To Mirror or Not to Mirror – Modeling Relationships in Social Trading

By Juye Shirley Lu
B.S. Electrical Engineering and Computer Science and B.S.
Mathematics, Massachusetts Institute of Technology (2015)

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Author: . . . .
Department of Electrical Engineering and Computer Science July 28, 2017

Certified by: . . . .
Alex "Sandy" Pentland, Toshiba Professor of Media Arts and Science
Thesis Supervisor
July 28, 2017

Certified by: . . . .
Esteban Moro
Thesis Co-Supervisor
July 28, 2017

Accepted by: . . . .
Christopher Terman, Chairman, Masters of Engineering Thesis Committee

To Mirror or Not to Mirror – Modeling Relationships in Social Trading

by
Juye Shirley Lu

Submitted to the Department of Electrical Engineering and Computer Science
On July 28, 2017
In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract

Many scholars have applied ecological principles to study the financial market. As early as 1940s, John Maynard Keynes coined the term “animal spirits” to describe human decision making under uncertainty. In modern economic terms, “animal spirits” are often used to describe the psychological factors that drive investors decision making during volatile market. Many scholars used Darwin’s evolutionary theory to explain evolution of investment strategies [5] However, few studied leader election, individual adaptation, and social dynamics in the financial market. This lack of research is mostly due to a lack of centralized research entities to implement large-scale experiments. Luckily, a new investment mechanism, social trading, where investors can interact with each other by mirroring and commenting on each other’s trade ideas, provided a new avenue to study evolution of a new financial system. We are able to observe how leaders become leaders, how followers choose their leaders, and how different groups interact with each other. Our research takes place on one of the biggest platforms of this kind, eToro, a retail social trading platform in foreign exchange and other asset markets. Treating this economic system almost as a new ecological environment, we begin with understanding who are the different players and how they interact with each other. We categorize traders based on their investing styles and observe how their types change over time. Interestingly, these profiles resemble major players in the financial market: diversified institutional investors, speculators, and specialized strategy (macro and value) funds. Then we try to understand why some leaders have more followers than others and train a model to predict whether a leader will get a new followers/unfollowers on a particular day. Build upon existing literature, we found that not only can leader’s style factors predict whether he gets new followers/unfollowers, popularity rank, average performance of his followers, and recent maximum gains also have predictive power. Our models are trained using SMOTE-balanced training sets and are able to achieve roughly 80%-90% accuracy. Lastly, we take a microscopic view of how followers follow. We claim that followers exhibit “foraging” pattern when choosing their leaders. Followers create people-portfolios, and foraging is essentially equivalent to diversification. By foraging, followers can prevent significant losses regardless of which type of investor they are. However, foraging would not lead to outstanding gains, or alpha per se. Traders who forage are analogous to index-following investors who track the market.

Thesis Supervisor: Alex "Sandy" Pentland
Toshiba Professor of Media Arts and Science
Thesis Supervisor: Esteban Moro
Associate Professor at Departamento de Matemáticas, Universidad Carlos III de Madrid, Spain

Acknowledgements

MIT has always had a special place in my heart because I met so many brilliant yet down-to-earth people through the institute. I would like to take this opportunity to acknowledge and express my sincerely gratitude to all of them here.

First, I would like to thank the EECS department at MIT and the Human Dynamics Group at the Media Lab. It was not an easy decision to leave my previous role at Citadel and come back to MIT to conduct research for my Master's of Engineering degree. However, the EECS department was incredibly supportive of my decision and provide tremendous financial support and mentorship. Additionally, the Human Dynamics Group has been supportive of my research efforts ever since when I was an undergraduate researcher (UROP) at the lab. Both my undergraduate research on Mobility in Innovation Teams and this research on Social Trading were funded by the group through research assistantship. My education would not have been complete and this paper would not have been possible without their generous support.

Next, I would like to thank my supervisors, Sandy Pentland and Esteban Moro, for providing such an engaging and interesting topic to work with and for providing guidance throughout the project. Behavioral finance has always been my personal interest and career direction. This research project has allowed me to apply the statistical and machine learning tools that I learned as an undergraduate and master's student at MIT to understanding an active area of research in the financial market. My two mentors not only taught me rigorous research methodology, but also encouraged me to think outside the box during the research process. Many interesting ideas were exchanged and many new areas of research emerged during our conversations. It is unfortunate that I only had one year to work with these two giants in their fields, but fortunately they have become life-long mentors. I am confident that our common passion in research will extend beyond MIT.

I would also like to thank Xiaowen Dong and Peter Krafft, two graduate students who have been supervising me on this project. There were many hurdles over the course of the paper, without their advice and encouragement, the project would not have been possible.

Finally, I would like to thank my parents for the sacrifices they made along the way to ensure the greatest education I can ever have. Thank you for always being there for me.

Contents

List of Figures

Chapter 1

Introduction to Social Trading and eToro

This chapter gives an overview of social trading and eToro as a trading platform. We first introduce the concept of social trading then explain how investors can use the eToro platform to manage relationships and social trade. We introduce the data we used for this paper and the prediction problem we are trying to solve. Finally, we describe how we cleaned, sampled, and split the data to solve our prediction problem.

1.1 Introduction

Applications of computational social science have gone beyond traditional social contexts such as Facebook and Twitter. These research has made significant contributions in public policy, healthcare, and technology, but few in finance [3]. Beginning 2016, for example, crowded trades were one of financial phenomena that cannot be explained by traditional economic consumption models; as many beta factors such as value and growth become known, trade ideas are becoming increasingly correlated. However, very few sell-side research firms had the empirical data to conduct any hypothesis testing in order to provide an explanation. Many investors, both private and institutional alike, have recognized the importance of orthogonal trade ideas and direction of crowd wisdom when making investment decisions.

Thanks to innovations in online trading platforms, we now have empirical data to study social dynamics, specifically the mirroring behavior, in financial systems. Similar to social media platforms, a social trading platform brings professional traders and enthusiasts on a single environment to connect and interact. Social traders not only create multi-asset portfolios but also build “people-based” portfolios. Participants on such sites can post information, relate experiences, and read the latest news; view and analyze the performance of others; and copy trades from others. This provides a unique opportunity to conduct large-scale social-economic research and draw inference on individual behavior.

This paper takes a big-data approach to address why and how investors social trade. If users come to social trading platforms due to cash incentives, then social trades should be more profitable than original trades, which we confirm to be true. We noticed that investors adopt the social trading concept through a “explorer-to-keeper” transformation. When one is first introduced to the concept, he/she follows and unfollows various leaders of the crowd to assemble the appropriate profile to follow. We discover that how actively one explores depends on his portfolio size, risk appetite, trade horizon, proportion of mirror trades in portfolio, daily trade volume, experience with social trading, and cognitive capacity. Then explorers either become keepers or skeptics. In the case of keepers, their mirror preferences converge to three common “leaders-of-the-crowd” profiles: active fund managers overseeing large, diversified portfolios, risk-seeking, short-term speculators betting on momentum, and traders specializing in a few sectors or strategies. This close resemblance between leaders on eToro and major market players provides a powerful piece of evidence to crowd wisdom’s role in shaping the market landscape today.

Despite our best effort at modeling, our data is not sufficient to claim that social trading caused a change in behavior. We propose an experiment to the eToro platform to assess whether more social trading makes one more social.

1.2 Social Trading

The social trading concept is not novel. Investors used to follow successful traders via a plethora of newsletters and newspaper columns, and later, email. There were also physical establishments such as investment clubs, where people met, pooled their funds, and debated investments. With the advent of social media and online brokerage platforms, today’s version of social trading became more sophisticated than the email newsletter, providing real time data and trades, phone apps and cutting-edge technology with the click of a mouse.

Social trading gives those with limited financial knowledge and capital an opportunity to participate in the market. It builds on the concept that the collective wisdom of thousands of traders is better than the wisdom of one, taking full advantage of user-generated content to generate trade ideas. Social trading sites provide their users with a variety of community-based tools in order to give them the information needed for making smarter investment and trading decisions. For example, a social trading platform would enable its users to see other users’ portfolios, read their news feeds, and look at their overall performance, to gain a better understanding of trading strategy. Anybody with $2,000 or sometimes considerably less can learn from star investors and piggy-back on their investment strategies. This decentralization of financial information and trading tools is one of the most significant shifts in retail investing. To date, $4 trillion market cap on forex market is estimated to be related to social trading [2]. Eighteen major stock exchanges are now accessible to social trading. Top sites are ZuluTrade, eToro, and Ayondo. Most platforms offer a range of investment portfolios to mirror – shares, indices, commodities and Forex seem the most common.

There are two ways investors can use social trading sites -- as a source of information or to mirror the trades of ‘experts’. Investors have monetary incentive to mirror and be mirrored in their trades. As existing literature documents, the financial systems are among the best systems to study collective intelligence and researchers are able to infer network properties of financial systems with newly developed tools [3] to understand the underlying connectivity from individual trades. In addition, new financial data with explicit social relationships are also becoming available, which encouraged new areas of research that focused on the social aspects of the financial system.

Research from the Massachusetts Institute of Technology (MIT) demonstrated that higher returns are possible using social trading [4]. Published by the Harvard Business Review, Media Lab researchers found that social explorers seek to “form connections with many different kinds of people and to gain exposure to a broad variety of thinking”. Social traders who found the ‘sweet spot’ – or in other words, the right balance of ideas from a diverse number of traders – were able to increase returns by up to 30%. The research found that the rate of idea flow is a critical measure of how well a social network functions in collecting and refining decision strategies, and this idea flow must come from both within the network and outside of the social network.

1.3 eToro

One of the largest providers of social trading is eToro, an online retail broker for foreign exchanges, index, and commodities trading. Investors can take long and short positions. With its CopyTrader feature, investors can look up others’ trades, portfolios, and past performance. They can trade on either their own or other’s ideas. These trades are classified as original/individual, copy, and mirror trades. There is no cap on leverage, so one can easily lose more than 100% of a position value in a single transaction. The site provides live streaming feeds of fundamental announcements, market news, trading activity of fellow traders, and advice from top performers. eToro’s key tagline is: “Join millions who've already discovered smarter investing by automatically copying the leading traders in our community, or get copied yourself to earn a second income!” The platform currently trades in 170 countries. CNBC named it among the hottest fintech startups to watch in 2015, and 50% of investors on the eToro platform copy other traders’ strategies, while 5% of investors are copied. Fig. 1 provides a screenshot of the eToro platform, and Tbl. 1 has a detailed documentation of the trading rules on eToro.

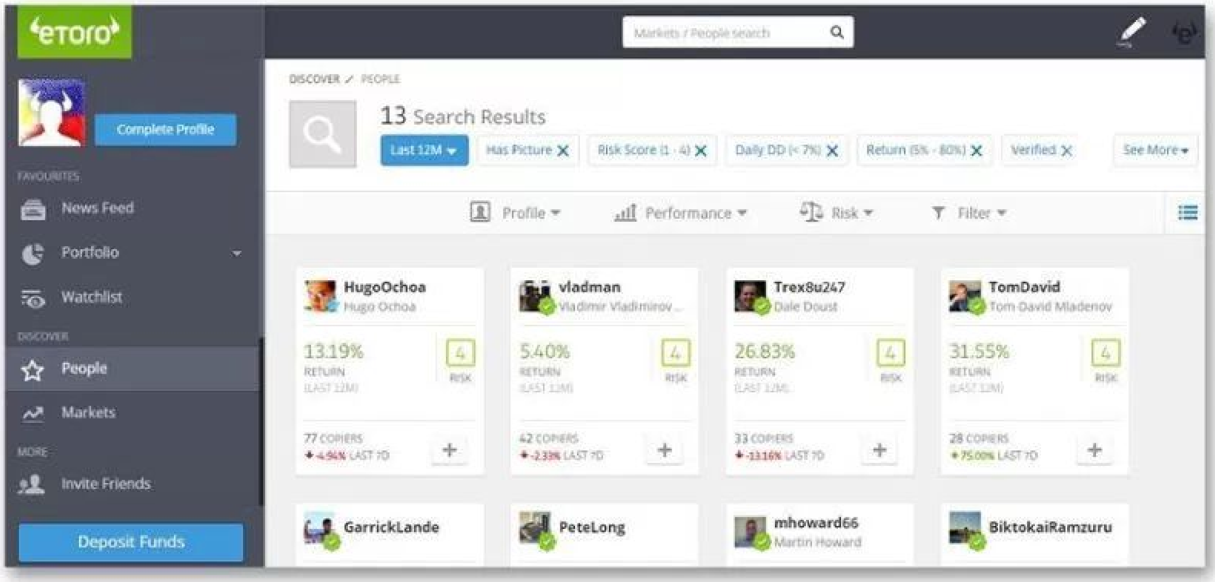


Fig. 1. eToro leader board screenshot

Minimum transaction size after leverage	1,000 Units (USD)
Minimum deposit	USD $50
Leverage	1:2, 1:5, 1:10, 1:25, 1:50, 1:100, 1:200, 1:400
Maintenance margin	No
Accepted Currencies for Deposits	USD/EUR/GBP
Base account currency	USD
Fees	Commission free deposits, trading, withdrawal

Tbl. 1. eToro trading rules

The platform does have impose certain constraints on social trading. One is constrained to 20 traders to mirror at a time. For example, if a portfolio was based on copying 20 traders with equal distribution of funds, capital allocated to each mirrored trade idea would be 5%. Traders have the flexibility to either mirror a trader completely, i.e. execute all open trades, or copy certain trades.

