

Firm Profiling and Competition Assessment: A Heterogeneous Occupation Network-based Method

Hao Zhong
 ESCP Business School
hzhong@escp.eu

Chuanren Liu
 The University of Tennessee, Knoxville
cliu89@utk.edu

Abstract

Extensive efforts have been made by both academics and practitioners to understand inter-firm competitive relationship owing to its profound impacts on multiple key business goals. However, it has never been an easy task to fully characterize firms and assess the competitive relationship among them mainly due to the challenge of information heterogeneity. In this regard, we propose a novel IT artifact for firm profiling and inter-firm competition assessment guided by Information System Design Theory (ISDT). We start by constructing a Heterogeneous Occupation Network (HON) using employees' occupation details and education attainments. Then we adopt a Methpath2Vec-based heterogeneous network embedding model to learn firms' latent profiles (embeddings). Using the firm embeddings as input, we train multiple supervised classifiers to assess the competitive relationship among the firms. Following the logic of design as a search process, we demonstrate the utility of our IT artifact with extensive experimental study and in-depth discussions. Our study also reveals that employees' occupation and education information significantly contribute to the identification of the focal firm's potential competitors.

Keywords: firm profiling, inter-firm competition assessment, heterogeneous network embedding, design science research, information system design theory.

1. Introduction

Competitor analysis, as an essential component of corporate strategy, has long been a critical yet challenging task to business. It is the process of gathering information about rival firms and evaluating their strengths and weaknesses in relation to your own company's position. This information can then be used to develop strategies for how to best compete against those rivals. As the competitive landscape of the business world becomes more dynamic, it is vital

for businesses to get a knowledge advantage through competition analysis. In particular, to the task of competitor analysis, competitor identification (CI) is a critical precursor and serves as the starting point of forming competitive strategies (Bergen & Peteraf, 2002). Timely and precise identification of potential competitors has considerable impacts on several key tactical and strategic business goals, e.g., company benchmarking and marketing strategy planning.

From managerial perspectives, two approaches have been proposed for the identification of competitors, i.e., *supply-based* and *demand-based* (Clark & Montgomery, 1999). The *supply-based* approach identifies competitors by assessing similar firms in terms of strategy, technology and products, while the *demand-based* approach recognizes potential competitors by leveraging customer attitudes and perceptions. Despite that the *demand-based* approach is popular in identifying competing products or services in marketing (Zhang et al., 2013), the *supply-based* approach is favored for the identification of competing firms in the literature (Li et al., 2006; Pant & Sheng, 2009, 2015). In line with the *supply-based* approach, firm resources have been studied extensively in understanding firm's competitive advantages, especially following the *resource-based view (RBV)* (Barney, 1991). Meanwhile, Chen (1996) argued that resource similarity is one of the two vital firm-specific factors in conducting competitor analysis (so as for competitor identification). Human resources, as a bundle of human capital in a direct employment relationship with the firm (Barney, 1991), can contribute to firm's sustained competitive advantages (Lado & Wilson, 1994). Thus, the assessment of the human resources proximity across multiple firms can serve as a tangible proxy in understanding their competitive relationship.

Following that, numerous studies have attempted to perform competitor analysis by leveraging human resources across firms (Li, 2017; Liu et al., 2020; Pennings & Wezel, 2007; Zhang et al., 2020). In this study, we approach this problem through a

different lens. The main goal of our study is to design an IT artifact which can consume employees' heterogeneous information across firms and perform the assessment of inter-firm competitions. Attributed to the highly recognized design science research literature (Hevner et al., 2004; Walls et al., 1992), we frame our study based on the popular *Information System Design Theory (ISDT)* proposed by Walls et al. (1992). This framework consists of four components, i.e., *kernel theory*, *meta-requirement*, *meta-design*, and *testable hypotheses*. We base our design on the following theory proposed by Pennings and Wezel (2007): *Inter-firm personnel mobility is a form of transformation of firms' competitive surroundings, which increases the similarity of both routines and resources across organizations, as well as their competitive interdependence*. The requirement of our IT artifact is to study how employees' heterogeneous information can project inter-firm competitions. Our design process is driven by a pivotal work of managerial competitors identification from (Clark & Montgomery, 1999). It proposed a schematic model for competitor identification, which consists of 1) forming firms' representations; 2) retrieving competitor categories from memory; and 3) evaluating similarity of target firms to category representations. Following their model, we first construct a *Heterogeneous Occupation Network (HON)* and learn the latent representations of firms using a *Metapath2Vec*-based heterogeneous network embedding model. Then we assess inter-firm competitions in two ways: 1) evaluating the representations similarities and 2) employing a broad set of machine learning classifiers. In contrast to Clark and Montgomery (1999)'s work, our design does not require explicit competitor categories and relies mainly on pairwise firm representations. Lastly, following the principle of *design as a search process* (Hevner et al., 2004), we propose several *testable hypotheses* to guide our IT artifact design and meanwhile to validate its efficacy.

