Industry, Committee, and Lobbying - Uncovering Congressional Stock Trading using Graph Data

Suyeol Yun [*]

May 22, 2023

Abstract

In this study, I delve into the dynamics of congressional stock investments, revealing abnormal and asymmetrical excess returns. This suggests the potential influence of privileged information. Contrary to prior studies, I found that a congressperson's stock portfolio significantly mirrors stocks tied to their committee assignments, leveraging a novel graph-structured dataset. By employing a graph neural network, I discovered that committee assignments and firms' lobbying activities are predictive of congresspersons' stock choices. This research prompts a reevaluation of the motivations underpinning congressional service, indicating that personal financial gains could significantly influence decision-making. Additionally, it uncovers intriguing behavioral variations in stock trading among congresspersons. These variations could reflect individual characteristics or systemic factors within our political and financial systems, offering an intriguing avenue for further research.

[*] PhD Student, Department of Political Science, MIT. Email: syyun@mit.edu

1 Introduction

In democratic societies, electoral accountability is a central mechanism through which politicians, vested with the power to make significant public decisions, remain answerable for their actions (Besley, 2006; Fearon, 1999; Ferejohn, 1986). However, the intermingling of public service and personal financial interests can blur the lines of accountability. A key question arises: how closely intertwined are a politician's personal financial gains and their public service role? Are these two realms neatly partitioned, or do they bleed into each other?

This issue of financial behavior among politicians has been explored from various angles. Legal scholars have examined the use of political intelligence for profit and the potential for insider trading within the political sphere (Jerke, 2010; Bainbridge, 2010). These studies highlight the challenges of designing legal and enforcement structures to prevent and deter such behavior.

While these studies provide valuable insights, there has also been attention given to anecdotal approaches that explore specific instances of insider trading by politicians. For example, Schweizer (2011) presents a series of case studies that highlight how politicians and their associates have profited from insider stock tips, land deals, and cronyism. These accounts, while informative, underscore the need for a more systematic and empirical investigation of stock trading behavior among politicians, moving beyond individual anecdotes to a broader analysis of patterns and trends.

Recognizing this need to a broader analysis of patterns and trends, researchers have turned to the excess return approach, which has been used in the broader literature on insider trading, to assess whether politicians achieve abnormal returns on their investments. The excess return approach provides a foundation for understanding how informational advantages may lead to excess returns in financial markets (Jeng et al., 2003; Ivković and Weisbenner, 2005; Seasholes and Zhu, 2009).

Building on this approach, several studies have specifically examined the stock trading activities of U.S. Congresspersons. Ziobrowski et al. (2004) and Ziobrowski et al. (2011) conducted pioneering research in this area, analyzing the common stock investments of members of the U.S. Senate and House of Representatives, respectively. Their findings suggest that members of Congress achieve abnormal returns on their stock investments, raising questions about the potential use of privileged information in their trading decisions.

However, the notion that members of Congress achieve excess returns on their investments has been challenged by other researchers. Eggers and Hainmueller (2013) call into question the consensus that members of Congress trade with an information advantage. They reinterpret existing studies of congressional stock trading between 1985 and 2001 and conduct their own analysis of trades in the 2004–2008 period. Their findings challenge the notion of widespread "insider trading" in Congress, concluding that in neither period do members of Congress trade with an information advantage. Furthermore, they conduct the first analysis of members' portfolio holdings, showing that between 2004 and 2008, the average member of Congress would have earned higher returns in a passive index fund.

While this methodology provides insights into the overall trading performance of Congress as a collective entity, it may not fully capture the individual-level behavior of specific politicians. By aggregating the data and calculating average returns, the approach may mask variations and nuances in the trading behavior of individual members of Congress. As a result, instances of potentially unethical trading behavior by specific politicians, who may leverage privileged information for personal financial gain, could be obscured within the average calculations.

In addition, previous research, such as the study conducted by Eggers and Hainmueller (2014), focused on estimating the impact of firm-level lobbying or committee-assignment of bills lobbied by specific firms on the increase of that specific firm's weight in a congressperson's portfolio. However, this approach does not account for a congressperson's specialization in industry-level knowledge, gained through their committee assignments, and the potential utilization of such industry-level knowledge in shaping their personal investment portfolio.

To address this gap, I collected data that captures diverse aspects of legislative activities - such as firms' lobbying on specific bills, bills assignment to specific committees, and committee membership of congresspersons - alongside with congresspersons' stock trading data. I sourced this data from various relevant platforms, such as Lobbyview (Kim, 2018), Senate & House's financial disclosures, and Congress.

In order to investigate the relationship between a congressperson's portfolio and the industry specialization of the committees they are assigned to, I utilized a graph-structured data format. This facilitated an analysis capable of revealing a clear resemblance between the two. To substantiate this relationship statistically, I conducted a one-sided paired t-test (Hsu and Lachenbruch, 2014) aimed at determining whether the mean cross-entropy value measuring the similarity between the industry distribution of a congressperson's stock transactions and the industry distribution of their assigned committees was significantly smaller than that of their unassigned committees.

The result of the one-sided paired t-test shows that a congressperson's stock trading pattern significantly resembles the industry distribution of their assigned committees more than that of non-assigned committees. This finding is in stark contrast to the conclusions drawn by Eggers and Hainmueller (2014), highlighting one of the novel contributions of this research in understanding the relationship between committee assignments and stock trading behavior among Congress members.

