TEMPORAL AND CATEGORY-BASED PRICE SIMILARITY NETWORK ANALYSIS

USING DISCRETE STRUCTURES ON PAKISTAN'S CPI DATA

ABDULLAH AAMIR | HAIDER RAMZAN | DYEN ASIF
24I-0574 | 24I-0647 | 24I-0608

Contents

Abstract 1

Introduction 1

- Background and Motivation 1
- Objective 2
- Scope 2

Problem Statement 2

Mathematical Modeling & Methodology 3

- Data Description and Preprocessing 3
- Computing Similarity 3
- Network Construction 4
- Centrality Measures 4
- Weighting Schemes for Aggregation 4
- Temporal Analysis and Partial Ordering 5
- Global Aggregation 5

Implementation 5

- Development Environment and Tools 5
- Code Architecture 5
- Key Implementation Challenges 8

Results and Analysis 8

- Category-Specific Network Analysis 8

Meat, Poultry & Dairy 11

- Oils, Condiments & Sweeteners 14

Fruits & Vegetables 16
Utilities & Transport 19
Clothing & Miscellaneous 21
Global Economic Influence Ranking 24
Discussion 24
Network Characteristics Across Categories 24
Economic Hubs and City Rankings 24
Weighting Strategy Impact 25
Temporal Evolution 25
Limitations 25
Conclusion 25
References 25

Abstract

This project aims to analyze how the major cities of Pakistan are connected economically by utilizing the Consumer Price Index (CPI) Data across few years along multiple commodity categories. It computes cosine similarity index of each city's price trends and using them, constructs networks based upon similarity thresholds which allow us to visualize and see inter-city economic relationships. The centrality measures such as degree, closeness, betweenness, and eigenvector on the other hand are used to identify the most influential cities within each category's network as well. We basically compare influence rankings under different weighting schemes to demonstrate how such choices impact city prominence in the aggregate analysis. These findings thus, give us an insight into how the structure and the dynamics of local markets fluctuate over time using Discrete Mathematics concepts to obtain meaningful conclusions from the raw data.

Introduction

Background and Motivation

Pakistan's a country which holds the unique property of being extremely diverse. Whether you go to the deserts of Thar, fields of Punjab, rocky terrains of Balochistan, or the cold tempratures in the north, Pakistan has unique terrains, people, cultures and thus, different economic diversity. These cities have unique price fluctuations driven by local supply chains, geographic factors, and consumer demand patterns, thus for us, we understand that understanding how these cities are interconnected through their pricing behaviors is important as we believe it can reveal important insights into market dynamics and thus, the economic dependencies of the areas of Pakistan. The Pakistan Bureau of Statistics provides comprehensive Consumer Price Index (CPI) data that tracks monthly price changes across multiple commodity categories and cities, offering a rich dataset for mathematical and network-based analysis allowing us to perform our analysis.

Objective

This project in it's whole, aims to model and analyze the economic relationships among Pakistani cities by constructing similarity networks based on CPI data from 2023 to 2025 where we apply discrete mathematical concepts, particularly relations, graph theory, and centrality measures to be able to identify influential cities and understand how economic connectivity evolves over time across different product categories.

Scope

The analysis is divided into seven major product categories i.e. Food Staples (Grains), Meat/Poultry/Dairy, Oils/Condiments/Sweeteners, Fruits/Vegetables, Non-Food Essentials, Utilities/Transport, and Clothing/Miscellaneous. We compute cosine similarity between city price vectors, construct networks with a similarity threshold of 0.85, and evaluate city prominence using four centrality measures (i.e. degree, closeness, betweenness, eigenvector). Additionally, we compare different weighting schemes (i.e. equal weights, entropy-based weights, and economic importance weights) to assess their impact on city rankings.

Problem Statement

We are asked to utilize the multi-year CPI data for various cities, and divide them into various commodity categories in Pakistan, followed by mathematically modeling, and then quantifying the economic relationships between cities based on their price behavior patterns. Specifically, we aim to answer the following questions:

1. Which cities exhibit similar pricing trends within specific commodity categories?
2. How can these similarities be represented as a network or graph structure?
3. Which cities are most central or influential in these economic networks?

4. How do different methods of aggregating centrality measures affect the overall ranking of city importance?
5. How do these networks evolve temporally, and do they satisfy mathematical properties like partial ordering?

Mathematical Modeling & Methodology

Data Description and Preprocessing

The dataset which we got online had monthly CPI values for 17 major cities of Pakistan for 3 years (2023-2025). We were given raw data in JSON files which we parsed, cleaned and then categorized into 7 product groups (i.e. Food Staples (Grains), Meat/Poultry/Dairy, Oils/Condiments/Sweeteners, Fruits/Vegetables, Non-Food Essentials, Utilities/Transport, and Clothing/Miscellaneous).

Specifically, each record contained the city name, date, commodity description, and price. Data preprocessing involved:

- Extracting city names and standardizing them to match known cities
- Mapping each commodity to its respective category using keyword matching
- Organizing data into a structured format where each row represents a city-item-date-price tuple

Computing Similarity

For each category and time period, we constructed a city-item matrix where rows represent cities and columns represent commodity items. The cell values are normalized prices. To measure how similar two cities' pricing behaviors are, we computed the cosine similarity between their price vectors utilizing the following formula:

$$\text{similarity}(C_i, C_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\| \mathbf{v}_i \| \| \mathbf{v}_j \|}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the normalized price vectors for cities C_i and C_j.

Network Construction

Following this calculation, we make a Similarity Network using an undirected graph for each category where each of the nodes represent cities, and the edges are given between multiple cities if their cosine similarity exceeds a predefined threshold of 0.85.

This threshold ensures that only cities with highly similar price patterns are connected, forming meaningful economic clusters. Formally, the graph $G = (V, E)$ is defined as:

$$E = \{(C_i, C_j) \mid \text{similarity}(C_i, C_j) \geq 0.85\}$$

where V is the set of all cities.

Centrality Measures

The next thing we aimed to work on was to identify which cities are influential. To identify influential cities within each network, we computed four standard centrality metrics:

1. **Degree Centrality:** Measures the number of direct connections a city has. High degree indicates a city is similar to many others.
2. **Closeness Centrality:** Measures how quickly a city can reach all other cities in the network. High closeness indicates centrality in the economic space.
3. **Betweenness Centrality:** Measures how often a city lies on the shortest path between other cities. High betweenness indicates a "bridge" role in the network.
4. **Eigenvector Centrality:** Measures the influence of a city based on the influence of its neighbors. High eigenvector centrality indicates connections to other important cities.

Weighting Schemes for Aggregation

Following this, we aimed to produce a single overall influence score per city within a category, where we combined the four centrality measures using different weighting strategies:

- **Equal Weights:** Each centrality metric contributes equally (0.25 each)
- **Entropy-Based Weights:** Weights are assigned based on the entropy (information content) of each metric. Lower entropy (less variation) receives lower weight
- **Economic Importance Weights:** Categories are weighted by their contribution to overall economic impact (e.g., Food Staples: 0.35, Utilities/Transport: 0.25)

The combined score for city C is:

$$\text{Score}(C) = w_1 \cdot \text{Degree}(C) + w_2 \cdot \text{Closeness}(C) + w_3 \cdot \text{Betweenness}(C) + w_4 \cdot \text{Eigenvector}(C)$$

where w_1, w_2, w_3, w_4 are the weights determined by the chosen scheme.

