Received 14 September 2014
Revised 30 September 2014
21 December 2014
7 January 2015
19 January 2015
Accepted 19 January 2015

Predicting investor funding behavior using crunchbase social network features

Yuxian Eugene Liang and Soe-Tsyr Daphne Yuan
Department of Management Information Systems, National Chengchi University, Taipei, Taiwan

Abstract

Purpose – What makes investors tick? Largely counter-intuitive compared to the findings of most past research, this study explores the possibility that funding investors invest in companies based on social relationships, which could be positive or negative, similar or dissimilar. The purpose of this paper is to build a social network graph using data from CrunchBase, the largest public database with profiles about companies. The authors combine social network analysis with the study of investing behavior in order to explore how similarity between investors and companies affects investing behavior through social network analysis.

Design/methodology/approach – This study crawls and analyzes data from CrunchBase and builds a social network graph which includes people, companies, social links and funding investment links. The problem is then formalized as a link (or relationship) prediction task in a social network to model and predict (across various machine learning methods and evaluation metrics) whether an investor will create a link to a company in the social network. Various link prediction techniques such as common neighbors, shortest path, Jaccard Coefficient and others are integrated to provide a holistic view of a social network and provide useful insights as to how a pair of nodes may be related (i.e., whether the investor will invest in the particular company at a time) within the social network.

Findings – This study finds that funding investors are more likely to invest in a particular company if they have a stronger social relationship in terms of closeness, be it direct or indirect. At the same time, if investors and companies share too many common neighbors, investors are less likely to invest in such companies.

Originality/value – The author's study is among the first to use data from the largest public company profile database of CrunchBase as a social network for research purposes. The author's also identify certain social relationship factors that can help prescribe the investor funding behavior. Authors prediction strategy based on these factors and modeling it as a link prediction problem generally works well across the most prominent learning algorithms and perform well in terms of aggregate performance as well as individual industries. In other words, this study would like to encourage companies to focus on social relationship factors in addition to other factors when seeking external funding investments.

Keywords Social network analysis, CrunchBase, Investor funding behavior, Link prediction

Paper type Research paper

1. Introduction

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

Many studies have attempted to understand investment behavior in general. Factors such as psychological and geographic differences, funding investment experiences and even genetics have been proposed as factors that spur investment.

Internet Research
Vol. 26 No. 1, 2016
pp. 74-100

1066-2243
DOI 10.1108/IntR-09-2014-0231

Funding behavior refers to the financing funding that typically have different funding round names[1] (e.g. seed rounds, angel rounds) related to the class of stock being sold and during which stage of the company when it sells stocks to investors.

Few research consider the role of social relationships in the funding investment behavior between investors and companies. Shane and Cable (2002) used fieldwork to provide some theoretical evidences that social ties could help overcome the information asymmetry problem of future profitable opportunities between entrepreneurs and seed-stage's investors and influence the investors' decisions on the selection of ventures to fund; the social ties here refer to direct/indirect personal relationships. This is because social ties help bring the expectation of trust and reciprocity into information exchanges and activate possible cooperative economic exchanges (Uzzi, 1996). In this study, our aim is to further identify the social relationship factors that can help prescribe the investor funding behavior. However, this paper does not address the issues of venture capital investment activities (Burchardt *et al.*, 2014) (e.g. selection, appraisal, contracting, monitoring and exiting of target companies).

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

This study accordingly explores predictions about funding investment behavior as a link prediction problem. We build a social network graph using data from CrunchBase, the largest public database with profiles about companies. Using this data set, we attempt to investigate if an investor will invest in a company based on certain identified factors of social relationship. That is, this study combines social network analysis with the study of investing behavior. We explore how similarity between investors and companies affects investing behavior through social network analysis. Also, our study is among the first to use data from CrunchBase as a social network for research purposes.

The research problem of predicting investor funding behavior through social network analysis is explored by combining multiple link prediction features to gain greater insight about 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 techniques can provide us with a holistic view of a social network. Our social network analysis also provides insights about how investors invest within a social network.

This paper is organized as follows. Section 2 reviews related literature. Section 3 focusses on the presentation of how we intend to solve the problem of predicting investor funding investment behavior by modeling it based on the classic link prediction problem. Section 4 then perform experiments and evaluates the effectiveness of our methods. In Section 5, we focus on the discussion of the use of social network features and our prediction method's performance, followed by discussion of the implications for research and practice. Finally, we summarize and conclude our study by outlining the important findings and various possibilities for extending the research presented here in Section 6.

2. Related work

The section discusses works on investment behaviors in general and works using link prediction to predict investment behavior.

2.1 Previous research on investment behaviors

Prior studies on investment behaviors in general can be categorized into six categories based on the type of factors that drive investment behaviors. The followings exemplify some existing studies:

- Personal opinions: Doran *et al.* (2010) studied the role of the personal opinions of finance professors on the efficiency of the stock market in the USA and found out that personal opinions do not affect investment behaviors. Rather, investment behaviors of financial professors are largely driven by the same behavioral factor that drive amateur investors.
- Funding investment experience: Giot *et al.* (2012) analyzed the differences in the investment behaviors among experienced and novice private equity firms and found out that novice firms tend to invest more slowly than experienced firms but the size and value of the funding of novice firms tends to be larger.
- Geographic identities: Grinblatt and Keloharju (2000) discovered that investment behaviors can be determined by the geographic identity of investors. For instance, 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.
- Online vs offline communities: Tan and Tan (2012) explored the roles played by online and offline communities and discovered that offline communities have greater influence on investing behaviors. This is expected since offline communities involve offline interaction which is, hence, more likely to take place in person, thus increasing the level of influence.
- Psychology: Bakker *et al.* (2010) investigated psychological factors that impact market evaluation and found out that trust and social influence affect the stability of investment markets.
- Genetics: Barnea *et al.* (2010) 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 the effect is not long lasting.

It is rare to see literatures investigating the role of social relationships between investors and companies excluding family firm investments (Bianco *et al.*, 2013). As mentioned in Section 1, Shane and Cable (2002) provided some fieldwork evidences about social ties of personal relationships influencing investors' funding decisions on the selection of ventures. For the recent trend of crowdfunding community like Kickstarter, Kuppuswamy and Bayus (2013) found that potential backers do not contribute to a project already receiving a lot of support based on the assumption of the others already providing the necessary funding; this collective phenomenon of the crowdsourcing investment behavior is generally different from the investor funding behavior by it not focussing on the business return.

Other relevant literatures include the topics such as how a company's social network position can affect its business performance (Grewal *et al.*, 2006; Vir Singh *et al.*, 2011;

Seaman *et al.*, 2014; Pahnke *et al.*, 2014), how to identify financially successful companies (Martens *et al.*, 2011), how different funding investments methods engender different kinds of company culture (Hamilton, 2001), etc. However, this study mainly focusses on how the investors can invest within a social network in order to further the study of Shane and Cable (2002) that showed social ties influence investors' funding decisions.

2.2 Previous research on link prediction

Common social network analysis topics and relevant techniques and applications include, but are not limited to, centrality analysis (Leskovec *et al.*, 2010), community detection (Girvan and Newman, 2010; Newman, 2006; Leskovec *et al.*, 2007; Bliss *et al.*, 2014), link label prediction (Gallagher *et al.*, 2008; Kajdanowicz *et al.*, 2010), information diffusion (Leskovec *et al.*, 2007; Kempe *et al.*, 2003; Fan *et al.*, 2013) and team formation (Lappas *et al.*, 2009; Kargar and An, 2011). Other related works include statistical features of networks (Liben-Nowell and Kleinberg, 2007; Newman, 2001b) such as information networks, collaboration networks, biological networks and social networks. The similarity between the above applications is that the use of social network analysis techniques often improve the performance of the solution for the given problem domain.