Our study focuses on *IT industry* due to its high inter-firm competition and rich labor market dynamics in modern business environment. We performed extensive experimental analyses on our model using a unique dataset sourced from *LinkedIn* and *Owler*¹. The contributions of our study are in three folds. First, we contribute to the literature of design science research given that our IT artifact is designed by following the framework of *Information System Design Theory (ISDT)*. Second, we also contribute to the emerging strand of work on inter-firm competition analysis. We design and implement an IT artifact

¹<https://corp.owler.com/>

for firm profiling and competitors identification by utilizing a heterogeneous network embedding model. To our knowledge, we are the first to apply heterogeneous network embedding models in tackling the problems alike. Third, our study contributes to validating the feasibility of using employees' heterogeneous information (occupation details and education attainments) for assessing general inter-firm competitions using data from *LinkedIn* and *Owler*.

2. Related Work

2.1. Inter-firm competition and human resources

As a major *demand-based* approach, a large body of work focused on understanding organizational behavior and sustained competitive advantages through the lens of *firm resources*, more broadly known as the *resource-based view (RBV)* of firms (Barney, 1991; Conner & Prahalad, 1996). The concept of *firm resources* is adopted from Daft (1983): all assets, capabilities, competencies, organizational processes, firm attributes, information, and knowledge that enable the firm to conceive of and implement strategies to improve its efficiency and effectiveness. According to Barney (2014), firm resources can be divided into four categories: financial capital, physical capital, human capital, and organizational capital. However, not all these resources contribute to firm's sustained competitive advantages. To possess this potential, a firm resource must be valuable, rare, inimitable, and have no adequate substitutes (Barney, 1991).

Is human resource a credible *proxy* through which we can assess inter-firm competitions? Lado and Wilson (1994) argued that the contributions of human resources originate from facilitating the development of competencies that are firm-specific and embedded in a firm's history and culture, and generate tacit organizational knowledge. Moreover, Wright et al. (1994) adopted the theoretical concepts from the *RBV* to affirm that human resources meet the aforementioned requirements: they are valuable, rare, inimitable, and non-substitutable. Along this line, one pivotal work by Pennings and Wezel (2007) discussed the transformation of a firm's competitive surroundings through inter-firm employee mobility. They argued that inter-firm mobility will cause organizational routines replication and therefore the rise of competitive implications. Their discussions and arguments lay a great theoretical foundation to our study, about which we will have more in-depth discussions in Section 3. From a different perspective, Li (2017) studied a unique competitive relationship between firms, i.e., competing for talents in the labor markets, and developed the concept of

labor market peer firms. The author proposed a unique measure which segments firms based on their labor market similarities, in contrast to using standard industry codes. She found that the identified labor market peer firms have higher potential in revealing economical linkages between firms. Thus, it is of sufficient evidence that human resource can serve as a credible proxy to understanding inter-firm competitions.

2.2. Data-driven inter-firm competition analysis

In this part, we will discuss various strands of work on data-driven inter-firm competition analyses. For example, Bao et al. (2008) proposed a novel algorithm, *CoMiner*, to automatically mine competitors from the Web. Their work viewed firm co-occurrences on media reports as evidence of competition. Later, Pant and Sheng (2009) addressed the competitor identification problem using some novel web metrics extracted from in-links, out-links as well as texts of websites. Furthermore, Zhang et al. (2012b) designed and developed a novel probabilistic generative model for latent business relationship mining. Their proposed method can discover evolving latent business networks over time. As a follow-up study, they proposed a semi-supervised method to identify business entities and their relationships (Zhang et al., 2012a).

