JULY 30, 2024

Resilient Knowledge Graph Representations for Federated Financial Data

By Srivatsan Ravi (University of Southern California Computer Science and Information Sciences Institute), Larry Harris (University of Southern California Marshall), Ke-Thia Yao (University of Southern California Information Sciences Institute)

This paper presents a new data paradigm that can facilitate analyses of critical financial issues. Specifically, the paper examines how the widespread use of resilient data structures could enhance the efficiency and stability of financial markets by allowing regulators and market participants to understand and better identify systemic risks. The ability to obtain value from data depends on how easily the data can be accessed for their intended use. A knowledge graph organizes federated data that lends itself to understanding relationships among entities such as market participants, exchanges, or instruments. Compared to other data structures, such as flat files or relational databases, knowledge graphs are more extensible, have lower barriers to access, and are uniquely suited to identifying relations within networks for visualization and analysis. The research shows that knowledge graphs also can be made resilient to attacks by malicious actors and physical failures. The paper demonstrates through examples how knowledge graphs can be leveraged to derive resilient meta statistics that financial regulators can use to identify abnormal behaviors and unusual variations in financial database characteristics over time.

INTRODUCTION

The efficient operation of the financial system depends on reliable financial data and the ability of all participants—including regulators—to access the data they need to make decisions and form judgments. When regulators cannot access data in useful forms for evaluating systemic risks, they cannot recognize when the financial system is vulnerable to severe disruption, nor can they best evaluate policies to address weaknesses when extreme events occur. This brief presents a new data paradigm that can facilitate analyses of important financial issues.

Data security and robustness are paramount in finance because databases hold records of trades and the resulting financial positions in multitudes of assets, contracts, and cash flows. Ensuring that these

Note: This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

records are consistent among entities is crucial to the productive operation of the financial system. Without such assurance, people would lose confidence in the financial system and shy away from using it, thus disrupting the flow of the many economic benefits that financial markets provide Americans. Economic activity would be severely impaired.

Informational inconsistencies in the financial system can arise anytime, such as when buyers and sellers record different terms for their trades or when one side records a trade and the other fails to do so. Such problems usually resolve in a process called clearing. The problem is most severe in bilateral trading markets, where buyers and sellers negotiate directly—but it can also arise in multilateral exchange markets, where a central party (typically an exchange or broker) arranges trades for its clients. Although clearing such trades is trivial because one entity arranged them, buyers and sellers may not receive the trades' reports, due to potential data transmission problems.

Informational inconsistencies can also arise when:

- traders record the wrong terms for their trades (e.g., by recording 1,000 contracts instead of 100);

- information transmission systems fail to transmit information and the failure is not recovered (e.g., when lines of communication are disrupted); andself-regulatory organizations, such as the Financial Industry Regulatory Authority (FINRA); and

- malicious actors try to disrupt the financial system (e.g., by hacking into systems to destroy, modify, or prevent the delivery of records; inject false records; or engage in denial-of-service attacks).[1]

Informational inconsistencies can also arise when data are lost, inconsistently processed, or fraudulently modified or created within companies. Double-entry accounting systems can catch many of these problems—but such problems may still arise when both data entries are wrong, as when a single entity simultaneously enters both records. Reconciliation processes can also catch these problems, but such processes may fail when only one side of a transaction (e.g., the change in cash position) is reconciled. In contrast, the other side (e.g., a future payment that depends on a formula defined on a not-yet-known quantity) may not be determined until a future date. Errors may also arise when the formula is misstated or currently known quantities upon which the formula is based are misstated. Such misstatements could result from transcription errors, misunderstandings, or fraudulent activities. Finally, financial auditors can try to catch these problems by examining and testing data systems and control procedures.

Such problems can also occur when:

- people make transcription errors (e.g., when they record a price as 9 instead of 6);

- people misunderstand what they have done (e.g., when they do not recognize that they have arranged a bond transaction on a dirty versus clean basis);

- programmers and other people involved in data processing and manipulation make mistakes (e.g., when they make undetected coding errors or fail to code for economically significant contingencies);

- internal or external entities engage in fraud (e.g., by misallocating assets or modifying activity records to affect payments made to them or their confederates).

Overall, this paper examines how the widespread use of resilient data structures could enhance the efficiency and stability of the financial markets by allowing regulators and other market participants to better identify and understand systemic risks. Users' ability to obtain value from data depends on how easily they can access it for their intended purposes. One way to optimize data access is through the use of data structures.

