Knowledge Graph-Based Modeling of Corporate Risk Factors for Financial Contagion Analysis

Warren Mendez
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
warren1977@ucf.edu

Kenneth Wunt
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
hellokenneth@ku.edu

Leon Thornton
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
thorntonleon@uab.edu

Abstract

Financial contagion research has traditionally focused on direct interbank exposures and market price co-movement, but the underlying corporate risk factor landscape remains underspecified. This paper develops a knowledge graph-based modeling framework for representing corporate risk factors extracted from regulatory filings, news, and structured firm attributes, and for analyzing how these factors propagate instability across economic networks. We discuss system-level design choices, including entity resolution, semantic normalization, relation extraction, and graph representation. The approach treats risk factors as first-class nodes linked to firms, sectors, geographies, and event types, allowing analysts to trace indirect transmission channels that are not visible in balance-sheet data. We examine structural trade-offs among coverage, precision, latency, and interpretability. We further consider governance, fairness, deployment, and regulatory acceptability. The paper argues that knowledge graph-based methods can complement established network models of contagion by enriching the causal and semantic fabric in which financial distress unfolds. It also identifies key infrastructure requirements for production deployment, including data lineage, auditability, and sustainability. Through this lens, the paper contributes to both financial stability research and the design of regulatory technology systems. Implications for future research and supervisory practice are highlighted.

Keywords

financial contagion, corporate risk factors, knowledge graphs, semantic modeling, systemic risk, regulatory infrastructure.

1. Introduction

The study of financial contagion has advanced substantially through network models that represent institutions as nodes and exposures or contractual relationships as edges. Early contributions established that the topology of interbank lending networks conditions the probability and severity of cascading defaults [1]. Subsequent work demonstrated that densely connected networks can absorb small shocks but may become fragile when a shock exceeds a critical threshold, revealing a non-monotonic relationship between connectivity and systemic stability [2]. Centrality-based indicators such as DebtRank have further shown that the systemic importance of a financial institution depends not only on its size but also on its position within a network of bilateral obligations [3]. These models, however, rely primarily on observable financial exposures and often leave the underlying corporate risk factor landscape implicit. This limitation is significant because contagion can propagate through shared exposures to similar business models, technological transitions, regulatory changes, or supply-chain vulnerabilities even when direct financial linkages are weak.

In parallel, knowledge graph technologies have matured as a means of representing heterogeneous entities and their semantic relationships at scale [4]. Knowledge graphs are now widely used for search, recommendation, and enterprise data integration, and have been extended to domain-specific analytical tasks in finance and beyond [5]. Corporate disclosure research has independently demonstrated that unstructured text in regulatory filings contains material information about risk. Dictionary-based and supervised approaches applied to 10-K filings show that changes in risk-related language correlate with future volatility and operational outcomes [6]. More recent methods infer dynamic risk themes and compare firms according to the similarity of their forward-looking statements [7]. Text-based network representations have also been used to identify competitive and product-market relationships from corporate filings [8]. These streams point toward an integrated modeling strategy in which risk factors extracted from text become nodes in a network alongside firms, sectors, countries, and events. Semantic anomaly detection applied to risk factor sections provides an additional mechanism for identifying firms whose disclosure structure deviates from historical patterns in ways that may signal emerging stress [9].

This paper presents a system-level perspective on knowledge graph-based modeling of corporate risk factors for financial contagion analysis. The discussion emphasizes architectural choices, structural trade-offs, governance, deployment, sustainability, and policy implications rather than narrow algorithmic description. The central argument is that a well-governed knowledge graph can serve as a shared operational substrate for financial supervisors, risk managers, and researchers. Such a substrate makes risk factor relationships explicit, supports attribution and audit, and enables the study of indirect contagion channels that remain hidden in conventional balance-sheet models. The remainder of the paper is organized as follows. Section 2 reviews conceptual foundations and related work. Section 3 describes the system architecture for constructing and maintaining the knowledge graph. Section 4 examines semantic modeling of corporate risk factors. Section 5 discusses contagion inference. Section 6 addresses governance, fairness, and policy implications. Section 7 considers deployment, sustainability, and operational robustness. Section 8 concludes.