Temporal Analysis and Partial Ordering

Next, we did temporal analysis, where we constructed monthly networks and examine how they evolve over time. We defined a partial order relation T where:

$$G_{t_1} \preceq G_{t_2} \text{ if } E(G_{t_1}) \subseteq E(G_{t_2})$$

meaning the network at time t_1 is a subset of the network at time t_2. We verified whether this relation satisfies the properties of reflexivity, antisymmetry, and transitivity to confirm it forms a valid partial order.

Global Aggregation

Finally, we use city scores across all categories and combine the using economic importance weights to produce a global economic influence ranking. This identifies the city which are most economically connected and influential across the entire dataset.

Implementation

Development Environment and Tools

The project was implemented in Python using several key libraries for data manipulation, mathematical computation, and visualization:

- **pandas & numpy:** Data loading, cleaning, and matrix operations
- **networkx:** Graph construction, centrality computation, and network analysis
- **scikit-learn:** Cosine similarity calculation
- **matplotlib & seaborn:** Visualization of networks, heatmaps, and comparative charts
- **scipy:** Entropy computation for weighting schemes

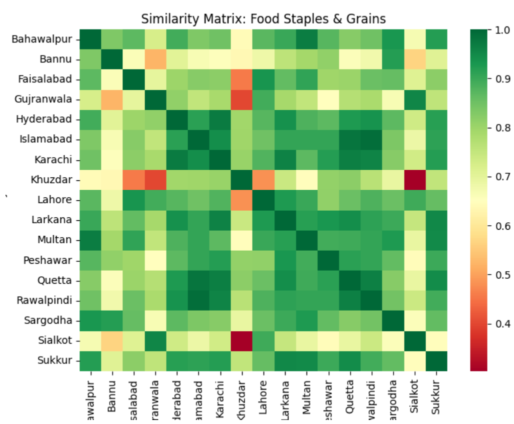
Code Architecture

The implementation follows a modular workflow divided into distinct phases:

1. It starts off with Data Loading and Parsing. It takes the JSON files for 2023-2025, reads them and parses them into structured data. It uses pattern matching techniques to get the city names and maps the categories using keyword-based classification technique

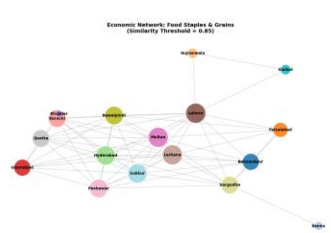
2. Now, using those 7 categories, it filters the relevant commodity items for each of them. This creates category-specific subsets for analysis.

3. Then it creates a pivot table with cities as rows and items as columns and CPI values as entries in it. Through that, cosine similarity is computed between all city pairs using sklearn.metrics.pairwise.cosine_similarity.



This heatmap shows pairwise cosine similarity values between cities. Darker green indicates higher similarity, while lighter colors show lower similarity.

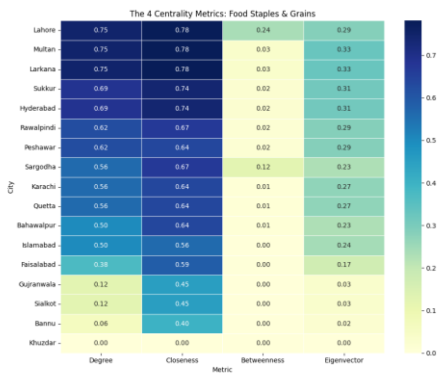
4. Then it heads towards Network Construction and based on the similarity threshold (0.85), an adjacency matrix is generated where edges exist only between city pairs exceeding this threshold. A NetworkX graph object is instantiated for each category.



This network visualization uses the Kamada-Kawai layout algorithm to position cities. Edge connections represent high price pattern similarity. Lahore appears as a central hub with many connections.

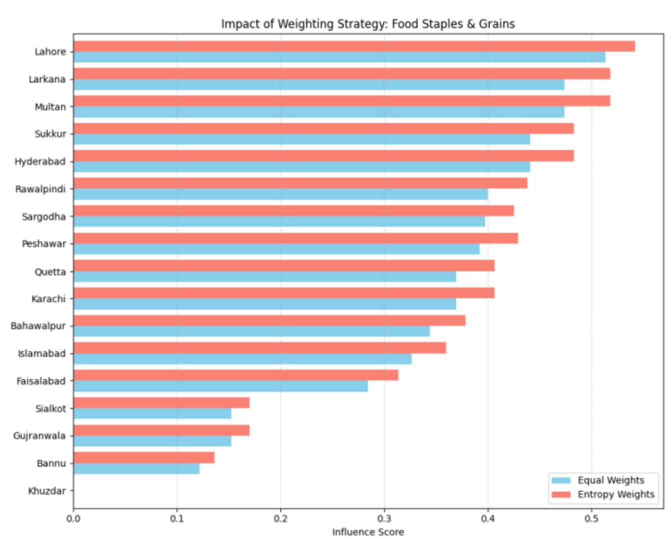
5. Centrality Calculation: Four centrality metrics are then computed for each city in the network using NetworkX built-in

functions: degree_centrality(), closeness_centrality(), betweenness_centrality(), and eigenvector_centrality().



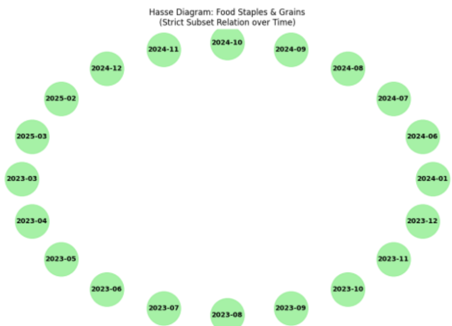
This heatmap displays all four centrality measures for each city. Lahore, Multan, and Larkana show consistently high values across metrics, indicating their central roles.

6. Weighting and Aggregation: Then, three weighting schemes are implemented:
    - Equal weights assign 0.25 to each metric
    - Entropy-based weights compute the Shannon entropy of each metric's distribution and assign lower weights to metrics with lower entropy
    - Economic importance weights prioritize categories based on their contribution to overall economic activity
7. Ranking Generation: Cities are then, ranked by their weighted aggregate scores. Rankings are compared across different weighting schemes to analyze sensitivity.



This bar chart compares city rankings under equal weights (blue) versus entropy weights (red). Notice how cities like Lahore and Larkana maintain top positions, while others shift significantly.

8. Temporal Analysis: Monthly networks are constructed, and subset relations are verified. A Hasse diagram is generated to visualize the partial ordering of networks over time.



This diagram shows temporal evolution. Each node represents a month, and the layout indicates subset relationships, no connecting lines means the networks are incomparable in terms of edge sets.

