# Capturing Mutual Fund Holdings: A Network Analysis with Machine Learning

(very preliminary draft)

Xingkong Wei*

May 12, 2018

## Introduction

### Motivation

Is it possible to predict mutual fund managers' behaviours with public information? It is common that some mutual fund managers pick up the same stocks for their portfolios. But there are also some stocks chosen by only a few mutual funds. For a given period, will these less chosen stocks be picked up by other mutual funds in the next period? If so, what kind of stocks are more likely to be selected by other mutual funds? If mutual funds managers are well-informed, those stocks with less mutual funds coverage may be a signal of undervalued stocks because those fairly priced stocks should have more mutual funds coverage and those trash stocks should have no mutual funds coverage.

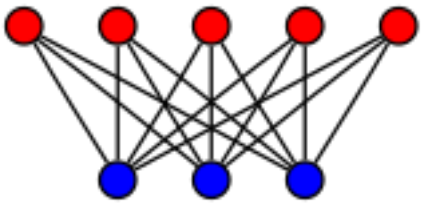
If we can predict which stocks would be covered by other mutual funds in the next period, we can build a strategy on that. In this project, I plan to use mutual funds holding networks to do some data mining and use machine learning techniques to predict whether some less covered stocks would be covered by other mutual funds. Just to clarify, these less covered stocks mean that they are selected by at least one mutual fund.

*Fintech Institute,Tsinghua PBCSF. email:weixk@pbcsf.tsinghua.edu.cn

# Networks

## Bipartie Networks

Bipartite graphs B = (U, V, E) have two node sets U,V and edges in E that only connect nodes from opposite sets. It is common in the literature to use an spatial analogy referring to the two node sets as top and bottom nodes.



## Network Features

**Betweenness Centrality** Betweenness centrality of a node $v$ is the sum of the fraction of all-pairs shortest paths that pass through $v$:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where $V$ is the set of nodes, $\sigma(s,t)$ is the number of shortest $(s,t)$-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node $v$ other than $s,t$. If $s=t$, $\sigma(s,t)=1$, and if $v \in s,t$, $\sigma(s,t|v)=0$

**Closeness Centrality** The closeness of a node is the distance to all other nodes in the graph or in the case that the graph is not connected to all other nodes in the connected component containing that node. Thus the closeness centrality for node $v$ in the two bipartite sets $U$ with $n$ nodes and $V$ with $m$ nodes is

$$c_v = \frac{m+2(n-1)}{d}, \text{for} v \in U,$$
$$c_v = \frac{n+2(m-1)}{d}, \text{for} v \in V,$$

where $d$ is the sum of the distances from $v$ to all other nodes. Higher values of closeness indicate higher centrality.

**Degree Centrality** Compute the degree centrality for nodes in a bipartite network.

The degree centrality for a node $v$ is the fraction of nodes connected to it.

**Page Rank** PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages.

**Clustering** Compute a bipartite clustering coefficient for nodes.

The bipartie clustering coefficient is a measure of local density of connections defined as [R158]:

$$c_u = \frac{\sum_{v \in N(N(v))} c_{uv}}{|N(N(u))|}$$

where $N(N(u))$ are the second order neighbors of $u$ in $G$ excluding $u$, and $c_{uv}$ is the pairwise clustering coefficient between nodes $u$ and $v$.

The mode selects the function for $c_{uv}$ which can be:

*dot*:

$$c_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

*min*:

$$c_{uv} = \frac{|N(u) \cap N(v)|}{min(|N(u)|, |N(v)|)}$$

*max*:

$$c_{uv} = \frac{|N(u) \cap N(v)|}{max(|N(u)|, |N(v)|)}$$

# Data

Table 1: 变量名称对应表

| Variable | 变量名称 |
|---|---|
| *holding_shares* | 持股占流通股比(%) |
| *coverage_funds* | 持有基金数 |
| *total_share_holdings* | 持股总量 |
| *seasonal_holding_change* | 季度持仓变动(万股) |
| *total_market_value* | 持股总市值(万元) |
| *buy_sec* | 券商评级：买入 |
| *hold_sec* | 券商评级：持有 |
| *neutral_sec* | 券商评级：中性 |
| *pledge_parity* | 股票质押方数量 |
| *pledge_shares* | 股票质押数量 |
| *pledge_market_value* | 股票质押市值 |
| *y* | 新增持有基金数 |
| *y_degree* | 是否有新增基金公司持有 |

