Potential Anomalies in Congressman Stock Trading

A Needle in a Data Haystack: Introduction to Data Science - 67978

Avinoam Nukrai, nukrai.avinoam@mail.huji.ac.il, 206997132, avinoam_nukrai
Dean Tahory, dean.tahory@mail.huji.ac.il, 319045837, dean.tahory
Uriel Ron, uriel.ron@mail.huji.ac.il, 318164118, urielron

The Hebrew University of Jerusalem

1. Introduction

In recent years, public attention has turned to the financial activities of U.S. lawmakers, with congressional stock trading raising concerns about unfair advantages and conflicts of interest. Lawmakers often have access to non-public information and make decisions that affect markets, which has fueled debates over whether they should be allowed to trade at all. While transactions are disclosed, identifying suspicious timing or coordinated trading remains challenging and requires systematic methods. Our project addresses this gap by focusing on two hypotheses. First, members of Congress may exploit non-public information from their roles to make profitable trades. Second, groups of politicians may act in coordination, showing similar trading patterns around the same time. To investigate these possibilities, we concentrate on about 90 members who executed stock purchase transactions during 2020. The analysis is based on multiple datasets, combining congressional trading disclosures, market data, and additional information collected through search-enabled LLM systems such as Gemini. Using these data sources, we construct networks of politicians and transactions and analyze them with graph-theoretic properties studied in class, which allows us to systematically detect unusual and potentially suspicious trading activity in both financial and political contexts.

2. Research Focus

We expect to find signs of insider information use among members of influential committees, as well as evidence of coordinated trading within smaller groups of lawmakers. By contrast, we do not anticipate large-scale coordination across Congress or consistent abnormal profits among all members. Beyond these expectations, we expand the analysis to explore links between politicians and sponsoring companies, assess potential gains tied to these connections, and investigate whether committee memberships correlate with trading outcomes. With the support of Gemini, a search-capable system that integrates retrieval with LLM reasoning, we generate structured politician profiles, identify potentially suspicious transactions, and detect collaboration signals across members. Our overarching aim is to surface as many meaningful and interesting patterns as possible at the intersection of politics and financial markets, shedding light on behaviors that may warrant closer scrutiny.

3. Data

Our data construction process unfolded in three main stages that together produced the final datasets we used. We began with Kaggle's "congressional-trading-inception-to-march-23" collection, which contains over 50,000 transactions disclosed by more than 1,000 U.S.

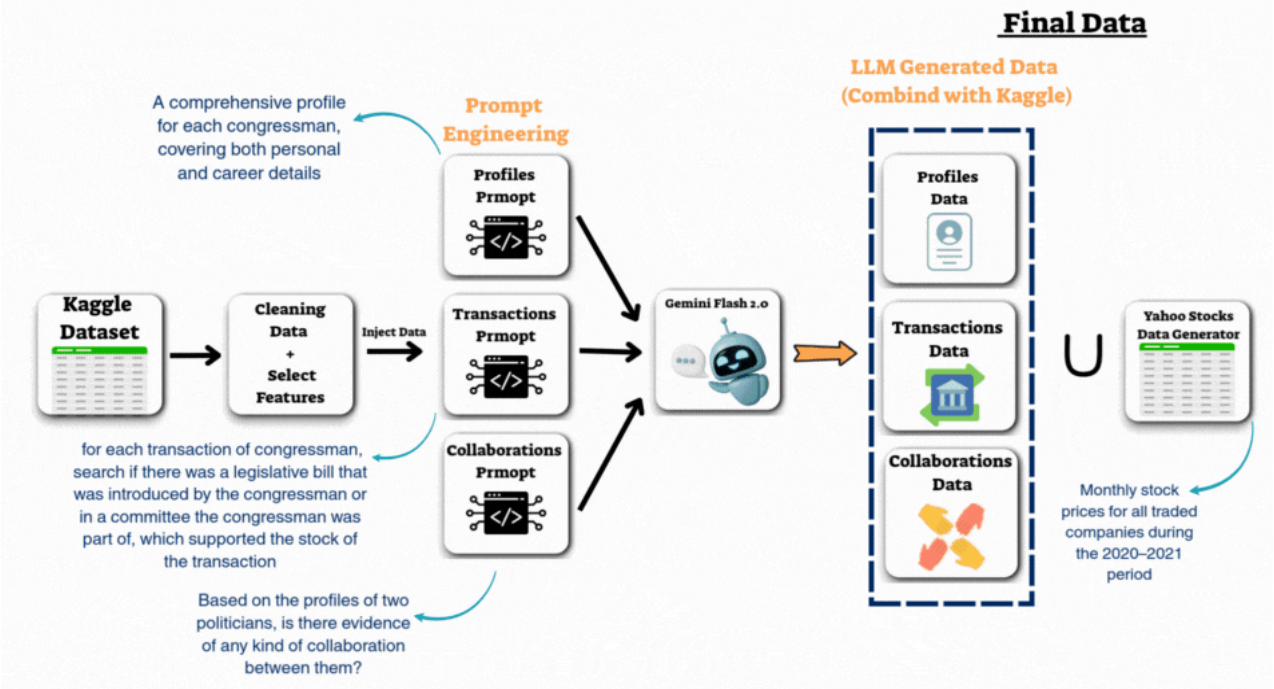
politicians between 2004 and March 2023 (about 50 MB). Each record includes politician details, transaction information, and political context such as committee memberships. From this source we focused on 2020–2021, due to resource limitation, we filtered only purchase transactions, limited the sample to forty trades per politician, and after preprocessing obtained ninety representatives and senators who purchased stocks in 2020.