1.4 Defining the Prediction Problem

As mentioned in section 1.1, we are interested in relating social relationships to trading decisions. Essentially we are trying to understand why one trader decides to follow or unfollow another trader. We define any trader that was copied by another trader is a leader and any trader that followed another trader is a follower. In our analysis, a trader is categorized as either a leader or a follower, not both. Essentially we want to answer the question of why one trader follow or unfollow another trader. However, the problem is complicated because there are multiple dimensions. There are two spaces of interest: trading and social space. There are also two perspectives: leaders' perspective and followers' perspective. Therefore, we break down our analysis into three sections, each representing a separate perspective across time:

1. Broad perspective on eToro ecosystem: Who are the leaders? Are there any common profiles? Do "like attract like"?
2. Leader perspective: What characteristics of a leader attract others to follow him? Why are some leaders more popular than others? Can we predict whether a leader will get new follower/unfollower on a particular day given our features?
3. Follower perspective: How does a follower decide who to follow? Are there any interesting behavioral patterns exhibited during the following/unfollowing process? If so, can we predict who is more susceptible to such behavioral patterns?

1.5 Outline of this Paper

- **Chapter 2** This chapter describes how we prepared the data, specifically, how we select our sampling periods and generate the descriptors used in our prediction models.
- **Chapter 3** This chapter discusses the techniques we used to get a general sense of the eToro environment. Specifically, we discuss how we discovered the three distinctive leader profiles and the clustering techniques (K-Means) we employed to arrive at our results. We then link our results to the overall market and provide detailed summary statistics for each cluster.
- **Chapter 4** This chapter analyzes following and unfollowing from leader's perspective. We first correlate the number of new followers and unfollowers a leader gets on a day with his/her trading heuristics. Then we build models for whether a leader will get a follower or unfollower on a particular day given trading heuristics for period 1 and 2 respectively. The chapter goes over the three feature selection techniques used to explore our prediction problem and trains a model for P1 follow, P1 unfollow, P2 follow, and P2 unfollow respectively.
- **Chapter 5** This chapter studies following and unfollowing from follower's perspective. We analyze a particular type of follower -- "networkers" by first give a strict definition of who they are. Then we study whether foraging would improve performance for each type of investors.
- **Chapter 6** This chapter concludes our findings and discusses what can be done next as the next steps in having a better understanding of investor following and unfollowing behavior.

Chapter 2

Data Preparation

In this chapter, we introduce the structure of our data, how we sampled, and how we defined our training and testing sets. We also talk about how we manipulate the data to generate features as input to our prediction problem.

2.1 Data

Our research focuses on over 87.5 million trade and 18.5 million mirroring relationship data collected between August 2011 and December 2013 (social trading features were launched at early 2011 at eToro). In this dataset, there were 17.2 million original trades, 70 million mirror trades. The most traded instruments are generally currencies, with commodities and indices as well (trade volume by currency listed in Tbl. 2-1).

Instrument	Trade Volume (in million)
EUR/USD	23.2
GBP/USD	11.8
AUD/USD	11.8
NZD/USD	11.8
GOLD	4.4
USD/CHF	4.1
EUR/JPY	3.5
USD/CAD	2.9
GBP/JPY	2.7
EUR/CHF	1.2
CHF/JPY	1.1

SILVER	1.1
OIL	0.88
SPX500	0.7

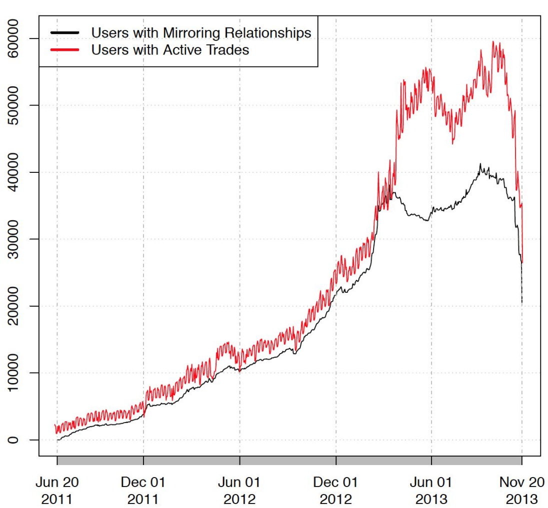
Tbl. 2-1. Number of trades in my dataset categorized by instruments (top 15)

Many may argue that our results may not be representative as eToro is a special type of brokerage platform. However, just as traders these days can converse on Bloomberg and many hedge funds disclose their holdings through 13F and 13G filings, we believe that mirror trading do exist in the real financial market, just in more subtle forms. Information flow, opinions and influence from other peers, and the eventual trading decisions are often largely constrained by the network connections of traders in a manner similar to the eToro user network.

2.2 Sampling

We plot the overall growth of the platform in terms of number of users and trade volume in Fig. 2-1. The platform expanded significantly since 2013.

Number of Traders Trend



Dollar GMV Trend

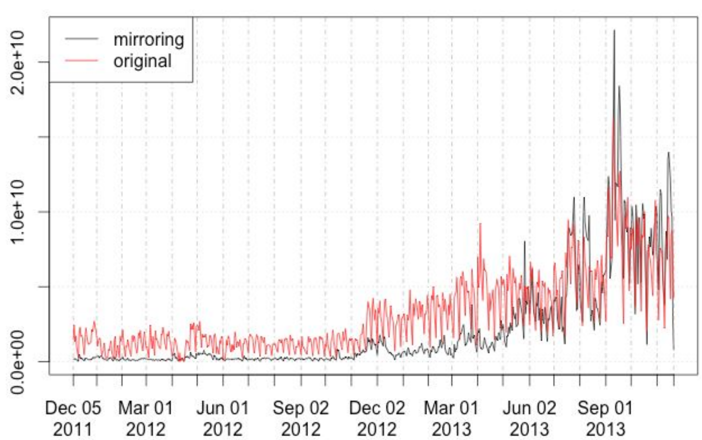


Fig. 2-1. Time-series illustration mirroring activities on eToro since 2011. Mirroring behavior is different between 2011 and 2013 with K-S test $p <$ 1e-15.

Q-Q plot of mirroring relationships before and after 2013 (in Fig. 2-2) shows that distribution of mirroring relationships are not the same on the right tail: hyper-active copy traders' behaviors are different before after 2013. This might be due to the new capacity limit of 20 in 2013.

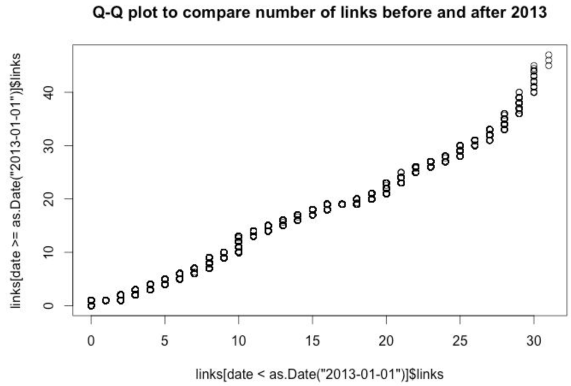


Fig. 2-2. Q-Q plot on mirroring relationships before and after 2013.

