GROUP H

PASTATION

DATA SCIENCE – COMP0047

Nayurvi Munogee
Tyler Supersad
Kumaraguru Thambidura
Sharanya John Alex
Pengrui Huang
Jianian Zheng
Abdullah Alotaishan
Eric Yew
Muhammad Zaharan
Elliot Ward
Artem Arshakyan

Fin_USA_SectorEquities: Dataset for Sector-Level Stock Analysis

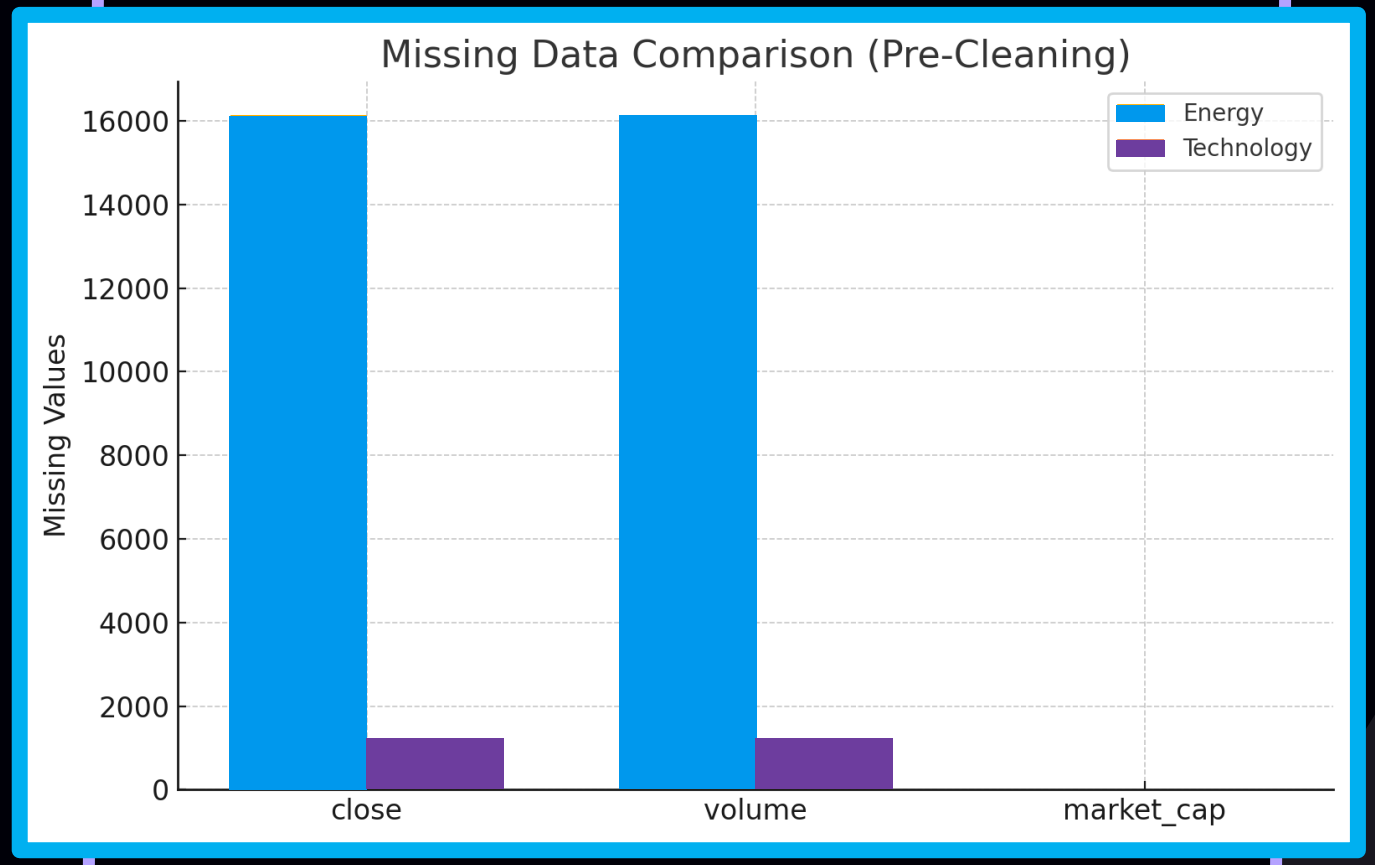
01 ☑ Two distinct datasets: Energy and Technology sectors.

02 ☑ Daily ticker-level time series from 2019 to 2024.

03 ☑ Three files covered per sector: close (prices), volume, and market cap.

04 ☑ ~58–60 tickers per sector, ~1,260 daily records per ticker.

DATASET CHALLENGES !



COMPLETENESS

Several tickers in Energy had 100% missing close and volume data, making them unusable.

Partial missing data in tickers like CHK, GPOR, CA, and CVG will require interpolation or removal.

SPARSITY

IMBALANCE

Some tickers had near-perfect data, while others had large gaps, complicating model fairness.

Missing close and volume data could distort return and volatility calculations if not handled properly.

RELIABILITY

PREPROCESSING (PART 1)

01 **Data Merging:** Combined close, volume, and market cap files on date and ticker using outer joins.

02 **Price & Volume Interpolation:** Filled missing close and volume values using linear interpolation per ticker for smooth continuity.

03 **Market Cap Imputation:** Used a 5-day rolling median per ticker to impute missing market_cap values without distorting trends.

MR. CLEAN YOUR DATA

PREPROCESSING (PART 2)

04

HIGH-MISSING TICKER REMOVAL

Dropped tickers with over 30% missing close prices to ensure modeling reliability.

05

CRITICAL ROW FILTERING

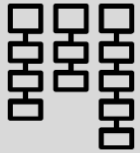
Removed remaining rows missing any essential field: close, volume, or market_cap.

06

FEATURE NORMALIZATION

Applied z-score scaling per ticker to standardize close, volume, and market_cap.

07



FINAL SORTING & DEDUPLICATION

Chronologically sorted data and dropped duplicate date-ticker pairs for consistency.

