Investors Are Social Animals: Predicting Investor Behavior using Social Network Features via Supervised Learning Approach

Yuxian, Eugene Liang
Management Information System
National Cheng Chi University
Taipei, Taiwan
eugene@liangeugene.com

Soe-Tsyr Daphne Yuan
Management Information Systems
National Cheng Chi University
Taipei, Taiwan
yuans@nccu.edu.tw

ABSTRACT

What makes investors tick? In this paper, we explore the possibility that investors invest in companies based on social relationships be it positive or negative, similar or dissimilar. This is largely counter-intuitive compared to past research work. In our research, we find that investors are more likely to invest in a particular company if they have stronger social relationships in terms of closeness, be it direct or indirect. At the same time, if there are too many common neighbors between investors and companies, an investor are less likely to invest in such companies. We use social network features such as those mentioned to build a predictive model based on link prediction in which we attempt to predict investment behavior.

Categories and Subject Descriptors

D.3.3 [**Social and Behavioral Sciences**]: Economics

General Terms

Human Factors, Algorithms, Experimentation, Economics.

Keywords

investment behavior, social network analysis, link prediction.

1. INTRODUCTION

With Facebook's IPO fresh in our minds and strings of startups, startup incubators popping up in Silicon Valley and around the globe, many entrepreneurs will have come across the act of raising investments from investors. Such behavior is not limited to startups: small medium enterprises or even large companies seek external investments as a way to enhance cashflow or meet various business objectives.

While the topic of investments is one of the most widely discussed topics in the realm of investing and business, there are limited studies that provide evidence as to how companies can raise investments from investors. One way to understand how companies can increase their chances of receiving investment from investors is to understand what investors are looking for, that is factors that affect investing behavior.

There are many studies that seek to understand investment behavior. Factors such as psychological, geographic differences, investment experiences and even genetics have been proposed as what spurs investments. However, most research fails to consider the role of social relationships between investors and companies.

Our main hypothesis is that investors have a tendency to invest in companies that have exhibit certain social relationships between them, be it similar or dissimilar between the investor and the company in question. For example, we might expect an investor to invest in a company that is "closer" (similar) socially, such as in terms of shortest path. At the same time, if there is a form of competitive (negative or dissimilar) relationship, such as having too many common neighbors between the investor and the company, we do not expect the investor to invest in that company in this case.

Our contributions to the literature are as follows:

Modeling prediction of investment behavior as a link prediction problem: We build a social network using data from CrunchBase, the largest public database with profiles about companies. Using this dataset, we attempt to predict if an *Investor* will invest in a *Company* based on their social relationship. To the best of our knowledge, our work is amongst the first to model investment behavior as a link prediction problem.

Combining multiple link prediction features to gain greater insight of social networks: Various link prediction techniques such as Common Neighbors, Shortest Path, Jaccard Coefficient and others provide useful insights as to how a pair of nodes may be related within a social network. Nonetheless, each technique only reveals certain aspects of a social network: for example Common Neighbors measures the number of neighbors that are common between two nodes in a social network while Shortest Path measures the shortest number of hops between two nodes in a social network. We believe that combining multiple approaches will provide us with a holistic view of a social network.

Marriage of social network analysis with investing behavior: We explore how similarity between investors and companies affect investing behavior through social network analysis. Also, our work is amongst the first to use data from CrunchBase as a social network for research purposes.

Appearing in Proceedings of the Workshop on Mining and Learning with Graphs (MLG-2013) Chicago, USA, 2013 by Yuxian, Eugene Liang and Soe Tsyr Daphne Yuan

Insights to collective behavior of investors: Our research also provides a collective overview as to how investors invest within a social network.

2. RELATED WORK

We have three parts for related work since our research is amongst the earliest that makes use of the CrunchBase dataset, and that our research focuses on the use of link prediction on investment behavior.

2.1 Related Research using the CrunchBase Dataset

Eugene and Daphne [1] performed descriptive data mining using the CrunchBase dataset and uncovered general rules for companies seeking investment. They considered features adapted from graph theories and social network analysis such as shortest path, Adamic/Adar, Jaccard coefficient, common neighbors, and preferential attachment. In general, the greater the similarity (in the case of shortest paths, Adamic/Adar and Jaccard Coefficient) the more investment activities were found. There were counter-intuitive results too: the greater the number of common of neighbors and that the greater the Preferential Attachment score, the less likely that Investors will invest in companies.

Guang, Zheng, Wen, Hong, Rose and Liu [12] performed studies using the CrunchBase dataset and predicted company acquisitions with factual and topic features using profiles and news articles on TechCrunch. Although they made use of a similar dataset as our work, their work did not make use of social relations as part of their feature set and focused on a different domain of mergers and acquisitions.

2.2 Previous Research on Investment Behaviors

Prior studies on investment behaviors can be categorized into 6 categories based on the type of factors that drive investment behaviors.

