Bring Me a Good One: Seeking High-potential Startups using Heterogeneous Venture Information Networks

Shengming Zhang
Rutgers University
shengming.zhang@rutgers.edu

Hao Zhong
ESCP Business School
hzhong@escp.eu

Yong Ge
The University of Arizona
yongge@arizona.edu

Hui Xiong
HKUST (Guangzhou)
xionghui@ust.hk

Abstract

The rapid acceleration of technology and the evolving global economy have led to a significant surge in high-potential startups, presenting immense opportunities for venture capital firms and investors to support and benefit from these innovative ventures. However, identifying startups with the highest likelihood of success remains a complex task, necessitating the examination of various information sources, including firm demographics, management team composition, and financial performance. The effectiveness of existing methodologies, such as feature-based and network-topological approaches, is limited for predicting high-potential startups. In response, we propose a novel Venture Graph Neural Network (VenGNN) model, leveraging Heterogeneous Information Networks (HIN) and Graph Neural Networks (GNN) techniques to address the prediction problem. Our experimental analysis reveals that VenGNN outperforms state-of-the-art models by 15-20% across a wide range of performance metrics.

Keywords: high-potential startups, heterogeneous information networks, graph neural networks

1. Introduction

Startup companies play an important role in today's economy. According to the U.S. Census Bureau, over 5 million startup companies were created in 2022, and on average, more than 4 million have been created every year in the United States over the past five years[1]. These numerous startups have significantly contributed to the U.S. economy by promoting innovation, increasing production, and creating job positions. Investors, such as venture capital firms and individuals, provide vital capital and advice to support the development of startup companies, and the success of these startups can bring substantial returns to investors. Unfortunately, over two-thirds of startups never deliver a positive return to investors[2]. Therefore, it poses a critical challenge for investors to make informed investment decisions. A promising way to tackle this challenge is to develop a framework to predict high-potential startups using historical data in the entrepreneurial ecosystem. This approach will, in turn, benefit multiple stakeholders, including investors, startup founders and employees, as well as the U.S. economy and society.

In recent years, researchers from different disciplines have studied the prediction of high-potential startups. Some involve processing startups' demographic information into features based on insights gleaned from interviews, surveys, and expert input (Gompers et al., 2009; Nanda et al., 2020). These features are then utilized in supervised learning models for predicting potential startups (Bargagli-Stoffi et al., 2021; Sharchilev et al., 2018; Zhong et al., 2018). Some use the interactions between startups to construct information networks, and then topological features extracted from these networks are employed to identify high-potential startups (Bonaventura et al., 2020; Gloor et al., 2013; Zhang et al., 2021). While the first group of methods neglects the heterogeneous yet useful connections among different entities (e.g., startups, founders, VC firms), the second group fails to consider the wealth of non-topological information (e.g., startup size and industry) relevant to startup success. Given the limitations of these prior studies, it becomes evident that a more comprehensive and integrative approach is needed to effectively address the high-potential startup prediction problem. Such approaches should consider not only static aspects of startups' profiles or topological features of startup networks, but also the heterogeneous relations between entities and the wealth of hidden information that exists within the

[1] https://www.census.gov/econ/bfs/index.html

[2] https://hbr.org/2021/05/why-start-ups-fail

URI: https://hdl.handle.net/10125/106905
978-0-9981331-7-1

entrepreneurial ecosystem.

To fill the research gap, we propose a novel Venture Graph Neural Network (VenGNN) approach for predicting high-potential startups. It consists of three major steps. In the first step, we construct a Heterogeneous Venture Information Network (HVIN) from a publicly available startup database. The constructed HVIN consists of various startup-related entities, such as startups and VC firms, and different connections among them. Second, we define and extract multiple types of meta-paths from the HVIN, each of which captures one type of heterogeneous connection between startups. These extracted meta-paths enable us to effectively measure the multifaceted relations among startups. Built upon the HVIN and meta-paths, third, we design a novel multi-head graph attention network for predicting high-potential startups. The designed multi-head graph attention network models node features and startup connections. In contrast to prior studies (Bonaventura et al., 2020; Gompers et al., 2009; Nanda et al., 2020; Zhang et al., 2021), our proposed VenGNN approach considers both attribute and structure information of multiple types of startup-related entities (e.g., startups, VC firms, founders, educational institutes) and effectively integrates all useful information.

We conduct intensive empirical evaluations to demonstrate the superiority of our developed VenGNN model. Specifically, we construct multiple evaluation datasets from Crunchbase and use them to evaluate VenGNN's performance on predicting startup success in three different funding rounds. We compare our model with multiple state-of-the-art benchmark methods that consist of three groups: feature-based classification approaches (e.g., XGBoost (Xu et al., 2022)), homogeneous graph-based ones (e.g., GAT (Veličković et al., 2018)), and heterogeneous graph-based methods (e.g., HAN (Wang et al., 2019)). The evaluation results across the three funding rounds show that our VenGNN significantly outperforms other benchmark methods in terms of various evaluation metrics. The results demonstrate how and to what extent different factors contribute to the prediction of startup success, providing useful insights for real-world practitioners.

2. Literature Review

2.1. Existing Methods for Identifying High-potential Startups

In this section, we investigate two categories of high-potential startup prediction models: feature-based models and network-topological models, and another set of recommendation-based models. Feature-based models are designed to identify key information from raw data and transform it into features that reveal essential patterns with predictive power. Specifically, in the context of predicting startup success, relevant information is gathered from data platforms like CrunchBase and LinkedIn, including firm demographics, past funding rounds, and founder/employee information (Sharchilev et al., 2018). These sources are then used to extract features such as the number of offices, total funding amount raised, gender diversity on the board, employee characteristics, and company milestones. Typically, a machine learning algorithm is trained using these features to predict the likelihood of a startup's success, which is commonly defined as 1) receiving follow-on funding, 2) being acquired, or 3) going public. The literature has a broad coverage of machine learning models, including Logistic Regression (Zbikowski & Antosiuk, 2021), Naive Bayes (Krishna et al., 2016), Support Vector Machine (Zbikowski & Antosiuk, 2021), Random Forests and Decision Trees (Krishna et al., 2016), Gradient Boosting Classifiers (Bargagli-Stoffi et al., 2021),and neural networks (Sharchilev et al., 2018). However, such feature-based models focus solely on static company characteristics and do not consider interactions among different entities, such as firm-firm connections and VC-firm investments.