One such data structure is the *knowledge graph* (KG), which organizes federated data in a way that lends itself to understanding relationships among entities

such as market participants, exchanges, and instruments. Compared with other data structures, such as flat files and relational databases, KGs are more extensible, have lower barriers to access, and are uniquely suited to identifying relationships within networks for visualization and analysis. Our work shows that KGs can also be made resilient to attacks by malicious actors and to physical failures. We give examples to demonstrate how KGs can be leveraged to derive resilient meta statistics that financial regulators can use to identify anomalous behaviors and unusual variation in financial database characteristics over time.

A SHORT ILLUSTRATIVE VIGNETTE

Suppose a malicious actor intercepts a process within a bank that transmits the terms of trades to the bank's archival systems. The actor changes terms whose implications for cash flows will not be discovered for many months. The attack continues for many weeks before the bank discovers it, at which time, the bank has no idea what its actual risk exposures are and how much money it owes or is due on thousands of contracts. Recovery will be costly because the bank must recover and process its primary records produced from the interception of the breach.

SYSTEMIC IMPLICATIONS

When information about the problem becomes well known, the bank's customers, counterparties, and regulators may lose confidence in its ability to accurately determine its financial condition. Entities may rush to close positions on which the bank owes money and may stop doing business with the bank. Both responses can lead to a liquidity crisis at the bank when nobody can reliably determine the bank's solvency.

Suppose the bank is large and systemically important. In that case, other banks will be concerned about which of their peers may be exposed to losses from contracts that the disrupted bank may be unable to honor during its liquidity crisis. Those other banks then will shy away from new contracts until the problems are resolved, resulting in a lockup of the financial system. A similar scenario occurred during the 2007–09 financial crisis, although for a different reason. At that time, banks were concerned about whether their swap contract counterparts would be able honor their commitments, due to uncertainty about the values of mortgage-backed bonds held or insured by those banks.

To some extent, banks mitigate these counterparty risks by requiring payments of variation margins, to remove risks associated with future contingencies as markets reveal information about those contingencies.[2] These payments help ensure that the banks quickly address informational inconsistencies that cause them to expect different cash flows. However, inconsistencies may still lead to long-term problems when information about future contingencies is not revealed. For example, contracts may depend on future binary events for which no markets exist to price their occurrences (so a variational margin scheme can be hard to define).

These concerns illustrate the importance of building and maintaining resilient data structures.

RESILIENT DATA STRUCTURES

The widespread use of resilient data structures—particularly KGs—could enhance the efficiency and stability of the financial markets by allowing regulators and other market participants to better identify and understand systemic risks.

The ability to obtain value from data depends on how easily users can access those data for their desired purposes. Therefore, good data structures have the following characteristics:

- They are *extensible*, meaning new types of information are easily added to them. This feature is particularly important in dynamic financial markets where participants regularly introduce new instruments.

- They are *efficiently accessible*, meaning their users know which data are available from them and can retrieve those data quickly and at a low cost. The large diversity of participants, instruments,

and information systems in the financial system complicates efficient data access.

- They facilitate the characterization of relationships and concepts that interest users. Systemic risks generally originate in the network of relationships among actors and instruments. For example, large concentrations, chokepoints, and centralized points of failure can contribute to financial instability.

- They are resilient to damage caused by system failures and attacks by malicious players. Financial stability critically depends on reliable data. If data were not trustworthy, markets would freeze up because market participants could not reliably assess risks they face, particularly counterparty risks.

Of the various data structures available, KGs are among the most easily extensible, are efficient to access, and naturally lend themselves to network analyses (Ilievski et al., 2020). Accordingly, KGs may be particularly valuable to financial practitioners and regulators—if they can be resilient.

In this, we provide examples of how the KG data structure can facilitate analyses of systemic vulnerabilities. We also explain how recently developed consensus protocols can make any KG resilient. In contrast, while system designers can apply consensus protocols to data structures like flat files and relational databases, doing so requires structural specifications specific to the contents of each database. Applying consensus protocols to any KG, regardless of its contents, makes KGs uniquely valuable.

KG tools can help regulators identify systemic risks and respond to potential threats to financial stability that may arise through network vulnerabilities stemming from central dependencies. KGs can also help regulators control money laundering and market manipulation. Finally, new easy-to-use extensible tools can help regulators detect irregularities and identify and fix corrupted data.[3] Since engineers can easily represent traditional flat files of financial transactions using KG methods, widespread adoption would be straightforward.