Link prediction is one of the most important topics in social network analysis. Link prediction seeks to predict changes in terms of edges or nodes of social networks over time. This type of prediction can be problematic in social networks. Liben-Nowell and Kleinberg (2007) performed extensive studies on link prediction in social networks and noted that there is no singular technique that can ensure the best performance. In fact, the techniques have shown limited performance. The techniques used for link prediction include PageRank (Page and Brin, 1998), HITS (Kleinberg, 1998), Adamic/Adar (Adamic and Adar, 2001), Jaccard Coefficient, shortest paths, etc. Moreover, Liben-Nowell and Kleinberg (2007) proposed that performance may be improved by taking node-specific information into account. More recently, link prediction has been applied to data sets in popular social networks, which include Twitter, Facebook and others as covered by Leskovec *et al.* (2007, 2010a, b) and Fire *et al.* (2011). These studies include the prediction of positive and negative links recommended to friends on Facebook by using computationally efficient topologic features.

The novelty of this study is the use of social relationships (represented by social network features) as the main way to predict whether or not funding investments will occur. For example, we attempt to predict if an investor will invest in a 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.

We opt to use link prediction rather than other social network analysis methods as a way to model investor behavior for the following reasons:

- We find that link prediction suits our problem as it seeks to predict new links within a social network as time progresses. This is very similar to how investors and start-up investors operate: as time progresses, will new links (investments) occur between different pairs of investors and companies? Link prediction usually focusses on the addition of links and does not take into account of removal of links, which suits our problem perfectly as we rarely see investors withdraw investment after an investment is made into a company.

- Link prediction allows us to input different characteristics of individual entities, which also reflects the reality of investment behaviors and transactions. Investors and companies both reflect different characteristics in terms of relationships and node information, both of which can be readily reflected using network structures (such as "closeness" using shortest paths and similarity using the Jaccard Coefficient) and investor/company information such as age and industries.
- Prediction models that use only a singular metric (such as common neighbors only) yield less than satisfactory results (Liben-Nowell and Kleinberg, 2007). By taking into account different metrics (shortest paths, Jaccard Coefficient, common neighbors, Adamic/Adar, preferential attachment and the number of shortest paths), we can derive a more complete perspective of the network we are dealing with.

3. Method of predicting investor funding behavior

This study models funding 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 they are in terms of their social relationship within a social network graph that is attained by crawling and analyzing data from CrunchBase. Kalampokis *et al.* (2013) addressed social network data analysis for prediction would involve the steps of collection and filtering raw data, computation of predictor variables, creation of predictive model and evaluation of the predictive performance. The following sections will then provide these step details.

3.1 The CrunchBase data set and social network

CrunchBase (www.crunchbase.com) is an open data set 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. The CrunchBase data set represents a rich multi-modal social network of investors and companies. For instance, each company shows a list of people who currently (or previously worked) for a company; drilling further we get to see the person's profile which states the list of companies (or financial organizations) which he/she is involved in.

Xiang *et al.* (2012) performed studies using the CrunchBase data set and predicted company acquisitions with factual and topic features using profiles and news articles on TechCrunch. Although they made use of a similar data set as our work, their work did not make use of social relations as part of their feature set and focussed on a different domain of mergers and acquisitions. In particular, they made use of node information, such as company age, the number of financing rounds and categories as well as news articles related to mergers and acquisitions to build machine learning features.

In this study, we make use of social relationships, represented by social network features, to predict acts of investment; mergers and acquisitions are not covered in this work. We gathered data related to companies, persons and financial organization. We chose Facebook as the seed node, and gathered people, companies and financial organizations found within social and funding investment relationships within four degrees of separation from Facebook. We selected Facebook as the seed node due to its meteoric rise in the social network industry and much hyped recent IPO. We chose four degrees of separation as a cutoff point as opposed to six degrees of separation (i.e., any two people are distanced by at most six friendship links) due to the fact that Backstrom *et al.* (2011a, b) showed recent advances in technology have reduced the degrees of separation between people and the world is even smaller than we expected (only about

four degrees of separation). In addition, there are limits to the "Horizon of Observability" (Friedkin, 1983) from the viewpoint of using Facebook as a seed node. Given our assumption that CrunchBase is a social network, a data set of up to four degrees of separation given any seed node is representative of CrunchBase. Our final data set contains 11,916 companies, 12,127 people and 1,122 financial organizations within four degrees of separation from Facebook. The entity and relationship types provided by CrunchBase are as follows:

(1) Entities types:

- People/person: people (person) refers to founders, executives and other persons working for a particular company or organization. Examples from our data set 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 funding investment in companies. Prominent examples in our data set include Accel Partners (which has invested in prominent companies such as Facebook, Dropbox, Groupon, Angry Birds and so on) and Digital Sky Technologies (an international investment firm which focusses solely on the internet sector and has also invested in Facebook and Groupon).
- Investor: investors consist of people, companies and financial organizations. This is due to the duality of roles played by people, companies and financial organizations in the CrunchBase data set. For example, companies like Microsoft play the role of a company yet performed the act of investment on other companies such as Facebook in the early days. Similarly, Peter Thiel is a person entity, yet he also invested in Facebook.

(2) Relationship types:

- Social: we define social 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) are recruited by a company or want to work for that particular company or financial organization in question, social relations are undirected. For instance, Bret Taylor has a social relationship with Google and Facebook since he has previously worked for both companies.
- Investment: funding investment relationships are created as a result of an act of funding investment by a person, company and/or a financial organization in a company. For example, Microsoft invested in Facebook, thus resulting in a funding investment relationship.

(3) CrunchBase social network:

Using the data set from CrunchBase, we build a CrunchBase social network based on the entity and relationship types, where nodes represent entities, while relationships represent edges. Since we are interested in the prediction of funding investment acts, we further simplify the network into only two node

types, i.e., a two-mode network (Latapy *et al.*, 2008) with two different set of nodes as the investors and the companies:

- Investors: investors are made up of people/person, financial organizations and companies. Note that companies can play the role of an investor; 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. We can also categorize the data set into two types of networks (social network and funding investment):
  - G_{Social}: G_{Social} is a social graph that is undirected and derived from social relationships. The edges represent social relationships only. In this study, social relationship is defined 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, we define social relations as undirected. Most importantly, we do not include the act of investment as part of social relations as investment behavior is what we are attempting to predict. Here, nodes represent investors or companies . a link is formed when there is an employment relationship between nodes.
  - G_{Investor}: G_{Investor} is an funding investment graph that is directed and derived from investor relationships. This means that the edges are made up of investor relationships only. Nodes represent investors and companies. A link is formed when an investor invests in a particular company.

Our final data set consists of G_{Social}, where edges represent social relationships and the nodes include investors and companies. G_{Investor} is used to provide ground truth labels. Using our final data set, we discover 5,341 investment activities. We define such investment activity as an instance when an investor invests in a company. For example, when an investment round occurs with three investors investing in a company, we take it that three investment activities have been discovered.

We also attempt to make a visualization of a subset of the diagram and we see that certain investors cluster companies that they have invested in based on their social relationships. In Figure 1, blue vertices represent financial organizations, green vertices represent companies and red vertices represent people. Blue edges represent investment relationships while black edges represent social relationships. Visually, we can see clusters of social and investment relationships, i.e., social relationships are sometimes present where investment relationships are present.

3.2 Modeling investment behavior as a link prediction problem

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

Meanwhile, investors consists of people, companies and financial organizations. This is due to the duality of roles played by people, companies and financial organizations in the CrunchBase data set. For example, Microsoft played a dual role of company and financial organization when it invested in Facebook.