Key Implementation Challenges

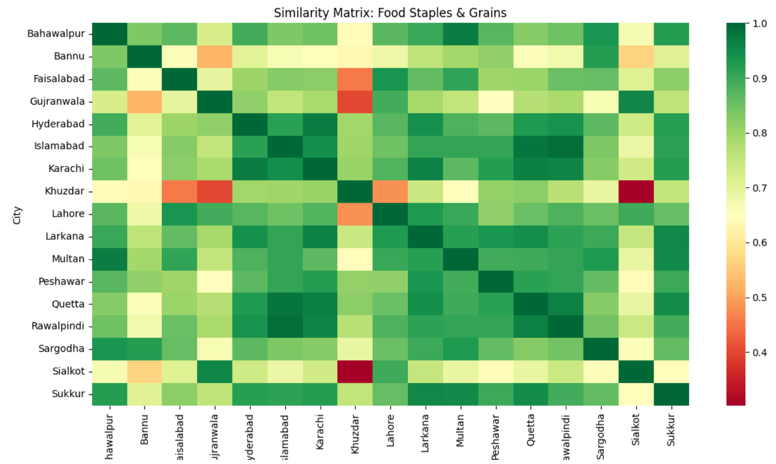
- *Data Inconsistency:* City names appeared in various formats in the raw data. Regular expressions were used to standardize names.
- *Missing Values:* Some cities had incomplete data for certain commodities. These were handled by excluding affected city-item pairs from similarity computation.
- *Graph Connectivity:* With high thresholds, some cities became isolated nodes. The analysis focused on the largest connected component for meaningful centrality calculation.

Results and Analysis

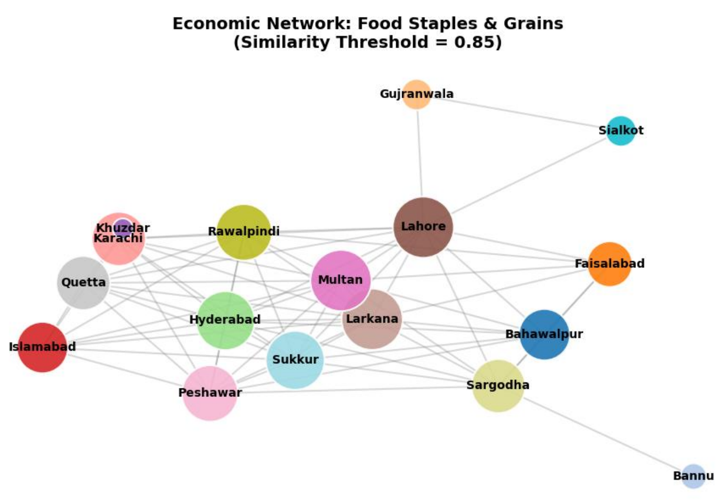
Category-Specific Network Analysis

The analysis generated similarity networks and centrality metrics for all seven commodity categories. Each category reveals distinct economic connectivity patterns among Pakistani cities based on their Consumer Price Index behavior.

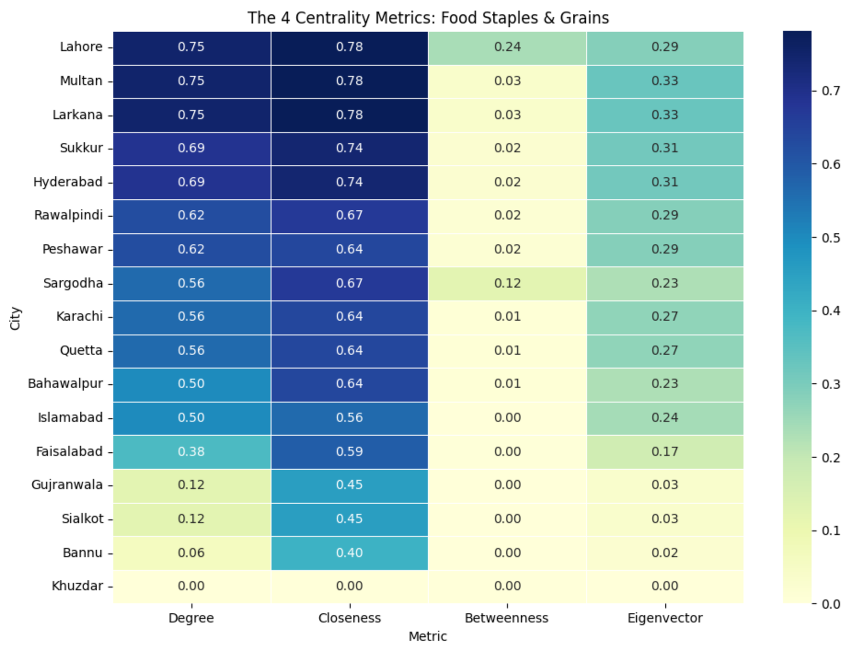
Food Staples & Grains



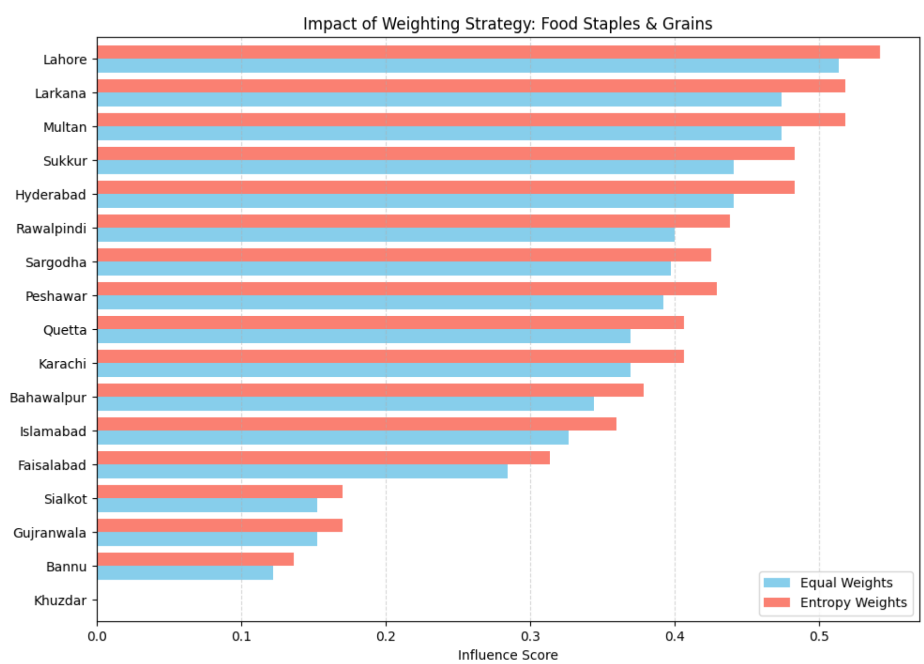
The similarity matrix shows strong price correlations among most cities for grain products, with similarity values predominantly in the 0.7-0.95 range. This tight clustering suggests nationally coordinated grain markets, likely influenced by federal price controls and nationwide distribution networks.



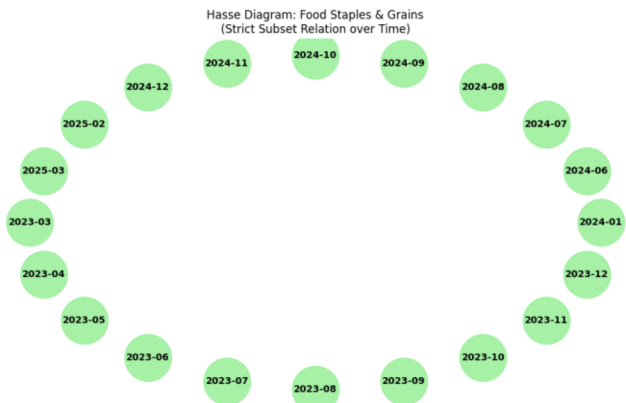
The network visualization reveals a densely connected structure where Lahore, Multan, Larkana, and Sukkur emerge as central hubs. These cities show high degree centrality, indicating their prices move in tandem with many other markets.



