MGT 6203 Final Report
Fall 2022 Team 81

Team Members

Luke Athans (lathans3)

Karl Hernes (khernes3)

Yujie Liang (hliang35)

John Stirling (jstirling7)

Table of content

Contents

Introduction

Members of the United States Congress, their spouses, and their direct business partners are required by law to disclose their business activities. These disclosures are required because members of the legislative body of the United States government could easily use their position for personal gain. Below are some, but certainly not every scenario where this could occur:

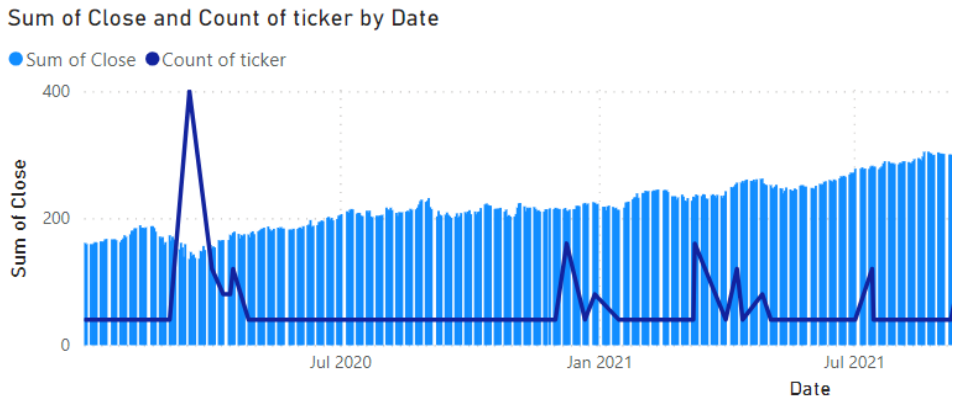
1. Using unreleased, or classified information to inform a stock purchase or sale before the general population can learn of this information. This would be a form of insider trading.
2. Investing in a company of a certain sector / market function, and then proposing or promoting a new law that would be beneficial to this sector as a whole or specific company.
3. Opposing legislation that would harm a company that a congressperson holds equity in.

In order to help prevent corruption in the branch of government responsible for drafting, proposing, and ratifying new laws, members of congress must submit disclosures of their business activities. These disclosures typically contain the following information:

- transaction date
- disclosure date
- owner of the equity (Self / Spouse / Joint)
- ticker (the stock market ticker code for the company/fund)
- asset description (the long-form name of the stock, ETF, etc.)
- asset type (i.e., public stock, private stock, bond, option, commodities/futures, others)
- type of transaction (a purchase, sale, partial sale)
- amount of money spent, typically denoted by tiered ranges
- URL to the transaction report

Problem Statement

Even with regulations in place to deter congressional representatives and senators from partaking in illegal trading activity, many still believe that there is evidence that these individuals abuse their office for financial gain. For example, in March 2020, the month that COVID-19 led to widespread “lockdowns”, you can see a huge spike in the number of trades reported by legislators:



Many consider this proof that legislators traded on knowledge that was not yet made available to the public. However, this conclusion cannot be proven by simply looking at the chart above.

Approach

To investigate whether this was true we analyzed the trading data that representatives and senators were required to disclose. We carried out this analysis by training several analytical models that aimed to identify hidden patterns within the trading activity. Our goal for most of these models was to predict the percentage increase or decrease in a stock's price, or the probability that a stock would see an increase in price based on congressional trading patterns. We also compare the stock performance against the historical major market index, aim to determine if Senators and Representatives have superior stock picking abilities against SP500.

Data overview

Data Sources

We utilized public records released by the United States House of Representatives, and the United States Senate outlining stock trade disclosures made by the members of each legislative body. We also referenced historic stock/fund data to show the fluctuation in price compared to the trading activity of congresspersons. In addition, to compare the performance against a major market index, we reference SP500 historical daily quote and monthly return.

Data Cleaning and Wrangling

Direct Disclosures

We first created a composite dataset of both the House and Senate disclosure data. We standardized some slight differences in schema between the two and created an indicator variable that denoted whether a row belonged to the House or Senate dataset. This data was filtered to include only dates after January 1, 2020.

