Leveraging Data Science for Supply Chain & Risk Mitigation

Diego Velázquez, Erjia Meng, Luca Laport, Zhefu Qin, Zhe Chen
Group 11 Rel8ed, DSC 383W
Goergen Institute for Data Science
University of Rochester

1. Introduction

Global supply chains operate within a web of complex and often opaque relationships, exposing them to risks stemming from disruptions, inefficiencies, and compliance failures. Traditional supply chain analysis tools are often hindered by fragmented data and limited visibility, which constrains effective monitoring and decision-making. To address this challenge, our project leverages Rel8ed's global shipment database and AI capabilities to map supplier-consignee relationships at both country and company levels. This approach aims to illuminate hidden dependencies and provide actionable insights for operational resilience and risk mitigation in modern logistics networks.

The primary goal of this project is to construct a detailed map of global supply chain connections and uncover embedded patterns of systemic risk. By analyzing transaction-level shipment records, we aim to identify structurally critical actors, assess their exposure across three key risk indicators—Country Risk Assessment (CRA), Sector Risk Assessment (SRA), and Debtor Risk Assessment (DRA)—and recommend data-driven strategies to enhance security and diversification within international trade operations. Our methodology integrates clustering, network analysis, and predictive modeling. First, we performed data cleaning and country extraction using asynchronous API calls to standardize geographic information. We then applied K-Means clustering on over 63,000 supplier records to group them by product category, country of origin, and sector risk, identifying meaningful risk profiles across the global supplier landscape. This was followed by network analysis, where we constructed directed graphs at both the country and company levels, using centrality metrics and community detection to highlight dominant nodes, trade asymmetries, and structural roles. Finally, we layered CRA, SRA, and DRA attributes onto these networks to visualize and quantify multi-dimensional risk exposure, enabling more nuanced diagnostics of supply chain vulnerabilities.

2. Dataset

Our analysis is on a selected subset of the Rel8ed global shipments database, this includes external geographic information such as addresses, city, and countries to support both clustering and network analysis techniques. By focusing on this sample we ensure our findings reflect real world shipping patterns.

2.1 Dataset Description

The original "Global Shipments" dataset comprises over 30 million individual shipment records, specifically the dataset has shipment records, each recording details such as shipper and consignee identifiers, descriptive text, HS codes and risk metrics such as Country Risk Assessment, Sector Risk Assessment, Debtor Risk Assessment, and a Late-Payment Index. Our sponsor decided to let us analyze 100 thousand individual shipment records out of the 30 million.

In order to create a working dataset, we applied a multistage deduplication process: first retaining a single record per unique shipper name, then per consignee and shipper pair, and finally per unique description when both shipper and consignee matched existing entries. From this we then drew a

random sample of 100 thousand transactions, using a fixed random seed. The resulting dataset then contains 100 thousand rows and 57 columns including the original shipment fields plus 13 geographic attributes like city latitude, longitude, and population joined from an external cities reference.

2.2 Data Preprocessing

To prepare our data for analysis, we began with a raw dataset consisting of approximately 100,000 rows. Each row included 'Shipper' and 'Consignee' fields containing unstructured and inconsistent location data. Our goal was to extract the respective countries from these fields to support downstream analysis. To accomplish this, we leveraged the Deepseek API, which is designed to understand and interpret natural language effectively.

To manage the workload efficiently, our team divided the dataset, with each member handling approximately 20,000 rows. A custom asynchronous Python script using the aiohttp library was developed to automate the extraction process. This script allowed us to process data in parallel, making multiple concurrent API calls. The code sent data to the API in batches of 100 rows, reducing overhead and helping to avoid hitting API rate limits. We also implemented retry logic to handle network or API failures, ensuring robustness and reliability.

Initially, we had used a synchronous version of the script, which processed API calls sequentially without any batching or retry logic. This early version proved to be slow and prone to interruptions. Processing 20,000 rows using the synchronous approach took approximately 4–5 hours. In contrast, the optimized asynchronous version completed the same task in about 2 hours, effectively doubling the processing speed. For example, one team member successfully processed their assigned 60,000 to 80,000 rows within a 2-hour window using the asynchronous script.

This performance improvement underscores the importance of choosing the right tools and programming paradigms for large-scale data tasks. Asynchronous processing not only boosted efficiency but also provided resilience against API errors, enabling us to produce a clean, standardized dataset with clearly defined 'Shipper Country' and 'Consignee Country' fields.

3. Exploratory Data Analysis

This section presents key insights derived from an exploratory analysis of the 100K shipment dataset provided. The aim is to uncover patterns in shipping behavior, geographic flows, and categorical distribution, as well as to assess the reliability of location data used in downstream risk assessments.

3.1 Identifying Major Shippers and Consignees

The analysis began by identifying the most active companies based on their shipment frequency. According to Fig.1. On the shipper side, Orient Express Container Co. Ltd. leads with 510 recorded shipments, followed by Kuehne Nagel AG Co. KG (290 shipments), FR Meyer's Sohn GmbH (258), and HECNY Shipping Limited (247). These companies represent significant contributors to outbound logistics, indicating their role as central players in supply chain activity.

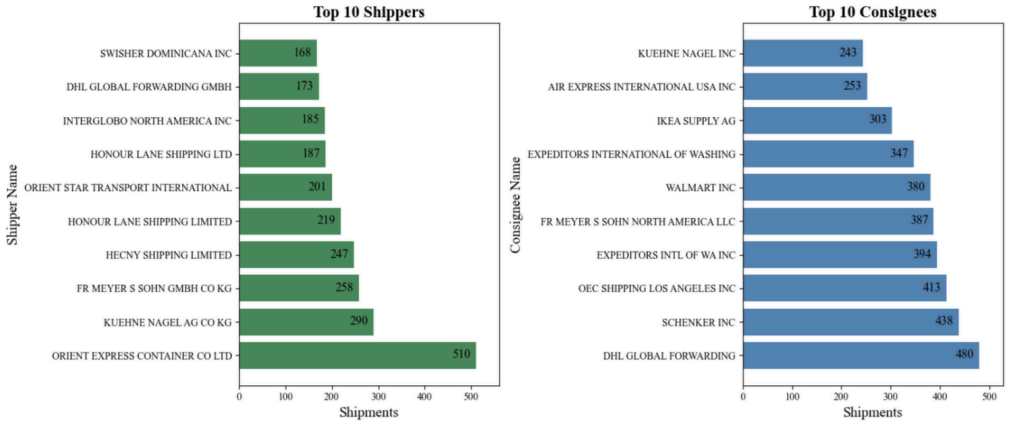

Fig.1 Top 10 Shippers & Consignees

On the receiving end, DHL Global Forwarding, while only ranked 10th among shippers, stands out as the top consignee with 480 shipments received. Other high-frequency consignees include Schenker Inc. (438 shipments), OEC Shipping Los Angeles Inc. (413 shipments), and Expeditors International, which appears multiple times due to name variations. When aggregated, Expeditors' entries represent over 700 shipments, reflecting the dominance of multinational logistics companies in global import operations.

3.2 Shipment Volume Trends Over Time

An analysis of shipment volumes over time reveals a notable peak in mid-2019, As shown in Fig.2, during which monthly shipments reached approximately 2.5 million. This period represents the highest level of recorded activity in the dataset. Notably, there is a complete absence of shipment records for the years 2020 and 2021. This gap is due to the structure of the dataset, which contains transactions exclusively from 2019 and 2022, with no data captured for the intervening period.

In 2022, shipment activity resumes but remains below the peak levels observed in 2019, suggesting a partial recovery in trade volumes. While this resurgence indicates some rebound in supply chain operations, it also reflects the continued impact of global disruptions on logistics flows during this time.