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| holding_shares | 1215.0 | 2.916783 | 4.332290 | 0.000007 | 0.207457 | 1.231583 | 3.618788 | 3.503806e+01 |
| coverage_funds | 1215.0 | 22.046091 | 59.820513 | 1.000000 | 2.000000 | 6.000000 | 16.000000 | 8.570000e+02 |
| total_share_holdings | 1215.0 | 3760.189396 | 11270.007651 | 0.005000 | 115.598350 | 740.184800 | 2788.640400 | 1.744305e+05 |
| seasonal_holding_change | 1215.0 | 1092.785143 | 3604.799098 | -6836.756000 | 5.000000 | 131.062400 | 816.770700 | 4.843970e+04 |
| total_market_value | 1215.0 | 73774.233220 | 266745.509142 | 0.052600 | 1770.224400 | 11105.829000 | 44352.233239 | 4.923366e+06 |
| closeness_fund | 1215.0 | 0.308970 | 0.047998 | 0.252714 | 0.275803 | 0.292397 | 0.327471 | 5.083333e-01 |
| betweenness_fund | 1215.0 | 0.000580 | 0.001754 | 0.000000 | 0.000002 | 0.000031 | 0.000249 | 1.888093e-02 |
| pagerank_fund | 1215.0 | 0.000375 | 0.000427 | 0.000113 | 0.000147 | 0.000212 | 0.000407 | 3.252711e-03 |
| coverage_fund | 1215.0 | 8.931687 | 13.341582 | 1.000000 | 2.000000 | 4.000000 | 10.000000 | 9.800000e+01 |
| clustering_fund | 1215.0 | 0.123512 | 0.035114 | 0.048889 | 0.101872 | 0.114503 | 0.134316 | 2.741882e-01 |
| degree_centrality_fund | 1215.0 | 0.078348 | 0.117031 | 0.008772 | 0.017544 | 0.035088 | 0.087719 | 8.596491e-01 |
| buy_sec | 1215.0 | 0.607536 | 0.294299 | 0.000000 | 0.466667 | 0.666667 | 0.812500 | 1.000000e+00 |
| hold_sec | 1215.0 | 0.383802 | 0.291972 | 0.000000 | 0.175192 | 0.333333 | 0.500000 | 1.000000e+00 |
| neutral_sec | 1215.0 | 0.008619 | 0.059694 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000e+00 |
| coverge_sec | 1215.0 | 10.076543 | 7.735212 | 1.000000 | 4.000000 | 8.000000 | 15.000000 | 3.900000e+01 |
| clustering_sec | 1215.0 | 0.166799 | 0.028386 | 0.084286 | 0.144685 | 0.165294 | 0.189484 | 2.424944e-01 |
| degree_sec | 1215.0 | 0.157534 | 0.123521 | 0.021277 | 0.063830 | 0.127660 | 0.234043 | 6.808511e-01 |
| closeness_sec | 1215.0 | 0.380176 | 0.058516 | 0.258773 | 0.334394 | 0.386454 | 0.429058 | 4.939287e-01 |
| betweenness_sec | 1215.0 | 0.000318 | 0.000444 | 0.000000 | 0.000015 | 0.000128 | 0.000461 | 3.131027e-03 |
| pledge_parity | 1215.0 | 4.332510 | 7.984936 | 0.000000 | 0.000000 | 1.000000 | 5.000000 | 9.100000e+01 |
| pledge_shares | 1215.0 | 11409.119102 | 35856.345619 | 0.000000 | 0.000000 | 1200.000000 | 10396.760000 | 8.702692e+05 |
| pledge_market_value | 1215.0 | 154751.272958 | 386529.408318 | 0.000000 | 0.000000 | 28516.000000 | 174321.459155 | 7.066903e+06 |
| y | 1215.0 | 3.639506 | 4.850152 | 0.000000 | 0.000000 | 2.000000 | 5.000000 | 3.400000e+01 |
| y_degree | 1215.0 | 0.715226 | 0.451492 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000e+00 |