To alleviate this issue, network-topological models incorporate interactions between various actors in entrepreneurial activities into network structures, aiming to gain further insights. An illustration of interaction networks can be observed in the context of relationship networks, in which startups and individuals are depicted as nodes and the presence of an interaction (e.g., the founding or management of a company by an individual) results in the creation of an edge between the respective nodes (Bonaventura et al., 2020). These studies usually construct a dynamic network of connections among startups, by linking two startups if they have at least one individual who holds, or has held, a professional position in both entities. Then, a centrality-based methodology is employed to compute the centrality metric ranking of each startup node, with a higher ranking indicative of an elevated probability of success (Gloor et al., 2013; Hadley et al., 2018). Alternatively, some studies have utilized the social networks of entrepreneurs and employees to study the connection between social proximity and the success of startups (Song & Vinig, 2012; Zhong et al., 2016). However, only one type of interaction is considered, and the centrality-based approach cannot fully capture interaction patterns, making prediction a challenge.

Another strand of recommendation-based approaches also takes into account various actors (primarily investors and startups) involved in entrepreneurial activities. However, different from directly predicting the outcome of a startup, these studies aim to identify the potential for investments from VCs to startups, which serve as an indirect gauge of startup success. In the context of network analysis, it is similar to the problem of link prediction, i.e., predicting the existence of a link between two nodes in a network. An early study by Stone (2014) framed the task as a recommendation problem, recommending startup candidates to investors which will eventually lead to real funding commitments. The author developed both an item-based k-Nearest Neighbor collaborative filtering (CF) approach and a latent factor model to make investment recommendations solely based on investors' preferences. Building on that, Zhong et al. (2018) proposed another recommendation model that considers both investors' preferences and risk-averse portfolio returns. Although these recommendation-based approaches have the potential to identify actual investment links, it has been found to be challenging due to low top-k precision or AUC scores shown in previous studies (Stone, 2014; Zhong et al., 2018). Additionally, the startups with high investment potential may not necessarily have a high probability of success in the end.

2.2. Computational Design Science Research

Our work adds to the growing body of computational design science research in information systems (IS) (Padmanabhan et al., 2022; Rai, 2017). This paradigm emphasizes an "interdisciplinary approach in developing novel data representations, computational algorithms, business intelligence, and analytics methods" (Rai, 2017). They utilized natural language processing (NLP) and deep learning models to analyze customers' social media activity. They worked with a leading apparel firm to mine this data and found that their algorithmic solution performs 7%-9% better at detecting misbehavior than traditional methods. Yang et al. (2023a) designed an IS artifact, DeepPerson, for text-based personality detection using advanced deep learning strategies and psycholinguistic theories. It incorporates novel transfer learning and hierarchical attention network methods, along with data augmentation and person-level linguistic information. Yang et al. (2023b) utilized a design science approach to create DeepVoice, a new system for predicting financial risk during quarterly earnings calls. DeepVoice's design analyzes both what and how managers speak during these calls to forecast risk and tackles several challenges in analyzing nonverbal communication. Likewise, our research adopts a design science approach to develop a novel IS artifact, VenGNN, for predicting startup success. Our contribution is distinguished by the utilization of graph neural networks, which effectively capture the heterogeneity of the focal business information network.

3. Research Methodology

In this section, we introduce our novel method, Venture Graph Neural Network (VenGNN), for predicting high-potential startups. An illustration of VenGNN is shown in Figure 1, which mainly includes three components. In the first component, we construct an HVIN from the collected Crunchbase data, where we define six types of nodes and eight types of connections. In the second one, we define eleven different types of meta-paths between startups based on related theory and literature from the constructed HVIN, and identify and operationalize other startup-related features. Each type of meta-path captures the proximity between startups based on one kind of composite relation. In the third component, we develop novel multi-head graph attention networks to predict startup success, where the proximity among startups based on different types of meta-paths is combined. Next, we will provide a breakdown of each individual component, along with its respective details.

3.1. Heterogeneous Venture Information Network (HVIN) Construction

There are various entities and relations in the Crunchbase data. We construct a Heterogeneous Venture Information Network (HVIN) to model them. Our HVIN-based prediction is grounded in the theoretical foundation of the *homophily principle* that originated from McPherson et al. (2001). The homophily principle indicates that connections tend to be formed between similar objects (such as connected people with shared demographic information). Prior studies have applied and examined the homophily principle in both homogeneous networks (e.g., social networks (McPherson et al., 2001)) and bipartite networks (e.g., startup-person graph (Bonaventura et al., 2020)). Through the constructed HVIN, we consider multiple-typed connections among startups and hypothesize that the startups connected together share a similar likelihood of success. Specifically, the constructed HVIN consists of six types of nodes: *Startup* nodes (**S**), *Person* nodes (**P**), *VC Firm* nodes (**V**), *Institute* nodes (**I**), *Location* nodes (**L**), and *Category* nodes (**C**); eight types of edges representing various relations between all types of nodes are defined: *Co-work* between Persons, *Investments* between Persons and Startups, *Investments* between VC Firms and Startups, *Employment* between VC Firms and Persons, *Foundation/Execution* between Persons and Startups, *Education* between Persons and Institutes, *Locate in*

