such as Bitcoin, as a new form of foreign exchange, challenges a country's foreign exchange management.

Therefore, financial risks in the big data era have broken the borders of traditional data, algorithms, and systems, generating more challenges than ever.

* Corresponding authors.
E-mail addresses: cxq@ict.ac.cn (X. Cheng), liushenghua@ict.ac.cn (S. Liu).

https://doi.org/10.1016/j.fmre.2021.08.017

1.1. Financial risks in the big data era

According to the "2020 China Internet Finance Development Report" issued by iResearch[1], investment in financial technologies was 112 billion yuan, an increase of 19.4%, whereas investment in AI in banking was 14.3 billion yuan, an increase of 28.8%. Furthermore, the three most preferred investment areas are smart risk control, smart insurance, and smart customer service, accounting for more than 70% of the entire investments.

A side effect of the admission policy for high-tech companies into the market is that a large number of non-traditional financial institutions without risk control mechanisms in place are engaged in high-risk businesses. Consequently, financial fraud affecting a wide range of investors and communities is a possibility. The well-known financial fraudster, eZubao, had a cumulative transaction volume of 70 billion yuan from 2014 to 2015.

As discussed previously, financial risks have different forms and depend on a range of factors in the big data era, such as abusing AI algorithms and technologies, digital currencies, and online business. To review these challenges for emerging financial risks, we follow a well-known taxonomy wherein financial risks are classified as market, credit, liquidity, volatility, operational, and financial crime-compliance (FCC) risks. However, our reviews mainly focus on credit, liquidity, volatility, and FCC risks Because their management and detection always require the analysis of large-scale and heterogeneously related data. Moreover, any type of risk can cause systemic risk[2] if managed improperly, which can result in the collapse of an entire financial system or an entire market, or it can even trigger an economic crisis. Thus, we also discuss systemic risk in this paper.

1.2. Challenges of financial risk management

The main challenges in combating financial risks in the big data era can be summarized as follows:

- **Needs to utilize multimodal data.** Big data provide a robust method to handle financial risks. Data are generally obtained from multiple sources or channels, such as financial reports, curves of sales data, non-traditional information from texts, pictures, or videos on social media, satellite images, and human mobility data. These data are semantically correlated and sometimes provide complementary information to each other, thus reflecting an early signal for a more reliable assessment of risks that are not visible when working with individual channels. However, consolidating heterogeneous, disconnected data from multiple channels is challenging given traditional models and statistical methods in financial risk management, such as linear regression, Naive Bayes methods, and classic Hidden Markov models (HMM).
- **Long-range and heterogeneous dependent nature.** Data objects are often related to each other; for example, companies have interdependent business, and money is transferred from one account to another. Furthermore, the relations for some data are heterogeneous, such as networks of shareholders and companies. Some important financial risks can only be identified through long-range dependency, whereas short-range dependent relations cannot reflect these risks, such as the subprime crisis in 2008. However, mining long-range and heterogeneous dependent patterns to detect risks is quite challenging, given the existing tens of billions of data relations.
- **Dynamic and real-time characteristics.** The global financial marketplace is highly dynamic and data, generated by financial organizations change in real time in terms of content and properties, thereby posing difficulty in tracing and analysis.
- **Adversary.** Financial fraud and money laundering of FCC risks rely on new advanced techniques and methods of camouflage, making their detection difficult.

1.3. Related research projects

In the past several years, related research projects have been funded to study the previously mentioned challenging problems.

The BIGDATA program of the National Science Foundation (NSF) has funded many research projects on data science. "Understanding the Financial Market Ecosystem" introduces new behavioral models of financial trading using big data techniques, as well as new metrics and data for the discipline of finance in economics. "Detecting Financial Market Manipulation: An Integrated Data- and Model-Driven Approach" applies innovative data-driven approaches to improve detection and deterrence of market manipulation. The main focus of the project is to use simulation and optimization to generate manipulation strategies from market data streams and identify these manipulation behaviors by extracting signatures and spoofing activities. "Network Analytics on Complex Economic Data Streams for Monitoring Financial Stability" focuses on identifying and predicting market participants that could endanger the overall financial systems, leveraging a wide array of the diverse quantitative financial data stream, metadata, and market announcements.

The National Natural Science of China (NSFC) has launched a major research plan, called "Research on Big Data Driven Management and Decision Making," which supports a series of basic research projects related to internet finance. "Big Data-Driven Financial Monitoring and Service Platform and Demonstration Application" carries out studies such as monitoring of Internet finance and construction of knowledge maps for Internet finance. "Value Analysis, Discovery and Collaborative Creation Mechanism of Financial Big Data based on Knowledge Association" focuses on finding the connection within the knowledge and building a large-scale knowledge graph using an innovative approach and then applying these methods and the knowledge graph on real-world problems. The implementation of these projects helps us better understand Internet finance and provides theoretical foundations and technologies for big-data-driven Internet financial monitoring and services.

1.4. Content organization

In the remainder of the paper, we provide a thorough introduction of the financial big data with respect to data collection, organization, and common types in Section 2. We describe the emerging technologies in Section 3. Detailed applications of big data are summarized and listed for the major categories of financial risks in Section 4, and financial crime compliance risk is specially discussed in Section 5. We then discuss some future directions in Section 6, and finally conclude the paper.

2. Data

2.1. Data collection

Currently, collecting financial big data requires delicate management of acquisition objects, channels, frequency, as well as target-driven customization, which is generally implemented using web crawlers in data centers or on rented cloud resources. By setting some targets or seed sites as initial nodes, web crawlers retrieve data iteratively along with webpage hyperlinks following the breadth-first or depth-first strategy. In addition, system logs, such as event logs of a loan application process in a bank, record various financial institutes' activities in large amounts that could contribute to careful business decisions in many fields, including risk management [6]. While retrieved structured data are ready for most big data algorithms, unstructured data require further processing. Many deep learning models have recently contributed to specific tasks for extracting information from unstructured data in finance. In addi-

[1] http://report.iresearch.cn/report/202009/3648.shtml
[2] https://en.wikipedia.org/wiki/Systemic_risk

tion, there are general frameworks that extract structured knowledge from massive unstructured data, such as TextCube [7].

2.2. Data organization

Knowledge graphs, as a rich and intuitive technique to express knowledge, have attracted widespread attention from the perspective of big data organization. It is essentially a semantic network wherein nodes represent entities or concepts and edges represent various semantic relationships between entities/concepts. For example, "Elon Musk is a visionary entrepreneur," where "Elon Musk" is an entity and "entrepreneur" is a concept. There is an "is-A" relationship between "Elon Musk" and "entrepreneur." According to its data sources, knowledge graphs can be divided into two categories: general web pages, such as Google knowledge graph and Microsoft's Bing, and relatively structured online encyclopedias, such as YAGO[3] and DBPedia[4]. Event is another important type of knowledge, in addition to entities/concepts and their relations. A knowledge graph comprising events and their participants is often called an event graph. Existing event graphs include ICEWS[5], GDELT[6] and the financial event graph of the Harbin Institute of Technology.

Knowledge graphs have become an important tool for financial risk identification. Using shareholding relationships among financial institutes, Lv et al. [8] constructed a knowledge graph for financial equities in China that contains more than 45 million entities/concepts and 145 million relationships. From the equity graph, the true beneficiaries and people acting in concert can be revealed to guide the stable development of the financial industry.

2.3. Data modality

After big data are carefully structured and represented, mining their value becomes a major task. In recent years, the financial industry has gone beyond traditional data, such as SEC filings and press releases, and has paid more attention to company sales records, social networks, credit card transaction information, positioning information, and satellite images, among others, because they contain valuable information on many issues, such as risk perception in financial markets. However, incorporating data from different modalities has brought challenges in exploring the data value. Owing to big data algorithms, these insights are now accessible to researchers and investors.

Text data, especially those from social media, are widely used in financial risk forecasting, detection, and management. For example, the form 10-k plays an important role for listed companies in textual disclosures, contributing to risk prediction [9]. Meanwhile, investigating news in the Wall Street Journal has proven the susceptibility of the stock market to rumors. Topic trends and sentiments on social media have also been studied [10], providing important seeds for further related tasks.