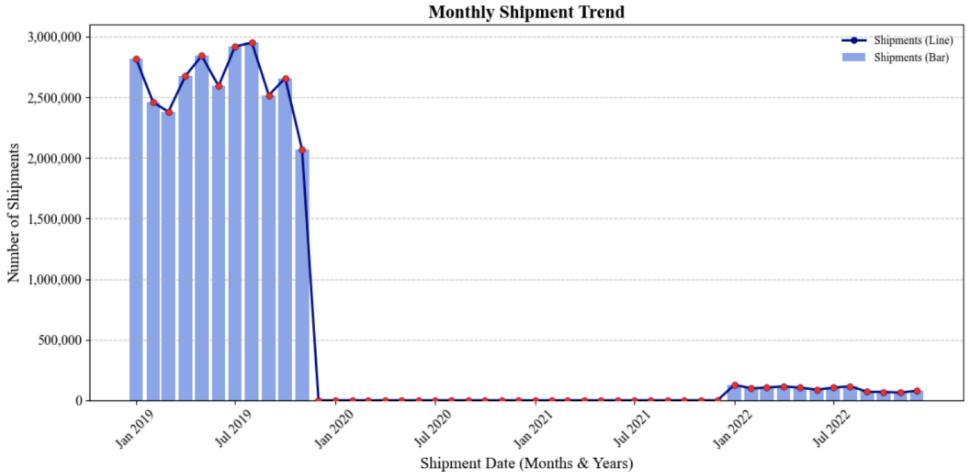

Fig.2 Monthly Shipment Trend

3.3 Common Shipment Categories

An examination of shipment content indicates a wide range of goods and services moving through the system. As shown in Fig.3, freight and transport services dominate the dataset, highlighting logistics providers' internal operations. In terms of physical goods, common categories include raw materials (e.g., wood, steel, aluminum), manufactured items such as furniture, and various industrial parts. This distribution reflects a robust business-to-business (B2B) logistics ecosystem, with movements covering both production inputs and consumer-ready products.

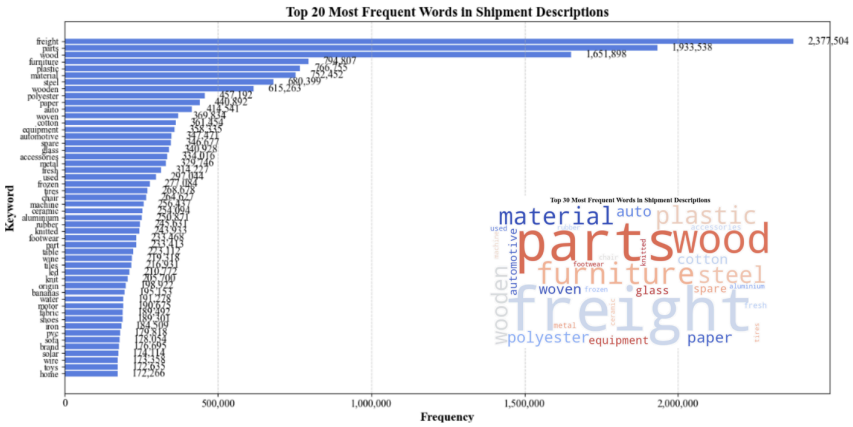

Fig.3 Top 20 Most Frequent Words in Shipment Descriptions

3.4 Geographic Distribution of Shipper & Consignee Countries

According to Fig.4, country-level analysis of shipping origins shows China as the leading shipper country, accounting for over 12,500 shipments, which reinforces its global manufacturing dominance. Other active Asian exporters include Hong Kong, Taiwan, India, and South Korea. In Europe, Germany and Italy emerge as primary origin countries. Interestingly, the United States ranks relatively low in terms of shipping volume—under 900 shipments—suggesting a more prominent role as an importer within this dataset.

On the destination side, the United States is by far the most frequent consignee, with over 25,000 inbound shipments, underscoring its status as a primary consumer and redistribution hub. Canada and Australia also record high per capita shipment volumes, suggesting strong economic reliance on international trade. In Europe, the United Kingdom, Germany, and the Netherlands stand out as key import destinations. Additionally, emerging markets such as India, Brazil, and Mexico show growing shipment volumes, indicating their expanding roles in global commerce.

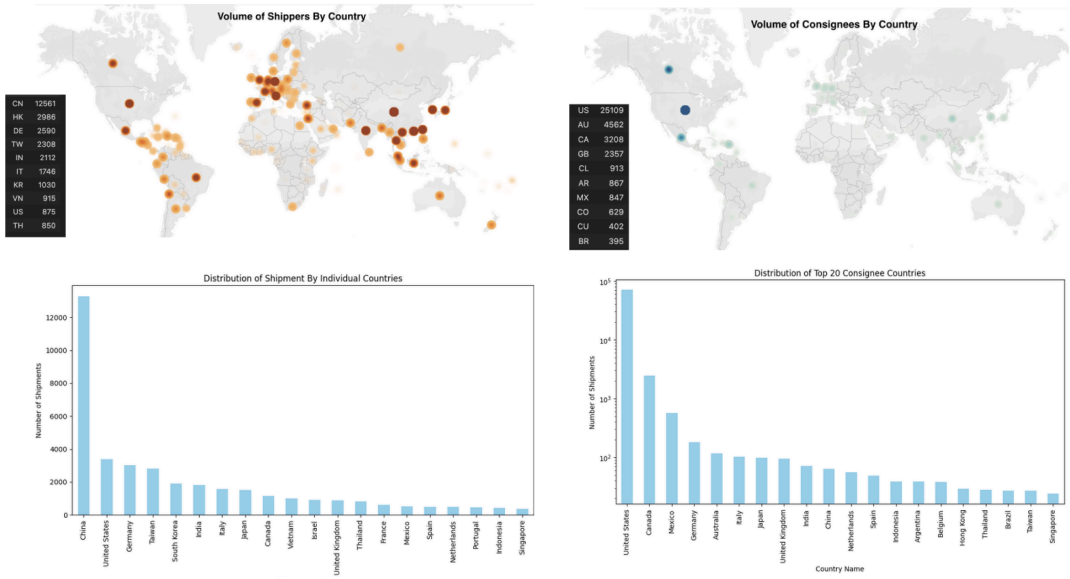

Fig.4 Volume & Distribution of Shippers & Consignees by Country

3.5 Manual Validation

To assess the accuracy of country assignments derived from business addresses, a manual validation exercise was conducted on 1,000 records—split evenly between shipper and consignee entries. The results in Fig.5 showed that approximately 5% of rows had incorrect country codes, while 8% contained missing country information, often recorded as 'UN'. The most common causes of error included ambiguous formatting and incomplete address fields. Although the DeepSeek API provided a reasonable level of automation, it struggled in cases lacking contextual clarity. These findings underscore the importance of manual validation as a critical step in ensuring data reliability. For downstream tasks such as geopolitical risk assessment or route optimization, even small location errors can lead to significant analytical distortions.

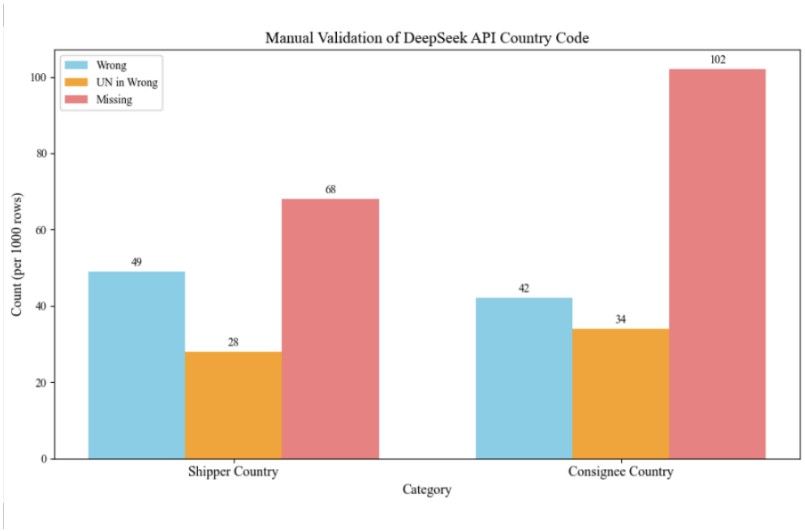

Fig.5 Manual Validation of DeepSeek API Country Code

3.6 Summary

The exploratory data analysis of the 100K shipment dataset revealed several key insights into global supply chain patterns. Major players such as Orient Express and DHL Global Forwarding emerged as top shippers and consignees, respectively, indicating the prominence of multinational logistics companies. Shipment volumes peaked in mid-2019 at around 2.5 million per month, with a data gap in 2020–2021 and partial recovery observed in 2022. The dataset also highlighted the dominance of logistics services and industrial goods in shipment categories, while geographic analysis confirmed China's role as the top exporter and the U.S. as the leading import destination. Finally, a manual validation of country assignments showed notable error rates—5% incorrect and 8% missing—emphasizing the continued need for human oversight in ensuring data accuracy for reliable downstream risk analysis.

4. Cluster Analysis

To better assess supplier vulnerability and segment global trade partners by risk, we implemented an unsupervised learning approach using K-Means clustering. This section outlines how we derived product categories, filled missing risk data, built the clustering model, and interpreted the resulting insights.

4.1 Data Preprocessing

4.1.1 Shipper Category

Our first step was to extract meaningful product-level data from each shipment's Harmonized System (HS) code. We isolated the first two digits of the HS code to classify suppliers into broad industry categories. This allowed us to assign each shipper a dominant product category, which later served as a key feature in clustering suppliers with similar business operations.