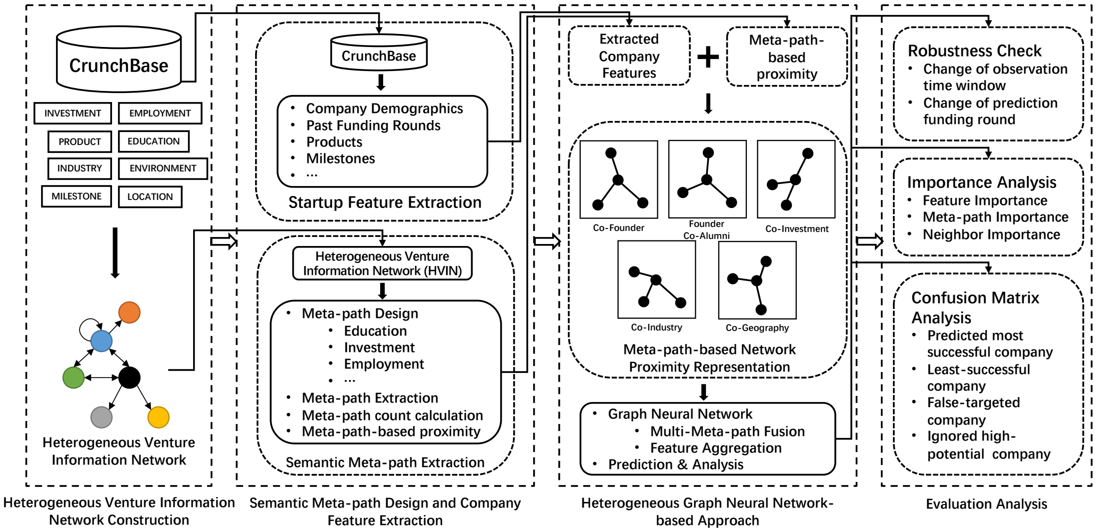
CrunchBase
INVESTMENT
EMPLOYMENT
PRODUCT
EDUCATION
INDUSTRY
ENVIRONMENT
MILESTONE
LOCATION
Heterogeneous Venture Information Network
Heterogeneous Venture Information Network Construction
CrunchBase
Company Demographics
Past Funding Rounds
Products
Milestones
...
Startup Feature Extraction
Heterogeneous Venture Information Network (HVIN)
Meta-path Design
Education
Investment
Employment
...
Meta-path Extraction
Meta-path count calculation
Meta-path-based proximity
Semantic Meta-path Extraction
Semantic Meta-path Design and Company Feature Extraction
Extracted Company Features
Meta-path-based proximity
Co-Founder
Founder Co-Alumni
Co-Investment
Co-Industry
Co-Geography
Meta-path-based Network Proximity Representation
Graph Neural Network
Multi-Meta-path Fusion
Feature Aggregation
Prediction & Analysis
Heterogeneous Graph Neural Network-based Approach
Robustness Check
Change of observation time window
Change of prediction funding round
Importance Analysis
Feature Importance
Meta-path Importance
Neighbor Importance
Confusion Matrix Analysis
Predicted most successful company
Least-successful company
False-targeted company
Ignored high-potential company
Evaluation Analysis

Figure 1: Overview of Our Proposed Framework

between Startups and Locations, and *Belong to* between Startups and Categories.

3.2. Meta-path Construction

Within the constructed HVIN, startups indirectly connect to each other through one or multiple types of nodes. Meta-path, which is used to capture the composite relations between two nodes on a graph (Sun et al., 2011), is a natural fit to measure the indirect relations between startups. Specifically, a meta-path between two startup nodes on the HVIN has a form of $Startup_1 \xrightarrow{R_1} A_1 \xrightarrow{R_2} \dots \xrightarrow{R_l} Startup_2$, where there is a composition of relations $R = R_1 \diamond R_2 \diamond \dots \diamond R_l$ between the two startups. $\diamond$ denotes the composition operation of relations, and the length of the meta-path is set as the number of composite relations, i.e., l. With the constructed HVIN, we define a total of 11 types of meta-paths that start from and end with startup nodes. We construct each type of meta-path based on the relevant literature. Each type of meta-path reflects a specific kind of composite relation and measures the proximity between startups. All the 11 types of meta-paths together capture the heterogeneous proximity between startups from different perspectives. We will build multiple adjacency graphs among startups based on these meta-paths and fuse the graphs using multi-head graph attention networks towards predicting startup success. Figure 2 shows a sub-graph of our constructed HVIN, which exemplifies all the elevent types of meta-paths.

3.3. Company Feature Construction

Our Heterogeneous Venture Information Network (HVIN) depicts the intricate relationship between various entities, such as firms, individuals, and institutes, among others. While these relationships can be helpful in predicting the success of a startup, certain characteristics of these entities can also play a crucial role in the outcome. For example, previous research (Sharchilev et al., 2018; Xu et al., 2022) has demonstrated that firm demographics, including the age of the company, its history of acquisitions, the number of products launched, and the company's online presence, can provide valuable insights into the current state of the firms. As a result, our study incorporates these characteristics to enhance the HVIN and deliver more precise predictions of startup success. More specifically, we have compiled a comprehensive list of 52 relevant features to facilitate our analysis, which is omitted due to the space limit. These features include key company demographics, such as firm age, size, industry, and number of employees. Our feature selection process is grounded in the latest research, ensuring that we are well-equipped to make informed predictions.

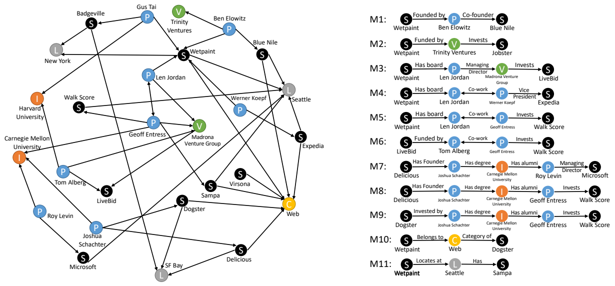
Gus Tai
Badgeville
Trinity Ventures
Ben Elowitz
Blue Nile
New York
Wetpaint
Len Jordan
Seattle
Harvard University
Walk Score
Werner Koepf
Geoff Entress
Madrona Venture Group
Expedia
Carnegie Mellon University
Tom Alberg
LiveBid
Virsona
Sampa
Roy Levin
Dogster
Web
Joshua Schachter
Microsoft
SF Bay
Delicious
M1: Wetpaint Founded by Ben Elowitz Co-founder Blue Nile
M2: Wetpaint Funded by Trinity Ventures Invests Jobster
M3: Wetpaint Has board Len Jordan Managing Director Madrona Venture Group Invests LiveBid
M4: Wetpaint Has board Len Jordan Co-work Werner Koepf Vice President Expedia
M5: Wetpaint Has board Len Jordan Co-work Geoff Entress Invests Walk Score
M6: LiveBid Funded by Tom Alberg Co-work Geoff Entress Invests Walk Score
M7: Delicious Has Founder Joshua Schachter Has degree Carnegie Mellon University Has alumni Roy Levin Managing Director Microsoft
M8: Delicious Has Founder Joshua Schachter Has degree Carnegie Mellon University Has alumni Geoff Entress Invests Walk Score
M9: Dogster Invested by Joshua Schachter Has degree Carnegie Mellon University Has alumni Geoff Entress Invests Walk Score
M10: Wetpaint Belongs to Web Category of Dogster
M11: Wetpaint Locates at Seattle Has Sampa