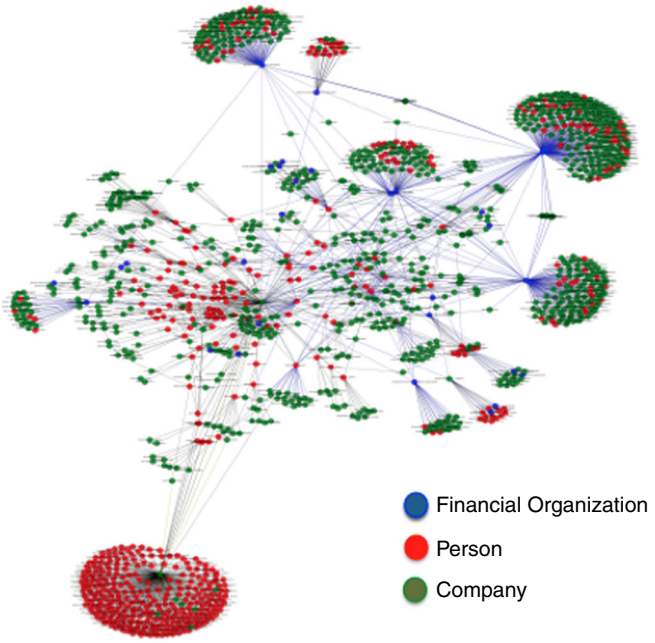

Figure 1. Network visualization of companies, people and financial organizations

3.3 Modeling social relationships

In order to determine the social similarity between an investor and a company, we use features based on node neighborhoods, 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 (named social features in this study). Similarity refers to the numerical measures of how alike two data objects are, and the selected social features are adapted from graph theories and social network analysis according to a former investigation (Liang and Yuan, 2012).

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 between 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 $\overline{\Gamma}(x)$. The descriptions of the selected social features are as follows.

3.3.1 Shortest paths. We simply consider the shortest path between investors and companies. General intuition about the shortest path in this context suggests that investors are more likely to invest in companies that are found within their "small world," in that investors and companies are related through short chains (Liben-Nowell and Kleinberg, 2007; McPherson *et al.*, 2001). We define score (x, y) to be the length of the shortest path between an investor and a company. It is hypothesized that the shorter the shortest path, the more likely that the investor will invest in that company (Liang and Yuan, 2012). The reasoning behind this is that the investor is "closer" to the company and hence it is much easier for them to reach each other.

3.3.2 Adamic/Adar. Adamic and Adar (2001) considered the similarity between two personal homepages by computing features of the pages and defining the similarity between two pages to be:

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

Where we consider the similarity feature, z to be the common neighbors, while x represents investor's features and y represents a company's features. For our purposes, we consider the similarity feature to be the common neighbors. Adamic/Adar weighed rarer features more heavily. The intuition of Adamic/Adar in our context is that investors are more likely to invest in companies that have greater similarity (Liang and Yuan, 2012).

3.3.3 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:

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

3.3.4 Common neighbors. Common neighbors are considered as the most direct implementation. According to Newman (2001b), 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 a greater number of common neighbors based on the previously justified reasons (Liang and Yuan, 2012). The score(x,y) for common neighbors is defined as the number of common neighbors of x and y:

$$|\Gamma(x) \cap \Gamma(y)| \tag{3}$$

3.3.5 Preferential attachment. Preferential attachment (Newman, 2001a) suggests that the probability that a new edge has node x as an endpoint is proportional to the current number of neighbors of x (Liben-Nowell and Kleinberg, 2007). This models the "rich get richer" phenomena where companies which already received investments should receive even more investments as time progresses. Results from Liang and Yuan (2012) suggest that investors are less likely to invest in companies with higher preferential attachment. When a company becomes more popular in terms of receiving investments, such companies are more likely to receive preferential treatment from investors. The score(x,y) for preferential attachment is defined as that the probability of collaboration of x and y being correlated with the product of the number of neighbors of x and y:

$$|\Gamma(x)| \bullet |\Gamma(y)| \tag{4}$$

3.3.6 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 among 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 can be more easily reached via multiple shortest paths.

Liang and Yuan (2012) have shown these selected social features have the following features. Being "closer" in terms of shortest path and greater similarity in terms of Adamic/Adar score generally leads to investment. However, greater similarity in terms of common neighbors, Jaccard Coefficient and preferential attachment does not lead to investment in general. In fact, the greater the dissimilarity in terms of common neighbors, Jaccard Coefficient and preferential attachment, the greater the chances of funding. Using these features, we can then formulate the funding investment prediction problem with some learning algorithms described in the next section.

3.4 Funding investment prediction learning algorithms

In this study, we chose three machine learning algorithms: Decision Tree (based on CART algorithm) (Breiman *et al.*, 1984), SVM (with rbf as the kernel) (Chang and Lin, 2012) and Naïve Bayes (Bernoulli Model) (Manning *et al.*, 2008). This is to make sure that social network features can indeed be used as reliable indicators for predicting investments.

We chose CART Decision Tree (Breiman *et al.*, 1984) algorithm as our primary learning method that can construct a classification and regression decision tree from class-labeled training tuples. This is because we want a model which is simple to understand so that companies seeking funding can gain a better understanding of an investor's behavior. More importantly, a model learnt using Decision Tree can be readily visualized; such information can be very useful for companies in gauging their chances of receiving funding investment from a particular investor.

To make sure that social features can indeed be used as reliable indicators for predicting investments, we also use SVM with rbf as the kernel (Chang and Lin, 2012) and Naïve Bayes with the Bernoulli Model (Manning *et al.*, 2008) as alternate learning algorithms (widely regarded as prominent classical supervised learning algorithms for classification and regression problems) as a way to cross validate our proposition. More importantly, we select 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.

3.5 Characteristics of the method

Our method presents several characteristics in light of previous related works in terms of the data set used, problem formulation/predictive model and the introduction of new factors for predicting funding investment behavior.

3.5.1 Richness and size of data set. We use a data set from CrunchBase and the size of our network consists of 11,916 companies, 12,127 people and 1,122 financial organizations within four degrees of separation from Facebook. This means that we have a total of 25,165 unique nodes in the network. In addition, our data set consists of very different entities, which include people, companies and financial organizations. These entities also consist of various demographic groups. These factors make our data set richer and larger as compared to previous works. For example, Grinblatt and Keloharju (2000) made use of financial data to predominantly focus on investments in Finland only. As another example, Doran *et al.* (2010) focussed their data on only 96 Taiwanese adults. Similarly Bakker *et al.* (2010) focussed on finance professors exclusively.

3.5.2 Problem formulation and predictive model. While previous work has shown merits, there has been a lack of generalizability in previous approaches. This might

be due to how the problem of predicting funding investment behavior is formulated; in our approach, we choose to model funding investment behavior as a classic link prediction problem. This allows us to build a model in which funding investment behavior can be predicted.

3.5.3 Social network features as a factor for predicting investment. Most previous work has focussed on financial data, psychology, experience, etc., as factors for predicting investments. We 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 of prediction.

4. Evaluations

Since we model our research problem as a link prediction problem, and more specifically, use machine learning techniques, we use a standard performance metrics for many binary classification tasks, the true positive rate and false positive rate (FPR). We also use the accuracy measured by the area under the ROC curve (AUC) (Cortes and Mohri, 2003) as our evaluation metric, which represents the trade-off between the true positive and false positive.

We also evaluate our results based on the performance metrics mentioned above on two levels, aggregate and industry. Aggregate evaluation is performed when all companies are taken into account regardless of their industry. On the industry level, companies within an industry are evaluated against the aggregate level to see if there are any wide differences in performance. Since CrunchBase indicates the industries ("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.

Using the above mentioned evaluation performance metrics and levels of evaluation, we compare the performance of these metrics across the three machine 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 funding investment behavior.

4.1 Experiment ground truth labels and baseline performance

Using our G_{Investor}, we discovered 5,341 funding investment activities. We define such funding investment activity as an investor investing in a company, and they are regarded as the ground truth labels. For example, when an funding investment round occurs with three investors investing in a company, we take it that three funding investment activities have been discovered.

We do not have prior results as a basis for comparison since to our knowledge, and no previous studies have modeled funding investment behaviors as a link prediction problem. In addition, the 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 the true positive rate (TPR – the fraction of true positives out of the total actual positives) baseline should be above 60 percent, and finally, the FPR (the fraction of false positives out of the total actual negatives) should be lower than 40 percent.

4.2 Experiment data split for training, testing and running

Experiments were performed using Amazon Web Services for data storage and partial computation, while PiCloud was used extensively to perform parallel computations related to calculation of network features. We made a 40 percent training data split for training, with the remaining data being used for testing purposes. This 40 percent training data split were heuristically derived from multiple data splits and experiment runs that provided stable results. This was applied to both aggregate and category experiments across the three learning algorithms.