Image and video data, meanwhile, can play a dominant role in detecting financial fraud. For example, outlet videos of more than 10,000 h irrefutably disclosed the fraud scandal of Luckin Coffee in 2020 and satellite images were used for identifying the fraud of Zhangzidao Group Co., Ltd. (one of the listed companies in China) by tracking the trajectories of fishing boats [11]. Audio data are also important for investor meetings and earnings conference calls. They contains not only text information but also sentiment clues, from which features can be extracted to predict stock volatility [12]. Moreover, structured data generated from supply chains, such as freight records, have shown value in alternative data utilization [13], and combining different banking application flows with IoT intelligence by analyzing users' data can remind customers before their credit cards are stolen [14]. However, different types of alternative data are complementary to each other for solving specific tasks. With the development of deep learning, the blending of multimodal information becomes possible through a unified model to predict financial risks.

2.4. Public data for financial research

Table 1 summarizes the available public data used in financial risk research.

3. Emerging technologies and methodologies

In this section, we review the emerging technologies and methodologies that learn from big data, or mine anomalies, patterns, and trends in big data, which are effective tools for handling financial risks in the big data era.

3.1. Deep learning and representation learning

The key to effectively utilizing multimodal financial data is to extract informative representations from it, which can be achieved by leveraging the power of deep learning. The key to the success of deep learning in this context is its ability to automatically learn high-quality representations from large-scale data.

Representation learning is the process of transforming raw data into a more appropriate form for machine learning. It can extract key patterns from the data and play an important role in machine learning applications in which performance may highly depend on the representation.

For example, as a milestone in computer vision, a convolutional neural network (CNN) learns convolution kernels automatically from data and outperforms conventional manual feature extraction methods. For graph data, graph neural networks [15,16], which incorporate the graph structure into the learning process, can learn high-quality representations for downstream tasks. With the help of pre-training frameworks, multimedia data can be jointly exploited to learn representations in a unified model. Such a unified model trained on multimodal data can provide richer information compared with standalone models.[7] [8] [9] [10] [11] [12] [13]

3.2. User profiling and behavior modeling

In a large system with multiple interacting agents, despite significant randomness and unpredictability of each individual's behavior, group behavior has a strong regularity. This provides a framework that can be used to examine group behaviors in such systems. Furthermore, user profiling, which aims to understand and analyze user behavior, can be highly useful for discovering valuable information.

User behavior modeling Traditional methods rely on classic Markov models to simulate user behaviors. Recent works have focused on the idea of learning compressed representations (following the *encoder-decoder* method) that can be used to reconstruct raw sequences. In recurrent neural networks (RNNs), recurrent structures such as long short-term memory (LSTM) are used to model sequences. In addition to the recurrent architecture, convolution and attention architectures have been introduced [17].

[3] https://yago-knowledge.org/
[4] https://www.dbpedia.org/
[5] https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html
[6] https://www.gdeltproject.org/
[7] https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions/
[8] https://www.kaggle.com/ellipticco/elliptic-data-set
[9] https://www.kaggle.com/mczielinski/bitcoin-historical-data
[10] https://www.kaggle.com/ealaxi/paysim1
[11] https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
[12] http://www.ulb.ac.be/di/map/adalpozz/data/creditcard.Rdata
[13] https://www.phishtank.com/index.php and https://www.alexa.com/topsites

Table 1
Public Datasets.

Dataset	Network Type	Data Size	Description
Czech Financial Data[7]	Bank transactions	1.05M	An anonymous transferring data of Czech bank released for Discovery Challenge.
Elliptic Data[8]	Bitcoin	234 K	A transaction graph collected from the Bitcoin blockchain.
Bitcoin Historical Data[9]	Bitcoin	4.86M	Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to March 2021.
Synthetic financial data for fraud detection[10]	Mobile money transactions	6.36M	Synthetic dataset generated by the PaySim mobile money simulator.
Huge Stock Market Data[11]	Stocks and ETFs	17.5M	Historical daily prices and volumes of all U.S. stocks and ETFs.
Credit-card Data[12]	Credit card transactions	284 K	Anonymized credit card transactions labeled as fraudulent or genuine.
Phishing website Data[13]	Uniform resource locator (URL)	3.52 K	2119 phishing sites from PhishTank and 1407 legitimate sites from Alexa Database.

Table 2
Summarization of research works on various applications*.

Paper	Theme	Subtheme	Year	Traditional ML	DL	Methods
Ha [39]	Credit risk		2016		✓	Neural network
Moradi [40]	Credit risk		2019	✓		Adaptive fuzzy network
Zhang [41]	Credit risk		2020		✓	GBDT, Neural Network
Yang [42]	Credit risk		2018		✓	Time-aware LSTM
Tavana [43]	Liquidity risk		2018		✓	ANN-BN
Guijarro [44]	Liquidity risk		2019	✓		linear regression
Sahin [45] Askari [46]	FCC risk	Card fraud	2013,17	✓		Decision Tree
Mubalaik [47]	FCC risk	Card fraud	2017		✓	MLP
Malini [48]	FCC risk	Card fraud	2017	✓		KNN
Martínez [49]	FCC risk	MM	2016			reinforcement learning
Shi [50]	FCC risk	MM	2019			graph mining
Humpherys [51]	FCC risk	FSF	2011	✓		naive Bayes, decision tree
Li [52] Sun [53]	FCC risk	AML	2020,21			graph mining
Shi [54]	FCC risk	AML	2019			sequence mining
Nyman [55]	Systemic risk		2021	✓		clustering
Zhou [56]	Systemic risk		2020		✓	CNN, BiGRU
Catullo [57]	Systemic risk		2015	✓		reinforcement learning
Yu [58]	Systemic risk		2020	✓		Network model, GBDT
Ahelegbey [59]	Systemic risk		2021	✓		VAR
Bianchi [60]	Systemic risk		2019	✓		MCMC
O'Halloran [61]	Systemic risk		2019	✓		Markov model

*ML: machine learning; DL: deep learning; FCC risk: financial crime compliance risk; FSF: financial statement fraud; AML: anti-money laundering; MM: Market manipulation.

User identity identification One single user may have multiple accounts in different networks. Thus, identifying cross-platform users is important for user profiling and financial risk management. Two main types of approaches for this are supervised methods and unsupervised methods. Compared with the former, unsupervised methods are less susceptible to subtle changes in network structure and are more robust. For example, PALE [18] learns robust embeddings from network structures in an unsupervised manner and then finds matching pairs according to nearby neighbors (see Fig. 1).

3.3. Knowledge graph

As mentioned in Section 2.2, knowledge graphs represent rich information of entities and relations among them, making them highly useful for financial applications. Knowledge graph (KG) is a graph-structured data model to represent knowledge, and knowledge graph reasoning (i.e., inferring missing facts in KG) is a basic task. Recently, translation-based models have been widely researched. These models often learn embeddings by translating a head entity to a tail entity through this relation. For example, TransE [19], a representative translation-based model, maps the entities and relations into the same vector space and forces the added embedding $\mathbf{h} + \mathbf{r}$ of a head entity h and a relation r that is close to the embedding of the tail entity $\mathbf{t}$, $\mathbf{t}$. SENN [20] integrates the prediction tasks of head entities, relations, and tail entities into a neural network with shared embeddings of entities and relations.

KGs can be also used to present sequences of events. When only considering the temporal relation between events, event graphs are often organized into a sequence of event subgraphs. Existing models first encode the subgraph in a time stamp and then model the temporal order information via a sequence model.

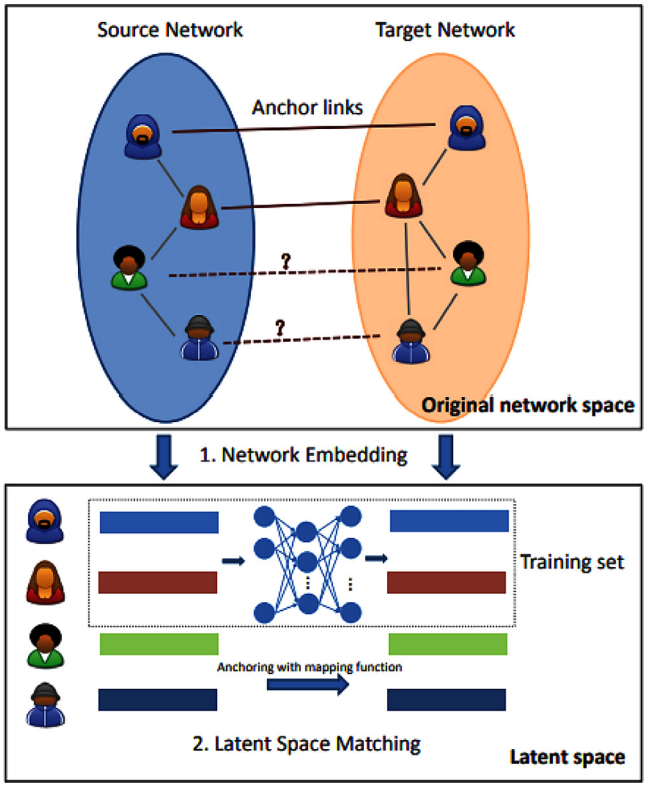


Fig. 1. PALE: Identifying anchor links across networks [18].