Next, we used Gemini search capabilities to scrap data regarding the connections between the politician and the stock's company and data about the politician itself. Specifically, we used two prompts. The first - profile prompt, extracts factual, biographical data about politicians, while the second, the transaction analysis prompt, investigates potential conflicts of interest

For the profile building, we asked for the following fields: full name, sex, party, birth date, state, city, university, seniority in congress,committees and companies sponsorships.

For each transaction, we recorded several fields in order to capture both political and financial context. The subcommittees field lists the specific congressional subcommittees the politician belongs to, defining their area of legislative influence. The supporting_agenda field is a boolean value that indicates whether the politician's stated goals and policies align with the company's business sector, while supporting_agenda_explanation provides a short summary explaining why the agenda does or does not support the industry. The direct_legislative_connection field is another boolean value that flags whether there is a provable, direct link between the politician's official legislative work and the company, and direct_legislative_connection_proof specifies the concrete evidence for such a connection, such as an entry in the Congressional Record or an official report. Similarly, the subcommittee_decision field checks whether a decision benefiting the company was made in the politician's subcommittee within a five-month window around the trade, and the subcommittee_decision_proof field supplies supporting documentation, such as a hearing transcript, to validate such a decision.

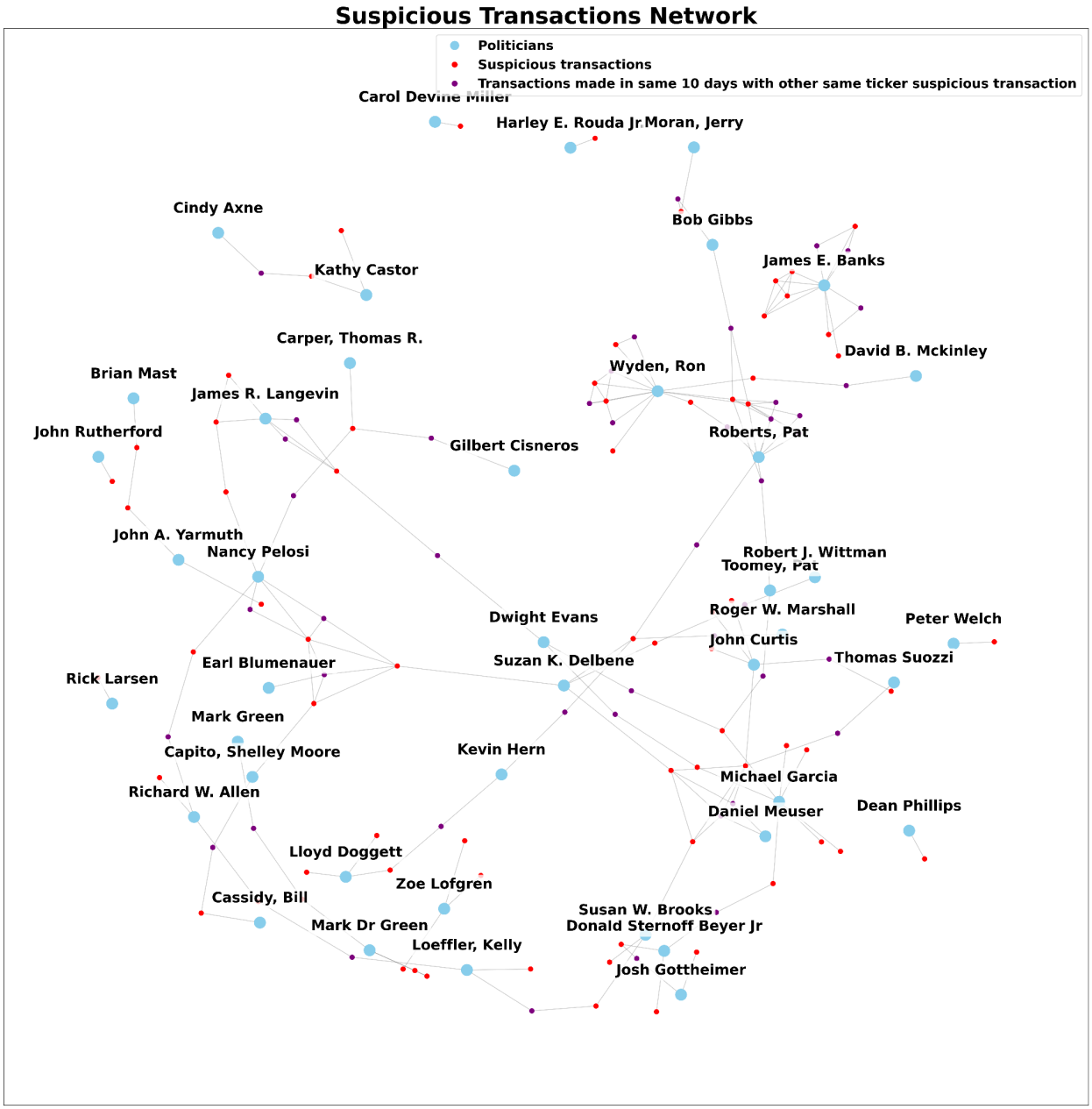
Finally, we supplemented this information with financial data collected via the Yahoo Finance API in Python, which provided monthly stock prices for every company traded by our sample in a one-year window around the transactions (about 250 KB). Covering 155 unique tickers, this dataset enabled us to measure market performance and compute relative gains or losses for each transaction. After cleaning, integrating, and enriching these sources, we consolidated them into three final datasets that served as the foundation of our analysis: transactions_with_analysis.csv (6.4 MB), containing the enriched transaction records; politician_profiles.csv (29 KB), containing structured lawmaker attributes; and stock_prices.csv (250 KB), containing historical price data for traded companies. These three files represent the definitive working datasets of our project. [Fig 1].