Most recently, studies on the use of human resource information for firm competitor identification have emerged. Liu et al. (2020) performed an inter-firm labor market competitor analysis by leveraging a human capital flow network constructed using a longitudinal employer-employee matched dataset. Their main focus is to predict future labor market competition. One major difference between this work and ours lies at the target of the study: Liu et al. (2020) focuses *uniquely* on the inter-firm competition of labor market and the objective of the model is to predict future human capital flow, while our analysis is to study how human resources can predict *general* inter-firm competitions. In addition, Liu et al. (2020)'s method is developed on the basis of *human capital flows* which are considered aggregated information of employees' occupational data. Whereas, with the aid of heterogeneous network embedding models, we are able to analyze the *raw* occupational data (as well as education information) and leverage their predictive capabilities on inter-firm competitions. The advantage of using heterogeneous network embedding models will be discussed in our experimental studies in Section 4. As another relevant example, Zhang et al. (2020) also utilized human capital flow (*talent flow*) to assess competitions between companies. Despite that they developed a more advanced model (*Talent Flow*

Embedding (TFE) to measure pairwise competitive relationship of firms, the underlying network still relies on the aggregated flows, alike Liu et al. (2020)'s work. And the *competitiveness* that their model attempts to assess is *self-defined* using human capital flow metrics and therefore subjective, which fails to reflect the actual inter-firm competitive relationship. Our work relies on real and credible competitor labels (from *Owler*) for the evaluation of our model utility.

2.3. Heterogeneous network embedding models

In our study, we mainly focus on *heterogeneous network embedding* models (Shi et al., 2016; Yang et al., 2020). As an overview, Yang et al. (2020) summarizes and evaluates existing studies on *heterogeneous network embedding (HNE)*, covering proximity-preserving methods, message-passing methods, and relation-learning methods. Proximity-preserving methods capture the network's topological information by preserving different types of proximity among nodes (e.g., *Metapath2Vec* (Dong et al., 2017) and *PTE* (Tang et al., 2015)). In contrast, message-passing methods aim to learn node embeddings by aggregating the information from neighbors, e.g., *HAN* (Wang et al., 2019) and *HetGNN* (Zhang et al., 2019). Relation-learning methods are widely adopted in knowledgebase (KB) embedding. It is to learn a scoring function which evaluates an arbitrary triplet and outputs a scalar to measure the acceptability of this triplet. Popular methods in this category include TransE (Bordes et al., 2013) and ConvE (Dettmers et al., 2018). Yang et al. (2020) performed rigorous and detailed evaluations of the aforementioned models for two major tasks in network analysis: node classification and link prediction. Their study showed that overall, *Metapath2vec* is superior or at least comparable to most other models, including message-passing methods and relation-learning methods. We therefore opt for the *Metapath2Vec-based* model to learn firm latent representations in our study.

3. Kernel Theory-based Design

3.1. Kernel theory

The *kernel theory* for our IT artifact design originates from (Pennings & Wezel, 2007). The authors pointed out that *inter-firm personnel mobility is a form of transformation of firms' competitive surroundings, which will cause organizational routines replication and therefore the rise of competitive implications.* *Routines* are patterned sequences of learned behavior involving multiple actors who are linked by relations of

communication and/or authority (Cohen & Bacdayan, 1994). There is a high propensity toward routine replications when individuals migrate from one firm to the other. Meanwhile, industry experience and educational attainment also act as surrogate indicators of the ability and competence of human capital (Pennings & Wezel, 2007). We therefore aim to investigate the extent to which employees' occupation and education details can distill inter-firm competitions.

3.2. Meta-requirements and meta-design

Meta-requirements are referred to as the class of goals to which the kernel theories apply. Our *meta-requirements* are to develop an effective system for assessing firms' competitive relationships based on collective information of their employees' occupation and education. Following our *kernel theory*, collective information about human capital and personnel mobility has significant implications for the firm's competitive surroundings. In other words, employees' occupation details presumably unveil firms' internal essentials and their positions in business competition environment.

In addition, *Meta-design* aims to construct a class of IT artifacts which can meet the *meta-requirements* in a rigorous manner. Our *meta-design* process operationalizes the theoretical framework of managerial competitors identification from (Clark & Montgomery, 1999). According to their theory, the identification of firms' competitors is achieved by 1) forming firms' representations; 2) retrieving competitor categories; and 3) evaluating the similarity of target firms to category representations. We adopt and improve their framework to realize our IT artifact design given the uniqueness of our *meta-requirements*. Following Clark and Montgomery (1999)'s model, obtaining firms' representations is the critical first step. But unfortunately, no explicit and measurable constructs have been developed in their study. We thus attempt to operationalize the notion of firms' representations as numeric vectors (i.e., firm embeddings) by leveraging the power of the heterogeneous network embedding models. Meanwhile, we assess inter-firm competition through the lens of human capital information, rather than as the similarity of industry or market categories. Thus, we do not need to include the explicit *category* in the operation (as the second step of Clark and Montgomery (1999)'s framework). Instead, our assessment of inter-firm competitions is in two ways: 1) by measuring the similarity of their firm embeddings; or 2) by developing predictive models using supervised classifiers. We elaborate the details in the following.