KNOWLEDGE GRAPHS FOR FEDERATED FINANCIAL DATA

WHAT ARE KNOWLEDGE GRAPHS (KGS)?

A KG is a data structure consisting of a list of triplets in which an edge connects two nodes. The *nodes* generally are objects of interest, while the *edges* represent the relationships between pairs of nodes. For example, if John weighs 150 pounds, a *triplet* represents this information in a KG as "John – 150 – Weight," where *John* and *Weight* are nodes and *150* (pounds) is their associated edge.

Analysts can easily convert flat-file tables to KG graphs by letting each row and column identifier be a node. The value in the table associated with a pair of row and column identifiers is their edge.

KGs can summarize information with greater flexibility and ease than flat files or a set of flat files linked in a relational database. Unlike flat files, KGs can link nodes without restriction or dependence on common keys to link the flat files of relational databases. In contrast, a flat file must be a two-dimensional (or more) table in which many indexed values may be missing, and relational databases must be structured with keys to link their various tables.

KGs provide a particularly efficient way of storing and analyzing information when the data are sparse, as is often the case when new fields are added to a database or a database is designed to represent diverse information not applicable to all observations. For example, customized swap contracts vary substantially in terms, and traders regularly create contracts with new terms. Accordingly, flat files that describe customized swap contract terms must have many fields, most of which do not apply most contracts, and new fields must be added whenever a contract depends on previously unused terms. Flat databases that would provide universal machine-readable descriptions of customized swap contracts would be large, sparse, and hard to manage.

The storage requirements and computational costs of analyzing uncompressed flat files depend only on the dimensions of the data structure and not on

how completely or sparsely the database is populated. Although compression can significantly reduce the storage costs of sparse flat files, data analyses of flat files usually require examining all points, populated or otherwise. In contrast, KGs only record known data. When the data are particularly sparse, KGs are much smaller than flat files, and the computational costs of tools used to analyze KGs do not depend on the sparsity of the data.

KGs facilitate understanding relationships among entities that may be market participants, exchanges, or instruments. Compared with other data structures, such as flat files or relational databases, KGs are extensible, have lower barriers to access, and are uniquely suited to identifying relationships within networks for visualization and analysis.

Finally, since the information embodied in a KG does not depend on where the data are located within a record (as it does in a flat file), unauthorized or accidental additions to or deletions from a record encoded in a KG are easy to identify if the changes are not consistently applied across all triplets related to the record. When the triplets are not stored physically adjacent to each other, unauthorized changes can be difficult to make surreptitiously.

In the following section, we explain how good design can make KGs resilient to attacks by malicious actors and physical failures. Standard implementations of distributed KGs are subject to the same lower bounds on data discrepancies that characterize all distributed database systems that do not employ robust mechanisms to bound these discrepancies.

BENEFITS OF EXTENSIBILITY IN FINANCE

The demand for derivative financial products to meet client-specific needs has produced a diverse and ever-changing universe of financial contracts. These bespoke contracts include swaps (as discussed previously), bonds, options, and commodities. Describing these financial contracts using a common data specification is difficult, due to the diversity of their features and the regular introduction of new contracts with complex and previously unseen features. Using KGs to describe such contracts is attractive because KGs permit users to specify self-defining features that other users can access. Accordingly, financial entities may increasingly use KGs to store important financial information. However, such use requires that the KGs be resilient to failures due to physical processes, programming problems, and malevolent actors, as we explain in the following sections.

HOW REGULATORS CAN USE KNOWLEDGE GRAPHS

Public- and private-sector financial regulators can apply various tools to KGs to surveil for irregularities, plus identify and fix corrupted data. These tools include spatial operators that characterize relationships among nodes and temporal diff operators that characterize differences in states. Perhaps most advantageously, regulators can apply these tools to any data stored in the KG structure, without regard to content. Also, analysts can easily create and implement new classes of spatial and temporal diff operators to address additional issues of interest to them.

Regulators using new robustness tools can confidently determine how vulnerable critical databases stored in the KG format are to data discrepancies. Moreover, using these tools, regulators can specify parameters to assure any level of reliability, subject to the associated cost.