FEATURE SELECTION (PART 1)

60 + 47 = 107 features (stocks)

3 benchmark stocks
Sharpe ratio/return

Compute the top 5 winner stocks and loser stocks

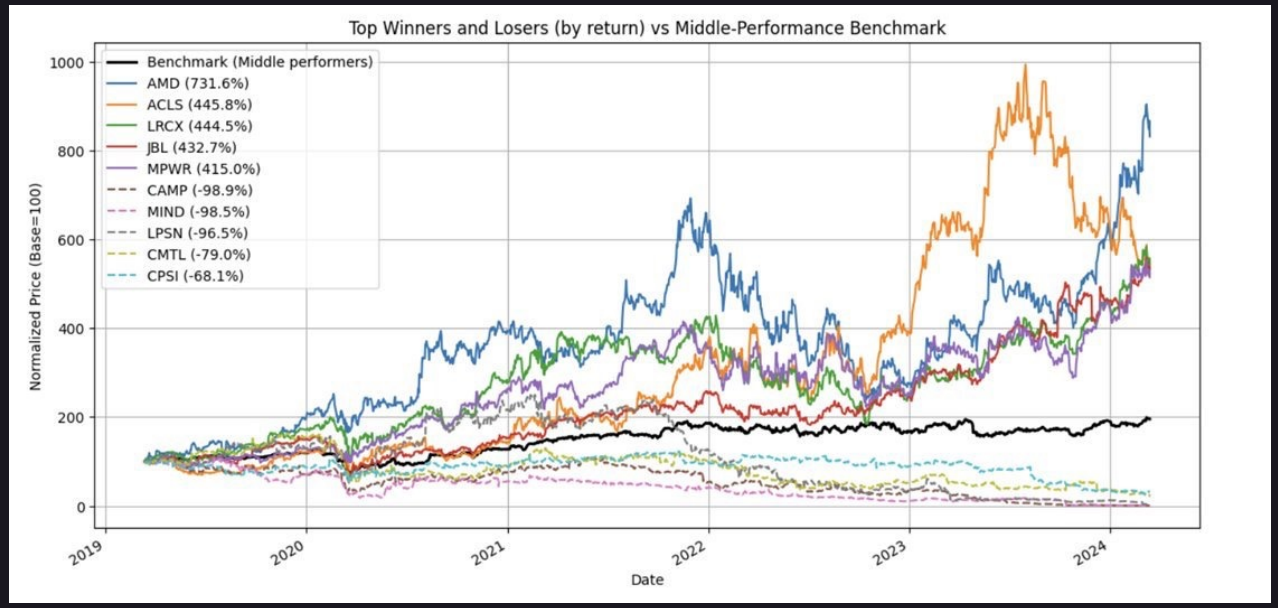
78% dimension reduction (technology)
72% dimension reduction (energy)

Energy Benchmark
Return: 50.69%
Volatility: 0.47

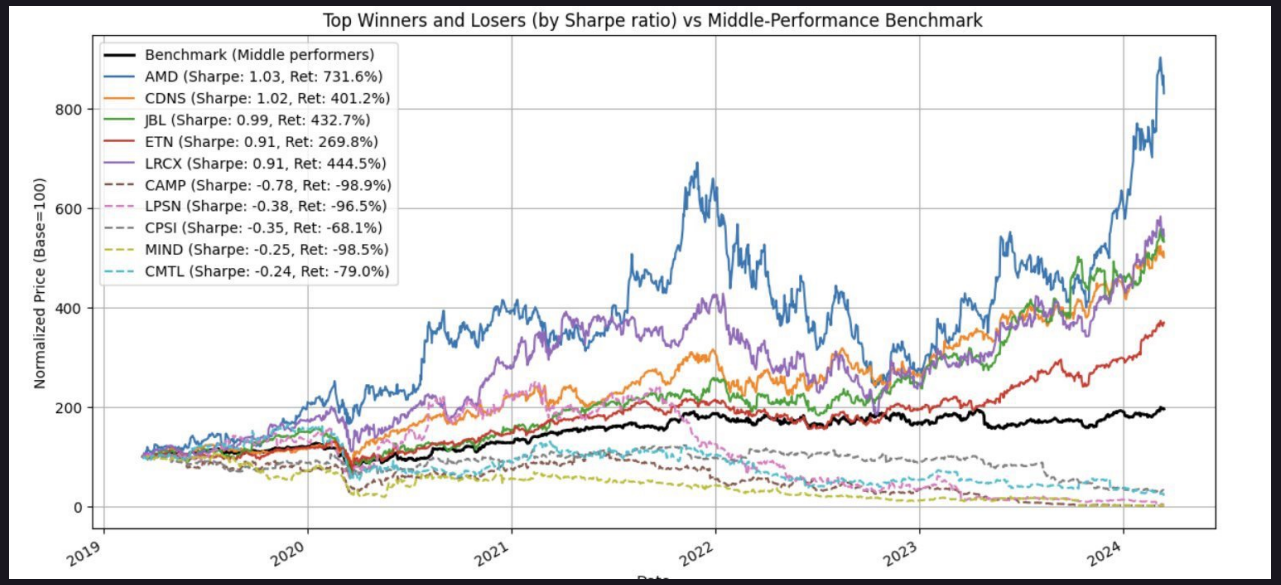
Technology Benchmark
Return: 95.78%
Volatility: 0.36

FEATURE SELECTION
(PART 2)

