VENTURE CAPITAL INVESTMENT: FROM RULE OF THUMB TO DATA SCIENCE

by

HAO ZHONG

A dissertation submitted to the Graduate School-Newark
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements for the degree of Doctor of Philosophy Graduate Program in Management written under the direction of Dr. Hui Xiong and approved by

Newark, New Jersey

May 2019
ABSTRACT OF THE DISSERTATION

Venture Capital Investment: From Rule of Thumb to Data Science

By HAO ZHONG

Dissertation Director: Dr. Hui Xiong

Recent years have witnessed the booming of venture capital market. Traditionally, venture investors (e.g., business angels, venture capitalists, private equity investors) make investment decisions based on past investment experiences, social relationship and/or qualitative assessment on startups. By offering capital and advice, venture investors could receive high financial returns once portfolio companies successfully exit, via acquisition or IPO (Initial Public Offering). Meanwhile, startups backed by venture capitals have higher opportunities to exit successfully, which are entrepreneurs’ striving goals at all times. It is thus critical for venture capitalists to find startups with high financial return potentials, and likewise for startups to get financing and business advice from right investors/mentors, especially at an early stage.

Extensive research has been conducted on venture capital investment analyses, mostly from finance and/or managerial perspectives in a quantitative manner. Studies based on post hoc methodologies (e.g., interviews and surveys) to understand venture capitalists’ natural decision-making process are doubtful. People’s retrospection is subject to rationalization and post hoc recall biases. Thus, from both academic and practical perspectives, the entrepreneurial finance industry has an active call for quantitative and methodologically sound studies on venture capital investments.

Which startups should venture capitalists invest in, when is the proper time to
fund, and what is the right amount? What are intrinsic hidden drivers for reaching investment deals? In this dissertation, we address these research problems by utilizing cutting-edge data-driven analytical methodologies. This dissertation starts with a background introduction and the scope of research problems, followed by an extensive literature review on state-of-the-art research. We then develop an analytical approach to assist venture capitalists to make better decisions on potential investment deals. We adopt recommender system techniques to learn VCs’ investment preferences and identify the right startup candidates at the screening stage. Our method mainly uses historical investment deals and additional firmographics, including startups’ geographic locations, their industry categories, historical acquisition records, leading products. We provide investment strategy, based on Modern Portfolio Theory, to maximize financial returns while suppressing investment risks. Also, we found venture capital investments and social relationships have a strong association. We propose and develop a probabilistic latent factor model to foresee venture capital investment deals using the information of social connections between members of VC firms and startups. We uniquely approach the problem – not through vague organizational social connection but directly using social information between members from both parties (VC firms and startups). To the best of our knowledge, we are the first to employ social information between venture capital firms and entrepreneurial companies for venture capital deals prediction and recommendation.
ACKNOWLEDGEMENTS

I would like to express my keen appreciation to all who motivated, inspired, assisted and supported me during my entire Ph.D. study.

Firstly and most importantly, I sincerely thank my advisor, Prof. Hui Xiong, for his persevering encouragement, forethoughtful guidance and inexhaustible support both intellectually and mentally. I enjoyed all my time studying under his advice and working with him as his research assistant. Without his guidance and support, I am not able to obtain all academic achievements, including this heavy and invaluable Ph.D. degree.

I would also like to thank my other committee members: Prof. Xiaodong Lin, Prof. Spiros Papadimitriou, and Prof. Lin Peng. I express all my appreciation for their critical instructions and constructive feedbacks on my research work and dissertation. Only with their help can I thoroughly inspect my work and improve the quality of my dissertation to a great extent.

Special thanks are to Dr. Jin Cao at Bell Labs, Dr. Maulin Patel at Philips Research - NA and Mr. Don Yan at DIA Associates, who get me involved in various projects and help me sharpen my skills and broaden my knowledge. Thanks are also to my colleagues and friends: Wenjun Zhou, Yong Ge, Keli Xiao, Zhongmou Li, Chuanren Liu, Bin Liu, Yanjie Fu, Zijun Yao, Konstantine Alexander Vitt, Jingyuan
Yang, Meng Qu, Can Chen, Farid Razzak, Junming Liu, Yanchi Liu, Qingxin Meng, Mingfei Teng, Hengshu Zhu, Bo Jin, Yuanchun Zhou, Xiaolin Li, Ling Yin, Hongting Niu, Bowen Du, Yanhong Guo, Yayao Zuo, Jie Liu, Li Huang, Chen Zhang, Deqing Wang, Pengpeng Zhao, Tong Xu, Guannan Liu, Liyang Tang, Qi Liu, Leilei Sun, Xinjiang Lu, Huang Xu, Jing Sun, and Chu Guan for their help and friendship.

I would also like to acknowledge our department and school at Rutgers for providing a perfect environment and facilities at all times. Thanks are also to Ms. Luz Kosar, Ms. Monnique Desilva, and Mr. Goncalo Filipe for all their help and support.

Finally, I would like to thank my parents, my sister and all my family members, for their love, support, and understanding. My work is dedicated to them, as always.
# TABLE OF CONTENTS

**ABSTRACT** .............................................................................................................. ii

**ACKNOWLEDGEMENTS** ............................................................................................ iv

**LIST OF TABLES** .................................................................................................... viii

**LIST OF FIGURES** ................................................................................................... ix

**CHAPTER 1. INTRODUCTION** ................................................................................... 1

1.1 Research Motivation .............................................................................................. 2

1.2 Related Works ........................................................................................................ 3

1.2.1 Traditional Venture Capital Research ............................................................... 3

1.2.2 Data Mining in Venture Capital Research ......................................................... 13

1.2.3 Recommender Systems ...................................................................................... 15

1.3 Contributions ......................................................................................................... 26

**CHAPTER 2. WHICH STARTUP TO INVEST IN: A PERSONALIZED PORTFOLIO STRATEGY** ............................................................................................................ 29

2.1 Introduction ............................................................................................................ 29

2.2 Overview ................................................................................................................. 31

2.3 Matching Investment Preferences ........................................................................ 33

2.3.1 Model Specification .......................................................................................... 35

2.3.2 Model Inference ............................................................................................... 38

2.3.3 Recommendation with Bayesian PLF Model ................................................... 42

2.4 Optimizing Investment Return and Risk ................................................................ 42

2.5 Results and Discussions ......................................................................................... 44

2.5.1 Dataset ............................................................................................................. 45

2.5.2 Implementation Details for Matching Investment Preferences ....................... 46

2.5.3 Performance Evaluation for Matching Investment Preferences ..................... 48

2.5.4 Managerial Implications on Investor’s Propensities ........................................ 52

2.5.5 Portfolio Optimization for Investment Strategy ................................................. 53
LIST OF TABLES

2.1 Mathematical Notations ..................................................... 35
2.2 List of Additional Properties of Startups ............................. 39
2.3 RMSE and MAE of our algorithm and baseline algorithms. .... 48
2.4 Precision and Recall comparisons of our algorithm and baseline algorithms. .................................. 50
2.5 MAP and NDCG@$N_{test}$ of our algorithm and baseline algorithms. ................................................................. 51
2.6 Investment return and risk comparison. ............................... 54
3.1 The set of membership and connection labels. ......................... 69
3.2 The significant labels of members and connections. ............. 88
# LIST OF FIGURES

1.1 The venture capital process ............................................. 6
1.2 Characteristics of institutional venture capital, business angel and cor-
porate venture capital .................................................. 7
1.3 Kaleidoscope of research in managerial venture capital .......... 9

2.1 Funding amounts from a random VC in Crunchbase dataset. The left
figure is the original distribution of funding amounts while the right
one is the distribution after log(·). ...................................... 34
2.2 Distribution of investor-startup distances in Crunchbase dataset. ...... 36
2.3 Estimated distribution of investor’s propensity to startup properties . 53

3.1 The illustration of CoRec model ........................................ 66
3.2 The percentage of investment in different data samples. ............ 67
3.3 The illustration of tensors $F$ and $C$. .................................. 73
3.4 The estimation process with synthetic data. ........................... 77
3.5 The normalized objective with different $r_1$. ......................... 80
3.6 The normalized objective with different $r_2$. ......................... 80
3.7 The normalized objective with different $\gamma$. ....................... 81
3.8 The normalized objective with varying noise $\sigma$. ................. 82
3.9 The member titles in VC firms and startups. ......................... 85
3.10 The connection groups from VC (startup) members to startup (VC)
members ................................................................. 85
3.11 RMSE, MAE and MAP on real-world data. .......................... 87
CHAPTER 1
INTRODUCTION

Recent years have witnessed the significant booming of startups. According to Global Entrepreneurship Monitor (GEM) 2014 Global Report\(^1\), the Total Early-Stage Entrepreneurial Activity (TEA) rate, which is the rate of adults in the process of starting a business or are running a new company, is more than 13.1% in the United States. Specific industries, like Software, Internet, Biotechnology, have the highest growth speed and are the most heated fields chased by numerous entrepreneurs. Seigle\(^2\) pointed out the main reason for the “startup explosion” is that the basic building blocks for digital services and products have become so evolved, cheap and ubiquitous that they can be easily combined and recombined. Additionally, the maturity of entrepreneurial financing market further fosters newly founded companies. According to the MoneyTree\(^{TM}\) Report by PricewaterhouseCoopers LLP (PwC) and the National Venture Capital Association (NVCA) in 2014\(^3\), venture capitalists (VCs) invested $48.3 billion in 4,356 deals in 2014, an increase of 61 percent in dollars and a 4 percent increase in deals over the prior year, based on data from Thomson Reuters.

Venture capital is a form of private finance provided in return for an equity stake in potentially high growth companies [Stone et al., 2013]. As one of the biggest capital

---

suppliers to entrepreneurial firms, their primary role is to fuel startups for facilitating operation, production, or expansion. Meanwhile, [MacMillan et al. 1989, Sapienza 1992, Hellmann and Puri 2002] stated that venture capitalists play an extra role as directors influencing developments on the organizations, such as devoting their knowledge and contacts, offering strategic and operational advice, helping optimize human capital structure, and planning stock options. By offering capital and advice, investors would receive high returns if their portfolio companies successfully exit, namely being acquired or going public.

In the following of this chapter, we first present our research motivation, followed by state-of-the-art literature reviews. We then highlight our contributions to addressing major and challenging research problems in this community. We give an overview of this dissertation in the end.

1.1 Research Motivation

Which startups should venture capitalists invest in, when is the proper time to fund, and what is the right amount? What are intrinsic hidden drivers for reaching investment deals? Studies on these venture financing questions have been undertaken for decades. In earlier times, studies were relying on personal interviews and/or questionnaires [Stuart and Abetti 1987, Fried and Hisrich 1994, Clarysse et al. 2005]. For instance, to predict success/failure of newly founded companies, Stuart and Abetti [1987] developed a 68-item questionnaire for 24 new technical ventures associated with the RPI Incubator or Technology Park. Sorensen [2007] studied the relationship between VCs’ investment experiences and the outcome of their investments, given
data from 1,666 sampled companies. Zacharakis and Meyer [1998] argued that studies based on post hoc methodologies (e.g., interviews and surveys) to analyze venture capitalists decision-making process are doubtful. In other words, people’s retrospection, which most of the past studies relied on, is subject to rationalization and post hoc recall biases. As a consequence, there’s an active call for more sophisticated and advanced quantitative approaches to address venture capital investment problems. In this dissertation, we address these research problems by utilizing cutting-edge data-driven analytical technologies to find more solid and theoretical basis.

1.2 Related Works

In this part, Section 1.2.1 introduces the development of venture capital research, mainly from managerial and entrepreneurial perspectives. It seeks to provide, about venture capital research, background information, domain knowledge, research problems, and traditional methodologies. Beyond these conventional techniques, Section 1.2.2 investigates the state-of-the-art research work particularly on venture capital research enhanced by recommendation techniques.

1.2.1 Traditional Venture Capital Research

Landström [Landström, 2007] provided a comprehensive summary about the development of the venture capital industry in the USA. Dated back to 1946, General Georges Doriot founded the first institutional venture capital firm, American Research and Development (ARD) in Boston. Between 1946 and 1977, the flow of money into venture capital funds was never over a few hundred million dollars annually (often much less). The venture capital market stagnated even more, at the
beginning of the 1970s, mainly due to a sharp rise in capital gains tax – from 25 to 49 percent – which reduced the potential profit on investments. Almost by the end of the 1970s, the venture capital industry was very small, homogeneous in strategy and practice, and competition for deals was weak. However, owing to the increase in investment opportunities and the introduction of tax-related incentives, the venture capital industry began to grow dramatically in the early 1980s. Specifically, venture capital firms increased both in number and size and, as a consequence, the market showed increased heterogeneity across firms, and greater specialization in investment stage, industry and region in the 1980s. After this period of growth, the development during the 1980s and 1990s had several ups and downs. However, there was renewed interest at the beginning of the 1990s – due to new possibilities on the initial public offerings (IPO) market and the successful exits of many experienced venture capitalists [Gompers and Lerner, 2010]. Afterward, the venture capital market in the US has declined due to the dot.com crash, where the drop was more significant than in many other countries. In recent years, another conspicuous prosperity of the venture capital industry was witnessed, as stated at the beginning of this chapter.

Types of venture capital markets

As for the categories of venture capital markets, there are mainly three: institutional venture capital, corporate venture capital, and informal venture capital [Landström 2007]. Institutional venture capital is defined as professional investments of long-term, unquoted, risk equity finance in new firms where the primary reward is eventual capital gain supplemented by dividends [Robbie and Mike 1998]. As indicated in the
definition by Mason and Harrison [Mason and Harrison, 1999], an institutional venture capital firm can take different organizational forms, depending on the ownership structure, but usually consists of:

- **Independent Limited Partnership**, in which the venture capital firm serves as the general partner, raising capital from limited partners such as institutional investors (for example, pension funds, insurance companies and banks);

- **Captive venture capital firms**, which are mainly funded by the internal resources of a parent organization – often a financial institution, such as a bank or insurance company, but sometimes by a larger non-financial company (so-called corporate venture capital);

- **Government venture capital organizations**, which are financed and controlled by government institutions.

Since 1980s, the limited partnership has emerged as the dominant organizational form in venture capital. In a limited partnership, the venture capitalists are general partners and control the funds activities, whereas the investors act as limited partners who are not involved in the everyday management of the fund (Fig. 1.1 [Landström, 2007] [Mason and Harrison, 1999]). Note that the definitions of institutional venture capital in Europe are somewhat different and venture capital is usually considered synonymous with *private equity* in a more general sense and includes investments in terms of early and expansion stage financing as well as those covering a range of other stages such as funding of management buy-outs, consolidations, turnarounds, and so
on. On the other hand, in the US, the term *venture capital* is narrower and refers to early-stage investments in growth-oriented companies [Landström, 2007].

Second, *corporate venture capital* is defined as equity or equity-linked investments in young, privately held companies, where the investor is a financial intermediary of a non-financial corporation, by Maula [Maula et al., 2001]. Thus, the main difference between *institutional venture capital* and *corporate venture capital* is the fund sponsor – in *corporate venture capital* the only limited partner is a corporation or a subsidiary of a corporation [Landström, 2007]. Third, *business angels*, the main players in informal venture capital market, are defined as private individuals who make investments directly in unlisted companies in which they have no family connections [Mason and Harrison, 2000]. Fig. 1.2 [Landström, 2007] summarizes the similarities and differences between *institutional venture capital*, *business angels* and *corporate venture capital*, from the perspectives of 1) source of funds, 2) legal form, 3) motive for investment, 4) investment, and 5) monitoring, respectively. In this literature review, we primarily concentrate on *institutional venture capital* for the following rea-
Figure 1.2: Characteristics of institutional venture capital, business angel and corporate venture capital

Strands of venture capital research

As the venture capital industry grew in scope and importance, more and more scholars are of great interests in it. Many scholars studied venture capital finance from different disciplines – mainly from management and entrepreneurship as well as from the field of finance and economics [Landström 2007]. One group of researchers focused on the
venture capital process – “managerial-oriented venture capital research”, a micro-
level focus, whose topics include fundraising, pre-investment activities, the exit of
the investment. On the other hand, another group of scholars has been attentive
to “market-oriented venture capital research”, a macro-level focus from perspectives
of finance and economics. Their studies consist of, but not limited to, the flow
of venture capital, its role in the development of new industries, regional aspects
of venture capital and the like [Landström, 2007]. Here we pay more attention to
the analysis of venture capital investment deals (venture capitalists’ decision-making
system, investment deals assessment, the relationship between venture capitalists
and startups, etc.) and thus our investigation of literature are more from managerial
venture capital research community.

After the development of venture capital research for decades, managerial venture
capital research evolves in a more theory-driven direction and have been developed
along different paths. Fig. 1.3 [Sapienza and Villanueva, 2007] represents the dimen-
sions by which the most common examples of past venture capital research might be
viewed and classified. By looking at the interior of the kaleidoscope, we see three
overlapping dimensions:

- type of venture capital (for example institutional venture capital – VCF, business
  angels – BA, or corporate venture capital – CVC);
- interests or perspectives being investigated (for example investor vs. entrepreneur);
- stages of the venture capital process (for example fundraising and selection).