4.1.2 Risk Score (SRA)

The SRA (Sector Risk Assessment) score measures the macro-level risk associated with a supplier's sector and country. However, many suppliers in our dataset were missing this value. To impute missing SRAs, we developed a conservative, multi-tiered method. Country + Category average: If a supplier's country and product category had sufficient data, we used the average SRA for that combination. Nearest match: If not available, we identified the closest available (Country, Category) pair using frequency proximity. Conservative default: If multiple SRAs were found for a shipper (e.g., across different records), we used the highest score to avoid understating risk.

4.2 Clustering Model

We applied K-Means clustering with k=4, selected using the elbow method to balance interpretability and performance. The model used a feature matrix constructed from: Product category, country of origin, and SRA risk score. We standardized all features and encoded categorical values to ensure numerical compatibility. The model was trained on over 63,000 global supplier records, resulting in four well-separated clusters, each with distinct risk and sourcing profiles (Fig. 6)..

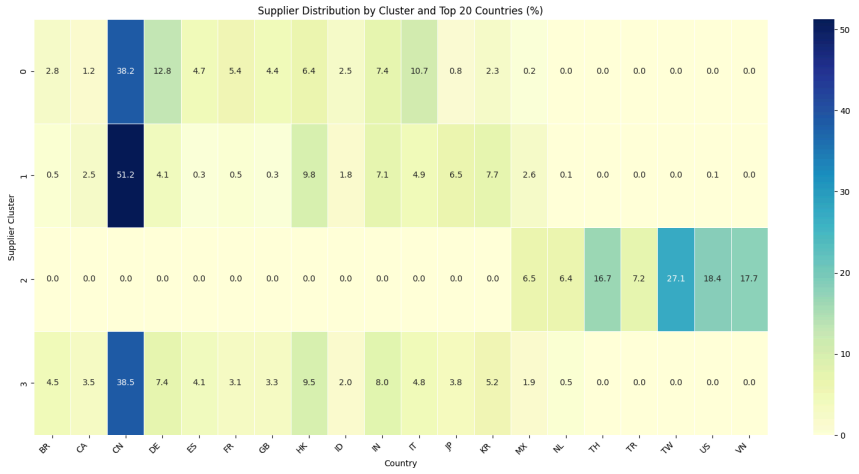

Fig. 6 Supplier Distribution by Cluster and Top 20 Countries

4.3 Cluster Profiles and Key Insights

These cluster insights provide a valuable foundation for identifying overexposed consignees, especially those heavily tied to high-risk manufacturing sources. They also allow for more nuanced supplier segmentation, enabling smarter procurement strategies and resilience planning in a globally volatile trade environment.

4.3.1 Cluster 0

Cluster 0 emerged as the highest-risk group, containing 19,315 suppliers with an average SRA score of 3.05, the highest across all clusters. This segment is dominated by heavy industrial and manufacturing sectors, such as machinery, furniture, plastics, vehicles, and iron/steel. The supplier base is geographically concentrated, with China alone accounting for 38.2% of all suppliers, followed by Germany (12.8%), India (7.4%), and Hong Kong (10.7%). This strong Asia–Europe sourcing pattern suggests significant geopolitical and operational risk exposure. Critically, we identified multiple consignees who sourced 100% of their suppliers from this high-risk cluster, even when they relied on more than one supplier (Fig. 7 and 8). This heavy dependence on Cluster 0 presents a substantial vulnerability to supply chain disruptions and reinforces the need for diversification strategies.

Top 10 Consignees Most Exposed to High-Risk Cluster 0		
Consignee ID	Cluster 0 %	Total Suppliers
2773803759	100	1
2773798673	100	1
2773787151	100	1
2773787145	100	1
2773752152	100	1
9417689	100	1
2773734593	100	1
2773730150	100	1
2773727636	100	1
2773725036	100	1

Fig. 7 Top 10 consignees entirely reliant on a single supplier from the highest-risk cluster

Consignees with >50% of suppliers from High-Risk Cluster 0		
Consignee ID	Cluster 0 %	Total Suppliers
23583405	100	2
2773763584	100	4
29975131	100	2
2663326607	100	2
2663121994	100	2
30712353	100	3
31873997	100	2
2654443670	100	2
2593121005	100	2
2773199017	100	3

Fig. 8 List of consignees where over 50% of their suppliers belong to the high-risk Cluster 0

4.3.2 Cluster 1

Cluster 1, by contrast, is the lowest-risk group with 15,574 suppliers and an average SRA of 1.93. While it is even more China-dependent than Cluster 0, with 51.2% of suppliers located in China, the industries represented in this cluster are generally more stable. These include electrical machinery, textiles, plastics, and furniture, all of which tend to face fewer regulatory or operational disruptions. The cluster also includes suppliers from Hong Kong, South Korea, India, and Japan. Despite the geographic concentration, the low sector risk scores suggest that these suppliers are relatively safe and reliable, making Cluster 1 a potential target for de-risking without drastically altering existing sourcing patterns.

4.3.3 Cluster 2

Cluster 2 is characterized by geographic diversification and moderate sector risk, with 13,878 suppliers and an average SRA of 2.79. The top supplier countries include Taiwan (27.1%), the United States (18.4%), Vietnam (17.7%), and Thailand (16.7%). This spread indicates a strong presence of exporters from Southeast Asia and North America. The industries involved, primarily machinery, electrical machinery, textiles, and plastics, are balanced in nature, offering a good middle ground between operational stability and market accessibility. As such, Cluster 2 represents a strategic opportunity for consignees seeking to mitigate risk while maintaining supplier diversity.

4.3.4 Cluster 3

Cluster 3 consists of 15,122 suppliers with an average SRA of 2.78, reflecting moderate risk. This cluster includes suppliers primarily involved in the automotive, mobility, and material sectors, including industries like plastics, vehicles, wood, and electrical machinery. The geographic concentration is again heavily tilted toward China (38.5%), followed by Hong Kong (9.5%), India (8.0%), and Germany (7.4%). While the overall sector risk is not as high as Cluster 0, the narrower industry profile and regional dependence could pose challenges, particularly for consignees aiming to reduce exposure to specific markets or sectors.

5. Network Analysis

5.1 Shipper-Consignee Country Network

We modeled global trade as both an undirected and a directed network built from shipper–consignee country pairs weighted by shipment volume. We began by extracting 215 country nodes and their shipper–consignee relationships from our shipment database, resulting in 4,580 directed edges. In constructing the undirected network, each pair of countries A–B was aggregated into a single edge

whose weight equaled the sum of shipments in both directions (computed via a pandas `groupby` and `sum`), yielding an average edge weight of approximately 12,300 metric tons. In parallel, we built the directed network by preserving the original shipper→consignee orientation and shipment volume on each edge, allowing us to distinguish import versus export roles. By comparing undirected and directed analyses, we capture both aggregate connectivity and the asymmetries of import/export flows. Core hubs such as the United States maintain dominance across all metrics and representations, while regional and broker roles become apparent only when directionality and community context are considered. This multi-metric, community-aware framework thus provides a comprehensive, multi-scale map of global trade roles, offering actionable insights for resilience planning and risk mitigation in international supply-chain networks.

5.1.1 Centrality Measurements

For both network representations, we computed four weighted centrality measures using NetworkX. Degree centrality (or in- and out-degree for the directed graph) was normalized by the maximum possible partner count. Betweenness centrality was calculated via Brandes' algorithm, measuring the fraction of all weighted shortest paths traversing each node. Closeness centrality was defined as the inverse of the average weighted shortest-path distance to all other nodes, capturing reachability. Finally, we applied a directed, weighted PageRank (damping factor = 0.85) to quantify each country's influence through its most important trading partners.

To uncover the meso-scale structure, we ran the Louvain community detection algorithm (via the `python-louvain` package) on the undirected graph. This partitioned the network into four geographically coherent modules (modularity Q = 0.48):

1. North America & East Asia (US, CA, MX, CN, TW),
2. Western Europe (DE, FR, GB, ES, IT, NL),
3. Southeast Asia & Oceania (AU, SG, MY, TH, ID),
4. Latin America, Middle East & Africa (BR, AR, SA, AE, NG, ZA).

   Within each community, we observed clear hierarchies—for example, Germany leads Western Europe in PageRank (≈ 0.022) and closeness (≈ 0.505), while Brazil tops Brazil–Middle East–Africa with PageRank ≈ 0.014.