BY RETURN



BY SHARPE



WINNERS

BY SHARPE: AMD, CDNS, JBL, ETN, LRCX

BY RETURN: AMD, ACLS, LRCX, JBL, MPWR

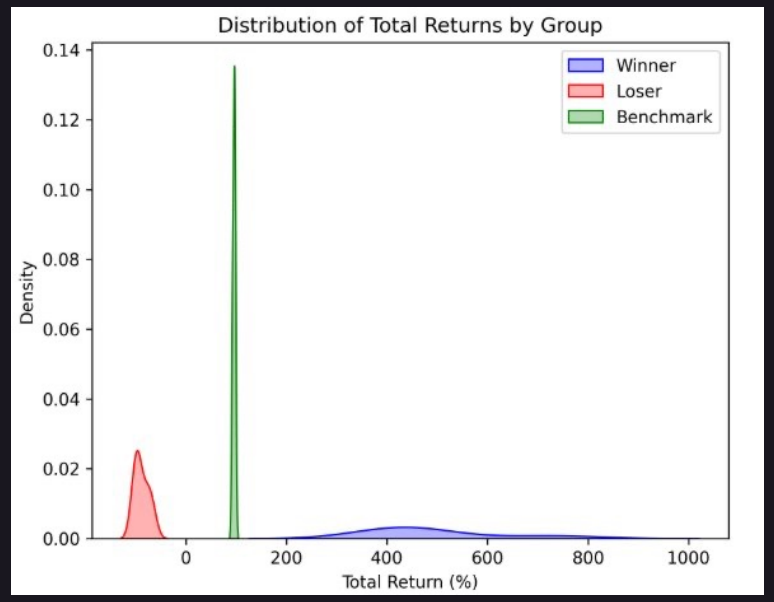
LOSERS

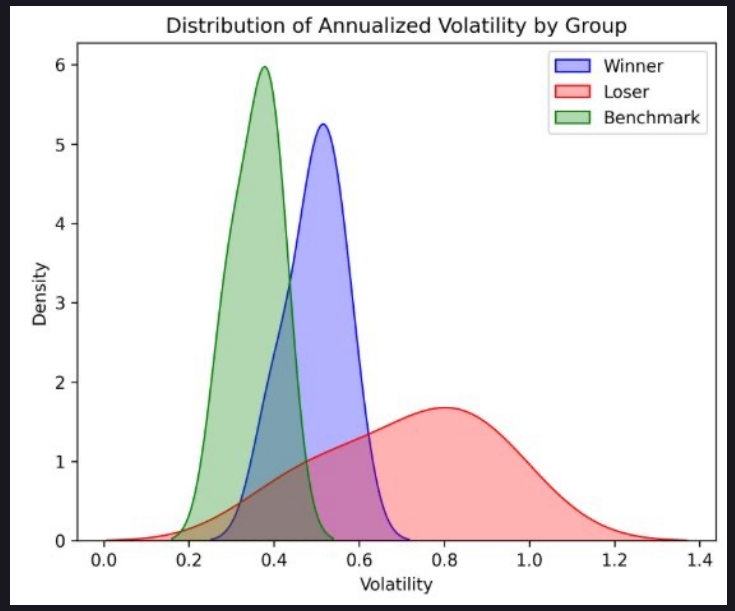
BY SHARPE: CAMP, LPSN, CPSI, MIND, CMTL

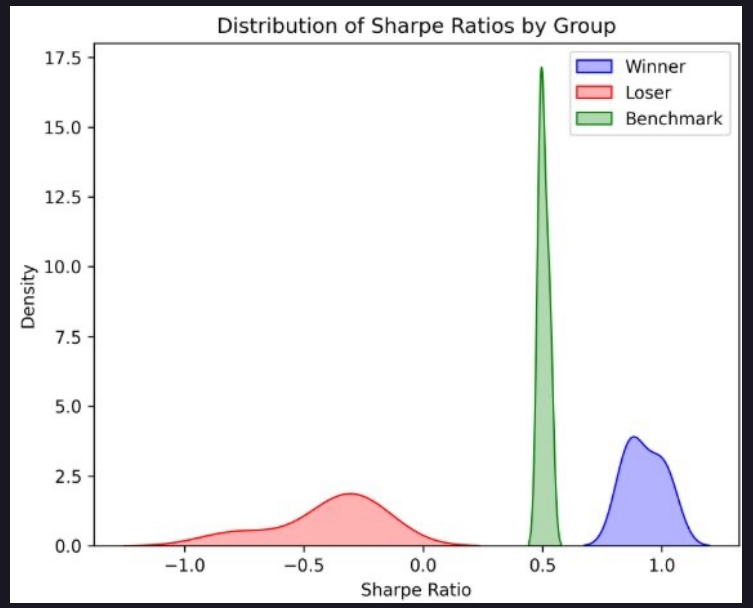
BY RETURN: CAMP, MIND, LPSN, CMTL, CPSI

WINNER-LOSER DISTRIBUTIONS

(PART 1)







TOTAL RETURNS — Higher upside for potential winners

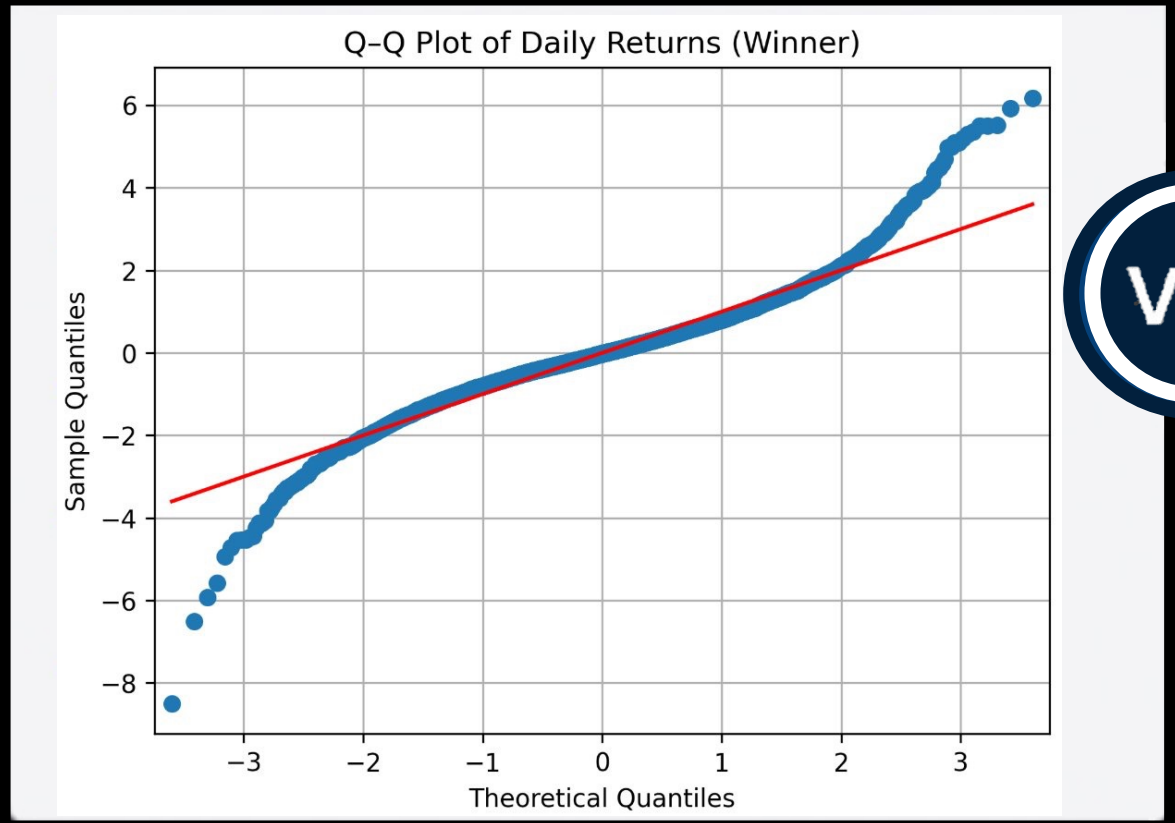
ANNUALIZED VOLATILITY — Winners show greater risk exposure

SHARPE RATIOS — Winners outperform on risk-adjusted basis

WINNER-LOSER DISTRIBUTIONS

(PART 2)