Personal Opinions. Doran, Peterson and Wright [6] studied the role of personal opinions of finance professors on the efficiency of the stock market in the United States and found out that personal opinions do not affect investment behaviors. Rather, investment behaviors found in financial professors were largely driven by the same behavioral factor as amateur investors.

Investment Experience. Hege and Schwienbacher [8] analyzed the differences in the investment behaviors of experienced and novice private equity firms and found out that novice firms tend to invest more slowly than experienced funds but the size and value of the funding size of novice firms tend to be larger.

Geographic Identities. Grinblatt and Keloharju [9] discovered that investment behaviors can be determined by the investors' geographic identity: foreign investors in Finland tend to purchase past winning stocks and sell past losers. On the other hand, domestic investors sell past winning stocks and purchase losing stocks.

Inline versus Offline Communities. Tan and Tan [5] explored the roles played by online and offline communities and discovered that offline communities are more influential over investing behaviors.

Psychology. Bakker, Hare, Khosravi and Ramadanovic [7] on the other hand investigated into psychological factors that impact market evaluation and found out that trust and social influence affects the stability of investment markets.

Genetics. Amir, Henrik and Stephan [4] investigated the relationship between genetics and investment behavior by studying the investment behaviors of identical and fraternal twins. They discovered that "a genetic factor" explains up to a third of twins' investing behavior, though not long lasting.

2.3 Previous Research on Link Prediction

Link prediction is one of the most important topics in social network analysis and in recommender systems. Link prediction seeks to predict the changes in terms of edges or nodes of social networks over time. Link prediction in social networks can be problematic: Nowell and Kleinberg [2] performed extensive studies on link prediction in social networks and noted that there is no singular technique that can ensure the best performance; the techniques used for the experiment shows limited performance. The techniques used for link prediction include PageRank[20], HITS[21], Adamic/Adar[3], Jaccard Coefficient, shortest paths etc. Moreover, Nowell and Kleinberg [2] proposed that performance may be improved by taking into account of node-specific information. More recently, link prediction has been applied to datasets in popular social networks, which includes Twitter, Facebook and others [18,19,22]. These studies include the prediction of positive and negative links to recommending friends on Facebook to using computationally efficient topologic features.

The originality of this paper is that we propose the use of social relationship (represented by social network features) as the main way to predict if investments will occur. For example, given an *Investor* and a *Company*, can we predict if the *Investor* will invest in that particular *Company* just by understanding their social relationships? We believe that this will be a much easier approach for companies seeking investments since they are more likely to understand their social relations with potential investors.

3. Investors Are Social Animals

3.1 Methodology

We model the investment behavior as a classic link prediction problem. In general, we compare every pair of *Investor* and *Company* and attempt to predict if the *Investor* will invest in that *Company* based on how similar or dissimilar in terms of their social relationship.

3.2 The CrunchBase Dataset

CrunchBase (http:/www.crunchbase.com) is an open dataset which contains information about startups, investors, founders, trends, milestones and other related information. It relies on the community to provide and edit most of its content. As of 16th May 2012, CrunchBase consists of profiles of 89,370 companies, 118,888 people, 7,759 financial organizations, 4,308 service providers, 28,109 investment rounds and 6,596 acquisitions.

We chose Facebook as the seed node, and gathered *People*, *Companies* and *Financial Organizations* found in its social and investment relationships within 4 degrees of separation from Facebook.

We selected Facebook as the seed node due to the company's meteoric rise in the social network industry and it's much hyped IPO recently. We chose 4 degrees of separation as a cutoff point as opposed to 6 degrees of separation due to the fact that recent advances in technology have reduced the degrees of separation between people as shown in [13]. In addition, there are limits to the “Horizon of Observability” [10] from the viewpoint of using Facebook as a seed node.

Our final dataset contains 11916 companies, 12127 people, and 1122 financial organizations within 4 degrees of separation from Facebook. The entity and relationship types that provided by CrunchBase are as follows:

3.2.1 Entity Types

People/Person. *People* (person) refer to founders, executives and other persons working for a particular company or organizations. Examples from our dataset include Mark Zuckerberg and Peter Thiel. A single *Person* has the same definition as *People* for our purposes.

Companies. Some popular examples of *Companies* include Google, Facebook and Microsoft.

Financial Organizations. *Financial Organizations* are organizations that typically perform the act of investment on *Companies*. Prominent examples in our dataset include Accel Partners and Digital Sky Technologies.

3.2.2 Relationship Types

Social. We define S*ocial relationship* as an instance where a *Person (People)* has previously or currently works for a particular *Company* or *Financial Organization*. Since there is no way of finding out if the *People* (person) is recruited by the company or wanted to work for that particular company or financial organization in question, *social relations* are undirected. For instance, Bret Taylor[1] has a social relationship with Google and Facebook since he has previously worked for both of the companies.