3.2.1. The design of Heterogeneous occupation network (HON) Following Yang et al. (2020)'s summary on heterogeneous networks, we formally present the notion of our unique *Heterogeneous Occupation Network (HON)* here.

Definition 1 (Heterogeneous Occupation Network)

A *Heterogeneous Occupation Network (HON)* is defined as $\mathbf{G}_O = (\mathbf{V}_O, \mathbf{E}_O, \mathbf{A}_O, \mathbf{C}_O)$. $\mathbf{A}_O = \{F, P, J, S\}$ where each denotes, respectively, Firm (F), Employee (P), Job (J) and School (S) and $\mathbf{C}_O = \{offer, hire, work, study\}$ which includes $F \longleftrightarrow J$ (offer), $F \longleftrightarrow P$ (hire), $P \longleftrightarrow J$ (work), and $P \longleftrightarrow S$ (study).

Here, *Firms* (F) can offer different *Jobs* (J) and hire a number of *Employees* (P). *Employees* (P) may have studied in one or more *Schools* (S) and are currently working or formerly worked on single or multiple *Jobs* (J). In such a network, a *metapath* traverses multiple nodes and edges. The formal definition of *metapath* is given below.

Definition 2 (Metapath) A *Metapath* \mathcal{P} is defined on the network schema G denoted in the form of $a_1 \xrightarrow{c_1} a_2 \xrightarrow{c_2} \dots \xrightarrow{c_L} a_{L+1}$, where $a_l \in \mathbf{A}$ are node types and $c_l \in \mathbf{C}$ are link types, respectively, and $1 \leq l \leq L$.

For simplicity, we can use a sequence of node types to denote the meta path, e.g., $\mathcal{P} = (a_1 a_2 \dots a_{L+1})$. As such, in our proposed HON network, one typical example of *metapath* is $F \xrightarrow{offer} J \xrightarrow{worked.by} P \xrightarrow{study} S$, which can be denoted ($FJPS$). Note that the links are *undirected* in our HON network.

3.2.2. Firm representation learning

Definition 3 (Heterogeneous Network Embedding)

For a given heterogeneous network $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A}, \mathbf{C})$, *heterogeneous network embedding* is a mapping function $\mathcal{F} : \mathbf{V} \mapsto \mathbb{R}^{|\mathbf{V}| \times d}$ ($d \ll |\mathbf{V}|$), which defines the latent representation of each node $v \in \mathbf{V}$ and captures network topological information in \mathbf{E} .

As for the heterogeneous network embedding task, we opt for a random walk-based model, *Metapath2Vec* (Dong et al., 2017), mainly due to its effectiveness and evidenced competency in a variety of applications (Yang et al., 2020). *Metapath2Vec* is a variant of *Node2Vec*, its homogeneous network embedding counterpart (Grover & Leskovec, 2016), both of which root in the idea of a word embedding model, *Skip-gram based Word2Vec* (Mikolov et al., 2013), from the natural language processing (NLP) research community. Simply put, given a sentence, the continuous Skip-gram model uses a focal word as

an input to a log-linear classifier with a continuous projection layer of Feedforward NNLM (Neural Net Language Model), and predicts words within a certain range (context window) before and after the focal word. The model’s final outputs are word-specific latent representations (word embeddings). Following this idea, *Metapath2Vec* first exploits metapath-guided random walks on the network to generate “sentences”. Given a metapath scheme $\mathcal{P} : a_1 \xleftrightarrow{c_1} a_2 \xleftrightarrow{c_2} \dots \xleftrightarrow{c_L} a_{L+1}$, the transition probability at step l is defined as:

$$P(v_{l+1} | v_l, \mathcal{P}) = \begin{cases} \frac{1}{|\mathcal{N}_l(v_l)|} & \phi(v_{l+1}) = a_{l+1}, \psi(v_l, v_{l+1}) = c_l \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathcal{N}_l(v_l) = \{u | \psi(u, v_l) = c_l\}$. The flow of the random walker is conditioned on the pre-defined metapath \mathcal{P} . These simulated “flows” are pseudo-sentences which are to generate skip-grams, inputs to the continuous Skip-gram model. The objective of *Metapath2Vec* (Dong et al., 2017) is

$$\mathcal{J} = \sum_{v \in \mathcal{V}} \sum_{u \in \mathcal{G}(v)} \log \frac{\exp(\mathbf{q}_u^T \mathbf{q}_v)}{\sum_{u' \in \mathcal{V}} \exp(\mathbf{q}_{u'}^T \mathbf{q}_v)}, \quad (2)$$