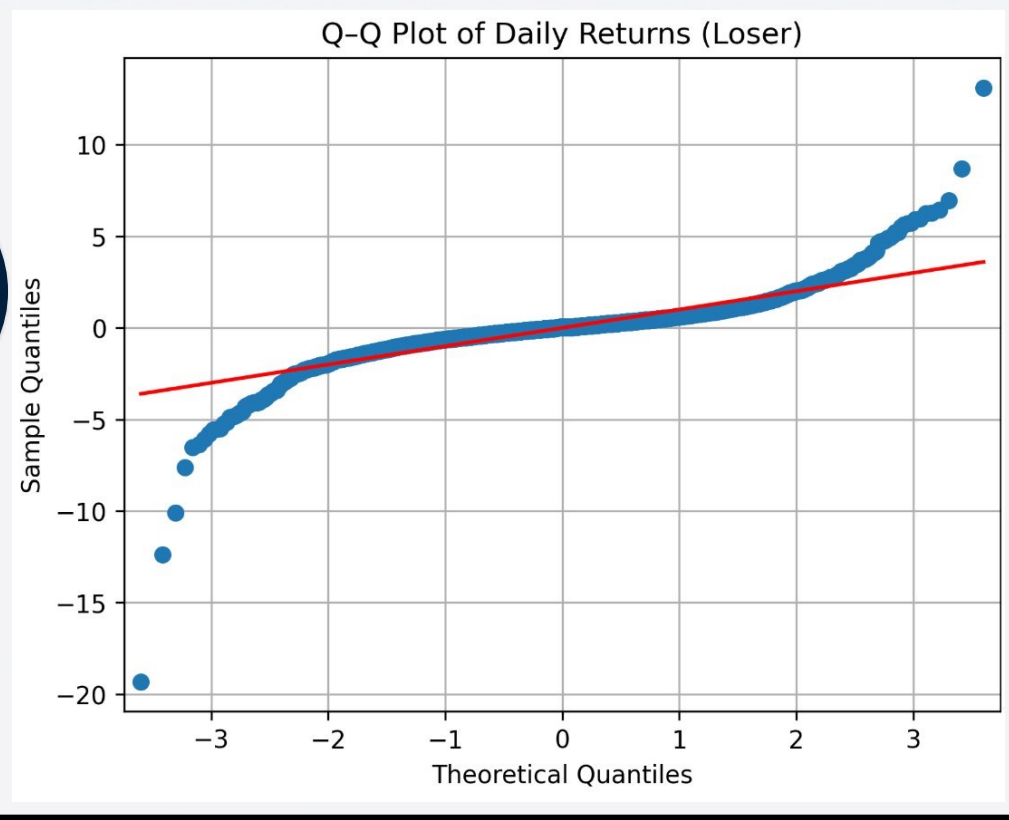
WINNER Q-Q PLOT



- The left tail dips.
- The right tail goes well above the line, reflecting those high daily gains.