Figure 2: Instances of the Constructed HVIN and the Theory-guided Semantic Meta-paths

3.4. Definition of Startup Success

We have framed the problem of startup success prediction as a node classification task. However, the definition of a successful startup still remains unclear. In this section, we will discuss multiple indicators of success and provide our definition of a successful startup. According to Salamzadeh and Kawamorita Kesim (2015), each series funding round is crucial to startups as it helps them maintain fast-growing momentum. However, obtaining additional series funding rounds can be challenging, making each successful round a significant achievement for a startup. Aside from funding rounds, other signs of success for startups are IPOs and mergers and acquisitions (M&A) events (Hegde & Tumlinson, 2014; Hochberg et al., 2007; Xu et al., 2022). Therefore, we can also view IPOs and acquisitions as key metrics for gauging a startup's success. On the other hand, time is another critical factor in evaluating startup success. Numerous research studies have emphasized the significance of time when evaluating the success of startup companies (Bonaventura et al., 2020; Hochberg et al., 2007; Sharchilev et al., 2018; Zhang et al., 2021). Therefore, our definition of success for startups is as follows: *If a startup secures initial or series funding and then receives additional funding, goes public, or gets acquired during a specified observation time interval, it is considered a success.* It is important to note that this definition takes into account both the timing and the stage of the startup, and therefore a startup must be evaluated multiple times throughout its lifecycle.

3.5. Venture Graph Neural Network for Startup Success Prediction

With the constructed meta-paths and startup-related features, we are now ready to present our developed Venture Graph Neural Network (VenGNN) for predicting the defined startup success. Figure 3 shows the architecture of our proposed VenGNN. The VenGNN extracts meta-path-based proximities and company features, and combines them with a closeness-centrality-based encoding, generating encoded features. The encoded features and meta-path proximities are fed to a fused heterogeneous graph attention network for proximity fusion, and a random-walk-based graph sampling is fed to a self-attention block to capture long-distance information. The graph attention output and self-attention output are combined and fed to a feed-forward network (FNN) to generate the final prediction. Additionally, we combine transfer learning with the VenGNN to further enhance the prediction of startup success.

Meta-path Knowledge Fusion. Given the eleven types of meta-paths constructed in Subsection 3.3, we introduce how to fuse all meta-paths and quantify the closeness between startups. Specifically, we utilize the *PathCount* (Xu et al., 2022; Zhang et al., 2021), which is essentially the number of meta-path instances between startups, to measure the proximity among startups. For instance, if there are two instances of the M1 meta-path between two startups, their proximity is set as two. With each type of meta-path, we calculate the *PathCount* between every pair of startups and then build a homogeneous graph of startups. The homogeneous graph consists of startups as nodes and edges among them. An edge exists only when the PathCount between two nodes is greater than zero, and the weight on each edge is the value of PathCount. We construct eleven such homogeneous graphs based on the eleven types of meta-paths. In the next section, we will introduce the second part of VenGNN (i.e., multi-head graph attention network) that merges the eleven homogeneous graphs toward predicting startup success.

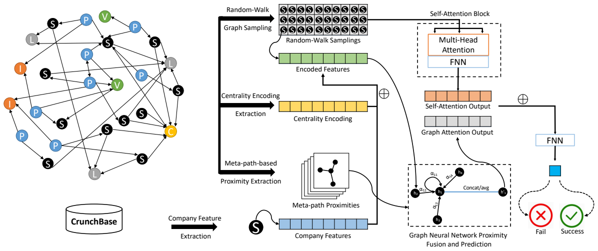

Figure 3: Architecture of Our VenGNN Model

Fused Heterogeneous Graph Attention Network. Graph Attention Network (GAT) (Veličković et al., 2018) is one of the state-of-the-art GNN models. The key feature of GAT is an attention mechanism that automatically learns the attentional weights among nodes. Each attentional weight indicates the importance of one node to another. Prior studies have shown that GAT outperforms most of the existing state-of-the-art GNNs on many prediction tasks (e.g., node classification and link prediction) with various heterogeneous graphs (Lv et al., 2021). In observation of the success of the attention mechanism, we develop a *fused heterogeneous attentional layer* that utilizes the attention mechanism to model the multiple homogeneous startup graphs that we have constructed based on the defined meta-paths. Specifically, given the startup node features $\mathbf{X}$ and M homogeneous startup graphs (M is 11 in our setting), we specify the fused heterogeneous attention layer as:

$$h_i = \mathop{\Big\|}_{m=1}^{M} \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_{m,i}} \alpha_{m,ij}^{k}\mathbf{W}_{m}^{k}X_j\Big), \qquad (1)$$

where $N_{m,i}$ denotes the i-th node's direct neighbors on the m-th homogeneous graph; K is the number of heads for stability concerns (Vaswani et al., 2017); h_i is the output representation of the i-th node; $\alpha()$ is the activation function; $||$ is the concatenation operation; $\alpha_{m,ij}^{k}$ and $\mathbf{W}_{m}^{k}$ represent the attention mechanism and the corresponding learnable parameters of the k-th head and m-th homogeneous graph, respectively. We stack two layers of the fused graph attention layer, where the second layer takes the first layer output as input.

Despite the success of the graph attentional layer shown in prior studies (Veličković et al., 2018; Wang et al., 2019; Zhang et al., 2021), it still faces a critical challenge, namely the over-smoothing issue, as do other GNN models (Oono & Suzuki, 2021). Over-smoothing causes the performance of GNN to decrease as the number of layers (e.g., attention layers) increases. It also indicates that ordinary GNN cannot capture and pass long-distance information over the graph.