Finally, we focused on the top 50 countries by undirected PageRank to classify their structural roles via hierarchical clustering (Ward linkage) on their four-dimensional centrality profiles. Three role categories emerged:

- Global Core Hubs (e.g., US, CA, MX) scored above 0.40 on degree, closeness, and betweenness, and exceeded 0.04 in PageRank.
- Regional Connectors (e.g., GB, KR, IN) maintained moderate degree (0.25–0.35) and closeness (0.50–0.60) but showed lower betweenness (0.02–0.03) and PageRank (0.01–0.02), reflecting strong intra-bloc integration.
- Peripheral Players (e.g., DK, GR, EC) registered below 0.15 in degree and below 0.005 in PageRank, marking them as marginal participants.

5.1.2 Visualizations

In ig.9, each node represents a country, colored by role: blue for shippers, orange for consignees. Edge thickness reflects the number of shipments, with directional arrows indicating export flow. The United States and China dominate as central nodes, with China acting primarily as an exporter and the U.S. as the primary importer. A small number of countries, including Mexico, Vietnam, and Taiwan, serve as significant intermediaries, while most others have sparse or asymmetric connections.

This network graph reveals the asymmetric nature of global trade. The U.S. dominates as a top importer, while China leads in exports, reinforcing its role as a manufacturing hub. Secondary players like Vietnam, Taiwan, and Germany show strong, often balanced flows. Mexico and Canada appear as key North American intermediaries.

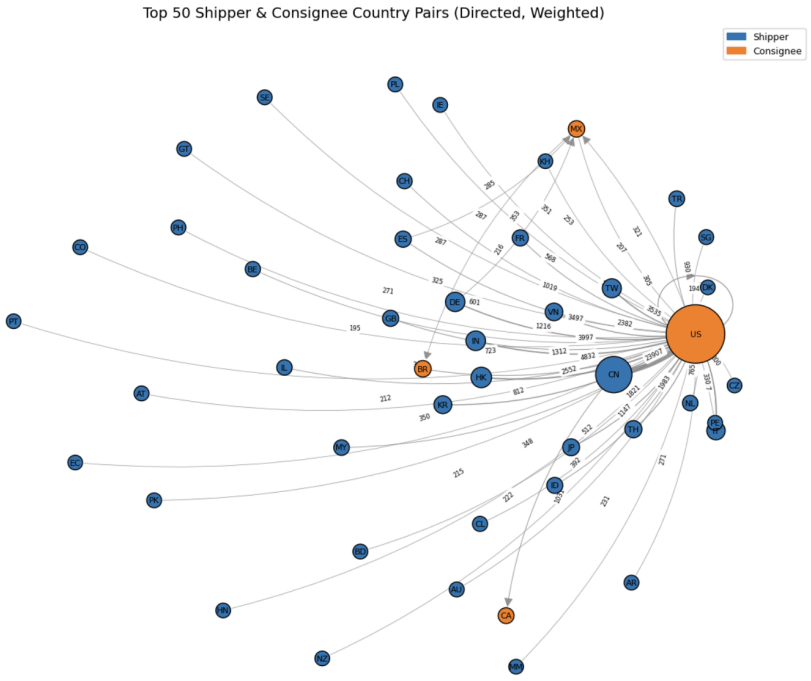

Fig.9 Top 50 shipper-consignee directed and weighted network graph

Panel 1a ranks countries by degree centrality, with the United States (0.924), Canada (0.589), and Mexico (0.563) emerging as the most connected hubs (See Fig.10). Panel 1b orders the same nodes by undirected PageRank; here, China's placement (0.048) surpasses Mexico's (0.046), indicating that its links to other highly central countries enhance its overall influence.

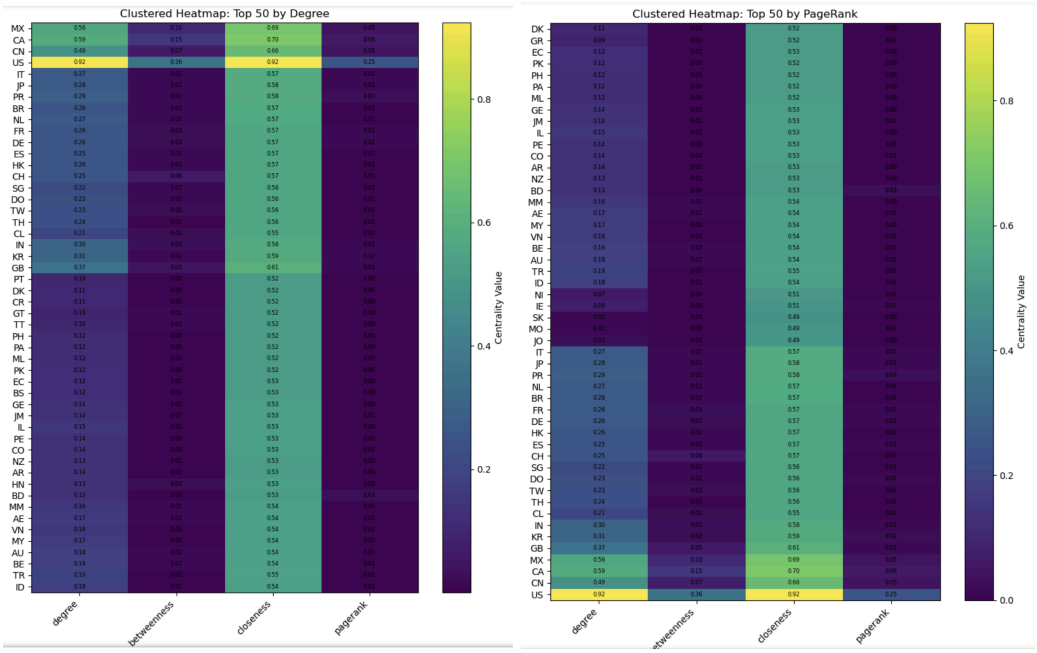

Fig.10 bar-chart comparisons of undirected centralities at country level

Fig.11 contrasts directed import and export roles. Panel 2a displays in-degree centrality, which identifies the United States (0.898), Canada (0.431), and China (0.203) as principal importers. In Panel 2b, out-degree centrality places China (0.467), Canada (0.452), and the United States (0.452) at the forefront of export activity. These asymmetries reflect China's production orientation and the United States' dual role as both a significant consumer and exporter.

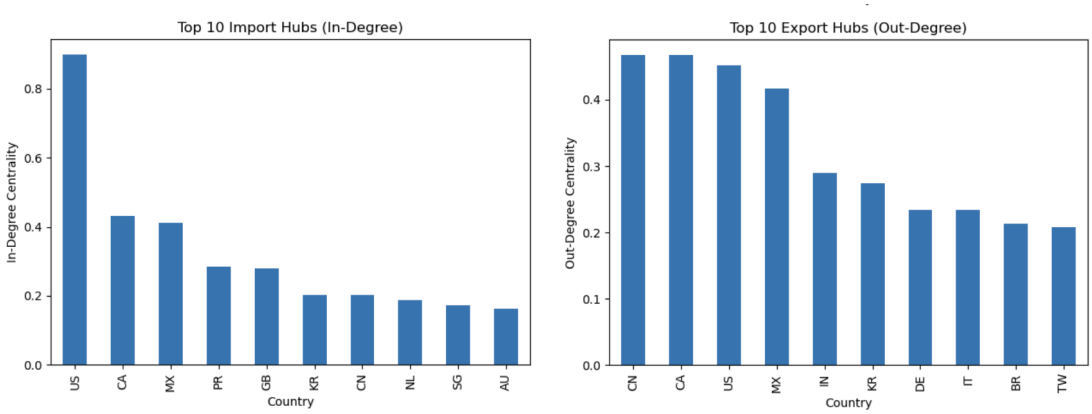

Fig.11 Top 10 country in-degree and out-degree comparison

Fig.12 employs bubble scatterplots to map trade footprints and flow roles. In Figure 3a, each point's coordinates (in- vs. out-degree) and bubble area (PageRank) reveal that North American and East Asian countries cluster in the high-import/high-export quadrant, whereas resource-export specialists and consumption-oriented economies occupy distinct off-diagonal regions. Figure 3b plots betweenness against closeness centrality, identifying the United States (betweenness = 0.255; closeness = 0.895) as the dominant broker and reach hub; Canada and Mexico form a secondary cluster, indicating strong reachability with slightly lower brokerage.

