2016 IEEE 16th International Conference on Data Mining

TO BE OR NOT TO BE FRIENDS: Exploiting Social Ties for Venture Investments

Hao Zhong*, Chuanren Liu†, Xinjiang Lu‡, Hui Xiong*
* Rutgers, the State University of New Jersey, Newark, NJ, USA
h.zhong31@rutgers.edu, hxiong@rutgers.edu
† Drexel University, Philadelphia, PA, USA, chuanren.liu@drexel.edu
‡ Northwestern Polytechnical University, Xi'an, China, xjlu@mail.nwpu.edu.cn

Abstract—**Recent years have witnessed the boom of venture capital industry. Venture capitalists can attain great financial rewards if their invested companies exit successfully, via being acquired or going IPO (Initial Public Offering). The literature has revealed that, from both financial and managerial perspectives, decision-making process and successful rates of venture capital (VC) investments can be greatly improved if the investors well know the team members of target startups. However, much less efforts have been made on understanding the impact of prominent social ties between the members of VC firms and start-up companies on investment decisions. To this end, we propose to study such social relationship and see how this information can contribute to foreseeing investment deals. We aim at providing analytical guidance for the venture capitalists in choosing right investment targets. Specifically, we develop a Social-Adjusted Probabilistic Matrix Factorization (PMF) model to exploit members social connections information from VC firms and startups for investment recommendations. Unlike previous studies, we make use of the directed relationship between any pair of connected members from the two institutions respectively and quantify the variety of social network groups. As a result, it brings in much more flexibility, and the modeling results inherently provide meaningful managerial implications for the operators of VC firms and startups. Finally, we evaluate our model on both synthetic and real-world data. The results demonstrate that our approach outperforms the baseline algorithms with a significant margin.**

I. Introduction

Recently, the prosperity of venture capital industry has caught great attention from the whole society. Venture capital (VC) is a form of private finance provided in return for an equity stake in potentially high growth companies [1]. By offering capital and mentoring, investors would receive high returns if their portfolio companies successfully exit, namely being acquired or going public. According to the MoneyTree$^{\text{TM}}$ Report by PricewaterhouseCoopers LLP (PwC) and the National Venture Capital Association (NVCA)[1], venture capitalists invested $48.3 billion in 4,356 deals in 2014, an increase of 61 percent in dollars and a 4% increase in deals over the prior year, based on data from Thomson Reuters.

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. ([2, 3]), a large portion of research paid significant attention on entrepreneurial teams. Tyebjee and Bruno [2] pointed out entrepreneur's personality and prior experience are assessed in particular by venture capitalists while they are reviewing investment deals. Vogel et al. [4] 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 accessible, benefited from several public data holders, like Crunchbase[2], SpokeIntel[3], Owler[4], 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[5], Twitter[6], About.me[7]) 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 [5, 6, 7] 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 [8, 9], since the definition of

[1] PWC & NVCA et al., "Moneytree Report Q4 2014", 2014.

[2] https://www.crunchbase.com/

[3] https://www.spokeintel.com/

[4] https://www.owler.com/

[5] https://www.facebook.com/

[6] https://twitter.com/

[7] https://about.me/

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.25

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 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) [10] 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 paper is organized as follows. Section II presents the statistical evidence of our motivations and the overview of our proposed method. Section III presents the details of our model, including the model specification and inferences. Section IV introduces our synthetic and real-world data for experiments as well as the evaluation results. The related work is presented in Section V and finally Section VI concludes our paper and envisions possible future work.

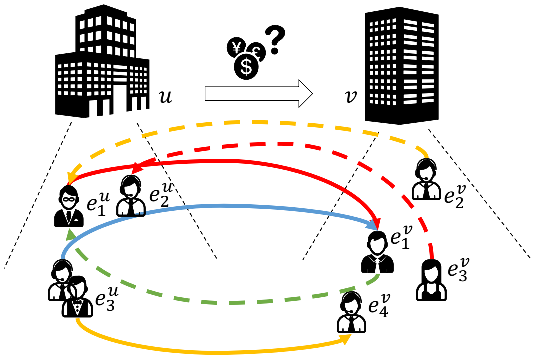

Fig. 1: The illustration of CoRec model.