where $\mathbf{q}_u, \mathbf{q}_v$ are embeddings for node u, v respectively and $\mathcal{G}(v)$ is the *contexts* (a.k.a. skip-grams) of v in \mathcal{P} . To illustrate the idea in our setting, we assume a metapath $(F_1 P_1 F_2 J_1 P_2 \dots)$ and a context window size of 2. If we attend to node F_2 , the *context* of node F_2 is thus $\mathcal{G}(F_2) = \{F_1, P_1, J_1, P_2\}$. Note that the initial *Metapath2Vec* model proposed by (Dong et al., 2017) does not deal with the scenario involving multiple metapaths. In our work, the model is revised to leverage multiple metapaths which are randomly traversed in a uniform distribution. Providing that *Metapath2Vec* is based on random walks, its node embedding results may fluctuate with varying random parameter initializations.

3.2.3. Assessment of inter-firm competition Recall that our *meta-requirement* is to mine the inter-firm competitive relationship based on information about employees’ occupations and education. We therefore define *firms’ competition* as follows.

Definition 4 (Firms’ Competition) *Given a HON $\mathbf{G}_O = (\mathbf{V}_O, \mathbf{E}_O, \mathbf{A}_O, \mathbf{C}_O)$, a competitive relation is defined between two firms $F_i, F_j \in \mathbf{V}_O$ for any $1 \leq i, j \leq |\mathbf{V}_O|$ and $i \neq j$. We state that competition exists between firm i and j if the relation $F_i \xleftrightarrow{\text{compete}} F_j$ is observed; otherwise not.*

Note that we will perform inter-firm competition assessment in two ways: 1) compute the similarity of pairwise embeddings as the measure of competition; or 2) leverage supervised classifiers to distill more granular competitive dynamics between firms.

3.3. Testable hypotheses

Our inter-firm competition assessment problem can be viewed as a link prediction problem on a homogeneous network in which firms are nodes and pairwise competitive relationships are edges. Link predictions aim to model the network link formation process by predicting missed or future relationships based on currently observed connections (Al Hasan & Zaki, 2011). By contrast, network representation learning aims to learn low-dimensional latent features from the given network data which encode a variety of structural and semantic information. It has been proven useful in many tasks of data mining and machine learning such as link prediction, node classification, and network reconstruction (Li & Pi, 2020). Given the different methods for addressing our target problem, we propose our first testable hypothesis.

Hypothesis 1. *Firm latent representations learned from our homogeneous/heterogeneous networks have stronger capabilities in assessing inter-firm competition than traditional network link prediction methods.*

Network representation learning methods can be categorized into two groups according to the underlying network types: *homogeneous network embedding* and *heterogeneous network embedding*. A homogeneous network contains single type of node and edge. Heterogeneous networks overcome the limitation of single types of node and edge and thus encompass more diversified and richer information on different entities and relationships. Heterogeneous network embedding could map different heterogeneous objects into a unified latent space and thus grasp more network essentials (Li & Pi, 2020). Thus, we argue that firm representations learned from HON can capture more competitive dynamics, as opposed to the representations distilled from a homogeneous network, which brings our second hypothesis.

Hypothesis 2. *Firm latent representations learned from HON capture more competitive dynamics than those learned from a homogeneous network constructed using talent mobility flows.*

Human capital theory distinguishes industry-specific from firm-specific human capital (Becker, 1964). Industry-specific human capital is knowledge about complicated business routines that can be developed through professional education and industry experience and can be transferred across firms in the same industry. Firm-specific human capital is knowledge about unique routines and procedures that have limited value outside the firm. For valuable know-how and expertise that are transferable to other firms along with the migration of employees, educational attainment and industry experience act as key surrogate indicators (Pennings

Table 1: HON network statistics

Node	Count	Edge	Count
Firm (F)	520	$F \xleftrightarrow{\text{hire}} P$	85,658
Employee (P)	63,279	$F \xleftrightarrow{\text{offer}} J$	20,139
Job (J)	3,388	$P \xleftrightarrow{\text{work}} J$	81,105
School (S)	5,752	$P \xleftrightarrow{\text{study}} S$	75,805
Total Nodes	72,939	Total Edges	262,707

& Wezel, 2007). We posit that the incorporation of information about employees’ past occupation and education is essential to uncover competitive message between firms. We thus argue that the integration of employees’ occupation and education information can help improve inter-firm competitive assessment. Our third hypothesis is defined below.

