Hierarchical Clustering of Financial Networks Using Generative Models

Ishaq Hamza, Rohit Jorige, Nagasai, Sathvik, Shankaradithyaa

November 15, 2024

Outline

1. Problem Statements
2. Data Sources
3. Model
   - Weighted Stochastic Block Model
     - Edge Existence Probability
     - Edge Weight Distribution
4. Hierarchical extension: Nested Block Model
5. Experiments
6. Challenges & Improvements
7. Contributions and References

Problem Statements

Identify and analyze natural trade circles among countries

- Detect potential trade alliances and unusual transaction patterns (e.g., Russia-North Korea exchanges)

Investigate if stock-price movement correlations reflect similar business practices

- Uncover clusters that may offer strategic economic insights and support policy decisions

Data Sources

- Data on countries' imports and exports to each other. wits.worldbank.org
- Stock market data from Yahoo Finance.

Model

Recap of Generative Equation from lecture notes:
Given the partition of the network $\mathbf{b} = \{b_i\}$ into groups, where $b_i \in [0, B-1]$ is the group membership of node i, we define a model that generates a network A with probability $P(A|\theta, b) = \prod_{i<j} P(A_{ij}|\theta, b_i, b_j)$, where A_{ij} is the edge between nodes i and j, and θ are the model parameters.
The likelihood that A was generated by a given partition b is obtained via the Bayesian posterior probability

$$P(b|A) = \frac{P(A|\theta)P(b,\theta)}{P(A)}$$

where $P(A|\theta) = \sum_b P(A|b,\theta)P(b)$ is the marginal likelihood of the network and $P(b,\theta)$ is the prior probability of the partition and the model parameters.

Model

We make a simplifying assumption that only one θ is compatible with the data, which simplifies the Bayesian posterior to

$$P(b|A) = \frac{P(A|\theta, b)P(b, \theta)}{P(A)}$$

which can be written as

$$P(b|A) = \frac{exp(-\Sigma)}{P(A)}$$

where $\Sigma = -\log P(A|\theta, b) - \log P(b, \theta) = \Sigma_{model} + \Sigma_{data|model}$ is called the *description length* of the network which is to be interpreted as the amount of the information required to describe the network given b, θ.
$\rightarrow$ This reduces the problem of finding the best partition b to minimizing the description length Σ. For which there are algorithms abstracted away in graph_tool.

Model

The probability of an edge existing between any two nodes depends solely on their community assignments:

$$P((i,j) \in E \mid b_i = r, b_j = s) = \omega_{rs}$$

Given that an edge exists between nodes i and j, the weight w_{ij} is modeled by a probability distribution conditioned on their community memberships.

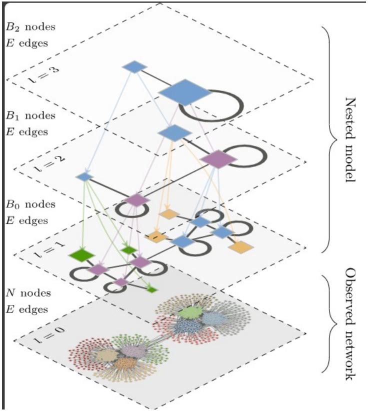
$$w_{ij} \, | b_i = r, b_j = s \sim P(w | \theta_{rs})$$

where θ_{rs} are parameters of the distribution for the weight between communities r and s.

Choice of distribution for $P(w \mid \theta_{rs})$: Exponential, the reason is we are modeling non-negative weights and larger weights are to be related with high frequency of shipments. The number of arrivals of packages can be modelled by a Poisson distribution, and the time between arrivals can be modelled by an exponential distribution.

$$\mathcal{L}(A \mid \{b_i\}, \{\omega_{rs}, \theta_{rs}\}) = \prod_{(i,j)\in E} P(w_{ij} \mid \theta_{b_i b_j} \times \prod_{(i,j)\notin E} (1 - \omega_{b_i b_j}))$$

Hierarchical extension: Nested Block Model



For each node i, multiple levels of community assignments $(b_i^{(1)} \dots b_i^{(L)})$ are considered, where each level corresponds to a different resolution of the network. To optimize Σ, multi-level markov chain monte carlo method is used, which samples from hierarchy of distributions increasing in complexity.