4. Methodology

We now turn to the core of our solution. Using graph analytics, community detection, and centrality measures, we analyze suspicious trading activity among U.S. lawmakers. The results are presented through a sequence of network visualizations and statistical evaluations, moving step by step from individual suspicious trades to broader collaboration patterns, sponsorship ties, and influence rankings.

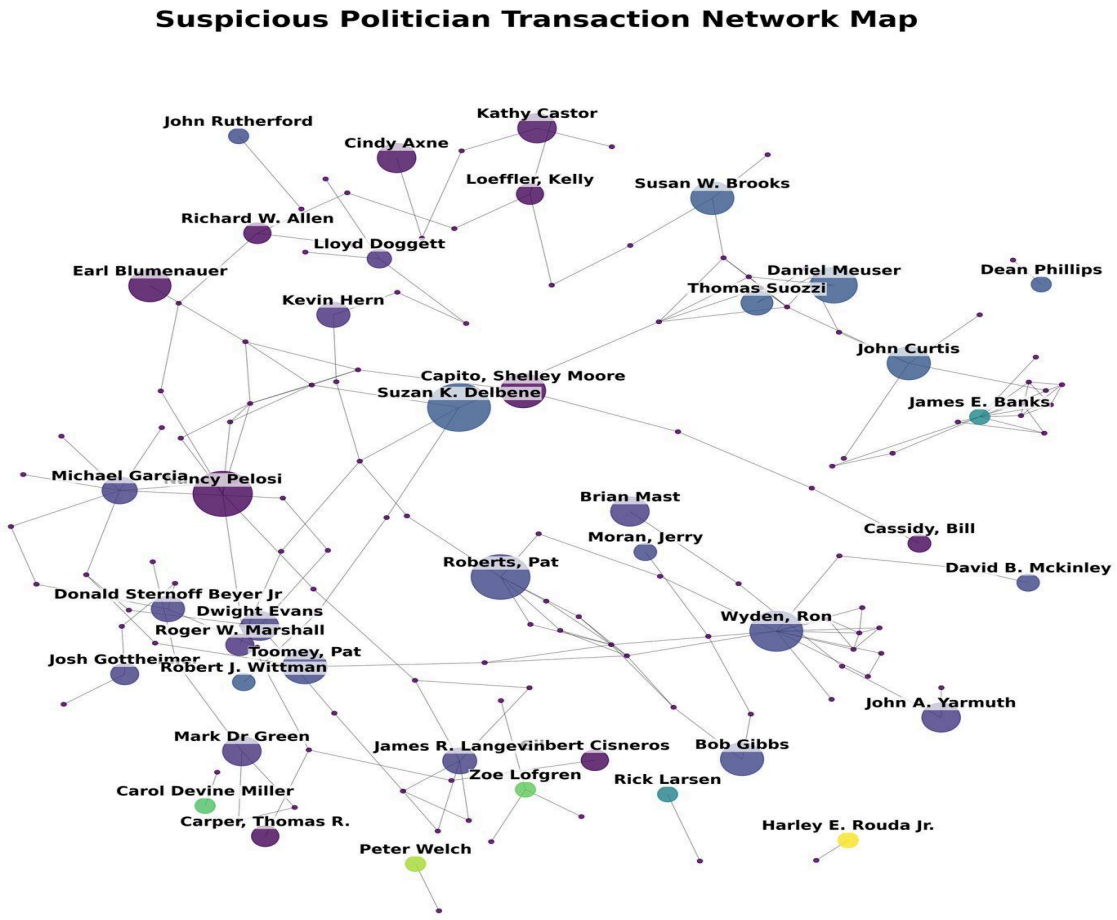
We begin by defining that each politician is a node and each transaction is also a node. Politicians are connected to their transactions, forming a bipartite structure. From this graph, we include only transactions flagged as either subcommittee_decision=True or direct_legislative_connection = True which says that the politician voted in a committee or subcommittee in favor of a decision that directly or indirectly benefited a company, all others are excluded. We call the remaining trades *suspicious*. To capture temporal proximity, every suspicious trade is connected to all other trades (even non-flagged ones) if they involve the same ticker and occur within a 10-day window. This projection allows us to build a politician and transactions network $G = (V, E)$, where edges represent overlapping trades tied to at least one suspicious event. [Fig 2].