Sampled Self-Attention. In order to further capture long-distance information and address over-smoothing, we propose a sampled self-attention design to learn additional node representations. The learned node representations will be combined with the outputs of the last graph attentional layer. The self-attention mechanism (Vaswani et al., 2017), a key design in sequence representation learning tasks, is formalized as:

$$\text{Self-Attention}(V) = \text{softmax}(\frac{VV^T}{\sqrt{d_v}})V, \qquad (2)$$

where $V \in \mathbb{R}^{L\times d}$ denotes the packed value matrix for self-attention; L is the sequence length; d is the feature dimension, and d_v is a scaling factor. As shown in Equation 2, the self-attention design passes messages globally over the value matrix. However, if such global self-attention is applied to all nodes on a graph, the computational cost can become exponentially expensive, even infeasible on large graphs with numerous nodes. To overcome this computational issue, we propose a random-walk-based sampling technique. Specifically, we conduct multiple-round random walk graph sampling starting with each startup node on the HVIN. Each round of random walk sampling results in a path over multiple startup nodes on the HVIN. Let us denote the multiple paths from the i-th startup node as $S_i = \{S_i^1, S_i^2, ..., S_i^b\}$, where b is the total number of paths. The length of each sampled path (i.e., the number of startup nodes on each path) is set as L. The i-th node's representation output through the self-attention can be denoted as:

$$g_i = \sum_{j=1}^{b}(W_i^2)^T \text{Self-Attention}(S_i^j W_i^1), \qquad (3)$$

where j denotes the j-th path; $W_i^1 \in \mathbb{R}^{d\times d'}$ is a learnable weight matrix that transforms node representations from the original feature space to the hidden space. The dimension of the hidden space is the same as that of h_i in Equation 1; $W_i^2 \in \mathbb{R}^{L\times 1}$ is another learnable parameter matrix that fuses the L nodes' representations of a path into one; the summation over b sampled paths ensures stability. The designed sampled self-attention passes messages over sampled node sequences, not only capturing long-distance information but also enhancing computational efficiency.

The combination of the output of the sampled self-attention layer (i.e., g_i) and that of the last attentional layer (i.e., h_i) is fed into a Feed-forward Neural Network (FNN). The last layer of the FNN is a Sigmoid activation function because our problem is a binary node classification problem. The output of the FNN can be denoted as: $y'_i =$ FNN $(g_i + h_i)$, where y'_i is the predicted likelihood of success for startup node i. Finally, we can formalize the learning objective of our VenGNN model with cross-entropy loss as:

$$\mathcal{L} = argmin_\Theta - \frac{1}{N}\sum_{i=1}^{N} y_i \, log(y'_i) + (1-y_i)\, log(1-y'_i), \qquad (4)$$

where N is the total number of nodes and y_i is the ground-truth label of the i-th node. Θ denotes all the learnable parameters of our VenGNN model. To solve the learning objective, we develop a learning algorithm based on widely used gradient descent and back-propagation techniques to estimate the optimum model parameters.

3.6. Transfer Learning for Enhancing Startup Success Prediction

The developed VenGNN model can be used for predicting startups' success across funding rounds and different time windows. The knowledge gained in previous predictions (e.g., success in Series-B funding round) could be useful for future predictions (e.g., success in Series-C). Therefore, we further apply transfer learning techniques to transfer knowledge across different startup success prediction tasks. For instance, when we train our VenGNN model for predicting startups' success in one funding round (e.g., Series-C), we use the parameters of the last FNN layer, learned in the previous training procedure (e.g., predicting startups' success in Series-B), as the initialization of the current training procedure. As revealed in prior studies (Torrey & Shavlik, 2010), machine learning models incorporating transfer learning techniques on related tasks start at a higher performance point and converge to a higher asymptote with a steeper slope. Thus, we expect that training our VenGNN model with transfer learning techniques could further enhance the prediction performance, which will be examined and demonstrated in the evaluation.

4. Empirical Evaluation

4.1. Experimental Setup

Data Description. We obtain the *Crunchbase 2013* snapshot, which is an instance of Crunchbase dataset when the database was updated to include information on companies and investors up to December 31, 2013. This snapshot serves as a historical record of the startup ecosystem at that time and is useful for analysts and researchers who want to understand how the industry was evolved over time. Our objective is to predict if a startup candidate could receive a next series funding round (e.g. Series-B), go IPO or get acquired within an observational time window (e.g., three years). Note that, to ensure sufficient information to capture the growth potential of startups, we will focus on startups that *have received at least one round of funding*. We also exclude any events or developments that occurred after the date of receiving funding to prevent information leakage. For example, we extract all the startups that received Series-A funding between Jan 1, 2008 and Dec 31, 2010. For each startup, we only preserve information happening on and before its date of receiving Series-A funding, and label this startup as positive if it receives Series-B, goes IPO, or gets acquired in the following three years, otherwise negative. We sort all targeted startups w.r.t. their dates of receiving Series-A funding, and select the earliest 60% startups as the training set, subsequent 20% as the validation set and the remaining 20% as the test set. This setting replicates real-world situations and reduces the risk of unintended information leakage.

To fully assess the efficacy of our proposed framework, we developed a range of datasets that concentrate on different time periods and funding rounds. Seven datasets were created, consisting of three sets focusing on performance across different funding rounds within the period of Jan 1, 2008 to Dec 31, 2013, and four sets focusing on the alteration of rolling time windows, each spanning six months, for a specific funding round (Series-A). In each dataset, we have presented a tabulation of the total count of different nodes, including investors, startups, persons, institutes, locations, and categories. In addition to nodes, we further provide the mean count of every meta-path (M1 to M11) identified in the dataset. The percentage of positive cases in each dataset is between 20% and 26%.

Benchmark Methods. We compare our approach against various benchmark methods as well as ablations of our own approach. Three groups of benchmarks are selected: (1) Feature-based approach that only considers startup features, including **Centrality** (Bonaventura et al., 2020), **LR** (Zbikowski & Antosiuk, 2021) and **XGBoost** (Xu et al., 2022); (2) Homogeneous GNN-based approaches, including **GCN** (Kipf & Welling, 2017) and **GAT** (Veličković et al., 2018); (3) Heterogeneous GNN-based approaches, including **HAN** (Wang et al., 2019) and **SHGMNN** (Zhang et al., 2021). Three ablated version of our approach is also included for comparison: **VenGNN-S** that removes the sampling-based self-attention module, **VenGNN-A** that removes the graph attention module, and **VenGNN-T** that removes the transfer learning module.