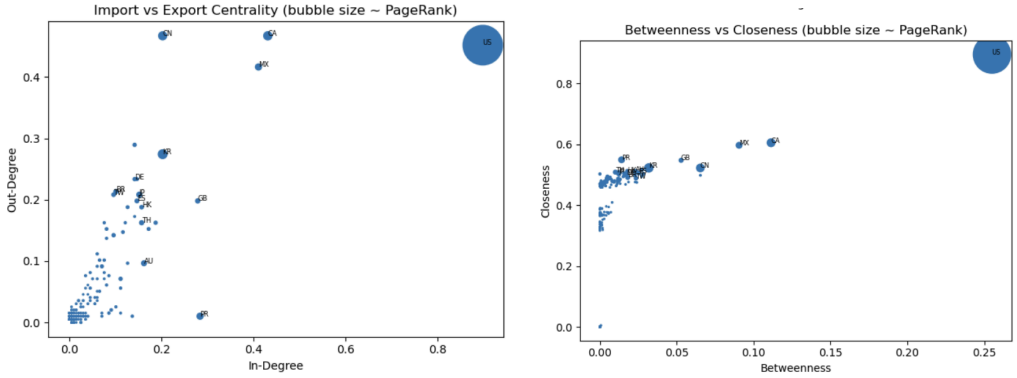

Fig.12 Scatterplots of country-level trade centrality metrics

Fig.13 layers Louvain community assignments onto the directed scatterplots. Community 0 (North America & East Asia) countries concentrate in the high-import/high-export region, Community 1 (Western Europe) occupies the moderate-import/moderate-export zone, and Communities 2–3 (Oceania; Latin America, Middle East & Africa) disperse into asymmetrical quadrants. A parallel betweenness–closeness plot demonstrates which communities serve as inter-bloc connectors versus internally cohesive modules.

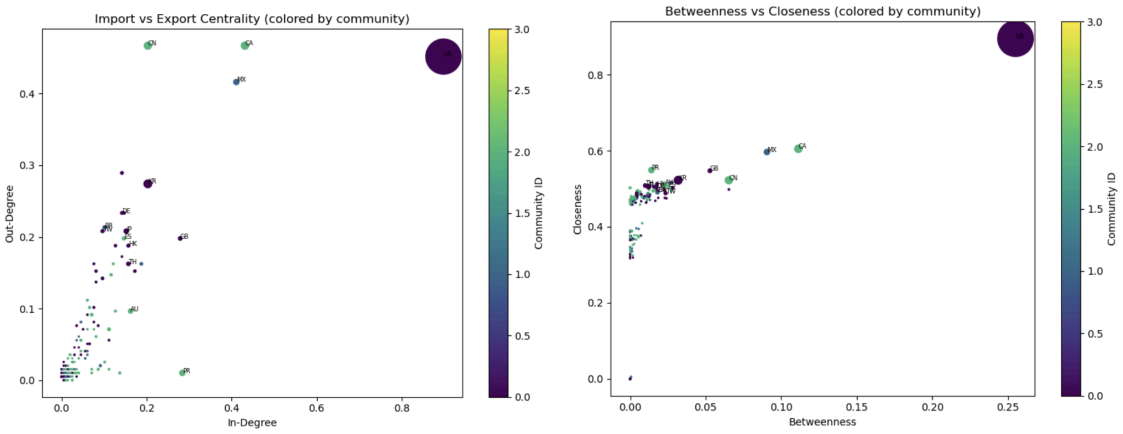

Fig.13 Trade centrality with Louvain community assignment

Fig.14 presents the correlation matrix of all five centrality measures. In‑degree and betweenness exhibit a strong positive correlation ($\rho \approx 0.92$), as do betweenness and PageRank ($\rho \approx 0.82$), whereas closeness shows a weaker association with PageRank ($\rho \approx 0.23$), underscoring the distinction between mere reachability and network influence.

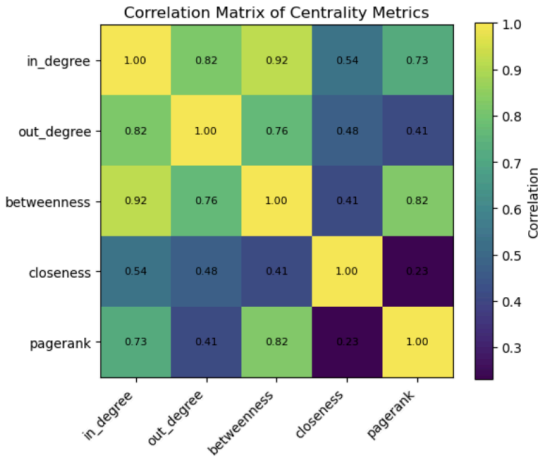

Fig.14 Correlation matrix of centrality metrics

Collectively, these analyses substantiate that the United States, Canada, and Mexico form the systemic backbone of global trade, dominating connectivity, brokerage, reachability, and influence in both undirected and directed contexts. China's elevated export centrality further emphasizes its production role. Community overviews and hierarchical role classification reveal the meso‑ and micro‑scale positions of regional blocs, specialized brokers, and peripheral economies. This comprehensive, multi‑metric framework offers a robust basis for resilience planning, targeted risk mitigation, and policy design in international supply‑chain networks.

5.2 Company-Level Network Analysis

To construct the company-level trade network, we extracted company-to-company shipment frequencies by grouping `shipper_name` and `consignee_name` from our full dataset,

counting each pair's occurrences. We then constructed both directed and undirected graphs from this dataset to capture the roles of companies in the global trade network.

5.2.1 Centrality Measurements

In the same vein, we calculated four core centrality metrics: degree, betweenness, closeness, and PageRank. Similarly, these metrics were computed using NetworkX, leveraging edge weights derived from transaction frequency. Degree centrality highlighted companies with the most direct connections; in the directed graph, this distinction allowed us to separate major importers from exporters. Betweenness centrality measures brokerage potential—identifying companies that act as intermediaries between otherwise disconnected trading clusters. Closeness centrality indicated how efficiently a company could reach all others in the network. PageRank, incorporating both edge direction and weight, captured overall influence, particularly favoring companies connected to other powerful entities.

To detect higher-order structure, we applied the Louvain method on the undirected graph, identifying hundreds of communities. These reflected clusters of retailers, logistics providers, or regional players. Among the top 50 most influential companies (by PageRank), we observed recurring patterns through hierarchical clustering: core hubs like IKEA Supply AG and Walmart Inc. stood out for their connectivity and influence, while others played more specialized regional or logistical roles.

We then focused on the 50 most central companies by undirected PageRank, clustering them via hierarchical Ward linkage based on their four-dimensional centrality vectors. This classification revealed three broad structural roles:

- Global Core Hubs (e.g., IKEA Supply AG, Walmart Stores Inc.) showed uniformly high degree, closeness, and PageRank, signifying central placement and broad integration.
- Strategic Brokers (e.g., Honour Lane Shipping Limited, Interglobo North America Inc.) had lower degree but elevated betweenness, indicating importance in bridging disconnected sub-networks.
- Peripheral Specialists (e.g., Jo Sung Sea & Air Co., Liberty Procurement Co) had lower centrality across all metrics, suggesting localized or niche participation within the global trade network.

This company-level network analysis uncovers the underlying topology of global trade relationships, allowing us to pinpoint structurally central companies, identify chokepoints in supply chains, and expose interdependent clusters. These insights offer a foundation for evaluating systemic risk and developing more resilient, data-informed logistics strategies.

5.1.2 Visualization & Interpretation

Fig.15 illustrates the top 50 shipper–consignee company pairs, emphasizing the directional structure of company-level trade. Blue nodes represent net shippers, orange nodes net consignees, and gray nodes companies with balanced flows. Key logistics providers like DHL and Interglobo act as central shippers, while companies like Amazon Logistics and Royal Caribbean emerge as major consignees. Thick edges reflect high-volume relationships, such as between Beijing Kang Jie Kong and Transworld Shipping. The structure highlights how trade flows are shaped by specialized roles—logistics companies drive distribution, while consignees anchor consumption.