Kolmogorov-Smirnov test also suggests that underlying distributions are different.

```
D = 0.1516, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Therefore, we use two sampling periods[1]:

Period 1: start = 2011-06-14, end = 2011-12-13

Period 2: start = 2013-04-01, end = 2013-10-01

2.3 Heuristics and Features Generation

We also define the following heuristics in both the trading space and the social space:

2.3.1 Trading Descriptors

We break down trading descriptors into three categories: style, performance, and risk.

2.3.1.1 Style

We represent individuals *i, j*'s trading behaviors on day *t* through portfolio size, leverage, investment horizon, diversification, and social trading experience. Portfolio size is measured by total capital (MV) and total position size (GMV). Leverage is dollar-weighted average leverage on open positions.[2] Investment horizon is measured by average trade age. Diversification counts the number of instruments in the portfolio. Experience with social trading is measured by i's trading age in days with eToro. Popularity/numberOfFollowees measures how popular a leader is, counting the number of followers that

---

[1] We adopt the following convention to timestamp any trade or mirroring relationship:

(open <= start & close >=start) | (open <= end & close >= end) | (open >= start & close <= end)

[2] eToro has a minimum leverage of 2 and maximum leverage of 400, 100, 100, and 5 for currencies, commodities, indices, and equities respectively.

one has on a particular day. We also created rank (both raw and normalized) metrics based on popularity. The above translate into the following descriptors[3]:

Style Descriptors

$\mathrm{MV}_{i,t}$ = dollar capital in i's portfolio on t
$GMV_{i,t}$ = dollar GMV in i's portfolio on t
$pos_{i,p,t}$ = number of i's active positions on t
$tr_{i,p,t}$ = number of trades i placed on t
$l_{i,t}$ = leverage of i's portfolio on t
$\rho_{i,t}$ = dollar weight of mirroring positions in i's portfolio on t
$\sigma_{i,t}$ = proportion of mirroring positions in i's portfolio on t
$\delta_{i,p}$ = average trade age in i's portfolio during p
$diversification_{i,p}$ = number of instruments traded by i during p
$experience_{i,p}$ = percentage rank of number of days i has been trading on eToro by the end of p
$numberOfFollowers_{i,t}$ = number of traders that follow trader i on t
$rank_{i,p}$ = i's popularity rank among all leaders on t, with higher rank indicating higher popularity
$rank\_norm_{i,p}$ = normalized rank i's popularity on t with 3 being very popular and -3 very unpopular

From the distribution of MirrorRatio in Fig. 2-3, we see that many traders came to the site to either take advantage of low transactional fees (original trade) or completely copy other traders (social trade). Very few traders do a mixture of both.

$\rho_{i,t}$ Distribution in Period 1 and 2

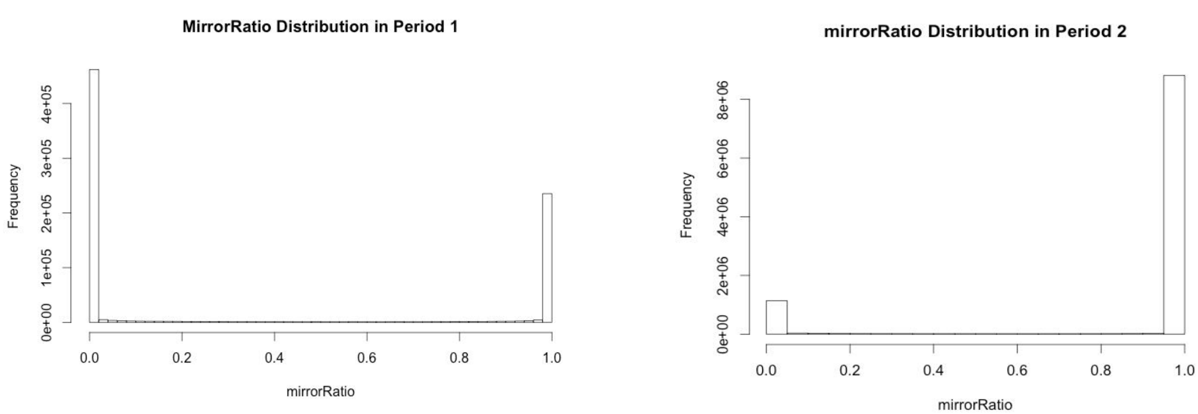


Fig. 2-3. Different mirroring activities between period 1 and 2 with K-S test $p <$ e-15

[3] Cross-sectional descriptors are generated at the end of day with the following datetime conventions: FX rates are snapped at 18:00 GMT; US equity and index prices are snapped at 21:00 GMT; mirror relationships are snapped at 18:00 GMT; trades fall on a given date if opened before 21:00 GMT. Our price data come from Quandl.

2.3.1.2 Performance

Previous literature has established that idea flow contribute to higher performance and that traders unfollow another trader based on the leader's performance metrics: the long-term risk adjusted return [12]. However, we want to supplement this conclusion with more fine-grained performance metrics. Additionally, we are curious whether certain cognitive bias such as salience is present in the following/unfollowing decision making. We hypothesize that a leader's maximum rally and drawdown matter more than his/her short term and long term return in predicting whether the leader gets new followers. Moreover, previous literature suggests that the number of consecutive days that a trader made positive returns help boost his/her popularity, so we included the number of days positive in our performance metrics. Lastly, we want to test whether there is a secondary effect linked to performance, as in does a leader's followers' good performance help the leader attract new followers? Does a leader's followers' bad performance make he/she lose followers? Therefore we have defined additional performance heuristics at different horizons: ROI_day, ROI_week, ROI_month, ROI_net, max_rally, max_drawdown, daysPositive, averagePerfFollowee_day, averagePerfFollowee_week, averagePerfFollowee_month, averagePerfFollowee_net.

Performance Descriptors

$\mathrm{ROI\_day}_{i,t}$ = daily return of i's portfolio on t
$ROI\_week_{i,t}$ = weekly return of i's portfolio as of t
$ROI\_month_{i,t}$ = monthly return of i's portfolio as of t
$ROI\_net_{i,t}$ = overall return of i's portfolio as of t
$daysPositive_{i,t}$ = number of consecutive days before t that trader i has had positive returns
$averagePerfFollowee\_day_{i,t}$ = average daily return of i's followers' portfolios on t
$averagePerfFollowee\_week_{i,t}$ = average weekly return of i's followers' portfolios as of t
$averagePerfFollowee\_month_{i,t}$ = average monthly return of i's followers' portfolios as of t
$averagePerfFollowee\_net_{i,t}$ = average net return of i's followers' portfolios as of t
$max\_rally_{i,t}$ = maximum daily rally of i's portfolio in the past week of day t
$max\_drawdown_{i,t}$ = maximum daily drawdown of i's portfolio in the past week of day t

2.3.1.3 Risk (Volatility)

We adopt the traditional risk measures here using weekly and monthly volatility of one's portfolio to represent how much risk the trader is taking:

Risk Descriptors

$\text{vol\_week}_{i,t}$ = standard deviation of previous week's daily return of i's portfolio as of t
$vol\_month_{i,t}$ = standard deviation of previous month's daily return of i's portfolio as of t

2.3.2 Social Descriptors

One's social activities are represented by his social capacity, frequency, duration, and target retention rate. This allows us to quantitatively model how actively one interacts with his social environment, and we combined them to one single metric, *activity,* where we normalize number of links created and destroyed by capacity. We also measure one's commitment level by duration of his mirroring relationships and number of times one revisits the same target trader. The above translate into the following descriptors:

Social Descriptors

$\text{numberOfMirrorRelationshipOpened}_{i,t}$ = number of traders start mirroring i on t
$numberOfMirrorRelationshipClosed_{i,t}$ = number of traders stop mirroring i on t
$hasNewFollow_{i,t}$ = indicator variable that numberOfMirrorRelationshipOpened is greater than 1 on t
$hasNewUnfollow_{i,t}$ = indicator variable that numberOfMirrorRelationshipClosed is greater than 1 on t
$\mu_{i,t}$ = total number of traders mirrored by i on t
$a_{i,t}$ = number of new traders mirrored by i on t compared to t-1
$d_{i,t}$ = number of traders un-mirrored by i on t compared to t-1

$$activity_{i,t} = \begin{cases} 0 & \text{if } \mu_{i,t} = 0 \\ (a_{i,t} + d_{i,t})/(2 * \mu_{i,t}) & \text{if } \mu_{i,t} \neq 0 \end{cases}$$

$\tau_{i,p}$ = average duration in days of i's mirror relationships during p
$f_{i,p}$ = average number of times i mirrors the same trader during p

2.3.3 Putting Everything Together

Having created trading and social heuristics, we can now proceed to more sophisticated statistical and machine learning modeling. Fig. 2-4 following shows a screenshot of the complete data frame which we will use for modeling in Chapter 4.

```
> head(social_modeling_df_rf)
         date     CID mui ai di activity        tau        f follower numberOfFollowers rank  rank_norm
1: 2011-06-27  180610   1  0  1      0.5   1.411029 1.071429        0                 3   45  0.9909212
2: 2011-06-27  632773   1  0  0      0.0  19.535903 1.000000        0                 1    1 -0.8417792
3: 2011-06-27 1411530   2  0  0      0.0  31.560671 1.333333        0                13   55  1.4074440
4: 2011-06-27 1421189   1  1  0      0.5 118.377439 1.300000        0                 1    1 -0.8417792
5: 2011-06-27 1500898   1  0  0      0.0   4.154969 1.500000        0                 3   45  0.9909212
6: 2011-06-28  436768   1  0  0      0.0   4.097093 1.200000        0                 1    1 -0.8848947
   averagePerfFollowee_net averagePerfFollowee_day averagePerfFollowee_week averagePerfFollowee_month
1:            1.555050e-03            0.000000e+00             0.000000e+00                         0
2:            6.499902e-04            0.000000e+00             0.000000e+00                         0
3:           -1.162244e+01            8.663292e-04             8.290461e-07                         0
4:           -2.534737e+01            5.349728e-05             0.000000e+00                         0
5:           -9.698376e-04            0.000000e+00             0.000000e+00                         0
6:           -1.899388e+02            0.000000e+00             0.000000e+00                         0
        ROI_net       ROI_day      ROI_week ROI_month daysPositive numberOfMirrorRelationshipOpened
1:   0.002210396  0.000000e+00  0.000000e+00         0            0                                3
2:   0.006768360  0.000000e+00  0.000000e+00         0            0                                1
3:  -0.001459515  1.826953e-03  2.980383e-05         0            1                               10
4: -11.643538652  3.357433e-03 -3.403172e-03         0            1                                1
5:  -0.001507777 -1.321658e-05  0.000000e+00         0            0                                3
6: -14.472072395 -1.670716e-03  0.000000e+00         0            0                                0
   numberOfMirrorRelationshipClosed     vol_week vol_month   max_rally max_drawdown          MV         GMV
1:                                0 0.0000000000         0 0.006828668 -0.013567349   92343.804  9270805.03
2:                                0 0.0041794310         0 0.009002800 -0.016264867    7174.495   722339.96
3:                                0 0.0009412722         0 0.018400381 -0.024293374  566924.790 17260232.75
4:                                0 0.0031122215         0 0.026236908 -0.021932546 3545426.097 14125621.14
5:                                0 0.0024546840         0 0.005251776 -0.006288197   12144.613  1186887.69
6:                                0 0.0125533132         0 0.035545682 -0.013567349    9683.250    16445.32
   totalPos   leverage meanTradeAge       exp diversification hasNewFollow hasNewUnfollow cluster
1:       53 100.394446    0.5537109 0.9056067              16            1              0       3
2:        6 100.681647    0.3278588 0.9006644               9            0              0       1
3:       17  30.445366    3.4355297 0.9909338              14            1              0       2
4:        9   3.984182    1.9640792 0.9930940              10            0              0       2
5:        3  97.729564    1.7466898 0.9342454               8            1              0       1
6:       10   1.698326    0.7339884 0.9818021              16            0              0       3
```

Fig. 2-4. Complete data frame from period 1 for modeling following and unfollowing

Chapter 3

eToro Ecosystem

In this chapter, we provide a high-level overview of the eToro ecosystem. Our goal is to understand who are the “leaders” in this ecosystem and what are the common leader profiles? Do these profiles evolve over time (aka different in sampling period 1 vs sampling period 2)? Do “like attract like” -- Do similar profile follow similar profile?

3.1 Methods for Classification

Because we do not currently know which profiles are present on the platform, classification is best achieved with unsupervised learning. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis to reveal hidden structure in data. The specific clustering technique that we used is k-means [9], which is an iterative partitioning algorithm that groups data into k coherent, user-defined clusters. The algorithm first chooses k (random) data points (seeds) to be the initial centroids, aka cluster centers. Then it assigns each data point to the closest centroid and re-computes the centroids using the current cluster memberships. And lastly, if a convergence criterion is not met, the algorithm repeats the data assignment and centroid selection process. However, in order to have an accurate depiction of the underlying structure of our data, we need to know 1) how many clusters there are and 2) what are the defining features and centroids for each cluster.

To address the first question, we use the elbow method to select the optimal number for k in each period. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

To address the second question, we run k-means clustering based on the optimal number of k we selected in the elbowing method. We compare the clusters between P1 and P2 to see if there is any universal profile that exist.

3.2 Features and Optimal Number of Clusters

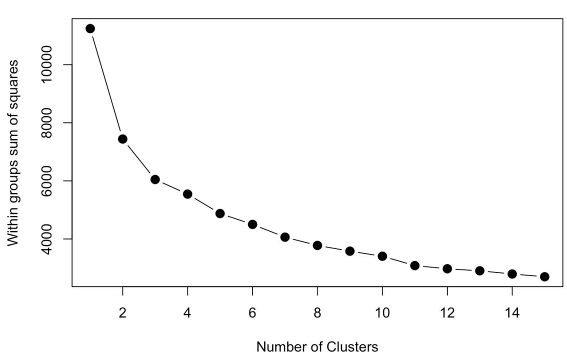
For the purpose of our study, we classify eToro leaders, which are the traders that were followed in that period, based on their trading styles, which include their investing objectives and constraints. Objectives being the type of return being sought and , while constraints encompasses investment horizon, liquidity and leverage etc. We use the following features (winsorized at 3std and standardized) from the leaders: MV, GMV, totalPos, leverage, meanTradeAge, exp, diversification as shown in Fig. 3-1.