Investment. Investment relationships are created as a result of an investment act of a Person, Company and or a Financial Organization on a Company. For example, Microsoft invested in Facebook, thus resulting in an Investment relationship.

3.3 The CrunchBase Social Network

Using the dataset from CrunchBase, we build a network based on the entity and relationship types, where nodes represent entities, while relationship represent edges.

3.3.1 Simplification of Network

Since we are interested in the prediction of investment acts, we further simplify the network into only 2 node types: Investors and Companies:

Investors. Investors are made of *People/Person*, *Financial Organizations* and *Companies*. Note that companies can play the role of an investor, take for instance companies like Google, Microsoft and Facebook make investments in smaller companies.

Companies. Companies are simply *Companies*, which may or may not have received any investments.

3.3.2 Types of Network

Going deeper into the dataset for our experiment purposes, we note that we can categorize the dataset into 2 types of network: Investment and Social Network:

G_{Social}: G_{Social} is an undirected graph derived from social relationships. This means that the edges are made up of social relationships only. We do not include the act of investment as part of *social relations* as investment behavior is what we are attempting to predict. Nodes represent *Investors* and *Companies*. If we take into account of social relationships only, features such as shortest paths, common neighbors and so on differ greatly to the graph that considers both social and investment relationships. For instance, G_{Social}’s shortest paths ranges from 1 to 19 hops.

$G_{Investor}$: $G_{Investor}$ is a directed graph derived from investor relationships. This means that the edges are made up of investor relationships only. Similarly, nodes represent *Investors* and *Companies*.

3.3.3 Network Used for Experiment

Our final dataset consists of $\mathrm{G_{Social}}$, where edges represent social relationship and nodes are made up of Investors and Companies. $\mathrm{G_{Investor}}$ is used to provide ground truth labels.

3.4 Problem Formulation

We define the problem of predicting investment as a link prediction problem: given an undirected Social Graph $G_{Social} = (V, E)$ where V represents either an *Investor* i or a *Company* c, and $e = <i,c> \in$ E represents social relationship between an *Investor* and a *Company* that occur at time $\mathrm{T_0}$, predict if an *Investor* will invest in a particular *Company* at $\mathrm{T_1}$

Note that *Investors* consists of *People*, *Companies* and *Financial Organizations*. This is due to duality of roles played by *People*, *Companies* and *Financial Organizations* in the CrunchBase dataset. For example, companies like Microsoft play a dual role of a *Company* and a *Financial Organization* when Microsoft invested in Facebook.

3.5 Modeling Social Relationship

In order to determine the social similarity between an *Investor* and a *Company*, we use features based on node neighborhood, graph distance and common node features between an *Investor* and a *Company*. Each of these features represents a form of similarity in a social sense. The following features were derived from $G_{Social.}$:

3.5.1 Social Network Features

All following features are adapted from graph theories and social network analysis. The algorithms used here for our analysis assign a score (x, y) to pairs of nodes <x, y>, based on the input graph G_{Social}. Nodes X and Y are defined as follows: Node X represents an *Investor*, while node Y denotes a *Company*. This is because we want to compare the similarities of *Investors* and *Companies* for the purposes for our research. No comparisons are made when node X equals node Y. We define the set of neighbors of node x to be $\Gamma(x)$.

Shortest Path. We simply consider the shortest path between *Investors* and *Companies*. The general intuition of shortest path in our context is that Investors are more likely to invest in *Companies* that are found within their “small world”, in which *Investors* and *Companies* are related through short chains.[2, 16]. We define score (x, y) to be the length of the shortest path between an *Investor* and a *Company*. We hypothesize that the smaller the shortest path, the more likely that the *Investor* will invest in that *Company* [1]. The reasoning behind is that the *Investor* is “closer” to the *Company* and hence much easier for them to reach each other.

Adamic/Adar. Adamic and Adar [2] considers similarity between two personal homepages by computing features of the pages and defining the similarity between two pages to be:

[1] Bret Taylor’s CrunchBase Profile. http://www.crunchbase.com/person/bret-taylor

$$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (1)$$

For our purposes, we consider the similarity feature to be the common neighbors. Adamic/Adar weighs rarer features more heavily. The intuition of Adamic/Adar in our context is that *Investors* are more likely to invest in *Companies* that are of greater similarity [1].