Although each dimension is composed of several elements, most studies center on
one element within each dimension Sapienza and Villanueva 2007. Take Shepherd et al. 2000 for instance, it examined how an investor type (the institutional venture capital firm) attempts to maximize returns (investor’s perspective) via decisions made during the selection stage of the venture capital cycle.

The elements in bold indicate that they caught significant attention of the majority of researchers. As shown, the most common studies focus on institutional venture capital firm type, from the investor’s perspective, in the selection and/or monitoring stages of the venture capital cycle Sapienza and Villanueva 2007, which is in line with our focus. There are several reasons why researchers shed more light on the issues on investors side, rather than startups side Sapienza and Villanueva 2007. On the one hand, venture capitalists are the immediate stakeholders for venture cap-
ital research. On the other, from a practical perspective, venture capitalists (perhaps except for angels) are more visible than entrepreneurs and can provide researchers with access to many ventures. Moreover, researchers focus more on selection stage since collecting information on selection criteria is particularly amenable to the questionnaire and interview techniques, favored by early researchers. Interestingly, the dominance of post-investment activities in time spent by investors caught great interests of scholars, which leads to extensive research on the monitoring stage. Due to various aspects of other stages of the venture capital process involve individuals outside venture capitalist-entrepreneur dyad (for example fund raising involves limited partners, and exit involves several external organizations), research designs on those phases are particularly complicated.

Investment decision-making process

In this section, we shed more light on the selection stage, which has caught great attention of the majority in venture capital research community. Zacharakis and Shepherd [Zacharakis and Shepherd, 2007] showed that improving the investment decision can improve the venture capitalist’s performance. Also, better understanding of how venture capitalists make decisions and more importantly, how they can improve their decision process will lead to more efficient use of their time and higher overall returns [Zacharakis and Meyer, 2000].

In early years, venture capitalists’ decision making has produced empirically derived lists of venture capitalists’ espoused criteria, including early seminal articles by Tyebjee and Bruno [Tyebjee and Bruno, 1984] and MacMillan and colleagues
However, Zacharakis and Meyer [Zacharakis and Meyer, 1998] later found that venture capitalists aren’t accurate in self-introspection. In other words, post hoc studies may not truly capture how venture capitalists use decision criteria [Zacharakis and Shepherd, 2007]. Then several studies, including [Sandberg et al., 1988, Hall and Hofer, 1993, Zacharakis and Meyer, 1995], attempted to overcome prior post hoc study flaws by using verbal. Verbal protocols are real-time experiments where venture capitalists think aloud as they are screening a business plan [Ericsson and Crutcher, 1991]. Venture capitalists aren’t required to retrospect about their thought processes which removes recall and post hoc rationalization biases [Hall and Hofer, 1993]. While verbal protocols are rich in the amount of data collected from each venture capitalist, they are time-consuming as the researcher needs to observe each venture capitalist as he/she reviews a plan [Zacharakis and Shepherd, 2007]. The method inherently suffered extreme small data sample size issue, which was later improved by conjoint analysis.

Conjoint analysis and policy capturing (a type of conjoint analysis) move beyond survey methods used to identify decision criteria and verbal protocols used to assess how and when criteria are used [Zacharakis and Shepherd, 2007]. Conjoint analysis is a technique that requires respondents to make a series of judgments, assessments or preference choices, based on profiles from which their captured decision processes can be decomposed into its underlying structure [Shepherd and Zacharakis, 1997]. Conjoint analysis and policy capturing allows us to gain a deeper understanding of the venture capital decision process [Shepherd and Zacharakis, 1999]. Not only can researchers capture how important each decision criterion is to the decision relative
to other decision standards, but also it allows for examining contingent decision processes [Zacharakis and Shepherd 2005]. Thus, the research in venture capital decision making has followed a natural progression from identifying decision criteria through post hoc surveys to understanding how that information is utilized during the actual decision via verbal protocols to conjoin analysis [Zacharakis and Shepherd 2007]. However, the characteristics of traditional methods, such as low information utilization and simplistic analytical techniques, confine their popularities and advances.

In spite of these limitations, some research in management and entrepreneurship fields have inspired our work. Kaplan and Strömberg [2000] analyzed how venture capitalists choose investments by taking into account different factors, such as market size, investment strategy, management team, etc. However, their analysis was more qualitative and descriptive, in which the conclusion cannot be well generalized and thus is less convincing. Jeng and Wells [2000] studied a variety of determinants of venture capital funding, including initial public offerings (IPOs), gross domestic product (GDP) and market specialization growth, labor market rigidities, etc., by utilizing linear regression models. They concluded that IPOs are the strongest driver of venture capital investment, which coincides with our idea of incorporating investment return and risk in the model (see Chapter 2).

Note that, in this section, we aim to sketch the research field of venture capital investment, which serves as domain knowledge for our further quantitative analyses of similar problems. We continue our discussion on more related work which utilizes advanced and sophisticated data mining techniques in Section 1.2.2.
1.2.2 Data Mining in Venture Capital Research

The venture capital research community has an active call for advanced and sophisticated information processing techniques given that much more venture capital investment information becomes publicly accessible. Opportunely, with accelerated development for decades, recommendation engines become a favorite tool to evaluate and filter redundant information in various scenarios. The startups screening and evaluation processes to VCs are similar to have startups selected in recommendation manner. Plus, recommendation system has demonstrated its efficiency and effectiveness via a variety of practical use scenarios [Linden et al., 2003, Miller et al., 2003, Billsus et al., 2002].

More specifically, data mining and predictive analytics are employed to improve venture capital investments. These research work can be further divided into two groups. One small group is startup-oriented approaches, intending to predict successful exits of startups (i.e., being acquired or going public). For example, Xiang et al. [2012] utilized profiles and news articles of companies and people to predict Mergers and Acquisitions (M&A) activities using topic modeling techniques. Sharchilev et al. [2018] addressed the problem of predicting success/failure of startup companies at their early development stages by utilizing web-based open data sources. This type of approach made use of investment utilities but has limited capability in terms of personalized recommendation.

The other group of methods is investor-oriented approaches, which aims to learn investors’ decision-making model or predict their overall investment performance
by identifying successful investors. Gupta et al. [2015] proposed the InvestorRank method to identify successful investors via understanding how an investor’s collaboration network change over time. The key idea is to measure if an investor gets increasingly close to some known exemplary successful investors as time passes. Zhang et al. [2017] predicted startup crowdfunding success by performing a longitudinal data collection and analysis of AngelList. Their analyses showed that active engagement on social media is highly correlated to crowdfunding success. The above work studied the relationship between investors and start-up companies from certain perspectives but failed to present insightful guidance for investments.

Zhang et al. [2015] formulated investor-startup pairing problem as link prediction problem in bipartite venture capital investment networks, whose solution relied solely on the topological structure of investment networks and failed to integrate other information. Stone et al. [2013] conducted an empirical study on recommending top-n startups for venture capital firms, which is the pioneering work to ours. They tested several collaborative filtering based recommendation algorithms on VentureSource dataset and demonstrated the efficacy of recommender systems in this domain. In Stone [2014], the author improved his work and propose revised algorithms based on SVD++ [Koren 2008] and Bayesian Personalized Ranking (BPR) [Rendle et al. 2009], respectively, with the incorporation of industry categories and company profiles information. Nonetheless, with maximum-a-posterior (MAP) alike method, their work failed to quantify uncertainties in the model fitting, which can be addressed by employing full Bayesian approach [Salakhutdinov and Mnih 2008].

Besides, in contrast to our work, they failed to take the expected investment return
and potential risk into consideration. Mohamedali [2018] developed a tool (VCWiz) to help inexperienced founders navigate founder-investor matching process and generated founders ranking based on communications between founders, investors and their mutual connections (VCWiz Email Graph). Their work will facilitate future research in the community.

1.2.3 Recommender Systems

In this dissertation, we aim at assisting venture capitalists to make better decisions in venture capital investment deals, namely to recommend the right startups to investors. Recommender systems inherently have the capability of guiding users in a personalized way to interesting or useful objects in a large space of possible options [Burke, 2002], therefore we resort to recommender systems, which have been well studied and have demonstrated its competence in addressing problems like information filtering.

Many researchers have studied the taxonomy of recommender systems, which we will summarize here briefly. Burke [2002] distinguishes between five different recommendation techniques: collaborative, content-based, utility-based, demographic, and knowledge-based. An early work in recommender systems Resnick and Varian [1997], which focuses on collaborative recommenders identifies five dimensions to place the systems in a technical design space. The dimensions characterize properties of users’ interactions with the recommender and the aggregation methods of users’ evaluations (ratings). A more recent survey on recommender systems Adomavicius and Tuzhilin [2005] classifies recommendation methods (omitting knowledge-based)
into three main categories: content-based, collaborative, and hybrid recommendation approaches. They are defined as, respectively [Adomavicius and Tuzhilin 2005]:

- Content-based: user will be recommended items similar to the ones the user preferred in the past;
- Collaborative: user will be recommended items that people with similar tastes and preferences liked in the past;
- Hybrid approaches: these methods combine collaborative and content-based methods.

Note that we adopt the above taxonomy in the dissertation.

**Content-based methods**

Content-based Information Filtering (IF) systems need proper techniques for representing items and producing user profiles, and some strategies for comparing user profile with item representation. The recommendation process is performed in three steps, each of which is handled by a separate component [Lops et al. 2011]:

- **Content Analyzer** – When information has no structure (e.g., text), some pre-processing step is needed to extract structured, relevant information.
- **Profile Learner** – This module collects data representative of the user preferences and tries to generalize this data, in order to construct the user profile. Usually, the generalization strategy is realized through machine learning techniques, which can infer a model of user interests starting from items liked or disliked in the past.
• **Filtering Component** – This module exploits user profile to suggest relevant items by matching the profile representation against that of items to be recommended.

Regarding **Content Analyzer**, items that can be recommended to the user are represented by a set of features, also called attributes or properties. In most content-based filtering systems, item descriptions are textual features extracted from Web pages, emails, news articles or product descriptions. Usually, a document is represented by a vector in an n-dimensional space, where each dimension corresponds to a term from the overall vocabulary of a given document collection, referred as **Vector Space Model (VSM)** or **bag-of-words** representation [Lops et al., 2011]. There are two issues to account for before the system starts to learn user profile: weighting terms and measuring feature vector similarity. The most commonly used term weighting scheme is **term frequency - inverse document frequency** (TF-IDF) weighting [Salton, 1989]. Supposed that $N$ is the total number of documents that can be recommended to users and that keyword $k_i$ appears in $n_i$ of them. Moreover, assume that $f_{i,j}$ is the number of times keyword $k_i$ occurs in document $d_j$. Then, normalized term frequency $TF_{i,j}$ of keyword $k_i$ in document $d_j$ is

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}},$$  \hspace{1cm} (1.1)$$

where the maximum is computed over the frequencies $f_{z,j}$ of all keywords $k_z$ that appear in the document $d_j$. Next, the inverse document frequency for keyword $k_i$ is usually defined as

$$IDF_i = \log \frac{N}{n_i}.$$  \hspace{1cm} (1.2)
Then, the TF-IDF weight for keyword $k_i$ in document $d_j$ is defined as

$$\omega_{i,j} = TF_{i,j} \times IDF_i.$$  

(1.3)

In addition, *cosine similarity* is the most widely used similarity measures for describing the proximity of two document vectors $\omega_c$ and $\omega_s$, defined as

$$sim(\omega_c, \omega_s) = \frac{\omega_c \cdot \omega_s}{||\omega_c||_2 \times ||\omega_s||_2},$$

(1.4)

where $K$ is the total number of keywords in the system.

In the phase of *Profile Learner*, let $Prof(c)$ be the profile of user $c$ containing tastes and preferences of this user. Different methods have been proposed to generate user profiles $Prof(c) = (\omega_c1, \ldots, \omega_{ck})$. For example, some averaging approach, such as Rocchio algorithm [Rocchio, 1971], can be used to compute the user profile as an “average” vector from an individual content vectors [Lang, 1995]. On the other hand, [Pazzani and Billsus, 1997] uses a Bayesian classifier in order to estimate the probability that a document is liked. The Winnow algorithm [Littlestone and Warmuth, 1994] has also been shown to work well for this purpose, especially in the situations where there are many possible features. Lastly, in the *Filtering Component* stage, supposed that TF-IDF representation of document $s$ is $Prof(s) = (\omega_s1, \ldots, \omega_sh)$, the utility function $u(c, s) = sim(Prof(c), Prof(s))$ can be computed using Equation (1.4), serving as a scoring heuristic and the relevant documents can be retrieved accordingly.

The content-based recommendation paradigm has several advantages when compared to the collaborative one (see Section 1.2.3). First, content-based recommenders exploit solely ratings provided by active user to build his/her own profile, namely no
information from other users is needed. Second, the system is transparent and explainable. Explanations on how the recommender system works can be provided by explicitly listing content features or descriptions that caused an item to occur in the list of recommendations, whereas the collaborative systems are relatively black boxes alike. Last, content-based recommenders are capable of recommending items not yet rated by any user, which does not suffer from the first-rater problem. Nonetheless, there are also some shortcomings for content-based methods. On the one hand, no content-based recommendation system can provide proper suggestions if the analyzed content does not contain enough information to discriminate items liked from the ones disliked by the user. On the other hand, content-based recommenders have no inherent method for finding something unexpected. This drawback is also called *serendipity* problem to highlight the tendency of content-based systems producing recommendations with a limited degree of novelty.

**Collaborative methods**

Other than content-based recommender systems, there is another big category based on collaborative filtering (CF), predicting the utility of items for a particular user given the items previously rated by other users [Adomavicius and Tuzhilin 2005]. Collaborative approaches overcome some of the limitations of content-based ones [Desrosiers and Karypis 2011]. For instance, items for which the content is not available or difficult to obtain can still be recommended to users given feedbacks of other users. Furthermore, collaborative recommendations are based on the quality of items as evaluated by peers, instead of relying on content that may be a bad
indicator of quality. Finally, unlike content-based systems, collaborative filtering ones can recommend items with fairly different content, as long as other users have shown interests for them.

**Neighborhood-based methods**

Collaborative methods can be further grouped into two categories: *neighborhood-based* (*memory-based* or *heuristic-based*) methods and *model-based* methods [Breese et al., 1998]. *Neighborhood-based* methods are heuristics that make rating predictions based on the entire collection of previously rated items by the users [Adomavicius and Tuzhilin, 2005], which can be done in two ways: *user-based* and *item-based*, explained as follows:

- *User-based* systems evaluate the interest of a user $u$ for an item $i$ using the ratings for this item by other users, called neighbors with similar rating patterns;

- *Item-based* systems predict the rating of a user $u$ for an item $i$ based on the ratings of $u$ for items similar to $i$.

Supposed that the user-item rating matrix is denoted as $R$ and there are two users $u$ and $v$. Let $r_u$ and $r_v$ be the $u^{th}$ and $v^{th}$ rows of matrix $R$, respectively. Without loss of generality, we discuss user-based recommendation method, whereas the item-based case is analogous. Thus, the recommendation procedure can be illustrated in two steps [Su and Khoshgoftaar, 2009]. First, the system calculates the similarity or weight, $\omega_{ij}$, which reflects distance, correlation, or weight, between two users. There are many different methods to compute similarity or weight between users. The vector cosine-based similarity and the Pearson correlation similarity are the most
frequently used ones [Su and Khoshgoftaar, 2009]. The definition of cosine similarity is analogous to Eq. (1.4), of which the only difference is it calculates the similarity between vectors of actual user-specified ratings, as

$$sim_{u,v} = \frac{r_u \cdot r_v}{||r_u|| \times ||r_v||},$$

(1.5)

The Pearson correlation similarity is defined as [Herlocker et al., 1999]

$$sim_{u,v} = \frac{\sum_i (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_i (r_{v,i} - \bar{r}_v)^2}}$$

(1.6)

At the second step, it produces a prediction for the active user by taking an aggregate of some other (usually, the $N$ most similar) users for the same item $i$ [Adomavicius and Tuzhilin, 2005]. Some examples of the aggregation function are:

(a) $r_{u,i} = \frac{1}{N} \sum_{u' \in S} r_{u',i}$,

(b) $r_{u,i} = \frac{1}{\sum_{u' \in S} |sim_{u,u'}|} \sum_{u' \in S} sim_{u,u'} \times r_{u',i},$

(1.7)

(c) $r_{u,i} = \bar{r}_{u'} + \frac{1}{\sum_{u' \in S} |sim_{u,u'}|} \sum_{u' \in S} sim_{u,u'} \times (r_{u',i} - \bar{r}_{u'})$,

where the average rating of user $u$, $\bar{r}_{u'}$, defined as

$$\bar{r}_{u'} = \frac{1}{|L_u|} \sum_{w \in L_u} r_{u,w}, \text{ where } L_u = \{w \in L_u | r_{u,w} \neq \emptyset\}.$$

(1.8)

There have been many collaborative systems developed in academia and industry [Adomavicius and Tuzhilin, 2005]. Grundy system [Rich, 1979] was considered the first recommender system, which proposed using stereotypes as a mechanism for building models of users based on a limited amount of information on individual user. Using stereotypes, the Grundy system would build individual user models to recommend relevant books to each user. Later on, the Tapestry system relied on each user
to identify like-minded users manually [Goldberg et al., 1992]. GroupLens [Konstan et al., 1997, Resnick et al., 1994], Video Recommender [Hill et al., 1995], and Ringo [Shardanand and Maes, 1995] were the first systems to employ collaborative filtering algorithms. Other examples of collaborative recommender systems include book recommendation system from Amazon.com, PHOAKS system that helps people find relevant information on the WWW [Terveen et al., 1997], and Jester system that recommends jokes [Goldberg et al., 2001].