II. Preliminary

Our key idea is illustrated in Figure 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 2. In this figure, we illustrate the investment statistics for the real-world data extracted from Crunchbase and About.me (refer to Section IV 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 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

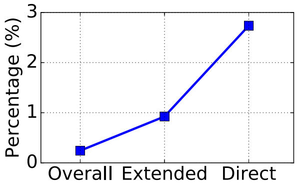

Fig. 2: The *percentage of investment* in different data samples.

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 prediction.

- 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 followed 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.

III. Methodology

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

A. Notations of Social Information

We have a set of VC firms $U = \{u_1, \cdots, u_N\}$, a set of startups $V = \{v_1, \cdots, v_M\}$ and member information of these VC firms and startups. Specifically, $\boldsymbol{u_n}$, $\boldsymbol{v_m} \in \mathbb{R}^{K\times 1}$ 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.

TABLE I: The set of membership and connection labels.

Set	Labels
$\mathcal{C}$	Twitter, Facebook, Insp, Love, Comp, Boldstart, Fave
$\mathcal{F}$	Founders/CEO, Partner, Managing Director, CFO, Others
$\mathcal{G}$	Founders/CEO, Partner, Chairman, CFO, Director, VP, CTO, Head, CPO, COO, Others

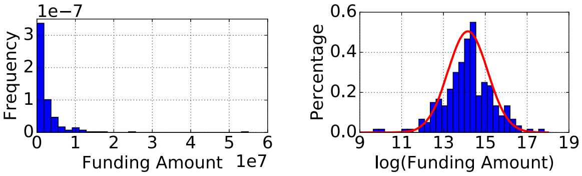

Fig. 3: The distribution of investment amounts from one randomly selected investor.

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

$$G(e|v) = \{\texttt{Founder}, \texttt{CTO}\}.$$

In our dataset, the universal membership label set is $\mathcal{F} = \cup_u \cup_{e\in u} F(e|u)$ and $\mathcal{G} = \cup_v \cup_{e\in v} G(e|v)$ for VC firms and startups (in Table I), 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

$$C(e_1, e_2) = \{\texttt{Twitter follower}, \texttt{Facebook follower}\},$$
$$C(e_2, e_1) = \{\texttt{Google+ follower}\}.$$

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 $\mathcal{C} = \cup_{e_1\in E} \cup_{e_2\in E} (C(e_1, e_2) \cup C(e_2, e_1))$ (see Table I).

B. The Social-Adjusted PMF

Given our real-world dataset (see Section IV), we observe the investment amounts R^o follow *log-normal* distribution, illustrated in Figure 3. 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

$$R_{nm} \sim \mathcal{N}((1 + S^{\gamma}_{nm})\langle \boldsymbol{u_n}, \boldsymbol{v_m}\rangle, \sigma^2_{nm}). \tag{1}$$

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

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

$$S_{nm} = \sum_{e_1 \in u_n} \sum_{e_2 \in v_m} W_{e_1 e_2} \sum_{f \in F(e_1|u_n)} \alpha_f \sum_{g \in G(e_2|v_m)} \beta_g, \quad (2)$$

where α_f and β_g are the *influence potentials* of label $f \in \mathcal{F}$ and $g \in \mathcal{G}$, respectively. $W_{e_1 e_2}$ 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

$$W_{e_1 e_2} = \sum_{\ell \in C(e_1, e_2)} p_\ell + \sum_{\ell \in C(e_2, e_1)} q_\ell. \quad (3)$$

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_ℓ or q_ℓ as per the direction.

C. 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:

$$\begin{aligned} \boldsymbol{u_n} &\sim \mathcal{N}(\boldsymbol{\mu_u}, \boldsymbol{\sigma_u}^2), \\ \boldsymbol{v_m} &\sim \mathcal{N}(\boldsymbol{\mu_v}, \boldsymbol{\sigma_v}^2), \end{aligned} \quad (4)$$

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

$$\begin{aligned} &\sum_f \alpha_f = \sum_g \beta_g = 1, \\ &\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q} \geq 0. \end{aligned} \quad (5)$$