Even though each table had the same naming conventions for the amount, type, and owner variables there were some text differences. We had to remove spacing, replace capital letters with lower case, and remove misspelled entries.

Once we had cleaned the Senate and House dataset, we merged it with the historical stock prices and returns. We joined the two datasets on the transaction date and stock ticker, so that we had a closing price and daily return for each transaction.

Along with the daily information, we also calculated a few other fields to be used as dependent variables:

- Stock return for the month that the transaction was completed
- Market return for the month that the transaction was completed (using the Dow Jones Industrial Average to measure market performance)
- Stock return over the 3 months following the transaction date
- Market return over the 3 months following the transaction
- Binary variables for both the stock and overall market indicating a positive return over a 1-month and 3-month time frame following the transaction date

Finally, in addition to the daily data, we created a monthly view of the data by aggregating the daily transactions by month and stock. This allowed us to train models based upon the net investment amount and total trades by Members of Congress and Senators over the course of a given month.

PDF Disclosures

During our exploration of the data, we noticed that some Senators had disclosures in two forms: a direct reporting of stock transactions, and a PDF disclosure that is less easily read and analyzed. This PDF version of disclosure seemed like a method for senators to obscure their trading activities, because a person would be forced to manually inspect every PDF instead of being given a clean CSV file. We decided to focus on one Senator, John Boozman. We discovered that since Aug 23rd, 2019, John Boozman submitted two disclosures directly and submitted 12 disclosures via PDF. A member of our team dug through these 12 PDF disclosures and manually created a CSV file containing this data. These PDF disclosures outlined 167 transactions from August 2019 to September 2022. These transactions involved 70 unique stock tickers and occurred on 30 unique days of trading.

Issues with this data:

- PDF disclosures often appeared messy and irregular. They often appeared to be hand-typed, with misspellings, abbreviations, and sometimes with the appearance that someone cut out strips of paper and photocopied them into one page.
- The stock ticker & typical designation of the type of investment was rarely used and had to be found manually
- Some of the mutual funds, ETFs, index funds, etc. were not able to be queried programmatically. This stood in the way of a large-scale algorithmic analysis of these more obscure funds.

Data Enrichment

The raw data for the majority of congressional stock disclosures was simple to acquire as a CSV file. However, it was very difficult to locate other pertinent information about a congressperson that gave a better picture of their activities and affiliations. If this information was available, it was never packaged in an easily processable form. The specific information we were seeking included the following: a congressperson's committee membership, bill sponsorships, lobbyist relationships, certain forms of private business dealings, among others. If this information were available, it would have allowed us to approach this problem from many additional perspectives.

Data Issue for Analysis and Modeling

One issue we have from data is the disclosure does not disclose the exact amount for the trades, it lists the amount range of the trade, see the figure below for trading amount range frequency table.

Table 1. Frequency of Transaction Amounts

Amount Range	**Frequency**
$1,000 - $15,000	12629
$15,000 - $50,000	3119
$50,001 - $100,000	1006
$100,001 - $250,000	776
$250,001 - $500,000	336
$500,001 - $1,000,000	205

$1,000,000 - $5,000,000	100
$5,000,001 - $25,000,000	19
$25,000,000 - 50,000,000	0
$50,000,001 +	1
Unknown	22

Our modeling would appreciate the exact amount of the trade, we construct the continuous variable by taking the lower bound of the range and mid-point of the range for our analysis and modeling.

Models

Linear Regression and Logistic Regression

In the first part of our modelling effort, we primarily focused on two types of models, Linear Regression and Logistic Regression. Our linear regression models were aimed at predicting the percent return of a stock or the market over a given period, while our logistic regression models aimed at predicting the probability that a stock or the market would see a positive return. To train these models, we used the following independent variables:

- Amount of the transaction (Positive for a purchase, negative for a sale)
- Net amount traded during a specified period (Over 1 month, 3 months, etc.)
- Gross volume of trades over a given period
- Net amount traded during a prior or post period
- Gross volume of trades over a prior or post period
- Gross volume of Sales and Purchases over a period
- The branch (Senate or Congress) of the representative making the trade
- The number of days between the transaction date and the disclosure date
- Whether the representative was among the top traders

And we combined these with the following dependent variables:

- Linear Regression
    - Stock return during the month of the transaction
    - Market return during the month of the transaction
    - Stock return over the 3 months following the transaction
    - Market return over the 3 months following the transaction
- Logistic Regression:
    - Did the stock see a positive return during the month of the transaction? (0 or 1)
    - Did the market see a positive return during the month of the transaction? (0 or 1)
    - Did the stock see a positive return over the 3 months following the transaction? (0 or 1)
    - Did the market see a positive return over the 3 months following the transaction? (0 or 1)

However, by the end of our analysis we were unable to build a model that accurately predicted either an individual stock's performance or the performance of the stock market as a whole. While we were able to produce some linear regression models with a high adjusted R-squared value, these models were for specific stocks that did not have enough data to be statistically significant. One of the better performing linear regression models that we were able to produce is pictured below, it used the trading activity of the prior 3 months (volume and dollars) to predict the three-month return for individual stocks. The adjusted R-squared value was only 0.22:

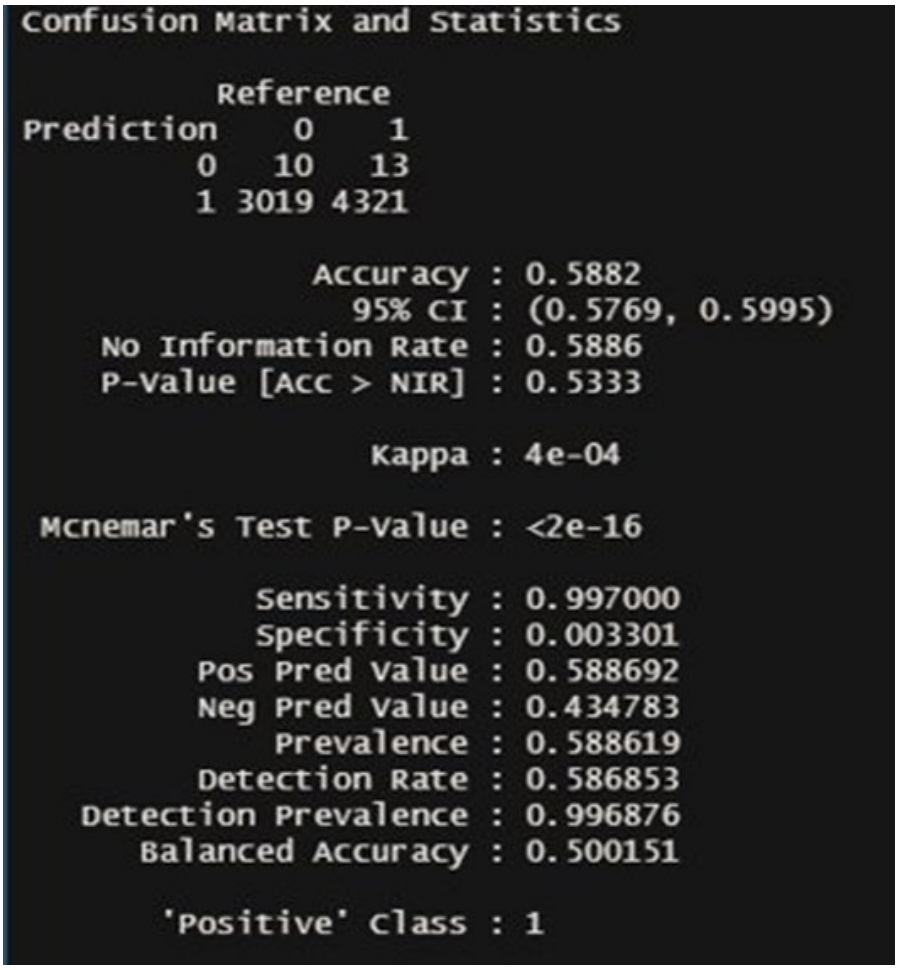
```
Call:
lm(formula = threeRet ~ volume + prior1vol + prior2vol + amount +
    prior1amt + prior2amt, data = stock_returns)

Residuals:
     Min       1Q   Median       3Q      Max
-19.5255  -7.3936   0.6277   5.9650  19.1499

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.464e+00  3.880e+00  -1.666   0.1087
volume       4.677e-01  3.236e-01   1.446   0.1612
prior1vol    3.130e-01  3.040e-01   1.030   0.3135
prior2vol    6.221e-01  2.869e-01   2.169   0.0403 *
amount      -1.639e-06  9.722e-07  -1.686   0.1048
prior1amt   -1.636e-06  1.005e-06  -1.627   0.1168
prior2amt   -2.582e-07  1.013e-06  -0.255   0.8010
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.58 on 24 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.3774,    Adjusted R-squared:  0.2217
F-statistic: 2.425 on 6 and 24 DF,  p-value: 0.05646
```