Experiments

Harmonic mean metric: Weight of edge from A to B

$$\left(\frac{1}{\text{A to B imports}}+\frac{1}{\text{B to A imports}}+\frac{1}{\text{A to B exports}}+\frac{1}{\text{B to A exports}}\right)^{-1}$$

Bidirectional Trade Intensity (BTI) metric: Weight of edge from A to B

$$\frac{\text{A to B exports}+\text{B to A exports}}{\text{Total A exports}+\text{Total B exports}}$$

Experiments

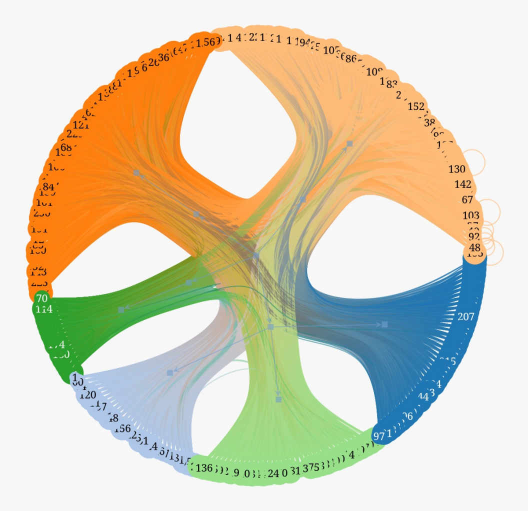


Figure: Output of minimize_nested_blockmodel_dl model after minimizing description length (using graph_tools on countries' data).

Experiments

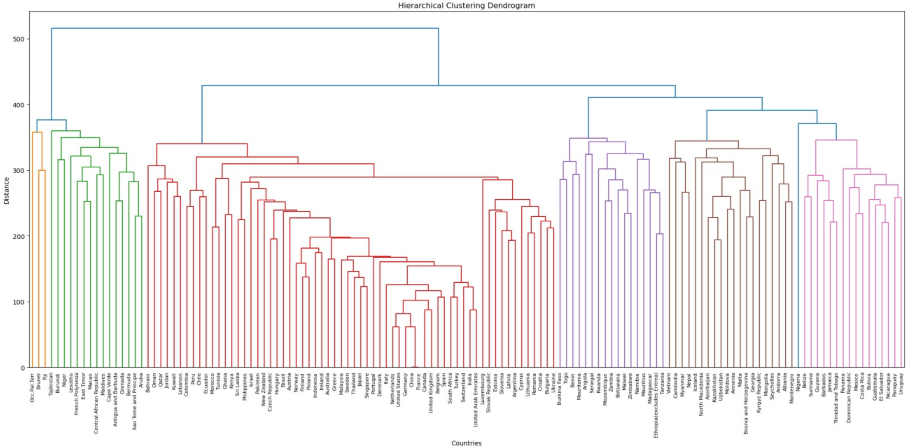


Figure: Lovain algorithm with harmonic mean metric for edge weights

Experiments

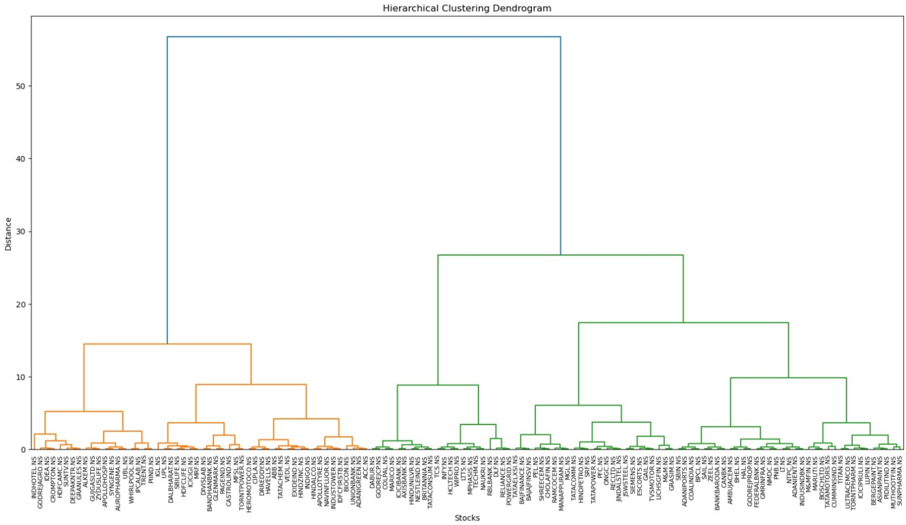


Figure: Dendrogram for correlation of stock price movement: Interval: 1 week, Duration: 1 year