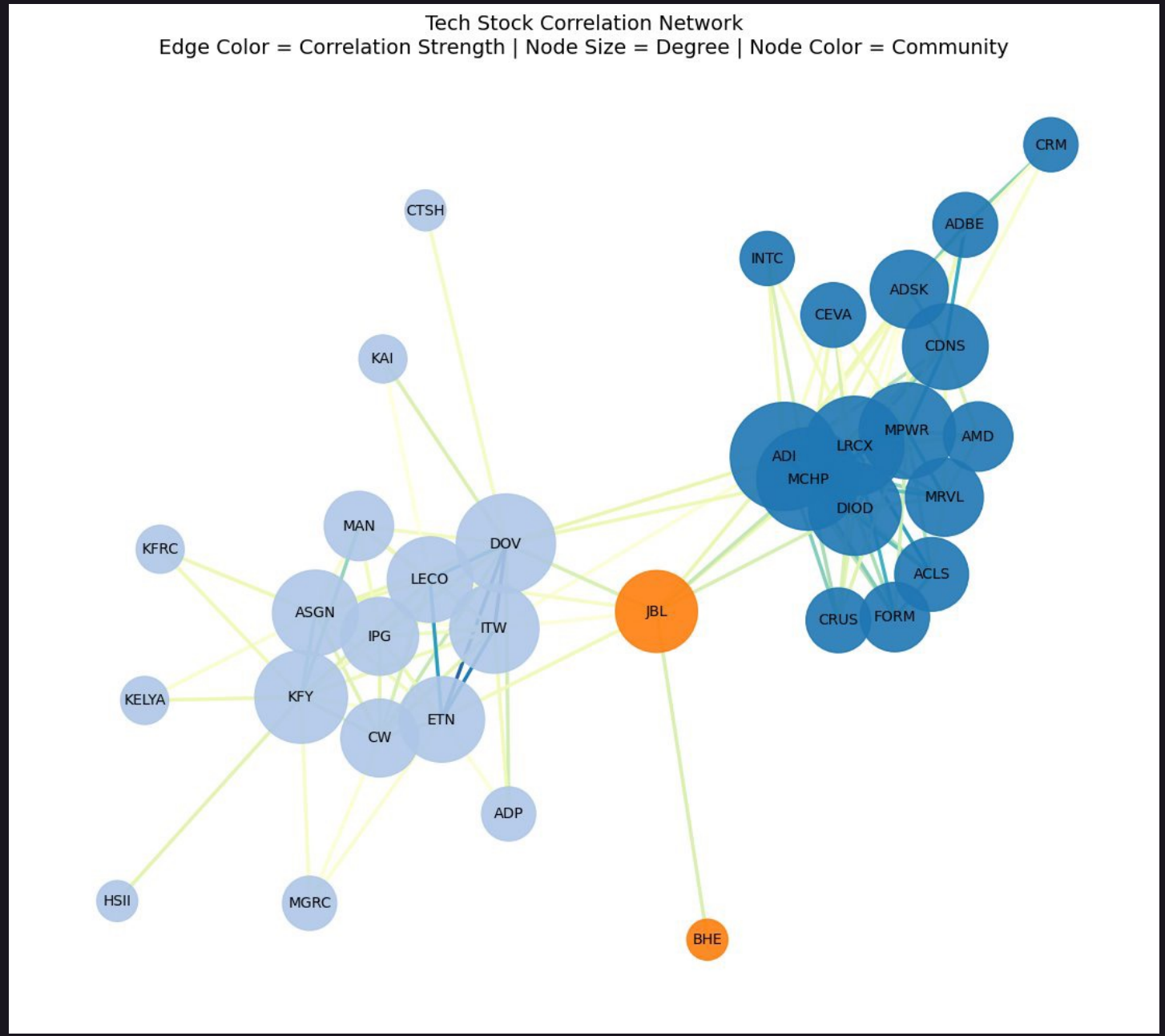
V/S

LOSER Q-Q PLOT

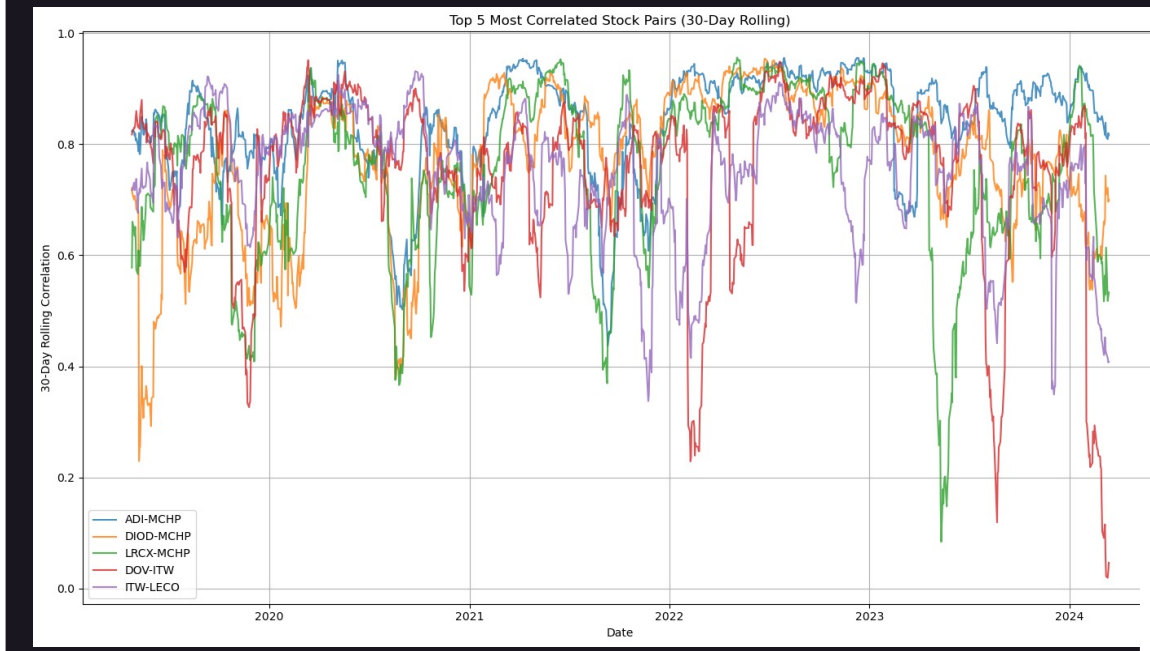


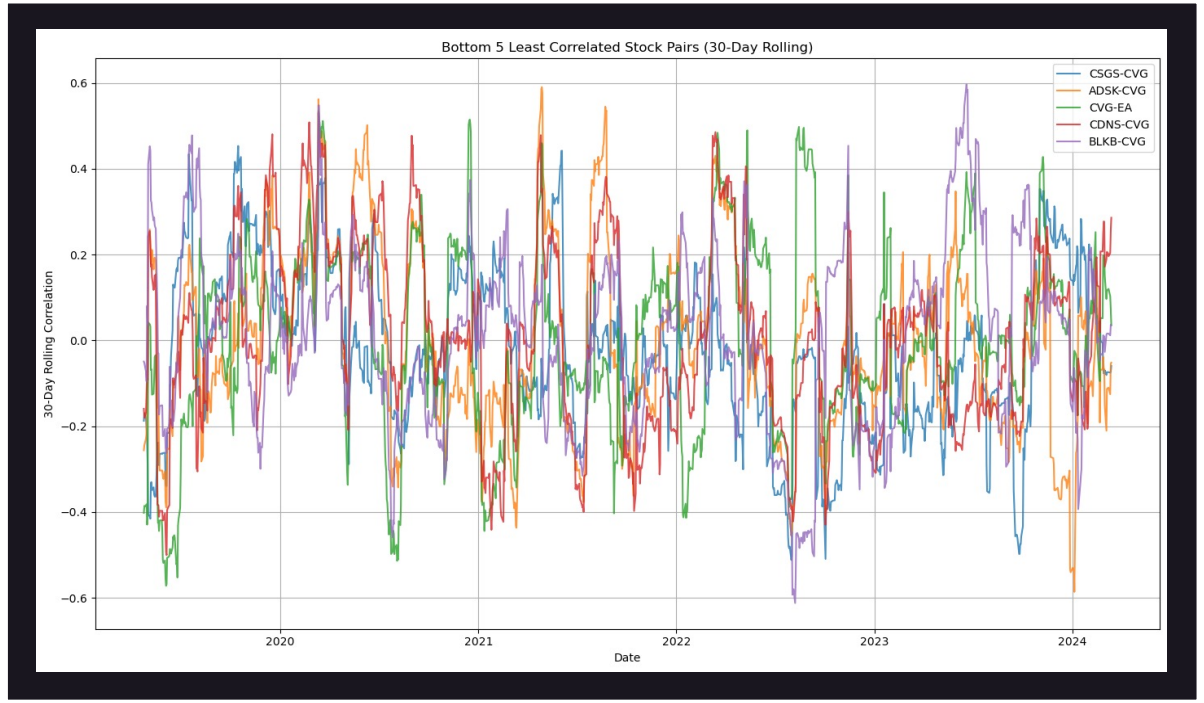
- Extreme dip occurs in the lower tail.
- The right tail extends above the line.