A KNOWLEDGE GRAPH USE CASE FOR FINANCIAL DATA

NETWORKS

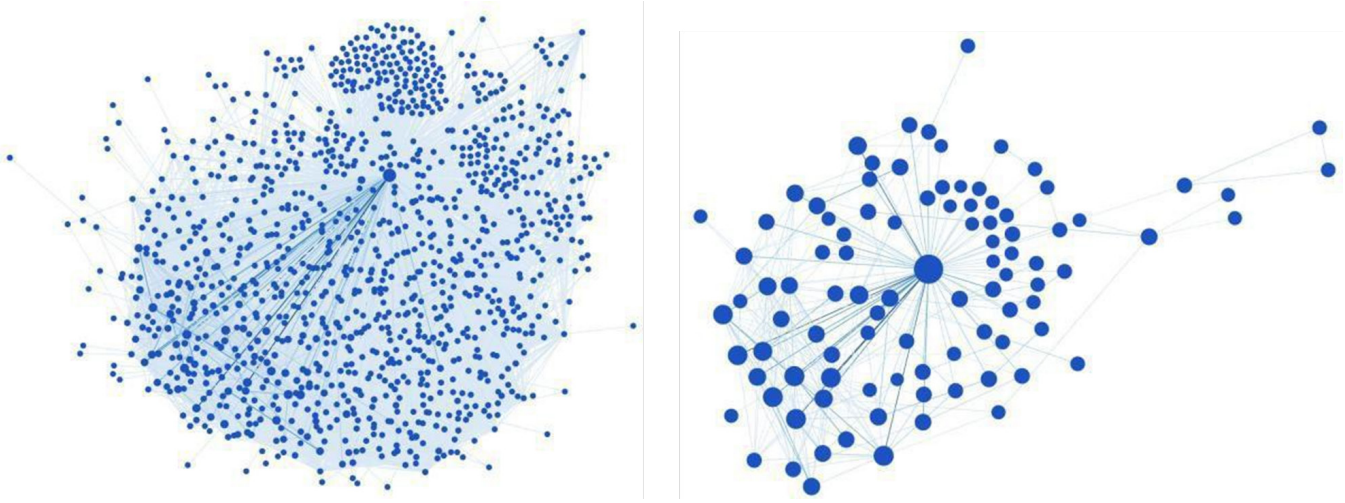
Systemic financial risks are due to various entities' dependence on other entities, such as traders, markets, and instruments. Interconnectedness is a source of systemic risk because failures involving highly connected central nodes can propagate through the network and significantly degrade performance and user confidence. Accordingly, characterizations of financial-systemic risk involve network analyses. KG representations of data are uniquely useful for identifying network characteristics such as central dependencies and local concentrations because KGs are defined on nodes where edges characterize connectedness.

We illustrate how KG analyses can help regulators better understand systemic risk using the Financial Industry Regulatory Authority (FINRA) Trade Reporting and Compliance Engine (TRACE) database. The TRACE data include records of all corporate bond trades dealers have reported to FINRA. The academic version of this database includes obfuscated dealer identities for all trades reported from 2002 through 2019, but it does not include customer identities. We translated these data from a flat file into a KG.

Our first analysis illustrates how traders relate to each other and all customers taken as a group (see **Figure 1**). The left chart presents the entire trading network observed in December 2011. The central node represents all customers. All other nodes represent individual traders. The size of the node*s* indicates the total dollar volume of each trader. Two nodes are connected if the two associated traders have traded with each other this month. The intensity of the line connecting them represents the aggregate trade volume between them; darker-blue colors indicate higher dollar volume. The right chart presents the trading network, focusing on the top 100 traders. Both charts show the market is not dependent on any central entity other than (obviously) all customers. Highly centralized trading exposes the market to greater systemic risk should one of the central dealers become incapacitated, so these charts help alleviate concerns about systemic risk in the corporate bond markets.

To characterize network centrality, we apply the PageRank network centrality measure. *PageRank*—the measure that Google uses to rank search results—characterizes the importance of an entity by how well connected it is with other entities. The PageRank of all traders sums to 1. The left chart of **Figure 2** shows that during December 2011, the aggregate customer accounted for 5.7% of PageRank. No dealer accounted for more than 3.1%. The right chart shows that during this period, the top 20 traders represented 24.6% and the top 100 traders represented 55.9%. These statistics indicate that trading is well distributed within this market, and they alleviate concerns about systemic risks due to the failure of a single trader. These results hold true for all other months.