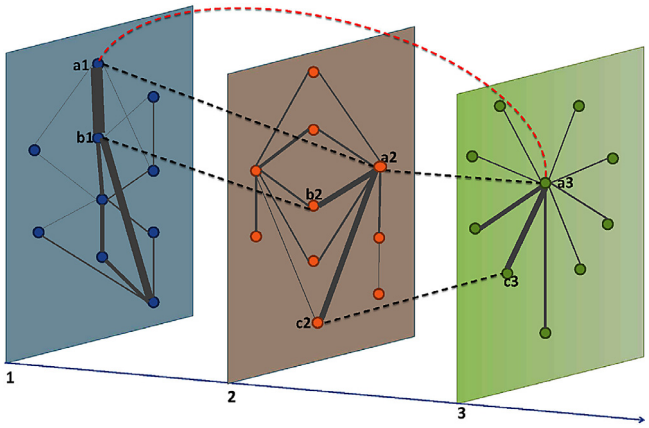


Fig. 2. Schematic of a multi-slice trading network. Three slices, each corresponding to one daily trading network, are shown. The same trader on different trading days is connected by inter-slice connections. Solid lines represent intra-slice connections, and dashed lines represent inter-slice connections [24].

3.4. NLP technologies

In financial markets, text data can provide rich information for analyzing risks. Thus, NLP, including entity extraction, sentiments analysis, machine translation, and representation learning (language model), is an extremely important set of technologies for understanding and mining financial information and knowledge. Furthermore, the success of multilingual machine translation technologies [17] based on big corpora has made the open financial market almost seamless across countries.

Pre-trained language model In recent years, large-scale pre-trained language models based on the transformer [17] architecture have made outstanding breakthroughs in NLP tasks, such as BERT [21] and GPT-3 [22]. They provide a stronger tool for information acquisition from texts in financial analysis, such as sentiment analysis of financial texts [23]. Moreover, GPT-3 exhibits great potential for natural language generation (NLG). Trained on large-scale corpora collected from the web, GPT-3 has shown impressive power to generate high-quality texts that are indistinguishable from human-written texts.

3.5. Graph and Network analysis

In modern finance, considering Not only individual entities and their attributes but also The relationships between them is important. Some data are already present in networks such as investment and transaction networks, whereas other data are implicitly and heterogeneously correlated, for example, data regarding co-founders, business, investment behaviors, and stock exchanges. In this review, we follow the convention that a graph is a mathematical representation of vertices (entities) and their edges (relationships) and use it interchangeably with the term network, which refers to specific instantiations of a graph.

3.5.1. Structure analysis

Networks are randomized and disordered at the microscopic level and often appear regular and ordered at the macroscopic level. This property has attracted researchers to explore and analyze the network mesostructure (i.e., subgraph) that connects micro and macro levels. For example, HoloScope [25] detects fraudulent entities and subgraphs in networks based on topology and spikes. At present, network mesostructure analysis is mainly focused on community structure. Motifs, known as small frequent subgraph patterns in networks, are also important for understanding the structure of large networks [26,27]. Fig. 2 proposes integrating all daily trading networks for stock into a multi-slice trading network([24]). *Communities and subgraphs* Communities exist in a large number of real-world complex networks. Generally, community structure refers to groups of nodes that are “tightly connected internally and sparsely connected externally”.

In the past, uniform metrics for the community are lacking. Subsequently, Newman proposed modularity [28] based on an intuitive understanding of the community structure which is characterized as a high density of connections within a community and relatively sparse connections between different communities. Because modularity provides a good formal definition of community partitioning based on edge density, many methods for community discovery based on modularity optimization have been proposed. Its flexibility also allows researchers to easily extend it to fit different kinds of networks, such as bipartite graphs. It is noteworthy that *spectral graph theory*, which has played an important role in modularity optimization, emerged as powerful tools for application-driven tasks beyond community discovery [29].

Evolution of communities The evolution of community structures is another important issue. Evolution is a fundamental characteristic of real-world networks and reveals interactions between network structures. Palla et al. studied community evolution based on their proposed community discovery method using complete subgraph percolation [30]. They concluded that the stability of small communities is a prerequisite to ensure their existence, whereas the dynamics of large communities is the basis for their existence. The relationship between community structure and network dynamics is closely related to the evolution of communities. The mainstream research method involves learning network dynamics properties through synchronization and diffusion processes. [31] studied the relationship between the diffusion process and community structures and pointed out that the local equilibrium states emerging from the diffusion process can reflect the community structure of networks. Accordingly, they proposed network conductivity to describe the relationship between network community and local equilibrium states and provided a community discovery method based on network conductivity optimization, with significant performance compared with modularity optimization methods.

3.5.2. Graph neural networks

Owing to the powerful representation capability of graphs, graph representation learning has attracted widespread research attention. Furthermore, in recent years, graph neural networks (GNNs) [15,16] have achieved great success in graph representation learning and subsequent tasks, such as node classification and link prediction. GNNs follow a neighborhood aggregation scheme, where the representation of a node is obtained by recursively aggregating and transforming the representations of its neighboring nodes. The success of GNNs can be attributed to their high expressive power of learning representations of nodes and graphs. Existing methods aim to design neural networks for graph data, for example, CNN on graphs, self-attention on graphs, and RNN on graphs. For example, GWNN [32] implements graph convolution operator via graph wavelet transform.

GNNs have achieved great improvement in financial risk detection. HGN [33] constructs a heterogeneous network of account devices and establishes a GCN to identify fake accounts. Furthermore, cash-out detection was investigated in [34]. The authors aggregated neighbor features based on meta-paths to obtain expressions and learned attention from each meta-path to obtain node representations to classify nodes. GRC [35] characterizes multiple types of relationships via the self-attention mechanism and employs a conditional random field to detect loan fraud.

3.5.3. Large-scale graph mining

With the explosive growth in data volume, the number of real-world networks is increasing. This poses great challenges to graph mining, analysis, and processing. Thus, multiple large-scale graph mining methods have been proposed. For example, SpecGreedy [36] efficiently detects multiple types of suspicous dense subgraphs in large networks based on spectral graph theory, which is useful for detecting financial

anomalies, such as subgraphs that suddenly become dense in a transaction network.

In contrast, the *graph summarization* technique, which reduces the input graph to a much smaller summary graph, can help handle large-scale graphs. For example, SWeG [37] finds a compact summary graph with edge corrections and boosted graph queries. DPGS [38] adopts the configuration model as a null model in graph summarization and saves both memory and time for GNN training, which facilitates graph mining tasks on large-scale financial networks.

4. Applications

We now review studies that solve specific financial risks by utilizing big data and emerging methods. The works are summarized in Table 2.

4.1. Credit risk

Credit risk assessment is the basis and key to risk management. While traditional financial credit information often contains only strong financial attributes (e.g., credit card, foreign exchange, private lending, and other financial transaction data), credit evaluation based on big data algorithms comprehensively considers financial data, government public service data, life, and social data, among others. Ha et al. [39] constructed a credit scoring model based on deep learning and feature selection and evaluated applicants' credit ratings according to the input characteristics of applicants. The test showed that deep learning is an effective method for managing high-dimensional credit feature data. Moradi and Rafiei [40] proposed an adaptive network-based fuzzy inference system that accommodates both customer profile data and fluctuating politico-economic factors to assess credit risk. The dynamic system could produce a table of bad customers on a monthly basis, and the results were highly compatible with real-life situations. Zhang et al. [41] designed an online integrated credit scoring model that can be updated in time based on an innovative learning framework called DeepGBM, which integrates the advantages of the gradient boosting decision tree (GBDT) and neural network to handle both categorical and categorical numerical tabular features [62]. As an example of non-traditional data, [42] built a DeepCredit model mainly based on time-aware LSTM, using a complete user clickstream dataset from a P2P lending platform that recorded more than 4 million financial activities of 10 thousand users. The model predicted both individual delinquencies and defaults with high accuracy.