```
##                MV        GMV   totalPos   leverage meanTradeAge      exp diversification
## [1,] -0.07328303 -0.4125360  0.7805935 -1.1989689    1.0192718 1.387937       1.7625714
## [2,]  8.02510235  0.8297714  3.7441940 -1.4341035    0.7282406 1.563852       2.8844953
## [3,] -0.28445615 -0.5971477 -0.5059996 -0.5772353    2.4260974 1.334713       0.1918778
## [4,] -0.26586094 -0.5482935  0.3588276 -1.3770788    1.6158905 1.307267       0.6406474
## [5,]  0.43586671  0.2314522  2.7652002 -0.1418683    1.9108525 1.544574       1.0894170
## [6,]  4.63936799  1.2776138  1.8950492 -1.2295267    1.1737685 1.539685       0.1918778
```

Fig. 3-1. Trading features used for clustering analysis

Using the features, we arrived at the following elbow plots for period 1 and period 2 respectively. Based on Fig. 3-2, k = 3, 3 clusters/common profile best characterizes our data.

Period 1 K-Means Clustering Elbow Plot



Period 2 K-Means Clustering Elbow Plot

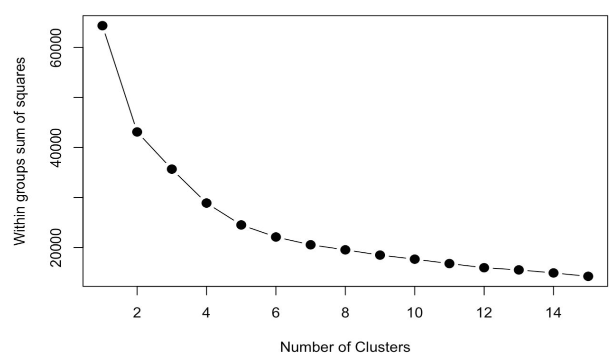


Fig. 3-2. K-means Clustering elbow plot for period 1 and period 2

3.3 Period 1 K-Means Clustering Results

Centroids For Three Crowd Leader Profiles in Period 1

Cluster	(1)	(2)	(3)
MV	-0.36	2.86	0.65
GMV	-0.14	4.61	0.44
totalPos	-0.09	0.15	0.88
leverage	0.37	-0.77	-0.72
meanTradeAge	-0.25	1.55	0.34
exp	0.01	1.46	0.61
diversification	-0.27	0.33	1.14
Cluster Size	505	133	469

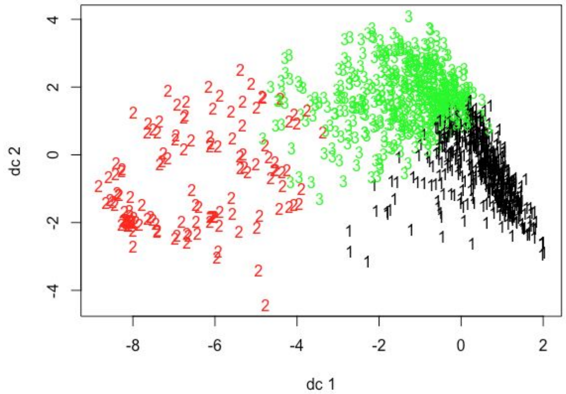


Fig. 3-3. Results from k-means clustering in period 1. Centroid plot against 1st and 2nd discriminant functions indicates clear cluster separation with BSS/TSS of 52%.

3.4 Period 2 K-Means Clustering Results

Centroids For Three Crowd Leader Profiles in Period 2

Cluster	(1)	(2)	(3)
MV	-0.22	1.92	-0.17
GMV	0.31	0.83	-0.41
totalPos	-0.03	2.68	0.15
leverage	0.52	-1.00	-0.62
meanTradeAge	-0.56	0.05	0.68
exp	-0.09	0.51	1.12
diversification	0.07	2.03	0.38
Cluster Size	2976	1656	3115

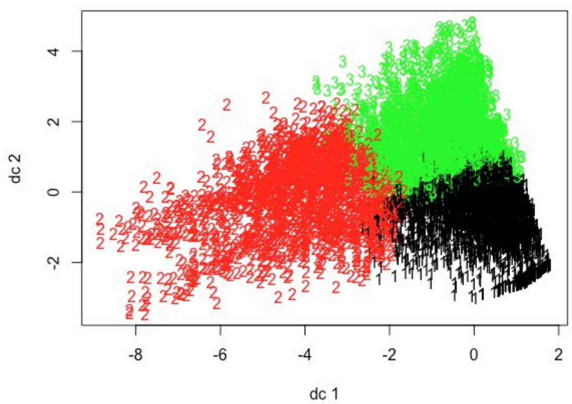


Fig. 3-4. Results from k-means clustering in period 2. Centroid plot against 1st and 2nd discriminant functions indicates clear cluster separation with BSS/TSS of 51%.

3.5 Common Profiles and Connection to the Market

Based on the centroids and distributions of these three clusters, we are confident that the three clusters or profiles are representative of common leaders on the platform across time. Interestingly, these profiles, as represented by their centroids, resemble common investor profiles. The eToro ecosystem is almost a miniature of the broad financial market.

3.5.1 Profile 1 -- High-Velocity, Risky Speculators

These speculators can be day-traders or short-sellers. They tend to hold small to medium sized portfolios and take on highly leveraged, risky positions in the expectation for significant, short-term gains. Their entry and exit in a stock are quite fast, therefore have low price impact. It is surprising that speculators constitute a major mirrored profile, as very little “wisdom” is involved in the investment thesis. However, considering the peculiar context of our experiment, the FX market, it is sensible that such profile has emerged. Liquidity and real-time price dissemination characteristic of the currency and public securities markets make speculative activities favorable.

3.5.2 Profile 2 -- Institutional Portfolio Managers

These investors hold many big positions in a variety of instruments. Trade ideas usually come from well-educated, seasoned investment professionals and typically take on low leverage. Because these are actively managed and well balanced portfolios, average age of active positions tends to be low. Mutual funds and long-term institutional investors would fall into this category. A particularly good example would be providers of smart beta strategies, systematically selecting, weighting, and rebalancing holdings on the basis of factors other than market capitalization.

3.5.3 Profile 3 -- Low-Velocity, Strategy Junkies

These investors hold few positions, typically mid-sized, in highly specialized areas, such as in one single segment or instrument. Similar to Profile 1, these investors are well trained and knowledgeable in their fields. They can take on decent amount of risk. Commodity traders, such as corn and gold traders are good examples that fit in this profile. Traders in this camp have specialized sector knowledge. For example, successful crude traders understand supply and demand dynamics and can model contagion using proprietary data. They take on long/short positions based on views on the market. Considering commodities is the second largest market that eToro trades in, it is not surprising that this profile emerged.

3.6 Does "Like Attract Like"?

In order to answer answer whether traders follow traders of similar trading styles, we need to classify followers based on the three profiles we defined in 3.1.5. We fit followers' trading features using clusters trained by leaders' trading features to get a classification of the followers. For each leader we then looked at the composition of their followers in Fig. 3-5. The main takeaway from our membership analysis is that the spetculator type makes up most of the follower population. In period 1, most followers followed the strategy type but in period 2, following is split mostly between institutional and strategy leaders. Additionally, in period 1 6876 out of 15142 mirroring relationships (45.4%) is between same-type traders, but in period 2, 152983 out of 253314 mirroring relationships (39.6%) is between same-type traders. It does seem like followers are seeking more diversity in period 2.

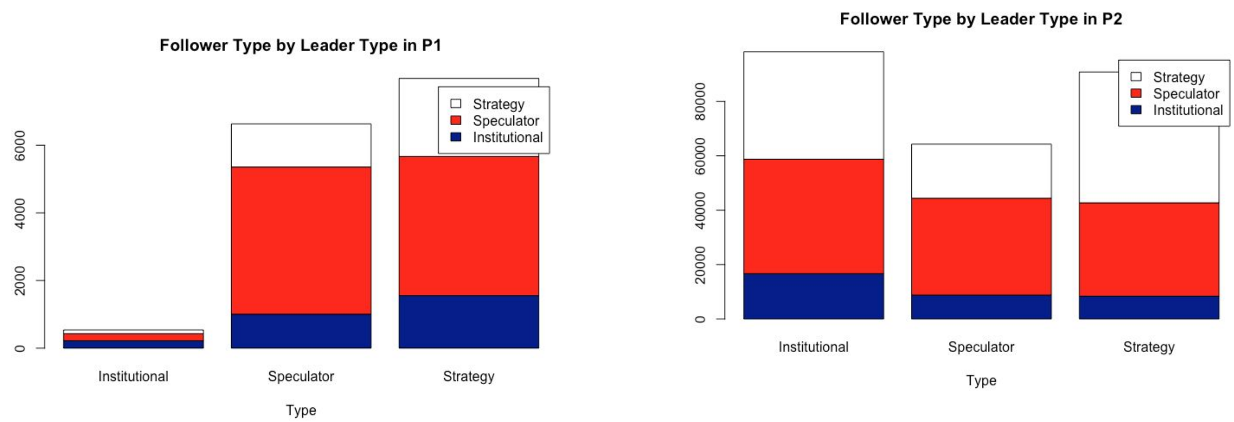


Fig. 3-5. Membership analysis of followers of different leader types

From an administrative point of view, psychologically-based profiling and classification systems help us understand preferences of the eToro community, which helps the site better dedicate its planning resources. From an individual investor point of view, knowing the category of crowd leader's personality will help choose the right kind of investments. If one wants to specialize in the biotechnology sector, it is essential to decide whether one wants to bet on long-term prospects of the technology or event-driven price dislocation. If the former, most of the value that a biotech company creates will not be realised for more than a decade, then one should mirror investors with long horizons and mature value-investing strategies; leaders who trade on short-term momentum would not be a great fit. In the case of latter, one should follow news forecasters and event-bettors, who either has channels for superior information on company-specific events such as clinical trials or have a view on the regulatory landscape.

This is analogous to picking asset managers in wealth management settings, where managers tailor their investment advisory services to clients needs and clients uses various screening tools to monitor performance. The most widely-adopted method is the "Barnwell Two-Way Model" where managers are "active" and "passive" [1]. More sophisticated models such as the "Bailard, Biehl and Kaiser Five-Way model" (BB&K) and the "Nine Money Personalities" by Kathleen Gurney also emerged where more subjective assessments of investing behaviors are performed to infer motivation and emotions. For example, in the BB&K model, level of confidence is reflected in emotional attachment associated with trade and asset allocation decisions. Investors may range from *confident* to *anxious*. Method of action is reflected in how methodical one is. This can range from *careful* to *impetuous* [11]. Additionally, there has been empirical research conducted to support such theories. There was a study performed in the Indian capital market, where researchers collected empirical evidence, questionnaires from Chennai city, to validate investment behaviors predicted by the BB&K personality types. For example, the study validated the hypothesis that there is significant relationship between an investor who is categorized as the Adventurer and invests in direct equity, Equity oriented mutual funds, pension schemes, Hedge Funds, PE Funds, VC funds [7].

The above three profiles are a few examples of how top performers on eToro can be characterised. We suspect other types of profile, such as index watchers and innovation punters, exist and are also popular profiles if we had more recent empirical data.

We would also like to highlight the importance of following the right leader. During the technology boom, many funds decided to invest in the technology sector and with healthcare becoming a hot sector, many trend-followed and invested significant capital in names favored by the masses, yet these strategy worked for some but not for all. Following other investors makes a lot of sense if the mirroring investment decision is coherent with one's investment thesis or offers diversity to the existing portfolio. This is why many of these trend-following investments failed, with the most well-known incident being feverish contrarian purchasing behavior of Valeant following Bill Ackman after the stock plummeted 86% despite the company missed earnings expectations. These followers got a bitter lesson when Valeant shares came crashing back down more than 10%, erasing much of their earlier gains. This is an example of how detrimental blindly mirroring can be.

3.7 Summary Statistics of Each Profile

Taking a longitudinal perspective, 2944 traders traded on eToro in both periods, with 297 (10%) of them stayed as leaders in both periods. Based on the summary statistics in Tbl. 3-1, we see that institutional and strategy types are more likely going to stay on the platform. Moreover, most of the traders that stayed on the platform converted to the strategy type, with strategy type having the highest conversion rate across all three types (65%, 60%, and 58% shown in Tbl 3-2). This suggests a convergence towards sector specialization, which agrees with what we witness in the broader financial market where portfolio managers cover a particular sector[4]. Lastly, we saw that speculators and strategy traders are the most social traders. This also confirms with our intuition.

Count of Each Type in Period 1 and 2

Period	Period 1 Count	Period 2 Count	In Both Periods
Speculator	**6611 (67%)**	49780 (58%)	1643 (25%)
Strategy	**2956 (30%)**	28972 (34%)	1052 (36%)
Institutional	**307 (9%)**	7559 (9%)	124 (40%)

Tbl. 3-1. Contingency table of types of traders on eToro in period 1 and period 2

Type Conversion Between Period 1 and 2

P1 Type	P2 Type	Number of Traders	Conversion Rate
Strategy	Strategy	680	65%
Speculator	Strategy	994	60%
Institutional	Strategy	72	58%
Speculator	Speculator	493	30%
Institutional	Speculator	26	21%
Institutional	Institutional	26	21%
Strategy	Speculator	220	21%
Strategy	Institutional	152	14%
Speculator	Institutional	156	9%

Tbl. 3-2. Type conversion from P1 to P2

In order to test randomness, we shuffle the types for all period 2 traders to create a control set. Our response variable is an indicator variable of whether the trader switched type (1=switched, 0=not switched). Tbl 3-3 shows contingency table if P2 trader types were randomly assigned, and we see the conversion rates are drastically different. We conduct K-S test to see if the two groups follow the same distribution and results are shown in Fig. 3-9.

[4] Conversion rate is calculated as (# of traders who switched from Type A to Type B in period 2)/(# of traders of Type A who stayed on eToro in period 2). For example, 65% conversion rate for Strategy-Strategy switch is calculated as 680/1052 = 65%.

P1 Type	P2 Type	Number of Traders	Conversion Rate
Strategy	Strategy	351	33%
Speculator	Strategy	553	34%
Institutional	Strategy	45	36%
Speculator	Speculator	953	58%
Institutional	Speculator	67	54%
Institutional	Institutional	12	10%
Strategy	Speculator	623	59%
Strategy	Institutional	78	7%
Speculator	Institutional	137	8%

Tbl. 3-3. Type switch of traders who traded on eToro in both periods if P2 trader types were randomly assigned

Then we test whether P1-P2 type switch is independent from P1-random type switch, using indicator variable of 1 when there is a type switch. Fig. 3-6 shows results from K-S test on each of the three types in period 1. As our p-values indicate, type switch for P1 speculators and strategists are statistically significant.