Hypothesis 3. *The integration of employees’ occupation and education information can enhance the model’s efficacy of inter-firm competition assessment.*

3.4. Design instantiation

This section discusses the instantiation of our IT artifact in four phases. First, heterogeneous data is collected, extracted and parsed from two online data sources (*LinkedIn* and *Owler*). Second, we utilize employees’ occupation and education information to construct a *Heterogeneous Occupation Network (HON)*. Four HON variants are investigated, whose metapath sets are summarized here: *BASE*: {*FPPF*}; *BASE+JOB*: {*FPPF, FJPJF*}; *BASE+EDU*: {*FPPF, FPSPF*}; and *OVERALL*: {*FPPF, FJPJF, FPSPF*}. In the third stage, the *Metapath2Vec* algorithm is used to learn company latent representations on several structured networks. Lastly, inter-firm competition assessment is performed by 1) computing the similarity of pairwise firm embeddings and 2) leveraging a broad set of supervised classifiers for prediction.

4. Design Evaluation

4.1. Data Collection

LinkedIn data. A typical LinkedIn profile mainly contains: i) *current and past employment records* and ii) *educational attainments* (e.g., school name, academic degree, specialized subject, start and end months, etc.). (e.g., employer names, job titles, job functions, start and end months, etc.) Our study focuses on the IT industry. We compute and present the key statistics of our HON in Table 1. Most employment records are between 1990 and 2018 and the majority of education records are situated in the 1970-2018 timeframe.

Original *job titles* in our dataset are self-reported and relatively cluttered. We leverage the Industrial and Professional Dataset (IPOD) released by Liu et al.

(2019) to normalize job titles. IPOD is a comprehensive corpus which consists of over 190,000 job titles drawn from over 56,000 profiles. Three domain experts created a gazetteer with domain-specific job terms, which are classified into *Responsibility*, *Function* and *Location*. To retain cleaner job titles, we preserve only the terms labelled as *Responsibility* or *Function*, resulting in 3,388 distinct job titles. Meanwhile, the school profile links (rather than school names) are used as school identifiers to avoid ambiguity.

Owler data. To obtain credible competitor labels, we use another well-recognized business intelligence database, *Owler*. Subscribed by FACTSET², *Owler* is considered the most comprehensive, accurate, and up-to-date source of business information available. For each focal company, top 10 competitors are presented in the site’s *Competitive Intelligence* section. And no evidence has shown that the competitors are selected given any specific criteria. Therefore, we argue that it is rational to view them as the credible labels of *general competitors*.

4.2. Evaluation Setup

This section details our evaluation setup. The above two databases are linked using *entity matching* of firm names through measuring their *Levenshtein distances*. From 520 company nodes, we have extracted 675 pairs of competitors from the *Owler* dataset. As a common practice in model building, we split the firm pairs into training/validation/testing sets with a ratio of 7:2:1. Given any company from a competitor-pair in the training and validation sets, we randomly sample another non-competitor counterpart to construct a negative sample, which ends up with a balanced dataset of *competitors* and *non-competitors*. The test set is augmented with all possible firm pairs to preserve the true percentage of competitors, which aims to reflect the real scenario. We then evaluate and compare the models in four different categories.

HMO link prediction models. We downgrade our HON into a homogeneous network (**HMO**), e.g., a network with single-type of node and edge. We preserve solely the focal *Firm (F)* nodes and generate edges between any two firms if common employees are identified in our dataset. Then three models are included in this category. The first one is *Common Neighbors (CN)*, with $score(u, v) = |\Gamma(u) \cap \Gamma(v)|$, where $\Gamma(\cdot)$ denotes the set of neighbors. The second is *Jaccard’s Coefficient (JC)*, with $score(u, v) = |\Gamma(u) \cap \Gamma(v)| / |\Gamma(u) \cup \Gamma(v)|$. The third, *Preferential Attachment (PA)*, with a score function as $score(u, v) = |\Gamma(u)| \cdot$

²<https://www.factset.com>

Table 2: Comparison of the link prediction models and embeddings-based models

	CN	JC	PA	HMO-COS	HON-COS
Pr@10	0.100	0.111	0.100	0.120	0.224
Re@10	0.004	0.004	0.004	0.005	0.009
F1@10	0.008	0.008	0.008	0.009	0.017
AP	0.027	0.027	0.025	0.018	0.035
AUC	0.706	0.710	0.693	0.709	0.727

$|\Gamma(v)|$ (Kunegis et al., 2013).