Jaccard Coefficient: The Jaccard Coefficient measures the probability that both x and y have a feature *f*, for a randomly selected feature f that either x or y has. Here, we take *f* to be neighbors in G_{Social}, leading us to the measure score:

$$\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \quad (2)$$

Common Neighbors. Common neighbors are considered as the most direct implementation. According to Newman [2], the general intuition is that the number of common neighbors of node X and node Y has a correlation with the probability that they will collaborate in the future, under the context of a collaboration network. For our purposes, investors are less likely to invest if the company in question has greater number of common neighbors due to reasons as explained in [1]. The score(x,y) for common neighbors is defined as follows:

$$|\Gamma(x)\cap\Gamma(y)| \quad (3)$$

Preferential Attachment. Preferential Attachment [15] suggests that the probability that a new edge has node x as an endpoint is proportional the current number of neighbors of x [2]. Results from [1] suggests that investors are less likely to invest in companies with higher preferential attachment. The score(x, y) for preferential attachment is defined as follows:

$$|\Gamma(x)|\cdot|\Gamma(y)| \quad (4)$$

Number of Shortest Paths between an Investor and a Company. We calculate the shortest path between an *Investor* and a *Company* and aggregate the number of paths with the same shortest path score. A node may appear more than once amongst these paths. The intuition here is that an *Investor* is more likely to invest in a *Company* if there are shortest paths connecting them. This is because more paths could mean that the *Company* or *Investor* is more easily reached via multiple shortest paths.

3.6 Learning Algorithms

In this experiment, we chose three learning algorithms: Decision Tree (based on CART algorithm), SVM (with rbf as the kernel) and Naïve Bayes (Bernoulli Model) algorithms. This is to make sure that social network features can indeed be used as reliable indicators for predicting investments.

We selected the Decision Tree learning algorithm as one of the learning methods as we wanted a simple to understand model so that companies seeking investment have a better understanding behind investor's behavior.

More importantly, the model learnt using Decision Tree can be readily visualized; such information can be very useful for companies to gauge their chances of receiving investment from a particular investor.

We also selected SVM and Naïve Bayes as they are widely regarded as classical supervised learning algorithms. More importantly, we selected RBF as SVM's kernel and the Bernoulli model for Naïve Bayes as our data's behavior appears to be more suited for such learning models.

3.7 Significance of Methodology

We believe that our methodology presents several advantages over previous work in terms dataset used, problem formulation/ predictive model and the introduction of new factors for predicting investment behavior.

3.7.1 Richness and size of dataset

We used a dataset from CrunchBase and the size of our network consists of 11916 companies, 12127 people, and 1122 financial organizations within 4 degrees of separation from Facebook. This means that we have a total of 25165 unique nodes in the network. In addition, our dataset consists of very different entities, which include people, companies and financial organizations. These entities also consist of various demographic groups. These factors make our dataset richer and larger as compared to previous works. For example, [9] made use of financial data predominantly while [10] focused on investments in Finland only. In addition, [6] focused their data on only 96 Taiwanese adults. Similarly [7] focused on finance professors exclusively.

3.7.2 Problem formulation and predictive model

While previous work presents merits, there is a lack of generalizability in their approach. This might be due to how the problem of predicting investment behavior is being formulated; in our approach, we chose to model investment behavior as a classic link prediction problem. This allows us to build a model in which investment behavior can be predicted.

3.7.3 Social network features as a factor for predicting investment

Most previous work focused on financial data, psychology, experience etc as factors for predicting investments. We would like to propose the use of social relationships in terms of similarity and differences not only as a factor for predicting investment, but also as a stable and sound possibility.

4. Experiment Setup

4.1 Evaluation Metrics

We use the standard metrics for many binary classification tasks, the true positive rate and false positive rate. We also used the area under the ROC curve (AUC) as our evaluation metric, which represents the trade-off between the true positive and false positives.

4.2 Aggregate and Industry Evaluation

We evaluate our results based on the metrics mentioned above on two levels: aggregate and industry level. Aggregate evaluation is performed when all companies regardless of their industry are taken into account. On an industry level, companies within an industry are evaluated against the aggregate level to see if there are any wide differences in performance. Since CrunchBase.com provides us with the industry ("category_code" field based on the JSON API) of the companies, we use these industry codes to differentiate them across industries. The industries are "web", "software", "mobile", "games_video", "ecommerce", "advertising", "enterprise", "legal", "consulting", "education", "biotech", "semiconductor", "security", "cleantech", "hardware", "search", "other" and "None". "None" occurs where the companies have no industry labels. We do not include "legal" and "None" in our final experiment results due to a lack of positive examples.

4.3 Cross Comparison of Performance Across Different Learning Algorithms

Using the above mentioned evaluation metrics and levels of evaluation, we also compare the performance of these metrics across three learning algorithms stipulated earlier: Decision Tree (CART), SVM (using rbf as the kernel) and Naïve Bayes (Bernoulli model). This is to ensure the soundness of social features as predictors of investment behavior.

4.4 Ground Truth Labels and baseline performance

Using our $G_{Investor}$, we discovered 5341 investment activities. We define such investment activity as an *Investor* investing in a *Company*. For example, when an investment round occurs with 3 investors investing in a company, we take it as 3 investment activities discovered.

We do not have prior results as a basis for comparison since to our knowledge, there are no previous studies that models investment behaviors as a link prediction problem. In addition, previous research and related work mentioned in section 2.2 do not provide baseline performance. Therefore, we regard an acceptable baseline performance for Area Under Curve (AUC) to be greater than 0.6, while True Positive Rate (TPR) baseline should be above 60% and finally, False Positive Rate (FPR) should be lower than 40%.