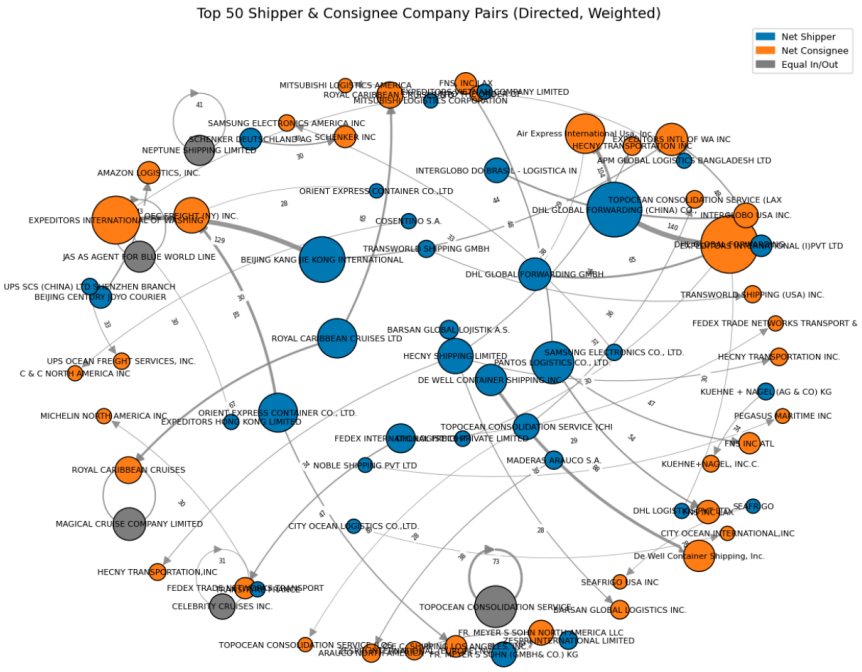

Fig.15 top 50 Shipper-consignee company network graph

Fig.16 shows the top 10 companies by degree centrality and PageRank. IKEA Supply AG leads both, reflecting wide reach and central importance. Honour Lane Shipping Limited ranks high in degree but lower in PageRank, indicating broad connections but less influence. Walmart Stores Inc. appears prominently across both metrics, showing strong structural positioning. The interplay of companies from both retail and logistics sectors points to a hybrid control model in international trade networks.

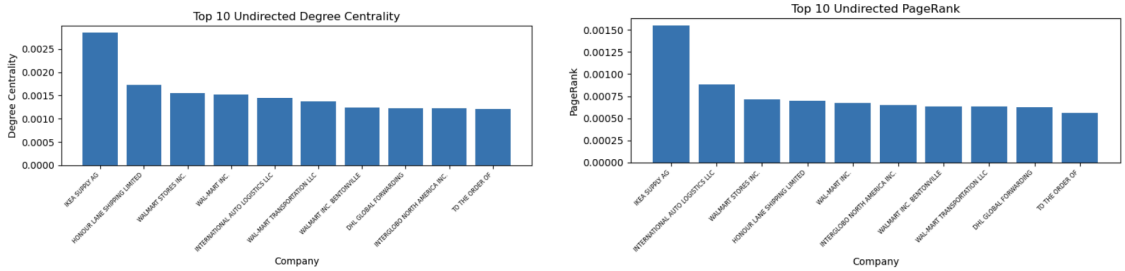

Fig.16 Top 10 undirected degree centrality and PageRank

Fig.17 breaks down degree centrality into in-degree (importers) and out-degree (exporters), highlighting distinct roles. IKEA Supply AG and Walmart lead as major import hubs, while Honour Lane Shipping and International Auto Logistics dominate export activity. The contrast reflects a network split between retail receivers and logistics dispatchers, with companies like DHL showing flexible but asymmetric positioning.

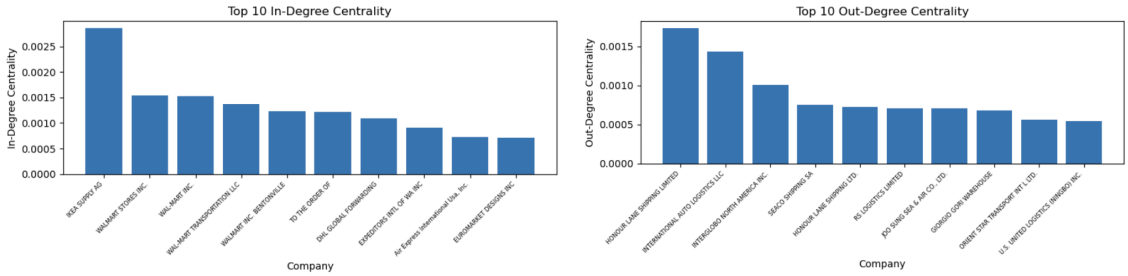

Fig.17 Top 10 In-Degree and Out-Degree Centrality

Figure 18 uses scatterplots to map company-level roles. Panel 18a shows in- vs. out-degree with PageRank as bubble size, revealing that most companies cluster near the origin, with only a few—like IKEA—exhibiting balanced, high centrality. Logistics companies generally skew toward the out-degree axis. Panel 18b, plotting betweenness vs. closeness, highlights the network's core-periphery structure: few companies are brokers or highly accessible, while most remain marginal.

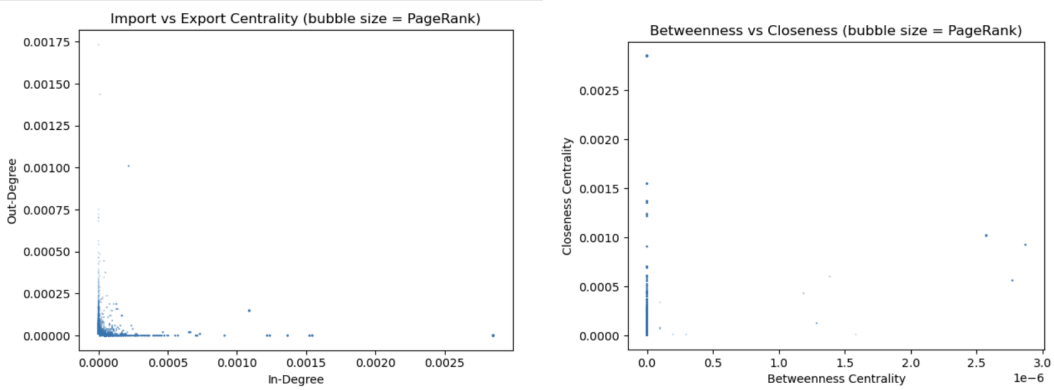

Fig.18 Company-level in-/out-degree and betweenness/closeness scatterplots

Figure 19 overlays Louvain communities onto the scatterplots, revealing that companies in the same module often share similar trade roles. In the import-export plot, certain communities cluster near the origin, indicating peripheral actors, while others occupy more central zones with balanced flows. The betweenness–closeness distribution is more diffuse, suggesting that influence and reachability are less tightly bound to community structure.

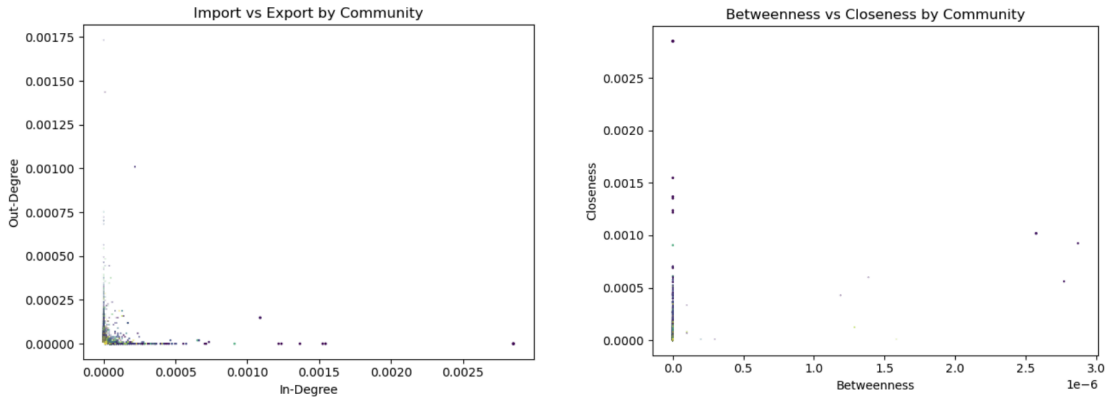

Fig.19 Company roles by Louvain community in centrality space

Fig.20 visualizes the top 50 companies based on their four-dimensional undirected centrality profiles. Three clusters emerge: global hubs like IKEA and Honour Lane Shipping score high across all dimensions; connector companies such as Schenker and UPS show strong degree and closeness but limited PageRank; and peripheral players exhibit uniformly low values. The heatmap illustrates how only a few companies occupy structurally central positions, while most remain marginal in influence and brokerage.