While this research arrives at a conclusion that is diametrically opposed to that of Eggers and Hainmueller (2014), it aligns with the broader trend in the study of congressional stock trading. The work of Eggers and Hainmueller (2014) is valuable in that it moves beyond the traditional excess return approach and instead seeks to identify meaningful associations between various factors that could influence a congressperson's choice of stocks, such as PAC donations, district-level connections, or committee assignments.

In a similar vein, this study emphasizes the need to understand how legislative activities in general, which encompass a vast and complex network of information flows and interactions, can impact a congressperson's choice of specific stocks. These activities include firms lobbying on bills of particular interest to them, the referral of these bills to specific committees based on statutory and historical jurisdiction, and the assignment of congresspersons to these committees based on their expertise.

In this regard, this research proposes a more fundamental approach to study these behaviors, directly tackling the problem from an information perspective. The aim is to test whether congressional activities, which are often centered around legislative activities, provide information to predict each Congressperson's specific ticker transactions. To achieve this, a predictive model is designed that takes into account a variety of factors tied to each Congressperson's legislative activities. This will include elements such as the committees they are assigned to, the bills being legislated through those committees, and the potential interests of various firms or industries related to those bills. The goal is to ascertain whether these factors can reliably predict a Congressperson's transaction with a specific ticker at a specific time.

In order to effectively capture the complex nature of congressional activities, I have collected and organized data using a heterograph, a type of graph structure that incorporates different types of nodes and edges. This graph-structured data is particularly useful for representing various entities and their relationships, which are inherent in the legislative process. These relationships include firms lobbying on bills of particular interest to them, the referral of these bills to specific committees, and the assignment of congresspersons to these committees.

Pursuing the approach of Eggers and Hainmueller (2014), this research endeavors to predict the specific stock transactions of given congresspersons. If these transactions can be forecast using the graph-structured data, it implies that congressional activities contain vital information that can explain their stock trading patterns. Such a finding is significant as it highlights a potentially strong association between a congressperson's legislative activities and their stock transactions.

In conducting this prediction task, the research successfully trained a Graph Neural Network (GNN) model (Zhou et al., 2020; Wu et al., 2020; Scarselli et al., 2008; Zhang et al., 2019), achieving an accuracy of 0.81 and an AUC-ROC score of 0.89. These results indicate that congressional activities provide substantial information to explain the stock choices of congresspersons, further supporting the hypothesis of a significant correlation between legislative activities and stock transactions.

To understand the varying contribution of different types of edges in the heterograph to the prediction task, an ablation study was also conducted. It systematically removed particular edge types from the training and calculated the Shapley values (Winter, 2002; Hart, 1989; Littlechild and Owen, 1973), which measure the contribution of each edge type to the prediction. This study found that a congressperson's committee assignments and firm-level lobbying on bills were the most important factors contributing to the model's predictive power. Interestingly, this finding contradicts Eggers and Haimuller's (2014) conclusion, which argued that there's no significant evidence that firm-level lobbying on bills and a congressperson's committee assignments explain the weight of specific stocks in their portfolio.

To summarize, this research contributes to the field in several significant ways. First, this study emphasizes the application of graph-structured data and implements a predictive analysis using a Graph Neural Network (GNN). The adoption of GNN enables a richer understanding of the relationship between congressional activities and stock transactions. Unlike conventional predictive modeling that merely relies on independent feature vectors, this approach integrates the topological properties of the network into the prediction task. Secondly, the research underscores the significance of a congressperson's committee assignments and firm-level lobbying in explaining the choice of stock transactions. This conclusion contradicts some previous studies, but aligns more closely with traditional literature on the influence of committee specialization and firm-level lobbying. Finally, this investigation redirects the traditional discourse on congressional motivation. Traditional research predominantly emphasizes the public career aspirations of congresspersons, such as re-election and the pursuit of sound public policies. However, the findings of this study reveal a marked similarity between a congressperson's legislative activities and their stock transaction patterns. This correlation necessitates an expansion of the current understanding of congresspersons' motivations, acknowledging the potential influence of personal financial success on their legislative behavior.

2 Graph-Structured Data for Representing Congressional Activities

[1] The data utilized in the following sections forms a large, complex network that is categorized as a heterograph or heterogeneous graph. This structure captures congressional activities through different types of nodes and edges, thereby encapsulating the multi-faceted nature of these activities. This heterograph encompasses information on congressional activities, such as committee assignments, bills being lobbied by firms, bill assignments to committees, and firms classified under specific NAICS codes. The detailed specifications of the node types can be found in Table 1, while the edge types are described in Table 2. The process of data collection from disparate sources and the subsequent disambiguation and merging of entities are elaborated upon in Appendix A. Additionally, Appendix B provides a detailed explanation of a more modern approach to extracting structured data from collected financial disclosure PDFs. In this approach, I utilized a Large Language Model (LLM) to aid in the extraction process. The specifics of how the LLM was employed are discussed in detail within the appendix.