Evaluation Metrics. We select three groups of evaluation metrics for comprehensive examining the approaches. The first group is overall correctness, including *AUC, AUPR, F-1* and *Accuracy*. The second and third group are *Precision@K* and *ROI@K*, where $K = 5, 10, 20$. In our context, *ROI@K* refers to the top-K ranked return over investment (ROI). In general, it is calculating the average ROI of top-K ranked predictions. Since we are predicting mainly on if startups could receive next round of funding, the ROI is defined as:

$$ROI = \begin{cases} \frac{\text{Post-Valuation - Pre-Valuation}}{\text{Pre-Valuation}} & \text{if receives next round} \\ -1 & \text{if not} \end{cases} \tag{5}$$

It is usually inaccessible to startups' valuation data, thus we estimate startup valuation using the money raised (accessible in the dataset) and estimated dilutions during a specific round:

$$\text{Post-Valuation} = \text{Raised Money}/\text{Estimated Dilution} \tag{6}$$

$$\begin{aligned} \text{Pre-Valuation} &= (1 - \text{Estimated Dilution}) \\ &\times \text{Raised Money}/\text{Estimated Dilution} \end{aligned} \tag{7}$$

Here the estimated dilution w.r.t. each targeted funding round (Series-B,C,D) is $19\%, 16.25\%, 13.5\%$, based on the mean value generated by Radicle, a disruptive research firm using a statistical model to estimates start-up valuations from their Funding Round stage and amount raised [3]. In our analysis, to ensure the robustness of the results, each experiment was repeated 10 times and the mean values of the metrics are reported.

4.2. Experimental Results

Next, we provide an in-depth discussion of the main experimental results. We evaluate these methods over seven datasets using a variety of performance metrics. To begin with, we showcase the performance of all methods on three of our main datasets ($\mathcal{D}_{\mathbf{A}\to\mathbf{B}}$, $\mathcal{D}_{\mathbf{B}\to\mathbf{C}}$, and $\mathcal{D}_{\mathbf{C}\to\mathbf{D}}$) in Table 1. We have made several noteworthy observations. First, it is important to note that our problem at hand is non-trivial. In particular, we examine the performance of LR and XGBoost, which are popular feature-based supervised models applied in various contexts. Upon closer inspection, we found their performance are poor, with AUC values hovering around 50%, barely distinguishable from a random classifier. This underscores the need for more sophisticated models to better tackle our focal problem's nuances.

[3] https://finerva.com/report/dilution-data-funding-rounds/

Dataset	$\mathcal{D}_{\mathbf{A}\rightarrow\mathbf{B}}$									
Metrics	**Overall Correctness**				**Precision@K**			**ROI@K**		
	AUC	$AUPR$	$Accuracy$	$F1$	P@5	P@10	P@20	ROI@5	ROI@10	ROI@20
Centrality	50.43	41.83	52.07	40.46	44.00	47.00	40.00	79.92	81.00	51.66
LR	55.85	45.41	55.85	45.15	54.00	49.00	48.00	52.63	47.13	53.65
XGBoost	55.20	46.19	56.53	**46.00**	50.00	52.00	48.50	69.20	66.20	78.11
GCN	58.54	49.35	57.74	33.52	58.00	70.00	62.50	75.30	187.97	121.56
GAT	62.60	53.22	60.19	36.94	68.00	67.00	**63.00**	59.31	80.84	65.71
HAN	59.19	46.89	58.48	35.47	60.00	60.00	**63.00**	99.15	121.99	123.53
SHGMNN	61.81	48.44	58.98	36.58	60.00	57.00	59.50	99.78	89.14	98.79
VenGNN-S	69.20	57.14	63.19	42.99	60.00	60.00	**63.00**	99.15	121.99	123.53
VenGNN-A	67.39	55.84	62.79	42.35	60.00	64.00	52.00	87.16	132.24	72.18
VenGNN	**69.57**	**58.85**	**63.22**	43.52	**82.00**	**81.00**	62.00	**258.74**	**237.52**	**152.36**

Dataset	$\mathcal{D}_{\mathbf{B}\rightarrow\mathbf{C}}$									
Metrics	**Overall Correctness**				**Precision@K**			**ROI@K**		
	AUC	$AUPR$	$Accuracy$	$F1$	P@5	P@10	P@20	ROI@5	ROI@10	ROI@20
Centrality	49.59	41.21	54.14	34.68	38.00	39.00	38.00	37.81	23.78	5.60
LR	54.07	45.39	55.56	36.69	44.00	51.00	50.50	16.50	16.57	31.95
XGBoost	49.86	41.63	53.84	34.24	46.00	39.00	38.00	2.40	-12.99	-14.19
GCN	60.24	52.25	64.04	42.67	68.00	66.00	54.50	109.60	92.88	46.59
GAT	62.12	54.15	62.73	38.98	76.00	56.00	49.00	131.80	73.63	45.02
HAN	57.80	51.23	59.49	33.65	36.00	39.00	47.50	167.55	96.92	**119.70**
SHGMNN	62.98	53.48	59.39	34.14	90.00	65.00	57.50	206.64	102.88	55.61
VenGNN-S	66.61	53.24	64.44	42.90	78.00	**79.00**	61.50	110.57	99.20	51.53
VenGNN-A	68.58	56.83	63.74	40.16	70.00	49.00	42.00	104.77	43.97	20.44
VenGNN-T	68.25	59.25	62.73	39.77	78.00	73.00	**62.00**	183.04	112.73	89.18
VenGNN	**69.19**	**60.71**	**65.32**	**43.44**	**96.00**	61.00	61.50	**233.99**	**132.09**	90.88