4.5 Data Split for Training and Testing

We took a 40% training data split for training, with the remaining data for testing purposes. The split of data is based on timestamps, especially for the case for true examples where investments occurs; investments are split by timestamps where earlier investments are used for training while latter investments are used for testing. False examples were split randomly. This was applied to both aggregate and category experiments across three learning algorithms.

5. Experiment Results

We ran the experiment using Decision Trees, SVM and Naïve Bayes algorithms. The results are as follows:

5.1 Aggregate Performance

On the whole, all three algorithms performed above baseline performance of 0.6 for AUC, 60% for TPR (with the exception of Naïve Bayes) and below 40% for FPR.

Figure 1 shows a summary of performance metrics based on AUC. Decision Tree's experiment produced an AUC of 0.77 while SVM produced an AUC of 0.79. Naïve Bayes produced an AUC of 0.77. All three learning algorithms performed better than the baseline performance in terms of aggregate results.

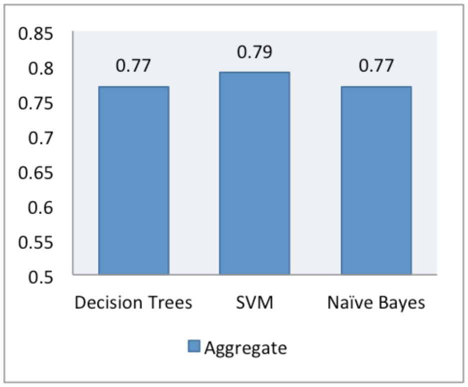

Figure 1. Area Under Curve (Aggregate).

Figure 2 shows a summary of performance based on TPR. The TPR for Decision Tree is 87.53%, SVM registered an aggregate TPR of 89.6%, while Naïve Bayes has an aggregate TPR of 54.8%.

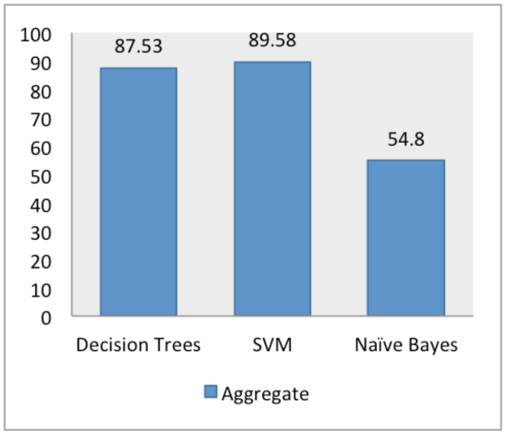

Figure 2. True Positive Rate (Aggregate)

Figure 3 shows a summary of performance based on FPR. The FPR for Decision Tree is 33.18%, SVM registered an aggregate FPR of 33.38%, while Naïve Bayes has an aggregate TPR of 0.05%.

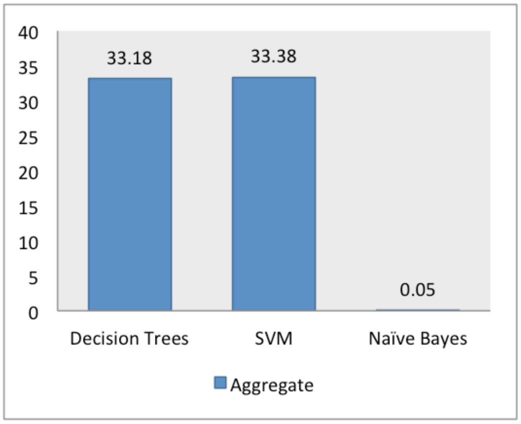

Figure 3. False Positive Rate (Aggregate)

5.2 Industry Performance

We repeated the experiment using by splitting the data by categories with the same 40% training data split and also obtained reasonable performance:

For most of the categories, AUC hovers between 0.63 to almost 0.80 for Decision Trees. Their TPR ranges from 56% to 91%

Similarly, the AUC ranges from 0.65 to 0.84 and the TPR ranges from 51% to 91% for SVM. The AUC ranges from 0.75 to 0.78 and the TPR ranges from 52% to 57% for Naïve Bayes.

The results are shown in Figures 4, 5 and 6.