We let $\Omega = \{\boldsymbol{\sigma}, \boldsymbol{\mu_u}, \boldsymbol{\mu_v}, \boldsymbol{\sigma_u}, \boldsymbol{\sigma_v}\}$ be the set of hyperperameters. The posterior probability of our model is

$$\begin{aligned} &P(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q} | \boldsymbol{r}, \Omega) \\ \propto &\prod_{n=1}^{N} \prod_{m=1}^{M} \Big[\frac{1}{\sigma} \exp\left(-\frac{(r_{nm} - (1 + S_{nm}^{\gamma}) \langle \boldsymbol{u_n}, \boldsymbol{v_m} \rangle)^2}{2\sigma^2} \right) \Big]^{I_{nm}} \\ &\times \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{1}{\sigma_u} \exp\left(-\frac{(u_{nk} - \mu_u)^2}{2\sigma_u^2} \right) \\ &\times \prod_{m=1}^{M} \prod_{k=1}^{K} \frac{1}{\sigma_v} \exp\left(-\frac{(v_{mk} - \mu_v)^2}{2\sigma_v^2} \right), \end{aligned} \quad (6)$$

where S_{nm} is drived by Equation 2 and Equation 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:

$$\begin{aligned} &\mathcal{J}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q}) = -\mathcal{L}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q} | \boldsymbol{r}, \Omega) \\ = &\frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} I_{nm} (r_{nm} - (1 + S_{nm}^{\gamma}) \langle \boldsymbol{u_n}, \boldsymbol{v_m} \rangle)^2 \\ + &\frac{1}{2\sigma_u^2} \sum_{n=1}^{N} \sum_{k=1}^{K} (u_{nk} - \mu_u)^2 + \frac{1}{2\sigma_v^2} \sum_{m=1}^{M} \sum_{k=1}^{K} (v_{mk} - \mu_v)^2. \end{aligned} \quad (7)$$

Therefore, by minimizing the objective function $\mathcal{J}(\cdot)$, 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 4, 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 $F \in \mathbb{R}^{N \times E \times |\mathcal{F}|}$: $F_{ni}^{f} = \begin{cases} 1 & f \in F(e_i|u_n) \\ 0 & \text{otherwise} \end{cases}$,
- Tensor $G \in \mathbb{R}^{M \times E \times |\mathcal{G}|}$: $G_{mj}^{g} = \begin{cases} 1 & g \in G(e_j|v_m) \\ 0 & \text{otherwise} \end{cases}$,
- Tensor $C \in \mathbb{R}^{E \times E \times |\mathcal{C}|}$: $C_{ij}^{\ell} = \begin{cases} 1 & \ell \in C(e_i, e_j) \\ 0 & \text{otherwise} \end{cases}$.

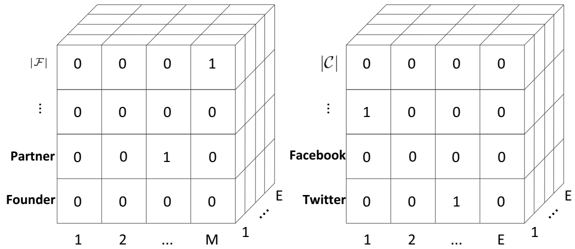

Fig. 4: The illustration of tensors F and C.

For the sake of simplicity, we also define the following tensor operators:

$$\begin{aligned} \kappa(\alpha, F) &= \sum_f \alpha_f F^f, & \kappa(\beta, G) &= \sum_g \beta_g G^g, \\ \kappa(p, C) &= \sum_\ell p_\ell C^\ell, & \kappa(q, C) &= \sum_\ell q_\ell C^\ell. \end{aligned}$$

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

$$\begin{aligned} W &= \kappa(p, C) + \kappa(q, C)^\top, \\ S &= \kappa(\alpha, F) \left(\kappa(p, C) + \kappa(q, C)^\top \right) \kappa(\beta, G)^\top. \end{aligned}$$

It is straightforward to show the computation results are equivalent with those of Equation 3 and Equation 2. Moreover, the gradients of matrix S with respect to social parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q}$ are as follows:

$$\begin{aligned}
\frac{\partial S}{\partial \alpha_f} &= F^f \left(\kappa(p, C) + \kappa(q, C)^\top \right) \kappa(\beta, G)^\top, \\
\frac{\partial S}{\partial \beta_g} &= \kappa(\alpha, F) \left(\kappa(p, C) + \kappa(q, C)^\top \right) (G^g)^\top, \\
\frac{\partial S}{\partial p_\ell} &= \kappa(\alpha, F) C^\ell \kappa(\beta, G)^\top, \\
\frac{\partial S}{\partial q_\ell} &= \kappa(\alpha, F) (C^\ell)^\top \kappa(\beta, G)^\top.
\end{aligned}$$

Now we give the gradients of the objective function:

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \boldsymbol{u_n}} &= -\frac{1}{\sigma^2} \sum_m I_{nm} \cdot d_{nm} (1 + S_{nm}^\gamma) \cdot \boldsymbol{v_m} + \frac{1}{\sigma_u^2} (\boldsymbol{u_n} - \boldsymbol{\mu_u}), \\
\frac{\partial \mathcal{J}}{\partial \boldsymbol{v_m}} &= -\frac{1}{\sigma^2} \sum_n I_{nm} \cdot d_{nm} (1 + S_{nm}^\gamma) \cdot \boldsymbol{u_n} + \frac{1}{\sigma_v^2} (\boldsymbol{v_m} - \boldsymbol{\mu_v}), \\
\frac{\partial \mathcal{J}}{\partial \xi} &= -\frac{\gamma}{\sigma^2} \sum_{n,m} I_{nm} \cdot d_{nm} \langle u_n, v_m \rangle S_{nm}^{\gamma - 1} \cdot \frac{\partial S_{nm}}{\partial \xi}.
\end{aligned}$$

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

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

D. Algorithm Implementation

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

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

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

$$\begin{aligned}
&\text{proj}_{\text{splx}}(x) = \arg\min_{y: y \geq 0, \|y\|_1 = 1} \|x - y\|^2, \\
&\text{proj}_{\text{nn}}(x) = \arg\min_{y: y \geq 0} \|x - y\|^2.
\end{aligned}$$

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

The initialization of $\boldsymbol{u}, \boldsymbol{v}$ are randomly sampled from their distribution priors. The other parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q}$) are initialized with data statistics. According to the simplex constraints in Equation 5, we initialize α_f as:

$$\alpha_f = \frac{\rho_f}{\sum_f \rho_f}, \text{ where } \rho_f = \frac{\sum_{n: f \in u_n} \sum_m I_{nm}}{\sum_{n: f \in u_n} M}. \tag{8}$$

Specifically, ρ_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 β_g is:

$$\beta_g = \frac{\rho_g}{\sum_g \rho_g}, \text{ where } \rho_g = \frac{\sum_{m: g \in v_m} \sum_n I_{nm}}{\sum_{m: g \in v_m} N}, \tag{9}$$

in which ρ_g is the percentage of VCs investing startups v_m, for $g \in v_m$.

We also compute the percentage of investments ρ_ℓ for $\ell \in \mathcal{C}$. Since $\rho_\ell \geq 0$, they are directly used to initialize the parameters p_ℓ and q_ℓ:

$$p_\ell = \frac{\sum_{n,m: u_n \overset{\ell}{\rightsquigarrow} v_m} I_{nm}}{N \times M}, \quad q_\ell = \frac{\sum_{n,m: v_m \overset{\ell}{\rightsquigarrow} u_n} I_{nm}}{N \times M}. \tag{10}$$

Here $u_n \overset{\ell}{\rightsquigarrow} v_m$ 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 $\boldsymbol{u_n}$ for $n = 1, 2, \cdots, N$ can be implemented in parallel, since the updating procedures are independent with each other at each iteration. Similarly, $\boldsymbol{v_m}$ 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.

Algorithm 1: The algorithm of Social-Adjusted PMF.

```
Initialize $\theta \in \{\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q}\}$:
Initialize $\boldsymbol{u}$ and $\boldsymbol{v}$ with Equation 4.
Initialize $\boldsymbol{\alpha}$ with Equation 8.
Initialize $\boldsymbol{\beta}$ with Equation 9.
Initialize $\boldsymbol{p}$ and $\boldsymbol{q}$ with Equation 10.
repeat
    for $n \leftarrow 1, 2, \cdots, N$ do
        $\boldsymbol{u_n} \leftarrow \boldsymbol{u_n} - \lambda \times \frac{\partial \mathcal{J}}{\boldsymbol{u_n}}$
    end for
    for $m \leftarrow 1, 2, \cdots, M$ do
        $\boldsymbol{v_m} \leftarrow \boldsymbol{v_m} - \lambda \times \frac{\partial \mathcal{J}}{\boldsymbol{v_m}}$
    end for
    $\alpha \leftarrow \alpha - \lambda \times \frac{\partial \mathcal{J}}{\alpha}$ /*** Update $\alpha$ ***/
    $\alpha \leftarrow \text{proj}_{\text{splx}}(\alpha)$
    $\beta \leftarrow \beta - \lambda \times \frac{\partial \mathcal{J}}{\beta}$ /*** Update $\beta$ ***/
    $\beta \leftarrow \text{proj}_{\text{splx}}(\beta)$
    $p \leftarrow p - \lambda \times \frac{\partial \mathcal{J}}{p}$ /*** Update $p$ ***/
    $p \leftarrow \text{proj}_{\text{nn}}(p)$
    $q \leftarrow q - \lambda \times \frac{\partial \mathcal{J}}{q}$ /*** Update $q$ ***/
    $q \leftarrow \text{proj}_{\text{nn}}(q)$
until Convergence.
```