There are several advantages of neighborhood-based recommendation methods [Desrosiers and Karypis, 2011]. First, neighborhood-based methods are intuitive and relatively simple to implement. The second is justifiability, i.e., providing a concise and intuitive justification for the computed predictions. Third, they require no costly training phases. Although the recommendation phase is usually more expensive than for model-based methods (see Section 1.2.3), the nearest-neighbors can be pre-computed in an offline step, providing near instantaneous recommendations. The last aspect is stability as they are little affected by the constant addition of users, items and ratings, which are typically observed in large commercial applications. For instance, once item similarities have been computed, an item-based system can readily make recommendations to new users, without having to re-train the system.

**Model-based methods**

Model-based CF approaches are based on prediction models that have been trained using the U-I matrix, in whole or in part, as input [Shi et al., 2014]. Examples of conventional model-based CF approaches include the Bayesian network model [Breese...
et al., 1998], which models the conditional probability between items; the latent semantic model [Basilico and Hofmann, 2004], which clusters users and items around latent classes of U-I interactions; and the mixture model [Si and Jin, 2003], which models probability distributions of items within each cluster of like-minded users.

Recently, matrix factorization (MF) techniques have attracted considerable attention due to their advantages concerning scalability and accuracy, as witnessed by the algorithms developed within the Netflix contest [Koren, 2010].

Mnih and Salakhutdinov [2007] presents Probabilistic Matrix Factorization (PMF) model which, including its variants, have been widely employed in all sorts of recommender systems. The main drawback of PMF model is the need for manual complexity control in training phase to make model generalize well. To this end, Salakhutdinov and Mnih [2008] developed a fully Bayesian treatment of the PMF model. They introduce priors for the hyperparameters and maximize the log-posterior of the model over both parameters and hyperparameters, which allows model complexity to be controlled automatically based on the training data [Mnih and Salakhutdinov, 2007].

Model-based methods with additional information

Model-based CF, Matrix Factorization approaches in particular, can be extended to incorporating additional information into recommender systems. Shi et al. [2014] grouped additional information for recommender systems into two types: rich side information about users and items, and interaction-associated information. Speaking of rich side information, two sources of information are important, social networks information (we will discuss it in Sec. 1.2.3) and user-contributed information. Specif-
ically, user-contributed information includes tags, geotags, multimedia content, and reviews and comments [Shi et al., 2014].

Extensive work has been devoted to extending model-based CF approaches to incorporating rich side information. Popescul et al. [Popescul et al., 2001] have extended the aspect/topic model of Hofmann et al. [Hofmann, 1999] to incorporate the item side information. Likewise, Wetzker et al. [Wetzker et al., 2009] proposed to extend probabilistic latent semantic analysis (PLSA) [Hofmann, 1999] to integrate item-tag relations with the U-I matrix for item recommendation. Apart from topic models, a particular family of model-based approaches, matrix factorization, has drawn the most attention in the recommender system research community. Singh and Gordon [2008] proposed collective matrix factorization (CMF), which simultaneously factorizes multiple matrices, including the U-I matrix and matrices containing the side information. Moreover, one of the most influential frameworks, regression-based latent factor models [Agarwal and Chen, 2009] were proposed to integrate attributes of both users and items with U-I preference data into a generalized linear model for preference prediction. Particularly, fLDA [Agarwal and Chen, 2010] can be regarded as a specific extension of regression-based latent factor models that targets the recommendation scenarios with rich side information. Further work that extends MF to incorporating side information is the localized matrix factorization (LMF) [Agarwal et al., 2011], which employs local latent factors for each entity under different types of side information, referred to as contexts.

On the other hand, algorithms for incorporating interaction-associated information can be divided into four groups [Shi et al., 2014]. The first is time-dependent
CF, which focuses on improving CF performance by modeling the dynamics of user preference over time. For example, in [Koren et al., 2009], latent factors of users and items are designed as decay functions of time and also linked to each other based on time. Then, the latent factors of users and items at different time are learned individually and fine-grained for improved prediction accuracy. The second group is tensor factorization (TF). Researchers in recommender systems have exploited the Tucker TF model for processing [user, item, tag, usage (boolean)] data for either item recommendation [Xu et al., 2006], or tag recommendation [Symeonidis et al., 2008], or both [Symeonidis et al., 2010]. Third, factorization machines [Rendle, 2010] are another group of important algorithms. The innovative idea of FM is to transform the multidimensional data associated with each U-I interaction into a real-valued feature vector. The last group is graph-based approaches. For instance, temporal recommendations are mined from the graph using random walk with restarts [Xiang et al., 2010]. Likewise, Lee et al. [2011] have proposed to use random walk with restarts on a graph involving users and items together with interaction-associated information sources, such as location and time.

Social recommendation

A social recommender system improves on the accuracy of traditional recommender system by taking social interests and social trusts between users in online social networks as additional inputs [Yang et al., 2014a]. Yang et al. [2014a] classifies CF-based social recommender systems into two main categories: Matrix Factorization (MF) based social recommendation approaches and Neighborhood-based social recom
recommendation approaches. In this paper, we place concentration on MF-based social recommender systems, in which user-user social trust information is integrated with user-item feedback history (e.g., ratings, clicks, purchases) as to improve the accuracy of traditional MF-based recommender systems, which only factorize user-item feedback data. The common rationale behind MF-based social recommender systems is that a user’s taste is similar to and/or influenced by her trusted friends in the social network. For example, in [Ma et al., 2008], trust between users in a social network is integrated into the recommender systems by factorizing the social trust matrix $S$. Moreover, Recommendation with the Social Trust Ensemble (STE) was introduced in [Ma et al. 2009], which is a linear combination of the basic matrix factorization approach [Mnih and Salakhutdinov 2007] and a social network based approach. Moreover, [Jamali and Ester 2010] proposed Social Matrix Factorization (SocialMF), which addresses the transitivity of trust in social networks, as the dependence of a user’s feature vector on the direct neighbors’ feature vectors can propagate through the network, making a user’s feature vector dependent on possibly all users in the network (with decaying weights for more distant users). Beyond that, [Ma et al. 2011] proposed social regularization as to incorporate social network information into the training procedure. They coined the term Social Regularization to represent the social constraints on recommender systems.

1.3 Contributions

In this dissertation, we identify several interesting yet challenging research problems in venture capital investments analysis, and introduce how we address these problems by
utilizing cutting-edge data mining techniques, specifically recommendation systems.

First, we develop an analytical approach to help venture capitalists make decisions on potential investment deals. We adopt recommender system techniques to learn the investment preferences of VCs to recommend the right investment candidates at the screening stage. Our adoption mainly utilizes historical investment records, with additional information, including startups’ locations, industry categories, acquisition records, products, etc. More importantly, the primary goal of venture capital investments is to maximize financial return with portfolio companies exiting via acquisition or going IPO (Initial Public Offering). Thus, expected investment returns and potential risks are critical factors concerned by VCs. Based on modern portfolio theory [Elton et al., 2009], we optimize investment strategy for the purpose of financial returns maximization and investment risks aversion, by ruling out unqualified startups and suggesting right investment amounts.

Along with this work, we found a strong association between venture capital investment and social relationships. We develop a probabilistic latent factor model to predict venture capital investment deals using social connections between members of VC firms and startups. In terms of methodology, we adopt Probabilistic Matrix Factorization (PMF) [Mnih and Salakhutdinov, 2007] framework, which has proved efficient and effective for recommendation-alike problems in the research community of recommender systems. PMF model is extended by incorporating member relationship information which consists of three different types. The first is job title which characterizes the position that the individual holds in the organization. The other two are type of social group and “follow” direction between any two of the members who
share a connection. In other words, social networks of the members are directed, and all social entities (nodes) and connections (edges) can associate with multiple labels. We are the first to utilize member social relationships from VC firms and startups for investment deals prediction.
2.1 Introduction

In this chapter, we develop a novel analytical approach to help venture capitalists make decisions on potential investment deals. We are primarily motivated by two observations. First, in recent years, information about startups and their fund-raising records are accessible much more easily, benefited from several public data holders, like Crunchbase1, SpokeIntel2, Owler3, etc. Declared as the world’s most comprehensive dataset of startup activity, Crunchbase has about 650k profiles of people and companies, in addition to financing history, operating status, and so forth. Another motivator for quantitative venture financing analysis is the advent of advanced analytical techniques and rapid development of computing power. Specifically, the emerging recommender systems have stimulated the abundance of practical applications which help users deal with information overload and provide personalized recommendations

Adomavicius and Tuzhilin 2005. The applications vary from products recommendation at Amazon.com Linden et al. 2003, movies recommendation by MovieLens

1https://www.crunchbase.com/
2https://www.spokeintel.com/
3https://www.owler.com/
The applications of recommender systems in a wide variety of communities manifest their potentials and advantages. Likewise, we can analyze the investment preferences of VCs and recommend them the startups best meeting their interests. Nonetheless, several challenges remain in the application of recommender systems for startups recommendation tasks. First, in the products/movies recommendation cases, one user could rate hundreds of movies whether or not he/she has watched those movies. On the contrary, a medium-size venture capital firm typically closes less than 100 investment deals on average each year. This could lead to high data sparsity which potentially causes the cold-start problem. Besides, compared with explicit rating information in movies recommendation case, we have merely investment amounts from investors to startups as implicit propensity estimates. Third, different from traditional recommendation problem, financial concerns are important in investment deals recommendation problem. In venture financing, investors usually fund the selected companies with a certain amount of money and expect to profit from these investments in several years, typically 5 to 10 years. Thereupon, further consideration of potential return and risk are crucial in our recommendation model.

To deal with the above challenges, we adopt recommender system techniques to learn the investment preferences of VCs to recommend the right investment candidates at the screening stage. Our adoption mainly utilizes the historical investment records, with the incorporation of additional information, including startups’ locations, industry categories, acquisition records, products, etc. More importantly, the primary goal
of venture capital investments is to maximize financial return with portfolio companies exiting via be acquired or going IPO (Initial Public Offering). Thus, expected investment returns and potential risks are critical factors concerned by VCs. Based on modern portfolio theory [Elton et al., 2009], we optimize the investment strategy for the purpose of financial returns maximization and investment risks aversion, by ruling out unqualified startups and suggesting the investment amounts.

The remaining of this chapter is structured as follows. Section 2.2 provides the overview and general idea of our methodology framework. We present the investors’ investment preferences model in Section 2.3 and the further incorporation of investment return and risk in Section 2.4. In Section 2.5, systematical evaluations of our method and several benchmarks are presented and discussed. Related works are discussed in Section 2.6 and finally Section 2.7 concludes our work and envisions possible future works.

### 2.2 Overview

In this section, we give an overview of our analytical investment approach. Generally, our goal is to assist venture capitalists to determine which group of entrepreneurial companies to invest in order to meet their investment preferences as well as generate maximized return by taking the least risks. In the current context, we refer venture capitalists specifically to institutional capitalists, as opposed to individual investors, like business angels. The reasons to aim at institutional capitalists are at least in two folds. On the one hand, in contrast to individual angel investors, venture capital institutions are primary funding sources to startups who signal financial de-
mands. On the other hand, venture capital institutes have more formal investment
decision-making process and retain more accessible investment records [Tyebjee and
Bruno 1984; Fried and Hisrich 1994]. Whereas, individual investors reach their final
investment decisions more informally and have less reliable information disclosed.

For formal institutional venture capitalists, the investment process can be modeled
in five steps: deal origination, deal screening, deal evaluation, deal structuring, and
post-investment activities [Tyebjee and Bruno 1984]. We aim to utilize rich venture
financing data to assist venture capitalists to screen startup candidates and evaluate
potential investment deals.

For screening startup candidates, we adopt the recommender systems techniques
to learn VCs’ investment preferences. The recommendation techniques can collec-
tively model VCs’ historical investment records and provide each VC with the best-
matched investing candidates. Matching VCs’ investment preferences is vital in ven-
ture financing markets because VCs will mentor their portfolio companies after capital
investment and different VCs have fairly different experiences and propensities when
assessing investment deals. Some might possess strong favoritism on Internet com-
panies while some others prefer biotech startups. Certain investors tend to value
team members’ technical skills; whereas others place more weights on managerial
capabilities of the company. Therefore, understanding and better leveraging various
preferences of different venture investors are the keys in recommending right venture
candidates.

For evaluating potential investment deals in venture financing market, since it’s
difficult to observe or estimate the expected returns of venture investments [Lerner
et al., 2007, MacMillan et al., 1989, Sapienza, 1992, Hellmann and Puri, 2002], we resort to startups *exiting probability* as their performance indicator. Generally, as the portfolio startups successfully exit, investors expect to gain financial returns, the amount of which depends on the issued stock price (via IPO) or the buy-out price (via acquisition), and the percentage of shares they hold [Landström et al., 2007]. With a rather accurate estimation of *exiting probability*, one VC can increase (or decrease) its share percentage for better investment outcomes. Here we predict startups future *exiting probabilities* using a nonparametric method. The nonparametric regression method can learn non-linear relationship between random variables and thus improves prediction performance. Plus, the nonparametric method here can simultaneously estimate investment risks.

## 2.3 Matching Investment Preferences

The key idea in matching VCs’ investment preferences using recommender system techniques is to factorize their historical investment records. Specifically, we regard startups and VCs as two distinct groups of entities. Let’s assume there are *M* VCs and *N* startups. Let \( U = \{u_1, u_2, \ldots, u_M\} \) be the set of VCs and \( V = \{v_1, v_2, \ldots, v_N\} \) the set of startups. As widely used in recommendation systems [Mnih and Salakhutdinov, 2007, Schmidt et al., 2009], we model VCs and startups as latent factors. Specifically, VCs are modeled by \( U = (u_1, u_2, \ldots, u_M) \) while startups by \( V = (v_1, v_2, \ldots, v_N) \).

In conventional recommender systems based on latent factor models (e.g., Probabilistic Latent Factor models [Mnih and Salakhutdinov, 2007]), it assumes that the observed funding amount \( y_{ij} \) from VC \( u_i \) to the startup \( v_j \) follows Gaussian distribu-
Figure 2.1: Funding amounts from a random VC in Crunchbase dataset. The left figure is the original distribution of funding amounts while the right one is the distribution after \( \log(\cdot) \).

\[
y_{ij} \sim N(x_{ij}, \alpha^{-1}),
\]

where \( N(\mu, \alpha^{-1}) \) denotes the Gaussian distribution with mean \( \mu \) and precision \( \alpha \).

Here, the distribution mean \( x_{ij} \) is the dot-product of two latent factors \( u_i \) and \( v_j \):

\[
x_{ij} = \langle u_i, v_j \rangle. \tag{2.1}
\]

In our case, however, based on our investigation on data availability and characteristics, we update the model as:

\[
y_{ij} \sim LN(x_{ij}, \alpha^{-1}), \tag{2.2}
\]

\[
x_{ij} = d_{ij} \cdot \left( \langle u_i, v_j + \gamma \cdot c^v_j \rangle + \langle \beta_i, z_j \rangle \right),
\]

where \( LN(\cdot) \) denotes log-normal distribution. The reason we opt for log-normal distribution is that, the log-transformed funding amounts from one VC approximately follow Gaussian distribution, one example of which is illustrated in Figure 2.1.

In the remaining of this section, we illuminate more model details (for \( x_{ij} \) in Equation 2.2), followed by model inference with uncertainty quantification. Table 2.1
Table 2.1: Mathematical Notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U$</td>
<td>$K \times M$</td>
<td>VC latent factor</td>
</tr>
<tr>
<td>$V$</td>
<td>$K \times N$</td>
<td>startup latent factor</td>
</tr>
<tr>
<td>$d$</td>
<td>$M \times N$</td>
<td>$d_{ij}$ is the geographic proximity between $u_i$ and $v_j$</td>
</tr>
<tr>
<td>$C$</td>
<td>$K \times S$</td>
<td>industry category latent factor</td>
</tr>
<tr>
<td>$Z$</td>
<td>$L \times N$</td>
<td>startup explicit properties</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$L \times M$</td>
<td>VC’s propensity to startup properties</td>
</tr>
</tbody>
</table>

summarizes the important mathematical notations.

2.3.1 Model Specification

In comparison with conventional recommender systems (Equation 2.1), we incorporate some new terms in Equation 2.2 to conform to several important perspectives of matching VCs’ investment preferences.