2. Conceptual Foundations and Related Work

Contemporary systemic risk measurement has moved beyond pairwise exposure data toward connectedness derived from market prices and variance decompositions. Diebold and Yilmaz proposed network measures of financial firm connectedness based on forecast error variance decompositions, allowing dynamic spillover analysis [10]. This line of research demonstrates that systemic importance is not static and that the structure of connectedness shifts during periods of market stress. Network science offers additional conceptual tools. Scale-free network models explain how a small number of highly connected nodes can dominate connectivity and resilience [11]. Small-world topologies illustrate how short path lengths can accelerate the spread of disturbances across an otherwise clustered network [12]. These structural insights are important for risk factor graphs because corporate risk factors often exhibit skewed distributions and dense semantic clusters. A knowledge graph that faithfully represents these properties can reveal non-obvious pathways through which localized shocks become systemic.

Research on cascading failures provides a direct bridge between network topology and financial instability. Bipartite graph models of banks and assets show how overlapping portfolios produce cascades that are difficult to anticipate using only direct exposures [13]. Survey and synthesis work on financial network contagion emphasizes the distinction between default cascades, liquidity spirals, and information-driven contagion, and highlights the importance of network completeness for policy analysis [14]. Theoretical treatments of financial networks further show that the integration of multiple asset classes and debt instruments can generate rich patterns of contagion and diversification that depend on the full network of cross-holdings [15]. These results collectively motivate a representation that goes beyond pairwise exposure matrices. Corporate risk factors create common exposures across firms, such as dependence on a specific input market, sensitivity to interest rates, or vulnerability to cyber incidents. When many firms share a common risk factor node, shocks to that factor can propagate to all linked firms. The knowledge graph thus operationalizes the bipartite and multiplex structures that theoretical models have shown to be central to systemic risk.

Knowledge graph research has also produced a robust set of relational learning techniques suitable for large heterogeneous graphs. Relational machine learning methods enable prediction of missing links and node properties by exploiting graph structure [16]. Knowledge graph embedding approaches map entities and relations into continuous vector spaces, supporting similarity search and downstream predictive tasks [17]. Industry-scale knowledge graphs have demonstrated the importance of schema design, entity resolution, provenance tracking, and incremental updates [18]. These engineering lessons are directly relevant to financial risk factor modeling, where entity resolution across firms, subsidiaries, suppliers, and geographies is difficult and where provenance is essential for regulatory acceptance. The present paper builds on these foundations by emphasizing the system-level integration of risk factor extraction, graph construction, and contagion reasoning.

3. System Architecture for Knowledge Graph Construction

A production knowledge graph for corporate risk factors must integrate multiple data sources with different update cycles, formats, and reliability characteristics. The architecture typically includes ingestion pipelines for regulatory filings, earnings call transcripts, news, supply-chain registries, and structured market data. Ingested documents must be normalized, deduplicated, and linked to firm identifiers. Entity resolution is a foundational challenge because corporate names, subsidiaries, and suppliers appear in many different forms across sources. The system should maintain a mapping from each source-level identifier to a canonical graph entity. Without robust entity resolution, risk factor links are assigned to incorrect nodes and the resulting graph misrepresents the network. This architectural choice creates a trade-off between coverage and precision. Aggressive matching may inflate the graph with false relationships, while conservative matching may omit important linkages. A sustainable production system requires a feedback loop in which human analysts review a stratified sample of entity matches and errors are used to improve matching rules and models.

The relation extraction layer transforms unstructured text into typed relationships between risk factors, firms, products, geographies, and regulatory events. Relation types might include exposedto macroeconomic variables, dependence on input commodities, sensitivity to regulatory change, vulnerability to technology transitions, or operational exposure to cyber incidents. Each extracted relation is stored as a triple consisting of a subject entity, a typed predicate, and an object entity or attribute value. In addition to the triple itself, the system records the source document, the extraction model version, the confidence score assigned by the model, and the date on which the relation was asserted. This provenance layer is not an optional auxiliary feature. In a supervisory context, any relationship that informs a systemic risk assessment may need to be reconstructed and justified. The confidence score should be interpreted as a measure of extraction reliability rather than as a probabilistic claim about real-world exposure. Analysts can filter the graph by confidence, source type, and temporal validity to produce different views of the network.