IV. Empirical Study

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

A. Synthetic Data

In this part, we randomly generate a set of $N = 50$ VCs and $M = 100$ startups with latent factors, and their friendship

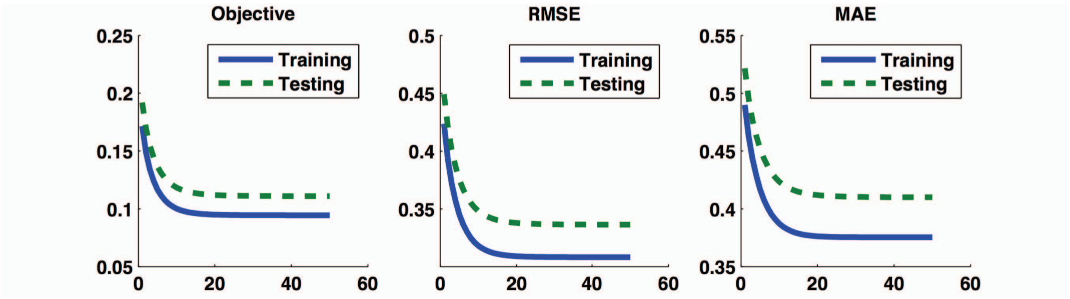

Fig. 5: The estimation process with synthetic data.

connections. Specifically, we draw $\boldsymbol{u_n}, \boldsymbol{v_m} \sim \mathcal{N}(\mathbf{0} \in \mathbb{R}^K, \boldsymbol{I} \in \mathbb{R}^{K\times K})$, 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 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 [11] for their detailed definitions). The objective function $\mathcal{J}(\gamma)$ is defined in Equation 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 [11].

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 5. 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 6, 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 7, 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 6 and Figure 7, we show the normalized metrics of our model with respect to that of PMF.

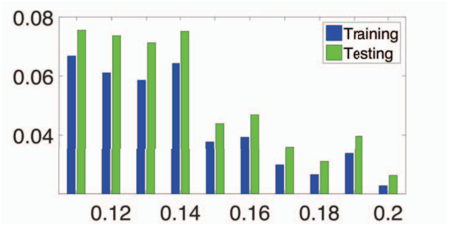

Fig. 6: The normalized objective with different r_1.

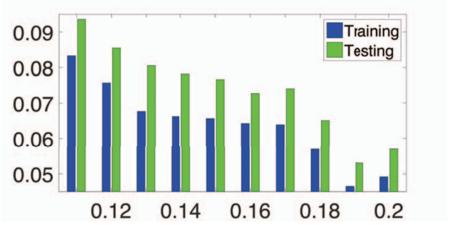

Fig. 7: The normalized objective with different r_2.

We further study the model performance by varying γ. Note that γ adjusts the influence of the social factor S_{nm} in our model. With bigger γ, social information would have more significant contribution to the investment recommendation

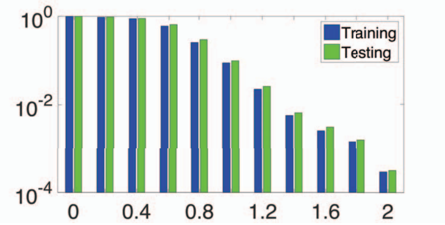

Fig. 8: The normalized objective with different γ.

model. Figure 8 shows the normalized objective with γ varying from 0 to 2. When γ is close to 0, our model performs similarly as the conventional PMF does. However, by increasing γ, 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 $\sim \mathcal{N}(0, \sigma_\epsilon^2)$ with varying σ_ϵ into R. Note that, in our experiment, the σ of the simulated R is around 21.39. Therefore, we vary σ_ϵ from 0.1 to 70 purposedly to investigate the model performance. Figure 9 presents the experimental results, i.e. the normalized objective with different σ_ϵ. We can see that the model performs rather well when adding noise with very small σ_ϵ. By slightly increasing σ_ϵ, the resulted normalized objective rises accordingly but still remains low. The trend slows down after $\sigma_\epsilon > 5$. Note that when $\sigma_\epsilon \approx 21.39$, our model still outperforms the conventional PMF with a significant margin. Such observation demonstrates the capability of our proposed model to tolerate addictive noise to certain extent.