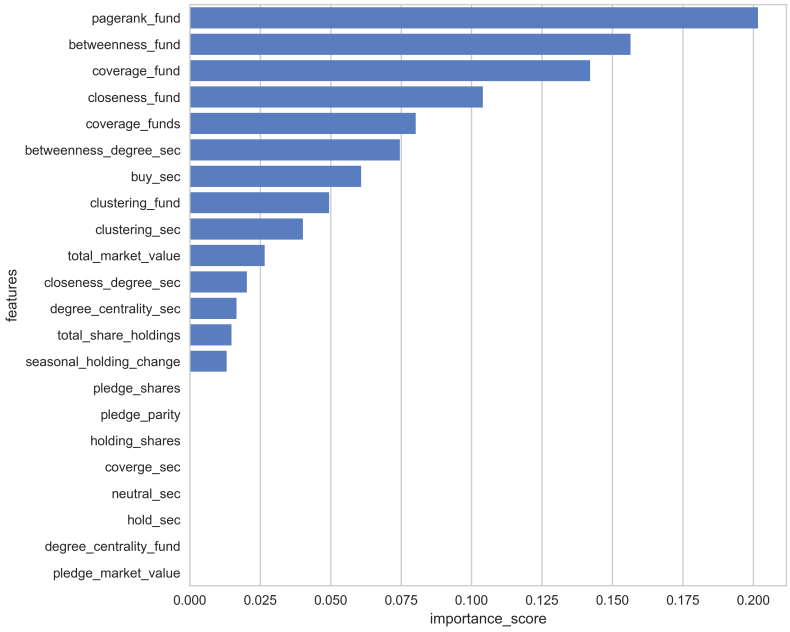
# Features Selection

## Random Forest Classifier

For those who are interested in algorithms, I suggest to look at the following link.

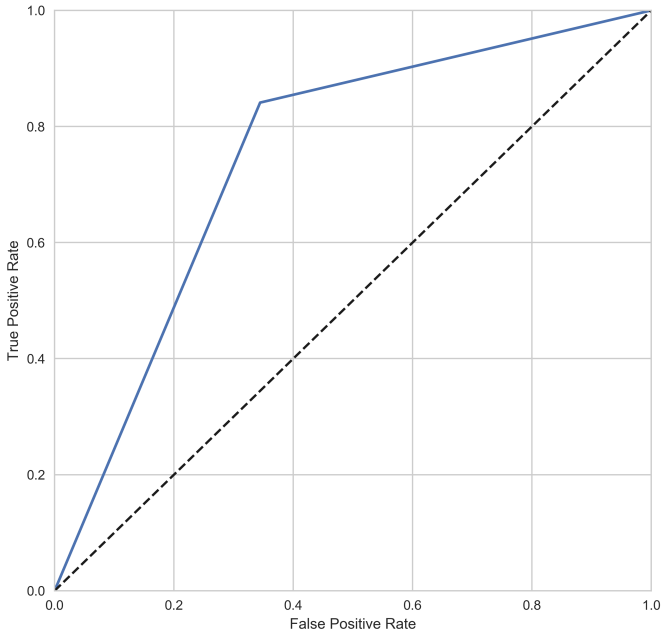
Random Forest : *https://en.wikipedia.org/wiki/Random_forest*

## Feature Importance




As we can see, the top five most important features come from mutual fund holding networks

## ROC Curve



# Potential mutual funds holding predictions

Table 2: Predictions with Machine Learning Algorithms

| Classifier | accuracy_score |
|---|---|
| *DGBT* | 0.75 |
| *GradientBoostingRegressor* | 0.744 |
| *XGBoost* | 0.786 |
| *NeuralNetwork* | 0.71 |
| *SVM* | 0.707 |
| *LogisticRegression* | 0.766 |
| *SDG* | 0.76 |
| *DecisionTree* | 0.786 |
| *RandomForest* | 0.76 |

# Empirical Strategy

$H_0$: Stocks' network features are related to potential mutual funds holdings.

$$y_i = \alpha + \beta_1 Network_i + \varepsilon_i$$

where $Network_i$ includes $pagerank\_fund$, $betweenness\_fund$, $closeness\_fund$, $betweenness\_degree\_sec$, $clustering\_fund$. The features with fund postfix stand for network features extracting from fund-stock network; the features with sec postfix stand for network features extracting from securities company-stock network.