CORRELATION NETWORK GRAPH

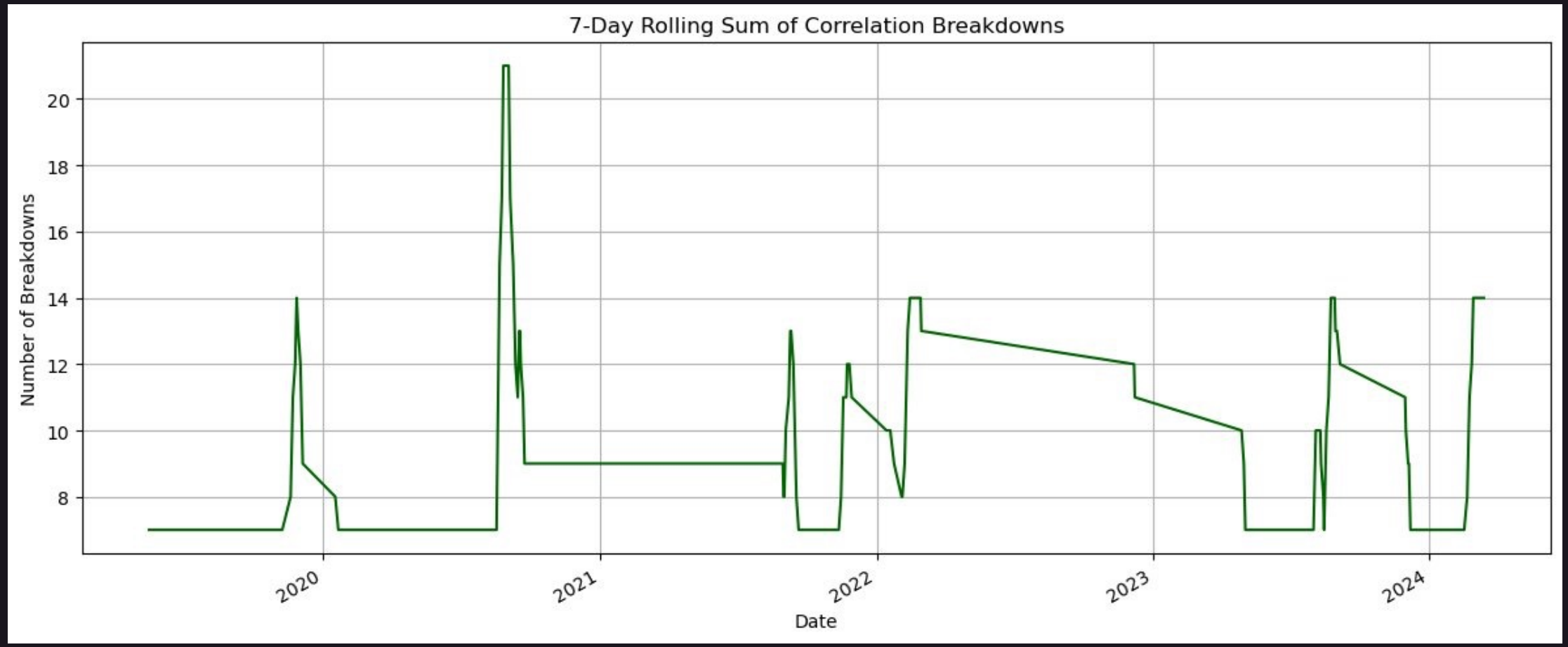


ROLLING CORRELATION

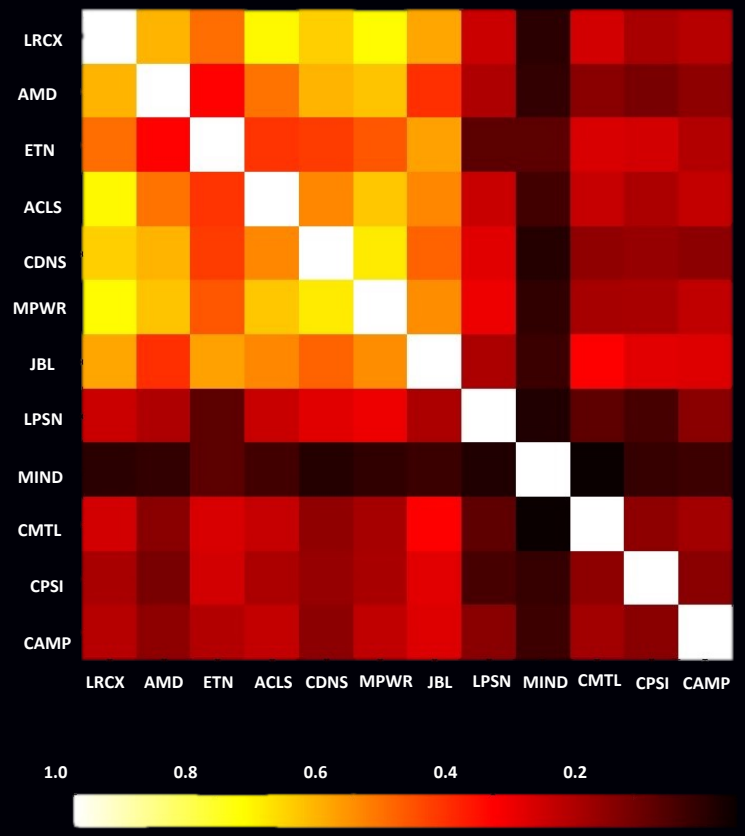




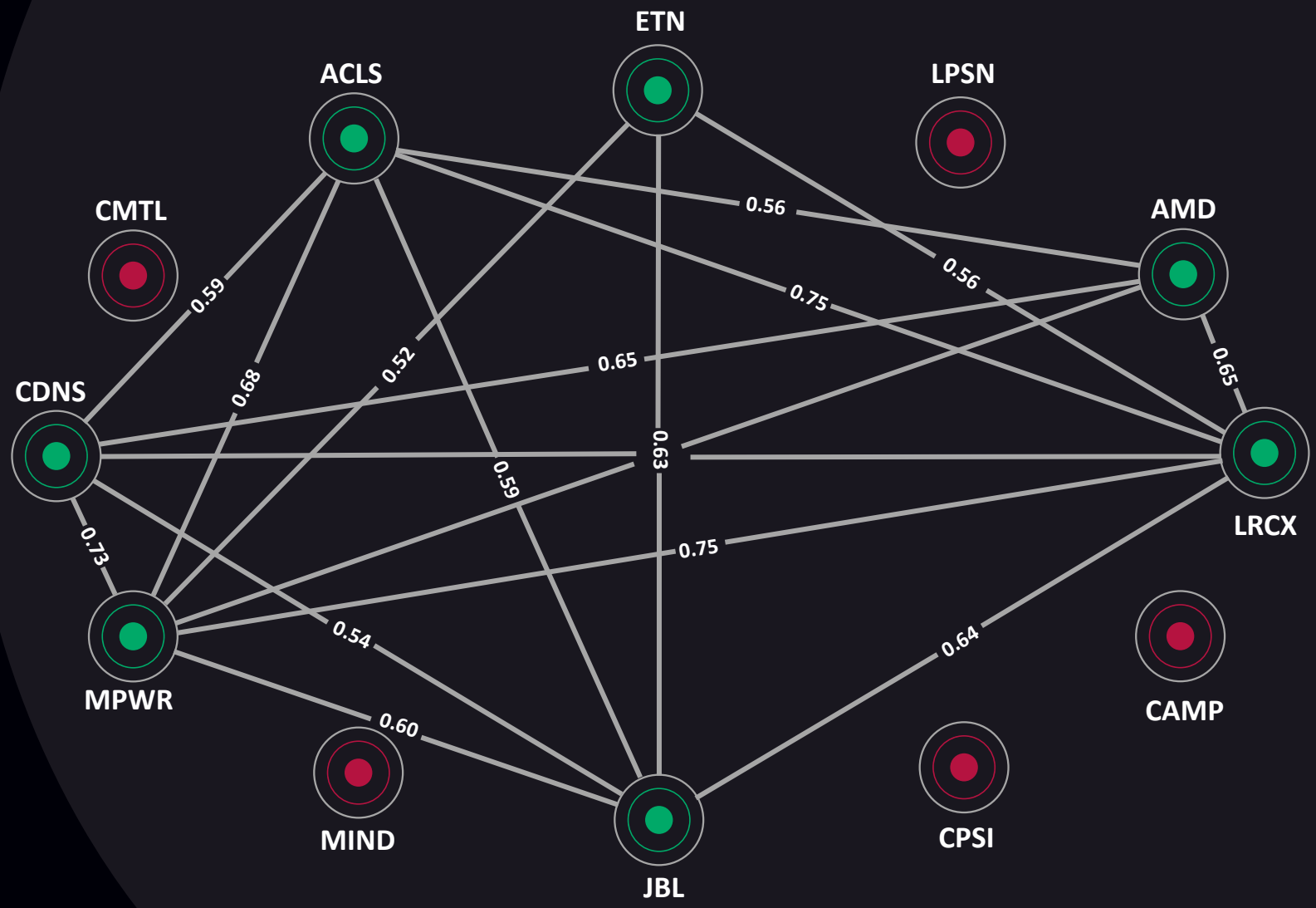
BREAKDOWN DETECTION

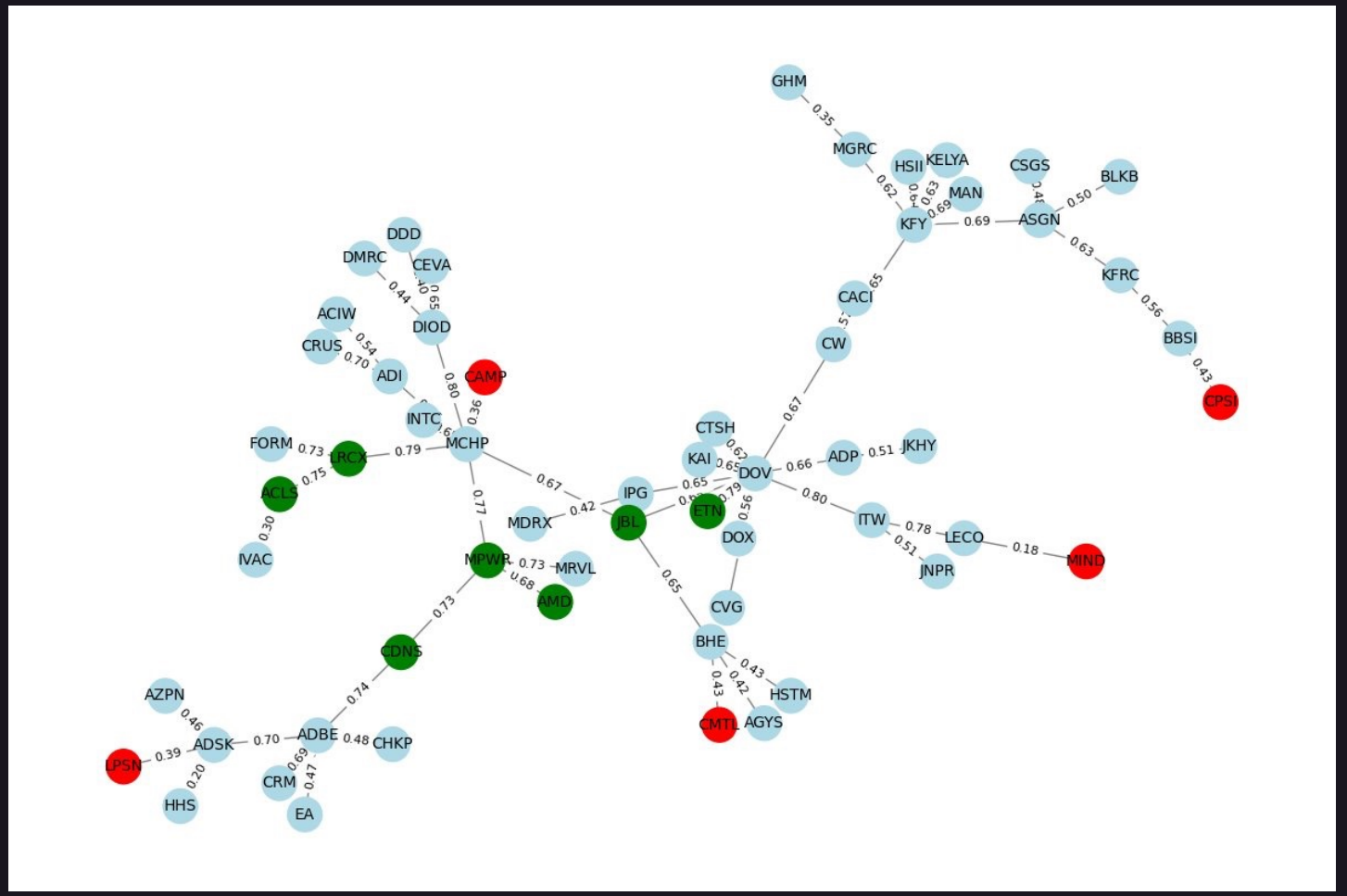