HMO embedding-based proximity model. We also employ a HMO node embedding model, *Node2Vec* (Grover & Leskovec, 2016), as an opponent. The *Node2Vec* model is trained on this HMO network and export latent representations for each node (firm). To measure inter-firm competition, *cosine similarities* of pairwise firm embeddings are computed. This model is denoted as **HMO-COS**.

HON embedding-based proximity model. Using HON-based firm embeddings, we bring in a similar proximity model as another baseline. *Metapath2Vec* is trained on top of the HON network and cosine similarity is computed for each pair of firm embeddings. We denote this model **HON-COS**.

HON embedding-based supervised classifiers. Given the pre-computed HON-based firm embeddings, we use supervised classifiers to enhance the model’s capability of inter-firm competition assessment. Multiple popular machine learning classification models are taken into account (Tan et al., 2016), such as *Logistic Regression* (**HON-LG**), *Support Vector Machine* (**HON-SVM**), *Random Forest* (**HON-RF**), and *Neural Networks* (**HON-NN**). All modeling parameters are tuned to achieve their best performances.

Note that the inter-firm competition assessment is a *binary classification* problem and the model calculates a score between 0 and 1 given any firm pair. A higher score indicates stronger inter-firm competition. For evaluation, we employ some common performance metrics from the field of *information retrieval* (*IR*). We begin with *Precision@k* (**Pr@k**), *Recall@k* (**Re@k**) and *F1-score@k* (**F1@k**), where @*k* denotes that only the items with the top-*k* highest scores are considered. Here we take $k = 10$ as it simply aligns with our ground-truth of top 10 competitors for each firm. We further include *Average Precision* (**AP**) as $AP = (1/N) \sum_{k=1}^N Pr@k$ for N candidate companies and *Area Under ROC Curve* (**AUC**). These two metrics better reflect a model’s overall performance over the full range of trade-offs between sensitivity and specificity.

4.3. Experimental Results

We have performed two experimental studies. In our first experiment, we apply four models (**CN**, **JC**, **PA** and **HMO-COS**) on the homogeneous network **HMO** and apply the **HON-COS** model on the heterogeneous

Table 3: Comparison of HON embedding-based models

		HON-COS	HON-LG	HON-SVM	HON-NN	HON-RF
BASE	Pr@10	0.150	0.000	0.000	0.220	0.230
	Re@10	0.006	0.000	0.000	0.009	0.009
	F1@10	0.012	0.000	0.000	0.017	0.018
	AP	0.026	0.009	0.009	0.020	0.025
	AUC	0.711	0.709	0.726	0.715	0.724
BASE+JOB	Pr@10	0.252	0.000	0.040	0.100	0.363
	Re@10	0.010	0.000	0.002	0.004	0.015
	F1@10	0.020	0.000	0.003	0.008	0.028
	AP	0.035	0.009	0.016	0.018	0.030
	AUC	0.744	0.715	0.774	0.755	0.741
BASE+EDU	Pr@10	0.250	0.000	0.040	0.130	0.262
	Re@10	0.010	0.000	0.002	0.005	0.011
	F1@10	0.019	0.000	0.003	0.010	0.021
	AP	0.035	0.009	0.013	0.021	0.029
	AUC	0.731	0.712	0.761	0.745	0.737
OVERALL	Pr@10	0.224	0.000	0.050	0.100	0.307
	Re@10	0.009	0.000	0.002	0.004	0.012
	F1@10	0.017	0.000	0.004	0.008	0.024
	AP	0.035	0.009	0.016	0.019	0.030
	AUC	0.727	0.712	0.772	0.750	0.742

network **HON** with the **OVERALL** setting. To accommodate the randomness of the embeddings learning algorithms (**HMO-COS** and **HON-COS**), we first run the two models for 30 times with random initialization of parameters, respectively. We then pick the top-ten best performers on validation set and report their average performance scores on the test set. Table 2 showcases the scores of **CN**, **JC**, **PA**, **HMO-COS** and **HON-COS** on the evaluation metrics. We observe that **HON-COS** outperforms other models on all performance measures, as expected. Meanwhile, for the sake of statistical rigor, we perform the Wilcoxon signed-rank test (Demšar, 2006) on *AUC* scores. Specifically, we test *AUC* score differences of the three non-embedding-based methods with 1) **HMO-COS** and 2) **HON-COS** and list the test results in Table 4. Looking at the first three columns, we find that nearly all three non-embedding-based link prediction models perform worse than embedding-based models (except **JC** vs. **HMO-COS**), which supports *Hypothesis 1*. Meanwhile, **HON-COS** beats the homogeneous network-based **HMO-COS** model, which thus supports our *Hypothesis 2*.