```
> speculator_KS

        Two-sample Kolmogorov-Smirnov test

data:  as.numeric(merged_all_shuffle[cluster.x == 1, type.x == type.y]) and as.numeric(merged_all_no_shuffle[cluster
.x == 1, type.x == type.y])
D = 0.28, p-value < 2.2e-16
alternative hypothesis: two-sided

> institutional_KS

        Two-sample Kolmogorov-Smirnov test

data:  as.numeric(merged_all_shuffle[cluster.x == 2, type.x == type.y]) and as.numeric(merged_all_no_shuffle[cluster
.x == 2, type.x == type.y])
D = 0.1129, p-value = 0.4081
alternative hypothesis: two-sided

> strategy_KS

        Two-sample Kolmogorov-Smirnov test

data:  as.numeric(merged_all_shuffle[cluster.x == 3, type.x == type.y]) and as.numeric(merged_all_no_shuffle[cluster
.x == 3, type.x == type.y])
D = 0.3127, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Fig. 3-6. K-S test on P1-to-P2 conversion vs P1-to-random conversion.

Lastly, we look at the average social activity level of each group in Tbl. 3-4. T tests on activity levels for each type between period 1 and 2 shows that the change in activity level is significant. Overall, traders are less socially active in creating and destroying relationships in period 2. It is expected that speculators are the most socially active traders followed by strategy traders and then institutional traders.

Average Activity of Each Type in Period 1 and 2

Period	Period 1 Average Activity	Period 2 Average Activity	ANOVA p val
Speculator	0.0569	0.0185	<2e-16
Strategy	0.0526	0.0155	<2e-16
Institutional	0.0219	0.0058	<2e-16
ANOVA p val	<2e-16	<2e-16	

Tbl. 3-4. Social activity level of each type of trader

Chapter 4

Leader Perspective -- Predicting Following and Unfollowing

In this chapter, we study why some leaders get more followers than others using cross-sectional (daily frequency) trading and social features of the leaders. Then we train separate models to predict whether a leader will have follower or unfollower on a particular day during each period following the traditional model building methodology: feature selection, train using training set, and evaluate using testing set. Because our dataset is highly unbalanced, in that we have more days where leaders do not receive new followers/unfollowers than days where they do, we use SMOTE (Synthetic Minority Oversampling Technique), which under-sample the majority class and oversample the minority class to create balanced training set to train our models [11]. We are able to achieve an accuracy of 84% (AUC 93%) for following and 90% for unfollowing (AUC 97%) in period 1 and 83% for following (AUC 94%) and 85% for unfollowing (AUC 95%) in period 2.

4.1 Correlation Analysis on Number of Links Opened and Closed

As there are many factors and noise that would be causing one leader to have more new followers than another leader on a particular day, it is difficult to draw causal inference. So we inspect the factors that relate to `numberOfMirrorRelationshipOpened` and `numberOfMirrorRelationshipClosed` instead of trying to predict these two variables. Correlation matrix for the two factors in period 1 and period 2 are included in Tbl. 4 bellow.

As `numberOfFollowers`, `rank`, and `rank_normalized` are highly correlated to each other, it makes sense that both `numberOfRelationshipOpened` and `numberOfRelationshipClosed` are correlated to these three variables. However, it is surprising that such relationship is not directional, which means that followers' opinion is not uniform: average followers have varying opinions on popular, highly ranked traders. `numberOfRelationshipOpened` and `numberOfRelationshipClosed` are highly correlated, which suggest that there might not be a need to distinguish the act of opening and closing of a relationship. What we are capturing might simply be activity. More volatile and leaders with significant rally attract more daily followers. Too many open positions in period 1 or large portfolio in period 2 are not favorable to followers. More experience helps a trader attract more daily followers.

numberOfMirrorRelationshipOpened	Period 1		Period 2	
Feature	Corr	P-val	Corr	P-val
numberOfFollowers	0.5384	0.0000	0.6278	0.0000
rank	0.3487	0.0000	0.1691	0.0000
rank_norm	0.3543	0.0000	0.1707	0.0000
averagePerfFollowee_net	0.0002	0.9934	-0.0006	0.9400
averagePerfFollowee_day	0.0103	0.6898	-0.0137	0.1130
averagePerfFollowee_week	0.0186	0.4696	-0.0076	0.3805
averagePerfFollowee_month	-0.0336	0.1912	0.0070	0.4135
ROI_net	-0.0209	0.3989	-0.0020	0.8134
ROI_day	0.0207	0.4047	-0.0141	0.0962
ROI_week	0.0392	0.1139	-0.0041	0.6344
ROI_month	0.0370	0.1353	0.0024	0.7812
daysPositive	0.0528	0.0330	0.0117	0.1444
numberOfMirrorRelationshipClosed	0.3371	0.0000	0.4949	0.0000
vol_week	0.0544	0.0311	-0.0184	0.0375
vol_month	0.1155	0.0000	-0.0003	0.9717
max_rally	0.2691	0.0000	0.0195	0.0641
max_drawdown	-0.0742	0.0101	-0.0411	0.0091
MV	-0.0494	0.0464	-0.0267	0.0008
GMV	-0.0737	0.0029	-0.0094	0.2462
totalPos	-0.1083	0.0000	-0.0190	0.0170
leverage	0.0457	0.0651	0.0110	0.1941
meanTradeAge	0.0751	0.0024	0.0002	0.9798
exp	0.1032	0.0000	0.0606	0.0000
diversification	-0.1462	0.0000	0.0355	0.0000
cluster	-0.1356	0.0000	-0.0050	0.5321

Tbl. 4-1. Correlation matrix with correlation and p-val for numberOfMirrorRelationshipOpened. Features with p-val < 0.0001 are highlighted

`numberOfFollowers`, `rank`, and `rank_normalized` are all correlated with `numberOfRelationshipClosed.` It agrees with our intuition that the more levered a leader becomes, the more followers he/she loses.

As aforementioned, activity from leader's perspective does not seem to be directional. What we are capturing seem to be social activity in general.

numberOfMirrorRelationshipClosed	Period 1		Period 2	
Feature	Corr	P-val	Corr	P-val
numberOfFollowers	0.6350	0.0000	0.5358	0.0000
rank	0.3313	0.0000	0.1456	0.0000
rank_norm	0.2722	0.0000	0.1451	0.0000
averagePerfFollowee_net	0.0103	0.6640	0.0010	0.9074
averagePerfFollowee_day	-0.0024	0.9198	0.0069	0.4015
averagePerfFollowee_week	0.0013	0.9573	-0.0073	0.3773
averagePerfFollowee_month	0.0019	0.9348	-0.0021	0.7998
ROI_net	0.0109	0.6236	0.0023	0.7750
ROI_day	0.0096	0.6641	-0.0078	0.3314
ROI_week	-0.0032	0.8862	-0.0052	0.5175
ROI_month	-0.0136	0.5389	-0.0115	0.1534
daysPositive	0.0275	0.2159	0.0147	0.0519
numberOfMirrorRelationshipOpened	0.3618	0.0000	0.4508	0.0000
vol_week	-0.0577	0.0108	0.0021	0.7998
vol_month	-0.0069	0.7660	0.0002	0.9852
max_rally	0.2777	0.0070	0.0098	0.3149
max_drawdown	0.0033	0.8991	-0.0229	0.0190
MV	-0.0515	0.0202	-0.0198	0.0089
GMV	-0.0673	0.0024	-0.0108	0.1580
totalPos	-0.0926	0.0000	-0.0061	0.4202
leverage	0.0703	0.0015	0.0404	0.0000
meanTradeAge	0.0234	0.2922	-0.0223	0.0032
exp	0.0481	0.0301	0.0259	0.0006
diversification	-0.1573	0.0000	0.0616	0.0000
cluster	-0.1502	0.0000	-0.0174	0.0213

Tbl. 4-2. Correlation matrix with correlation and p-val for numberOfMirrorRelationshipClosed. Features with p-val < 0.0001 are highlighted

4.2 Feature Selection

Because our data has lot of highly correlated predictors, the precision of the estimated regression coefficients can be compromised due to extraneous predictors. So we need to reduce multicollinearity in the predictors. We present three methods for feature selection. The first one is simple correlation analysis of the predictors. The second is random forest using Gini index. The third is Logit with lasso regularization.

4.2.1 Selection via Correlation Analysis

In order to predict following and unfollowing, we begin with studying the underlying correlation structure of our predictors. We have shown the features provided to us in the data set in Chapter 2. However, many of the features contribute little in predicting whether a leader will get a new follower/unfollower or not. As such, it is important to select the features that are relevant to feed into our model and to not overfit.

A few things stand out from the correlation analysis. As Fig. 4-1 shows, leader's performance is highly correlated with followers' performance, which validated that these social traders were indeed copying trades of the leaders. Moreover, number of followers is highly correlated with popularity rank. Maximum rally and drawdown are highly correlated with portfolio volatility. We also confirmed our hypothesis on salience. Number of followers is highly positively correlated with maximum rally and negatively correlated with maximum drawdown but not with ROIs -- subjects pay more attention and more likely to react to things that are obvious, and in our case, maximum rally and drawdowns.

Correlation Matrix for Period 1 and 2

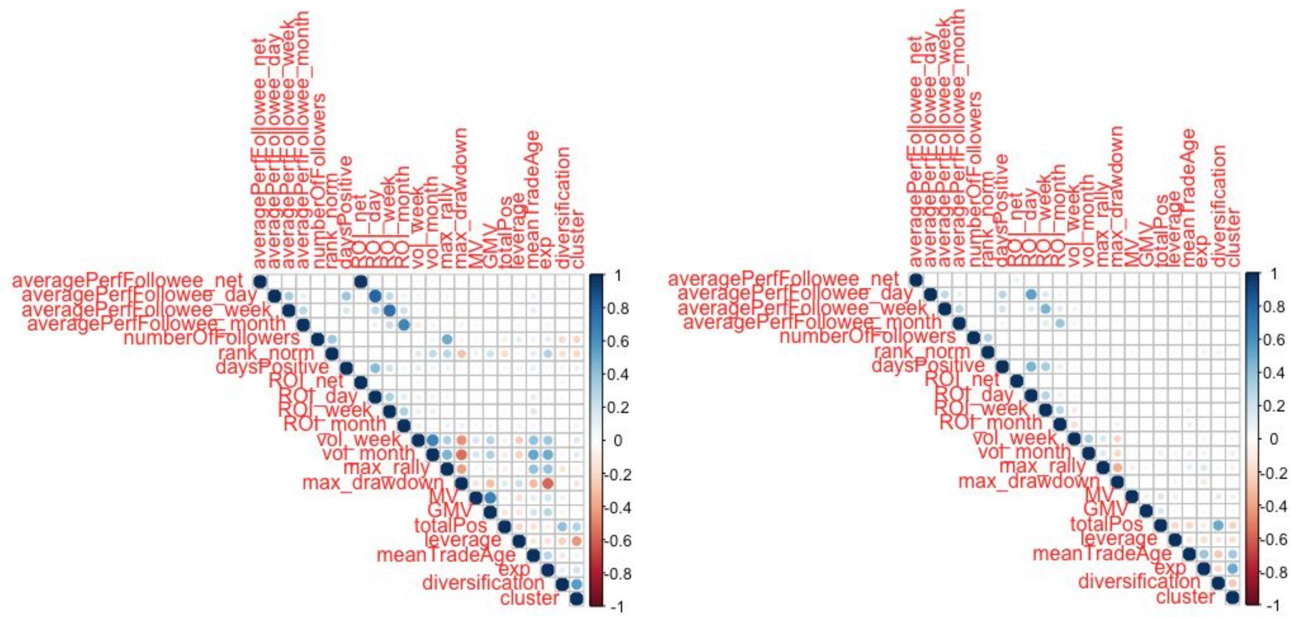


Fig. 4-1. Spearman correlation matrix of predictors

4.2.2 Selection via Random Forest

Random forest is an ensemble learning method for classification that trains a model by creating a multitude of decision trees and outputting the mode of the classes returned by the individual trees. The algorithm applies the technique of bootstrapping aggregation, also known as bagging, to tree learners. Given a training set and responses, the algorithm repeatedly selects a sample at random with replacement from the training set and fits trees to the samples. The algorithm is good for feature selection because tree-based strategies ranks each predictor by how well they improve the purity of the node, which is essentially accuracy of the classification tree.