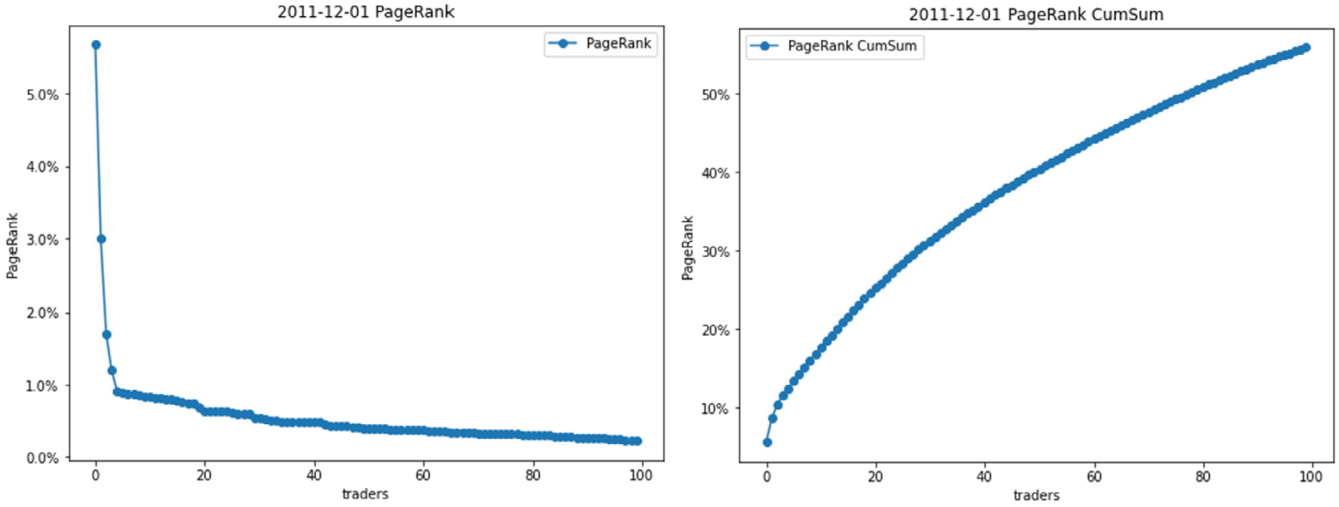
Figure 1. Corporate Bond Dealer Trading Network



Notes: The left chart represents all traders for December 2011. The largest node near the center represents all customers taken together. A small cluster of traders near the top deals only with the aggregate customer. A larger cluster of traders near the bottom-left deals with the customer and each other. The right chart presents a truncated version of the left chart that shows only the largest 100 traders and the edges for which their aggregated volume is above the 90th percentile. Again, the largest node represents all customers taken together. Two dealer clusters are apparent. A cluster of dealers that primarily deals only with customers is on the right. The second, larger cluster on the left deals with both customers and each other. Both charts show the market is not dependent on any central dealer or small collection of dealers.

Source: Authors' calculations from the Academic TRACE database

Figure 2. Dealer Concentration



Notes: The left chart plots the PageRank of traders against those ordered by their PageRank. The right chart plots the cumulative PageRank of the top traders.

Source: Authors' calculations from the Academic TRACE database

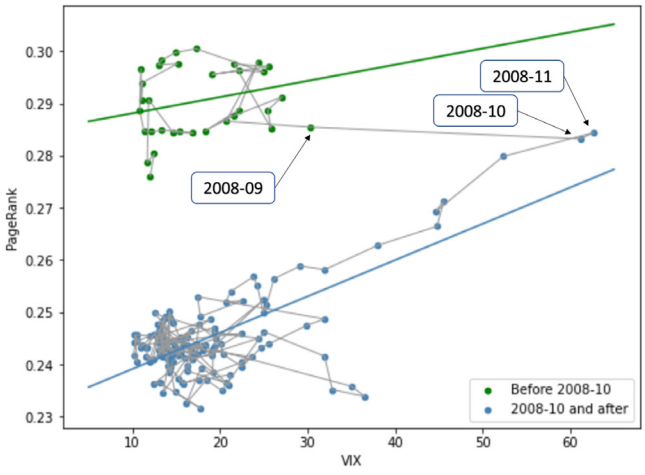
A HISTORICAL PERSPECTIVE: THE LEHMAN COLLAPSE

To illustrate how centrality, and thus systemic risk, changes over time, we computed the total page rank for the top 20 dealers for each month from 2005 through 2019, and we plotted it against the VIX index (see **Figure 3**).[4] VIX is a measure of equity volatility obtained from S&P 500 options price-implied volatilities. VIX rises when markets are stressed.