Dataset	$\mathcal{D}_{\mathbf{C}\rightarrow\mathbf{D}}$									
Metrics	**Overall Correctness**				**Precision@K**			**ROI@K**		
	AUC	$AUPR$	$Accuracy$	$F1$	P@5	P@10	P@20	ROI@5	ROI@10	ROI@20
Centrality	49.98	46.97	51.88	46.28	46.00	45.00	42.50	-9.21	-11.97	-14.01
LR	55.05	50.27	53.33	47.91	46.00	50.00	51.50	-13.82	-5.78	-4.30
XGBoost	54.27	47.61	52.92	47.44	40.00	41.00	44.00	-18.61	-18.26	-12.46
GCN	55.86	51.01	58.33	49.03	60.00	63.00	67.00	-2.89	9.92	21.09
GAT	62.34	56.40	56.88	38.13	74.00	66.00	57.50	31.62	18.48	5.61
HAN	56.47	54.27	53.54	35.48	58.00	56.00	61.00	-2.14	-0.80	8.92
SHGMNN	52.85	51.64	52.29	34.12	62.00	62.00	58.00	29.31	24.14	14.51
VenGNN-S	63.80	**61.12**	56.46	39.54	78.00	69.00	**69.00**	64.75	**42.32**	**35.06**
VenGNN-A	63.06	59.03	58.13	39.03	74.00	61.00	52.00	40.50	17.82	3.71
VenGNN-T	66.62	58.42	60.63	44.88	**84.00**	**71.00**	66.50	52.24	36.44	34.61
VenGNN	**66.86**	59.46	**61.42**	**45.87**	**84.00**	66.00	55.00	**75.30**	38.86	12.90

Notes: The best results are highlighted in bold. All numbers are shown in % format.

Table 1: Overall Performance of Various Models

Second, we observe that our proposed VenGNN consistently outperforms all the benchmark methods regarding multiple performance metrics. Aside from the feature-based methods, stronger benchmark methods include homogeneous GNNs (GCN and GAT) and heterogeneous GNNs (HAN and SHGMNN). Our VenGNN model achieves the best AUC and AUPR, with the largest margin being 20% compared with the second best model (SHGMNN). Similar observations can be made for other performance metrics, such as F-1 and Accuracy. Our model also outperforms other models by at least 30% in terms of Precision@K. Meanwhile, attributed to its high predictive power, our VenGNN model can achieve an ROI that is on average twice the return by other models. This strongly demonstrates the superiority of our model over other benchmark methods.

We conducted ablation studies to better understand the effectiveness of different modules in our model. We first analyze the two main designs in our model architecture: the sampled self-attention and the graph attention modules. Upon examining the performance of the models while excluding the two modules respectively (denoted as VenGNN-S and VenGNN-A), we found that they exhibit comparable performance. However, when compared with VenGNN, both models have slightly lower AUC or AUPR values and significantly lower Precision@K and ROI@K. We can conclude with confidence that the sampled self-attention and graph attention modules are indispensable and crucial components of our model architecture.

Another effort we have made on improving our model performance is to incorporate transfer learning into our model. Transfer learning allows our model to learn from the models built for previous funding rounds and better predict future funding rounds. To demonstrate the effectiveness, we compare the performance of our VenGNN model with and without transfer learning module (denoted as VenGNN and VenGNN-T, respectively). Note that the transfer learning module does not provide any added value to the prediction from Series-A to Series-B investment rounds (in the case of $\mathcal{D}_{\mathbf{A}\rightarrow\mathbf{B}}$), as there are no knowledge of prior models to learn from. But for the other two cases ($\mathcal{D}_{\mathbf{B}\rightarrow\mathbf{C}}$ and $\mathcal{D}_{\mathbf{C}\rightarrow\mathbf{D}}$ in Table 1), we can observe that the transfer learning module indeed improves the performance of our model, especially in terms of Precision@K and ROI@K.

4.3. Robustness Check

When investigating the investment activities, we have been focusing on a fixed time period in our

Metrics	**AUC**				**Precision@5**				**ROI@5**			
Dataset	$\mathcal{D}^{-4}_{\mathbf{A\to B}}$	$\mathcal{D}^{-3}_{\mathbf{A\to B}}$	$\mathcal{D}^{-2}_{\mathbf{A\to B}}$	$\mathcal{D}^{-1}_{\mathbf{A\to B}}$	$\mathcal{D}^{-4}_{\mathbf{A\to B}}$	$\mathcal{D}^{-3}_{\mathbf{A\to B}}$	$\mathcal{D}^{-2}_{\mathbf{A\to B}}$	$\mathcal{D}^{-1}_{\mathbf{A\to B}}$	$\mathcal{D}^{-4}_{\mathbf{A\to B}}$	$\mathcal{D}^{-3}_{\mathbf{A\to B}}$	$\mathcal{D}^{-2}_{\mathbf{A\to B}}$	$\mathcal{D}^{-1}_{\mathbf{A\to B}}$
Centrality	47.34	47.65	48.78	47.15	34.00	42.00	38.00	60.00	9.04	17.78	121.49	106.66
LR	57.03	52.03	57.96	57.53	40.00	42.00	52.00	62.00	17.32	42.26	360.52	68.75
XGBoost	55.11	51.29	55.61	50.78	32.00	52.00	40.00	50.00	12.10	59.63	82.91	59.85
GCN	60.40	56.40	63.02	53.61	58.00	64.00	62.00	56.00	39.12	79.06	120.53	157.30
GAT	61.70	50.90	58.52	56.50	52.00	70.00	54.00	68.00	29.00	96.23	97.20	107.16
HAN	55.80	53.70	58.49	57.99	50.00	58.00	62.00	64.00	45.20	68.20	146.77	149.53
SHGMNN	58.80	57.00	53.58	56.92	56.00	50.00	**68.00**	52.00	67.07	54.59	93.56	55.97
VenGNN-S	64.10	60.60	60.50	60.30	60.00	72.00	48.00	62.00	89.10	97.20	118.10	214.00
VenGNN-A	60.70	61.10	63.56	61.59	**76.00**	78.00	62.00	84.00	121.42	142.24	208.25	205.25
VenGNN-T	N/A	**61.90**	63.41	65.40	N/A	74.00	**68.00**	76.00	N/A	119.80	**500.90**	280.58
VenGNN	**66.50**	61.30	**63.58**	**65.91**	**76.00**	**80.00**	**68.00**	**90.00**	**151.15**	**384.91**	379.40	**366.36**

Notes: The best results are highlighted in bold. All numbers are shown in % format.