4.2. Liquidity risk

The provision of liquidity is key to all theories of financial intermediation [44]. In traditional models, liquidity risk is measured by considering changes in some indicators over a specific time horizon for comparison [63]. However, plurality, multiplicity, and diversity of accounts make it quite difficult and time-consuming to calculate cash flows, and thus these data are challenging to obtain in a short period of time.

Nowadays, big data algorithms can be used to better evaluate liquidity risk and analyze the key factors and their interconnections. Tavana et al. [43] used neural networks to analyze and evaluate liquidity risk and key factors. The proposed ANN-BN two-stage model primarily aimed to facilitate the systematic analysis of bank-specific measures based on balance sheet ratios and was proven to be flexible enough to be applied to any loan-based scenario. Guijarro et al. [44] analyzed the impact of social media on financial market liquidity by performing a regression analysis with the S&P500 index after scoring it based on Twitter content sentiment. The results showed that after adding a two-day moving average volume, investor sentiment is significantly and positively correlated with the effective spread of liquidity.

4.3. Volatility risk

Volatility risk in financial markets is the likelihood of fluctuations in portfolios' price under the effect of the changes of certain risk factors. Currently, one of the principal measurements to quantify financial risk is value at risk (VaR), in which the prediction of market volatility plays a crucial role. As price return is the key to understanding and modeling market volatility, research on volatility risk mainly focuses on price returns, where sequential data are common and widely used. In previous studies, researchers have found that the distribution of price returns displays a fat tail (Fig. 3) [64]. The price return does not have linear autocorrelation, whereas the absolute value of the price return displays a long-range memory [65].

In the big data era, the traditional analysis methods for sequential data generally fall into three categories: historical simulation, analytical method, and Monte Carlo simulation; however, these methods often fail to produce accurate results. Conversely, big data algorithms can overcome this difficulty. By employing the LSTM network, Fischer and Krauss [66] achieved a better prediction of the directional movements for the S&P 500 stocks over a long time span, compared with logistic regression, random forest, and even standard deep neural networks. They further showed that the fluctuation of the portfolio guided by LSTM networks due to common risk factors was significantly reduced.

4.4. Market manipulation

Financial markets provide a platform for companies to raise capital by allowing investors to trade stocks, and under normal circumstances, the prices of financial products reflect common judgments about the value of companies. However, market prices can be manipulated by misinformation and fraudulent trading, affecting investor information and disrupting the market to the detriment of financial market development. Martínez et al. [49] used a reinforcement learning framework within the full and partial observability of Markov decision processes and analyzed the underlying behavior of the perpetrators by finding the causes of what encourages these traders to perform fraudulent activities. In addition to machine learning methods, some network analysis methods have been applied to the detection of market manipulation. Some studies have attempted to detect trading-based manipulation in the Chinese stock market by analyzing trading networks [67], and the manipulated stocks were effectively distinguished by a degree-strength correlation approach [68]. Furthermore, by constructing a multi-slice trading network [24], they successfully identified abnormal traders in the stock market. In addition, they analyzed the clique of the trading network to detect colluded stock manipulation [50].

4.5. Systemic financial risk

The different types of risks discussed previously could trigger a collapse in a certain industry or economy. Therefore, appropriate risk management is required to treat systemic risk, especially from a regulatory perspective. Traditional systemic risk management has mainly relied on institutional micro-indicators and macro-economic indicators, such as asset adequacy, asset quality, and liquidity, laying the foundation for risk monitoring methods, including weighted average or structural models. However, limited by the update mechanism or intrinsic property, traditional indicators often suffer from hysteresis, low frequency, and other disadvantages, and incorporating alternative data could alleviate the problem. Nyman et al. [55] analyzed financial market text-based data to assess the effect of emotional content on the development of a financial system, revealing the formation of exuberance before the global financial crisis as well as the subsequent collapse.

With the development of big data algorithms with high computational power, a series of new data-driven methods have emerged. Nonlinear methods such as an SVM and neural networks provide higher prediction accuracy for systemic risk as a nonlinear problem. Zhou

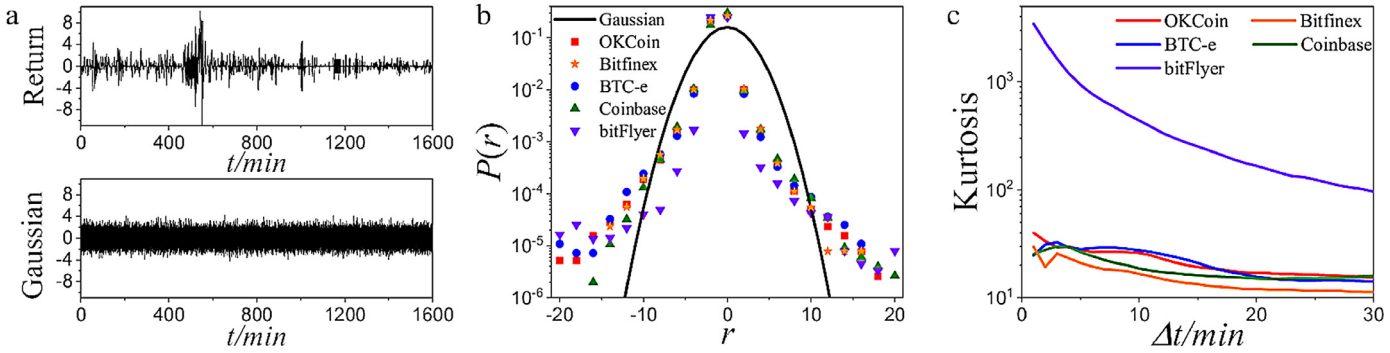


Fig. 3. Fat-tail of the price return distribution. (a) The top pane is an example of the normalized price return on the OKCoin platform from 14:04 on March 3, 2017, to 22:45 on March 4, 2017, and the bottom pane is the noise signal sampled from a Gaussian distribution. (b) The distribution of normalized price returns in different Bitcoin platforms compared with Gaussian distribution when $\Delta t = 2$ min. **c**, Kurtosis κ versus Δt in different Bitcoin platforms [50].

et al. [56] proposed a hybrid deep learning model based on a CNN and BiGRU to predict systemic financial risk, achieving superior performance in feature learning.

After the 2008 financial crisis, a large number of dynamic models based on emerging economic and financial theories described the evolution of systemic risk. Catullo et al. [57] reproduced the macroeconomic dynamics generated from the interaction between heterogeneous banks and enterprises in the endogenous credit network, based on which the authors defined the early warning indicators of crises. Yu et al. [58] simulated an interbank network using a complex network approach to predict systemic risk contagion and provided a method to select the optimal management policy for handling systemic risk. Additionally, researchers in other major branches have adopted data-driven statistical learning methods to build risk evolution models. Systemic risk indicators and macroeconomic variables are used as state variables, and vector auto-regression (VAR) and Markov chain are used as a more general form of VAR to obtain a state transition matrix based on historical data. Ahelegbey et al. [59] proposed a network VAR model to assess the financial impact of the COVID-19 pandemic. Recently, many AI approaches have been applied to systemic risk evolution research. O'Halloran and Nowaczyk [61] designed a systemic risk engine based on the Open Source Risk Engine and incorporated various risk metrics; they also adapted a simulation technology to assess the impact of regulations on the financial system in general.

5. Financial crime compliance risk

Financial crime compliance risk relates to losses that may arise when a company or institution fails to comply with the laws and regulations relevant to financial crimes in their respective jurisdictions. Typical financial crimes include identity theft, market manipulation, money laundering, and financial statement fraud. In this section, we will introduce the emerging big data approaches for the management of financial crime compliance risk.[14]

5.1. Identity theft

Identity theft refers to criminals using someone else's identity and other relevant information in unauthorized ways. It often results in immediate financial loss, and victims may suffer from a series of credit as well as other problems. To reduce the risk, Wang et al. [69] used a probabilistic generative model to detect identity theft in mobile social networks. Such crime conducted through the internet are called "phishing." Rao et al. [70] detected phishing websites by employing a random forest classifier based on heuristic features extracted from the URL, source codes, and third-party services. In [71], Benavides et al. provided a detailed description of deep learning approaches for tackling phishing attacks.