The storage layer must support heterogeneous nodes, typed edges, and flexible schema evolution. Two broad choices are available. Property graph stores represent nodes and edges as records with key-value properties and are well suited to traversal-heavy workloads. Resource Description Framework stores provide a formal semantics layer that supports reasoning over class hierarchies and property domains. In practice, financial risk factor graphs often begin as property graphs for performance and ease of modeling and later incorporate lightweight ontology reasoning where needed. Graph database systems offer indexes for pattern matching and path queries, but they also introduce operational complexity. A production deployment must therefore coordinate multiple storage engines, a metadata catalog, a query interface, and a versioned schema. Data matching research [19] provides established methods for record linkage and entity resolution that reduce duplication across heterogeneous source systems, yet these methods require careful tuning when applied to global corporate identifiers and subsidiaries.

A key architectural trade-off is between real-time ingestion and analytical consistency. Real-time news feeds and market data require fast incremental updates, whereas quarterly regulatory filings permit slower, more careful curation. A common design separates streaming ingestion from a batch reconciliation layer. The streaming path updates the graph immediately but may introduce duplicates and provisional relations. The batch path re-processes complete filings and reconciles streams against authoritative records. This two-layer design improves robustness, but it also creates a period during which the graph contains provisional knowledge. Governance policies must specify how provisional knowledge is labeled and how downstream consumers are informed. The architecture should also support rollback and replay, so that changes to extraction models or entity resolution rules can be propagated without losing provenance.

4. Semantic Modeling of Corporate Risk Factors

Corporate risk factor language is characterized by variability, ambiguity, and forward-looking statements. A risk factor may be described with different terms in different filings, across industries, and over time. The knowledge graph model must therefore separate lexical surface forms from canonical concepts. This separation is achieved through a semantic layer that maps extracted phrases to normalized risk factor nodes. The semantic layer may use rule-based normalization, controlled vocabularies, or learned representations. Learned sentence representations [20] can capture semantic similarity between risk factor phrases that do not share exact terminology. However, purely distributed embeddings are difficult to audit, and they may conflate risks that are economically distinct but linguistically similar. A hybrid approach combines symbolic normalization with embedding-based similarity. In this design, analysts can inspect the graph at the symbolic level, while similarity scores guide clustering and retrieval. The semantic layer should also record when a mapping was made and with what confidence.

A central concern in semantic modeling is temporal validity. Risk factors change as regulations evolve, supply chains shift, or business models transform. The knowledge graph should represent risk factor relevance as a time-bound relation rather than as an unqualified assertion. For example, a firm may be linked to a currency risk factor for a specific fiscal year. The relation should carry a validity interval and the date of the filing from which it was derived. Without temporal annotation, the graph cannot distinguish current exposures from historical ones. This distinction is critical for contagion analysis because a shock propagating through a risk factor node should affect only those firms whose exposure to that factor is currently active. Temporal modeling also supports retrospective analysis in which supervisors can reconstruct the state of the graph at any point in the past.

Cross-lingual and cross-jurisdictional variation presents another semantic challenge. Multinational firms file disclosures in different languages and under different reporting regimes. The semantic layer must accommodate local legal concepts while maintaining global comparability. A single global taxonomy may be too rigid, whereas a fully decentralized vocabulary may prevent cross-border aggregation. A federated semantic architecture allows local extensions to a shared core vocabulary. Local regulators can define additional risk categories while mapping them to shared higher-level nodes. This approach balances local specificity with system-level comparability. It also mirrors the institutional structure of financial supervision, in which national authorities retain significant discretion but participate in international coordination. The knowledge graph can serve as a technical intermediary in this coordination process.

5. Contagion Inference on Risk Factor Graphs

Once the knowledge graph is constructed, contagion analysis proceeds by interpreting firms and risk factors as nodes in a bipartite or multiplex network. Direct firm-to-firm edges represent known contractual relationships, whereas firm-to-risk-factor edges represent semantic exposure extracted from disclosures. Projecting the bipartite firm-risk-factor graph onto the firm layer yields a similarity network in which two firms are linked if they share one or more risk factors. The strength of such a link can be weighted by the number of shared factors, the specificity of those factors, and the temporal overlap of exposures. Contagion can then be modeled as a propagation process in which distress in one firm increases the probability of distress among connected firms. Network models of financial contagion [21] show that cascade behavior depends on the degree distribution, correlation structure, and the presence of highly exposed intermediaries. The risk factor graph adds a layer of common exposure that may not be visible in direct exposure data.