In the above visualization, red nodes represent transactions that we confidently classify as suspicious, while purple nodes are not necessarily fraudulent but remain potentially problematic and worth investigating further. Instead of issuing a direct LLM query for each transaction to ask explicitly whether it constitutes fraud, or manually verifying them one by one, our approach leverages the analytical methods studied in class. By applying graph-based tools such as community detection, clustering, and centrality analysis, we aim to uncover meaningful patterns and generate data-driven insights about potentially coordinated or unethical behavior, without relying solely on binary classification. For example, Congressman Parker promoted legislative initiatives related to green energy that indirectly benefited Tesla, while Pelosi purchased Tesla stock within 10 days of Parker's trade. Although Pelosi was not part of the relevant subcommittee, their long-standing ties, shared party affiliation, and

history of joint legislative activity suggest that her trade may still raise suspicion. Cases like this demonstrate why network-based analysis is critical, as it captures indirect links and potential collaboration beyond explicit committee assignments.

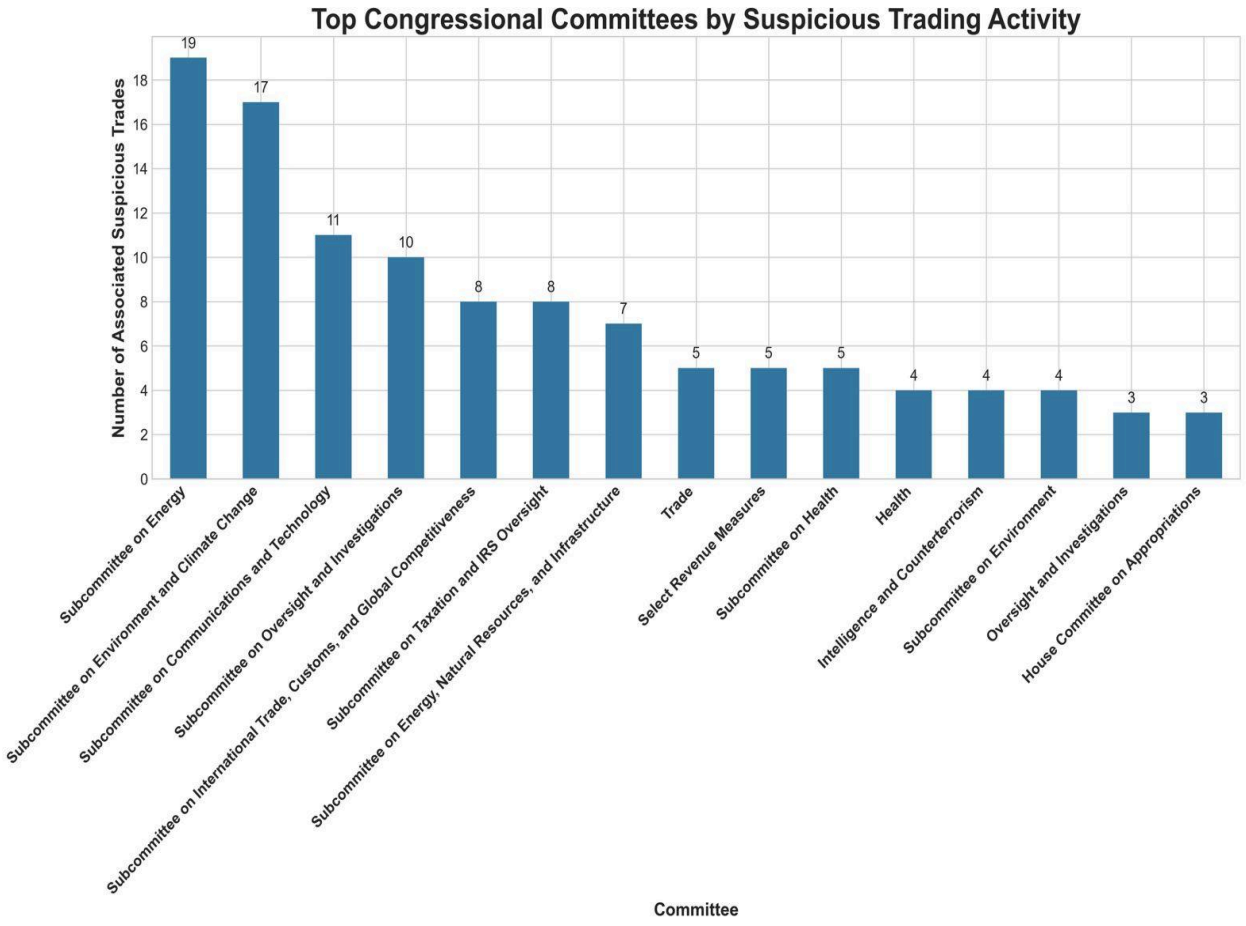
Next we detect communities with the Louvain algorithm. In the next graph, each large node represents a politician, with node size reflecting the weighted degree, meaning the number of overlapping suspicious trades they are involved in. Colors correspond to communities detected by the Louvain algorithm. The edges represent connections formed when two politicians traded the same stock within a 10-day window, with at least one flagged as suspicious. The smaller nodes connected to the politicians represent the transactions themselves. Each transaction acts as a bridge linking multiple politicians who participated in it. When several politicians share connections to the same transaction nodes, this indicates shared trading patterns and the possibility of indirect coordination.Communities are formed by linking trades made in the same stock within a short time frame. Therefore, a community represents a group of politicians who showed a coordinated, suspicious trading pattern in the same financial assets. [Fig 3]



This visualization highlights not only the central politicians but also communities of politicians