Random forest follows the general scheme of bagging but differs by using a modified tree-learning algorithm that selects, at each iteration of the learning process, a random subset of features. We use random forest to do feature selection because of its ability to rank the importance of the variables. We use GINI importance as a metric because it measures the average gain of purity by splits of a given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes. Permuting a useful variable, tend to give relatively large decrease in mean gini-gain. In the following subsections, we run random forest to find the most important features for each period and for following and unfollowing separately. We rank the variables based on their GINI importance and plotted importance values.

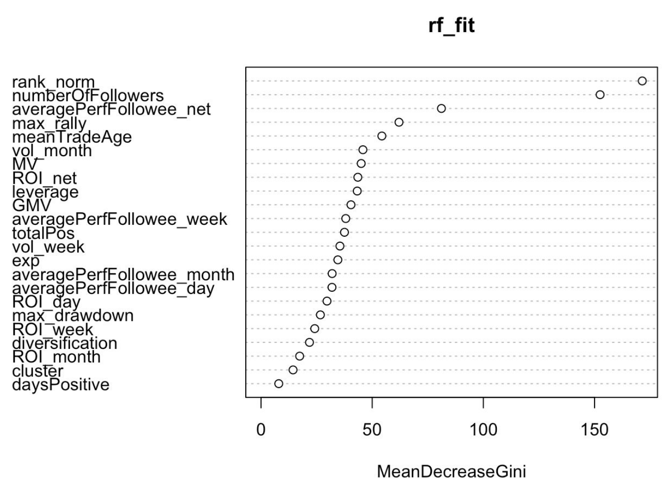
4.2.3 Selection via Logistic Regression

We also use generalized linear model, logistic regression, combined with cross validation for feature selection. Lasso is a shrinkage regression method that constrains the sum of the absolute values of the regression estimates. This is convenient when dealing with highly correlated predictors, where standard regression will usually have regression coefficients that are too large. Through cross validation, the algorithm computes mean squared error (MSE) for different penalization parameter, lambda. It then returns the lambda with the smallest MSE and returns the variables that have regression coefficients not shrunken to 0 as meaningful predictors.

4.3 Selected Variables

4.3.1 Period 1 Following

Predictor	MeanDecreaseAccuracy	MeanDecreaseGini
rank_norm	36.26	171.46
numberOfFollowers	34.37	152.51
averagePerfFollowee_net	42.99	81.14
max_rally	27.95	62.05
meanTradeAge	30.63	54.35
vol_month	31.25	45.83
MV	20.62	45.06
ROI_net	25.00	43.55
leverage	25.65	43.27
GMV	23.81	40.46
averagePerfFollowee_week	23.91	38.11
totalPos	28.24	37.53
vol_week	20.02	35.49
exp	24.70	34.56
averagePerfFollowee_month	23.56	31.99
averagePerfFollowee_day	15.88	31.91
ROI_day	19.80	29.67
max_drawdown	21.66	26.70
ROI_week	16.61	24.15
diversification	17.92	21.80
ROI_month	17.46	17.38
cluster	15.88	14.41
daysPositive	8.12	7.98



Tbl. 4-3. Importance values returned (left) and graph of importance values (right) for Period 1 following

Variables selected by lasso with cross validation:

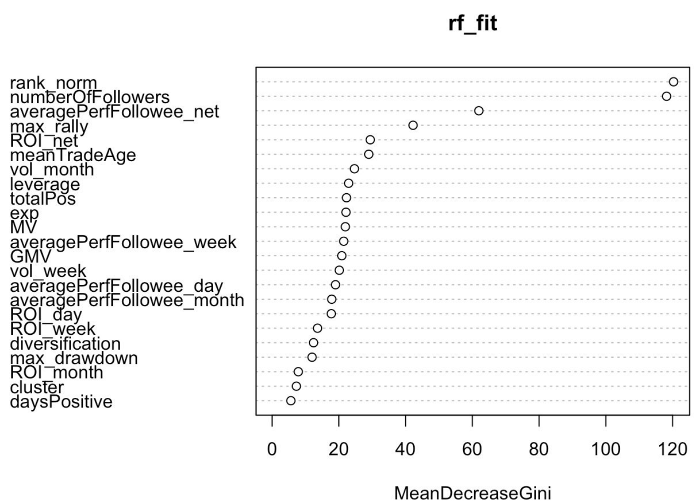
```
## [1] "numberOfFollowers" "rank_norm"         "ROI_month"
## [4] "vol_week"          "max_rally"         "max_drawdown"
## [7] "GMV"               "meanTradeAge"      "cluster"
```

Based on all three selection metrics, we use the following variables for P1 following modeling:

```
Rank_norm
numberOfFollowers
averagePerfFollowee_net
ROI_net
Vol_month
Max_rally
Leverage
meanTradeAge
MV
GMV
```

4.3.2 Period 1 Unfollowing

Predictor	MeanDecreaseAccuracy	MeanDecreaseGini
rank_norm	41.36	120.30
numberOfFollowers	36.10	118.20
averagePerfFollowee_net	41.74	61.95
max_rally	20.81	42.24
ROI_net	20.86	29.41
meanTradeAge	23.90	28.97
vol_month	22.72	24.67
leverage	16.04	22.94
totalPos	22.31	22.29
exp	18.11	22.14
MV	12.07	21.95
averagePerfFollowee_week	18.31	21.46
GMV	15.30	20.89
vol_week	16.76	20.13
averagePerfFollowee_day	11.53	19.00
averagePerfFollowee_month	13.30	17.87
ROI_day	11.99	17.74
ROI_week	11.06	13.62
diversification	12.79	12.42
max_drawdown	17.31	11.98
ROI_month	6.98	7.85
cluster	9.34	7.28
daysPositive	9.31	5.63



Tbl. 4-4. Importance values returned (left) and graph of importance values (right) for Period 1 unfollowing

Variables selected by lasso with cross validation:

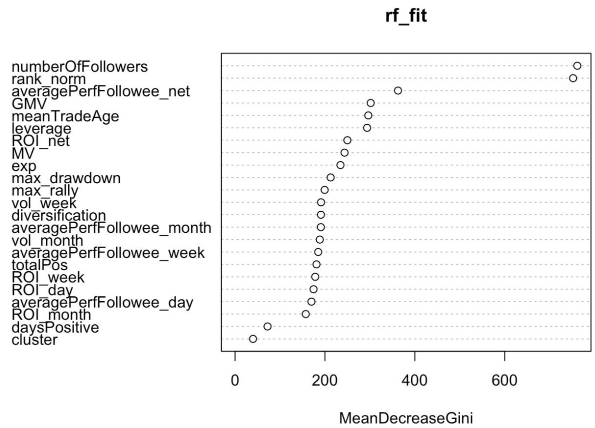
```
##  [1] "numberOfFollowers" "rank_norm"         "daysPositive"
##  [4] "ROI_month"         "vol_week"          "vol_month"
##  [7] "max_rally"         "max_drawdown"      "MV"
## [10] "GMV"               "meanTradeAge"      "exp"
## [13] "cluster"
```

Based on all three selection metrics, we use the following variables for P1 unfollowing modeling:

```
Rank_norm
numberOfFollowers
averagePerfFollowee_net
ROI_net
GMV
Leverage
cluster
```

4.3.3 Period 2 Following

Predictor	MeanDecreaseAccuracy	MeanDecreaseGini
numberOfFollowers	61.79	761.46
rank_norm	69.88	752.14
averagePerfFollowee_net	62.49	362.57
GMV	30.87	301.78
meanTradeAge	51.13	296.49
leverage	42.36	293.75
ROI_net	55.82	249.94
MV	29.91	243.62
exp	48.60	234.58
max_drawdown	59.41	212.69
max_rally	69.66	199.38
vol_week	51.66	191.34
diversification	50.40	191.06
averagePerfFollowee_month	48.78	190.91
vol_month	43.12	188.52
averagePerfFollowee_week	50.21	185.03
totalPos	45.81	181.22
ROI_week	43.25	178.38
ROI_day	27.37	174.64
averagePerfFollowee_day	34.58	169.96
ROI_month	34.51	157.31
daysPositive	17.16	72.33
cluster	19.74	39.99



Tbl. 4-5. Importance values returned (left) and graph of importance values (right) for Period 2 following

Variables selected by lasso with cross validation:

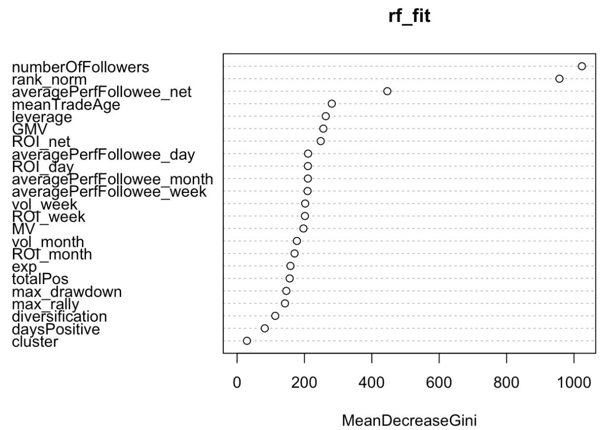
```
## [1] "numberOfFollowers" "rank_norm"         "vol_month"
## [4] "max_rally"         "leverage"          "meanTradeAge"
## [7] "exp"               "diversification"
```

Based on all three selection metrics, we use the following variables for P2 following modeling:

```
Rank_norm
numberOfFollowers
averagePerfFollowee_net
averagePerfFollowee_month
ROI_net
ROI_day
vol_month
Max_rally
Max_drawdown
Leverage
meanTradeAge
GMV
Exp
```

4.3.4 Period 2 Unfollowing

Predictor	MeanDecreaseAccuracy	MeanDecreaseGini
numberOfFollowers	58.77	1023.56
rank_norm	62.14	956.81
averagePerfFollowee_net	68.77	446.01
meanTradeAge	56.02	281.20
leverage	33.59	263.35
GMV	30.82	256.04
ROI_net	48.49	248.36
averagePerfFollowee_day	35.88	210.75
ROI_day	32.17	210.34
averagePerfFollowee_month	36.49	210.23
averagePerfFollowee_week	55.11	209.24
vol_week	42.11	202.20
ROI_week	55.33	201.42
MV	27.70	197.10
vol_month	30.49	177.63
ROI_month	32.88	170.76
exp	44.30	158.48
totalPos	51.51	156.17
max_drawdown	44.10	146.17
max_rally	50.60	142.31
diversification	43.05	113.40
daysPositive	23.16	82.26
cluster	19.79	29.38



Tbl. 4-6. Importance values returned (left) and graph of importance values (right) for Period 2 unfollowing

Variables selected by lasso with cross validation:

```
## [1] "numberOfFollowers" "rank_norm"         "ROI_day"
## [4] "vol_week"          "vol_month"         "max_rally"
## [7] "totalPos"          "leverage"          "meanTradeAge"
## [10] "exp"              "diversification"
```

Based on all three selection metrics, we use the following variables for P2 unfollowing modeling:

```
Rank_norm
numberOfFollowers
averagePerfFollowee_net
averagePerfFollowee_month
meanTradeAge
ROI_net
ROI_day
GMV
Leverage
Exp
Diversification
vol_month
```

4.4 SMOTE Sampling

We follow the conventional approach creating 67:33 train-test split. However, a simple confusion matrix shows that our dataset is highly imbalanced. There are significantly more days where leaders do not get new followers/unfollowers than days leaders do get them. SMOTE is able to over sample days with new follower/unfollower and under sample days without in our training set before testing in the realistic imbalanced settings in our testing set [11]. It mitigates the problem of overfitting caused by simple replication of data points.

4.5 Modeling With Lasso Logit Regression

For our classification problem of predicting a new follow/unfollow for a leader on a particular day, logistic regression is the classic model for binary classifications. It is usually the go-to for users and offers a baseline for other machine learning algorithms. Logistic regression doesn't perform well with a large number of features or categorical features with a large number of values, but it still predicts well when working with correlated features. However, we still use lasso to put a constraint on the regression coefficients to mitigate problems that could be caused by multicollinearity. We list regression results.