A few of our logistic regression models proved a bit more successful, however we were still unable to produce any statistical significance. At first it appeared that some of our logistic regression models were accurate at predicting market performance, however upon further investigation we noticed that the models were almost always predicting a positive market return. Since the market saw positive returns for most of the period that we analyzed, the model's accuracy was overstated. When we applied that model to other periods, or sub-selections of the dataset, the accuracy fell sharply. See the image below. This model was showing upwards of 75% accuracy on the training dataset, but once we tested the model on the test dataset the accuracy dropped to 58%, and we can see the model predicted a positive return in 99% of cases:



The data set we used that combined the Senate and House data was used in the initial models. Some of the initial models were simple linear regression. Using our initial hypotheses, we thought we could find some connection between the timing of when the transaction and then the disclosure took place. Using the difference in the time in days between the transaction and the disclosure we regressed that with what type of ownership the security had with a joint ownership being the prevailing type. The amount that was transacted, which was put into bands because the exact amount was not disclosed. And the type of security that was transacted. Running the first model showed nothing of significance. Here is the summary:

```
================================================
                        Dependent variable:
                    ----------------------------
                            time_trans_disclo
------------------------------------------------
owner_indicator               -13.666***
                               (0.725)

amount_indicator               -5.385***
                               (0.780)

type_indicator                  1.868*
                               (0.955)

Constant                       77.353***
                               (2.433)

------------------------------------------------
Observations                    18,191
R2                               0.024
Adjusted R2                      0.023
Residual Std. Error     118.040 (df = 18187)
F Statistic          146.364*** (df = 3; 18187)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

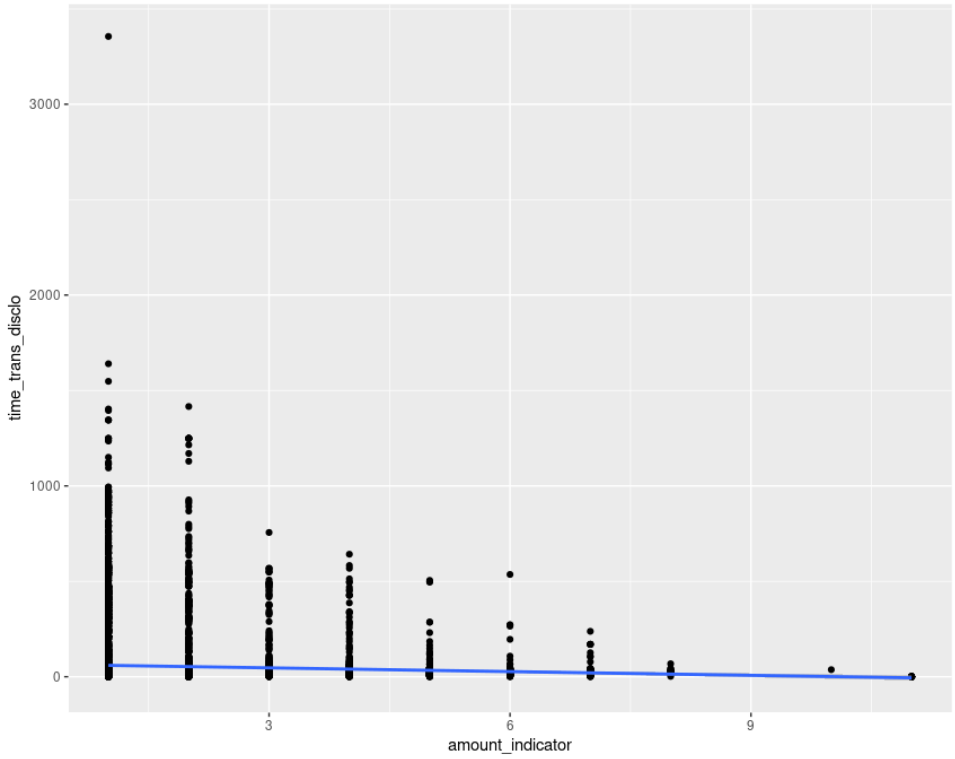
Next step was to regress using only 2 variables the time and the amount and see if there was any significance. Here is this summary:

```
================================================
                        Dependent variable:
                    ----------------------------
                            amount_indicator
------------------------------------------------
time_trans_disclo              -0.001***
                               (0.0001)

Constant                       1.628***
                               (0.010)

------------------------------------------------
Observations                    18,213
R2                               0.004
Adjusted R2                      0.004
Residual Std. Error      1.170 (df = 18211)
F Statistic           75.114*** (df = 1; 18211)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Here is a visual:



The idea that the longer the time between the transaction and the disclosure would indicate a higher amount of the transactions does is not significant. A challenge with this model it that when a disclosure is made many securities are listed and thus they all have the same amount of time.

A future model may include selecting only the security that has the largest amount in a transactions.

PDF Modelling

The approach we took to investigate Senator John Boozman's PDF disclosures was to see if the trades made by John Boozman came from the same distribution as the remainder of the data. We grouped the trades into categories, grouped by stock ticker symbol, transaction type, and spending amount. Then we compared the counts found in the PDF disclosures compared to the Direct Disclosures. None of the p-values of the chi-squared tests were significant. Therefore, we fail to reject the null hypothesis and carry on with the assumption that John Boozman's trades submitted by PDF follow the same distribution as the rest of the Senate disclosures.

Table 2. Chi-Squared Goodness of Fit Test

Grouped by stock ticker symbol, transaction type (Purchase or Sale), and spending amount (categorical ranges)

Chi-Squared Test Statistic	155.928
p-value	0.386

alpha	0.05
Degrees of freedom	105

Table 3. Chi-Squared Goodness of Fit Test

Grouped by stock ticker symbol, transaction type (Purchase or Sale)

Chi-Squared Test Statistic	219.43
p-value	0.25
alpha	0.05
Degrees of freedom	92

Performance Against SP500

In the last part of our modeling effort, we aim to determine if the stocks traded by members of the Congress have superior overall performance against the market index SP500.

We look at the three types of trade (purchase, sale full, and sale partial), and we look at both short term (2 trading days) and long term (1 month return). We define 2 trading days as short term and 1 month as long term mainly because generally a trade cannot be considered as insider trading if it is made 2 days after public release. We define 1 month as long term for this analysis because if there were insider trading activity, 1 month is long enough to realize the gain from information advantage.

The six performance comparisons we look at are: purchase within 2 trading days, purchase within 1-month, full sale within 2 trading days, full sale within 1-month, partial sale within 2 trading days, and partial sale within 1 month.

In this part of the analysis and modeling, the stock performance takes the amount traded into account. In a layman term, if the member of Congress purchases $1000 on Stock A, we compare the return of such trade against the return of SP500 within two trading days, and 1 month. If it is a sale (full sale or partial sale), same comparison, but when we interpret the results of sale and purchase separately because the "performance" has the opposite indication, that is—if one believes there is illegal trading activities, the stocks purchased by members of Congress should overperform SP500, the stocks sold should underperform SP500.

The return of SP500 is known, but the amount traded is from our data, therefore we conduct two-sample t test instead of one sample t test, we do not assume the two performances have the same variance.

Purchase – 2 days

Testing parameters:	mean of the ticker return within 2 trading days μ_1 mean of the SP500 return within 2 trading days μ_2
Null Hypothesis:	$\mu_1 - \mu_2 = 0$
Alternative Hypothesis:	$\mu_1 - \mu_2 > 0$
Testing Results:	p-value = 0.2745, do not reject null hypothesis

(No evidence purchased stocks overperformed SP500)