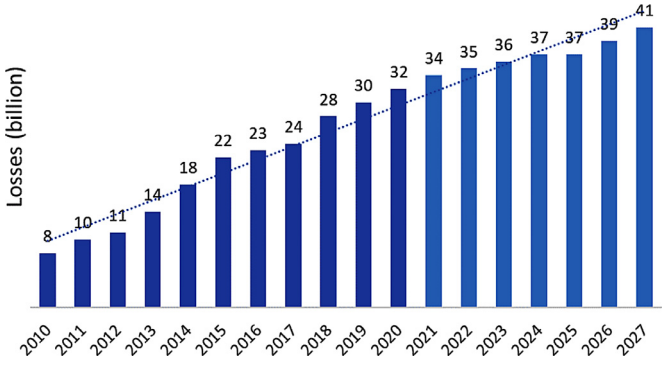


Fig. 4. Global losses from payment fraud have tripled from $9.84 Billion in 2011 to $32.39 in 2020. Payment fraud is expected to continue increasing and is projected to cost $ 40.62 billion in 2027. The data after 2020 are the predicted values[14].

One form of identity theft is credit card fraud, which typically only affects one or more of the victim's open credit card accounts[15]. As shown in Fig. 4, fraudulent transactions via credit cards could be as high as $ 41 billion by 2027. Many approaches have been designed for such an anti-fraud urgency. Malini and Pushpa [48] adopted the KNN algorithm and outlier detection methods, combined with oversampling and PCA techniques, to optimize the fraud detection rate for bank credit card fraud detection. Askari and Hussain [46] proposed a FuzzITree algorithm based on an ID3 decision tree using fuzzy logic to discover fraudulent transactions.

5.2. Financial statement fraud & insurance fraud

Financial statement fraud Financial statement fraud refers to deliberate misrepresentation of the financial condition of an enterprise accomplished by intentional misstatement or omission of amounts or disclosures in financial statements[16]. Losses from financial statement fraud are often quite large and sometimes even catastrophic to institutions. Therefore, financial statement fraud has been a serious concern for auditors, investors, and regulators. Various big data algorithms have been developed to address this risk. A common workflow consists of feature selection from diverse data (including report text and financial indicators) and classification for fraud prediction. After employing a wide

[14] https://nilsonreport.com/mention/1313/1link

[15] https://creditcards.usnews.com/articles/credit-fraud-vs-identity-theft-whats-the-difference

[16] https://www.acfe.com/article.aspx?id=4294967876

range of big data algorithms for both feature selection and classification to detect fraud, Hajek and Henriques [72] found that ensemble methods perform best for fraudulent companies and Bayesian belief networks (BBN) for non-fraudulent companies. With similar considerations, Yao et al. [73] compared several hybrid methods and found the best combination of XGBoost and random forest for statement fraud detection. Aided by deep learning, Craja et al. [74] achieved improved results compared to many current approaches by adopting a hierarchical attention network (HAN) for feature extraction to better reflect document structures and concentrate on both the content and context of the text. [75] used multiple data mining techniques to identify companies that resort to financial statement frauds. In addition, [76] evaluated multiple methods to predict financial statement frauds.

Insurance fraud Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for financial gain[17]. According to the FBI, non-health insurance fraud costs more than $40 billion per year as estimated and causes more premium burdens on the U.S. family[18]. With the burst of the COVID-19 pandemic, the number of potential fraud claims is believed to double[19]. However, traditional methods usually fail to handle textual information in claims, which could provide a valuable reference for insurance fraud detection. By leveraging deep learning, Wang and Xu [77] combined traditional numeric features with text features extracted by LDA for neural network training to detect automobile insurance fraud, and the results outperform widely used models such as SVM and random forest.

5.3. Anti-money laundering

Money laundering is the behavior of concealing the source of money achieved through illegitimate activities. According to the United Nations The Office on Drug and Crime, the estimated amount of money laundered globally in one year is 2–5% of global GDP or $800 billion - $2 trillion in current US dollars[20]. Therefore, anti-money laundering (AML) is of critical significance to national financial stability.

The most classic approach used in the bank for AML is rule-based classification. By using bid/ask orders, price returns could provide an innovative and effective method to detect abnormity in Bitcoin platforms [54]. However, these rule-based algorithms rely heavily on expert knowledge, are easy to evade by fraudsters, and cannot be used to discover new types of money laundering behaviors.

Machine learning algorithms are also applied for detecting money laundering activities, which can be effectively interpolated on new scenarios without the constraints of fixed rules. SVM was previously applied to process large data sizes and achieved high accuracy. Stavarache et al. [79] proposed a deep learning–based method trained for an anti-money laundering tasks using customer-to-customer relations. However, these algorithms detect money laundering activities in supervised or semi-supervised manners, suffering from imbalanced classes and lack of adaptability. In addition, clustering-based methods have been applied to the detection of money laundering activities by grouping suspicious transactions into clusters. However, these methods do not consider the interaction between accounts, resulting in a high false-positive rate of detection.

Graphs and networks are powerful representations for analyzing the inner-dependency among suspicious accounts involved in money laundering. Fig. 5 shows a real example of money laundering transfers in a bank, containing a two-step flow from source to middle to destination accounts, where three types of accounts are denoted by A, M, C respectively. Colladon and Remondi [80] utilized social network analysis to reveal the underlying roles and organization structures. However, these methods do not perform flow tracking or provide theoretical guarantees. A flow-based scalable approach FlowScope [78] was proposed for detecting the complete transfer flow of money from the source to the destination, which had a theoretical bound. It modeled the transactions using a multipartite graph and designed a novel anomaly metric to be optimized. The extensive experiments provided in this paper showed that Flowscope is effective and robust in detecting different adversarial money laundering behaviors. On this basis, Sun et al. [53] modeled the transfer flow as two coupled tensors, considering multiple attributes of transactions(e.g., time).

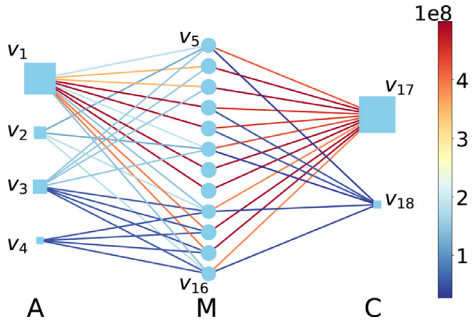


Fig. 5. Example of money laundering transfers in a bank, creating a dense tripartite subgraph [78]. Edge color and node size indicate the amount of money transferred.

5.4. Trafficking & sanctions

Trafficking Anti-trafficking policies often play an important role in policymaking. In recent years, various technologies have been considered to combat trafficking. Data such as financial transactions, mobile phone calls, and text messages contribute significantly to anti-trafficking work. Musto [81] explored the possibility of using big data to make anti-trafficking more effective.

Sanctions The retrogression of globalization and the intensification of the geopolitical struggle lead to occasional economic sanctions placed over countries. For example, sanctions on Iran and Russia have made people begin to evaluate the subsequent loss using algorithms. Tregub et al. [82] used the OLS algorithm to estimate the economic loss due to sanctions imposed by the European Union on Russian Federation and Russian counter-sanctions. Furthermore, sanctions can play a role in predicting bankruptcy [83].

5.5. Anti-bribery and corruption (ABC)

Corruption is the abuse of public power motivated by personal gain, an act that diverts tax revenues from national development. It not only erodes trust in government and institutions but also reduces the effectiveness and fairness of public policies. According to the International Monetary Fund (IMF), approximately $ 1 trillion in global government revenues are lost due to corruption each year[21].

Advances in information technology have made it possible to use machine learning methods to prevent and detect corruption. In some countries, relevant techniques were already deployed to fight against government corruption. For example, in the UK, Exiger and Transparency International (TI) attempted to identify corruption risks by analyzing public records. IBM has also been working with the Kenyan government to identify the key drivers of bribery through algorithms. In 2020, Microsoft announced the launch of its anti-corruption technology and solutions (ACTS), which will employ technologies such as cloud com-

[17] https://www.iii.org/article/background-on-insurance-fraud
[18] https://www.fbi.gov/stats-services/publications/insurance-fraud
[19] https://knowledge.friss.com/en-us/2020-insurance-fraud-report
[20] https://www.unodc.org/unodc/en/money-laundering/overview.html
[21] https://www.oxfordinsights.com/ai-for-anti-corruption

puting, AI, and machine learning to detect and deter corruption in the coming decades[22].

6. Future directions

We summarize several important research directions, where open problems in a wide range of domains, in addition to finance, can be critical challenges facing the combat with financial risks in the big data era.