**Geographic proximity** ($d_{ij}$). Venture capital investors have a strong tendency for local businesses. Florida and Smith [1993], Florida and Smith Jr [1992] observed that VC firms in the high-tech market favor the companies geographically close. More than half of the biotech firms in the USA received most fundings from local venture capital firms, according to Powell et al. [2002]. Figure 2.2 shows the distribution of investor-startup geographic distances utilizing the data from Crunchbase. We can
Figure 2.2: Distribution of investor-startup distances in Crunchbase dataset.

observe that the probability of investor \( u_i \) choosing startup \( v_j \) overall decays with increasing distances between them. One particular case is the peak at the distance of 3,000 miles shown in the figure which is mostly the accumulation of the investment fundings from VCs in New York City (in the east coast) to the startups in Bay Area (in the west coast). In this regard, we embed geographic distance as an essential factor in our model. Explicitly, we denote \( l^u_i \) and \( l^v_j \) as the locations of investor \( u_i \) and startup \( v_j \), respectively. Accordingly, the geographic proximity [Liu et al., 2013] between \( u_i \) and \( v_j \) is derived in the form:

\[
d_{ij} = \frac{d_0}{d_0 + \text{dist}(l^u_i, l^v_j)},
\]

where \( \text{dist}(l^u_i, l^v_j) = ||l^u - l^v|| \) is the Euclidean distance and \( d_0 \) is a heuristic parameter.

**Market Specialization (\( c^*_j \)).** As stated by Tyebjee and Bruno [1984], ventures’ market sectors are important factors in the phase of investment deals evaluation. Venture capital firms usually have different propensities in regards to industry divisions due to their fund managers’ diverse areas of specialty. Therefore, we include
the market sector as an important feature for better understanding of capitalists’ investment propensity.

In our dataset, each startup \( v_j \) is associated with a list of industry categories \( \Phi_j \subseteq C \) characterizing its market specialization, where the complete set of industry categories is termed as \( C = \{c_1, c_2, \ldots, c_S\} \). To capture implicit industry preference of VCs, we further assume each industry \( l \in C \) is associated with the latent vector \( c_l \in \mathbb{R}^K \). Then, we compute market specialization of startup \( v_j \) as:

\[
e_j = \frac{1}{|\Phi_j|} \sum_{l \in \Phi_j} c_l, \quad (2.4)
\]

and overall latent factors associated with startup \( v_j \) in Equation 2.2 is

\[
\hat{v}_j = v_j + \gamma \cdot e_j
\]

where \( \gamma \) is a weighting parameter. As such, our model incorporates startups’ market specialization to learn the industry preference of VCs.

**Additional Properties (\( z_j \)).** Aside from the concerns above, we leverage some other properties of startups to enhance our deal recommendation model. The features we utilize include, but not limited to, board members, current and past employees, company acquisition records, and sub-organizations. Table 2.2 lists the properties with their mathematical symbols and corresponding descriptions. As a convention, \( N_{\cdots} \) denotes the variables related to *counts* and \( F_{\cdots} \) refers to the *frequency* of certain events on monthly-base. For example,

\[
F_{\text{acquisitions}} = \frac{N_{\text{acquisitions}}}{N_{\text{month}}},
\]

where \( N_{\text{month}} \) is the number of months from the founded date of the startup as of
today and $V_{\text{monthly\_amount}}$ is defined as the amount of funds received by the startup per month in average. We let $Z$ be the set of all variables, each of which $z_j \in \mathbb{R}^L$ is the vector containing all properties of startup $v_j$, where $L$ is the number of property variables.

### 2.3.2 Model Inference

As a Probabilistic Latent Factor (PLF) model, we also place Gaussian priors on the latent factors of Equation 2.2:

$$
P(U|\mu_U, \Lambda_U) = \prod_{i=1}^{M} \mathcal{N}(u_i|\mu_U, \Lambda_U^{-1}) = \prod_{i=1}^{M} (2\pi)^{-K/2}|\Lambda_U|^{1/2} \exp\left(-\frac{1}{2}(u_i - \mu_U)'\Lambda_U(u_i - \mu_U)\right)
$$

$$
P(V|\mu_V, \Lambda_V) = \prod_{j=1}^{N} \mathcal{N}(v_j|\mu_V, \Lambda_V^{-1}) = \prod_{j=1}^{N} (2\pi)^{-K/2}|\Lambda_V|^{1/2} \exp\left(-\frac{1}{2}(v_j - \mu_V)'\Lambda_V(v_j - \mu_V)\right)
$$

where $\mu_U, \mu_V$ are the Gaussian mean vectors, and $\Lambda_U, \Lambda_V$ are the Gaussian precision matrices, for $U$ and $V$ respectively. Similarly, we assume:

$$
P(C|\mu_C, \Lambda_C) = \prod_{s=1}^{S} (2\pi)^{-K/2}|\Lambda_C|^{1/2} \exp\left(-\frac{1}{2}(c_s - \mu_C)'\Lambda_C(c_s - \mu_C)\right)
$$

$$
P(\beta|\mu_\beta, \Lambda_\beta) = \prod_{i=1}^{M} (2\pi)^{-K/2}|\Lambda_\beta|^{1/2} \exp\left(-\frac{1}{2}(\beta_i - \mu_\beta)'\Lambda_\beta(\beta_i - \mu_\beta)\right)
$$

Then, inferencing in this model can be performed by maximizing the log-posterior over the latent factors $\Theta$ with fixed hyper-parameters $\Omega$, where:

$$
\Theta = \{U, V, C, \beta\}
$$

$$
\Omega = \{\mu_U, \Lambda_U, \mu_V, \Lambda_V, \mu_C, \Lambda_C, \mu_\beta, \Lambda_\beta\}$$
Table 2.2: List of Additional Properties of Startups

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{acquisitions}$</td>
<td>Number of acquisitions</td>
</tr>
<tr>
<td>$F_{acquisitions}$</td>
<td>Frequency of acquisitions</td>
</tr>
<tr>
<td>$N_{rounds}$</td>
<td>Number of funding rounds</td>
</tr>
<tr>
<td>$V_{monthly_amount}$</td>
<td>Monthly funding amount</td>
</tr>
<tr>
<td>$N_{founders}$</td>
<td>Number of founders</td>
</tr>
<tr>
<td>$N_{past_members}$</td>
<td>Number of past members</td>
</tr>
<tr>
<td>$F_{members_quit}$</td>
<td>Frequency of member quitting</td>
</tr>
<tr>
<td>$F_{news}$</td>
<td>Frequency of news on public media</td>
</tr>
<tr>
<td>$N_{investors}$</td>
<td>Number of investors</td>
</tr>
<tr>
<td>$N_{employees}$</td>
<td>Number of employees</td>
</tr>
<tr>
<td>$N_{competitors}$</td>
<td>Number of competitors</td>
</tr>
<tr>
<td>$N_{products}$</td>
<td>Number of products</td>
</tr>
</tbody>
</table>
However, these hyper-parameters are often difficult to determine in making the model
generalize well, particularly due to the high uncertainties on sparse and imbalanced
data sets [Salakhutdinov and Mnih 2008]. Therefore, to improve the model generality
and quantify the data uncertainty, we employ a fully Bayesian inference procedure
of the PLF model using Markov Chain Monte Carlo (MCMC) method, in particular,
the *Gibbs sampling* algorithm [Carlo 2004].

The Gibbs sampling algorithm cycles through the latent variables (including both
$Θ$ and $Ω$), each of which is sampled conditionally on the current values of all other
variables. To make the sampling process easy to implement, we further place on $Ω$
the Gaussian-Wishart priors, which are conjugate with the Gaussian distribution of
$Θ$. Specifically, we let $Ω_θ = \{µ_θ, Λ_θ\}$ for $θ \in Θ = \{U, V, C, β\}$ and:

$$Ω_θ = \{µ_θ, Λ_θ\} \sim N(µ_0|µ_0, (λ_0Λ_θ)^{-1}) \cdot W(Λ_θ|W_0, ν_0).$$

Here $W$ is the Wishart distribution with $ν_0$ degrees of freedom and a $K \times K$ scale
matrix $W_0$. For convenience and symmetry, in our experiments we set $ν_0 = D$, $W_0$
to the identify matrix, $µ_0$ to be all zeros, and $λ_0 = 1$. In this way, the conditional
distribution of $Ω_θ$ is given by the Gaussian-Wishart distribution:

$$P(µ_θ, Λ_θ|θ) = N(µ_θ|µ_0^*, (λ_0^*Λ_θ)^{-1}) \cdot W(Λ_θ|W_0^*, ν_0^*)$$  (2.7)
where

$$\bar{\theta} = \frac{1}{|\theta|} \sum_{i=1}^{|	heta|} \theta_i \quad \bar{\theta'} = \frac{1}{|\theta|} \sum_{i=1}^{|	heta|} \theta_i'$$

$$\lambda_0^* = \lambda_0 + |\theta| \quad \nu_0^* = \nu_0 + |\theta|$$

$$\mu_0^* = (\lambda_0^*)^{-1}(\lambda_0 \cdot \mu_0 + |\theta| \cdot \bar{\theta})$$

$$(W_0^*)^{-1} = (W_0)^{-1} + |\theta| \cdot \bar{\theta'} + (\lambda_0^*)^{-1}\lambda_0|\theta| \cdot (\mu_0 - \bar{\theta})(\mu_0 - \bar{\theta})'$$

The notation $|\theta| = M, N, S, M$ for $\theta = U, V, C, \beta$, respectively.

Now we can derive the conditional distributions in the cyclic Gibbs sampling for $\theta \in \Theta = \{U, V, C, \beta\}$. In particular, we have:

$$P(U|Y, \Theta_{-U}, \Omega_U) = \prod_{i=1}^{M} P(u_i|Y, \Theta_{-U}, \Omega_U)$$

$$P(u_i|Y, \Theta_{-U}, \Omega_U) = N(u_i|\mu_i^*, (\Lambda_i^*)^{-1})$$

where $\Theta_{-U} = \Theta \setminus U$ and

$$\Lambda_i^* = \Lambda_U + \alpha \sum_{j: I_{ij}=1} d_{ij} \cdot \hat{v}_j \hat{v}_j'$$

$$\mu_i^* = (\Lambda_i^*)^{-1} \left( \Lambda_U \mu_U + \alpha \sum_{j: I_{ij}=1} d_{ij}(\log(y_{ij}) - d_{ij}\langle \beta_i, z_j \rangle) \cdot \hat{v}_j \right)$$

where $\hat{v}_j = v_j + \gamma \cdot c_j^i$. Note that, $I_{ij}$ is the indicator function that $I_{ij} = 1$ if we have observed data $y_{ij}$, and otherwise $I_{ij} = 0$. The above sampling of $u_i$ are independent for individual $i \in \{1, \cdots, M\}$, and the overall $U$ can be sampled in parallel. The sampling equations for $V, C$ and $\beta$ are similar and given in Appendix. The overall Gibbs sampling algorithm is shown as follows:

1: Initialize latent factors $\Theta^1 = \{U^1, V^1, C^1, \beta^1\}$

2: for $t \leftarrow 1, 2, \cdots, T$ do
3: Sample $\Omega_{\theta}^{t+1} = \{\mu_\theta^{t+1}, \Lambda_\theta^{t+1}\}$ for $\theta \in \Theta$ with $P(\mu_\theta, \Lambda_\theta | \theta^t)$ in Equation 2.7.

4: Sample $U$ with $P(U | Y, \Theta_{-U}^t, \Omega_U^t)$ in Equation 2.8.

5: Sample $\beta$ with $P(\beta | Y, \Theta_{-\beta}^t, \Omega_\beta^t)$ in Equation 5.1.

6: Sample $V$ with $P(V | Y, \Theta_{-V}^t, \Omega_V^t)$ in Equation 5.2.

7: for $s \leftarrow 1, 2, \cdots, S$ do

8: Sample $c_s$ with $P(c_s | Y, \Theta_{-c_s}, \Omega_C)$ in Equation 5.3.

9: end for

10: end for

2.3.3 Recommendation with Bayesian PLF Model

Suppose we have estimated the parameters of PLF model, we can compute the expected latent preference $x_{ij}$ in Equation 2.2 of VC $u_i$ for the startup $v_j$:

$$x_{ij} = \frac{1}{T} \sum_{t=1}^{T} x_{ij}^t,$$

where $x_{ij}^t$ is computed with $\Theta^t$ in the Gibbs Sampling Algorithm 3.3.4. Then for a specific VC $u_i$, we can sort the startups $V = \{v_1, v_2, \cdots, v_N\}$ with respect to the values $x_{ij}$ in descending order. The candidates ranked at the top will be recommended to the VC.

2.4 Optimizing Investment Return and Risk

In traditional recommendation task, generating a list of item candidates based on users’ preferences or tastes is the sole objective. However in our problem, recommending startups to investors solely considering their propensities seldom meet their objectives which are mostly to maximize financial returns with suppressed potential...
risks. Thus, in this section, we optimize the investment’s expected return and risk with the modern portfolio theory from Wang [2009].

In this regard, supposed that $\Pi_i$ is the set of startups recommended to VC $u_i$, we optimize the investment proportion $\omega_j$ for startup $v_j \in \Pi_i$. Subject to $\omega_j \geq 0$ and $\sum_j \omega_j = 1$, the optimization objective is to maximize the risk-averse expected return:

$$O_i = \text{Return}_i - \tau \text{Risk}_i,$$

(2.9)

where $\tau$ is a heuristic parameter, and

$$\text{Return}_i = E(\Pi_i) = \sum_{j: v_j \in \Pi_i} \omega_j r_j,$$

(2.10)

$$\text{Risk}_i = \text{Var}(\Pi_i) = \sum_{j: v_j \in \Pi_i} \omega_j^2 \sigma_j^2.$$

Note that, the negligible correlation between different startups are ignored.

In Equation 2.10, the variable $r_j$ indicates the success/failure of startup $v_j$,

$$r_j = \begin{cases} 1 & \text{if IPO or acquisition}, \\ 0 & \text{if closed}, \end{cases}$$

(2.11)

where $j = 1, 2, \cdots, N$. For some startups that we have unknown success/failure indicator as they are still private, their exiting probabilities will be estimated. To this end, we train a predictive model using the company properties in Table 2.2. Specifically, we employ kernel regression [Nadaraya, 1965, Guo et al., 2016], which can learn the non-linear relationship between random variables. Given the observed $r_j$ of $N^{obs}$ startups, the performance of a new startup is estimated as

$$\hat{r} = \sum_{j=1}^{N^{obs}} \kappa(z, z_j) r_j,$$

(2.12)
where

\[ \kappa(z, z_j) = \frac{K(\frac{z - z_j}{h})}{\sum_{k=1}^{N_{obs}} K(\frac{z - z_k}{h})}, \]  

(2.13)

where the parameter \( h (h > 0) \) is called the “bandwidth” which determines the proportion of local versus remote information used in the summation, and \( K(\cdot) \) is a kernel function which, in our case, is specified as multivariate Gaussian,

\[ K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}||x||^2). \]  

(2.14)

Using kernel regression has several benefits: 1) it is a weighted non-linear instance-based regression method, accompanied with regression uncertainty measures; 2) it regresses class labels by assigning higher weights on more similar startups controlled by customizable kernels; 3) the parameter \( h \) in the kernel regression can be automatically learned based on the observations [Nadaraya, 1965].

It is rather straightforward to use the weighted average of similar startups’ performance to predict the investment return of a new startup. However, for risk prediction, we cannot take such a weighted average directly, since the risk values of past startups are not quantified. As a result, we quantify the investment risk of a new startup via the weighted variance:

\[ \hat{\sigma}^2 = \sum_{j=1}^{N_{obs}} \kappa(z_j) (\hat{r} - r_j)^2. \]  

(2.15)

### 2.5 Results and Discussions

In this section, we present our experimental results in detail. We first introduce our test dataset from Crunchbase. Then we list several baseline algorithms for performance comparison. Detailed experimental settings are summarized afterwards,
followed by performance evaluations of our proposed and baseline algorithms.

2.5.1 Dataset

The dataset used in our work comes from Crunchbase, the world’s most comprehensive dataset of entrepreneurial activities. There are several advantages of using this dataset compared to other alternatives. The first is accessibility. Subject to the limit of API requests frequency, all of its data is accessible to the public with no charge. Second, it has various information about startup and entrepreneurial financing, including companies, people, financial organizations, funding rounds, acquisitions, products, news articles, and so on. Third, it contains a large volume of startup-related data since people in the web community can collaboratively contribute to its content. This character, however, could not guarantee data density due to the inherent randomness of information generation, which hence requires further data processing and analytics.

There are different kinds of information about startups in this dataset, which is mainly grouped into two divisions, basic properties and historical investment records. In the data about startup basic properties, startup name serves as the only identifier for each company. Offices locations are denoted as, in order, country - region - state - city. Industry categories are formed by a list of industry-specific terms. Other than these, the dataset also provides details about the company’s board and executive members, competitors, products, acquisition records, and operating public websites/media. Regarding investment records, more information is available such as venture capital investors, funding rounds, funding amount at each round from each investor, and so forth. Generally, one startup would have multiple funding rounds.
At each round, multiple VCs are engaged, and each provides certain amount of money to constitute one-round capital. In some cases, one single VC can meet the startup’s financing demand by supplying sufficient capitals. Along with funding amount information, the dataset provides the date when funding rounds are closed.