The experiments can be summarized as follows:

- Data collection phase: data is first collected from CrunchBase using its APIs. Data is stored on Amazon EC2 via MongoDB for ease of data slicing.
- Data preparation phase: the data collected in the previous phase are checked or incorrectly structured JSON data. We select Facebook as the seed node, and collected entities are based on a network built using its social and investor relationships. Using the network above, we perform pairwise computation of each pair of investor and company to derive scores for shortest paths, Adamic/Adar, Jaccard Coefficient, common neighbors, preferential attachment and number of shortest paths. These scores are dumped into a MongoDB node hosted on Amazon EC2 for ease of manipulation.
- Computation phase: to reduce the monetary cost of computation, we offload computation to PiCloud, with data streaming via the MongoDB node on Amazon EC2 instances to PiCloud. In order to obtain stable results, we conduct experiments with different proportions of true and false data and use different types of sampling for the data. We run over 1,000 rounds of experiments (including categorical experiments) to derive the current split of 40 percent of the data for training with the remainder used as testing data. The results are stored as text files in Amazon S3.
- The results of the computation phase are then used to derive the ranking of investors or companies for use in the final IT artifact.

4.3 Experiment results

We run the experiment using Decision Trees, SVM and Naïve Bayes algorithms. The results are shown in the following subsections.

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

Figure 2 shows a summary of performance metrics based on AUC. The Decision Tree experiment produces an AUC of 0.77 while SVM produces an AUC of 0.79. Naïve Bayes produces an AUC of 0.77. All three learning algorithms perform better than the baseline performance in terms of aggregate results.

Figure 3 shows a summary of performance based on TPR. The TPR for Decision Tree is 87.53 percent, SVM registers an aggregate TPR of 89.6 percent, while Naïve Bayes has an aggregate TPR of 54.8 percent.

Figure 4 shows a summary of performance based on FPR. The FPR for Decision Tree is 33.18 percent, SVM registers an aggregate FPR of 33.38 percent, while Naïve Bayes has an aggregate FPR of 0.05 percent.

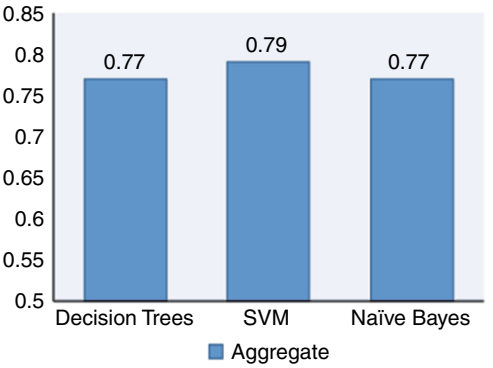

Figure 2.
Area under curve (aggregate)

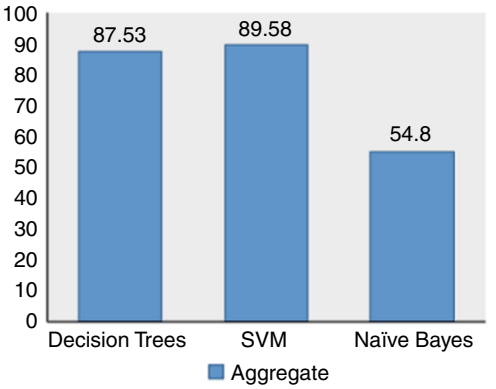

Figure 3.
True positive rate (aggregate)

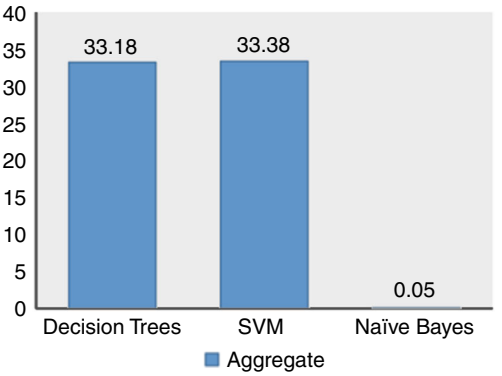

Figure 4.
False positive rate (aggregate)

Overall, SVM and Decision Tree perform above the baseline for both AUC and TPR, while Naïve Bayes performs under the baseline in terms of TPR. But, Naïve Bayes performs better than SVM and Decision Trees in terms of FPR.

4.3.2 Industry performance. We repeated the experiment by splitting the data by category with the same 40 percent training data split and also obtained reasonable performance.

For most of the categories, AUC hovers between 0.63 and almost 0.80 for Decision Trees. Their TPR ranges from 56 to 91 percent. Similarly, the AUC ranges from 0.65 to 0.84 and the TPR ranges from 51 to 91 percent for SVM. The AUC ranges from 0.75 to 0.78 and the TPR ranges from 52 to 57 percent for Naïve Bayes. The results are shown in Figures 5, 6 and 7.

We see Naïve Bayes performing better than Decision Trees and SVM in terms of FPR, but it underperforms in comparison to Decision Trees and SVM in terms of TPR. Interestingly enough, Naïve Bayes outperforms both Decision Trees and SVM in terms of AUC across various categories. On the whole, it appears that Naïve Bayes is generally weak in terms of TPR while it outperforms or is on par with the other algorithms when it comes to AUC and FPR. On the other hand, SVM and Decision Tree have above baseline performance when it comes to TPR and AUC. Since we are interested in predicting true outcomes, having a higher TPR with reasonable AUC and FPR is more important to us.

4.3.3 Summary of categorical performance. In terms of categorical performance, we notice that SVM generally performs well in different categories. Naïve Bayes shows unstable performance – while Naïve Bayes performs best for AUC in general and