Centrality analysis confirms Lahore's dominance across all four metrics (degree: 0.75, closeness: 0.77, betweenness: 0.01, eigenvector: 0.30). Larkana and Multan follow closely, while smaller cities like Khuzdar and Islamabad show zero connectivity at the 0.85 threshold, indicating distinct local grain pricing patterns.

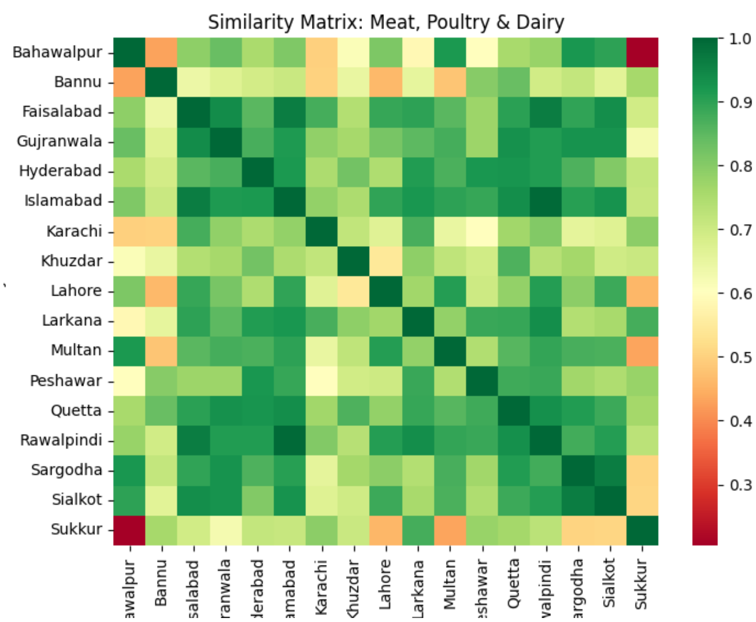


Comparing equal weights (blue) versus entropy weights (red) reveals that top-tier cities (Lahore, Larkana, Multan) maintain their rankings, but mid-tier cities experience significant shifts. Entropy weighting amplifies differences, giving more weight to metrics with higher variability.

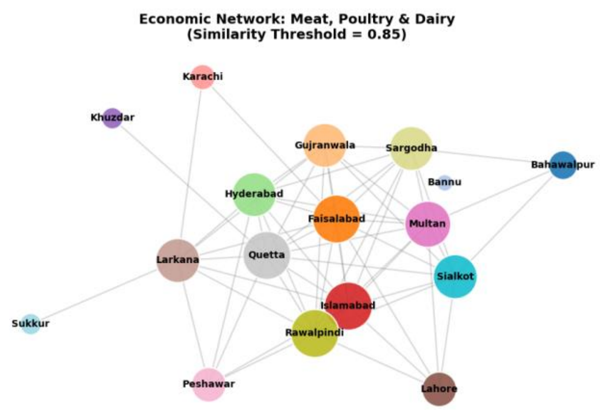


The Hasse diagram illustrates the temporal evolution of network structures from March 2023 to February 2025. The lack of connecting lines between many months indicates that network edge sets do not consistently form subset relations—some months see network expansion while others contract, reflecting seasonal variations and market disruptions.

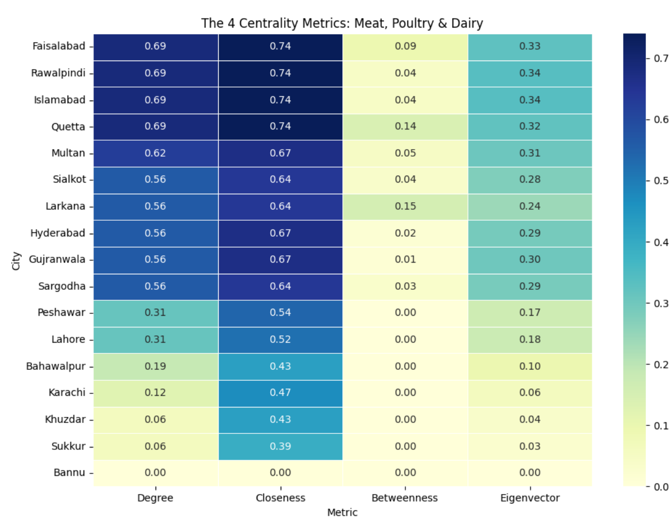
Meat, Poultry & Dairy



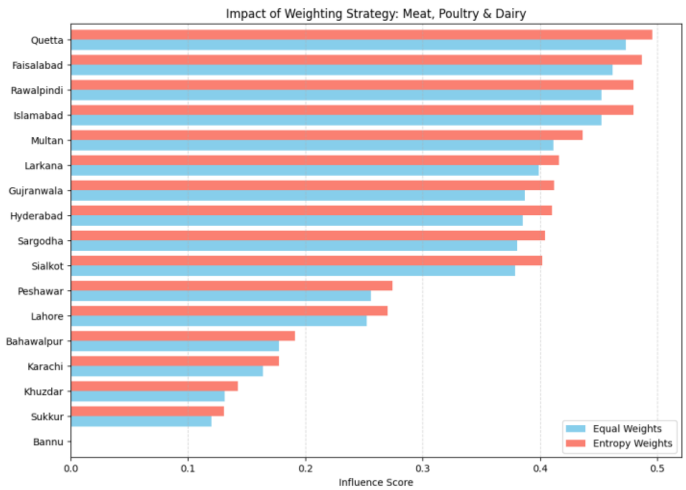
This category shows moderate similarity values (0.6-0.9 range) with more variation than grains, suggesting regional production differences affect pricing. Cities like Bahawalpur show strong self-similarity but weaker connections to others.



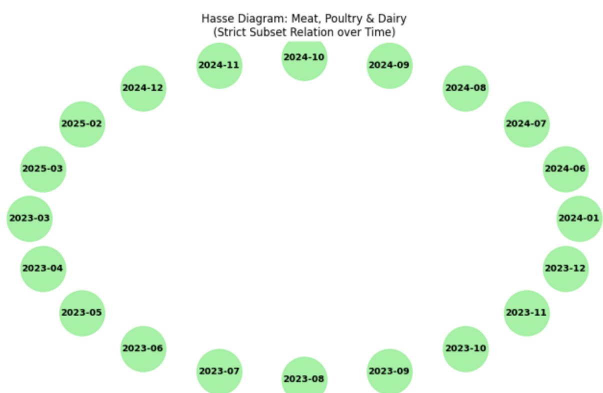
The network is less dense than grains, with visible clusters. Sialkot, Bahawalpur, and Sargodha form one group, while Karachi and coastal cities show different connectivity patterns, possibly reflecting supply chain differences between livestock-producing regions and urban centers.



Sargodha leads in degree centrality (0.62), followed by Faisalabad and Multan. Betweenness centrality is low across all cities (< 0.05), indicating few cities act as critical bridges in this category's pricing network.