CORRELATION MATRIX OF BENCHMARK, WINNERS AND LOSERS



NETWORK GRAPH FOR WINNERS AND LOSERS WITH THRESHOLD 0.5

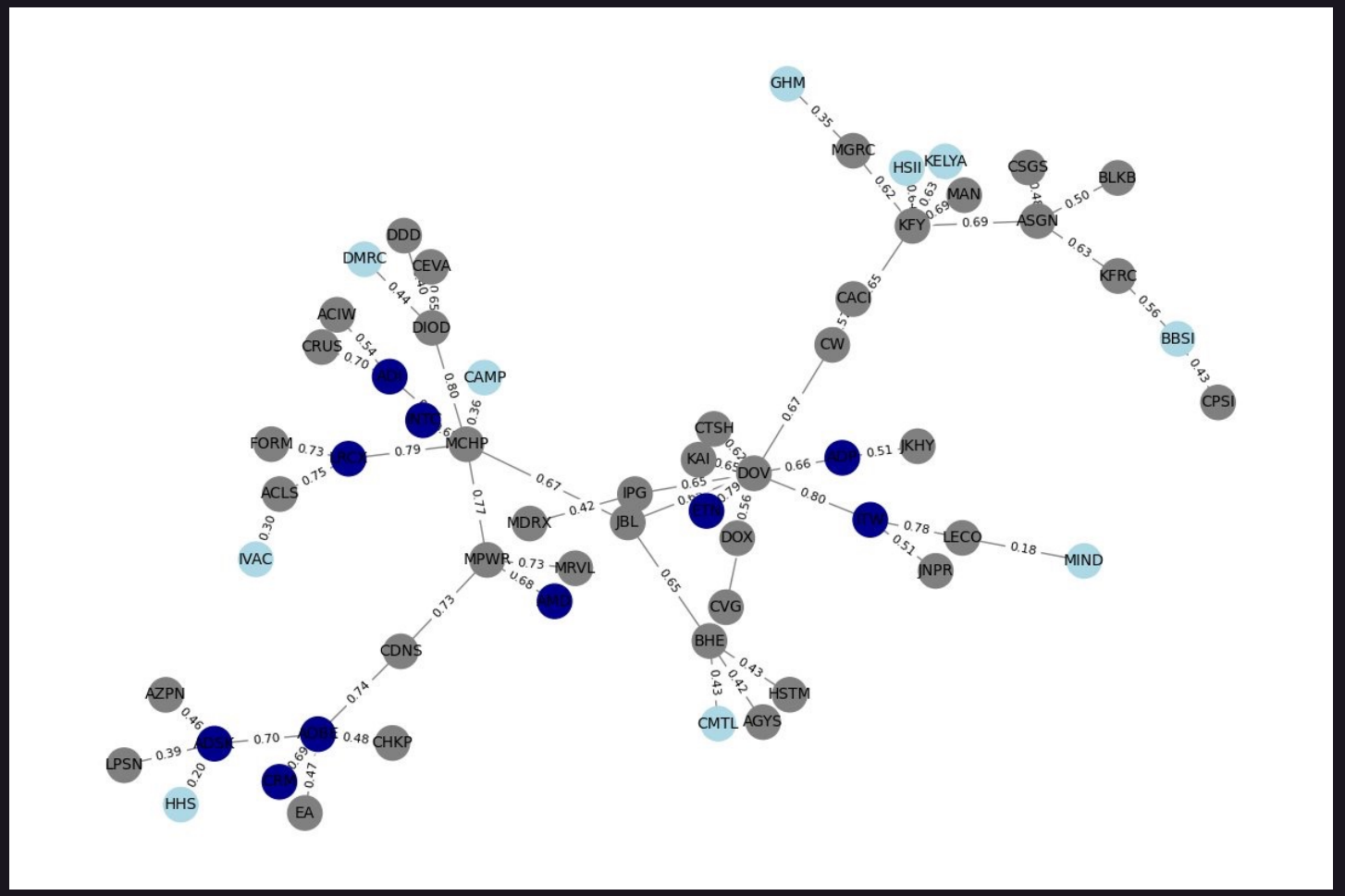




MAXIMUM SPANNING TREE

Green = Winners