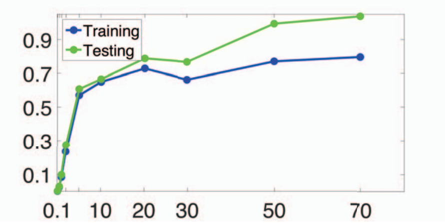

Fig. 9: The normalized objective with varying noise σ.

B. 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 converage. 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[8], 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 subsubsection IV-B2.

1) 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 confomed individuals. Note that, to the best of our knowledge, we are the first to attempt to bridge Crunchbase and About.me databases.

2) 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 10 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 11, 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

[8] https://en.wikipedia.org/wiki/About.me

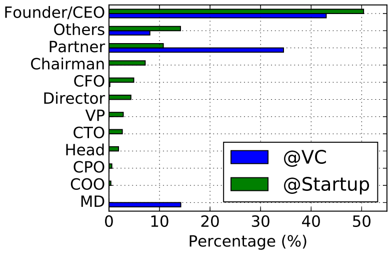

Fig. 10: The member titles in VC firms and startups.

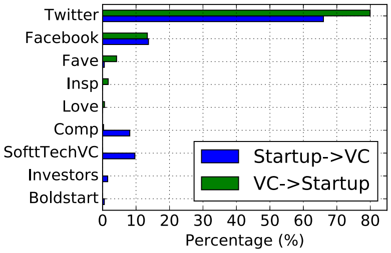

Fig. 11: The connection groups from VC (startup) members to startup (VC) members.

other platforms (e.g., Twitter, Facebook), and the timestamps were recorded at the event of imports. Such information is still of great usefullness 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.

3) 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:

$$S_{n_1 n_1} = \sum_{\substack{e_1 \in u_{n_1} \\ e_2 \in n_{n_2}}} \Big(|C(e_1, e_2)| + |C(e_2, e_1)| \Big) \times \Big| F(e_1 | u_{n_1}) \Big| \times \Big| F(e_2 | u_{n_2}) \Big|. \qquad (11)$$

Then the recommendation will be based on the estimation:

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

TABLE II: The significant labels of members and connections.

Parameter	Labels
α	`Founder/CEO`, `MD`
β	`Founder/CEO`, `Partner`, `Head`, `Others`
p	`Twitter`, `Comp`
q	`Twitter`, `Comp`

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 [12].

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

$$R_{nm} \sim \mathcal{N}((1 + \tau \cdot S^{\gamma}_{nm}) \langle u_n, v_m \rangle,\ \sigma). \qquad (12)$$

where

$$S_{nm} = \sum_{\substack{e_1 \in u_n \\ e_2 \in v_m}} \Big(|C(e_1, e_2)| + |C(e_2, e_1)| \Big) \times \Big| F(e_1 | u_n) \Big| \times \Big| F(e_2 | v_m) \Big|. \qquad (13)$$

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

$$S^{\gamma}_{nm} \to [S_{nm} > 0] = \begin{cases} 1 & S_{nm} > 0 \\ 0 & \text{otherwise} \end{cases}.$$

We use this simplified model to demonstrate the necessity of the parameters α, β, p, and q in our full model.

4) 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 IV-A, which measures the value prediction accuracy. The third one is *Mean Average Precision* (MAP) [11], to measure the recommendation performance of our model.