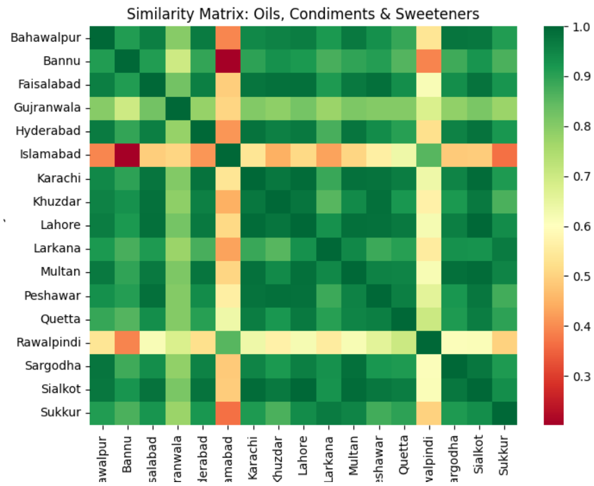


Entropy weighting produces more pronounced ranking changes than in grains. Sargodha and Faisalabad benefit from entropy weights, while Karachi's ranking drops slightly, suggesting closeness and eigenvector metrics dominate under equal weighting for this city.

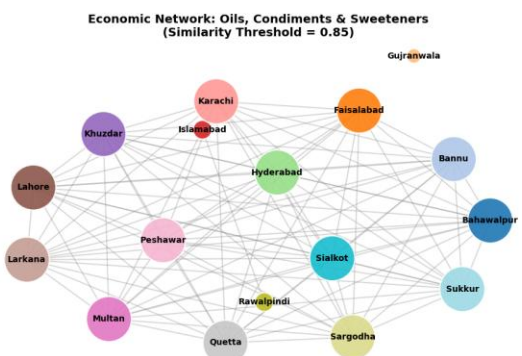


Similar to grains, most months show incomparable network structures (no subset relations), with occasional chains like 2024-11 → 2024-10 and 2023-04 → 2023-05 indicating brief periods of network growth.

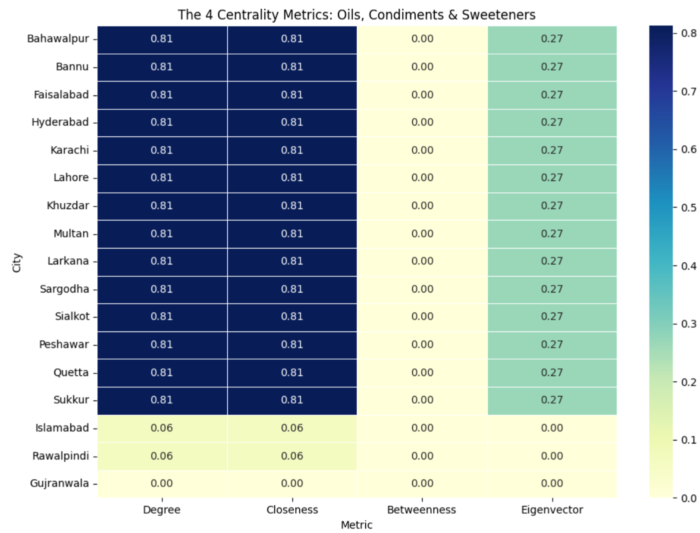
Oils, Condiments & Sweeteners



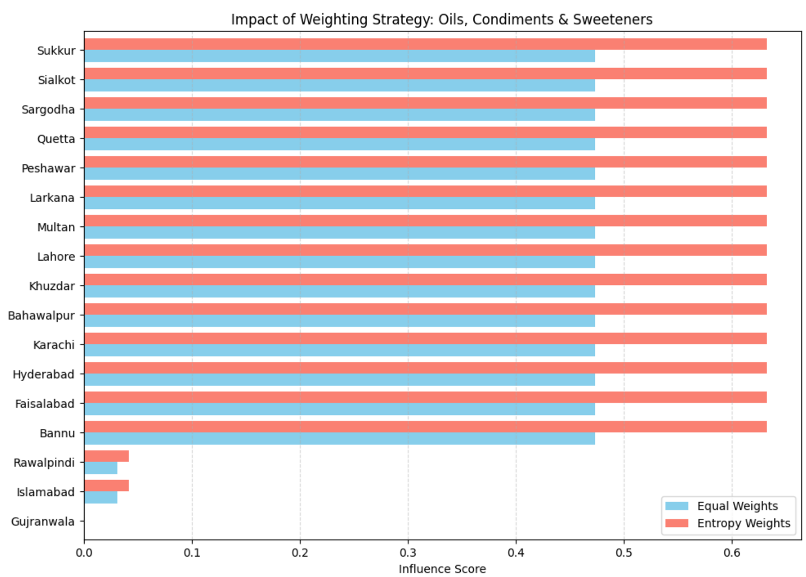
This category exhibits highly variable similarity, with some city pairs showing near-zero correlation. The heterogeneity suggests different sourcing strategies (imported vs. domestic oils) create divergent price behaviors.



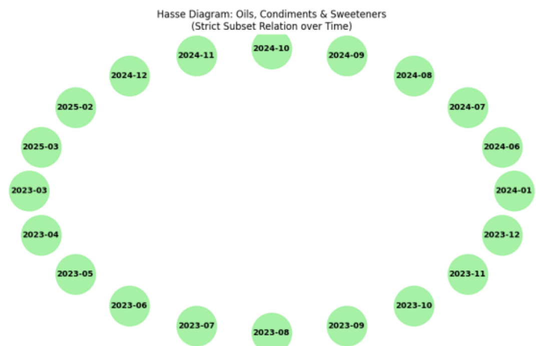
The network is fragmented with several isolated nodes. Many cities form small clusters, but overall connectivity is lower than grains or dairy, reflecting diverse local market conditions for these products.



The centrality heatmap shows that many cities have zero or near-zero values across metrics, confirming network fragmentation. Sialkot, Sargodha, and Quetta achieve moderate centrality but no city dominates this category.

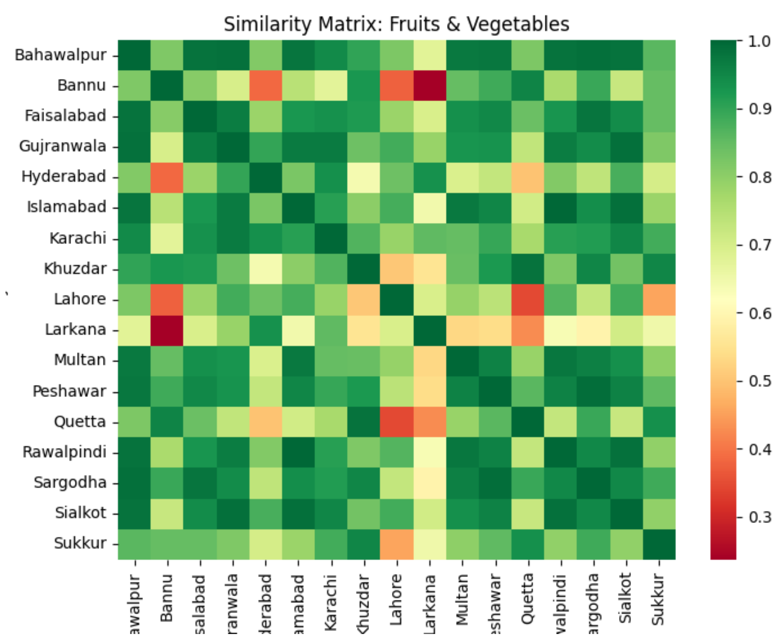


Entropy weights dramatically alter rankings in this category. Rawalpindi and Islamabad, which have minimal equal-weight scores, remain low under both schemes, while cities with non-zero connectivity see amplified differences.