A central analytical advantage of the knowledge graph is its ability to support path-based inference. A shock may travel from an originating firm to a shared risk factor and then to other firms linked to that factor. Longer paths may pass through geographic regions, commodities, or regulatory events. Path enumeration and shortest-path analysis can reveal indirect channels that are not apparent in pairwise exposure matrices. Community detection can identify clusters of firms that share dense sets of risk factors. Such clusters are potential amplification zones because a common shock may affect many members simultaneously. The graph also enables conditional reasoning in which an analyst restricts the network to firms with specific characteristics, such as high leverage or exposure to a given region, and then recomputes centrality and propagation indicators.

The contagion inference layer should distinguish between first-order exposure and higher-order feedback. First-order exposure captures the immediate set of firms linked to a risk factor. Higher-order feedback captures the possibility that distress among those firms generates additional stress through their counterparty relationships or through asset sales. Graph traversal algorithms can compute successive neighborhoods, but full feedback dynamics require simulation. System-level analysis must therefore couple the knowledge graph with quantitative financial models. The graph provides the structural scaffolding, while the simulation layer handles time-varying quantities such as capital buffers, liquidity, and market prices.

6. Governance, Fairness, and Policy Implications

The deployment of a knowledge graph for financial risk analysis raises significant governance questions. The graph is a sociotechnical artifact whose construction embodies choices about which data sources to include, which entities to resolve, and which relationships to extract. These choices can introduce systematic bias. Natural language processing models may perform differently across industries, firm sizes, and disclosure styles. Small firms with less standardized language may be underrepresented or misclassified, which could distort systemic risk rankings. Fairness in machine learning [22] requires attention to distributional differences and to the ways in which model errors are distributed across populations. In a financial stability context, unfairness may not be a matter of individual harm but of distorted resource allocation and supervisory attention. A firm that is systematically associated with spurious risk links may face unjustified capital requirements or reputational damage.

Explainability and auditability are essential for regulatory acceptance. The knowledge graph should expose every link with its provenance, including the source document, extraction model, confidence score, and human review status. This design supports a process of contestation in which firms can challenge erroneous links. Without such a process, automated risk assessment risks becoming opaque and unaccountable. Explainable artificial intelligence methods can provide local explanations for specific link predictions, but they must be integrated into a broader governance framework. Principles alone cannot guarantee ethical AI [23]. Effective governance requires institutional procedures, independent validation, and clear accountability for decisions based on the graph.

Policy implications extend beyond individual fairness to systemic stability. Regulators may use the graph to monitor emerging concentrations of risk before they materialize in market prices. This early warning function is valuable but also introduces new risks of overreach. A supervisor that observes a common risk factor across many firms must decide whether to intervene. The knowledge graph does not itself provide a policy threshold. It provides a structured evidence base that can support deliberation. The threshold for intervention remains a policy judgment involving trade-offs between stability, efficiency, and innovation. The graph can also support regulatory coordination by enabling different authorities to operate on a shared factual substrate while retaining their own analytical frameworks.

7. Deployment, Sustainability, and Operational Robustness

Production deployment of a financial risk factor knowledge graph requires continuous investment in data quality, model maintenance, and infrastructure. Machine learning systems accumulate technical debt [24]. Extraction models degrade as disclosure language evolves, entity distributions shift, and regulatory reporting formats change. Operational sustainability therefore requires a monitoring framework that tracks extraction quality, coverage, and latency. When a model no longer meets target performance, the system should trigger retraining with new labeled data. The cost of maintaining labeled data is significant. Supervisory institutions may pool resources through shared infrastructure, but data confidentiality requirements constrain such sharing. Federated learning and privacy-preserving analytics may offer partial solutions, yet they introduce additional technical complexity and governance challenges.