We make use of the data from Crunchbase with around 24k investors, 55k startups and 95k investment deals. We have the following data pre-processing procedures to clean the raw data for later-on analysis. First, the dataset covers startups and investors across the world, 85% of which are within USA. As we learned that USA-based startups have more comprehensive information compared with those outside, we limit our data scope on only USA-based companies. Then, we prune entries with too many missing values and retain only the investors having made at least three investment deals for data sparsity reduction. As a result, we reduce the size of data down to 1,467 VCs, 4,007 startups and 17,485 investment deals in total. The corresponding investor-startup matrix sparsity is 99.7%, which is still relatively high compared with that in common recommendation scenarios.

2.5.2 Implementation Details for Matching Investment Preferences

For experiments, we split our dataset into training, validation, and testing sets and evaluating our algorithms from different perspectives. To simulate the real use case, which is to recommend startups to VCs at specified dates, we cannot directly adopt conventional cross-validation scheme on our temporal-sensitive data. Otherwise, the information of some investments in later rounds might have been used for training model to predict the investments in earlier rounds, which goes against reality. There-
fore, we sort investment records in temporal order. Then three pairs of training-validation sets are generated by setting the proportion of training-validation sets 7:3, 8:2, and 9:1, respectively. On the other hand, the testing set is considered unseen data, used for model evaluation.

Our proposed model (BPLF+Portfolio) is integrated with preference (by Bayesian Probabilistic Latent Factor model, BPLF, in Section 2.3) from a portfolio perspective (by Portfolio optimization in Section 2.4). We compare our model with other state-of-the-art recommendation methods in learning VCs’ preference in the investment decision-making process. The first is Singular Vector Decomposition (SVD), which is a well-known method for matrix factorization that provides the best lower rank approximations of the original matrix [Adomavicius and Tuzhilin, 2005, Sarwar et al., 2000]. We also include an item-based collaborative filtering algorithm (ItemCF) [Sarwar et al., 2001] to study the performance difference between memory-based and model-based recommendation techniques. Besides, we take into account non-negative matrix factorization (NMF) [Lee and Seung, 2001] and probabilistic matrix factorization (PMF) model [Mnih and Salakhutdinov, 2007], both of which have found plenty of applications in recommendation scenarios. In addition, we incorporate two naive algorithms, BestU (best VC) and BestV (best startup). To put it briefly, BestU recommends the startups invested by the most successful VC while BestV recommends the startups with the most successful investment return records.
2.5.3 Performance Evaluation for Matching Investment Preferences

We evaluate our proposed algorithm as well as other state-of-the-art algorithms from different perspectives. The first metric is the prediction accuracy. As mentioned earlier, investors’ implicit investment propensities on certain startups are quantified by funding amount. Accordingly, we use RMSE (Root of the Mean Square Error) and MAE (Mean Absolute Error) to measure how well the specific algorithm predicts the investment amount from particular investor to a specific startup, defined as follows:

\[
\text{RMSE} = \sqrt{\frac{\sum_{i,j} (y_{ij} - \hat{y}_{ij})^2}{N_{all}}}, \\
\text{MAE} = \frac{\sum_{i,j} |y_{ij} - \hat{y}_{ij}|}{N_{all}},
\]

where \(y_{ij}\) and \(\hat{y}_{ij}\) denote the observed and predicted value respectively, and \(N_{all}\) is the number of all test instances.

Table 2.3: RMSE and MAE of our algorithm and baseline algorithms.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SVD</th>
<th>NMF</th>
<th>ItemCF</th>
<th>PMF</th>
<th>BPLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.7574</td>
<td>0.6995</td>
<td>0.8426</td>
<td>0.6774</td>
<td>0.6163</td>
</tr>
<tr>
<td>MAE</td>
<td>0.5516</td>
<td>0.5376</td>
<td>0.6075</td>
<td>0.5202</td>
<td>0.4592</td>
</tr>
</tbody>
</table>

Table 2.3 presents the results of all algorithms. As shown, our BPLF gives the best overall prediction performances regarding RMSE and MAE. Among all, ItemCF presents the poorest performance as it only makes use of local information (the most similar items). Note that, SVD model was trained to minimize sum-of-squared error of the observed entries in the target matrix \(Y\), with no regularization on the
latent factors. Thus SVD model can be seen as a special case of our model trained using maximum likelihood (ML) point estimation. With fixed prior regularizations, PMF model computes MAP (maximum a posteriori probability) point estimation. While the regularizations on the latent factors in PMF somehow improve testing performance compared to SVD, the optimal parameters in the prior regularizations are expensive to search. In BPLF, the Bayesian estimation procedure integrates out such regularization parameters which leads to less overfitting and better modeling generality. In other words, by averaging over all sampling estimations that are compatible with the data, BPLF deals with uncertainty more effectively than the non-Bayesian alternatives. NMF also gives poor performance for one reason that the non-negative constraint is not optimal to explain our data, which can be well modeled with log-normal distribution as proposed in our method.

The second concern in model evaluation is the algorithm’s performance in Top-K recommendation. We employ the commonly-used recommendation metrics, including Precision@K, Recall@K and MAP (Mean Average Precision) [Stone, 2014]. In our context, Precision@K, Recall@K and MAP [Christopher et al., 2008] are defined as

\[
\text{Precision@K} = \frac{N^{\text{rec}}}{K}, \quad \text{Recall@K} = \frac{N^{\text{rec}}}{N^{\text{test}}},
\]

\[
\text{MAP} = \frac{1}{M^{\text{test}}} \sum_{i=1}^{M^{\text{test}}} \left( \frac{1}{N_i^{\text{rec}}} \sum_{j=1}^{N_i^{\text{rec}}} \text{Precision@Rank}_{ij} \right),
\]

(2.17)

where \(N^{\text{rec}}\) is the number of recommended startups which are indeed invested by the VC, \(N^{\text{test}}\) is the number of startups appeared in the test set which are invested by the VC, and \(M^{\text{test}}\) is the number of VCs in the test set. Besides, \(\text{Rank}_{ij}\) denotes the rank of startup \(v_j\) in the recommendation list for investor \(u_i\). Thus, MAP is interpreted as
Table 2.4: Precision and Recall comparisons of our algorithm and baseline algorithms.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>@N</th>
<th>SVD</th>
<th>NMF</th>
<th>ItemCF</th>
<th>PMF</th>
<th>BestU</th>
<th>BestV</th>
<th>Portfolio</th>
<th>BPLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec</td>
<td>@1</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td><strong>0.0023</strong></td>
</tr>
<tr>
<td></td>
<td>@5</td>
<td>0.0005</td>
<td>0.0002</td>
<td>0.0009</td>
<td>0.0005</td>
<td>0.0021</td>
<td>0.0000</td>
<td>0.0021</td>
<td><strong>0.0035</strong></td>
</tr>
<tr>
<td></td>
<td>@10</td>
<td>0.0006</td>
<td>0.0002</td>
<td>0.0008</td>
<td>0.0002</td>
<td>0.0012</td>
<td>0.0005</td>
<td>0.0011</td>
<td><strong>0.0032</strong></td>
</tr>
<tr>
<td>Rec</td>
<td>@1</td>
<td><strong>0.0012</strong></td>
<td>0.0001</td>
<td>0.0002</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>@5</td>
<td>0.0013</td>
<td>0.0001</td>
<td>0.0011</td>
<td>0.0008</td>
<td>0.0029</td>
<td>0.0000</td>
<td>0.0029</td>
<td><strong>0.0042</strong></td>
</tr>
<tr>
<td></td>
<td>@10</td>
<td>0.0031</td>
<td>0.0001</td>
<td>0.0020</td>
<td>0.0008</td>
<td>0.0029</td>
<td>0.0030</td>
<td>0.0029</td>
<td><strong>0.0083</strong></td>
</tr>
</tbody>
</table>

the integrated average recommendation precision for each VC.

In addition, we utilized NDCG (Normalized Discounted Cumulative Gain) as an evaluation metric to measure ranking performance of recommendation algorithms. Specifically, the discounted cumulative gain (DCG@K) is given by

\[
DCG_K = rel_i + \sum_{i=2}^{K} \frac{rel_i}{\log_2(i)},
\]

(2.18)

where \(rel_i\) is the graded relevance of the result at position \(i\). Accordingly, NDCG@K is defined as

\[
NDCG_K = \frac{DCG_K}{IDCG_K},
\]

(2.19)

where \(IDCG_K\) is the ideal \(DCG_K\).

Note that, in this experiment, we focus on evaluating the algorithm’s capability of recommending non-invested startups, which means the previously-invested startups

---

4http://en.wikipedia.org/wiki/Discounted_cumulative_gain
Table 2.5: MAP and NDCG@$N_{test}$ of our algorithm and baseline algorithms.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SVD</th>
<th>NMF</th>
<th>ItemCF</th>
<th>PMF</th>
<th>BestU</th>
<th>BestV</th>
<th>Portfolio</th>
<th>BPLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.0030</td>
<td>0.0012</td>
<td>0.0024</td>
<td>0.0017</td>
<td>0.0036</td>
<td>0.0030</td>
<td>0.0029</td>
<td>0.0053</td>
</tr>
<tr>
<td>NDCG@$N_{test}$</td>
<td>0.1156</td>
<td>0.1107</td>
<td>0.1178</td>
<td>0.1136</td>
<td>0.1234</td>
<td>0.1236</td>
<td>0.1181</td>
<td>0.1239</td>
</tr>
</tbody>
</table>

have been excluded. Table 2.4 lists precision and recall measures of top-k recommendation by each algorithm. From the results, we see that our method BPLF outperforms other baseline algorithms with a significant margin. SVD, NMF, ItemCF and PMF perform much poorer compared with BPLF for one main reason that they solely depend on the investment record matrix of investors and startups while our proposed model BPLF is enhanced with rich additional information. Table 2.5 presents the MAP and NDCG measures of all models. The overall results indicate that our portfolio-free model can characterize investors’ decision-making model generally better.

In Table 2.4 and Table 2.5, for comparison purpose, we further include two return-oriented algorithms, BestU and BestV, as well as one portfolio-based algorithm Portfolio. We can see that these algorithms perform slightly better compared with other portfolio-free algorithms, which may imply that selecting potentially high return startups partially complies with VCs’ investment decision-making preferences. However, compared with our context-enhanced BPLF model, they inevitably reveal their weakness in explaining investors’ investment preferences.
2.5.4 Managerial Implications on Investor’s Propensities

To aim for better understanding VCs’ propensities on different properties of startups at the stage of investment deals evaluation, in this section we study the estimated value of $\beta$. Figure 2.3 shows the box plot of the estimated values of $\beta$ at all Gibbs sampling values. We identify several interesting findings. At one point, $V_{\text{monthly,amount}}$ (defined in Table 2.2) appears to be the most important factor for investors’ decision making. The observation is sound since sufficient monthly-averaged funding amount received by the startup plays an essential role in implying that the startup operates well. $N_{\text{acquisitions}}$ appears to be the least decisive aspect in assessing investment deals. On the one hand, it might hurt the startup’s cash flow if the acquisition activities are too aggressive. On the other hand, few startups has acquisition records, resulting in a large portion of missing values, which makes this indicator rather plausible. Interestingly, we see the factor $N_{\text{founders}}$ has less influence on investors’ decision as its corresponding parameter is close to 0. It is reasonable since usually the development of the startup depends more on the capabilities of its founders, rather than the headcount. We can also tell $N_{\text{news}}$ has a positive impact on investors’ decision. More positive exposure on public media implies the startup’s prosperous growth. Regarding $N_{\text{competitors}}$, we see that investors prefer the startup with more competitors as more competitors may convey the message that the startup resides in a booming market.
2.5.5 Portfolio Optimization for Investment Strategy

Table 2.6 gives the expected investment return and the potential risk of several algorithms. As seen, our portfolio-free algorithm BPLF and BestU present relatively poor performances. The performance of BPLF is following expectation as it fails to incorporate either return or risk in the model. On the other hand, compared to other algorithms with better performances, the poor performance of BestU might attribute to its indirect modeling of return and risk. As we can anticipate, BestV shows the best performance in terms of return, as well as, not surprisingly, its poor capability of suppressing potential risk, since it solely maximizes return with little concern on risk. In contrast, Portfolio has better capability of reducing risk and maintains desirable
Table 2.6: Investment return and risk comparison.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>BPLF</th>
<th>BestU</th>
<th>BestV</th>
<th>Portfolio</th>
<th>BPLF+Portfolio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Return</td>
<td>0.7891</td>
<td>0.7825</td>
<td>0.8899</td>
<td>0.8508</td>
<td>0.8365</td>
</tr>
<tr>
<td>Risk</td>
<td>0.00007</td>
<td>0.0011</td>
<td>0.0034</td>
<td>0.0000</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

return. Although our proposed algorithm BPLF+Portfolio does not supply the best performance in terms of return and risk, it remains at the top tier.

Recall that, in Table 2.4 and Table 2.5, BestU, BestV and Portfolio are inherently incapable of capturing VCs’ investment preferences which our proposed model captures well. Overall, we argue again that investment preference modeling is as important as that of investment return and risk. The theory behind is that, in addition to providing financing support, the investors play additional roles influencing developments on the start-up organizations, such as devoting their knowledge and contacts, offering strategic and operational advice, helping optimize human capital structure, and planning stock options. Therefore, we conclude that investors have subjective investment preferences significantly affecting their investment decisions.

2.6 Literature Review

In our research, we provide an integrated method to determine the next startup to invest by considering investor’s preference and startup’s potential possibility to successfully exit. Thus, our investigation on early research is grouped into two categories. One category consists of traditional analysis on how venture capital investors make
their investment decisions. Another category focuses on the popular recommendation techniques and startups performance estimation methods as well as how they can help investors on decision-making in venture financing markets.

2.6.1 Investment Decision-making Process

In the first category, analyses on decision-making behaviors of venture capitalists has been in active development at both micro-level and macro-level \cite{Cornelius2006, Landstrom2007}. At the micro-level, research work is rooted in management and entrepreneurship field \cite{Tyebjee1984, MacMillan1986, Robinson1988}. From management and entrepreneurship perspectives, there are three main research strands \cite{DaRin2011}, including: 1) questions dealing with interactions between startups and VC (communication, investment, etc.); 2) questions related to interactions between VC fund and its investors (fund-raising, compensation structure, etc.); and 3) questions about the organization of VC firms. As highlighted earlier, we bounded our scope on the research problems in regards to venture capital investment analysis (the 1st strand). In this vein, \cite{Tyebjee1984} first described a 5-step model for VC investment process, which are deal origination, deal screening, deal evaluation, deal structuring, and post-investment activities. Based on two datasets collected through telephone survey and by mail asking VCs to rank the importance of various criteria, they further articulated four critical factors that venture capitalists normally counted on: market potential, management, competition and product feasibility. Later, \cite{Fried1994} interviewed 18 representative startups and provided subsequent structured questionnaires to identify
industries, geographic location and the stage of financing as important elements in VCs’ decision model during firm-specific screening stage. However, Zacharakis and Meyer [1998] pointed out that the studies relying on interviews and/or questionnaires are subject to rationalization and post hoc recall biases.

To address this issue, research work, like Sandberg et al. [1988], Hall and Hofer [1993], Zacharakis et al. [1999], utilized verbal protocols which are real time experiments where VCs “think aloud” as they were screening a business plan [Ericsson and Crutcher, 1991]. Verbal protocols can capture the criteria that VCs considered while sieving companies and allow researchers to discover the importance of different criteria by observing how VCs ranked different factors and how much time they devoted on each [Landström et al. 2007]. Conjoint analysis Shepherd and Zacharakis [1997] is generally employed to study the verbal protocols by decomposing the decision making process into underlying structure, which allows investigators to gain a deeper understanding. Nonetheless, studies based on verbal protocols and conjoint analysis suffer the deficiency of small sample size due to the considerable time and efforts required when gathering information.

Regarding the work at macro-level, the researchers are interested in applied analysis of finance and economics Cooper and Carleton [1979], Chan [1983], Sahlman and Stevenson [1986], Jeng and Wells [2000] studied a variety of determinants of venture capital funding, including gross domestic product (GDP) and market specialization growth, labor market rigidities, etc., by utilizing linear regression models. Indeed, we agree that venture capital industry is not isolated from the macro economic/financial environment. Macro economic environment will have impacts on venture capital
market as a whole, such as the amount of capital into the market, the proportion
of successful startups, etc. Nonetheless, our work mostly focuses on the study of
investors’ investment preference, i.e. which startups the investors are most likely to
invest and in the meantime gain the highest potential return. In other words, macro
economic effects are weighted rather lower in our recommendation model.