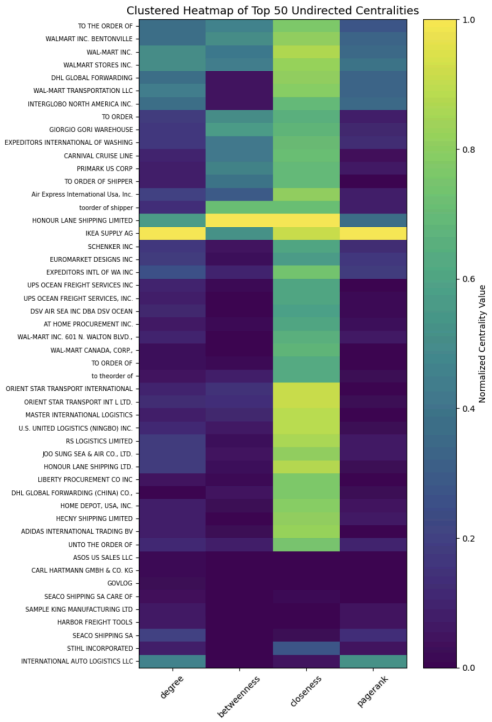

Fig.20 Clustered Heatmap of Top 50 Undirected Centralities

Fig.21 presents the Pearson correlation matrix of the four undirected centrality measures. Degree and PageRank exhibit a near-perfect correlation ($\rho = 0.93$), indicating that companies with broad connectivity also tend to wield high influence in the network. Betweenness is moderately correlated with both degree ($\rho = 0.42$) and PageRank ($\rho = 0.37$), suggesting that companies acting as intermediaries often—but not always—have many connections. Closeness centrality, by contrast, is weakly correlated with all other measures ($\rho < 0.25$), reinforcing its conceptual distinctiveness: reachability does not necessarily imply influence or brokerage power.

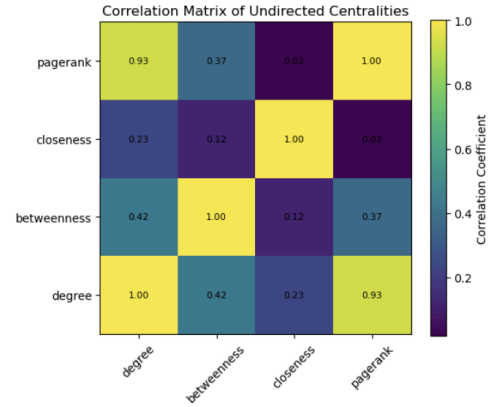

Fig.21 Correlation Matrix of Undirected Centralities

5.3 Network Analysis with Risk Indicators

To evaluate the underlying risk exposure across the global supply chain, we analyzed three key indicators: Country Risk Assessment (CRA), Sector Risk Assessment (SRA), and Debtor Risk Assessment (DRA). These metrics collectively capture geopolitical, industrial, and company-specific vulnerabilities, respectively. Country Risk Assessment (CRA) reflects a country's macroeconomic and political stability, ranging from A1–A2 (very low risk) to C–D (high risk). Most firms in our dataset operate in A1–A2 countries, suggesting stable geopolitical contexts, though links to B–D countries reveal pockets of elevated sovereign risk. Sector Risk Assessment (SRA) scores industries from 1 (very low) to 4 (high) based on volatility. The network skews toward stable sectors like consumer goods and logistics, with limited exposure to high-risk domains. Debtor Risk Assessment (DRA) measures firm-level creditworthiness on a 0–10 scale. While most firms score above 6, a notable subset falls below 5, indicating financial fragility. The broader spread in DRA underscores the need to integrate company-level credit data with country and sector risk to fully assess supply chain vulnerabilities.

Taken together, these risk distributions suggest that while global shipping networks are generally concentrated in low-risk countries and industries, financial fragility at the company level is more unevenly distributed. Therefore, a multilayered risk framework—one that incorporates CRA, SRA, and DRA in tandem—offers a more comprehensive and operationally actionable understanding of exposure across the network.

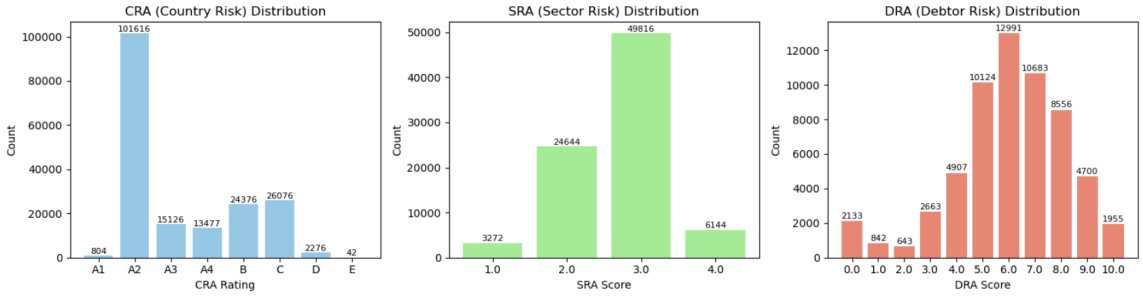

Fig.22 CRA/SRA/DRA Distribution

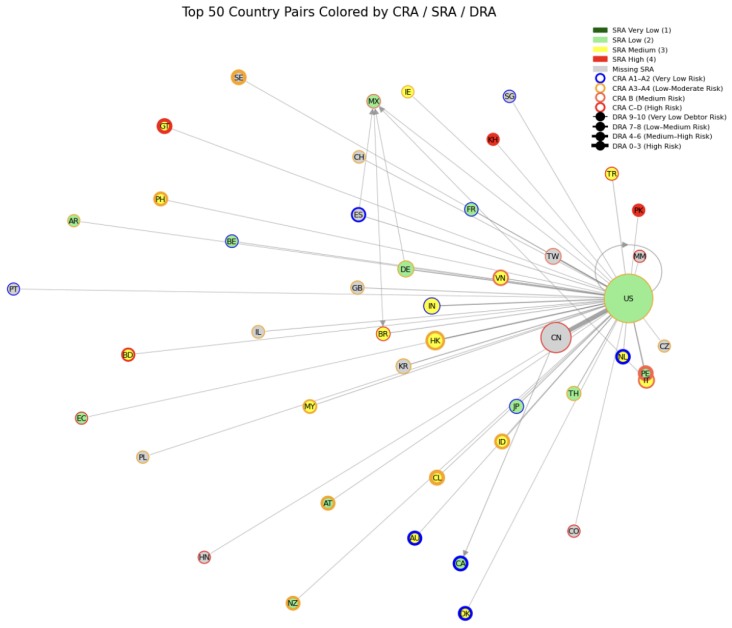

Fig.23 Top 50 Country Pairs Colored by CRA/SRA/DRA

The country-level network graph visualizes the top 50 directed shipping and consignee country pairs with three risk dimensions: CRA, SRA, and DRA. Node fill color reflects sectoral risk (SRA), border color indicates country risk (CRA), and border thickness encodes debtor risk (DRA), with thicker borders denoting higher financial default risk.SRA categories range from very low (green) to high (red), CRA is split into five tiers from A1 (very low) to D (very high), and DRA is inversely scaled so that higher risk results in thicker borders. By integrating these three risk indicators into a single directed graph, this figure presents a holistic overview of both supply chain connectivity and embedded financial exposure.

The U.S. and China are the most central trade hubs, but their risk profiles diverge. The U.S. shows high connectivity with low sector and country risk, while China trades heavily with higher-risk partners like Bangladesh and Pakistan, signaling greater exposure. Many Latin American and Southeast Asian countries face moderate to high risks across SRA, CRA, and DRA. Vietnam and India maintain strong flows despite financial or sector vulnerabilities, whereas Germany and Mexico combine high connectivity with relatively low risk. This analysis highlights that trade centrality does not guarantee stability. Key economies often engage with risk-prone partners, reinforcing the need for multi-dimensional risk assessment in global supply chain planning.