The graph shows a discontinuity in the results between September and October 2008. The data points clustering in the upper left are from before October 2008; all other points are from later months. The discontinuity likely was related to the collapse of Lehman Brothers, which occurred on September 15. VIX then jumped to its highest values in October and November. Immediately following the collapse, the total page rank did not change much, perhaps because Lehman was not a significant dealer in corporate bonds.[5]

Treating the two sample periods separately, the data show that centrality was higher during known periods of market uncertainty and, most obviously, during the 2007–09 financial crisis. That accounts for the largest VIX values on the right side of the graph. However, the decrease in centrality following the financial crisis may have been due to dealers diversifying their trading relationships over time, rather than the decrease in VIX. In addition, the increasing use of all-to-all trading platforms and the growth in proprietary firms dealing in corporate bonds (following the decrease in corporate bond

Figure 3. VIX vs. Top 20 Traders



Note: The chart depicts total page rank for the top 20 corporate bond dealers by month, scattered on average VIX for the month.

Source: Authors' calculations from the Academic TRACE database

dealing by the largest investment banks) contributed to the decreased cumulative page ranks of the largest dealers.

MAKING KNOWLEDGE GRAPH REPRESENTATIONS RESILIENT

The importance of financial data to our economy requires that users secure such data against loss. Accordingly, if entities employ KGs to store important data, systems must be available to ensure that entities can fully reconstruct KGs with confidence should local losses occur.

Entities responsible for settling trades and maintaining account balances usually employ multiple redundant systems to secure critical information and deliver that information when necessary. They use onsite and offsite information storage systems connected by independent and separated communications pathways. Also, they distribute their workloads to multiple redundant systems to help ensure that most of their work will continue unaffected should one system fail.

While most systems have regularly maintained data integrity, legitimate fears exist about their capacity to withstand increasingly sophisticated hacking attacks. Accordingly, employing new methods to further secure them and identify when they have been breached would increase confidence in the financial system and potentially prevent extraordinarily costly attacks. System designers can apply recent developments in algorithmic techniques for building resilient distributed systems to build resilient KGs.

For example, suppose we lose 10% of all triplets at random in a KG that multiple entities maintain. With properly designed consensus database systems, trusted participants can fully reconstruct the full KG with high probability, despite these losses. With such systems, recent research shows that more than 33% of KG triplets would have to be lost before the KG would lose the ability to produce consistent information when queried.

Here is a brief description of the steps necessary to construct this resilient consensus protocol (see Lembke, Ravi, Roman, and Eugster (2020) and Pires, Ravi, and Rodrigues (2018) for the theoretical underpinnings of this proposed approach):

1. A reliable broadcast mechanism must periodically distribute the state of the KG stored in each computing node to a large subset of other computing nodes that maintain redundant states of the KG.

2. To update the KG or retrieve the state of the KG via some query operator, the protocol proceeds as follows. Every node has one or more roles: proposer, acceptor, or learner. The proposers propose a new state change to the acceptors. The acceptors collectively decide whether a state change is valid. Once accepted, they share the change with learners, who use the data. Algorithmically, this process proceeds in two phases.

3. In the first phase, a proposer seeks permission from the majority of acceptors to assume a leadership role to organize a consensus of a specially defined majority of (Non malicious) acceptor nodes. The majority must be large enough to make the process secure from manipulation. Recent mathematical results provide a formula for determining the size of this majority.

4. In the second phase, the leader proposes a change and waits for acceptance from the majority of the acceptor nodes. In both phases, scenarios exist where a leader can lose its status (e.g., if another proposer is accepted as a leader or if an adequate number of acceptors do not respond quickly enough).

The two-phase protocol in steps 3 and 4 ensures that queries applied to the KGs stored at each node will produce equivalent results if no more than 33% of the data is corrupted.