Experiments

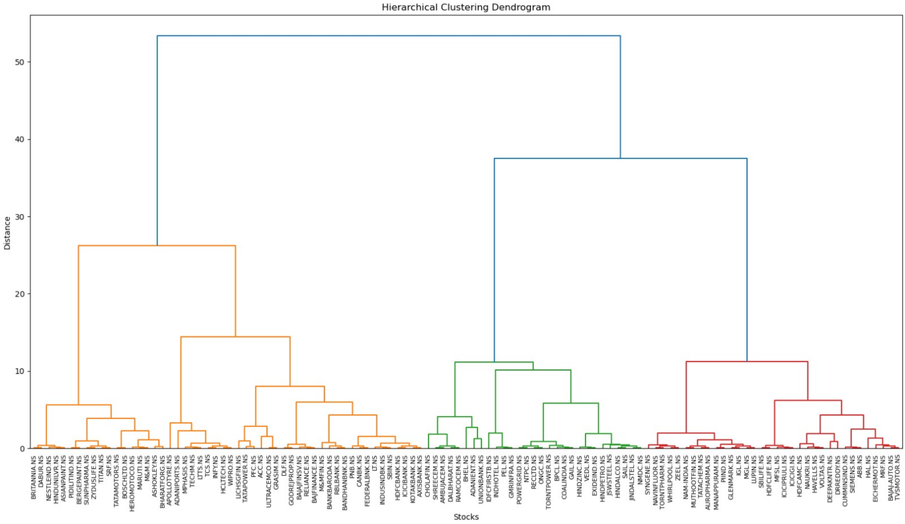
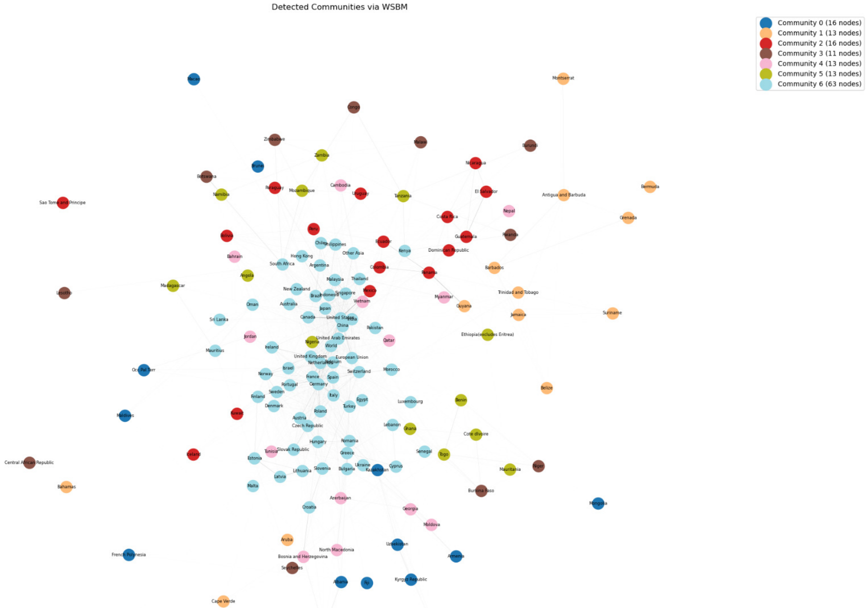


Figure: Dendrogram for correlation of stock price movement: Interval: 5 days, Duration: 5 years

Experiments


Detected Communities via WSBM

Stock Communities obtained

- POWERGRID.NS, NTPC.NS, RECLTD.NS, ONGC.NS, COALINDIA.NS, GAIL.NS, BPCL.NS
- RELIANCE.NS, BAJFINANCE.NS, BAJAJFINSV.NS, HDFCBANK.NS, ICICIBANK.NS, KOTAKBANK.NS, AXISBANK.NS, INDUSINDBK.NS, SBIN.NS, LT.NS, CANBK.NS
- ULTRACEMCO.NS, GRASIM.NS, ACC.NS, SHREECEM.NS, AMBUJACEM.NS, DALBHARAT.NS, RAMCOCEM.NS, ADANIENT.NS
- ADANIPORTS.NS, LICHSGFIN.NS, TATAPOWER.NS, PFC.NS, BHEL.NS
- TCS.NS, INFY.NS, HCLTECH.NS, WIPRO.NS, TECHM.NS, LTTS.NS, MPHASIS.NS
- HINDUNILVR.NS, ASIANPAINT.NS, NESTLEIND.NS, BRITANNIA.NS, DABUR.NS, PIDILITIND.NS, BERGEPAINT.NS, SUNPHARMA.NS
- APOLLOTYRE.NS, ASHOKLEY.NS, BHARATFORG.NS, MARUTI.NS, M&M.NS, HEROMOTOCO.NS, TATAMOTORS.NS

Challenges & Improvements

- Support for computation of probabilities of edges, in order to detect the anamolies we desired.
- Cleaner and elaborate datasets, providing **transactions** between countries.
- Volatility measures in market to determine interval of stock growth calculation.
- Robust implementations for Weighted Stochastic Block Method.

Contributions

Ishaq: Problem statement, clustering method.

Rohit: Experiments

Nagasai: Data scraping, Slides

Sathvik: Experiments

Shankar: Data processing

Data for import and export: world bank 2022 data.

- Stock data: Yahoo Finance.
- graph_tool: WSBM