Table 1: Heterograph (Nodes)

Node Type	N	Period	Source
Firm (Ticker)	4,202	-	Lobbyview & Finance Disclosure
Bills	47,767	110-117th Congress	Lobbyview
Congressperson	2,431	113-118th Congress	Lobbyview & Finance Disclosure
Committee	556	-	Lobbyview
NAICS code	744	-	naics.com
Total	55,700	-	-

[1] Reproducible code for this section is available at `https://github.com/syyunn/gnnex/blob/main/data/graph.ipynb`

Table 2: Heterograph (Edges)

Edge Types	N	Period	Source
Congressperson- Buy/Sell- Firm (Ticker)	24,675	[2013-01-24, 2023-03-08]	Finance Disclosure
Firm (Ticker) - Lobby On - Bill	148,487	[2016-01-02, 2022-02-24]	Lobbyview
Ticker- Classified as - NAICS Codes	4,147	-	Finance Disclosure & naics.com
Bill- Referred to - Committee	75,626	[2016-01-05, 2021-12-17]	Lobbyview
Congressperson- Assigend to - Committee	11,698	115-117th Congress	Finance Disclosure & Lobbyview
Total	264,633	-	-

3 Predicting Congressional Stock Transactions using Graph Neural Networks

In the previous section, therefore, I discussed the limitations of the linear prediction model used by Eggers and Hainmueller (2014), which employed a binary encoding of lobbying and committee assignments to predict the weight of a specific firm's stock in a congressperson's portfolio. I pointed out that this model did not fully capture the complex interactions between different entities involved in congressional activities, particularly at the industry level. However, I acknowledge that the model was an attempt to explain congressional stock transactions using potentially explanatory components such as district, PAC, lobbying, and committee assignments.

It's important to note that while the cross-entropy approach in Section **??** revealed a clear resemblance between Congresspersons' stock trading behavior and their assigned committees' industry-level specialization, this approach does not directly answer whether this resemblance originates from knowledge gained through congressional activities or from the Congresspersons' expertise and experience before their congressional tenure.

In this section[2], I propose to use a graph neural network (GNN) (Zhou et al., 2020; Wu et al., 2020; Scarselli et al., 2008; Zhang et al., 2019) to predict congressional stock transactions using the information embedded in the congressional activities captured in the data explained in Section 2. The GNN approach is uniquely equipped to handle this task because it can model the complex relationships among various entities involved in congressional activities, all of which are naturally structured as a graph.

By using GNN, we can design a model that directly consumes the congressional graph that captures legislative-related activities of different entities, thereby enabling us to test the predictability of

[2] Reproducible code for this section is available at `https://github.com/syyunn/gnnex/blob/main/hetero/train_kfold_auto.py`

congressional trading behavior based on these activities. This will help us isolate the influence of congressional activities on stock trading from pre-congressional expertise and other confounding factors, which is a significant step forward in our understanding of the interplay between committee assignments, congressional activities, and stock transactions. By leveraging a graph representation of the relationships between firms, bills, committees, and congresspersons, we can train a GNN to predict whether a congressperson is likely to buy a particular stock.

3.1 Designing a Binary Classifier with Graph Neural Networks

To predict congressional stock transactions using a graph neural network (GNN) approach, I design a binary classifier that takes as input a graph G, a congressperson and a ticker (stock symbol). The classifier, denoted as $f(G, \text{congressperson}, \text{ticker})$, will output a binary prediction of either 0 or 1, indicating whether an edge (a buy or sell relationship) exists between the given congressperson and the ticker.

The hidden representations (Rauber et al., 2016; Das et al., 2020)o f the congressperson and the ticker, denoted as $h_{\text{congressperson}}$ and h_{ticker} respectively, are obtained as outputs of the GNN model. The main task in this approach is to train the GNN model to learn a computational graph that generates "good" representation of the congressperson and the ticker, $h_{congressperson}$ and h_{ticker}, which involves how to effectively encode the information embedded in the network to perform the downstream task of binary classification (Féraud and Clérot, 2002).

To design the classifier, a probabilistic modeling approach is used that comprises of a sigmoid function applied to the logit, which is the output of the model. The logit of the model is obtained by passing the representation learned by the GNN, $h_{congressperson}$ and h_{ticker}, to an MLP (Multi-layer perceptron) (Gardner and Dorling, 1998; Tang et al., 2016) that maps the representations of the congressperson and the ticker to a single logit. In other words, the MLP takes as input the representations of the congressperson and the ticker learned by the GNN, and outputs a logit that will be used to compute the probability of the existence of edge between them. MLP is simply an affine transformation over the concatenation of two representations, $h_{congressperson}$ and h_{ticker}, follwed by a non-linear activation function (Lu and Lu, 2020), which is ReLU (Agarap, 2018) in this case.

Formally, the logit of the classifier is defined as:

$$\begin{aligned}
\text{logit} &= \text{MLP}\,(x) \text{ where} \\
\text{MLP}(x) &= \text{ReLU}(Ax+b) \\
&\quad x = \text{concat}\,(h_{\text{congressperson}}, h_{\text{ticker}}) \\
&\quad A \in \mathbb{R}^{(d+d)\times 1} \\
&\quad b \in \mathbb{R}^{1}
\end{aligned}$$

• logit: This is the output of the model. It's a transformed version of the probability that a congressperson would invest in a particular stock. In this binary classification problems, the logit (also known as log-odds) is the logarithm of the odds p/(1-p) where p is the probability of a positive event that congressperson trades such stock.

• MLP: This stands for Multi-Layer Perceptron, a type of artificial neural network. In this case, it's a function that takes the concatenated embeddings of a congressperson and a ticker as input and produces a logit as output.

The sigmoid function is then applied to the logit to obtain a probability value:

$$\text{prob } = \sigma(\text{ logit })$$

where $\sigma(x)$ is the sigmoid function. The probability value indicates the likelihood of an edge existing between the given congressperson and the ticker. If the probability value is above a certain threshold, we predict that an edge exists between them, otherwise we predict that there is no edge.