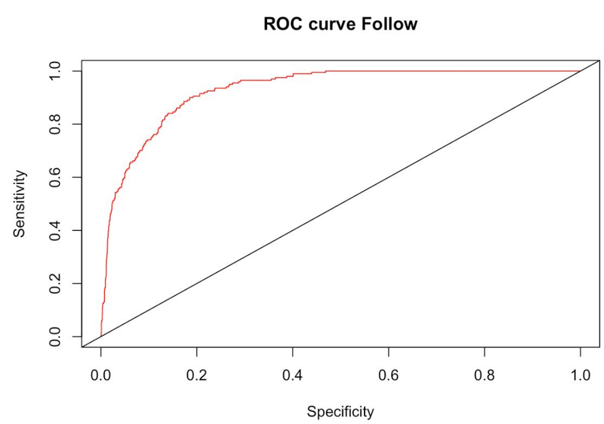
4.5.1 Period 1 Following

hasNewFollow ~ rank_norm + numberOfFollowers + averagePerfFollowee_net + ROI_net + vol_month + max_rally + leverage + meanTradeAge + MV + GMV

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                      1
## (Intercept)              -2.435696e+00
## rank_norm                 2.631736e+00
## numberOfFollowers         9.345609e-04
## averagePerfFollowee_net   2.004642e-07
## ROI_net                   .
## vol_month                -9.925570e+01
## max_rally                 5.460178e+01
## leverage                  1.610167e-03
## meanTradeAge             -4.608027e-02
## MV                       -2.084546e-08
## GMV                      -9.849001e-11
```

Tbl. 4-7. Logit coefficients for P1 following

```
## Confusion Matrix and Statistics
##
##
## lasso_pred    0    1
##          0 2344   29
##          1  437  172
##
##                Accuracy : 0.8437
##                  95% CI : (0.8302, 0.8566)
##     No Information Rate : 0.9326
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3598
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8429
##             Specificity : 0.8557
##          Pos Pred Value : 0.9878
##          Neg Pred Value : 0.2824
##              Prevalence : 0.9326
##          Detection Rate : 0.7860
##    Detection Prevalence : 0.7958
##       Balanced Accuracy : 0.8493
##
##        'Positive' Class : 0
##
```




```
auc.perf = performance(pred, measure = "auc")
auc.perf@y.values
```

```
## [[1]]
## [1] 0.928969
```

Fig. 4-2. P1 following model evaluation

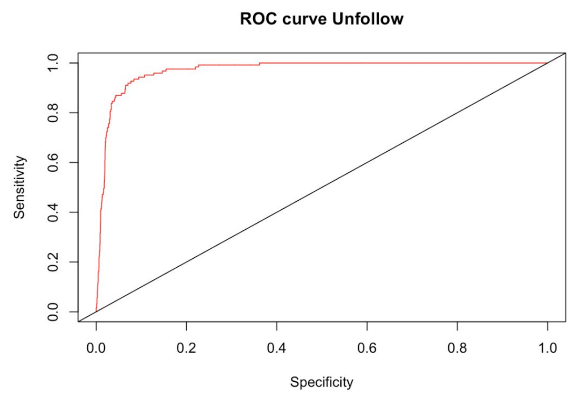
4.5.2 Period 1 Unfollowing

hasNewUnfollow ~ rank_norm + numberOfFollowers + averagePerfFollowee_net + ROI_net + GMV + leverage + cluster

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)             -8.852875e+00
## rank_norm                8.099874e+00
## numberOfFollowers       -2.359149e-04
## averagePerfFollowee_net  3.777533e-06
## ROI_net                  .
## GMV                     -6.618534e-09
## leverage                -4.059482e-03
## cluster                 -2.451047e-01
```

Tbl. 4-8. Logit coefficients for P1 unfollowing

```
## Confusion Matrix and Statistics
##
##
## lasso_pred    0    1
##          0 2590    8
##          1  269  115
##
##                Accuracy : 0.9071
##                  95% CI : (0.8961, 0.9173)
##     No Information Rate : 0.9588
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4172
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9059
##             Specificity : 0.9350
##          Pos Pred Value : 0.9969
##          Neg Pred Value : 0.2995
##              Prevalence : 0.9588
##          Detection Rate : 0.8685
##    Detection Prevalence : 0.8712
##       Balanced Accuracy : 0.9204
##
##        'Positive' Class : 0
##
```




```
auc.perf = performance(pred, measure = "auc")
auc.perf@y.values
```

```
## [[1]]
## [1] 0.9710314
```

Fig. 4-3. P1 unfollowing model evaluation

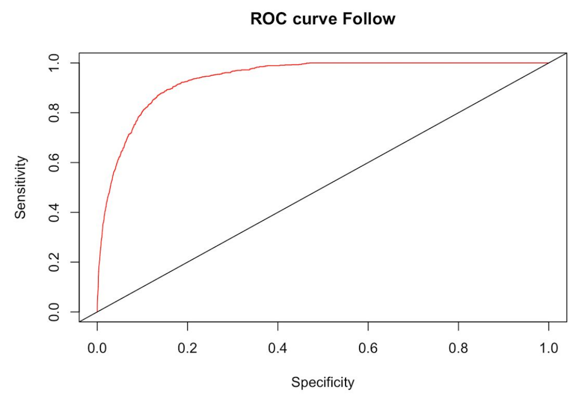
4.5.3 Period 2 Following

hasNewFollow ~ rank_norm + numberOfFollowers + averagePerfFollowee_net + averagePerfFollowee_month + ROI_net + ROI_day + vol_month + max_rally + max_drawdown + leverage + meanTradeAge + GMV + exp

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)               -1.886634e+00
## rank_norm                  3.290288e+00
## numberOfFollowers          6.024042e-03
## averagePerfFollowee_net    2.714595e-09
## averagePerfFollowee_month  1.354595e+00
## ROI_net                    7.116953e-09
## ROI_day                    9.961546e+00
## vol_month                 -2.243824e+02
## max_rally                  2.665063e+01
## max_drawdown               3.817675e+01
## leverage                   1.251247e-03
## meanTradeAge              -5.332028e-03
## GMV                       -9.531976e-19
## exp                       -1.197649e+00
```

Tbl. 4-9. Logit coefficients for P2 following

```
## Confusion Matrix and Statistics
##
##
## lasso_pred     0     1
##          0 36622    90
##          1  7814   916
##
##                Accuracy : 0.8261
##                  95% CI : (0.8225, 0.8295)
##     No Information Rate : 0.9779
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1546
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8242
##             Specificity : 0.9105
##          Pos Pred Value : 0.9975
##          Neg Pred Value : 0.1049
##              Prevalence : 0.9779
##          Detection Rate : 0.8059
##    Detection Prevalence : 0.8079
##       Balanced Accuracy : 0.8673
##
##        'Positive' Class : 0
##
```




```
auc.perf = performance(pred, measure = "auc")
auc.perf@y.values
```

```
## [[1]]
## [1] 0.9382322
```

Fig. 4-4. P2 following model evaluation

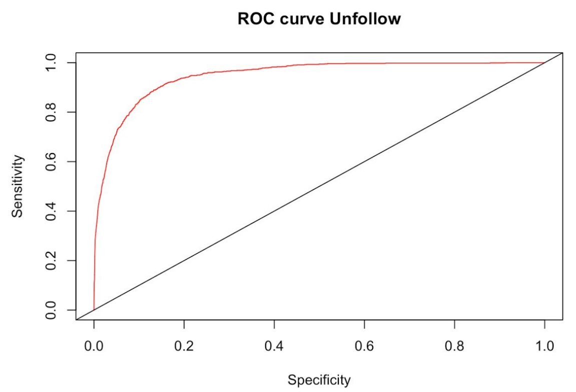
4.5.4 Period 2 Unfollowing

hasNewUnfollow ~ rank_norm + numberOfFollowers + averagePerfFollowee_net + averagePerfFollowee_month + meanTradeAge + ROI_net + ROI_day + GMV + leverage + exp + diversification + vol_month

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                       1
## (Intercept)                -3.114870e+00
## rank_norm                   3.041616e+00
## numberOfFollowers           1.217306e-02
## averagePerfFollowee_net     4.703079e-09
## averagePerfFollowee_month   6.813308e+00
## meanTradeAge               -3.802146e-03
## ROI_net                     .
## ROI_day                     1.033047e+01
## GMV                        -1.095107e-18
## leverage                    2.900376e-03
## exp                        -5.880896e-01
## diversification             9.678801e-03
## vol_month                  -1.243244e+02
```

Tbl. 4-10. Logit coefficients for P2 unfollowing

```
## Confusion Matrix and Statistics
##
##
## lasso_pred     0     1
##          0 37827    98
##          1  6595   922
##
##                Accuracy : 0.8527
##                  95% CI : (0.8494, 0.856)
##     No Information Rate : 0.9776
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1837
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8515
##             Specificity : 0.9039
##          Pos Pred Value : 0.9974
##          Neg Pred Value : 0.1227
##              Prevalence : 0.9776
##          Detection Rate : 0.8324
##    Detection Prevalence : 0.8346
##       Balanced Accuracy : 0.8777
##
##        'Positive' Class : 0
##
```




```
auc.perf = performance(pred, measure = "auc")
auc.perf@y.values
```

```
## [[1]]
## [1] 0.9467367
```

Fig. 4-5. P2 unfollowing model evaluation

4.6 Discussion

Based on the above results, we draw the following observations:

1. Popularity relative to other leaders (aka normalized rank) matters more than current number of followers
2. Preferential attachment exist in leader's popularity -- popular leaders get more popular
3. A leader's "performance" in social trading is not just measured by his own ROI; average return of his followers matters more

4. Salience plays a role in follower's decision making: rare occurrences such as maximum drawdown and rally have larger and more significant beta in predicting whether a leader will get new followers.
5. Leverage has a significant, nontrivial beta in both periods. Given forex traders take highly levered risky positions, this variable is significant by the nature of the trading environment
6. Signs of regression coefficients do not seems to differ between following and unfollowing. This seems to suggest that follow/unfollow decision is random, and it would be better to model "social activity".

Chapter 5

Follower Perspective -- Does Foraging Improve Performance?

In this chapter, we solve the following/unfollowing problem from a follower's perspective, taking a microscopic view of how followers follow and whether followers exhibit common behavioral patterns. We claim that when choosing who to follow, traders show animalistic characteristics, in that they forage for their leaders. We claim that different types of eToro traders forage differently, and foraging is a hedge against significant loss but rarely lead to outstanding gains.

5.1 Definition of Networkers

Taking a close look at *activity* revealed an interesting following/unfollowing pattern shown in Fig. 5-1.

Sample Explorer-type Mirroring Behavior

	date	links	linksCreated	linksDestroyed
1:	2011-12-04	1	1	0
2:	2012-01-05	1	0	1
3:	2012-01-17	1	1	0
4:	2012-01-23	1	0	1
5:	2012-01-31	1	1	0

Fig. 5-1. Explorers create and destroy the same number of links at short intervals, averaged at 4.8 days, $p < 0.0001$.

One explanation for the above behavior is that people have limited social capacity [8]. When looking for the right target to mirror, they simply cannot create infinitely many connections due to system restrictions, attention span, or risk appetite. Therefore, these traders create a set of following links and destroying all of them at a later date, then creating another set of following links and destroying all of them again. This pattern repeats until the social trader has built the right profile to follow. Only then their exploratory activities decrease and their social network stabilizes.

We call these traders networkers:

> ***Networkers*** *is one category of social traders who explore the eToro ecosystem through frequent following and unfollowing activities. Networkers start to follow a number of leaders on one day, wait for a few days, unfollow all these leaders. They then start to follow another set of leaders, wait for a few days, and then unfollow all these leaders. This pattern continues. If a trader exhibit this behavior (aka having activity = 0.5) for more than 10 days in a period, then he is a networker.*

We found 81 networkers in period 1 and 168 networker in period 2. There was only 1 trader who foraged in both periods.