In addition, Kaplan and Strömberg [2000] analyzed how venture capitalists choose
investments by taking into account different factors, such as the market size, the strat-
egy, the management team, etc. However, their analysis was more qualitative and
descriptive, in which the conclusion cannot be well generalized and thus is less con-
vincing. Baron and Markman [2000] investigated what makes some entrepreneurs
more successful than others in starting new ventures. They found that a high level of
social capital built on direct personal contacts is helpful to entrepreneurs in gaining
access to venture capitalists. Potentially, incorporating the investors’ and startup
members’ background and relationship information, such as the relationship of being
classmates in MBA program, into the prediction model can bring benefits. Such back-
ground information might unveil the implicit causes leading to the success/failure of
some investment deals. However, we decided not to take such information into con-
sideration primarily due to its insufficiency in our main data source, i.e. Crunchbase.
After all, such fine-grained information is privacy-sensitive and therefore difficult to
gather for a large volume of investors and startups.
2.6.2 Startup Recommendation Approaches

The traditional research follows a natural progression \[\text{Landström et al., 2007}\], from discovering decision criteria of venture financing via post hoc surveys to understanding how information is utilized during actual decision and the relative importance of different criteria. However, the characteristics of traditional methods, such as low information utilization and simplistic analytical techniques, confine their popularities and advances.

Therefore, advanced and sophisticated information processing techniques have a strong demand given that more and more venture capital investment information becomes accessible. Opportunely, with accelerated development and improvement for decades, recommendation engine becomes a sophisticated tool to screen and evaluate redundant information in various cases. The startups screening and evaluation processes to VCs are similar to have startups selected from a recommendation perspective. Plus, recommendation system has demonstrated its effectiveness via a variety of practical applications \[\text{Linden et al., 2003, Miller et al., 2003, Billsus et al., 2002}\].

More specifically, in the category of research on modern recommendation techniques, data mining and predictive analytics are utilized to optimize the expected outcome of venture capital investments. These research work can be further divided into two groups. One small group is startup-oriented approaches, intended to predict successful exits (i.e., being acquired or going public) of startups. For example, \[\text{Xiang et al., 2012}\] utilized profiles and news articles of companies and people to predict Merger and Acquisition (M&A) activities using topic modeling techniques. This type
of approach made use of investment utilities but were incapable of personalized recommendation. The other group of methods is investor-oriented, which aims to learn the investors’ decision making model or predict their overall investment performance by identifying successful investors. Gupta et al. [2015] proposed the InvestorRank method to identify successful investors by understanding how an investor’s collaboration network change over time. The key idea is to measure if an investor gets increasingly close to some known exemplar successful investors as time passes. Stone et al. [2013] conducted an empirical study on recommending top-n startups for VC firms, which is the pioneering work most close to ours. They tested several collaborative filtering based recommendation algorithms on VentureSource dataset and demonstrated the effectiveness of recommender systems in this application scenario. In Stone [2014], the authors extended his work to propose an SVD++-like algorithm with the incorporation of industry categories and company profiles information. Nonetheless, their work, by using SVD-based method, may suffer the over-fitting problem as they did not regularize latent factors and quantify the uncertainties in the model fitting. In contrast to our work, they failed to take investment expected return and risk into consideration.

2.7 Summary

We developed a unified model for the task of investment portfolio recommendation for venture capitalists by considering investors’ investment preferences, expected returns and potential risks. Our method first learned investors’ investment decision-making propensities by adopting Bayesian Probabilistic Latent Factorization framework with
the incorporation of additional investment information, including market specialization, geographic location, and other startup-specific properties. Beyond that, we accommodated modern portfolio theory, concerning expected investment returns and the corresponding potential risks, to enhance our utility function for quantifying the rank of startups in the recommendation list. With extensive experiments on the Crunchbase dataset and comprehensive comparison with other state-of-the-art algorithms on various evaluation metrics, our method manifested the best performance overall.

There are several advantages of our proposed approach. First, we utilized hybrid recommendation approach by considering both collaborative constraints and contextual information of startups, which alleviate the problems that each part may possess. Second, we integrated both investors’ preferences and portfolio utilities in a unified manner which provides additional parameter tuning flexibility as a user-friendly benefit. Third, our work demonstrated the effectiveness of accommodation of probabilistic factor models in addressing investor-specific startups recommendation problem, which paves the way for future research in this direction.
CHAPTER 3

TO BE OR NOT TO BE FRIENDS: EXPLOITING SOCIAL TIES FOR VENTURE INVESTMENTS

3.1 Introduction

Over the years, researchers in finance and management communities remain great interests in discovering the key factors that are highly associated with venture capitalists’ investment decision-making process. Other than the work focusing on characteristics of products/services, market backgrounds, financial dynamics, geographic location of startups, etc. (Tyebjee and Bruno [1984], Berchicci et al. [2011]), a large portion of research paid significant attention on entrepreneurial teams. Tyebjee and Bruno [1984] pointed out entrepreneur’s personality and prior experience are assessed in particular by venture capitalists while they are reviewing investment deals. Vogel et al. [2014] studied how the startup team diversity affects investors’ decision-making. These work showed that it is of great importance for investors to well know the startup team members, and vice versa. However, the above studies, mostly utilizing verbal protocols and conjoint analysis, can suffer the deficiency of small data sample due to the considerable time and efforts being required in gathering information.

Recently, information about startups and their fund-raising records are more ac-
cessible, benefited from several public data holders, like Crunchbase\footnote{https://www.crunchbase.com/}, SpokeIntel\footnote{https://www.spokeintel.com/}, Owler\footnote{https://www.owler.com/}, etc. Crunchbase, declared as the world’s most comprehensive dataset of startup activity, has about 650K profiles of people and companies, in addition to financing history, operating status, and so forth. Besides, the prosperity of online social communities (e.g., Facebook\footnote{https://www.facebook.com/}, Twitter\footnote{https://twitter.com/}, About.me\footnote{https://about.me/}) provided the foundation for tackling this issue. The last concern is how to close the information gap between venture capital investments and social relationship.

In this regard, we propose a novel approach to study the association between venture capital investment and social relationships. Specifically, we develop a probabilistic latent factor model to predict venture capital investment deals using social connections between members of VC firms and startups. The model is unique in the following ways. Unlike other state-of-the-art work Ma et al. [2008], Jamali and Ester [2010] in the field of social recommender systems, we do not approach the problem by exclusively accessing the social connections within the network of investors and quantifying the information as social influences. Plus, we do not exploit the so-called social connection between institutions, as proposed in Eugene and Yuan [2012], Yuxian and Yuan [2013], since the definition of organizational social connection is vague if no actual individual social relationship is retained. Instead, we approach this problem in a unique way – by directly utilizing social connection information between

\footnote{1https://www.crunchbase.com/} \footnote{2https://www.spokeintel.com/} \footnote{3https://www.owler.com/} \footnote{4https://www.facebook.com/} \footnote{5https://twitter.com/} \footnote{6https://about.me/}
members from both parties, namely, VC firms and startups. To the best of our knowledge, we are the first to employ social information between venture capital firms and entrepreneurial companies for venture capital deals prediction and recommendation.

In terms of methodology, we adopt the Probabilistic Matrix Factorization (PMF) \cite{Mnih2007} framework, which has proved efficient and effective for recommendation-alike problems in the research community of recommender systems. We extend PMF model by incorporating member relationship information which consists of three different types. The first is job title which characterizes the position that the individual holds in the organization. The other two are the type of social group and the “follow” direction between any two of the members who share friendship. In other words, the social network of the members are directed, and all the social entities (nodes) and connections (edges) can associate with multiple labels. To sum up, our work possesses three contributions, as follows.

- We are the first to utilize member social relationships from VC firms and startups for venture capital investment deals prediction. In the research field of recommender systems, our problem setting is rather unique: we have member sets associated with the VC (the “user” in conventional recommender systems) and the startup (the “item” to be recommended). Not only are the members tagged (with job titles), but also their connections are labeled and directed. Our model provides effective recommendations by integrating these complicated information with historical investment records.

- We are the first to generate the dataset with member connection via a third-
party online social network communities (About.me) for VC firms and startups data from Crunchbase. Note that, the information of social connections between the firm members are not available in the data from Crunchbase. We thus downloaded 64K social profiles from the platform About.me, which provides rich information of social profiles/connections of its users. After entity matching, we extract 1.3K social profiles for the members in the Crunchbase dataset.

- The effectiveness of our proposed model is assessed with extensive empirical studies. We first use the synthetic data to analyze the relationship between our algorithmic performances and the data properties. We then employ real-world data to demonstrate the advantages of our approach in comparison with competing methods. In addition, the modeling results can provide intuitive managerial implications and actionable knowledge for venture capitalists and startups. Overall, our model can capture and quantify the shared “trust” between the entities in venture financing market.

This chapter is organized as follows. Section 3.2 presents the statistical evidence of our motivations and the overview of our proposed method. Section 3.3 presents the details of our model, including the model specification and inferences. Section 3.4 introduces our synthetic and real-world data for experiments as well as the evaluation results. The related work is presented in Section 3.5 and finally Section 3.6 concludes our work and envisions possible future work.
3.2 Preliminary

Our key idea is illustrated in Figure 3.1, the CoRec (Connected Recommendation) system. Suppose that we have a venture capital firm \((u)\) and a potential portfolio company \((v)\). We need to determine whether the investor \((u)\) should finance the company \((v)\) who is actively seeking funding. In our model, we exploit the members and their social relationships from each organization. The members are shown as the avatars at the bottom of the figure. Briefly, we consider directed connection between members \((e^u)\) from \(u\) (investor) and members \((e^v)\) from \(v\) (startup) as a “trust” relationship. More specifically, we utilize three different types of information from this social network. The first is the job title of each member in his/her organization (VC firm or startup), shown as different icons in the figure. Besides, we differentiate various social groups where the social relationship is retained. The distinct social groups are illustrated as curves in different colors; for example, orange curve means Twitter followers, red curve Facebook friends, blue curve Google+ readers, etc. The third type of embedded social information is the direction of the connection, which is denoted by the solid line (any connection from \(e^u\) to \(e^v\)) and dotted line (any connection from \(e^v\) to \(e^u\)) in the figure.

We believe that such social information is a significant indicator to potential investment deals. This hypothesis is confirmed in Figure 3.2. In this figure, we illustrate the investment statistics for the real-world data extracted from Crunchbase and About.me (refer to Section 3.4 for more details). The statistic of interest is the percentage of investment, defined as the number of investment deals \(L\) divided by
$N \times M$, where $N$ and $M$ are the number of VC firms and startups, respectively. In Figure 3.2, the leftmost marker summarizes the overall investment data from Crunchbase, in which the percentage of investment is around 0.3%. Then, by exclusively considering the directly connected members of VC firms and startups, the percentage of investment is boosted to 2.7%, as indicated by the rightmost marker. We also consider extended connection between two members. Two members have an extended connection if there exists another member connected with both of them. For such members with extended connections, the percentage of investment is around 1%, still considerably higher than that of the overall data.

With the above observations and motivations, however, several challenges need to first be coped with in order to effectively model the social relationship for investments
Figure 3.2: The *percentage of investment* in different data samples.

- Both the observed investment records and social connections are quite sparse in the collected data. Consequently, the developed model would be fitted with only the observed data while a large portion of entries are missing in the VC-startup investment matrix. Moreover, regularization should be included in the modeling process to avoid *over-fitting* issue.

- The statistics discussed above show that the connections between members from different parties are related to investment decisions and outcomes. However, members are associated with various labels. The influence of different member types in the investment decision-making process may be prominently different, thus should be properly quantified in the model.

- In addition, the friendship connections between members are directed and also labeled. In tuition, the direction of such social connections is critical in predicting the future decisions. For example, a VC member following a startup member signifies proactive tendency for investments, while the VC member fol-
lowed by one startup member is rather passive. The model has to capture such
difference to provide accurate investment predictions and recommendations.

In the following section, we formulate the recommendation problem and develop our
model to address these challenges.

3.3 Methodology

In this section, we present our proposed model for venture capital investments predic-
tion by incorporating the social relationship information of members from VC firms
and startups.

3.3.1 Notations of Social Information

We have a set of VC firms \( U = \{u_1, \ldots, u_N\} \), a set of startups \( V = \{v_1, \ldots, v_M\} \)
and member information of these VC firms and startups. Specifically, \( u_n, v_m \in \mathbb{R}^{K \times 1} \)
denote latent vectors, which represent VC’s and startup’s latent preferences,
respectively. We use \( E \) to denote the set of all members, and we have collected their
social information.

**Member Titles.** Each VC firm \( u \in U \) or startup \( v \in V \) consists of a group of
members, respectively. We let \( e \in u \) (or \( e \in v \)) if the individual \( e \in E \) is a member
of VC firm \( u \) (or startup \( v \)). The membership in our dataset is also annotated by
labels. We use \( F(e|u) \) (for VC firm \( u \)) and \( G(e|v) \) (for startup \( v \)) to denote the label
information. For instance, if the person \( e \) is the founder of startup \( v \) and acting as
the CTO, we let

\[
G(e|v) = \{\text{Founder}, \text{CTO}\}.
\]
Table 3.1: The set of membership and connection labels.

<table>
<thead>
<tr>
<th>Set</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{C})</td>
<td>Twitter, Facebook, Insp, Love, Comp, Boldstart, Fave</td>
</tr>
<tr>
<td>(\mathcal{F})</td>
<td>Founders/CEO, Partner, Managing Director, CFO, Others</td>
</tr>
<tr>
<td>(\mathcal{G})</td>
<td>Founders/CEO, Partner, Chairman, CFO, Director, VP, CTO, Head, CPO, COO, Others</td>
</tr>
</tbody>
</table>

In our dataset, the universal membership label set is \(\mathcal{F} = \bigcup_{u} \bigcup_{e \in u} F(e|u)\) and \(\mathcal{G} = \bigcup_{v} \bigcup_{e \in v} G(e|v)\) for VC firms and startups (in Table 3.1), respectively.

**Social Connections.** For two individuals \(e_1, e_2 \in E\), we have the directed and labeled connection, denoted by the label set \(C(e_1, e_2)\). For example, if \(e_1\) follows \(e_2\) on Twitter and Facebook and \(e_2\) follows \(e_1\) on Google+, we let

\[
C(e_1, e_2) = \{\text{Twitter follower, Facebook follower}\},
\]

\[
C(e_2, e_1) = \{\text{Google+ follower}\}.
\]

Note that since the connection is directed, generally

\[
C(e_1, e_2) \neq C(e_2, e_1).
\]

If there is no connection form \(e_1\) to \(e_2\), we let \(C(e_1, e_2) = \emptyset\). In our data set, the universal connection label set is \(\mathcal{C} = \bigcup_{e_1 \in E} \bigcup_{e_2 \in E} (C(e_1, e_2) \cup C(e_2, e_1))\) (see Table 3.1).
3.3.2 The Social-Adjusted PMF

Given our real-world dataset (see Section 3.4), we observe the investment amounts \( R^o \) follow log-normal distribution, illustrated in Figure 2.1. We then use \( R = \log(R^o) \) as the input to our model, which follows Gaussian distribution, in line with PMF model settings. We further assume that social connections between members are predictive of investment decisions from VC firms to startups. Therefore, we enhance the conventional latent factor model for investment recommendations with the incorporation of social information. Specifically, we write

\[
R_{nm} \sim \mathcal{N}((1 + S_{nm}^\gamma)(u_n, v_m), \sigma_{nm}^2). \tag{3.1}
\]

Here, \( S_{nm} \) is the prior investment interest inferred from the social connections information and \( \gamma \) is the scaling parameter of social interest.