Then remaining task is how to design a GNN model that can effectively learn the representations of the congressperson and the ticker, $h_{\text{congressperson}}$ and h_{ticker}, respectively, which can be used to train the classifier. In the following section, I will discuss the design of the GNN model.

3.2 Design of the Graph Neural Network Architecture

To obtain the representations $h_{\text{congressperson}}$ and h_{ticker}, I use a GNN approach that is designed to handle the complexity and dynamics of the congressional graph. The GNN approach is based on the idea of message passing and updating (Zhou et al., 2020; Wu et al., 2020), which is a process of aggregating information from the neighbors and updating the representation of each node accordingly.

In the case of the congressional graph, I use an edge-conditioned convolution GNN model (Gilmer et al., 2017; Simonovsky and Komodakis, 2017), which takes into account the edge attributes, such as the date, to better capture the complex relationships in the graph. The message passing, aggragation and updating in this model is defined as:

$$\mathbf{h}_i' = \mathbf{\Theta}\mathbf{h}_i + \sum_{j \in \mathcal{N}(i)} \mathbf{MLP}\left(\mathbf{e}_{i,j}\right) \cdot \mathbf{h}_j$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are the representations of nodes i and j, respectively, $\mathbf{e}_{i,j}$ is the edge attribute between nodes i and j, $\mathcal{N}(i)$ is the set of neighbors of node i, $\mathbf{\Theta}$ is a learnable matrix of size $d \times d$, where d is the dimension of the representation space, and $\mathbf{MLP}$ takes the edge attribute $\mathbf{e}_{i,j}$ as input and outputs a weight matrix of size $d \times d$. This weight matrix is then multiplied with the representation $\mathbf{h}_j$ of the neighbor node j to obtain a message $\mathbf{m}_{i,j} = \mathbf{MLP}\left(\mathbf{e}_{i,j}\right) \cdot \mathbf{h}_j$. In the updating step, the message from each neighbor node is aggregated by summing them up, and the resulting sum is added to the current representation $\mathbf{h}_i$ of node i multiplied by the learnable parameter matrix $\mathbf{\Theta}$ to obtain the updated representation $\mathbf{h}_i'$.

3.3 Training & Evaluation of the GNN

3.3.1 Dataset Preparation

In the context of our GNN architecture, the goal is to predict the existence of edges between two nodes, a task commonly known as link prediction. To train the GNN for this task, the dataset must be prepared for training and evaluation (test). The dataset consists of a total of 24,675 edges, which represent the relationship (congressperson, buy-sell, ticker). To create a balanced dataset for the link prediction task, the dataset is divided into a train and test set with an 8:2 ratio, resulting in 19,740 instances for training and 4,935 instances for testing. The network is then trained using the 19,740 instances and its performance is evaluated on the 4,935 test instances.

In addition, to ensure a balanced dataset, the same number of randomly sampled negative edges (Yang et al., 2020) is prepared. These negative edges are created by randomly selecting pairs of nodes (congressperson and ticker) that do not have a connection in the original dataset. This results in a total of 39,480 edges for training and 9,870 edges for testing. Including both positive and negative examples in the training process helps the model to better differentiate between true and false existence of edges between congressperson and tidcker nodes, improving its ability to predict links in the graph.

3.3.2 Training of the GNN

For the training of the GNN, a two-layer GNN architecture, as $l = 2$, is employed. Additionally, node embeddings h_i are represented as vectors in a 64-dimensional space ($h \in \mathbf{R}^{64}$). This hyperparameter is also set experimentally.

In order to measure the performance of the model during the training process, binary cross-entropy loss is used as the loss function. Binary cross-entropy loss is particularly suitable for binary classification problems (Ruby and Yendapalli, 2020), such as link prediction (Zhang and Chen, 2018), where the goal is to differentiate between the presence and absence of a connection between two nodes. This loss function quantifies the difference between the predicted probabilities and the true labels, and penalizes the model for incorrect predictions. Formally, the binary cross-entropy loss for a set of samples is defined as:

$$L = -\sum_{i=1}^{N} \left(y_i \cdot \log\left(p_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - p_i\right) \right)$$

where L represents the total binary cross-entropy loss. N is the total number of samples. y_i is the true label for the ith sample (1 for the presence of a connection, and 0 for the absence of a connection). p_i is the predicted probability of a connection existing between two nodes for the ith sample.

By minimizing the binary cross-entropy loss, the GNN learns to accurately predict the existence or non-existence of links in the network, ultimately improving its performance on the link prediction task. For the minimization, the Adam optimizer (Kingma and Ba, 2014) with stochastic gradient descent (SGD) is utilized. SGD is an iterative optimization algorithm that updates the model's parameters based on a random sample (or minibatch) of training data in each iteration (Amari, 1993). This approach helps in converging faster and reduces the impact of noisy gradients, thus improving the optimization process. Adam is an adaptive learning rate optimization algorithm, combining the advantages of two other popular optimization methods, AdaGrad and RMSProp (Kingma and Ba, 2014). This optimizer is well-suited for large-scale problems and is known for its ability to efficiently handle noisy and sparse gradients, making it a suitable choice for training GNNs.

To obtain a more robust estimation of the model's performance and uncertainty, a 5-fold crossvalidation (Hastie et al., 2001) is performed. In this approach, the entire dataset is randomly split into five equal-sized chunks. For each fold, one chunk is used as the test set, while the remaining chunks are combined to form the training set. This process is repeated five times, with each chunk being used once as the test set. This technique allows for a better understanding of the model's performance across different subsets of the dataset and provides uncertainty statistics of overall prediction performance.