5.2 Demographics of Networkers

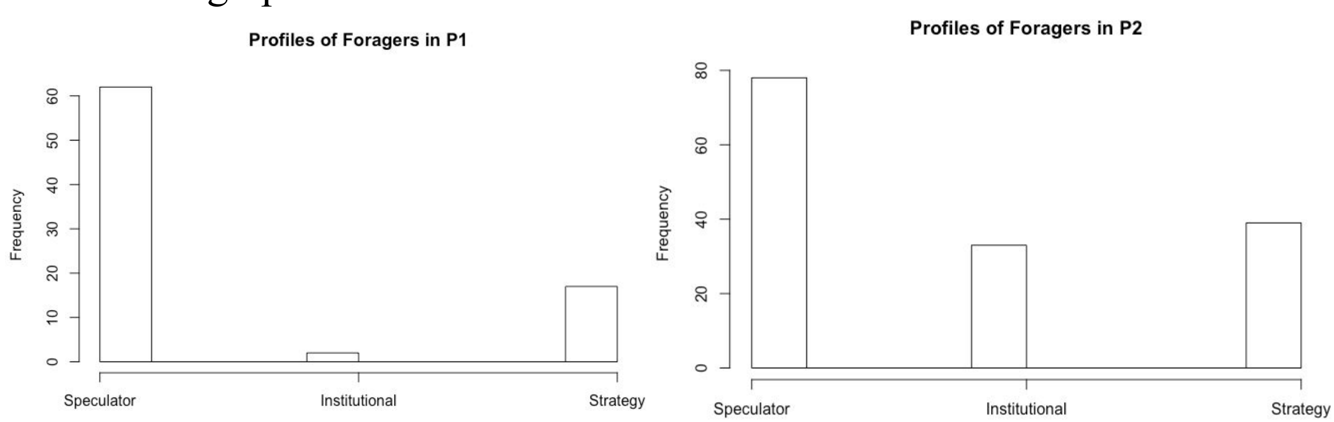


Fig. 5-2. Profiles of networkers in period 1 and 2

5.3 Does Foraging Improve Performance?

The most relevant question is, therefore, whether foraging as a treatment improves performance? Does it improve performance for all trader types? We calculate per period ROI of all traders on the platform, and there's the summary statistics in Tbl. 5-1 and Tbl. 5-2:

P1

Type	**Institutional**		**Speculator**		**Strategy**	
Forage Or Not	**No Forage**	**Forage**	**No Forage**	**Forage**	**No Forage**	**Forage**
Min	-1.48E+10	0.98	-2.02E+13	-1.47E+09	-1.67E+12	-7.98E+10
1st Qu.	-45.00	60.16	-2.00	-2.00	-444.00	-9.00
Median	60.00	119.30	-1.00	-1.00	-1.00	-4.00
Mean	-1.62E+08	119.30	-3.88E+09	1.82E+09	-3.88E+09	-4.64E+09
3rd Qu	2381999.00	178.50	-1.00	-1.00	3181999.00	12.00
Max	6.53E+09	237.70	1.44E+13	1.14E+11	8.51E+11	9.63E+08
ANOVA p val	0.88		0.89		0.95	

Tbl. 5-1. Period 1 ROI summary for different types of traders with and without foraging

P2

Type	Institutional		Speculator		Strategy	
Forage Or Not	No Forage	Forage	No Forage	Forage	No Forage	Forage
Min	-8.94E+14	-1.87E+10	-1.58E+14	-8.66E+11	-3.78E+14	-1.33E+11
1st Qu.	-3.05E+09	-1.58E+09	-1.00	-8.23E+08	-3.00	-6.12E+06
Median	-1.00	-6.36E+07	-1.00	-1.00	-1.00	-1.00
Mean	-1.35E+11	3.96E+08	-1.04E+09	-9.27E+09	-5.57E+09	9.90E+09
3rd Qu	5.46E+09	1.16E+09	-1.00	-1.00	-1.00	1.01E+09
Max	5.33E+14	2.94E+10	1.19E+14	2.28E+11	5.10E+14	2.76E+11
ANOVA p val	0.95		0.94		0.99	

Tbl. 5-2. Period 2 ROI summary for different types of traders with and without foraging

Although foraging seems to produce higher mean ROI for institutional traders in both periods and is beneficial for speculators in period 1 and strategists in period 2, ANOVA p values are not significant enough for us to make a confident claim that foraging would improve performance for one type vs another. The key takeaway is that foraging would help mitigate loss and that it is a good hedging mechanism, but would not generate outstanding returns or alpha. If we consider social traders as investors betting on people-portfolios, then foraging would be a good way to hedge against a single terrible leader. Wisdom of the crowd would at least bring a trader's performance close to the mean. This is the benefit of exploiting the environment and maintaining highly diversified people portfolio. Wisdom of the crowd, by construction (20% return shared between 1 person vs shared between thousands), would not lead to outstanding gains. So if a trader believes that he/she has skills, or alpha, it is better to not forage.

Chapter 6

Closing Remarks

In this chapter, we end the paper with a discussion on the results from Chapter 4 and Chapter 5. We then discuss on a high level how our conclusion is relevant to understanding phenomenon in the real market. Finally, we talk about areas in which further research could be done.

6.1 Conclusion

6.1.1 Applications of Our Research

Our research has shown that given a novel financial ecosystem, there will always be leaders and followers and the crowd tends to follow similar types of leaders. This falls in the same line as investor profiling where certain personalities emerge as leaders.

Our research can also be used as an algorithm for eToro to identify potential leaders and weed out bad traders to prevent significant losses. Lastly, our findings confirm many behavioral finance principles such as preferential attachment, endowment effect, and salience. In this regard the research can be further refined for behaviorists to extend the scope of our study.

6.1.2 Relating to the Field of Behavioral Finance

Traditional finance focuses on how people should make decisions while behavioral finance deals with how they actually reach their decisions. By combining cognitive psychological theory with conventional economics, behavioral finance researchers are able to generate new insights to decode human intuition [10]. For example, most researchers suspect that the financial crisis of 2008 triggered a behavioral change, especially amongst younger investors, to shift away from risky assets such as stocks and toward safer and more liquid investments. But to what extent did the crisis have an impact and how can one measure the effect? Traditional finance theory cannot explicate such phenomenon because it believes that market fundamentals matter but investor experiences do not. Behavioral finance serves as a better descriptor and predictor to rationalize intuition and market movements.

This paper covers the big-data approach at one staggering field of research in behavioral finance -- herding. The emergence of new trading strategies such as "twitter trades" is evidence of investors' tendency to follow the crowd and act irrationally [6]. Investing solely based on company fundamentals no longer suffice. Successful investment thesis should also incorporate sentiment into valuation. Successful investors are the ones who takes a multidisciplinary approach to see the market from different angles and identify hidden opportunities blind to the average.

Both traditional and behavioral finance provide valuable contributions and should be viewed as complementary rather than mutually exclusive. In fact, observation of actual behavior informs the development of good theory. As Fox commented, "while behaviorists and other critics have poked a lot of holes in the edifice of rational market finance, they haven't been willing to abandon that edifice." No matter how good the dynamics model become, it simply cannot predict the next move of Federal Reserve or energy prices. One have to pay attention to the specific context when applying these new principles.

6.2 Future Work

6.2.1 Event Study to Prove Robustness

Many events happened in 2013 on eToro, which serves as great opportunities for natural experiments. If keepers of social trading really exist, i.e. traders adopt the social trading concept, then internal and external shocks to the platform should not have any impact on investors' social trading behaviors. Even if there is, there should only be a temporary shift in behavior. Keepers of the concept should revert to previous behaviors shortly after the shocks.

6.2.1.1 Internal Shock

One trader, LifeForge, was reaching 986.8% annual return on March 13th, 2013 before experiencing a significant drawdown on March 29, 2013. His number of followers went up significantly. This might be due to LifeForge not closing his trades and only continuously adding money to his portfolio. When his loss exceeded his GMV, the loss are revealed to his followers.

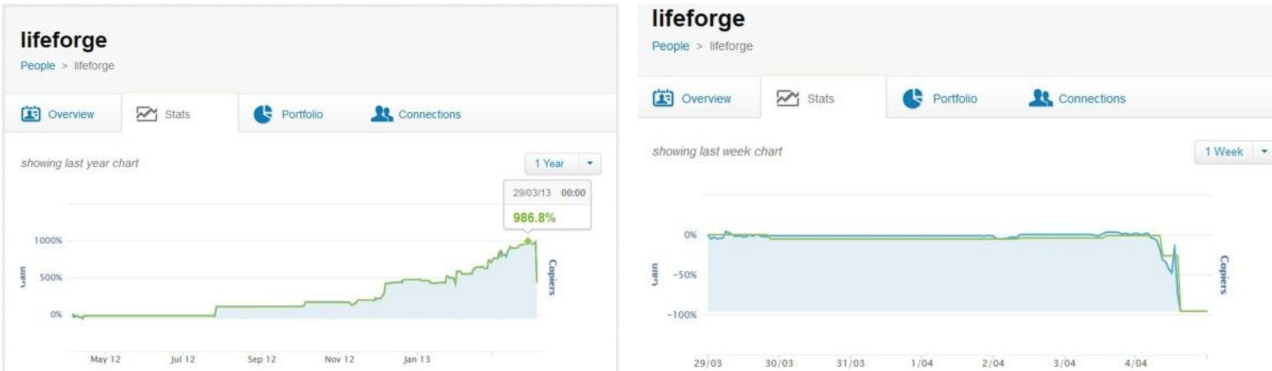


Fig. 6-1. LifeForge performance in 2013. The trader was generating significant returns of close to 1000% before losing all of his capital. One hypothesis is due to the way eToro calculate returns, the trader did not close up his non-profitable trades. Once he has lost all of his capital (MV) his losses were revealed, which triggered significant un-mirror behaviors from his followers and many left the site.

Traders who included "lifeforge" as part of their 20 people portfolio do not suffer such drastic losses as opposed to those eToro traders copying "lifeforge" exclusively or in larger percentage.

6.2.1.2 External Shock

Cyprus bailout and US censorship where two external triggers that the eToro network experienced in 2013 March. Significant volatilities in the number of users and mirroring resulted. In addition to studying individual motivation for mirror, we want to study how external events to the system triggers fundamental, structural shift in the mirror network. How long does it take for the shock to propagate through the network. If social trading is really a resilient system, how long does it take to reestablish equilibrium?

6.2.2 Does eToro Make Traders More Social?

So far we have built a model to predict following and unfollowing. But is following behavior caused by social trading? What are the odds that these behaviors happen due to mere chance? Does more social trading makes one more social?

We therefore propose the following field experiment to eToro: Randomly sample certain number of private traders and introduce a promotional program, which gives them a cash incentive, to only social trade for 6 months. After the trial period, the platform terminate the incentive program and observe how many of these traders revert back to their old trading behavior, trading only on their original ideas. In introducing a treatment of social trading and collecting results after the treatment we can have more confidence in our hypothesis that the platform is the reason for one to become a "social trader". However, with any experimental design, caveats exist. The experiment does not control for any factors external to the platform. For example, market beta is not accounted for. During times of economic growth, which was the case between 2011 and 2013, income effect may have encouraged more investment activities in general. If this experiment is implemented, researchers need to bear in mind other covariances when interpreting the result.

Bibiography

[1] "Behavioral Finance and Investment Processes”. *Finance Strokes*: n. pag. Web. 15 Dec. 2016.

[2] “Social Trading”. http://www.social-trading.worldfinance.com/.

[3] Mohan, Akshay. Olguin Olgufn, Daniel. Waber, Daniel. Kim, Taemie. Ara, Koji. Pentland, Alex. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(1):43-55, 2009.

[4] Pentland, Alex. “Beyond the Echo Chamber.” *Harvard Business Review*, 20 Mar. 2017, hbr.org/2013/11/beyond-the-echo-chamber.

[5] Pentland, Alex. *Social Physics: How Good Ideas Spread-the Lessons from a New Science*. New York: Penguin, 2014.

[6] Fang, Rui. Liu, Xiaomo. Li, Quanzhi. Shah, Sameena. “Tweet Sentiment Analysis by Incorporating Sentiment-Specific Word Embedding and Weighted Text Features”. IEEE/WIC/ACM International Conference on Web Intelligence (WI) Conference, DOI: 10.1109/WI.2016.0097, October 2016.

[7] Rajendran, G. Thomas, TC. ‘Behavioural Aspects of Capital Market Investors in India’, *Journal of Science*, Technology and Management, vol. 5, no.4, pp.25-33.

[8] Miritello, Giovanna et al. "Limited Communication Capacity Unveils Strategies For Human Interaction". Scientific Reports 3 (2013): n. pag. Web. 15 Dec. 2016.

[9] MacQueen, J. “Some methods for classification and analysis of multivariate observations.” *MacQueen : Some methods for classification and analysis of multivariate observations*, The Regents of the University of California, projecteuclid.org/euclid.bsmsp/1200512992.

[10] Nofsinger, John. Baker, Kent. *Behavioral Finance-Investors, Corporations, and Markets*. Hoboken, NJ: John Wiley & Sons, Inc.

[11] Chawla, Nitesh. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 853–867. Springer, (2005).

[12] Pan, Wei. “Reality hedging : social system approach for understanding economic and financial dynamics.” DSpace@MIT, Massachusetts Institute of Technology, hdl.handle.net/1721.1/97971.