Since each VC firm (or startup) usually consists of multiple members, we further define

\[
S_{nm} = \sum_{e_1 \in u_n} \sum_{e_2 \in v_m} W_{e_1e_2} \sum_{f \in F(e_1 | u_n)} \alpha_f \sum_{g \in G(e_2 | v_m)} \beta_g, \tag{3.2}
\]

where \( \alpha_f \) and \( \beta_g \) are the influence potentials of label \( f \in F \) and \( g \in G \), respectively. \( W_{e_1e_2} \) is the connection potential between two members \( e_1 \) and \( e_2 \). Since there might be multiple (directed and labeled) connections between two members, we compute

\[
W_{e_1e_2} = \sum_{\ell \in C(e_1, e_2)} p_\ell + \sum_{\ell \in C(e_2, e_1)} q_\ell. \tag{3.3}
\]

In other words, the connection potential between two members depends on the directions and labels of their social relationships. Each label is quantified by \( p_\ell \) or \( q_\ell \) as per the direction.
3.3.3 Parameter Estimation

In order to estimate the unknown parameters, we utilize the Maximum a Posterior (MAP) approach. Specifically, we use the following priors for the latent factors:

\[ u_n \sim N(\mu_u, \sigma_u^2), \]
\[ v_m \sim N(\mu_v, \sigma_v^2), \]

(3.4)

where \( \mu_u, \mu_v \in \mathbb{R}^{K \times 1} \) are the means and \( \sigma_u^2, \sigma_v^2 \in \mathbb{R}^{K \times K} \) are the variances of \( u_n \) and \( v_m \), respectively. In addition, we use non-negative constraints on all parameters in \( \{\alpha, \beta, p, q\} \) for better interpretation. Plus, we apply simplex constraints for \( \alpha \) and \( \beta \) to avoid ambiguous solutions. Note that in Equation 3.2, the results will not change if a constant is multiplied to \( \alpha_f \) and divided from \( \beta_g \). Therefore, the overall constraints are

\[ \sum_f \alpha_f = \sum_g \beta_g = 1, \]
\[ \alpha, \beta, p, q \geq 0. \]  

(3.5)

We let \( \Omega = \{\sigma, \mu_u, \mu_v, \sigma_u, \sigma_v\} \) be the set of hyper-parameters. The posterior probability of our model is

\[
P(u, v, \alpha, \beta, p, q | r, \Omega) \propto \prod_{n=1}^{N} \prod_{m=1}^{M} \left[ \frac{1}{\sigma} \exp \left( - \frac{(r_{nm} - (1 + S_{nm}) (u_n, v_m))^2}{2 \sigma^2} \right) \right] I_{nm} \]
\[ \times \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{1}{\sigma_u} \exp \left( - \frac{(u_{nk} - \mu_u)^2}{2 \sigma_u} \right) \]
\[ \times \prod_{m=1}^{M} \prod_{k=1}^{K} \frac{1}{\sigma_v} \exp \left( - \frac{(v_{mk} - \mu_v)^2}{2 \sigma_v} \right), \]

(3.6)

where \( S_{nm} \) is driven by Equation 3.2 and Equation 3.3. \( I_{nm} \) is the indicator function such that \( I_{nm} = 1 \) if and only if we observed the investment \( R_{nm} \).
We have the negative log-posterior as our objective function:

\[
J(u, v, \alpha, \beta, p, q) = -L(u, v, \alpha, \beta, p, q | r, \Omega) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} I_{nm}(r_{nm} - (1 + S_{nm}^\gamma)(u_n, v_m))^2 + \frac{1}{2\sigma^2_u} \sum_{n=1}^{N} \sum_{k=1}^{K} (u_{nk} - \mu_u)^2 + \frac{1}{2\sigma^2_v} \sum_{m=1}^{M} \sum_{k=1}^{K} (v_{mk} - \mu_v)^2.
\]

(3.7)

Therefore, by minimizing the objective function \( J(\cdot) \), we can estimate the unknown parameters. We then apply the alternative gradient decent algorithm. In order to give the gradients in concise forms, we define the matrix \( R = (r_{nm}) \in \mathbb{R}^{N \times M} \) where the entry at the \( n \)-th row and \( m \)-th column, i.e. \( r_{nm} \), is the observed investment from VC \( u_n \) to startup \( v_m \). In addition, we define three tensors \( F, G, \) and \( C \), representing the social information. As illustrated in Figure 3.3, \( F \) encodes the labeled memberships in VCs, and \( C \) encodes the labeled social connections between all members. We omit \( G \) for startup members in the figure since it conceptually coincides with \( F \). The rigorous definitions are as follows:

- Tensor \( F \in \mathbb{R}^{N \times E \times |F|} \): \( F^f_{ni} = \begin{cases} 1 & f \in F(e_i|u_n) \\ 0 & \text{otherwise} \end{cases} \)

- Tensor \( G \in \mathbb{R}^{M \times E \times |G|} \): \( G^g_{mj} = \begin{cases} 1 & g \in G(e_j|v_m) \\ 0 & \text{otherwise} \end{cases} \)

- Tensor \( C \in \mathbb{R}^{E \times E \times |C|} \): \( C^{\ell}_{ij} = \begin{cases} 1 & \ell \in C(e_i, e_j) \\ 0 & \text{otherwise} \end{cases} \)
For the sake of simplicity, we also define the following tensor operators:

\[
\kappa(\alpha, F) = \sum_f \alpha_f F^f,
\]

\[
\kappa(\beta, G) = \sum_g \beta_g G^g,
\]

\[
\kappa(p, C) = \sum_\ell p_\ell C^\ell,
\]

\[
\kappa(q, C) = \sum_\ell q_\ell C^\ell.
\]

As a result, \(\kappa(\alpha, F) \in \mathbb{R}^{N \times E}\), \(\kappa(\beta, G) \in \mathbb{R}^{M \times E}\), \(\kappa(p, C) \in \mathbb{R}^{E \times E}\), and \(\kappa(q, C) \in \mathbb{R}^{E \times E}\).

Then, they follow that

\[
W = \kappa(p, C) + \kappa(q, C)^\top,
\]

\[
S = \kappa(\alpha, F) (\kappa(p, C) + \kappa(q, C)^\top) \kappa(\beta, G)^\top.
\]

It is straightforward to show the computation results are equivalent with those of \textbf{Equation 3.3} and \textbf{Equation 3.2}. Moreover, the gradients of matrix \(S\) with respect to
social parameters $\alpha, \beta, p, q$ are as follows:

$$
\frac{\partial S}{\partial \alpha} = F_f (\kappa(p, C) + \kappa(q, C)^\top) \kappa(\beta, G)^\top,
$$

$$
\frac{\partial S}{\partial \beta} = \kappa(\alpha, F) (\kappa(p, C) + \kappa(q, C)^\top) (G^g)^\top,
$$

$$
\frac{\partial S}{\partial p} = \kappa(\alpha, F) C^\ell \kappa(\beta, G)^\top,
$$

$$
\frac{\partial S}{\partial q} = \kappa(\alpha, F) (C^\ell)^\top \kappa(\beta, G)^\top.
$$

Now we give the gradients of the objective function:

$$
\frac{\partial J}{\partial u} = -\frac{1}{\sigma^2} \sum_m I_{nm} \cdot d_{nm} (1 + S_{nm}^\gamma) \cdot v_m + \frac{1}{\sigma^2_u} (u_n - \mu_u),
$$

$$
\frac{\partial J}{\partial v} = -\frac{1}{\sigma^2} \sum_n I_{nm} \cdot d_{nm} (1 + S_{nm}^\gamma) \cdot u_n + \frac{1}{\sigma^2_v} (v_m - \mu_v),
$$

$$
\frac{\partial J}{\partial \xi} = -\frac{\gamma}{\sigma^2} \sum_{n,m} I_{nm} \cdot d_{nm} \langle u_n, v_m \rangle S_{nm}^{\gamma-1} \cdot \frac{\partial S_{nm}}{\partial \xi}.
$$

where $\xi \in \{\alpha_f, \beta_g, p_\ell, q_\ell\}$ and the prediction residual is defined as:

$$
d_{nm} = R_{nm} - (1 + S_{nm}^\gamma) \langle u_n, v_m \rangle.
$$

3.3.4 Algorithm Implementation

Given the above partial derivatives $\frac{\partial J}{\partial \theta}$, the modeling parameters:

$$\theta \in \{u, v, \alpha, \beta, p, q\}$$

can be optimized with the alternating gradient descent procedure in Algorithm 1. In the algorithm, $\lambda$ is the learning rate which is determined by the line-search procedure when updating each parameter. The two operators proj_{splx}(\cdot) and proj_{nn}(\cdot) are simplex
and non-negative projections, respectively:

\[
\text{proj}_{\text{splx}}(x) = \arg \min_{y: y \geq 0, \|y\|_1 = 1} \|x - y\|^2,
\]

\[
\text{proj}_{\text{nn}}(x) = \arg \min_{y: y \geq 0} \|x - y\|^2.
\]

Note that, the non-negative projection \(\text{proj}_{\text{nn}}(\cdot)\) simply replaces negative values with zeros.

The initialization of \(u,v\) are randomly sampled from their distribution priors. The other parameters \((\alpha, \beta, p, q)\) are initialized with data statistics. According to the simplex constraints in Equation 3.5, we initialize \(\alpha_f\) as:

\[
\alpha_f = \frac{\rho_f}{\sum_f \rho_f}, \quad \text{where} \quad \rho_f = \frac{\sum_{n:f \in u_n} \sum_m I_{nm}}{\sum_{n:f \in u_n} M}.
\]

(3.8)

Specifically, \(\rho_f\) is the percentage of investments of investors having members titled as \(f\), i.e., \(f \in u_n\) if and only if at least one member of VC \(u_n\) is titled as the label \(f\). Likewise, the initialization of \(\beta_g\) is:

\[
\beta_g = \frac{\rho_g}{\sum_g \rho_g}, \quad \text{where} \quad \rho_g = \frac{\sum_{m:g \in v_m} \sum_n I_{nm}}{\sum_{m:g \in v_m} N},
\]

(3.9)

in which \(\rho_g\) is the percentage of VCs investing startups \(v_m\), for \(g \in v_m\).

We also compute the percentage of investments \(\rho_\ell\) for \(\ell \in \mathcal{C}\). Since \(\rho_\ell \geq 0\), they are directly used to initialize the parameters \(p_\ell\) and \(q_\ell\):

\[
p_\ell = \frac{\sum_{n,m:u_n \overset{\ell}{\rightarrow} v_m} I_{nm}}{N \times M}, \quad q_\ell = \frac{\sum_{n,m:u_n \overset{\ell}{\rightarrow} v_m} I_{nm}}{N \times M}.
\]

(3.10)

Here \(u_n \overset{\ell}{\rightarrow} v_m\) indicates that there exists member in \(u_n\) following member in \(v_m\), and the connection between them is labeled by \(\ell \in \mathcal{C}\).
When the modeling scale is large \((N, M \gg 0)\), the updating of \(u_n\) for \(n = 1, 2, \cdots, N\) can be implemented in parallel, since the updating procedures are independent with each other at each iteration. Similarly, \(v_m\) for \(m = 1, 2, \cdots, M\) can also be updated in parallel for better computing efficiency. The algorithm convergences and learning performances will be discussed later in the empirical study.

1: Initialize \(\theta \in \{u, v, \alpha, \beta, p, q\}\):
2: Initialize \(u\) and \(v\) with Equation 3.4
3: Initialize \(\alpha\) with Equation 3.8
4: Initialize \(\beta\) with Equation 3.9
5: Initialize \(p\) and \(q\) with Equation 3.10
6: repeat
7:  for \(n \leftarrow 1, 2, \cdots, N\) do
8:      \(u_n \leftarrow u_n - \lambda \times \frac{\partial J}{u_n}\)
9:  end for
10: for \(m \leftarrow 1, 2, \cdots, M\) do
11:    \(v_m \leftarrow v_m - \lambda \times \frac{\partial J}{v_m}\)
12: end for
13: \(\alpha \leftarrow \alpha - \lambda \times \frac{\partial J}{\alpha} \quad /\text{*** Update } \alpha \text{ ***/}
14: \(\alpha \leftarrow \text{proj}_{\text{splx}}(\alpha)\)
15: \(\beta \leftarrow \beta - \lambda \times \frac{\partial J}{\beta} \quad /\text{*** Update } \beta \text{ ***/}
16: \(\beta \leftarrow \text{proj}_{\text{splx}}(\beta)\)
17: \(p \leftarrow p - \lambda \times \frac{\partial J}{p} \quad /\text{*** Update } p \text{ ***/}
3.4 Empirical Study

We evaluate the effectiveness of our approach on both synthetic data and real-world venture financing market data. We use synthetic data to analyze the relationship between our algorithmic performances and data properties. We use real-world data to demonstrate the advantages of our approach in comparison with competing methods. In addition, we provide intuitive managerial implications derived from the modeling results.

3.4.1 Synthetic Data

In this part, we randomly generate a set of $N = 50$ VCs and $M = 100$ startups with latent factors, and their friendship connections. Specifically, we draw $u_n, v_m \sim \mathcal{N}(0 \in \mathbb{R}^K, I \in \mathbb{R}^{K \times K})$, where $K = 3$. We design two types of members for VCs, two types
of members for startups, and three types of directed connections between them. The associated parameters are $\alpha_1 = 0.2, \alpha_2 = 0.8, \beta_1 = 0.4, \beta_2 = 0.6$ for members, and $p_1 = 1, p_2 = 2, p_3 = 4, q_1 = 10, q_2 = 1, q_3 = 0.1$ for connections. Accordingly, we create $E = 50$ members and each of them is associated with a VC/startup if $r < r_1$ where $r \sim \mathcal{U}(0, 1)$ is a uniform random real number in the range of $[0, 1]$. The member association is randomly labeled with types. Then, for each (directed) pair of member entities, we generate an edge if $r < r_2$ where $r \sim \mathcal{U}(0, 1)$. The generated edge is also randomly labeled with types. Finally, with the above VCs, startups, and their connections, we use Equation 3.1 ($\gamma = 1$) to draw the observation matrix $R \in \mathbb{R}^{N \times M}$.

To better demonstrate the modeling generality, we simulate another observation matrix $T$ for testing. In other words, our model is trained with data $R$ and evaluated on both $R$ and $T$. We use three performance metrics, normalized objective function, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) (refer to Gunawardana and Shani [2009] for their detailed definitions). The objective function $\mathcal{J}(\gamma)$ is defined in Equation 3.7 with the ground truth $\gamma = 1$, and computed with optimization solution at convergence. As aforementioned, by setting $\gamma = 0$, our model degenerates to conventional PMF model. Therefore we define the normalized objective function as

$$\text{Normalized objective function} = \frac{\mathcal{J}(\gamma)}{\mathcal{J}(0)}.$$  

Likewise, the normalized RMSE and MAE are defined as

$$\text{Normalized RMSE} = \frac{\text{RMSE with our model}}{\text{RMSE with PMF}},$$

$$\text{Normalized MAE} = \frac{\text{MAE with our model}}{\text{MAE with PMF}}.$$
where RMSE and MAE are defined in [Gunawardana and Shani, 2009].

First we set \( r_1 = r_2 = 0.1 \) and study the convergence behavior of our modeling algorithm. The optimization learning curves on both training data and testing data are shown in Figure 3.4. In the left panel of the figure, we see the objective function converges quickly within about 20 iterations. Interestingly, the convergence is observed on not only the training data but also the testing data, though the fitting is better on training data (the solid curve is under the dashed curve). This indicates that our model generalizes well on unseen data. Moreover, we compute RMSE and MAE along the optimization process and plot the corresponding curves in the middle and right panels, where we have consistent observations and conclusions in line with that of the objective functions.

We change the simulation parameters \( r_1 \) and \( r_2 \) to analyze the relationship between algorithmic performances and data properties. Specifically, we first increase \( r_1 \) from the current value 0.1 up to 0.2 with step 0.01, and at each step, we compute the normalized objective function on both training data and testing data. The results are shown in Figure 3.5 where we observe that, the model fits both the training data and the testing data better with more members associated with each VC and startup. This observation supports our research motivation, that the connections between members of VCs and startups can help predict investment decisions, and our proposed model can leverage such observation for investment recommendations. Also, in Figure 3.6, we observe a similar pattern with increasing \( r_2 \). In other words, the more connections between members of VCs and startups, the better our model performs in comparison with conventional PMF model. Note that, in both Figure 3.5.
and Figure 3.6, we show the normalized metrics of our model with respect to that of PMF.

![Figure 3.5: The normalized objective with different $r_1$.](image1)

![Figure 3.6: The normalized objective with different $r_2$.](image2)

We further study the model performance by varying $\gamma$. Note that $\gamma$ adjusts the influence of the social factor $S_{nm}$ in our model. With bigger $\gamma$, social information would have more significant contribution to the investment recommendation model. Figure 3.7 shows the normalized objective with $\gamma$ varying from 0 to 2. When $\gamma$ is close to 0, our model performs similarly as the conventional PMF does. However, by increasing $\gamma$, i.e. enhancing the influence of social information, our model performance is being improved steadily.
Besides, we perform analysis on how sensitive our model is wrt. noise. Specifically, we add Gaussian noise $\sim \mathcal{N}(0, \sigma^2)$ with varying $\sigma$ into $R$. Note that, in our experiment, the $\sigma$ of the simulated $R$ is around 21.39. Therefore, we vary $\sigma$ from 0.1 to 70 purposely to investigate the model performance. Figure 3.8 presents the experimental results, i.e. the normalized objective with different $\sigma$. We can see that the model performs rather well when adding noise with very small $\sigma$. By slightly increasing $\sigma$, the resulted normalized objective rises accordingly but still remains low. The trend slows down after $\sigma > 5$. Note that when $\sigma \approx 21.39$, our model still outperforms the conventional PMF with a significant margin. Such observation demonstrates the capability of our proposed model to tolerate additive noise to certain extent.