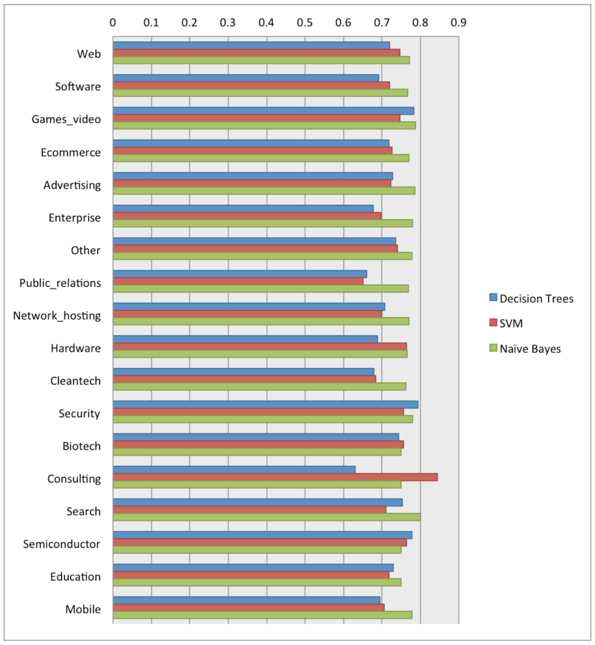

Figure 4. Area Under Curve by Categories.

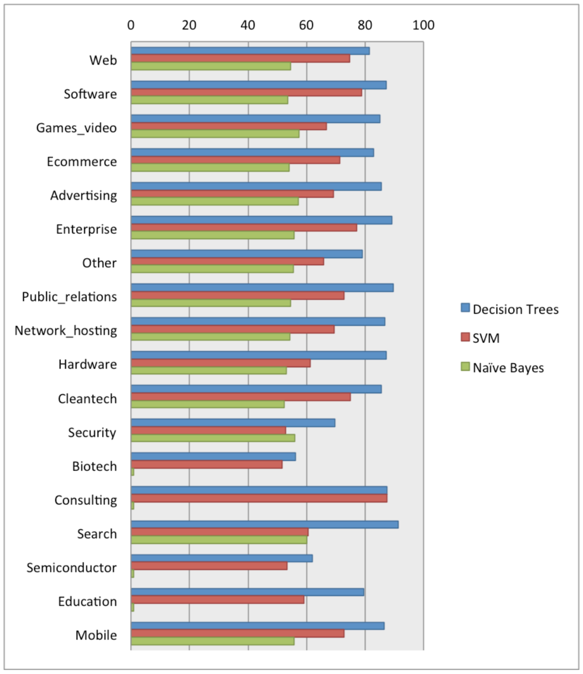

Figure 5. True Positive Rates by Categories.

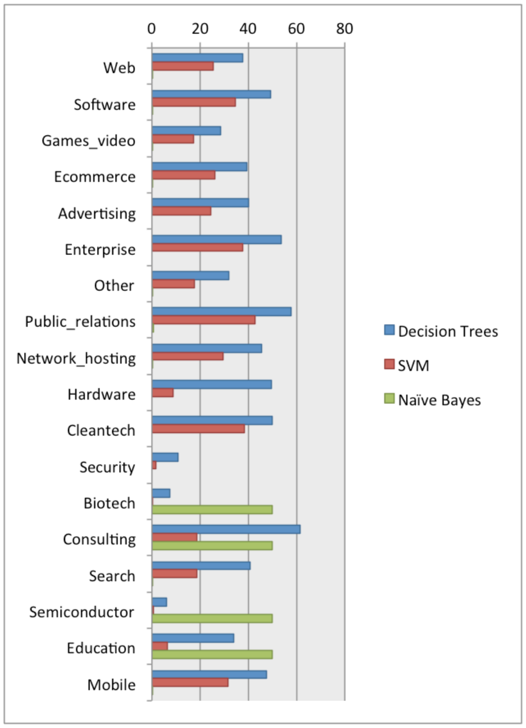

Figure 6. False Positive Rates by Categories.

5.3 Soundness of Social Network Features as Investment Behavior Indicators

5.3.1 General Performance

As shown in the sections 5.1 and 5.2, SVM and Decision Trees performed above baseline levels in general, while Naïve Bayes failed to meet the baseline performance for AUC and TPR.

The results for SVM and Decision Trees are generally encouraging: on an aggregate level, both learning algorithms produced an AUC of over 0.77 and TPR of over 60%. In addition to aggregate performance, the strategy also performed well in terms of individual industry categories: most industries across all three learning algorithms performed above the baseline performance of 60% and 0.6 respectively, with most TPRs ranging from 56% to 91% and AUC ranging from 0.63 to 0.80.

Naïve Bayes faired the worse with all of its aggregate and categorical experiments failing to achieve the baseline TPR and AUC, although it's FPR were well below 40% for most categories apart from biotech, education, consulting and semiconductor.

This infers that given suitable learning algorithms, using social network features generally provide consistent performance across different learning algorithms not only in terms of aggregate results, but also in terms of industries. More importantly, given the richness, diversity of the dataset and that it's above baseline performance, the prediction model reasonably used to predict *Investors*' behavior.