3.3.3 Evaluation & Ablation Study

In this study, we conducted a link-prediction (Zhang and Chen, 2018) task to predict the existence of an edge between a congressperson and a ticker, symbolizing the trade relationship - whether the given congressperson would sell or buy a particular stock. This task was performed using a variety of edge types, and the performance was evaluated using two metrics: accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The results of this evaluation are depicted in Figures 1 and 2. With all edge types included, the model achieved an accuracy of approximately 81% and an AUC-ROC of 0.89. These results indicate that the model was generally effective at predicting the stock transactions of congresspersons.

To further understand the importance of each edge type, I conducted an ablation study, where I systematically removed each edge type from the training and testing data and observed the resulting performance drop. The most significant drop in performance was observed when the edge type ('congressperson', 'assignment', 'committee') was removed. This resulted in a decrease in accuracy from 81% to 67%, and a decrease in AUC-ROC from 0.89 to 0.76. This suggests that the ('congressperson', 'assignment', 'committee') edge type carries significant information for predicting a congressperson's stock transactions.

In comparison, the removal of other edge types, such as ('bill', 'assigned_to', 'committee'), or ('ticker', 'lobbies_on', 'bill'), resulted in less dramatic performance drops. This further underscores the relative importance of the ('congressperson', 'assignment', 'committee') edge type in this prediction task.

To further quantify the importance of each edge type, I employed the concept of Shapley values (Winter, 2002; Hart, 1989; Littlechild and Owen, 1973), a concept borrowed from cooperative game theory. In this context, each edge type can be considered as a player in a cooperative game, where the "payout" is the performance of the model.

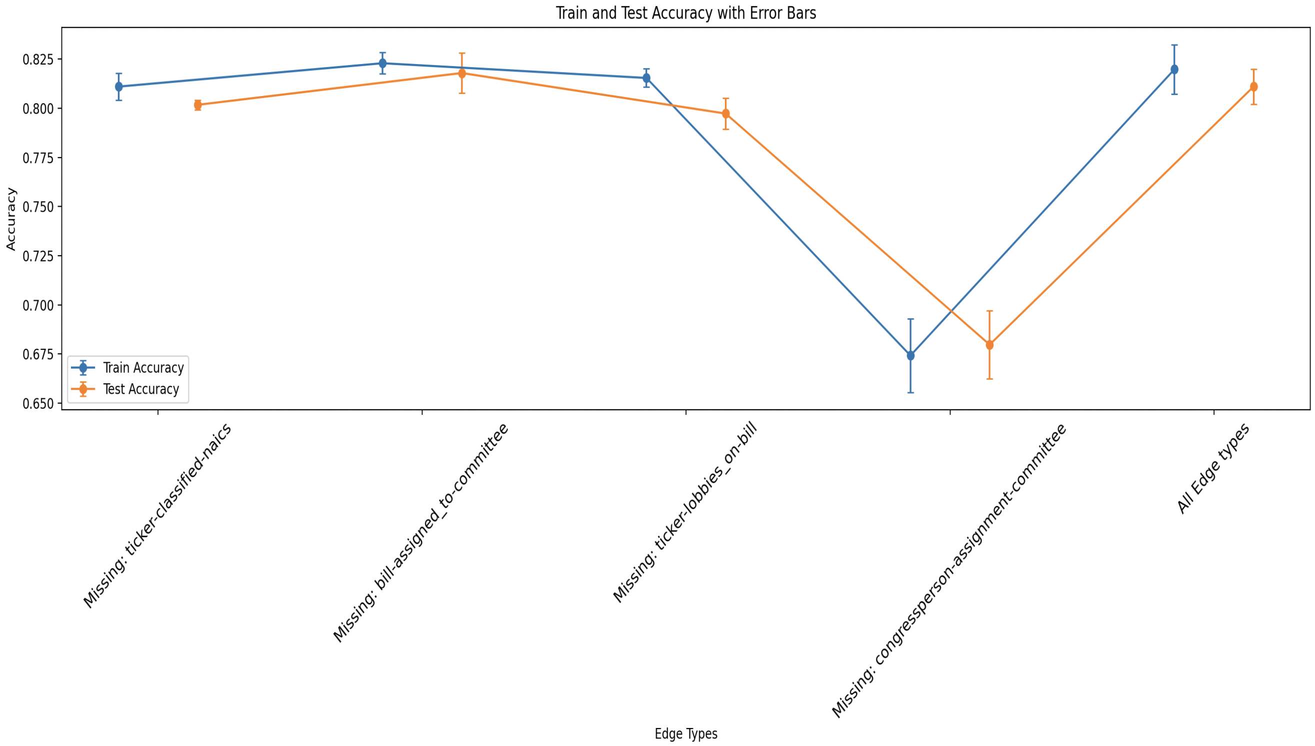

Figure 1: **Accuracy drop for different edge types.** The figure shows the accuracy of the model with all edge types included and with each edge type removed one at a time. With all edge types included, the model achieved an accuracy of approximately 81%. The most significant drop in accuracy, to 67%, was observed when the edge type (‘congressperson’, ‘assignment’, ‘committee’) was removed. This suggests that the (‘congressperson’, ‘assignment’, ‘committee’) edge type carries significant information for predicting a congressperson’s stock transactions.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