### 3.4.2 Real-world Data

To test our model performance in real scenario, we utilize the data from Crunchbase as the main source of information about VC firms, startups, and venture capital investment deals. As the world’s most comprehensive dataset of startup activities, there are several benefits to employ Crunchbase dataset compared to other alternatives, such as
accessibility, data volume, and information coverage. Although Crunchbase provides social networking feature which allows people to follow each other, it is far from an full-fledged social network community, like Twitter, Facebook, About.me, etc., and the amount of connection information is rather scarce compared with others. Alternatively, we resort to the data from About.me, a personal web hosting service founded in October 2009, to compensate the missing social information between members of VC firms and startups, mainly for two reasons. First, it links to multiple popular social networking websites, such as Facebook, Twitter, Pinterest, Google+, etc. We can extract much more social information on this single social network hub. More importantly, to the best of our knowledge, About.me is the only social networking website which releases the information about the time one individual followed or was followed by another one. In fact, temporal information about “following” is essential in our model settings, which will be discussed in Section 3.4.2.

https://en.wikipedia.org/wiki/About.me
Data acquisition

As mentioned above, our real-world data is gathered from two different data sources, Crunchbase and About.me. Specifically, we first crawled the data about startup profiles, VC firm profiles, and investment records from Crunchbase. We then extracted the member list from each organization with each individual’s name and the corresponding organization name. Given these lists of members, we searched for their profiles on About.me. As it is inescapable to have duplicates in search results due to the naturality of names duplication, we then applied a heuristic entity matching method, by comparing individual’s name and the corresponding organization if available. In this manner, we have the two databases linked via the conformed individuals. Note that, to the best of our knowledge, we are the first to attempt to bridge Crunchbase and About.me databases.

Dataset metadata and statistics

Specifically, the dataset we exported from Crunchbase consists of 2,462 investors, 8,817 startups, and 56,296 investment records in total. The closing time of investment deals ranges from 11/1990 to 05/2015. With the list of VC firms and startups, we extracted all current team members, past team members, and current board members and advisors, to generate the pool of all members. All members were then searched and matched by confirming the individuals’ affiliated institutions consistent in Crunchbase dataset and About.me profiles. After entity matching, we concentrate on 1.3K individuals from either VC firms or startups. By utilizing the social information on About.me, we are able to discover the (directed) social connections between
members and have their connections divided into different social groups as recorded in their profiles.

Figure 3.9 shows the frequency of members with different job titles in either VC firms or startups. As shown, founder/CEOs are in the majority, which is reasonable since the leader of an organization is inclined to attract more attention by publicizing his/her profile in online communities. We also find there are more members with the title Partners in VC firms, in line with the reality. We present the friendship connection types in Figure 3.10 which shows that most of social connections come from Twitter and Facebook while the rest consume a relatively small portion.

More importantly, social connections established between members should be unquestionably before the closing dates of investment deals of interest, so as not to invert their possible causal relationship. Note that, the timestamps from About.me are not accurate in measuring the time when the social connections were established. The reason is that, many of the connection records on About.me were imported from other platforms (e.g., Twitter, Facebook), and the timestamps were recorded at the event of imports. Such information is still of great usefulness as it guarantees the actual friendships were established no later than the recorded timestamps. Our statistics on the data show that, about 80% of the investment deals were closed after the social connections establishment. It means, for much more than 80% of the data, we confirm that the members from VC firms and the members from the corresponding startups knew each other in one way or the other before they subsequently reached their investment deals. We therefore extracted this data portion as our real-world dataset.
Baseline

Several baseline algorithms are introduced conceptually as follows.

**UserMean, ItemMean** The naive UserMean and ItemMean methods simply predict the missing values of $r_{nm}$ in the matrix $R$ by computing the row-wise and column-wise average, respectively.

**Social-CF** Another baseline is the collaborative filtering with social information.
about the firm members. Specifically, the collaborative filtering algorithm requires similarity measures between any two users (in our case, VCs). One natural choice is $S_{n_1 n_2}$ defined as follows:

$$S_{n_1 n_1} = \sum_{e_1 \in u_{n_1}} \sum_{e_2 \in u_{n_2}} \left(|C(e_1, e_2)| + |C(e_2, e_1)|\right)$$

$$\times \left|F(e_1|u_{n_1})\right| \times \left|F(e_2|u_{n_2})\right|.$$  \hfill (3.11)

Then the recommendation will be based on the estimation:

$$r_n = \sum_{n' \neq n} S_{nn'} r_{n'}.$$  

**PMF** The third baseline is the simple PMF without social supervision:

$$R_{nm} \sim \mathcal{N}(\langle u_n, v_m \rangle, \sigma).$$

We use the implementation developed in [Salakhutdinov and Mnih 2008](#).

**SimCoRec** The last baseline is a simplified version of our model. Specifically, we let $\alpha = \beta = p = q = 1$ and only learn the parameter $\tau$ in:

$$R_{nm} \sim \mathcal{N}(1 + \tau \cdot S_{nm}^\gamma \langle u_n, v_m \rangle, \sigma).$$  \hfill (3.12)

where

$$S_{nm} = \sum_{e_1 \in u_n} \sum_{e_2 \in v_m} \left(|C(e_1, e_2)| + |C(e_2, e_1)|\right)$$

$$\times \left|F(e_1|u_n)\right| \times \left|F(e_2|v_m)\right|.$$  \hfill (3.13)

Note that, when $\gamma \to 0$, it follows that:

$$S_{nm}^\gamma \to [S_{nm} > 0] = \begin{cases} 1 & S_{nm} > 0 \\ 0 & \text{otherwise} \end{cases}.$$
We use this simplified model to demonstrate the necessity of the parameters $\alpha$, $\beta$, $p$, and $q$ in our full model.

Results

In this part, we present the experimental results on the real-world data by comparing our proposed algorithm with other baseline algorithms. Our evaluation is conducted on two types of data, which are dataset with direct friendship and dataset with extended friendship. We employ three different evaluation metrics in our empirical study. The first two are RMSE and MAE, introduced in subsection 3.4.1, which measures the value prediction accuracy. The third one is Mean Average Precision (MAP)
Table 3.2: The significant labels of members and connections.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Founder/CEO, MD</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Founder/CEO, Partner, Head, Others</td>
</tr>
<tr>
<td>$p$</td>
<td>Twitter, Comp</td>
</tr>
<tr>
<td>$q$</td>
<td>Twitter, Comp</td>
</tr>
</tbody>
</table>

Gunawardana and Shani [2009], to measure the recommendation performance of our model.

From Figure 3.11a and Figure 3.11b, we see PMF-based models are overall performing better than the other three baseline algorithms with a significant margin. In particular, our proposed model CoRec outperforms all other competing algorithms, which demonstrates the incorporation of member social information indeed improves the model predictive capability. Note that, the performance of our simplified Sim-CoRec algorithm, worse than CoRec though, is still better than all the rest. On the other hand, in Figure 3.11c, we present the recommendation capabilities of different models. Likewise, two of our CoRec-based algorithms demonstrate their competence with better performance than all other competing algorithms. Our proposed model CoRec, overall, demonstrates its relative better predictive and recommendation capabilities based on our experimental results.

In the mean time, inherently in the model, we are able to discover the relative importance of the embedded factors, quantified by the learnt parameter values. This
can help in understanding what those positions are in the organizations and which contributes more towards reaching investment deals. As shown in Table 3.2 for VC firms, by observing the parameter $\alpha$, we get two non-zero entries corresponding to the labels, Founder/CEO and MD, respectively. We see consistent responses by checking the parameter $\beta$, such as Founder/CEO and Partners. It shows that social relationships between leaders of VC firms and startups can help reach investment deals. We can also learn what social groups are more important when considering their influences on reaching investment deals. As seen, for either connection direction ($p$ or $q$), Twitter is a critical social platform in this matter. One explanation might be Twitter is the dominant label in our label set. Besides, Comp (“complimented me” in About.me) is also another important group since it signifies explicit tendency for connection.

### 3.5 Related Work

In our research, we approach the problem of investment prediction from the perspective of social connections between members of VC firms and startups, by integrating the social information in a generalized recommendation system. In this regard, early related work we have investigated can be grouped into two categories. The first category is about general recommendation systems with the utilization of social information. The second category, mainly from the finance and management point of view, is regarding how traditional venture financing and entrepreneurship scholars relate social capital to venture capital investments.

In the first category, a variety of recommender systems have been developed in the past, such as content-based approaches, collaborative filtering approaches, and hy-
brid approaches which are the combination of the first two methods [Adomavicius and Tuzhilin, 2005]. In this dissertation, our latent factor model is one instance of the collaborative filtering approaches. Indeed, several latent factor models have been developed and applied in different application domains, and were often enhanced by incorporating additional context information or constraints. Such examples include context-aware recommender systems [Adomavicius and Tuzhilin, 2011], temporal recommender systems [Xiong et al., 2010], recommender systems with social constraints [Yang et al., 2014b], etc. Particularly on social recommender systems, [Ma et al., 2008] proposed a social recommendation (SoRec) model, in which trust between users in a social network is integrated into the recommender systems by factorizing the social trust matrix. In [Ma et al., 2009], Social Trust Ensemble (STE) model was introduced, which is a linear combination of the basic matrix factorization approach. Moreover, Social Matrix Factorization (SocialMF), proposed by [Jamali and Ester, 2010], incorporates social trust by making a user’s feature vector dependent on the direct neighbors’ feature vectors. [Ma et al., 2011] added social network information into the model training procedure as regularization terms. Recently, [Yang et al., 2014b] presented a survey on collaborative filtering based social recommender systems and concluded that a social recommender system improves on the recommendation accuracy of the traditional systems by taking social interests and social trusts between users in a social network. However, there exists substantial differences between our case and other social recommender system settings. In fact, all social recommender systems above address the social relationships between users, which is typical in traditional recommendation scenario. On the contrary, we studied another type of social
connections which is between users (VCs) and items (startups). To the best of our knowledge, we are the first to utilize this type of social network information in such unique problem settings.

From the financial and managerial perspectives, scholars have studied how social connections can improve venture capital investments. Hochberg et al. [2007] found that better-networked VC firms experience significantly better fund performance, as measured by the proportion of investments that are successfully exited. However, this literature concentrated on social connections between VC firms, different from our scenario. Eugene and Yuan [2012] applied social network analysis to the field of investing behaviors. They utilized Crunchbase and Facebook data and found that investors have a tendency to invest in companies that are socially similar to them. Although this aligns with our idea that VC firms are inclined to place investment deals on startups which they are socially related to, this study is more of a descriptive analysis. Yuxian and Yuan [2013] further utilized predictive analysis and demonstrated that investors are indeed more likely to invest in a particular company if they have stronger social relationships in terms of closeness, be it direct or indirect. Compared with their study, we attempt to address such problem by employing more sophisticated data mining techniques. As for evaluation dataset, the social network connections in our data are fine-grained, directed, and labeled with multiple semantics. The rich data and advanced analytical techniques are combined to provide not only better predictive performances for recommendation but also actionable managerial implications for business decision making.
3.6 Summary

We developed a novel approach for venture capital investments prediction based on probabilistic factorization model with the incorporation of member social information between VC firms and startups. Specifically, we took into consideration several types of information, including the members’ job titles, and their directed and labeled social connections. We tested our proposed model on both synthetic and real-world datasets. The empirical results not only demonstrated the effectiveness of our approach but also its applicability to real-world scenarios.

More directions are worth exploring beyond our current work. For example, to study more thoroughly how social interactions impact the investment deals, we can reach for more specialized social networks, such as alumni networks, family/friends networks, etc. Additionally, for better investment deals predictive model, we can combine social relationship information with other vital information, such as organization profiles, market environments, etc., which can potentially enhance the proposed model.
CHAPTER 4

CONCLUSIONS AND FUTURE WORK

This dissertation presents several works on addressing different problems in venture capital investment research.

In Chapter 2, in order to recommend an investment portfolio to venture capitalists, we developed a unified model by integrating investors’ investment preferences, expected returns and potential investment risks. We utilized Bayesian Probabilistic Latent Factorization framework to learn investors’ decision-making propensities, with additional firmographics and historical information incorporated, including market specialization, geographic location, leading products, etc. To enhance our utility function for quantifying the rank of startups by concerning not only expected investment returns but also potential investment risks, we accommodated Modern Portfolio Theory. We showed our model outperforming other state-of-the-art algorithms on various evaluation metrics with extensive experiments on Crunchbase dataset.

In Chapter 3, given the social relationship between members from VC firms and start-up companies, we developed a novel approach based on probabilistic factorization model to foresee venture capital investments. We incorporated different types of information, including members’ job titles/positions, directed and labeled social connections. With experiments on both synthetic and real-world datasets, results not only showed the effectiveness of our approach but also demonstrated its applicability.
to real-world applications.

Beyond our current work, more directions are worth further exploring. From social relationship perspective, to study more thoroughly how social interactions influence investment deals, we can seek for more specialized social networks, such as alumni networks, family/friends networks, etc. for better investment deals predictive model. We can also fuse social relationship information with other vital information, such as organization-specific profiles, market environments, etc., which can potentially enhance our proposed model.

As stated earlier, the VC cycle includes five stages: fundraising, screening/selection, negotiation/investing, monitoring/advising and exit. Our research mainly focuses on screening/selection stage, assisting investors to make the right decisions on investment deals. In order to close investment deals, startup valuation is an equally important yet challenging problem. Various factors should be taken into consideration, such as industry sector, founding team, product/service, business model, etc. Investors’ personal experience and judgement may drive assessment of company values, which largely varies from deal to deal. More quantitative and data-driven valuation method is needed for both investors and startups in the stage of negotiation/investing. Another research strand that attracts significant attention is “exit” of investment. The typical questions include 1) predicting the possibility of startups to be successful (acquisition or IPO) at the end of the investment cycle, 2) when is best for investment exiting, 3) what the startups are valued (pricing for acquisition or IPO) when exiting. For the questions above, it ought not to be constrained to current algorithms or models. Deep learning models can be applied to improve valuation estimation perfor-
mance, in which RNN can model sequential investments and capture more dynamic information than traditional models.

Other than corporate venture capital investors, informal investors (e.g. angel investors) also play an important role in the early stage of venture capital investments. It is lack of attention for this research strand in the community, which ought to be improved with data availability and cutting-edge analytics technologies. Meanwhile, macro economic/financial environment should be taken into account for studies on venture capital investments, which requires integration of different subjects, including but not limited to economy, finance, management, and data analytics in general.
CHAPTER 5
APPENDIX

Some distributions

The conditional sampling distribution of $\beta$:

$$P(\beta|Y, \Theta_{-\beta}, \Omega_\beta) = \prod_{i=1}^{M} P(\beta_i|Y, \Theta_{-\beta}, \Omega_\beta)$$

$$P(\beta_i|Y, \Theta_{-\beta}, \Omega_\beta) = \mathcal{N}(\beta_i|\mu^*_i, (\Lambda^*_i)^{-1})$$

$$\Lambda^*_i = \Lambda_\beta + \alpha \sum_{j: I_{ij}=1} d_{ij}^2 \cdot z_j z_j'$$

$$\mu^*_i = (\Lambda^*_i)^{-1}(\Lambda_\beta \mu_\beta + \alpha \sum_{j: I_{ij}=1} d_{ij} (\log(y_{ij}) - d_{ij} \langle u_i, \hat{v}_j \rangle) \cdot z_j)$$

The conditional sampling distribution of $V$:

$$P(V|Y, \Theta_{-V}, \Omega_V) = \prod_{j=1}^{N} P(v_j|Y, \Theta_{-V}, \Omega_V)$$

$$P(v_j|Y, \Theta_{-V}, \Omega_V) = \mathcal{N}(v_j|\mu^*_j, (\Lambda^*_j)^{-1})$$

where

$$\Lambda^*_j = \Lambda_V + \alpha \sum_{i: I_{ij}=1} d_{ij}^2 \cdot u_i u_i'$$

$$\mu^*_j = (\Lambda^*_j)^{-1}(\Lambda_V \mu_V + \alpha \sum_{i: I_{ij}=1} d_{ij} (\log(y_{ij}) - \gamma d_{ij} \langle u_i, c_j \rangle - d_{ij} \langle \beta_i, z_j \rangle) \cdot u_i)$$

The conditional sampling distribution of $C$:

$$P(c_s|Y, \Theta_{-c_s}, \Omega_C) = \mathcal{N}(c_s|\mu^*_s, (\Lambda^*_s)^{-1})$$
where

\[ \Lambda_s^* = \Lambda_C + \alpha \sum_{i,j: I_{ij} = 1 \atop s \in \Phi_j} \left( \frac{d_{ij}}{\left| \Phi_j \right|} \right)^2 \cdot u \cdot u_i' \]

and

\[ \mu_s^* = (\Lambda_s^*)^{-1} \left( \Lambda_C \mu_C + \alpha \sum_{i,j: I_{ij} = 1 \atop s \in \Phi_j} \frac{d_{ij}}{\left| \Phi_j \right|} (\log(y_{ij}) - d_{ij} \langle u_i, v_i \rangle + \frac{\gamma}{\left| \Phi_i \right|} \sum_{l: l \neq s, l \in \Phi_j} c_l) - d_{ij} \langle \beta_i, z_j \rangle \cdot u_i \right) \]
References


Luca Berchicci, Joern Block, and Philipp Sandner. The influence of geographical proximity and industry similarity in a business angels investment choice. *Available at SSRN 1964618*, 2011.


Bart Clarysse, Mirjam Knockaert, and Andy Lockett. How do early stage high technology investors select their investments. Technical report, Ghent University, Faculty of Economics and Business Administration, 2005.


Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of


Dean A Shepherd and Andrew Zacharakis. Conjoint analysis: A window of opportu-


Eugene Liang Yuxian and Soe-Tsyr Daphne Yuan. Investors are social animals: Predicting investor behavior using social network features via supervised learning approach. 2013.