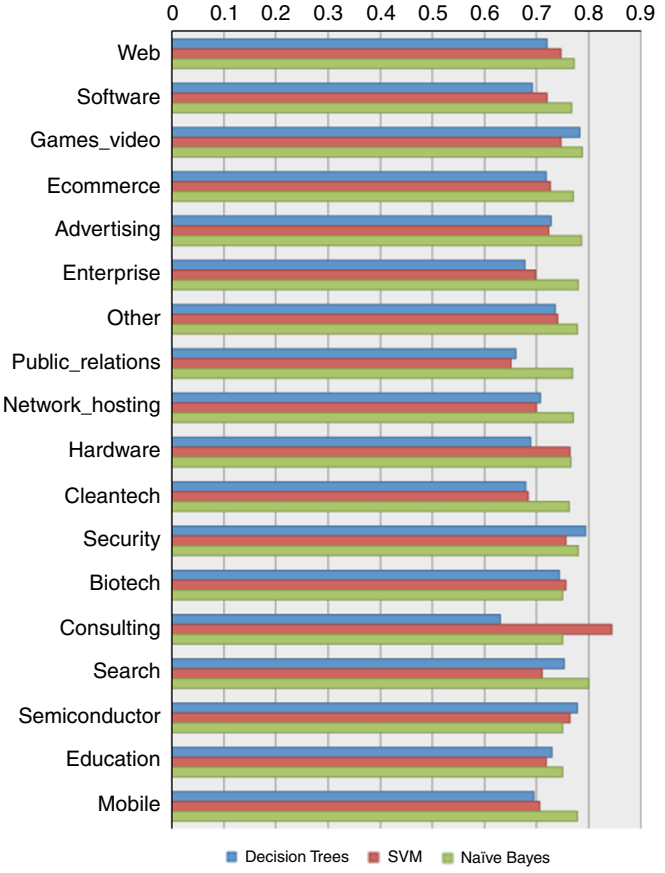

Figure 5.
Area undercurve by categories

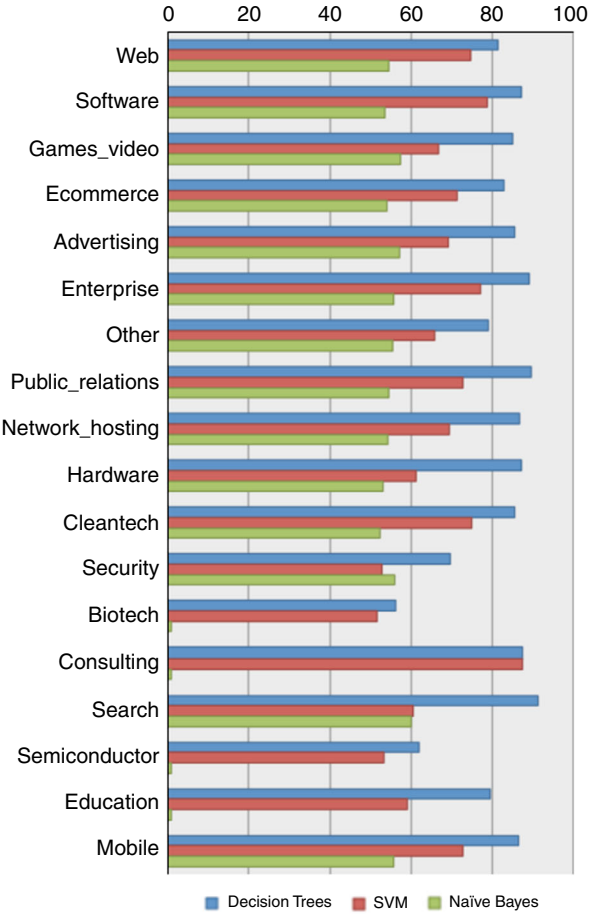

Figure 6.
True postive rates by categories

underperforms for most categories when it comes to TPR. On the other hand, Decision Trees performs best in TPR, but it has the worst performance in terms of FPR.

An important observation is that both SVM and Decision Trees generally perform within baseline levels, when there are enough training data. For instance, in the "web" category, we see SVM and Decision Tree performing above baseline in terms of TPR and performing below baseline level for FPR. However, Naïve Bayes underperforms by a wide margin when there are limited training data, which can be seen for the "consulting," "education," "biotech" and "search" categories.

The important implication is that the prediction model not only works on an aggregate level, it also generally works in terms of individual categories, with the exception of categories with limited data such as "security," "semiconductor," "education" and so on. Nonetheless, we see stable performances for Decision Trees and SVM across individual categories in terms of predicting true outcomes, hinting that the prediction model can work across aggregate and individual category levels.

4.3.4 Visualizing the decision process. An important aspect of this study is to understand the decision making process of investors. Decision Tree plays an important

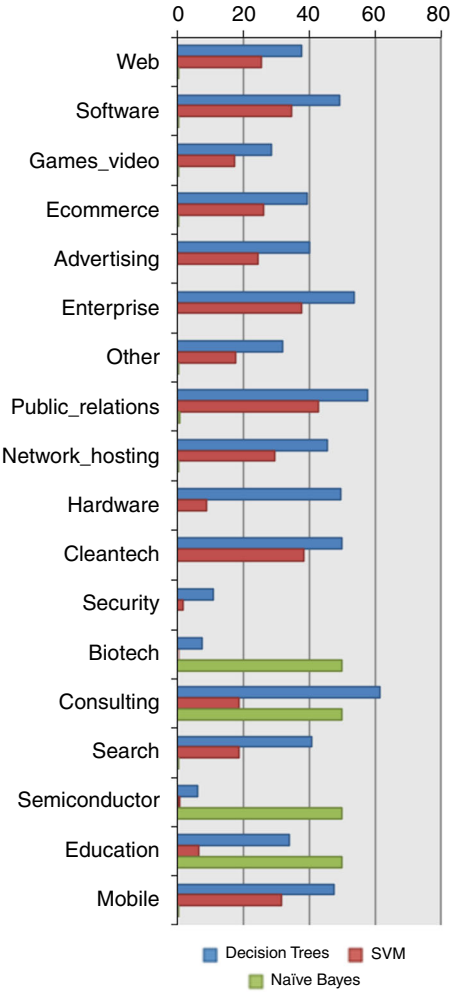

Figure 7.
False postive rates by categories

role in this respect since it is not only straightforward to understand, but it can also be visualized. Figure 8 shows a partial Decision Tree that visualizes the decision-making process.

Notice that the tree begins to split at the root using preferential attachment as a core factor of consideration, with shortest path playing an important role at the second level. Various splits typically revolve around preferential attachment and shortest paths; this echoes our general rules for investment which suggest that preferential attachment and shortest paths play an important role in investment behaviors.

5. Discussion

5.1 Soundness of social network features as funding investment behavior indicators

5.1.1 General performance. The main purpose of carrying out the experiments across multiple learning algorithms is to ensure the soundness of using social features as main

Figure 8.
Visualizing the decision process

predictors of investment behavior. This echoes what mentioned in Section 1 about our aim of identifying certain social relationship factors that can help prescribe the investor funding behavior. As shown in the previous subsections, Decision Trees and SVM perform above baseline levels in general, while Naïve Bayes fails to meet the baseline performance for AUC and TPR. 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 among 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.

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

- Suitability of Learning Algorithms: predicting funding investment behavior is a highly complex problem and we understand that there are more factors than

those which are discussed and implemented in this study. More importantly, the problem is a non-linear one. Intuitively, we know that investors do not make funding 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 is an independent feature model, thus not reflecting the true nature of the problem we are dealing with. Hence, the Naïve Bayes learning algorithm is expected to have lower TPR as compared to the experiment results of the Decision Tree and SVM experiments. Similarly, the Decision Tree learning algorithm and SVM reflect investor behavior more accurately. For instance, investors often start seeking out companies that fit one or more factors such as having a certain threshold of 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 Decision Trees and SVM better reflect our problem.

- Differences in Number of Samples: we notice that the available samples vary widely among different categories, thus resulting in a wide range of performances between categories across different learning algorithms. For instance, the number of true examples for the category "web" is 1,600, 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 9 shows the count of true examples for each category.

5.1.3 Other interesting findings. We perform additional descriptive mining based on the social feature of the number of shortest paths between an investor and a company. Adding to the findings of the previous work (Liang and Yuan, 2012), we uncover interesting trends related to funding investment behavior:

- More shortest paths is correlated with less investment: we aggregate the number of shortest paths for each investor and company pair in G_{Social} and take note of the

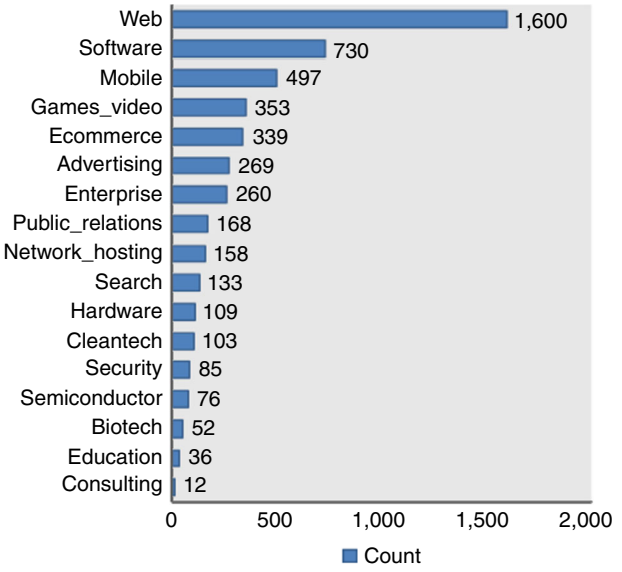

Figure 9.
Number of examples for each category

score for pairs where funding investment occurs. We then plot a best fit line which shows that as the number of shortest paths connecting investors and companies increases, funding investment activities decrease. This is shown in Figure 10. While this may seem counterintuitive, it makes sense if we consider the competitive relationships between investors and the fact that investors are more likely to make investments if there are of less hops (closer) to the companies. Within these paths, there exists one 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.

- The decision making process: an important aspect of our work is to help startups or companies seeking funding investment better understand the funding investment process. Decision Trees can be readily visualized and we notice that common neighbors and the length of shortest paths appear to play an important part of the decision making process (Figure 8). The results reflect the previous findings (Liang and Yuan, 2012) showing that investments are less likely to occur due to the possible of increased competition from similar companies. Similarly, the smaller the number of hops between an investor and a company, the "closer" their relationship.