6.1. Deepfake techniques and detection

Deep learning has achieved great success in the field of NLP and computer vision. However, GPT-3's power to generate high-quality text can be used to produce fake news or opinions. Deepfake techniques based on deep learning have been applied to create fake images and synthesize fake videos or speech [86,87]. These techniques, trained on large images and video datasets, can produce fake news and political rumors by tampering or replacing the face information of the original videos and synthesizing fake speeches. Financial fraud has been successfully carried out using the deepfake technique [88]. Meanwhile, public opinion can be swapped by fake news and videos. When not adequately responded to, this can impact companies' reputation, influence consumers' behaviors, and even affect stock prices, eventually endangering the entire financial market. However, detecting and identifying fakes remains an open problem.

6.2. Adversarial attacks

Researchers have found that deep neural networks can be easily fooled by the so-called "adversarial samples," which are obtained by adding imperceptible-to-human perturbations to the original samples [89,90]. Such "adversarial attack" techniques evade detections and pose potential threats to the applicants, especially in the financial field, which can result in highly critical consequences. For example, in credit scoring systems, fraudsters can fake a friendship connection with others to evade fraud detection models [91].

In recent years, researchers are focused on designing defense methods against adversarial attacks [84,85,92]. A provably robust neural network [84] was proposed for node classification via low-pass message passing. It is theoretically upper-bounded under adversarial attacks, with an easy-to-plugin module for GNNs, which is as robust as linear attack budgets, and as accurate as neural networks. Fig. 6a shows that guarded by the low-pass message passing mechanism, Node 1's embedding only has a slight shift, whereas the embedding from the other baseline shifts significantly. Certified defense [85] provided the worst-case performance bound for a given attack, and adversarial immunization was proposed to improve the certifiable robustness against any admissible adversarial attack, as shown in Fig. 6b.

These defense methods can be used to address potential crime threats, such as misleading the judgment of investments and recommendations by deep learning models. However, from the application perspective, no one has yet designed a powerful defense algorithm that can resist a wide variety of adversarial example attack algorithms.

6.3. Causality and interpretability

Currently, big data algorithms applied to financial risks are mainly based on statistics that emphasize the correlation of factors. Such models would be vulnerable to environmental changes if they were built upon correlations rather than underlying causal relationships. Modeling causal relationships is valuable for making financial predictions and decisions, which can improve prediction robustness, provide interpretable results, and enable counterfactual inference.

Traditional studies on causality require randomized controlled trials (RCTs) to determine the exact causal relationship. However, big data brings us another approach, that is, mining causal relationships inside observational data. Following the methodology of the structured causal model (SCM) [93], causal relationships are represented by a directed acyclic graph (DAG), where nodes are factors and edges are direct cause-effect relationships.

Causal discovery is a promising technique for mining causal relationships in financial data. Many effective methods for causal discovery exist for stationary data. Nevertheless, future challenges for modeling causal relationships in finance mainly lie in the issues of hidden factors. Hidden factors are critical, but unobservable, factors and may act as confounders of other observable factors, leading to spurious correlations.

6.4. Privacy

As mentioned previously, big data can facilitate many financial applications. However, there still remain privacy concerns, especially in the finance field. Data used in financial applications often contain sensitive information about users, while companies tend to keep data to themselves. Therefore, companies that collect users' locations through weather forecast apps could have a better ability to predict users' creditworthiness than traditional credit bureaus [94].

In addition to normalizing the use of users' shared information, global coordination of data-sharing from different countries or sectors is crucial to identify risks across individual borders. However, such cross-border data-sharing is always challenging due to the complex data-sharing policies of different countries or varying interests and regulators within an individual country. This results in greater fragmentation of the global digital economy and obstructs the combating of financial risks[23].

Therefore, a question can be raised: How can we use data related to users' sensitive information properly and safely and at the same time eliminate the fragmentation of economic data? Algorithms such as differential privacy [95], which has been listed as one of "10 breakthrough technologies 2020" by MIT Technology Review[24], is one of the key directions to combat financial risks in the future.

6.5. Multi-agent simulation

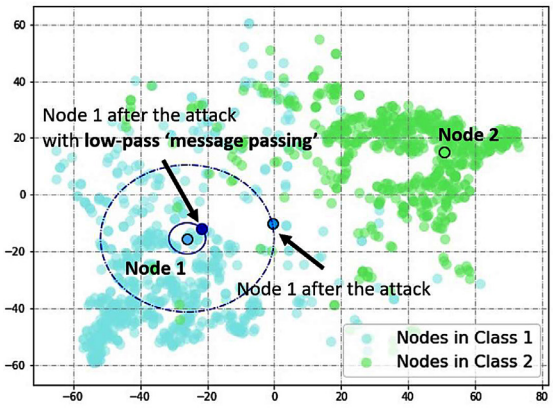
Most macro- and micro-studies on financial problems focus on inconsistent topics, with rare connections between them. Data from individual sensors provide a micro-view of financial behaviors, whereas statistics provide a macro-view. Recently, the simulation of multi-agent reinforcement learning has shown surprising results, for example, AlphaZero, as a single agent, played with another copy as an opponent at the same time, became a master of Go [96], and succeeded in playing a multi-agent game Dota [97].

The agent-based simulation exhibits the ability to find a better solution or forecasting than humans, given the proper environments and awards. Therefore, close-to-real environments of the real-world economy, multi-agent simulations, and learning can hopefully be a systematic solution to deeply understand our physical economy system and address emerging risks within it. Although the concept of "digital twins" was initiated for physical model simulation of spacecraft, it has become

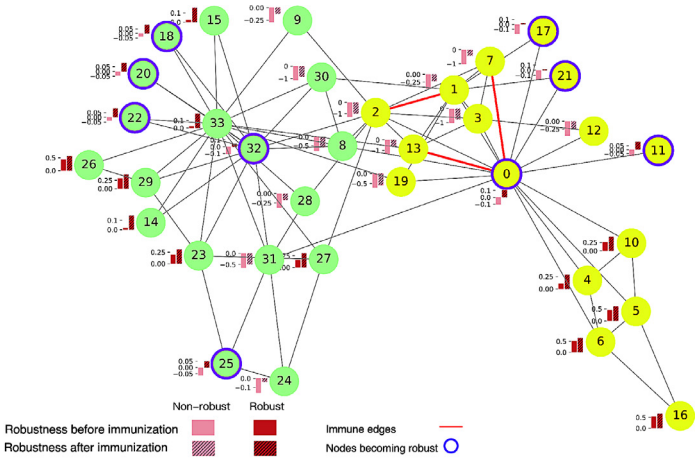
[22] https://venturebeat.com/2020/12/09/microsoft-launches-effort-to-fight-corruption-with-ai-and-other-emerging-technologies/

[23] https://www.imf.org/external/pubs/ft/fandd/2021/03/how-to-build-a-better-data-economy-carriere.htm

[24] https://www.technologyreview.com/10-breakthrough-technologies/2020/



(a) Node classification with adversarial attacks [84]



(b) Adversarial immunization on Karate club network [85]

Fig. 6. (a) Embeddings (i.e. representations) of nodes from two classes. Guarded by low-pass "message passing," node 1's embedding is slightly shifted under the attack compared with the unguarded method. (b) Two classes of nodes with different colors. Two vertical bars beside nodes represent node's robustness before and after immunization. The node is certified as robust (red) when its robustness is > 0; otherwise, it is non-robust (pink). Purple circle indicates the node that becomes robust through immunization. The red edges are immune edges.

an emerging topic in the digital economy[25]; however, the complexity of the real world makes the digital twin far from being realized.

7. Conclusion

In this paper, we presented a perspective review on the studies that focus on addressing emerging financial risks in the big data era, which has brought us challenges in terms of utilizing big and multimodal data; analyzing with long-range and heterogeneous dependent, dynamic, and real-time nature of data; and the adversary of financial fraud. We proposed a reviewing framework to classify the related works:

- what data to utilize,
- how to empower big data with the emerging tools that can analyze or learn from, and
- highlighting how successful the research works have been in various applications.

Finally, in addition to discussing methods that can address privacy issues, we list what we should do in the future in terms of handling risks emanating from deepfake techniques, adversarial attacks on deep models, cause-effect methods, and simulation of physical world, which are still open problems and gaining momentum.

Declaration of Competing Interest

The authors declare that they have no conflict of interest in this work.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 91746301, 61772498, 61802370, and 61902380.