The knowledge graph must be integrated into existing supervisory workflows rather than deployed as a standalone analytics tool. This integration requires application programming interfaces, standard data formats, and clear versioning policies. Regulators and firms need to understand which version of the graph was used to produce a given assessment. Immutable versioning and time-stamped snapshots support reproducibility. Cloud-native architectures offer scalability and elasticity, but they also raise concerns about data sovereignty and vendor lock-in. Hybrid deployment models may be necessary to satisfy legal requirements while preserving analytical flexibility. RegTech developments have emphasized the reconceptualization of financial regulation through technology [25]. The knowledge graph model aligns with this trajectory by enabling regulators to process heterogeneous disclosures at scale while maintaining traceability.

Sustainability also has an institutional dimension. The value of a knowledge graph accumulates over time as historical data accumulate and link prediction improves. This creates a natural incentive for long-term institutional ownership. However, if the system is bespoke and poorly documented, it can become fragile when key personnel leave. Open standards, modular design, and periodic external audits mitigate this risk. The system should also be designed for graceful degradation. If an extraction pipeline fails or a data source becomes unavailable, downstream analyses should continue with clearly marked degraded inputs. Operational robustness and analytical honesty are thus closely linked. A system that quietly propagates missing data can produce dangerously misleading risk assessments.

8. Conclusion

This paper has presented a system-level framework for knowledge graph-based modeling of corporate risk factors in financial contagion analysis. The approach represents risk factors as first-class nodes linked to firms, sectors, geographies, and events, thereby making explicit the semantic exposures that are implicit in regulatory filings. We argued that this representation complements established network models of financial stability by exposing indirect contagion channels, supporting path-based analysis, and preserving provenance. The architectural discussion emphasized integration of heterogeneous data sources, entity resolution, semantic normalization, and dual-path ingestion. We examined the trade-offs between coverage and precision, real-time streaming and batch reconciliation, and symbolic auditability and embedding-based similarity. The governance analysis underscored the need for contestability, explainability, and institutional accountability. Deployment considerations highlighted sustainability, technical debt, and integration into supervisory workflows. Future research should empirically evaluate the predictive value of risk factor graph features relative to market-based and balance-sheet-based indicators. Further work is also needed on privacy-preserving collaboration, dynamic risk factor taxonomies, and the development of policy thresholds for graph-derived indicators. By treating corporate risk factors as a networked semantic infrastructure, the knowledge graph model offers a path toward more transparent and structurally informative financial stability analysis.

References

1. Allen, F., & Gale, D. (2000). Financial contagion. Journal of Political Economy, 108(1), 1–33.

2. Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. American Economic Review, 105(2), 564–608.

3. Battiston, S., Puliga, M., Kaushik, R., Tasca, P., & Caldarelli, G. (2012). DebtRank: Too central to fail? Scientific Reports, 2, 541.

4. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71.

5. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

6. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.

7. Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Journal of Accounting and Economics, 58(1), 1–33.

8. Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423–1465.

9. Sun, F., He, S., Wang, R., Ke, L., Shen, H., & Liao, Q. (2026). Modeling Structural Deviation in 10-K Risk Factors: A Semantic Anomaly Detection and Explainable AI Approach. Risks, 14(4), 87.

10. Diebold, F. X., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics, 182(1), 119–134.

11. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

12. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. Nature, 393(6684), 440–442.

13. Caccioli, F., Shrestha, M., Moore, C., & Farmer, J. D. (2014). Stability analysis of financial contagion due to overlapping portfolios. Journal of Banking & Finance, 46, 233–245.

14. Upper, C. (2011). Simulation methods to assess the danger of contagion in interbank markets. Journal of Financial Stability, 7(3), 111–125.

15. Eisenberg, L., & Noe, T. H. (2001). Systemic risk in financial systems. Management Science, 47(2), 236–249.

16. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11–33.

17. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724–2743.

18. Noy, N. F., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges. Communications of the ACM, 62(8), 36–43.

19. Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.

20. Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, 32, II-1188–II-1196.

21. Gai, P., & Kapadia, S. (2010). Contagion in financial networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 466(2120), 2401–2423.

22. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. MIT Press.

23. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

24. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28.

25. Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech, RegTech, and the reconceptualization of financial regulation. Northwestern Journal of International Law & Business, 37(3), 371–413.