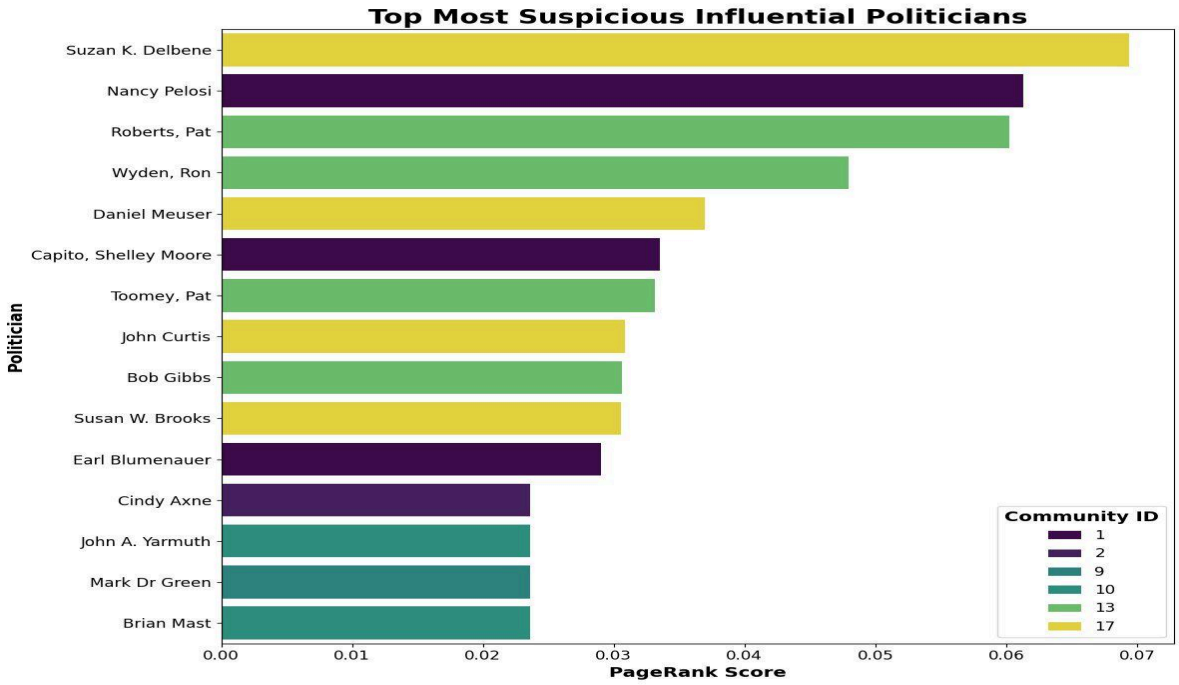
engaged in similar suspicious transactions, providing an additional perspective on collective suspicious trading activity.

To further understand why these suspicious overlaps emerge, we examine the congressional committees to which these politicians belong. The following chart shows the top committees ranked by the number of suspicious trades associated with their members [Fig 4].