References

[1] X. Cheng, H. Mei, W. Zhao, Data science and computing intelligence: concept, paradigm, and opportunities, Bull. Chin. Acad. Sci. 35 (12) (2020) 1470–1481.
[2] M. Cebrian, The past, present and future of digital contact tracing, Nat. Electron. 4 (1) (2021) 2–4.
[3] D. Donaldson, A. Storeygard, The view from above: applications of satellite data in economics, J. Econ. Perspect. 30 (4) (2016) 171–198.
[4] T.-M. Choi, J.H. Lambert, Advances in risk analysis with big data, 2017,
[5] M.M. Hasan, J. Popp, J. Oláh, Current landscape and influence of big data on finance, J. Big Data 7 (1) (2020) 1–17.
[6] C. Moreira, E. Haven, S. Sozzo, A. Wichert, Process mining with real world financial loan applications: Improving inference on incomplete event logs, PLoS One 13 (12) (2018) e0207806.
[7] J. Han, From unstructured text to textcube: automated construction and multidimensional exploration, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 5–6.
[8] L. Huakui, H. Liang, M. Feicheng, Constructing knowledge graph for financial equities, Data Anal. Knowl. Discov. 4 (5) (2020) 27–37.
[9] F. Mai, S. Tian, C. Lee, L. Ma, Deep learning models for bankruptcy prediction using textual disclosures, Eur. J. Oper. Res. 274 (2) (2019) 743–758.
[10] S. Liu, X. Cheng, F. Li, F. Li, Tasc: Topic-adaptive sentiment classification on dynamic tweets, IEEE Trans. Knowl. Data Eng. 27 (6) (2014) 1696–1709.
[11] S. Qiu, Y. Luo, H. Guo, Multisource evidence theory-based fraud risk assessment of China's listed companies, J. Forecast.
[12] Y. Qin, Y. Yang, What you say and how you say it matters: predicting stock volatility using verbal and vocal cues, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 390–401.
[13] A.H. Alizadeh, N.K. Nomikos, Ship finance: hedging ship price risk using freight derivatives, in: Blackwell Companion to Maritime Economics, first ed., Wiley-Blackwell, 2012, pp. 433–451.
[14] D. Vemula, G.R. Gangadharan, Towards an internet of things framework for financial services sector, 3rd IEEE international conference on Recent Advances in Information Techanology, 2016.
[15] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations (ICLR), 2017.
[16] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Advances in Neural Information Processing Systems, 2017, pp. 1024–1034.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.
[18] T. Man, H. Shen, S. Liu, X. Jin, X. Cheng, Predict anchor links across social networks via an embedding approach, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1823–1829.
[19] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Advances in neural information processing systems, 2013, pp. 2787–2795.
[20] S. Guan, X. Jin, Y. Wang, X. Cheng, Shared embedding based neural networks for knowledge graph completion, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 247–256.
[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[22] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
[23] Y. Yang, U.Y. Mark Christopher Siy, A. Huang, FinBERT: apretrained language model for financial communications, 2020, 2006.08097
[24] X.-Q. Sun, H.-W. Shen, X.-Q. Cheng, Y. Zhang, Detecting anomalous traders using multi-slice network analysis, Physica A 473 (2017) 1–9.
[25] S. Liu, B. Hooi, C. Faloutsos, Holoscope: topology-and-spike aware fraud detection,

[25] https://www.gray.com/insights/digital-twins-an-emerging-force-in-the-digital-economy/