5.1.4 Verification of prediction model. While the experiments in the previous subsections yielded satisfactory results, we go on to further verify the model by using a subset other than the Crunchbase data set. We want to verify if our experiment will work even if it is applied to a different data set, especially if the parties involved generally come from different cultures. This time round, we select RenRen.com as the seed node and repeat the data collection process of collecting all persons, financial organizations and companies. We select RenRen.com due to its status as China's Facebook. Thus, it can

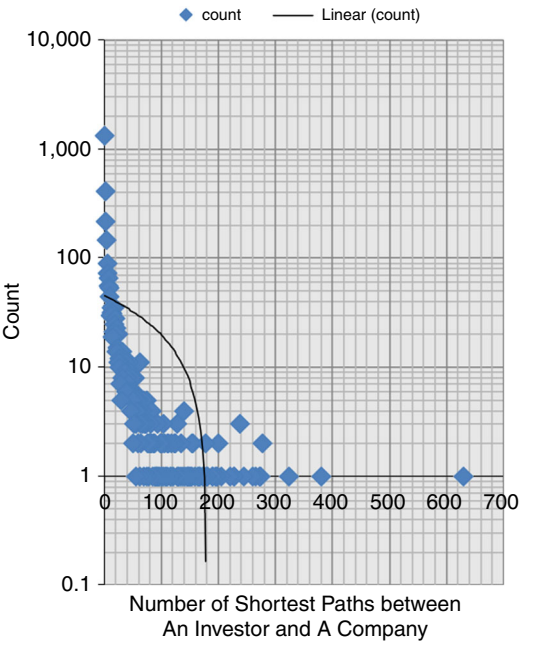

Figure 10.
Relationship between number of shortest paths and occurrences of investments

serve as a distinct data set in terms of seed node and other forms of entities. The network statistics based on RenRen are as such – we discovered 3,582 companies, 721 financial organizations and 3,386 persons within the small world of RenRen.

We keep the experimentation process in sync as much as possible, from the data collection process and number of hops to derive our final data set to the processing of data. Despite the different data set sizes using different seed node as Facebook and RenRen, the experiment results for the verification phase reveal a similar trend as compared to that of Facebook's. That is, we observe similar performances, especially in terms of aggregate performance, for AUC, TPR and FPR. We notice that the performances of Naïve Bayes become unstable or unpredictable in the face of limited data, which is evident when the Naïve Bayes is used to perform prediction for categories with limited data. On the other hand, Decision Tree and SVM show stable performance in general. Having stable performance, especially for Decision Tree, is a big deal for not only us, but also for investors and companies in general. This is because that we know that Decision Tree is relatively effective in its predictive power, and more importantly, we can visualize the decision making process.

5.1.5 Utility of our method and system. Our method explores how much similarity between investors and companies affects investing behavior through social network analysis. Moreover, this study is among the first to use data from CrunchBase as a social network for research purposes. Our method is also implemented into a service system that can allow users to explore relationships and search for investors/companies. For instance, when a user enters "Goldman" and "Facebook," he/she will first see the screen of a network diagram showing the connections between Goldman Sachs and Facebook (Figure 11). The user can also see the various scores such as shortest paths, number of shortest paths and so on.

Users can also search for a potential list of investors or companies, depending on their role they. If a user is viewing as a company, he/she is able to search for potential investors, as exemplified in Figure 12(a). However, if a user is viewing from an investor's point of view, he/she can see a list of companies to invest in, as exemplified in Figure 12(b). Users

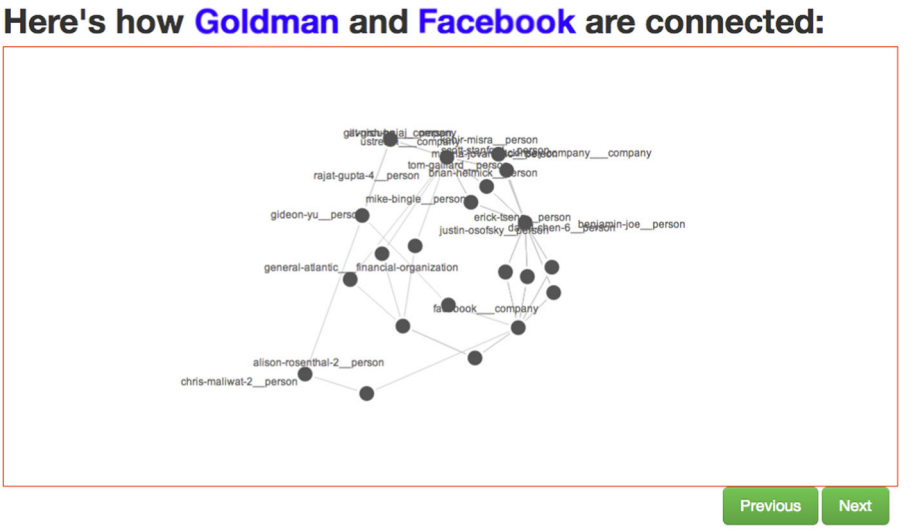

Figure 11.
A network diagram that shows the connection between Goldman Sachs and Facebook

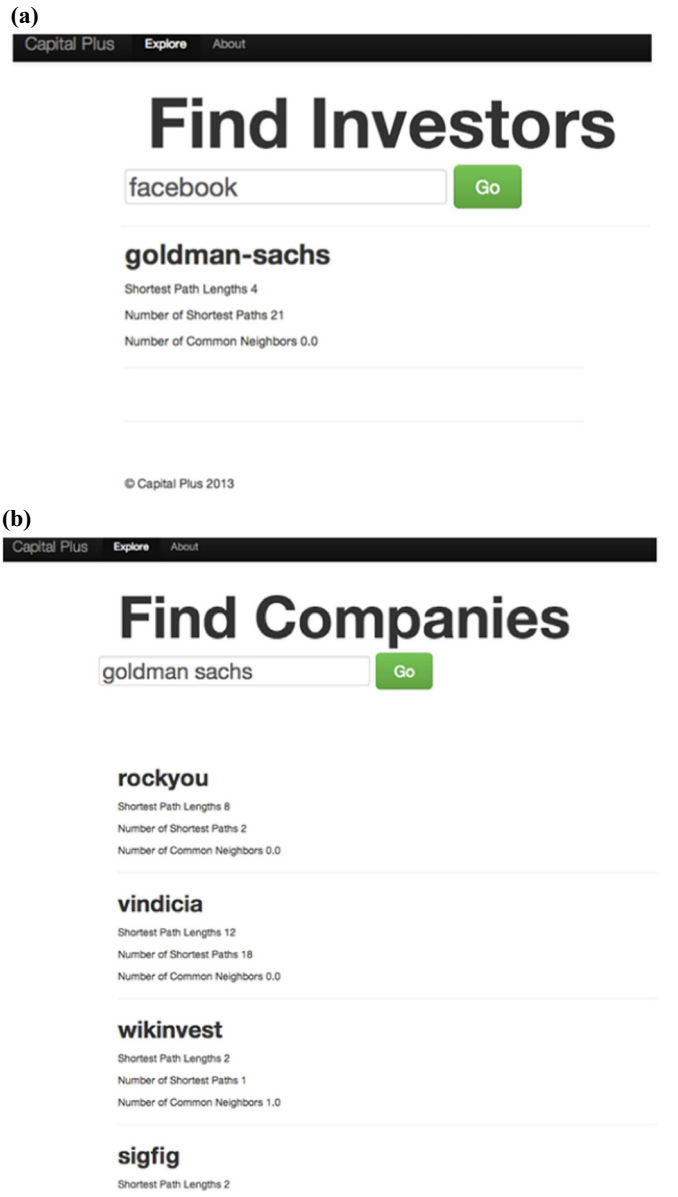

Figure 12.
Our method and system's exemplified runs

Notes: (a) Goldman Sachs is recommended if we are viewing from Facebook's point of view; (b) finding a list of companies to invest impersonating as Goldman Sachs

can visualize the relationships between investors and companies. For instance, Figure 13 shows the global relationships using the data set derived using Facebook as the seed node.