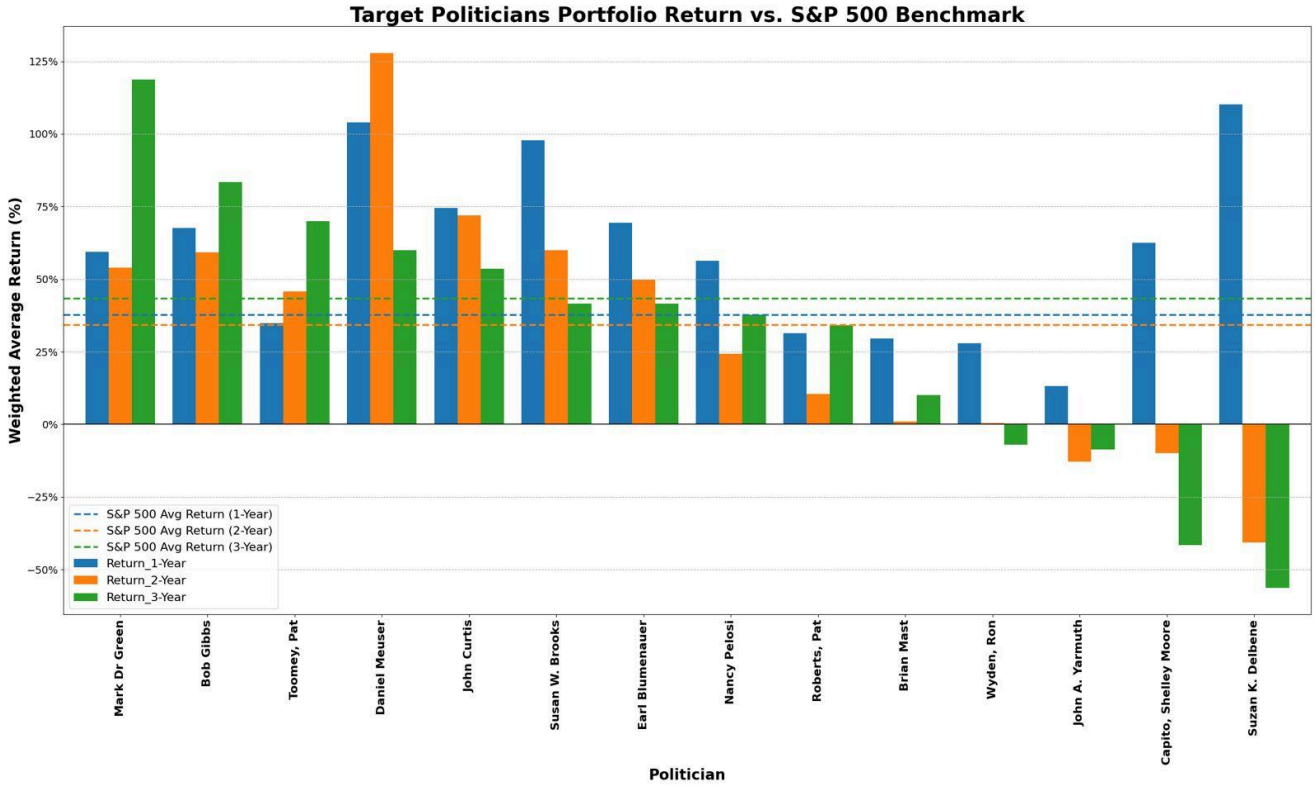


The results reveal that committees with jurisdiction over sectors like Energy, Environment and Climate Change, and Technology are consistently linked with the highest volumes of suspicious trades. This is not coincidental: these committees oversee industries that are both highly regulated and highly profitable, giving their members access to sensitive legislative information with clear financial implications. By linking trading activity back to committee memberships, we gain additional evidence that political responsibilities and financial incentives are tightly intertwined.

The next step was to compute PageRank, not to identify politically influential figures in terms of network structure, but rather to rank politicians by the frequency of their suspicious trading activity. The results revealed that a small group of politicians, including Pelosi and DelBene, consistently appeared at the top because they were involved in the highest number of suspicious transactions. [Fig 5].

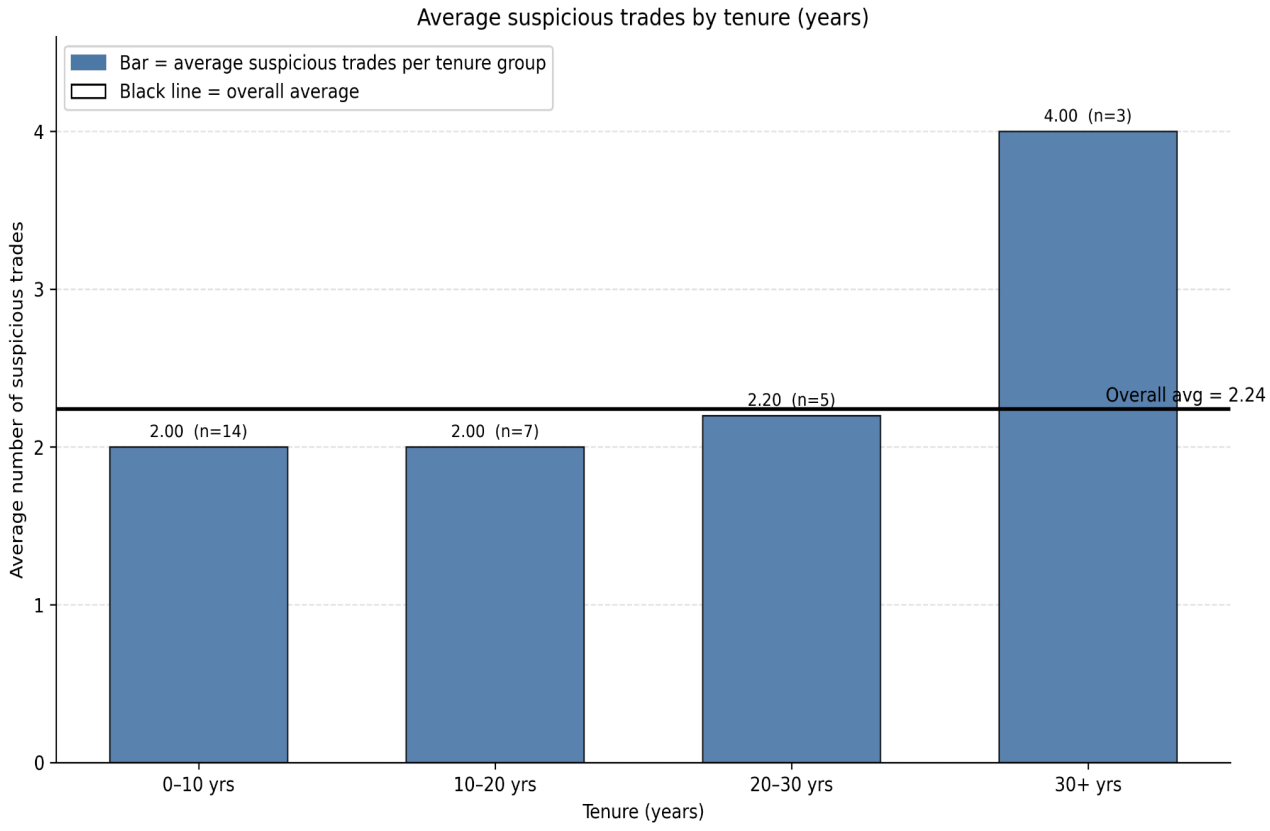


Now, we want to examine the potential returns of the suspicious politicians over time intervals of one, two, and three years, based on the total of their suspicious transactions and compared against their actual prices retrieved from Yahoo. As a baseline, we use an index that tracks the S&P 500. [Fig 6].