5.3.2 Differences in Performance

While the general performance is above baseline, we noticed differences in performance in terms of both TPR and FPR especially between Naïve Bayes versus SVM/Decision Tree algorithms. The Naïve Bayes learning algorithm generally produced a lower TPR as compared to SVM and Decision Trees. The reasons are as follows:

5.3.2.1 *Suitability of Learning Algorithms*

Predicting investment behavior is a highly complex problem and we understand that there are more factors than what is being discussed and implemented in our research work. More importantly, the problem is a non-linear one: intuitively, we know that investors do not make investment decisions based on a single factor but rather on a plethora of factors. At the same time, these factors may or may not be independent.

On the other hand, the underlying probability model of certain learning algorithms such as Naïve Bayes's is an independent feature model, thus not reflecting the true nature of the problem we are dealing with. Hence it is expected that Naïve Bayes learning algorithm had lower TPR as compared to the experiment results of the Decision Tree and SVM experiments.

Similarly, the Decision Tree learning algorithm and SVM reflects more accurately on investor behavior. For instance, investors often start seeking out companies that fit one or more factors such as having a certain threshold users, or a certain team make-up.

What we can deduce here is that Naïve Bayes is not a suitable learning algorithm for our problem while SVM and Decision Trees better reflect our problem.

5.3.2.2 *Differences in Number of Samples*

We noticed that the available samples varied widely amongst different categories, thus resulting in a wide range of performance between categories across different learning algorithms.

For instance, the number of true examples for the category "web" is 1600, while there are only 260 true examples for the "enterprise" category. Moreover, each category may or may not exhibit similar characteristics as compared to the aggregate data. Figure 7 shows the count of true examples for each category.

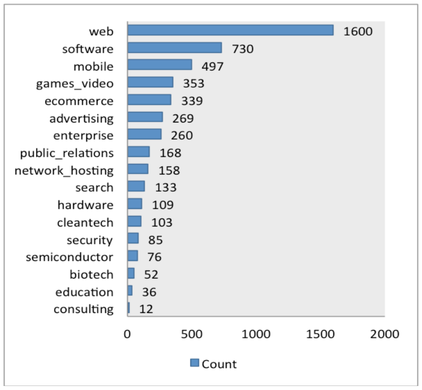

Figure 7. Number of Examples for Each Category.

5.3.3 *Other Interesting Findings*

We performed descriptive mining on the social network features and our results were similar to [1]. Since we had a new feature: number of shortest paths between an *Investor* and a *Company*, we decided to perform descriptive mining to uncover interesting trends related to investment behavior.

5.3.3.1 *More Shortest Paths is Correlated with Less Investment*

We aggregated the number of shortest paths for each pair of Investors and Companies in G_{Social} and took note of the score for pairs where investment occurred. We than plotted a best fit line and noticed that as the number of shortest paths connecting Investors and Companies increases, investment activities decreases. This is shown in Figure 8: the Y-axis represents the occurrences of shortest paths of a certain number between an *Investor* and a *Company*, while the X-axis represents the number of shortest paths between an *Investor* and a *Company*.

While this may seem counter intuitive, this makes sense if we consider competitive relationships between Investors and the fact that Investors are more likely to make investments if they are of less hops (closer) to the companies: within these paths, there exists 1 or more alternate investors; this may result in increased competition for the Investor. Similarly, since there are alternate investors within these paths, they are in fact closer to the company in question.

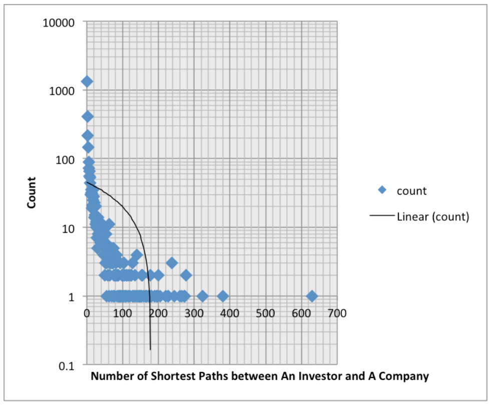

Figure 8. Relationship between Number of Shortest Paths and Occurrences of Investments.

5.3.3.2 *The Decision Making Process*

An important aspect of our work is to help startups or companies seeking investment better understand the investment process. Decision Trees can be readily visualized and we noticed that common neighbors and the length of shortest paths appears to play an important part of the decision making process.

Investors appears to be more avid in making investments when there are less than 3 to 4 common neighbors, and have a tendency to make investments when the length of shortest paths is less than 6 or 7 hops.

The results reflect the findings of [1], where investments are less likely to occur due to possible increased competition from similar companies. Similarly, the smaller the number of hops between an Investor and a Company, the "closer" their relationship.

6. CONCLUSION

In this research, we modeled investment behavior as a link prediction problem based on social network features and have obtained above baseline results across various learning methods and evaluation metrics. We discovered the following:

6.1 Implications for Startups and Companies seeking Investment

Our experiment results show that it is possible to predict investment behavior based on social relationships. Startups or companies seeking investments should take into consideration of their social relationships with a prospective investor.