Table 2: Overall Performance of Various Models (Varying Time Window)

experimental analysis. To further demonstrate the robustness of our proposed approach, we conduct experiments by changing the time period in the form of rolling windows with a targeted funding round (Series-A to Series-B). More specifically, we constructed four datasets, i.e., $\mathcal{D}^{-1}_{\mathbf{A\to B}}$, $\mathcal{D}^{-2}_{\mathbf{A\to B}}$, $\mathcal{D}^{-3}_{\mathbf{A\to B}}$, and $\mathcal{D}^{-4}_{\mathbf{A\to B}}$, with each rolling back six months consecutively. Using these datasets, we evaluate all models and report the AUC score, Precision@5 and ROI@5, as shown in Table 2. In comparison of our VenGNN model with other models, we have similar observations as earlier and our model consistently outperforms other models, especially in terms of Precision@K and ROI@K.

We see notable improvements of performance from the ablated models (i.e., VenGNN-S and VenGNN-A) to VenGNN, with a margin of 10%. This reinforces the importance of the two modules. The ablation analysis for the transfer learning module is conducted slightly different from the previous case. Essentially, the idea is to leverage the knowledge obtained from models built using earlier datasets and apply it to the later ones. When comparing VenGNN-T to VenGNN, it is important to note that the performance improvement is not as significant as in the previous case. However, it still demonstrates the effectiveness of incorporating transfer learning into the startup success prediction process across consecutive time windows.

5. Conclusion

In this paper, we introduce a novel venture graph neural network (VenGNN) approach to predict high-potential startups, with multiple contributions. First, we construct a heterogeneous venture information network (HVIN) from a publicly available startup database and define multiple types of meta-paths based on relevant theory and findings. The constructed HVIN and meta-paths serve as the basis for our developed VenGNN approach. Second, we design a novel fused heterogeneous attentional layer for modeling multi-graph data and integrate centrality encoding and sampled self-attention techniques for addressing the over-smoothing issue. Third, we further employ transfer learning techniques for transferring useful knowledge across different model training processes. We validate the superiority of our methodological designs using intensive empirical evaluations with a unique dataset from Crunchbase and demonstrate the interpretability of our VenGNN model with insightful analysis.

References

Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2021). Supervised learning for the prediction of firm dynamics. In *Data science for economics and finance: Methodologies and applications* (pp. 19–41). Springer International Publishing Cham.

Bonaventura, M., Ciotti, V., Panzarasa, P., Liverani, S., Lacasa, L., & Latora, V. (2020). Predicting success in the worldwide start-up network. *Scientific Reports*, *10*(1), 1–6.

Gloor, P. A., Dorsaz, P., Fuehres, H., & Vogel, M. (2013). Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure. *International Journal of Organisational Design and Engineering*, *3*(1), 67–85.

Gompers, P., Kovner, A., & Lerner, J. (2009). Specialization and success: Evidence from venture capital. *Journal of Economics & Management Strategy*, *18*(3), 817–844.

Hadley, B., Gloor, P. A., Woerner, S. L., & Zhou, Y. (2018). Analyzing vc influence on startup success: A people-centric network theory approach. In *Collaborative innovation networks* (pp. 3–14). Springer.

Hegde, D., & Tumlinson, J. (2014). Does social proximity enhance business partnerships? theory and evidence from ethnicity's role in us venture capital. *Management Science*, *60*(9), 2355–2380.

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance*, *62*(1), 251–301.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. *IEEE International Conference on Data Mining Workshops*, 798–805.

Lv, Q., Ding, M., Liu, Q., Chen, Y., Feng, W., He, S., Zhou, C., Jiang, J., Dong, Y., & Tang, J. (2021). Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1150–1160.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Nanda, R., Samila, S., & Sorenson, O. (2020). The persistent effect of initial success: Evidence from venture capital. *Journal of Financial Economics*, *137*(1), 231–248.

Oono, K., & Suzuki, T. (2021). Graph neural networks exponentially lose expressive power for node classification.

Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Editor's comments: Diversity of design science research. *MIS Quarterly*, *46*(1), iii–xix.

Rai, A. (2017). Editor's comments: Diversity of design science research. *MIS Quarterly*, *41*(1), iii–xviii.

Salamzadeh, A., & Kawamorita Kesim, H. (2015). Startup companies: Life cycle and challenges. *4th International conference on employment, education and entrepreneurship (EEE), Belgrade, Serbia*.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based startup success prediction. *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2283–2291.

Song, Y., & Vinig, T. (2012). Entrepreneur online social networks–structure, diversity and impact on start-up survival. *International Journal of Organisational Design and Engineering 3*, *2*(2), 189–203.

Stone, T. R. (2014). *Computational analytics for venture finance* (Doctoral dissertation). University College London.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, *4*(11), 992–1003.

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI global.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network. *Proceedings of The Web Conference*, 2022–2032.

Xu, R. R., Chen, H., & Zhao, J. L. (2022). Sociolink: Leveraging relational information in knowledge graphs for startup recommendations. *Journal of Management Information Systems*, *Forthcoming*.

Yang, K., Lau, R. Y. K., & Abbasi, A. (2023a). Getting personal: A deep learning artifact for text-based measurement of personality. *Information Systems Research*, *34*(1), 194–222.

Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023b). Unlocking the power of voice for financial risk prediction: A theory driven deep learning design approach. *MIS Quarterly*, *47*(1).

Zbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using crunchbase data. *Information Processing & Management*, *58*(4), 102–555.

Zhang, S., Zhong, H., Yuan, Z., & Xiong, H. (2021). Scalable heterogeneous graph neural networks for predicting high-potential early-stage startups. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2202–2211.

Zhong, H., Liu, C., Lu, X., & Xiong, H. (2016). To be or not to be friends: Exploiting social ties for venture investments. *IEEE International Conference on Data Mining (ICDM)*, 699–708.

Zhong, H., Liu, C., Zhong, J., & Xiong, H. (2018). Which startup to invest in: A personalized portfolio strategy. *Annals of Operations Research*, *263*(1), 339–360.