In short, in addition to the research contributions our method, the proposed system also demonstrates practical values for the real-world problem of funding

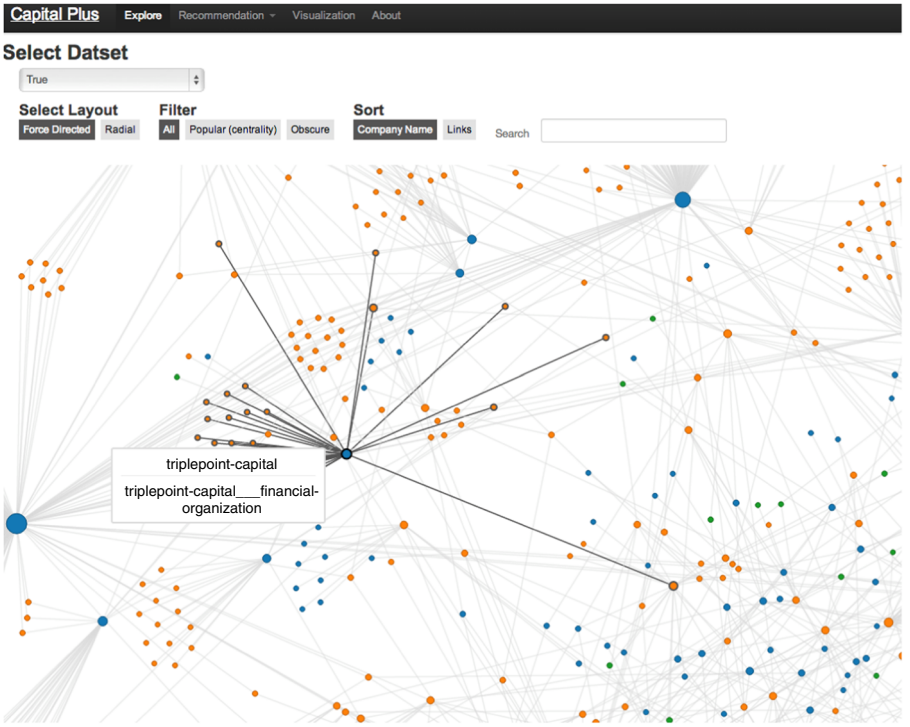

Figure 13.
Visualizing the connections between investors and companies

investment decisions. We hope that our work can help companies better understand how and when investors invest, and thus help companies be better prepared when they are attempting to seek external funding. We also hope that our work provides a fresh look at what factors drive investment behavior. Most importantly, we would like to encourage companies to focus on social relationships in addition to other factors when seeking external funding. The implemented service system derived from this research work will serve as a starting point for companies, individuals, financial organizations, governments or investors who are seeking funding or looking for investment targets.

6. Conclusion

In this study, we model funding investment behavior as a link prediction problem based on social network features and obtain above baseline results across various machine learning methods and evaluation metrics. Our study is among the first to use data from CrunchBase as a social network for research purposes. Our contributions can be summarized as three folds.

First, social features are reasonable features for predicting funding investment behavior: our experiment results show that it is possible to predict funding investment behavior based on social relationships. Startups or companies should take their social relationships into consideration when seeking investments from a prospective investor .We use social features based on common neighbors, shortest path, common neighbors, Jaccard Coefficient, preferential attachment and Adamic/Adar and other node-wise

features to compute social similarity between an investor and company pair and discover that these features can be used to predict funding investment behavior. Not only can social information can be used to predict funding investment behavior, it is also a reliable and sound strategy for predicting funding investment behavior. Our prediction strategy based on social features and modeling it as a link prediction problem generally works well across the most common learning algorithms, including Decision Tree and SVM. Not only does it perform well in terms of aggregate performance, it also performs well in terms of individual industries.

Second, multiple link predictors can be used to gain deeper and broader insight into a network. We believe that we obtain good performance because we combine multiple link predictors as our learning feature. Since each link predictor, such as shortest path or common neighbor, measures different aspects of a social network, combining multiple link predictors allows us to gain deeper and broader insight into a network. In our case, companies seeking funding investment can use multiple social relationship indicators to gain a deeper understanding of their potential investors.

Third, there are general rules of thumb based on social relationship for when an investor is most likely to invest in a company. For instance, being "closer" in terms of shortest path generally leads to investment, while having more common neighbors does not usually lead to investments. These general rules will be useful for companies seeking investments. A summary of the rules is as follows: first, the greater the similarity between an Investor and a Company, the more likely that investments will occur. This is especially true in the case of lower shortest path scores and Adamic/Adar scores leading to investment, as visualized by the Decision Tree. Second, however, if the metric means competitive relationships, it would lead to less investment occurrences when there is greater similarity. This is especially true in terms of common neighbors and total number of shortest paths, where more common neighbors and more paths point to greater competition for funding or growth and it simply reflects market realities. We hope that our work can help companies better understand how and when investors invest and thus help 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 social relationship factors drive funding investment behavior. Most importantly, we would like to encourage companies to focus on social relationships in addition to other factors when seeking external investments.

This study also has its limitations. To strengthen the value of this study, the issues of venture capital investment activities (e.g. selection, appraisal, contracting, monitoring and exiting of target companies) can be further examined. We also need to understand the macro aspects of the evolution of investors. In this study, we are interested in only the macro-level behaviors of investors. In particular, we are interested in the general trends between social relationships and investment relationships and thus create a social network graph consisting of both investors and companies while taking both social relationships and investment into account. However, there could be "super stars" within this network which governments subsidize to kick start investing activities. Or, if high degree nodes in terms of social relationships generally coincide with high degree nodes in terms of investment relationships, companies or governments can attempt to enhance investment activities by creating networking events. Meanwhile, this study does not differentiate big investors or companies from the small ones, and thus it cannot answer the question like – if big investment like Total Elf making acquisition of SunPower for $1.30B or Meetic making acquisition of Match.com for $7M is similar to acquisition of Fondu by Airbnb for $575K?

Further investigations are also required on the question – can the prediction model, trends and guidelines be readily replicated in other countries with differing cultures and ways of life? Comparisons are often made between the USA and China (or possibly other countries), due to differing cultures, racial makeups, socio-political systems and to a certain extent the economic systems. Therefore, a good way to test the generalizability of the results of this study is to apply the research results to China's startup environment. This not only allows us to test the generalizability of our study, it also serves as a basis for governments or organizations looking to strengthen industries. This could be particularly useful for a government that wants to find out what it takes to create an ecosystem like Silicon Valley in China. China and other countries are also interested in how to kick start their start up and or investing environments. Thus, governments or organizations may want to make use of the prediction models, trends and varying guidelines and create policies that foster a healthy environment for both investors and companies in their own countries.

Finally, we hope our study points toward a better way for researching startups and investment that will allow us to tackle the problem in predictable and quantifiable manner. With prediction models and general trends in place, organizations, governments, individuals, startup owners and investors will have a better sense as to how investments in startups evolve, what their criteria may be, and most importantly, create an environment conducive to startups. We hope that our study serves as an important step for researchers concerned about startup investment and related issues.

Note

1. Seed round (typically the first round of investment) where company insiders provide start-up capital. Angel rounds occur where early outside investors buy common stock. Series A, B, C and so on generally refers to significant funding rounds that are meant to capitalize the company in question for six to 24 months as if further develop its products and more. Such rounds occurs after seed and angel rounds.

References

Adamic, L.A. and Adar, E. (2001), "Friends and neighbors on the web", *Social Networks*, Vol. 25 No. 3, pp. 211-230.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J. and Vigna, S. (2011), " Four degrees of separation" available at: http://arxiv.org/abs/1111.4570v3 (accessed September 2014).

Bakker, L., Hare, W., Khosravi, H. and Ramadanovic, B. (2010), "A social network model of investment behavior in the stock market", *Physica A: Statistical Mechanics and its Applications*, Vol. 389 No. 6, pp. 1223-1229.

Barnea, A., Cronqvist, H. and Siegel, S. (2010), "Nature or nurture: what determines investor behavior?", *Journal of Financial Economics*, Vol. 98 No. 3, pp. 583-604.