From Figure 12a and Figure 12b, 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 **SimCoRec** algorithm, worse than **CoRec** though, is still better than all the rest. On the other hand, in Figure 12c, 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,

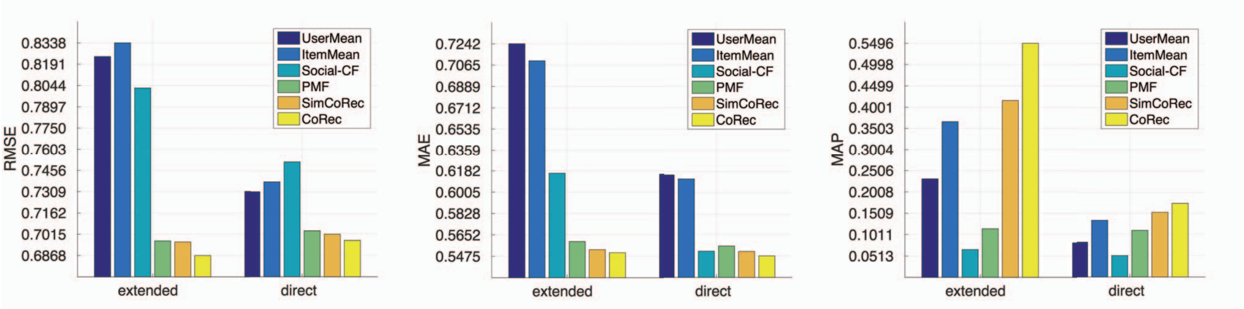

Fig. 12: RMSE, MAE and MAP on real-world data.

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 II, for VC firms, by observing the parameter α, we get two non-zero entries corresponding to the labels, `Founder/CEO` and `MD`, respectively. We see consistent responses by checking the parameter β, 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.

V. 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 hybrid approaches which are the combination of the first two methods [13]. In this paper, 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 [14], temporal recommender systems [15], recommender systems with social constraints [16], etc. Zhong et al. [17] developed a personalized portfolio model to assist investment decision-making utilizing recommender system techniques. They demonstrated the proposed model's effectiveness by employing multiple sources of information, which however failed to take members social relationship into account. On the other hand, in the realm of social recommender systems, Ma et al. [5] 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 [6], 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 [7], incorporates social trust by making a user's feature vector dependent on the direct neighbors' feature vectors. Ma et al. [18] added social network information into the model training procedure as regularization terms. Recently, Yang et al. [16] 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. [19] 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 [8] 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 [9] 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.

VI. Conclusion

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.

Acknowledgment

This research was supported in part by Natural Science Foundation of China (71329201). The contact author is Professor Hui Xiong (hxiong@rutgers.edu).

References

[1] T. Stone, W. Zhang, and X. Zhao, "An empirical study of top-n recommendation for venture finance." ACM Press, 2013, pp. 1865–1868.

[2] T. T. Tyebjee and A. V. Bruno, "A model of venture capitalist investment activity," *Management science*, vol. 30, no. 9, pp. 1051–1066, 1984.

[3] L. Berchicci, J. Block, and P. Sandner, "The influence of geographical proximity and industry similarity in a business angels investment choice," *Available at SSRN 1964618*, 2011.

[4] R. Vogel, T. X. Puhan, E. Shehu, D. Kliger, and H. Beese, "Funding decisions and entrepreneurial team diversity: A field study," *Journal of Economic Behavior & Organization*, vol. 107, pp. 595–613, 2014.

[5] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 931–940.

[6] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 203–210.

[7] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 135–142.

[8] L. Y. Eugene and S.-T. D. Yuan, "Where's the money? the social behavior of investors in facebook's small world." IEEE, Aug. 2012, pp. 158–162.

[9] E. L. Yuxian and S.-T. D. Yuan, "Investors are social animals: Predicting investor behavior using social network features via supervised learning approach," 2013.

[10] A. Mnih and R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2007, pp. 1257–1264.

[11] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *The Journal of Machine Learning Research*, vol. 10, pp. 2935–2962, 2009.

[12] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 880–887.

[13] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *TKDE'05*, vol. 17, no. 6, pp. 734–749, 2005.

[14] ——, "Context-aware recommender systems," in *Recommender systems handbook*. Springer, 2011, pp. 217–253.

[15] L. Xiong, X. Chen, T.-K. Huang, J. G. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization." in *SDM*, vol. 10. SIAM, 2010, pp. 211–222.

[16] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1–10, 2014.

[17] H. Zhong, C. Liu, J. Zhong, and H. Xiong, "Which startup to invest in: a personalized portfolio strategy," *Annals of Operations Research*, pp. 1–22, 2016.

[18] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 287–296.

[19] Y. V. Hochberg, A. Ljungqvist, and Y. Lu, "Whom you know matters: Venture capital networks and investment performance," *The Journal of Finance*, vol. 62, no. 1, pp. 251–301, 2007.