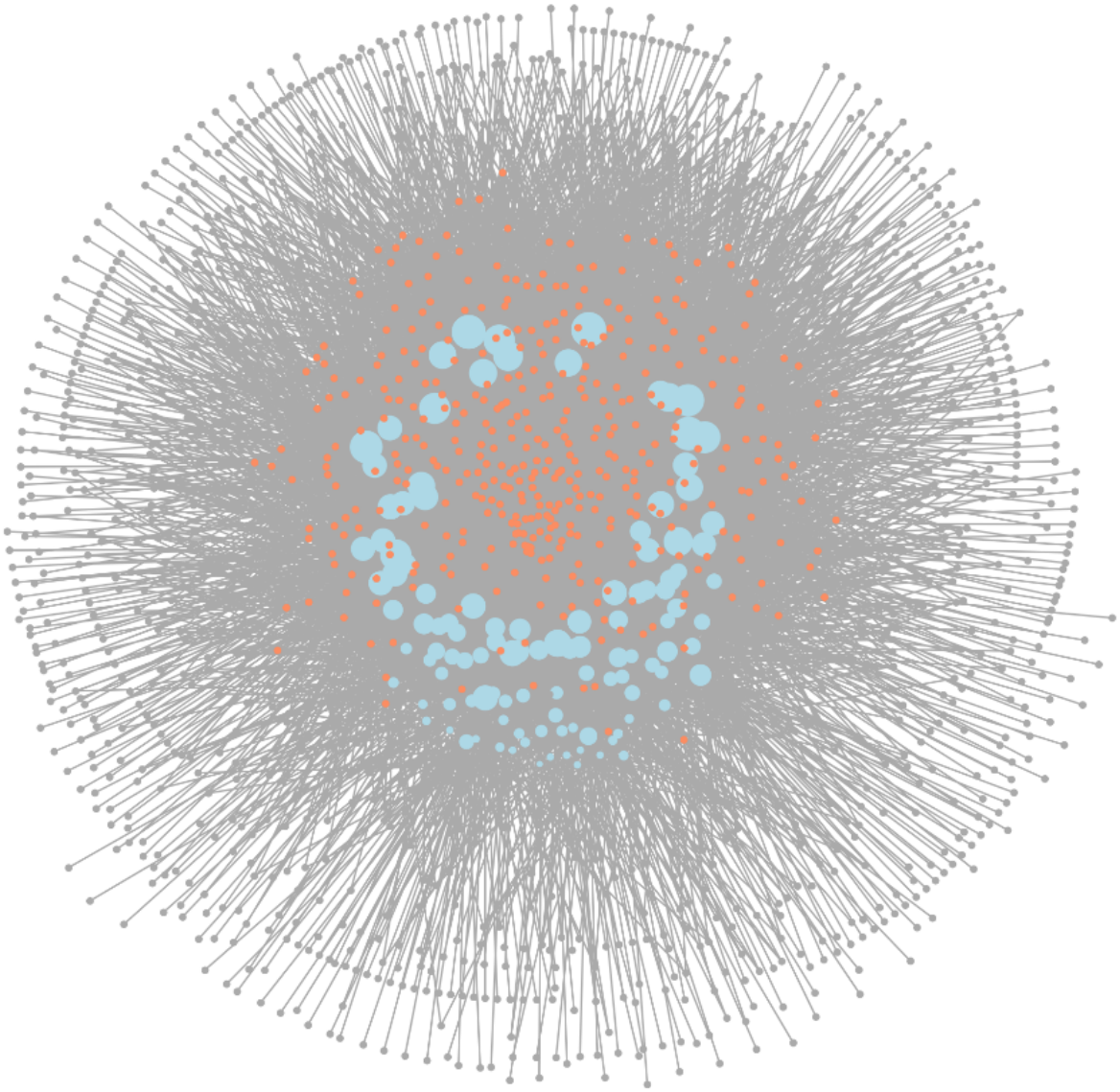
Table 3: Network features and potential mutual fund holdings

| | y | | | | |
|---|---|---|---|---|---|
| pagerank_fund | 5,007.708*** | 15,662.260*** | 10,677.290*** | 7,656.825*** | 7,384.123*** |
| | (525.578) | (943.085) | (2,484.185) | (2,527.049) | (2,514.690) |
| betweenness_fund | | −2,771.450*** | −2,227.508*** | −1,865.879*** | −1,805.174*** |
| | | (218.405) | (332.349) | (338.464) | (337.424) |
| closeness_fund | | | 27.320** | 27.940** | 26.694** |
| | | | (11.978) | (11.665) | (11.705) |
| betweenness_degree_sec | | | | 2,568.256*** | 2,528.955*** |
| | | | | (447.270) | (447.640) |
| clustering_fund | | | | | −6.043*** |
| | | | | | (2.022) |
| Constant | 1.762*** | −0.626*** | −7.513** | −7.599*** | −6.388** |
| | (0.183) | (0.200) | (2.972) | (2.893) | (2.992) |
| $N$ | 1,215 | 1,215 | 1,215 | 1,215 | 1,215 |
| $R^2$ | 0.194 | 0.319 | 0.322 | 0.357 | 0.358 |
| Adjusted $R^2$ | 0.194 | 0.318 | 0.321 | 0.355 | 0.356 |
| Residual Std. Error | 4.355 | 4.006 | 3.998 | 3.896 | 3.894 |
| F Statistic | 292.870*** | 283.886*** | 192.036*** | 167.762*** | 134.960*** |

*Notes:*

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Here we use feature importance score from random forest algorithm to help us select independent variables in the regression.



Blue nodes are mutual funds, gray nodes are stocks hold by mutual funds, orange nodes are stocks that are hold by more than 7 mutual funds