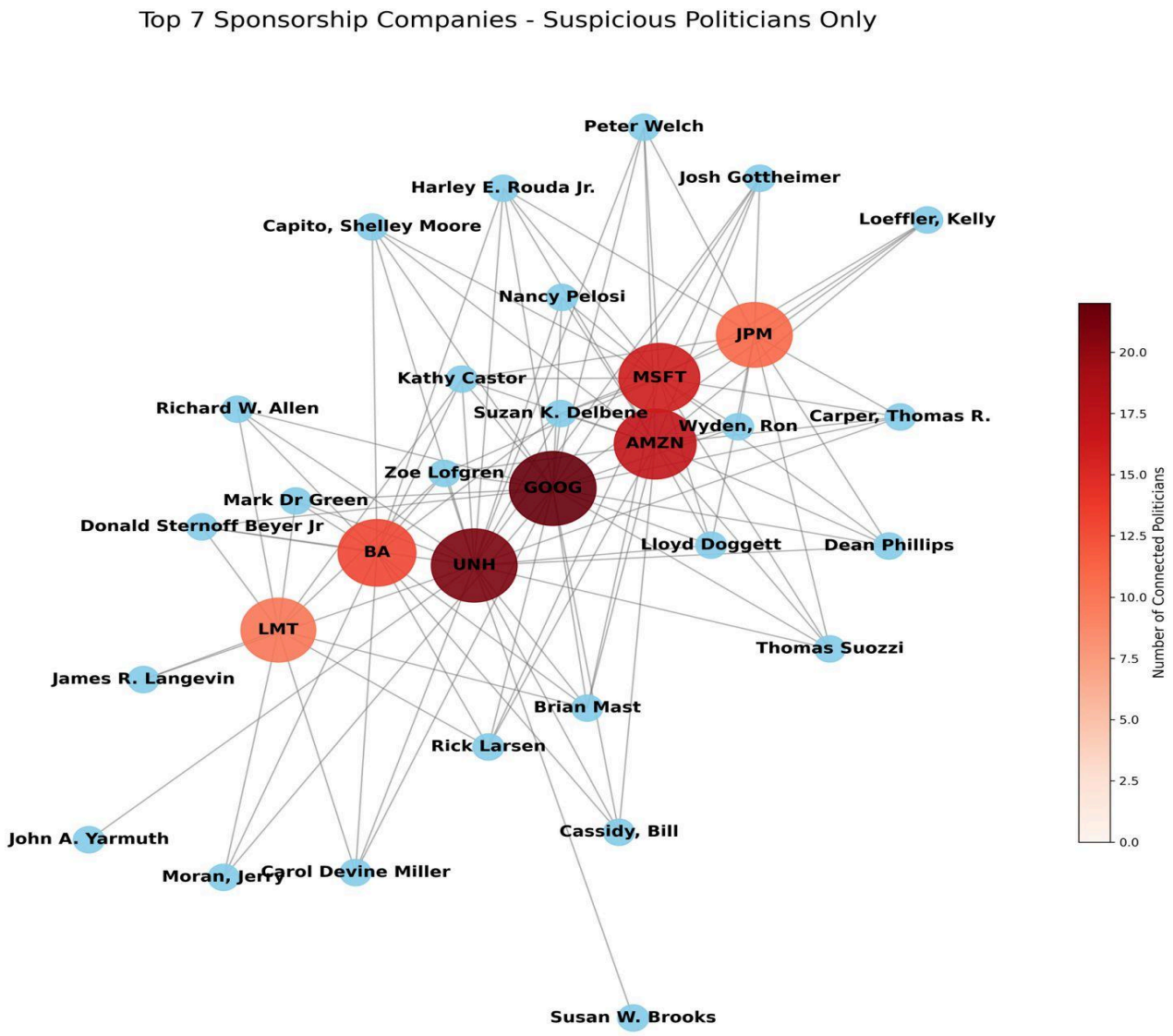
The results showed that politicians with the most suspicious trades consistently outperformed the S&P 500 benchmark across different time horizons, while less active or peripheral politicians often lagged behind the market and in some cases even incurred losses. This consistent pattern of abnormal gains strongly suggests the use of insider information.

To further profile the individuals most frequently involved in suspicious activity, we examine whether congressional tenure is associated with trading behavior. The next plot shows the average number of suspicious trades per tenure group in Congress. Although the absolute values appear low (2–4 per politician), this is due to strict preprocessing filters: we include only flagged politicians, limit the period to 2020–2021, focus solely on purchases, and cap trades at forty per politician. A trade is labeled suspicious only if a direct legislative connection is documented within five months of the transaction. Despite the low counts, a clear trend emerges: senior politicians with more than 30 years in Congress average nearly twice as many suspicious trades as less experienced members, suggesting that tenure is correlated with higher involvement in potentially suspicious financial activity. [Fig 7].