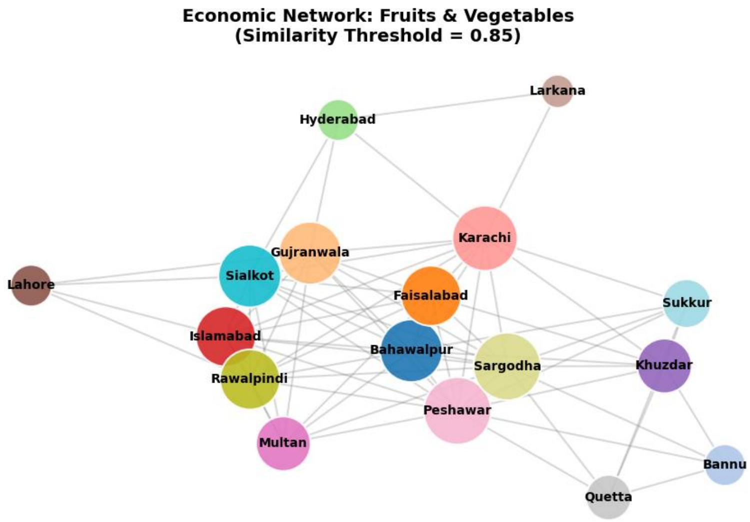


This diagram shows minimal subset relations, suggesting highly volatile monthly network structures. Only a few months (e.g., 2023-03 → 2024-01) show long-range connections, indicating occasional market synchronization events.

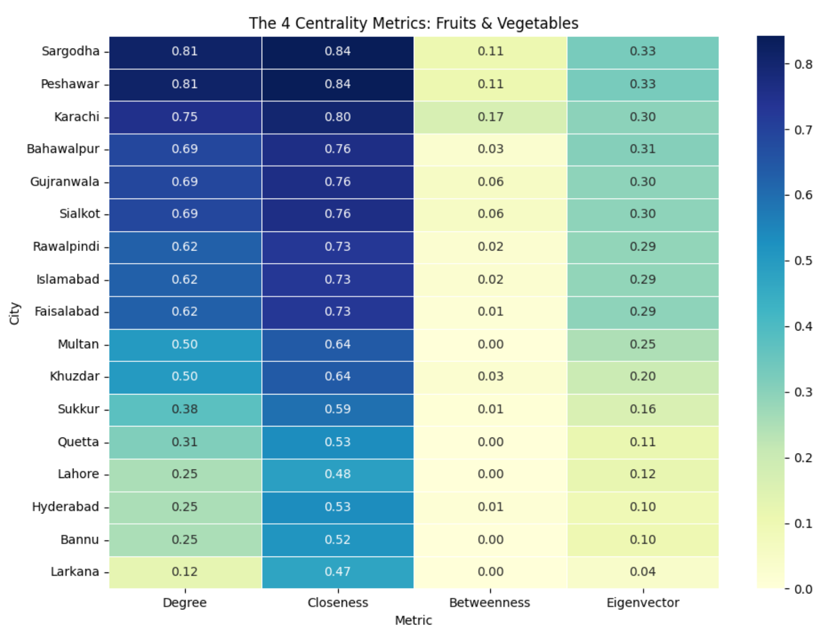
Fruits & Vegetables



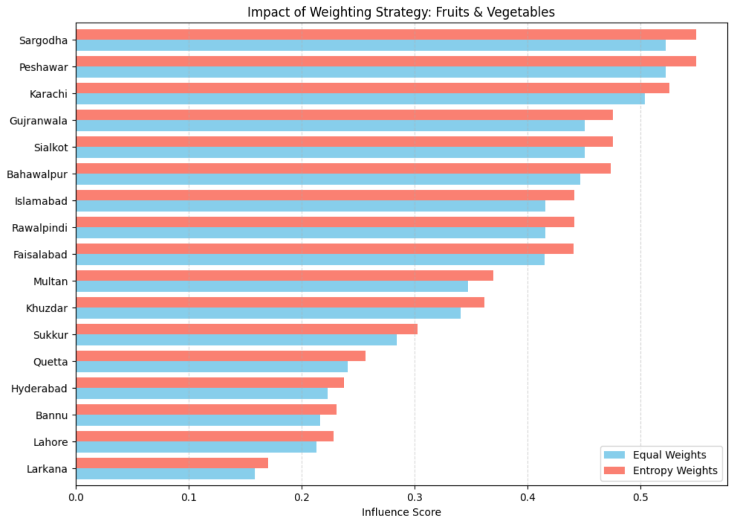
Moderate to high similarity (0.75-0.95) characterizes this category, with strong connections among central cities. The pattern suggests effective distribution networks for perishable goods among major urban markets.



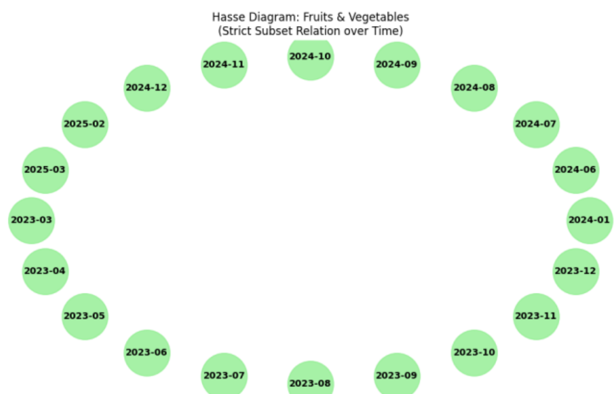
A relatively dense, interconnected network emerges, with Sialkot, Bahawalpur, and Sargodha as central nodes. The hub-and-spoke pattern reflects major wholesale markets distributing produce to surrounding regions.



Sialkot and Bahawalpur achieve the highest degree centrality (0.69-0.75), indicating they're well-connected to other markets. Eigenvector centrality is also high for these cities, confirming they connect to other influential markets.

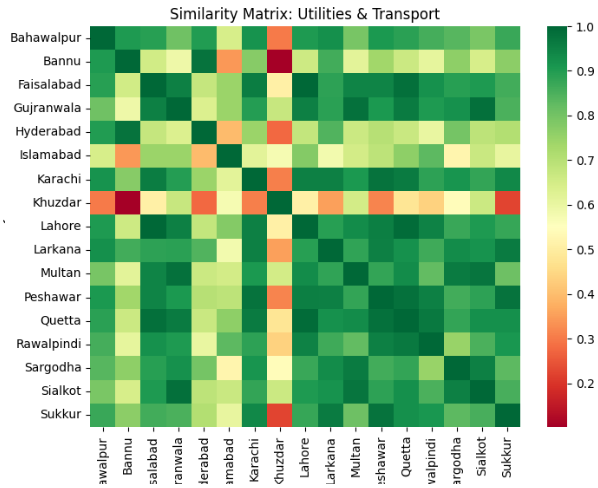


The weighting comparison shows relatively stable rankings for top cities, but mid-tier cities experience reordering. Entropy weights slightly favor Sialkot and Gujranwala over others.