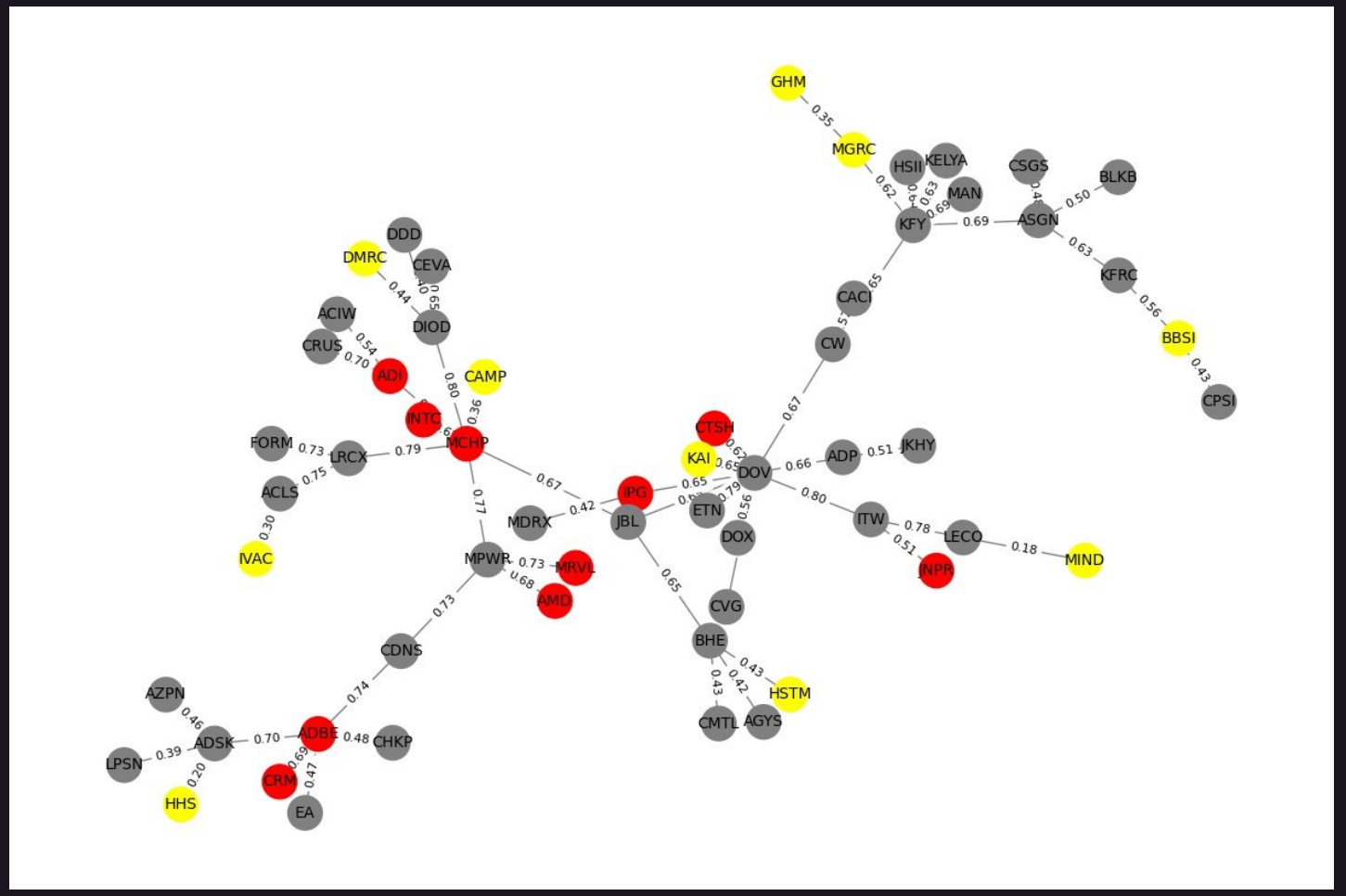
Red = Losers



MAXIMUM SPANNING TREE

Dark Blue = High Market Cap

Light Blue = Low Market Cap



MAXIMUM SPANNING TREE

Red = High Volume

Yellow = Low Volume

THANK YOU FOR LISTENING