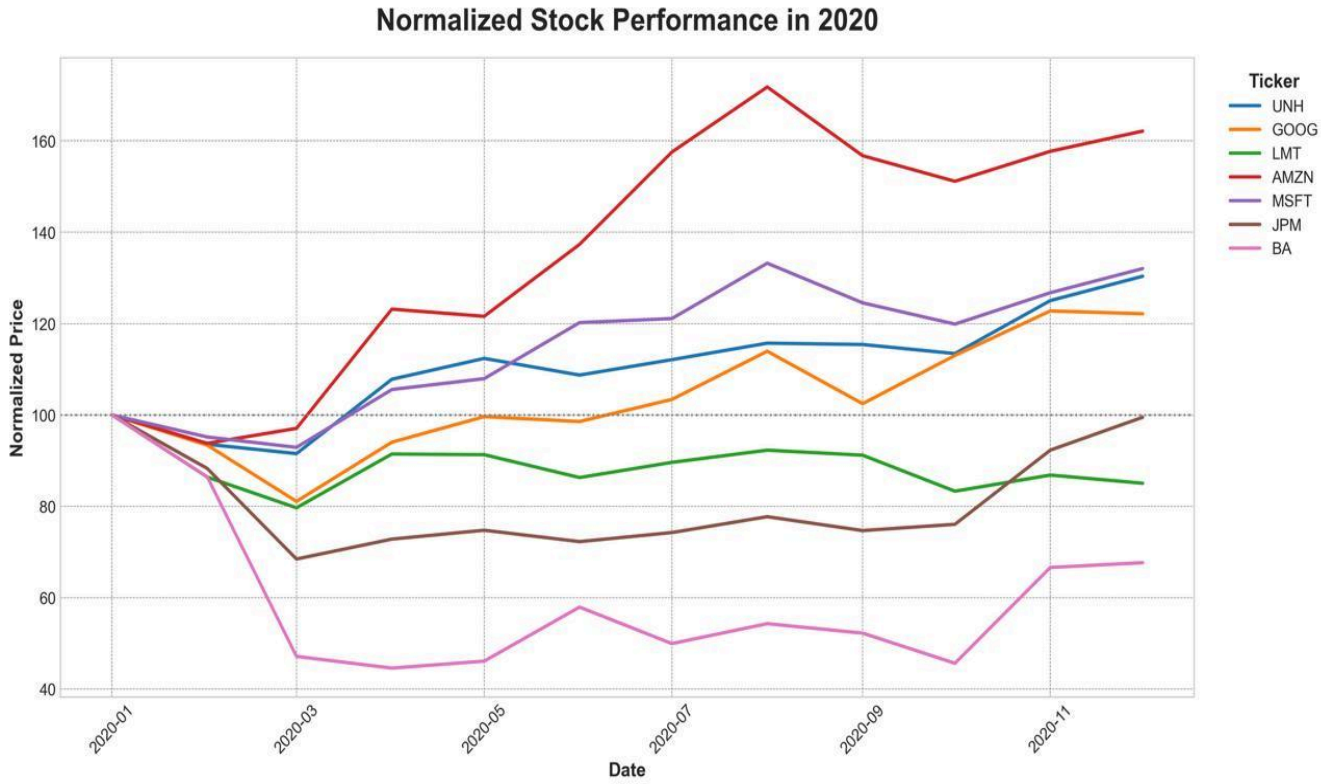


The next step was to investigate a potential link between a company's political influence (by sponsoring politicians) and its market value. We constructed a bipartite graph mapping sponsorship ties between corporations and politicians. The data for these connections was sourced from public disclosures scraped by our LLM. In this framework, a company's influence

is measured by the number of politicians it sponsors. This allows us to identify the most politically-connected corporations and then analyze whether there is a correlation between the breadth of their political support and their own stock's performance [Fig 8].



Our analysis of 2020 market performance shows that the leading companies with the sharpest stock price growth are the same firms we identified as providing the broadest political support. For example, a look at the firms most frequently appearing in suspicious trades reveals that top performers like Google, UnitedHealth Group, and Amazon clearly stand out, showing sharp price growth during the year and establishing themselves as dominant market players. Those findings reinforce the interpretation that financial and political advantages are closely intertwined. [Fig 9].



5. Evaluation

To evaluate our findings, we defined success as the ability to achieve a significant excess return over the S&P 500 benchmark - using percentage returns for a fair comparison. We constructed a network graph to identify trading communities and a bipartite graph to map corporate-politician sponsorship ties based on public data. Our findings, visualized through these graphs and time-series charts, revealed that many politicians achieved excess returns and showed a connection between a company's political support and its stock performance. The main issues we encountered were ensuring data integrity. We solved this by using the Adjusted Close Price for financial data to correct for skews from corporate actions and by validating potentially unreliable LLM scraped data through a rigorous process and performing continuous manual sanity checks online against reliable sources to find errors or outliers. These layered checks help ensure that our findings are not artifacts of arbitrary parameter choices but rather reflect genuine and systematic patterns in the data.

6. Future Work

Future extensions can explore several directions. First, analyzing results by party and chamber could reveal systematic differences between Democrats and Republicans, or between Senators and House members. Second, the timing of trades could be studied not only against legislative actions but also against broader signals such as public statements, tweets, or agenda themes, to capture subtler triggers for suspicious activity. Third, sector-specific committee analysis could test whether members systematically outperform in industries they directly oversee, such as health or energy. Finally, developing profile similarity measures such as Jaccard

similarity on committees, sectors, and transaction patterns could identify clusters of lawmakers with overlapping behaviors, adding another layer of evidence for coordinated or insider-like activity.

7. Conclusion

Our analysis highlights a consistent pattern in which a small subset of U.S. lawmakers engage disproportionately in suspicious stock trades. These politicians not only appear central in the trading networks but also achieve abnormal returns that exceed market benchmarks. We further find that tenure in Congress and committee memberships are strongly associated with higher involvement in such trades, and that the most profitable companies are often those with the strongest political sponsorship ties. Taken together, these findings suggest that financial and political advantages are tightly intertwined, raising concerns about conflicts of interest and the potential misuse of insider information.