6.2 Social features are reasonable features for predicting investment behavior.

We used social features based on Shortest Path, Common Neighbors, Jaccard Coefficient, Preferential Attachment and Adamic/Adar and other node-wise features to compute social similarity between a pair of *Investor* and *Company* and discovered that these features can be used to predict investment behavior. Not only can social information can be used to predict investment behavior, it is also a reliable and sound strategy to predict investment behavior: our prediction strategy based on social features and modeling it as a link prediction problem works well generally across the most common learning algorithm including Decision Tree, Naïve Bayes and SVM. Not only it performs well in terms of aggregate performance, it also performs well in terms of individual industries.

6.3 Multiple link predictors can be used to gain deeper and broader insight to the network.

We believe that we obtained good performance due to combining multiple link predictors as our learning feature: since each link predictors such as Shortest Path or Common Neighbor measures different aspects of a social network, combining multiple link predictors will allows us to gain a deeper and broader insight of a network. In our case, companies seeking investment can use multiple social indicators to gain a deeper understanding of their potential investors.

We hope that our work can help companies better understand how and when investors invest, thus helping companies be better prepared when they are attempting to seek external investment. We also hope that our work provides a fresh look as to what factors drives investment behavior. Most importantly, we would like to encourage companies to focus on social relationships in addition to other factors when seeking external investments, as investors are social animals.

7. REFERENCES

[1] Yuxian Eugene Liang, Daphne Yuan Sor-Tsyr. “Where’s the Money? The Social Behavior of Investors in Facebook’s Small World.” 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[2] David Liben-Nowell, Jon Kleinberg. “The Link Prediction Problem for Social Networks.” Journal of the American Society for Information Science and Technology, 58(7):1019 - 1031, May 2007.

[3] Lada A Adamic and Eytan Ada.” Friends and Neighbors on the Web.”

[4] Amir Barnea, Henrik Cronqvist and Stephan Siegel. “Nature or Nurture: what determines investor behavior ?”. Journal of Financial Economics 98 (2010) 583–604

[5] Tan, W.-K., Tan, Y.-J. “An exploratory investigation of the investment information search behavior of individual domestic investors.” Telemat. Informat. (2011), doi:10.1016/j.tele.2011.09.002

[6] James S.Doran, David R.Peterson, Colby Wright. “Confidence opinions of market efficiency and investment behavior of finance professors.” Journal of Financial Markets13 (2010) 174–195.

[7] L. Bakker, W. Hare,, H. Khosravi and B. Ramadanovic. “A social network model of investment behavior in the stock market.” Physica A 389 (2010) 12231229.

[8] Pierre Giot, Ulrich Hege, Armin Schwienbaher. “Expertise of Reputation? The Investment Behavior of Novice and Experienced Private Equity Funds.”

[9] Mark Grinblatt and Matti Keloharju. “The Investment Behavior and performance of various investor types: a study of Finland’s unque data set.”

[10] Friedkin, Noah. “Horizon of Observability and Limits of Informal Control in Organizations”, Social Forces 1981.

[11] J. Leskovec, D. Huttenlocher, J. Kleinberg. “Predicting Positive and Negative Links in Online Social Networks”. ACM WWW International conference on World Wide Web (WWW), 2010

[12] Guang Xiang, Zeyu Zheng, Miaomiao Wen, Jason Hong, Carolyn Rose, and Chao Liu. “A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch.” ICWSM'12, 2012.

[13] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna. “ Four Degrees of Separation”

[14] Mark S. Granovetter. “The Strength of Weak Ties”, American Journal of Sociology, Volume 78, Issue 6 (May, 1973)

[15] M. E. J Newman, “Clustering and preferential attachment in growing networks”, Physical Review E, Volume 64, Issue 2, 11 April 2001

[16] Miller McPherson, Lynn Smith-Lovin, James M Cook, “Birds of a Feather: Homophily in Social Networks”, Annual Review of Sociology, Volume 27: 415-444, August 2001

[17] Jeffrey Travers and Stanley Milgram, “An experimental study of the small world problem”, Sociometry, Volume 32 No 4 December 1969, Page 425 – 443

[18] L. Backstrom, J. Leskovec ,”Supervised Random Walks: Predicting and Recommending Links in Social Networks”, ACM International Conference on Web Search and Data Mining (WSDM), 2011.

[19] J. Leskovec, D. Huttenlocher, J. Kleinberg , “Signed Networks in Social Media” by. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), 2010.

[20] Larry Page and Sergey Brin, “The Anatomy of a large-scale hypertextual web search engine”, Proceedings of the Seventh International Conference on World Wide Web.

[21] Jon M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.

[22] Michael Fire, Lena Tenenboim, Ofrit Lesser, Rami Puzis, Lior Rokach and Yuval Elovici, “Link Prediction in Social Networks using Computationally Efficient Topological Features”, SocialCom 2011.