Bianco, M., Bontempi, M.E., Golinelli, R. and Parigi, G. (2013), "Family firms' investments, uncertainty and opacity", *Small Business Economics*, Vol. 40 No. 4, pp. 1035-1058.

Bliss, C.A., Frank, M.R., Danforth, C.M. and Dodds, P.S. (2014), "An evolutionary algorithm approach to link prediction in dynamic social networks", *Journal of Computational Science*, Vol. 5 No. 5, pp. 750-764.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Burchardt, J., Hommel, U., Kamuriwo, D.S. and Billitteri, C. (2014), "Venture capital contracting in theory and practice: implications for entrepreneurship research", *Entrepreneurship Theory and Practice*.

Chang, C.C. and Lin, C.J. (2012), *LIBSVM: A Library for Support Vector Machines",* Department of Computer Science, National Taiwan University, Taipei.

Cortes, C. and Mohri, M. (2003), "AUC optimization vs. error rate minimization", *Proceedings of the Advances in Neural Information Processing Systems (NIPS'2003)*, *British Columbia*.

Doran, J.S., Peterson, D.R. and Wright, C. (2010), "Confidence opinions of market efficiency and investment behavior of finance professors", *Journal of Financial Markets*, Vol. 13 No. 1, pp. 174-195.

Fan, L., Wu, W., Lu, Z., Xu, W. and Du, D.Z. (2013), "Influence diffusion, cmmunity detection, and link prediction in social network analysis", *Dynamics of Information Systems: Algorithmic Approaches*, Springer, New York, NY, pp. 305-325.

Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L. and Elovici, Y. (2011), "Link prediction in social networks using computationally efficient topological features", *Third IEEE International Conference on Social Computing, SocialCom. MIT*, *Boston, MA*.

Friedkin, N. (1983), "Horizon of observability and limits of informal control in organizations", *Social Forces*, Vol. 62 No. 1, pp. 54-77.

Gallagher, B., Tong, H., Eliassi-Rad, T. and Faloutsos, C. (2008), "Using ghost edges for classification in sparsely labeled networks", *ACM SIGKDD Conference on Knowledge Discovery and Data Mining Conference*, *Las Vegas, NV*.

Giot, P., Hege, U. and Schwienbaher, A. (2012), "Expertise of reputation? The investment behavior of novice and experienced private equity Funds", *29th International Conference of the French Finance Association (AFFI)*, March.

Girvan, M. and Newman, M.E.J. (2010), "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, Vol. 99 No. 12, pp. 7821-7826.

Grewal, R., Lilien, G.L. and Mallapragada, G. (2006), "Location, location, location: how network Embeddedness affects project success in open source systems", *Management Science*, Vol. 52 No. 7, pp. 1043-1056.

Grinblatt, M. and Keloharju, M. (2000), "The investment behavior and performance of various investor types: a study of Finland's unique dataset", *Journal of Financial Economics*, Vol. 55 No. 1, pp. 43-67.

Hamilton, R.H. (2001), "E-commerce new venture performance: how funding impacts culture", *Internet Research*, Vol. 11 No. 4, pp. 277-285.

Kajdanowicz, T., Kazienko, P. and Doskocz, P. (2010), *Label-Dependent Feature Extraction in Social Networks for Node Classification*, Springer Berlin Heidelberg, pp. 89-102.

Kalampokis, E., Tambouris, E. and Tarabanis, K. (2013), "Understanding the predictive power of social media", *Internet Research*, Vol. 23 No. 5, pp. 544-559.

Kargar, M. and An, A. (2011), "Discovering top-k teams of experts with/without a leader in social networks", *ACM Conference on Information and Knowledge Management*, *Glasgow*.

Kempe, D., Kleinberg, J. and Tardos, E. (2003), "Maximizing the spread of influence through a social network", *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, *Washington, DC*.

Kleinberg, J. (1998), "Authoritative sources in a hyperlinked environment", *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA*.

Kuppuswamy, V. and Bayus, B.L. (2013), "Crowdfunding creative ideas: the dynamics of project backers in Kickstarter", UNC Kenan-Flagler Research Paper.

Lappas, T., Liu, K. and Terzi, E. (2009), "Finding a team of experts in social networks", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 467-476.

Latapy, M., Magnien, C. and Del Vecchio, N. (2008), "Basic notions for the analysis of large two-mode networks", *Social Networks*, Vol. 30 No. 1, pp. 31-48.

Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010a), "Predicting positive and negative lnks in online social networks", *ACM WWW International Conference on World Wide Web (WWW), Raleigh, NC.*

Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010b), "Signed networks in social media", *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Atlanta, GA.*

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., Briesen, Jeanne, V. and Glance, N. (2007), "Cost-effective outbreak detection in networks", *ACM SIGKDD Conference on Knowledge Discovery and Data Mining Conference, San Jose, CA.*

Liang, Y.E. and Yuan, S.T. (2012), "The social behavior of investors", *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul.*

Liben-Nowell, D. and Kleinberg, J. (2007), "The link prediction problem for social networks", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 7, pp. 1019-1031.

McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001), "Birds of a feather: homophily in social networks", *Annual Review of Sociology*, Vol. 27, pp. 415-444.

Manning, C.D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Vol. 1, Cambridge University Press, Cambridge.

Martens, D., Vanhoutte, C., De Winne, S., Baesens, B., Sels, L. and Mues, C. (2011), "Identifying financially successful start-up profiles with data mining", *Expert Systems with Applications*, Vol. 38 No. 5, pp. 5794-5800.

Newman, M.E.J. (2001a), "Clustering and preferential attachment in growing networks", *Physical Review*, Vol. 64 No. 2, pp. 025102-025116.

Newman, M.E.J. (2001b), "The structure of scientific collaborative network," available at: www.researchgate.net/publication/12179403_The_structure_of_scientific_collaboration_networks (accessed September 2014).

Newman, M.E.J. (2006), "Modularity and community structure in networks", *Proceedings of the National Academy of Sciences, Vol. 103 No. 23*, pp. 8577-8582.

Page, L. and Brin, S. (1998), "The Anatomy of a large-scale hypertextual web search engine", *Proceedings of the Seventh International Conference on World Wide Web, Brisbane.*

Pahnke, E., McDonald, R., Wang, D. and Hallen, B. (2014), "Exposed: venture capital, competitor ties, and entrepreneurial innovation", *Academy of Management Journal, amj-2012.*

Seaman, C., McQuaid, R. and Pearson, M. (2014), "Networks in family business: a multi-rational approach", *International Entrepreneurship and Management Journal*, pp. 1-15.

Shane, S. and Cable, D. (2002), "Network ties, reputation, and the financing of new ventures", *Management Science*, Vol. 48 No. 3, pp. 364-381.

Tan, W.K. and Tan, Y.J. (2012), "An exploratory investigation of the investment information search behavior of individual domestic investors", *Telematics and Informatics*, Vol. 29 No. 2, pp. 187-203.

Vir Singh, P., Tan, Y. and Mookerjee, V. (2011), "Network effects: the influence of structural social capital on open source project success", *MIS Quarterly*, Vol. 35 No. 4, pp. 813-829.

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C. and Liu, C. (2012), "A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch", *Sixth International AAAI Conference on Weblogs and Social Media, Trinity College*, *Dublin*.

Uzzi, B. (1996), "The sources and consequences of embeddedness for the economic performance of organizations: the network effect", *American Sociological Review*, Vol. 61, pp. 674-698.

Further reading

Backstrom, L. and Leskovec, J. (2011), "Supervised random walks: predicting and recommending links in social networks", *ACM International Conference on Web Search and Data Mining (WSDM)*, *Hong Kong*.

About the authors

Yuxian Eugene Liang was a Graduate Student of Information Management in Commerce College of the National Chengchi University, Taiwan. His research interests include service science and social network analysis.

Soe-Tsyr Daphne Yuan is a Professor of Information Management in Commerce College of the National Chengchi University, Taiwan. Her research interests include service science, service design and e-commerce. Professor Soe-Tsyr Daphne Yuan is the corresponding author and can be contacted at: daphneyuans@gmail.com