in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, in: CIKM '17, Association for Computing Machinery, 2017, pp. 1539–1548.
[26] A. Paranjape, A.R. Benson, J. Leskovec, Motifs in Temporal Networks, WSDM '17, Association for Computing Machinery, 2017, pp. 601–610.
[27] D. Koutra, U. Kang, J. Vreeken, C. Faloutsos, Vog: summarizing and understanding large graphs, in: Proceedings of the 2014 SIAM international conference on data mining, SIAM, 2014, pp. 91–99.
[28] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[29] X. Cheng, H. Shen, Community structures in complex networks, Complex Syst. Complex. Sci. 8 (01) (2011) 57–70.
[30] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664–667.
[31] X.-Q. Cheng, H.-W. Shen, Uncovering the community structure associated with the diffusion dynamics on networks, J. Stat. Mech. 2010 (04) (2010) P04024.
[32] B. Xu, H. Shen, Q. Cao, Y. Qiu, X. Cheng, Graph wavelet neural network, in: International Conference on Learning Representations, 2019.
[33] Z. Liu, C. Chen, X. Yang, J. Zhou, X. Li, L. Song, Heterogeneous graph neural networks for malicious account detection, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 2077–2085.
[34] B. Hu, Z. Zhang, C. Shi, J. Zhou, X. Li, Y. Qi, Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 946–953.
[35] B. Xu, H. Shen, B. Sun, r. An, Q. Cao, X. Cheng, Towards consumer loan fraud detection: graph neural networks with role-constrained conditional random field, AAAI, 2021.
[36] W. Feng, S. Liu, D. Koutra, H. Shen, X. Cheng, Specgreedy: unified dense subgraph detection, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2020.
[37] K. Shin, A. Ghoting, M. Kim, H. Raghavan, Sweg: lossless and lossy summarization of web-scale graphs, in: The World Wide Web Conference, 2019, pp. 1679–1690.
[38] H. Zhou, S. Liu, K. Lee, K. Shin, H. Shen, X. Cheng, Dpgs: degree-preserving graph summarization, SDM (2021).
[39] V.S. Ha, H.N. Nguyen, K. Abou-El-Hossein, Credit scoring with a feature selection approach based deep learning, Matec Web Conf. 54 (2016) 05004.
[40] S. Moradi, F.M. Rafiei, A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks, Financ. Innov. 5 (1) (2019) 1–27.
[41] Z. Zhang, K. Niu, Y. Liu, A deep learning based online credit scoring model for p2p lending, IEEE Access 8 (2020) 177307–177317.
[42] Z. Yang, Y. Zhang, B. Guo, B.Y. Zhao, Y. Dai, Deepcredit: exploiting user cickstream for loan risk prediction in p2p lending, in: Proceedings of the International AAAI Conference on Web and Social Media, 12, 2018.
[43] M. Tavana, A.-R. Abtahi, D. Di Caprio, M. Poortarigh, An artificial neural network and Bayesian network model for liquidity risk assessment in banking, Neurocomputing 275 (2018) 2525–2554.
[44] F. Guijarro, I. Moya-Clemente, J. Saleemi, Liquidity risk and investors' mood: linking the financial market liquidity to sentiment analysis through twitter in the s&p500 index, Sustainability 11 (24) (2019) 7048.
[45] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Syst. Appl. 40 (15) (2013) 5916–5923.
[46] S.M.S. Askari, M.A. Hussain, Credit card fraud detection using fuzzy ID3, in: 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 446–452.
[47] A.M. Mubarek, E. Adalı, Multilayer perceptron neural network technique for fraud detection, in: 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 383–387.
[48] N. Malini, M. Pushpa, Analysis on credit card fraud identification techniques based on KNN and outlier detection, in: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017, pp. 255–258.
[49] E. Martínez-Miranda, P. McBurney, M.J. Howard, Learning unfair trading: a market manipulation analysis from the reinforcement learning perspective, in: EAIS '16, IEEE 2016, 2016, pp. 103–109.
[50] F.-B. Shi, X.-Q. Sun, H.-W. Shen, X.-Q. Cheng, Detect colluded stock manipulation via clique in trading network, Physica A 513 (2019) 565–571.
[51] S.L. Humpherys, K.C. Moffitt, M.B. Burns, J.K. Burgoon, W.F. Felix, Identification of fraudulent financial statements using linguistic credibility analysis, Decis. Support Syst. 50 (3) (2011) 585–594.
[52] X. Li, S. Liu, Z. Li, X. Han, C. Shi, B. Hooi, H. Huang, X. Cheng, Flowscope: spotting money laundering based on graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 4731–4738.
[53] X. Sun, J. Zhang, Q. Zhao, S. Liu, J. Chen, R. Zhuang, H. Shen, X. Cheng, Cubeflow: money laundering detection with coupled tensors 78–90. arXiv preprint arXiv:2103.12411.
[54] F.-B. Shi, X.-Q. Sun, J.-H. Gao, L. Xu, H.-W. Shen, X.-Q. Cheng, Anomaly detection in bitcoin market via price return analysis, PLoS One 14 (6) (2019) e0218341.
[55] R. Nyman, S. Kapadia, D. Tuckett, News and narratives in financial systems: exploiting big data for systemic risk assessment, J. Econ. Dyn. Control 127 (2021) 104119.
[56] Y. Zhou, J. Yan, A hybrid deep learning approach for systemic financial risk prediction, in: International Conference on Computational Science and Its Applications, Springer, 2020, pp. 859–874.
[57] E. Catullo, M. Gallegati, A. Palestrini, Towards a credit network based early warning indicator for crises, J. Econ. Dyn. Control 50 (2015) 78–97.
[58] J. Yu, J. Zhao, Prediction of systemic risk contagion based on a dynamic complex network model using machine learning algorithm, Complexity 2020 (2020).
[59] D.F. Ahelegbey, P. Cerchiello, R. Scaramozzino, Network based evidence of the financial impact of covid-19 pandemic, Available at SSRN 3780954 (2021).
[60] D. Bianchi, M. Billio, R. Casarin, M. Guidolin, Modeling systemic risk with Markov switching graphical SUR models, J. Econ. 210 (1) (2019) 58–74.
[61] S. O'Halloran, N. Nowaczyk, An artificial intelligence approach to regulating systemic risk, Front. Artif. Intell. 2 (2019) 7.
[62] G. Ke, Z. Xu, J. Zhang, J. Bian, T.-Y. Liu, DeepGBM: a deep learning framework distilled by GBDT for online prediction tasks, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 384–394.
[63] M.M. Cornett, J.J. McNutt, P.E. Strahan, H. Tehranian, Liquidity risk management and credit supply in the financial crisis, J. Financ. Econ. 101 (2) (2011) 297–312.
[64] J. Osterrieder, J. Lorenz, A statistical risk assessment of bitcoin and its extreme tail behavior, Annals of Financial Economics 12 (01) (2017) 1750003.
[65] Y. Jiang, H. Nie, W. Ruan, Time-varying long-termmemory in bitcoinmarket, Finance Research Letters 25 (2018) 280–284.
[66] T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, Eur. J. Oper. Res. 270 (2) (2018) 654–669.
[67] X.-Q. Sun, X.-Q. Cheng, H.-W. Shen, Z.-Y. Wang, Distinguishing manipulated stocks via trading network analysis, Physica A 390 (20) (2011) 3427–3434.
[68] X.-Q. Sun, H.-W. Shen, X.-Q. Cheng, Z.-Y. Wang, Degree-strength correlation reveals anomalous trading behavior, PLoS One 7 (10) (2012) e45598.
[69] C. Wang, B. Yang, J. Luo, Identity theft detection in mobile social networks using behavioral semantics, in: 2017 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, 2017, pp. 1–3.
[70] R.S. Rao, A.R. Pais, Detection of phishing websites using an efficient feature-based machine learning framework, Neural Comput. Appl. 31 (8) (2019) 3851–3873.
[71] E. Benavides, W. Fuertes, S. Sanchez, M. Sanchez, Classification of phishing attack solutions by employing deep learning techniques: a systematic literature review, developments and advances in defense and security(2020) 51–64.
[72] P. Hajek, R. Henriques, Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods, Knowl.-Based Syst. 128 (2017) 139–152.
[73] J. Yao, J. Zhang, L. Wang, A financial statement fraud detection model based on hybrid data mining methods, in: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, 2018, pp. 57–61.
[74] P. Craja, A. Kim, S. Lessmann, Deep learning for detecting financial statement fraud, Decis. Support Syst. 139 (2020) 113421.
[75] E. Duman, M.H. Ozcelik, Detecting credit card fraud by genetic algorithm and scatter search, Expert Syst. Appl. 38 (10) (2011) 13057–13063.
[76] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decis. Support Syst. 50 (2) (2011) 491–500.
[77] Y. Wang, W. Xu, Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, Decis. Support Syst. 105 (2018) 87–95.
[78] X. Li, S. Liu, Z. Li, X. Han, C. Shi, B. Hooi, H. Huang, X. Cheng, Flowscope: spotting money laundering based on graphs, AAAI, 2020.
[79] L.L. Stavarache, D. Narbutis, T. Suzumura, R. Harishankar, A. Žaltauskas, Exploring multi-banking customer-to-customer relations in AML context with poincar\'e embeddings, arXiv preprint arXiv:1912.07701
[80] A.F. Colladon, E. Remondi, Using social network analysis to prevent money laundering, Expert Syst. Appl. 67 (2017) 49–58.
[81] J. Musto, The limits and possibilities of data-driven antitrafficking efforts, Ga. St. UL Rev. 36 (2019) 1147.
[82] I.V. Tregub, K.A. Dremva, Estimating the consequences of Russia's and the EU's sanctions based on OLS algorithm, Int. J. Mach. Learn.Comput. 9 (4) (2019) 496.
[83] S.V. Kaledin, Changes of Methodology in Assessment of Probability of Bankruptcy of the Russian Companies in the Conditions of Economic Sanctions, Cambridge International Academics, 2018.
[84] Y. Wang, S. Liu, M. Yoon, H. Lamba, W. Wang, C. Faloutsos, B. Hooi, Provably robust node classification via low-pass message passing, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 621–630.
[85] S. Tao, H. Shen, Q. Cao, L. Hou, X. Cheng, Adversarial immunization for certifiable robustness on graphs, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, in: WSDM'21, 2021.
[86] S. Suwajanakorn, S.M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Trans. Graph. 36 (4) (2017) 1–13.
[87] T.T. Nguyen, C.M. Nguyen, D.T. Nguyen, D.T. Nguyen, S. Nahavandi, Deep learning for deepfakes creation and detection: asurvey, arXiv preprint arXiv:1909.11573
[88] M. Schreyer, T. Sattarov, B. Reimer, D. Borth, Adversarial learning of deepfakes in accounting, arXiv preprint arXiv:1910.03810
[89] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572
[90] D. Zügner, A. Akbarnejad, S. Günnemann, Adversarial attacks on neural networks for graph data, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, in: KDD '18, 2018, pp. 2847–2856.
[91] W. Jin, Y. Li, H. Xu, Y. Wang, J. Tang, Adversarial attacks and defenses on graphs: a review and empirical study, arXiv:2003.00653(2020).
[92] T. Pang, K. Xu, C. Du, N. Chen, J. Zhu, Improving adversarial robustness via promoting ensemble diversity, in: International Conference on Machine Learning, PMLR, 2019, pp. 4970–4979.
[93] J. Pearl, Causality, Cambridge university press, 2009.
[94] T. Berg, V. Burg, A. Gombović, M. Puri, On the rise of fintechs: credit scoring using digital footprints, Rev. Financ. Stud. 33 (7) (2020) 2845–2897.
[95] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Found. Trends Theor. Comput.Sci. 9 (3–4) (2014) 211–407.

[96] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, D. Hassabis, Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.
[97] C. Berner, G. Brockman, B. Chan, V. Cheung, S. Zhang, Dota 2 with large scale deep reinforcement learning(2019).

Xueqi Cheng is a professor in the Institute of Computing Technology, Chinese Academy of Sciences, and the director of the CAS Key Laboratory of Network Data Science and Technology. His main research interests include network science, web search and data mining, big data processing and distributed computing architecture. He has published more than 200 publications in prestigious journals and conferences, including IEEE Transactions on Information Theory, IEEE Transactions on Knowledge and Data Engineering, Journal of Statistical Mechanics, Physical Review E., ACM SIGIR, WWW, ACM CIKM, WSDM, AAAI, IJCAI, ICDM, and so on.

Shenghua Liu is an Associate Professor at the Institute of Computing Technology, Chinese Academy of Sciences. He received his Ph.D. degree from the Computer Science and Technology Department, Tsinghua University. He once visited at the University of California, Los Angeles and at Carnegie Mellon University as a visiting scholar respectively. His current research interests are designing intelligent and automated algorithms for big data mining problems, related to big graphs and series. Two related publications have separately been recognized as “best paper” award and candidate.

Huawei Shen is a professor in the Institute of Computing Technology, Chinese Academy of Sciences. He received his PhD degree from the Institute of Computing Technology in 2010. His major research interests include network science, social media analytics and recommendation. He has published more than 80 papers in prestigious journals and top international conferences, including in Science, PNAS, Physical Review E, WWW, AAAI, IJCAI, SIGIR, CIKM, and WSDM. He is an Outstanding Member of the Association of Innovation Promotion for Youth of CAS. He received the Top 100 Doctoral Thesis Award of CAS in 2011 and the Grand Scholarship of the President of CAS in 2010.