Purchase – 1 month

Testing parameters: mean of the ticker return in a month $\mu 1$
mean of the SP500 return in a month $\mu 2$
Null Hypothesis: $\mu 1 - \mu 2 = 0$
Alternative Hypothesis: $\mu 1 - \mu 2 > 0$
Testing Results: p-value = 0.9611, do not reject null hypothesis
(No evidence purchased stocks overperformed SP500, additional testing suggests purchased stock underperformed SP500)

Full Sale – 2 days

Testing parameters: mean of the ticker return within 2 trading days $\mu 1$
mean of the SP500 return within 2 trading days $\mu 2$
Null Hypothesis: $\mu 1 - \mu 2 = 0$
Alternative Hypothesis: $\mu 1 - \mu 2 < 0$
Testing Results: p-value = 0.9948, do not reject null hypothesis
(No evidence purchased stocks overperformed SP500)

Full Sale – 1 month

Testing parameters: mean of the ticker return in a month $\mu 1$
mean of the SP500 return in a month $\mu 2$
Null Hypothesis: $\mu 1 - \mu 2 = 0$
Alternative Hypothesis: $\mu 1 - \mu 2 < 0$
Testing Results: p-value = 0.0211, reject Null Hypothesis
(Evidence suggests purchased stocks underperformed SP500)

Partial Sale – 2 days

Testing parameters: mean of the ticker return within 2 trading days $\mu 1$
mean of the SP500 return within 2 trading days $\mu 2$
Null Hypothesis: $\mu 1 - \mu 2 = 0$
Alternative Hypothesis: $\mu 1 - \mu 2 < 0$
Testing Results: p-value = 0.0271, reject null hypothesis
(Evidence suggest purchased stocks overperformed SP500)

Partial Sale – 1 month

Testing parameters: mean of the ticker return in a month $\mu 1$
mean of the SP500 return in a month $\mu 2$
Null Hypothesis: $\mu 1 - \mu 2 = 0$
Alternative Hypothesis: $\mu 1 - \mu 2 < 0$
Testing Results: p-value = 0.8564, do not reject null hypothesis
(No evidence Purchased stocks overperformed SP500)

In summary, our findings are in Table 4.

Table 4. Performance Comparison between Traded Stocks and SP500

Overall Performance Against SP500	Within 2 Trading Days	Within 1 Month
Purchase	No evidence for overperformance	Underperformed SP500
Full Sale	Overperformed SP500	Underperformed SP500 *

Partial Sale	Underperformed SP500 *	No evidence for underperformance

* Out of the six scenarios, only two suggest the testing results align with the claim of illegal trading.

Based on the results, there is no consistent evidence in our data suggests the stocks traded by member of congress underperformed SP500 when they are purchases, or they underperformed SP500 when they are sales. Some testing results above even suggest members of Congress are not good at picking stocks.

Conclusion

Ultimately, we were unable to find a significant correlation between congressional trading activity and the performance of the stock market. It appears there is not a widespread abuse of power among legislators when it comes to trading based on the data we have. Additionally, we learned that using statistical methods to identify trends for individual representatives is challenging due to the limited data available and the small sample sizes.

With more time we would like to look at individual representatives, especially those who have large amounts of trading. The challenge would be to find news events around the time of the transaction and the disclosure. Locating a specific source for the news is challenging. If we were able to find the news data and create models and find any significance, in theory we may be able to forecast a security's movement.

Furthermore, we suggest further investigation into the relationship between a congressperson's committee membership, bill sponsorship, campaign donations, interaction with lobbyists, among other behaviors. This data was difficult to track down, and when located it was very cumbersome to analyze. As part of the bill mandating congressional disclosures of business activities, amendments should be made to standardize the format of this information into an easily downloadable CSV file.

References

Gnägi, M., & Strub, O. (2020). Tracking and outperforming large stock-market indices. *Omega (Oxford)*, *90*, 101999–. https://doi.org/10.1016/j.omega.2018.11.008

Li, M., Liu, D., Peng, H., & Zhang, L. (2022). Political connection and its impact on equity market. *Research in International Business and Finance*, *60*, 101593–. https://doi.org/10.1016/j.ribaf.2021.101593

McCoy, M. S., Bonci, M., Joffe, S., & Kanter, G. P. (2021). Historical trends in health care-related financial holdings among members of Congress. *PloS One*, *16*(7), e0253624–e0253624. https://doi.org/10.1371/journal.pone.0253624