Here, $\varphi_i(v)$ is the Shapley value for edge type i, representing the average marginal contribution of edge type i to the performance of the model, considering all possible combinations of edge types. N is the set of all edge types, not the total number of edges. In our case, there are four edge types, so N is 4. S is a subset of N that does not include edge type i, $|S|$ is the number of edge types in subset S, and n is the total number of edge types. $v(S \cup i)$ and $v(S)$ represent the performance of the model when edge type i is added to and excluded from the subset S of edge types, respectively. This means that the Shapley value indicates how much each edge type contributes to the performance of the model, which in our case is measured by prediction accuracy or AUC-ROC. These findings indeed underline the crucial role of the (congressperson’, assignment’, ‘committee’) edge in predicting congresspersons’ stock transactions. In addition, firm-level lobbying and industry-level classification provided by the edge types (‘ticker’, ‘classified_as’, ‘naics’) and (‘ticker’, ‘lobbies_on’, ‘bill’) are more also informative to predict

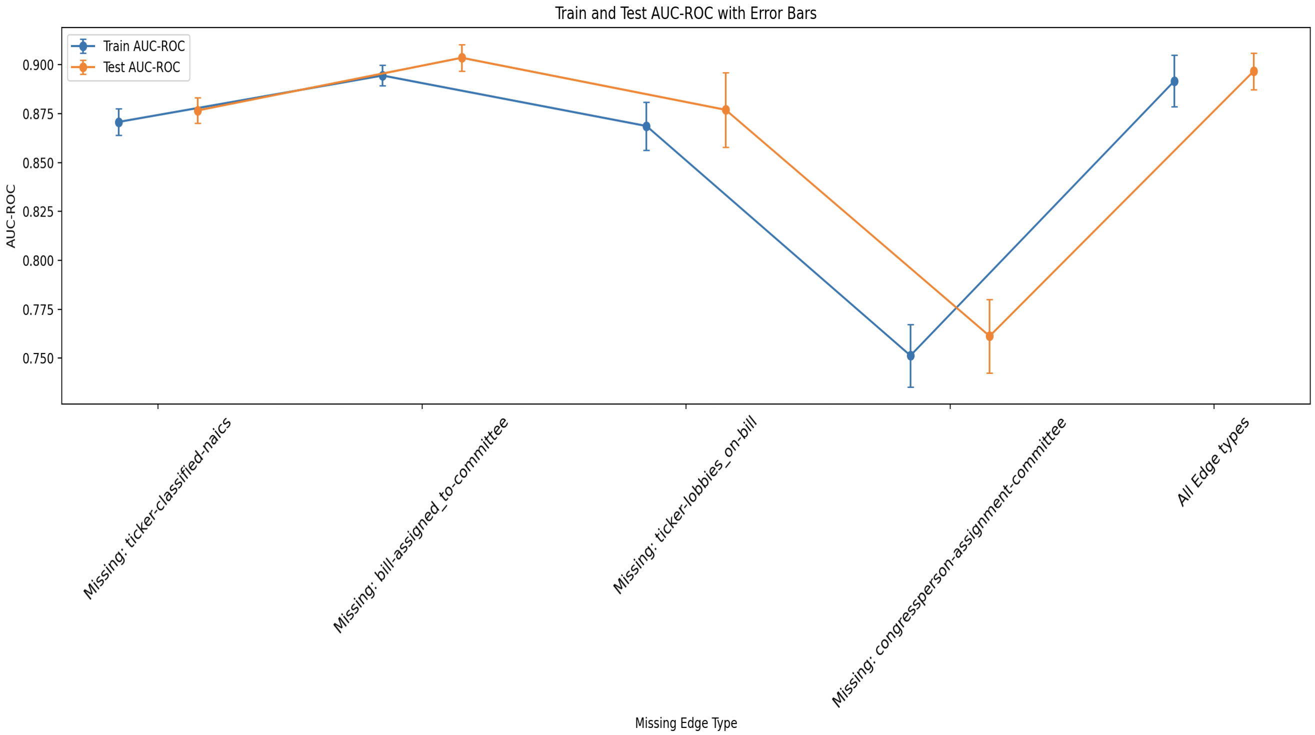

Figure 2: **AUC-ROC drop for different edge types.** The figure shows the AUC-ROC of the model with all edge types included and with each edge type removed one at a time. With all edge types included, the model achieved an AUC-ROC of approximately 0.89. The most significant drop in AUC-ROC, to 0.76, was observed when the edge type ('congressperson','assignment', 'committee') was removed. This suggests that the ('congressperson', 'assignment', 'committee') edge type carries significant information for predicting a congressperson's stock transactions.

congrespepople's stock trading. These edge types may allow the model to more easily discern patterns in company behavior and use this information to make accurate predictions.

4 Conclusion and Future Directions

In this study, I expanded on the work of Eggers and Hainmueller (2014) by using the graph neural network to determine how possibly relevant factors, such as congressional activities captured in the graph data, predict congresspersons' stock transactions. The results showed that, contrary to Eggers and Hainmueller (2014), the committee assignment of congresspersons and lobbying activities of firms are the most important features predicting their stock selections. In addition, congressional activities in general, including firm-level lobbying and firm's industry class code, are informative for predicting stock transactions of congresspeople.