This consensus protocol is like those at the heart of most blockchain protocols used by cryptocurrencies like Bitcoin (Malavota et al. (2017); Saad, Anwar, Ravi, and Mohaisen (2021); and Bitcoin (2023)) and

Ethereum (Ethereum, 2023). What is innovative about the KG framework is that it minimizes the number and size of states employed in the broadcast phase of the protocol when updates to the KG are distributed. Thus, the protocol makes the overall resilience mechanism feasible in practical federated deployments.

Most importantly, unlike with flat files and relational databases, transforming a centralized KG implementation to a distributed implementation for resiliency is particularly easy because system designers do not need to separately tailor the protocol for every application.

CONCLUSION

Risks to the financial-information infrastructure include physical and software failures and attacks by malevolent parties that introduce data inconsistencies. Controlling these risks generally involves minimizing potential single failure points, and controlling access to those single points is essential for operating the system. Redundant independent systems can reduce the number of single failure points at the potential cost of introducing new potential failure points into the systems required to coordinate the independent systems.

In this brief, we address a critical need for financial risk management: formalization of financial state and its associated (spatial and temporal) causal dependencies. We use a semantically rich data structure: the knowledge graph. Within this structure, we identify tools for analyzing inconsistency and indicators of erroneous activities in financial datasets that would otherwise be challenging. We also describe innovative methodologies for maintaining consistent trades and balances across federated accounts. Preserving information is essential to having the robust financial systems necessary for economic growth and prosperity.

REFERENCES

Bennett, Mike. 2013. "The financial industry business ontology: Best practice for big data." *Journal of Banking Regulation* 14, no. 3 (July): 255–268.

Bitcoin Project. 2023. https://bitcoin.org/en/.

Ethereum. 2023. https://ethereum.org/en/.

Ilievski, Filip, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Rongpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, and Pedro A. Szekely. *Kgtk: A toolkit for large knowledge graph manipulation and analysis*. International Semantic Web Conference, November 2020: Information Sciences Institute.

Lembke, James, Srivatsan Ravi, Pierre-Louis Roman, and Patrick Eugster. *Consistent and Secure Network Updates Made Practical*. In Silva, D. D., and R. Kapitza, eds. *Middleware '20: Proceedings of the 21st International Middleware Conference*, December 2020, Delft, The Netherlands: Association for Computing Machinery. 149–162. https://dl.acm.org/doi/proceedings/10.1145/3423211/.

Malavolta, Giulio, Pedro Moreno-Sanchez, Aniket Kate, Matteo Maffei, and Srivatsan Ravi. *Concurrency and Privacy with Payment-Channel Networks*. In Thuraisingham, B. M., D. Evans, T. Malkin, and D. Xu, eds. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, October–November 2017, Dallas, TX: Association for Computing Machinery. 455–471. https://dl.acm.org/doi/proceedings/10.1145/3133956/.

Pires, Miguel, Srivatsan Ravi, and Rodrigo Rodrigues. 2018. "Generalized Paxos Made Byzantine (and Less Complex). *Algorithms* 11, no. 9 (September): 141. https://doi.org/10.3390/a11090141/.

Saad, Muhammad, Afsah Anwar, Srivatsan Ravi, and David Mohaisen. *Revisiting Nakamoto Consensus in Asynchronous Networks: A Comprehensive Analysis of Bitcoin Safety and ChainQuality*. In Kim, Y., J. Kim, G. Vigna, and E. Shi, eds. *CCS '21: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, November 2021,

Virtual Event, Republic of Korea: Association for Computing Machinery. 988–1005. https://dl.acm.org/doi/10.1145/3460120.3484561/.

ENDNOTES

1 Malicious actors may also try to steal information from which they can profit. Although such activities are of great concern to practitioners and regulators, they do not affect data consistency and thus are beyond the scope of this study..

2 Variational margin payments transfer money from the losing side of a contract to its winning side when the value of the contract changes—typically, when the observed prices of instruments upon which the contract is based change. These payments effectively settle up the contract on a daily basis so that the final settlement of the contract does not involve a large payment that an insolvent counterparty might not be able to pay.

3 The tools we describe are available in an open-source KG library, KGTK (https://kgtk.readthedocs.io/en/latest/).

4 We did not go back to the 2002 beginning of the TRACE data because trade reports were notoriously error prone in the first few years.

5 The Academic TRACE dataset does not identify dealers by name. However, we identified a dealer that stopped reporting corporate bond trades in the Academic TRACE data around the Lehman collapse. This dealer had a page rank averaging only 0.6% in the days before the collapse.