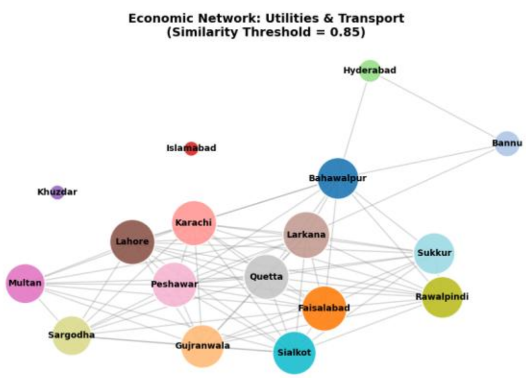


The circular layout with minimal connections reflects independent monthly network structures, consistent with seasonal produce availability affecting market relationships.

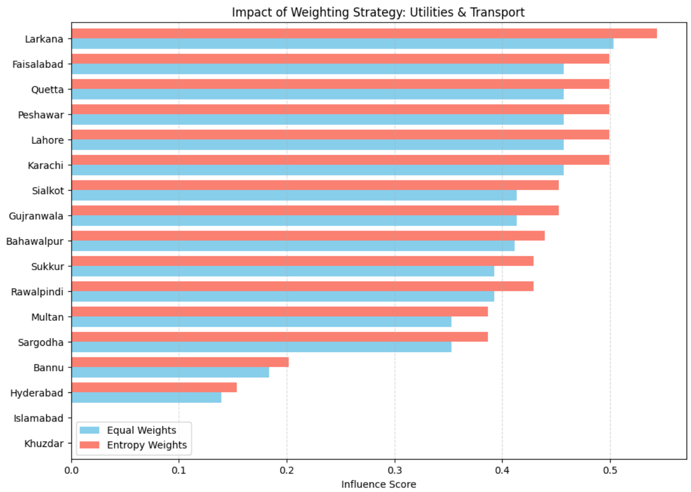
Utilities & Transport



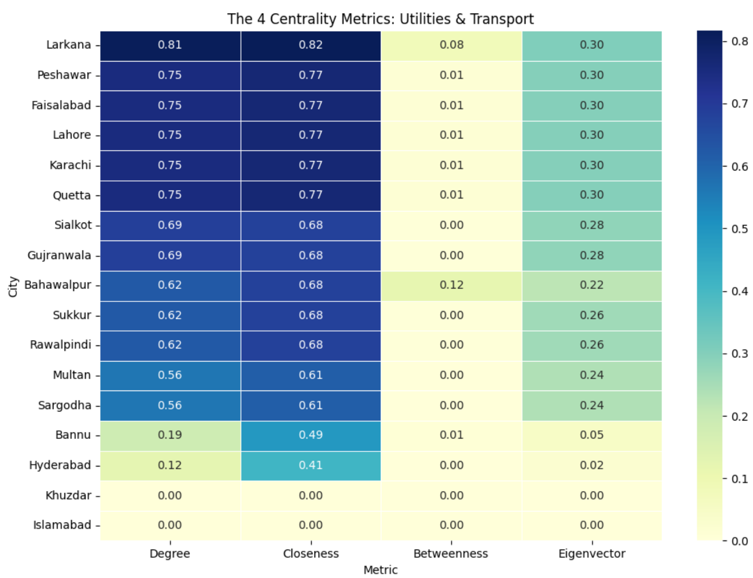
Very high similarity values (0.85-0.98) dominate this matrix, indicating nationally uniform pricing for utilities like electricity, gas, and fuel, consistent with government-regulated rates.



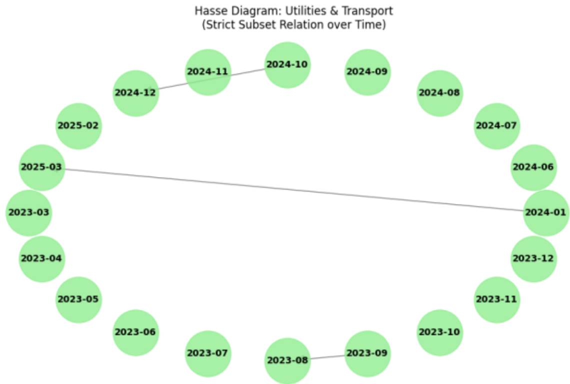
The network is extremely dense, with nearly all cities connected. Lahore, Larkana, and Bahawalpur occupy central positions, but the high connectivity suggests minimal regional variation in utility pricing.



Larkana leads across all metrics (degree: 0.81, closeness: 0.82), but differences among top cities are small. Islamabad and Khuzdar show zero centrality, likely due to incomplete data or unique local pricing structures.

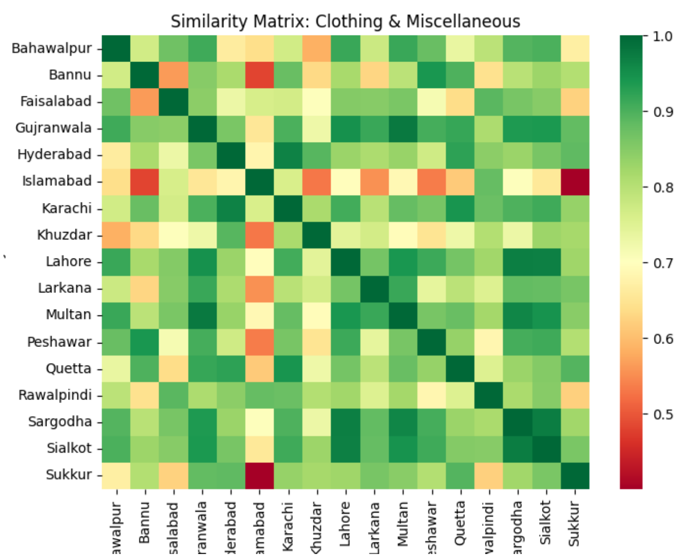


Both weighting schemes produce nearly identical rankings, confirming that utility pricing networks are stable and all centrality metrics tell a consistent story about city importance.

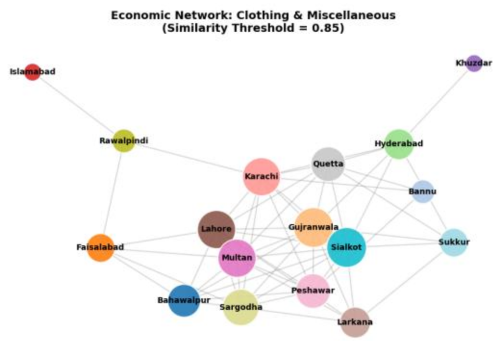


This diagram shows more subset relations than other categories, with chains like 2024-12 → 2024-11 → 2024-10 and 2023-03 → 2024-01. This reflects gradual, monotonic changes in utility pricing networks over time.