Adding onto these findings, it's worth examining the broader implications of our results. On a macro level, this research invites us to reassess our understanding of the motivations underpinning congressional

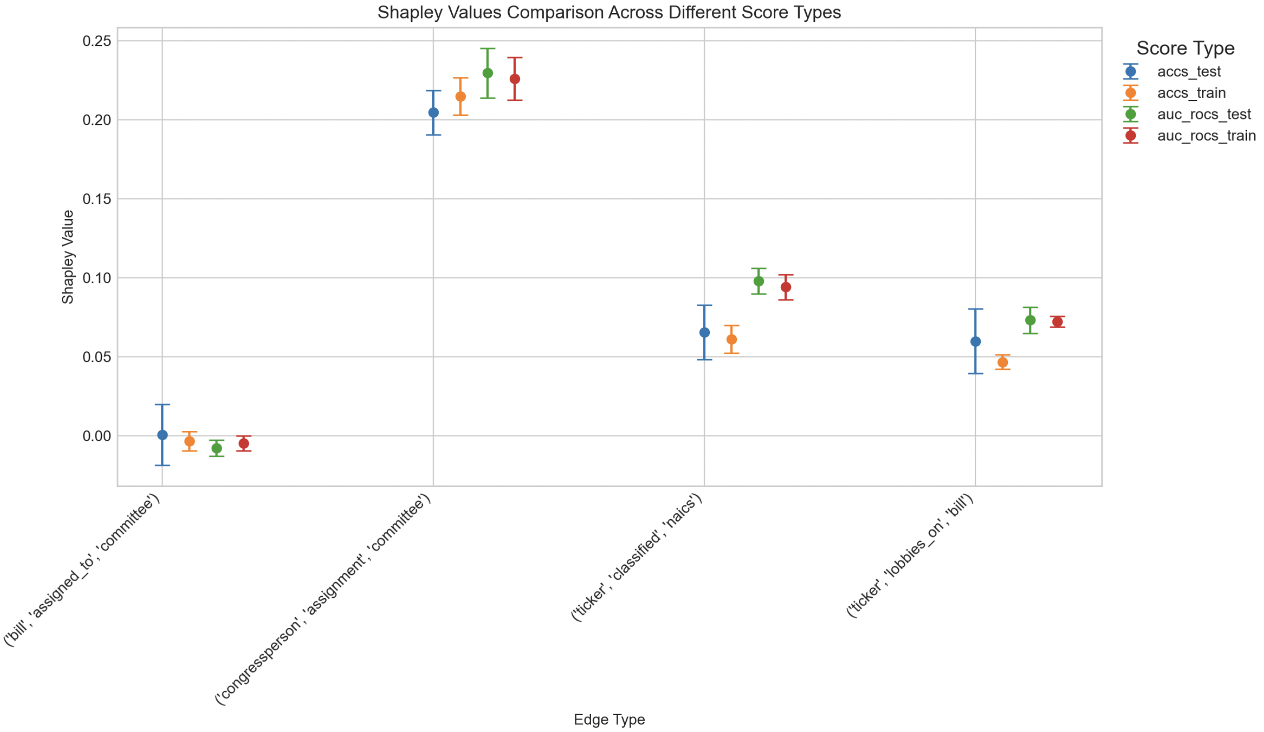

Figure 3: **Shapley values for different edge types.** The figure shows the Shapley values for each edge type, computed over all $16(2^4)$ possible combinations of the four different edge types. The Shapley value for an edge type represents its average marginal contribution to the performance of the model, considering all possible combinations of edge types. The most important feature, according to the Shapley value analysis, was ('congressperson', 'assignment', 'committee'), followed by ('ticker', 'classified_as', 'naics') and ('ticker', 'lobbies_on', 'bill'). This further reinforces the conclusion that the ('congressperson', 'assignment', 'committee') edge type plays a crucial role in predicting congressperson's stock transactions.

service. The common perception is that congresspersons are primarily motivated by the goal of reelection, with their actions driven by a desire to serve their constituents and deliver policy outcomes that align with their promises and their party's platform. However, our findings suggest that the picture may be more complex, with financial considerations also playing a significant role. The evidence that congressional activities are predictive of their investments points to the potential for personal financial gain to be a motivating factor in their decisions and actions in their congressional services. This raises profound questions about the alignment of incentives in our political system and the possibility of conflicts of interest.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

Amari, S. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4):185–

196.

Bainbridge, S. (2010). Insider trading inside the beltway. Law-Econ Research Paper 10-08, UCLA School of Law.

Besley, T. (2006). *Principled Agents?: The Political Economy of Good Government*. Oxford University Press, Oxford, UK.

Das, L., Sivaram, A., and Venkatasubramanian, V. (2020). Hidden representations in deep neural networks: Part 2. regression problems. *Computers & Chemical Engineering*, 139:106895.

Eggers, A. and Hainmueller, J. (2013). Capitol losses: The mediocre performance of congressional stock portfolios, 2004-2008. *Journal of Politics*, 75.

Eggers, A. C. and Hainmueller, J. (2014). Political capital: Corporate connections and stock investments in the u.s. congress, 2004-2008. *ERN: Models of Political Processes: Rent-Seeking*.

Fearon, J. D. (1999). Electoral accountability and the control of politicians: Selecting good types versus sanctioning poor performance. In Przeworski, A., Stokes, S. C., and Manin, B., editors, *Democracy, Accountability, and Representation*, pages 55–97. Cambridge University Press, Cambridge.

Féraud, R. and Clérot, F. (2002). A methodology to explain neural network classification. *Neural networks*, 15(2):237–246.

Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice*, 50(1/3):5–25.

Gardner, M. W. and Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32:2627–2636.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. *CoRR*, abs/1704.01212.

Hart, S. (1989). *Shapley value*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hsu, H. and Lachenbruch, P. A. (2014). Paired t test. *Wiley StatsRef: statistics reference online*.

Ivković, Z. and Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments. *The Journal of Finance*, 60(1):267–306.