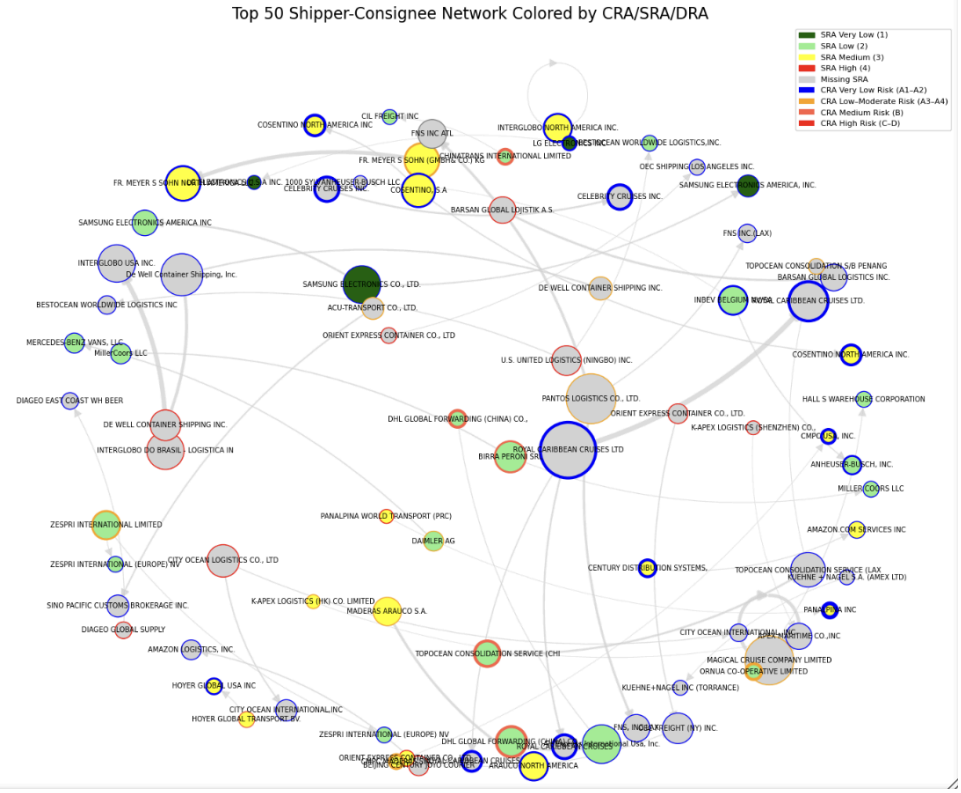

Fig.24 Top 50 Shipper-Consignee Network Colored by CRA/SRA/DRA

The network visualization in Fig.24 offers a company-level mapping of the top 50 shipper-consignee pairs with risk information from three dimensions of risk. Each node represents a unique company, while directed edges indicate transactional relationships weighted by shipment volume. Similarly, node size reflects total trade flow, while fill color indicates sector risk (SRA), border color denotes country risk (CRA), and border thickness represents debtor risk (DRA).

Large, centrally positioned nodes such as Samsung Electronics, Royal Caribbean Cruises, and DHL Global Forwarding highlight companies with high shipment volume and wide connectivity. However, centrality does not always correspond to low risk. Several key players are linked to medium or high country risk (CRA B–D), reflecting geopolitical exposure despite operational importance. Sector risk (SRA) shows a mix of low- and high-risk companies across the network. While core nodes tend to be in stable sectors (SRA 1–2), notable exceptions—such as freight and consolidation companies—operate in higher-risk industries, revealing potential points of vulnerability. Debtor risk (DRA), indicated by border thickness, adds another layer: financially fragile companies are not confined to the periphery but appear within important trade corridors, raising concerns about cascading risk. The color-coded CRA also shows that many companies, especially those connected to Southeast Asia and Latin America, operate in jurisdictions with elevated sovereign or regulatory risk. Even companies based in low-risk countries often transact heavily with higher-risk partners, underscoring the need for cross-border risk mitigation strategies. Overall, the graph illustrates that trade centrality does not guarantee resilience. Multiple types of risk—geopolitical, financial, and sectoral—converge across key actors, making them critical targets for enhanced monitoring, due diligence, and contingency planning.

6. Predictive Models

To evaluate predictive model performance across distinct risk categories—CRA (Credit Risk Assessment), SRA (Spending Risk Assessment), LPI (Late Payment Index), and DRA (Default Risk Assessment)—we compared four machine learning models: Random Forest (RF), Logistic Regression, XGBoost (XGB), and LightGBM (LGB). Two key evaluation metrics were used: Accuracy and Macro F1-score, visualized respectively on the left and right subplots.

Model	CRA Accuracy	CRA F1 (Macro)	SRA Accuracy	SRA F1 (Macro)	LPI Accuracy	LPI F1 (Macro)	DRA Accuracy	DRA F1 (Macro)
RF	0.999	1.000	0.780	0.688	0.990	0.690	0.921	0.534
Logistic	1.000	1.000	0.750	0.596	0.986	0.497	0.922	0.433
XGB	0.999	1.000	0.747	0.598	0.986	0.598	0.928	0.490
LGB	0.999	0.996	0.731	0.559	0.985	0.637	0.924	0.460

Table.1 Predictive Models

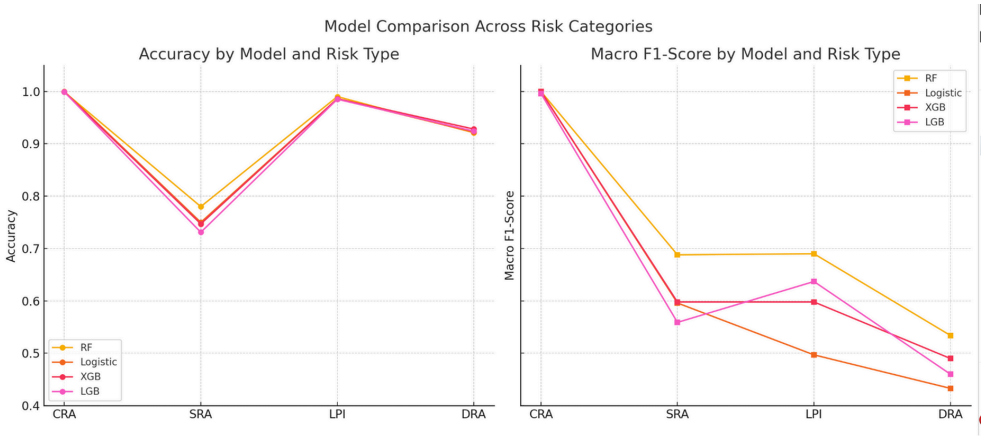

Fig.25 Model Comparison Across Risk Categories

6.1.1 Accuracy Comparison

As shown in both the table and the left plot of figure above, all models achieved near-perfect accuracy for CRA and LPI tasks ($\geq$ 0.985), indicating these problems are well-learned and potentially less complex or more linearly separable. For DRA, all models also performed well, with accuracies ranging from 0.921 to 0.928.

SRA stands out as the most challenging task, with all models experiencing a notable dip in accuracy. RF achieved the highest performance (0.780), followed by Logistic (0.750), XGB (0.747), and LGB (0.731). This suggests that the SRA dataset may be more complex, possibly due to higher class imbalance or less informative features.

6.1.2 Macro F1-Score Comparison

Macro F1-score further highlights the effect of class imbalance across tasks. While all models performed exceptionally on CRA with F1-scores of ~1.00, the scores dropped significantly for SRA and DRA. In particular, Logistic Regression showed poor performance on SRA (0.596) and DRA (0.433), reflecting difficulties in correctly predicting minority classes.

6.2 Overall Observations

Random Forest consistently demonstrated the most robust performance across all risk categories, particularly in handling imbalanced and multi-class tasks such as Spending Risk Assessment (SRA) and Default Risk Assessment (DRA). Its ensemble structure enables it to capture complex patterns that simpler models may miss.

Logistic Regression performed well on the Credit Risk Assessment (CRA) task but struggled with more complex datasets, likely due to its linear nature. XGBoost and LightGBM delivered similar performance, with each showing slight advantages depending on the task.

For the Late Payment Index (LPI), Random Forest also performed the best overall, demonstrating strong ability to capture non-linear relationships in binary classification.

Task difficulty, based on performance metrics, can be ranked from most to least challenging as: SRA, DRA, LPI, and CRA. These results highlight the importance of evaluating models using both accuracy and macro F1-score, especially under class imbalance.

While the results are promising, the models are still under development and haven't been tested on real-world operational data yet, and refining them for practical application will be a key focus of our future work.

7. Conclusion & Future works

This project successfully developed a robust risk classification framework for small and medium-sized enterprises (SMEs) by leveraging CRA, SRA, DRA, and payment behavior data. Through data preprocessing, feature engineering, and clustering techniques, we established meaningful business groupings and laid the foundation for risk stratification. The use of visual analytics and initial modeling efforts demonstrated the potential to translate complex datasets into actionable insights for the Rel8ed platform.

Looking ahead, several enhancements can elevate the effectiveness of the risk framework. Integrating external risk signals such as sanctions lists, ESG scores, and macroeconomic indicators will broaden the risk context. Developing a weighted composite risk score and incorporating time series analysis will further refine prediction accuracy and capture temporal dynamics. Ultimately, translating analytical findings into strategic decisions will enable Rel8ed to deliver scalable, data-driven risk management solutions.

Reference

[1] DeepSeek. (2024). DeepSeek Location Extraction API. Retrieved from https://deepseek.ai.

[2] Newman, M. E. J. (2018). Networks. Oxford University Press.

[3] Ponomarov, S. Y., & Holcomb, M. C. (2009). Understanding the concept of supply chain resilience. The international journal of logistics management, 20(1), 124-143.