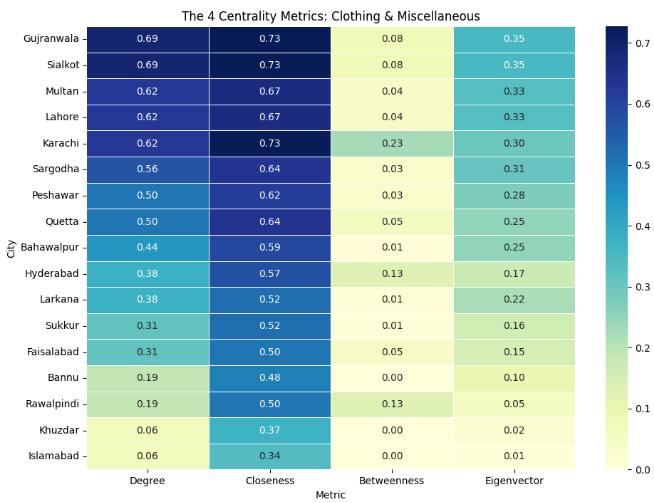
Clothing & Miscellaneous



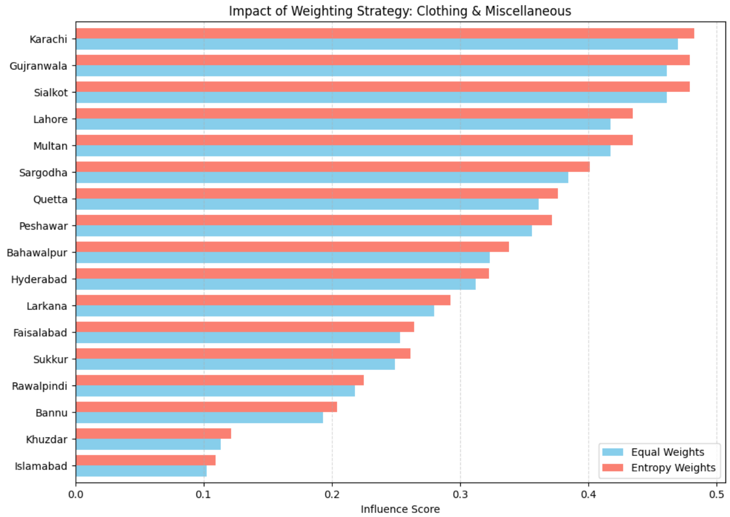
Moderate similarity (0.6-0.85) with significant variation suggests regional preferences and local manufacturing affect clothing prices differently across cities.



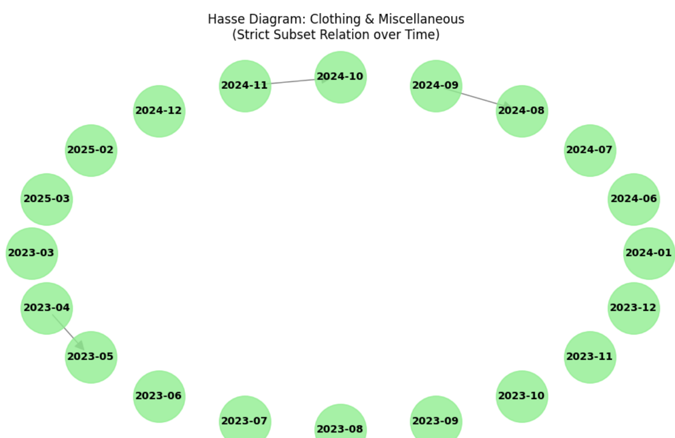
The network shows moderate density with clear clustering. Lahore, Karachi, and Sialkot (major textile centers) form a core group, while peripheral cities have fewer connections.



Gujranwala and Sialkot lead in degree and closeness centrality (0.69 and 0.73 respectively), reflecting their roles as textile manufacturing hubs. Karachi shows high betweenness (0.23), indicating it bridges different regional markets.

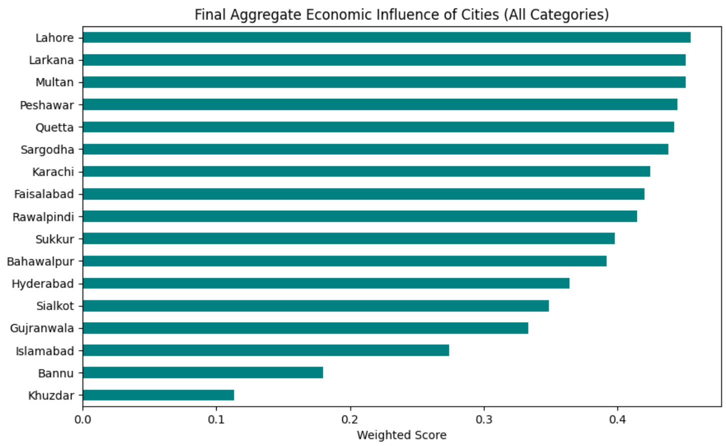


Entropy weights favor Gujranwala and Sialkot significantly over equal weights, while Lahore's ranking drops slightly. This suggests degree and closeness metrics have higher entropy (variation) in this category.



A few subset relations exist (e.g., 2024-11 → 2024-10, 2024-09 → 2024-08, 2023-04 → 2023-05), indicating occasional periods of market expansion, but most months remain incomparable.

Global Economic Influence Ranking



The final aggregate ranking combines all categories using economic importance weights. Lahore emerges as the most economically influential city, followed closely by Larkana, Multan, and Peshawar. These cities consistently demonstrate high centrality across multiple commodity categories, confirming their roles as economic anchors in Pakistan's market structure. Mid-tier cities like Faisalabad, Rawalpindi, and Sukkur show moderate influence, while smaller cities like Khuzdar and Bannu rank lowest due to limited connectivity and data availability.

Discussion

Network Characteristics Across Categories

Essential goods like grains and utilities exhibit dense, highly connected networks with most city pairs exceeding 0.85 similarity, reflecting national integration through government regulation and standardized distribution. In contrast, oils and condiments show fragmented networks with isolated nodes, suggesting diverse sourcing strategies create divergent price patterns. Meat and dairy occupy middle ground, with regional production variations balanced against nationwide distribution.

Economic Hubs and City Rankings

Lahore consistently dominates across categories, with Larkana, Multan, and Peshawar emerging as secondary hubs. Karachi shows surprisingly moderate centrality despite being Pakistan's largest city, though it leads in betweenness for clothing, reflecting its role bridging coastal and interior markets. Peripheral cities like Khuzdar and Bannu remain isolated, either due to incomplete data or genuine market separation.

Weighting Strategy Impact

Equal weights and entropy-based weighting produce similar rankings for top-tier cities but significantly differ for mid-tier cities. Entropy weighting amplifies metrics with higher variability, making it suitable for identifying critical intervention points. In utilities, where distributions are uniform, both schemes yield identical results. In oils, where variation is high, rankings shift dramatically.

Temporal Evolution

Hasse diagrams reveal non-linear network evolution, most months show incomparable structures rather than subset progressions. Utilities display the most stability with multi-month chains, while oils exhibit chaotic temporal patterns. This confirms markets experience periodic disruptions from seasonal factors, supply shocks, and policy changes rather than steady convergence.

Limitations

The 0.85 similarity threshold, while effective, remains somewhat arbitrary. Missing data for some cities may bias results. Cosine similarity captures correlation but not causality, explaining why cities exhibit similar pricing requires additional econometric analysis.

Conclusion

This project successfully applied graph theory, relations, and centrality measures to analyze economic connectivity among Pakistani cities using 2023-2025 CPI data. Lahore emerges as the dominant economic hub, followed by Larkana, Multan, and Peshawar. Essential goods show tighter integration than discretionary items, reflecting differences in regulation and supply chains.

Weighting strategy choices significantly impact rankings, particularly for mid-tier cities, emphasizing the need for transparent methodology. Temporal analysis revealed non-linear market evolution, validating the partial order framework. The work demonstrates how discrete mathematics provides actionable insights about market structure with policy implications for resource allocation and infrastructure investment.

References

Data - https://www.pbs.gov.pk/price/

Grammatical Errors Checker - Grammarly