Jeng, L. A., Metrick, A., and Zeckhauser, R. (2003). Estimating the returns to insider trading: A performance-evaluation perspective. *Review of Economics and Statistics*, 85(2):453–471.

Jerke, B. W. (2010). Cashing in on capitol hill: Insider trading and the use of political intelligence for profit. *University of Pennsylvania Law Review*, 158:1451–1523.

Kim, I. S. (2018). Lobbyview: Firm-level lobbying and congressional bills database. Working Paper.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Littlechild, S. C. and Owen, G. (1973). A simple expression for the shapley value in a special case. *Management Science*, 20(3):370–372.

Lu, Y. and Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing distributions. *CoRR*, abs/2004.08867.

Rauber, P. E., Fadel, S. G., Falcao, A. X., and Telea, A. C. (2016). Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):101–110.

Román, J. H., Hulin, K. J., Collins, L. M., and Powell, J. E. (2012). Entity disambiguation using semantic networks. *Journal of the American Society for Information Science and Technology*, 63(10):2087–2099.

Ruby, U. and Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Schweizer, P. (2011). *Throw Them All Out: How Politicians and Their Friends Get Rich Off Insider Stock Tips, Land Deals, and Cronyism That Would Send the Rest of Us to Prison*. Houghton Mifflin Harcourt (HMH).

Seasholes, M. S. and Zhu, N. (2009). Individual investors and local bias. *Journal of Finance*. Forthcoming.

Simonovsky, M. and Komodakis, N. (2017). Dynamic edge-conditioned filters in convolutional neural networks on graphs. *CoRR*, abs/1704.02901.

Tang, J., Deng, C., and Huang, G.-B. (2016). Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):809–821.

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Yang, Z., Ding, M., Zhou, C., Yang, H., Zhou, J., and Tang, J. (2020). Understanding negative sampling in graph representation learning. *CoRR*, abs/2005.09863.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 793–803.

Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. *CoRR*, abs/1802.09691.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

Ziobrowski, A. J., Boyd, J. W., Cheng, P., and Ziobrowski, B. J. (2011). Abnormal returns from the common stock investments of members of the u.s. house of representatives. *Business and Politics*, 13(1):1–22.

Ziobrowski, A. J., Cheng, P., Boyd, J. W., and Ziobrowski, B. J. (2004). Abnormal returns from the common stock investments of the u.s. senate. *The Journal of Financial and Quantitative Analysis*, 39(4):661–676.

A Data Merging and Entity Disambiguation

One of the key challenges to make this graph-structured data is the effective disambiguation of entities, as the data is collected from multiple sources, including LobbyView, Senate/House Financial Disclosures, and naics.com. In this graph-structured dataset, entities such as Congresspersons and firms may appear under different names or expressions. For example, "Ron Wyden" may also be referred to as "Ron L. Wyden", and "Apple" may appear as "Apple Inc.". To accurately disambiguate these differing text representations of entities, it is essential to establish a unique identifier for each entity, regardless of the variations in their names.

Theoretically, matching entities based on text similarity between two datasets with n and m rows has a computational complexity of $O(nm)$ (Román et al., 2012). Therefore, as the datasets grow larger, this complexity becomes prohibitively expensive. For instance, matching 70,000 firm names from LobbyView to 4,000 firm names appearing in the ticker table would require 280,000,000 times of computations for text similarity. To address this challenge, I developed a novel approach that leverages URLs as unique identifiers for entities.

The approach involves acquiring the corresponding URL for each entity through Google searches, such as `https://en.wikipedia.org/wiki/Ron_Wyden` for Ron Wyden and `https://www.apple.com/` for Apple, Inc. A key advantage of using URLs as unique identifiers is that they facilitate effective entity disambiguation. For example, if two different expressions, "Ron Wyden" and "Ron L. Wyden" are both assigned the same URL `https://en.wikipedia.org/wiki/Ron_Wyden`, we can confidently recognize that these two expressions refer to the same entity. This approach allows us to accurately consolidate information about entities that may be represented in various ways across different data sources. Additionally, this method reduces the computational complexity to $O(n+m)$, as only one query is required for each row of data. To further scale up this process, I parallelized the URL acquisition process by batching queries and distributing them across multiple servers available through commercial cloud services like AWS.

B Effective Parsing Technique for Financial Disclosures

Financial Disclosures from the House are provided as encrypted PDF files. While text can be extracted from these files, the encryption results in irregular patterns, particularly in the tables that contain information about Congresspersons' stock buying and selling activities. These irregular patterns make it challenging to parse the data using manually coded patterns, as the deviations are difficult to anticipate and account for. To address this challenge, I utilized OpenAI's APIs, specifically the GPT-3.5 Turbo language model, to parse the PDFs into a CSV format that includes information such as when and who bought or sold which ticker, and how much.

The process involves querying the Large Language Model (LLM) with the extracted text from the PDFs and instructing the model to convert the irregularly formatted tables into structured CSV data which includes columns such as the date of the transaction, the name of the Congressperson, the ticker symbol of the stock, the type of transaction (buy or sell), and the amount of the transaction.

By leveraging the capabilities of the GPT-3.5 Turbo language model, I was able to effectively parse information contained in PDF files that would normally require manual human labor. This approach significantly streamlines the data extraction process and ensures the accuracy and consistency of the parsed data.

In summary, this innovative approach to entity disambiguation through URL acquisition and parallelization enables efficient data merging from diverse sources, ensuring the accuracy and scalability of the analysis.