Table 4: AUC differences between HMO-COS and other models

	CN	JC	PA	HMO-COS
w.r.t. HMO-COS	-0.0032**	0.0008	-0.0162**	
w.r.t. HON-COS	-0.0208**	-0.0168**	-0.0338**	-0.0177**

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In the second experiment, we turn our attention to the HON-based models. We build four supervised classifiers on top of the HON-based firm embeddings. We take firm pairs as instances and their concatenated embeddings as the input to train the classifiers. Since inter-firm competition is undirected in our study, the concatenated vector of the swapped embeddings for any firm pair is also valid for model training. The

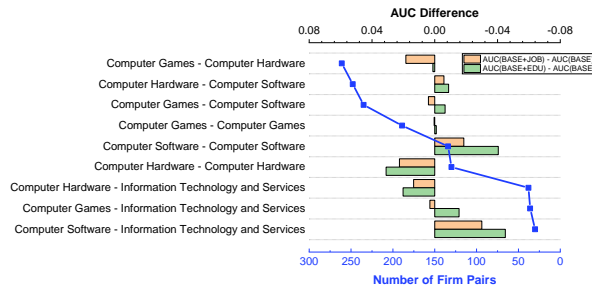


Figure 1: AUC differences by different industry pairs performances of each model on the test set are presented in Table 3. Most of the HON embedding-based supervised classifiers perform better than **HON-COS**. Among all, **HON-SVM** achieves the best AUC scores but performs poorly in terms of *precision*, *recall* and *F1-score*. By contrast, **HON-RF** demonstrates more balanced and stronger performance on almost all metrics. We therefore take **HON-RF** as our default model for the following analyses and discussions. On the other hand, we can observe from Table 3 that the models’ performances are generally better on HON networks with *JOB* and/or *EDU* information. The Wilcoxon test results in Table 5 further affirm this argument, which therefore supports *Hypothesis 3*.

Table 5: AUC differences of models built on BASE-based HON and other HONs

	BASE+JOB	BASE+EDU	OVERALL
w.r.t. BASE	0.0314***	0.0225***	0.0259***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Discussions and Implications

Added value of JOB and EDU. To better understand the added-value of occupation and education information, we group the firm pairs based on their industry sectors and examine the model’s performance separately. Given each industry pair, we calculate AUC scores of all firm pairs for model **HON-RF** built on the **BASE**, **BASE+JOB** and **BASE+EDU** networks. Then the AUC differences between them are presented in Figure 1. Note that we excluded the industry pairs with a small number (< 10) of firm pairs and 9 pairs remain. It is prominent that the efficacy of JOB and EDU information varies across different industry pairs.

To understand the possible causes to these disparity, we reviewed an emerging stream of work on general vs. specific human capital and transferability of skills (Gathmann & Schönberg, 2010; Shaw, 1987). It is broadly recognized that human capital is partially transferable across occupations (Gathmann & Schönberg, 2010) and that the probability of movement to a new occupation is positively related to the transferability of skills required in both positions (Shaw, 1987). One interesting pattern

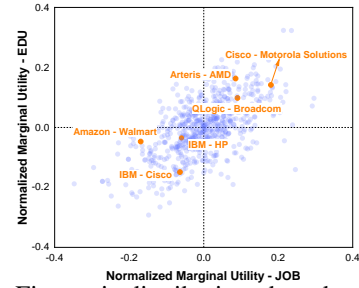


Figure 2: Firm-pair distributions based on normalized marginal utilities

in Figure 1 is that, for *Computer Software*-related pairs, the added JOB information leads to certain degradation of the model’s differentiability on inter-firm competition, while a prominent improvement is evident in *Computer Hardware*-related pairs. We posit that job title variability is less informative in the *Computer Software* or *Information Technology and Services* industries, as the required technical skills (e.g., computer programming) are more transferable across occupations. On the contrary, the required skills for jobs in the *Computer Hardware* industry (e.g., computer chips, circuit boards, or manufacturing and production) are more unique and function-specific.

Marginal utility of JOB and EDU. We attempt to understand the marginal utility of JOB and EDU to the problem. In specific, we compute the ranks of all firm pairs according to their competitive scores estimated by **HON-RF** built on the aforementioned three networks, i.e., **BASE**, **BASE+JOB** and **BASE+EDU**. To quantify how the added information can help reveal more subtleties in inter-firm competition, we develop the notion of *Marginal Utility (MU)* for a given competitive firm pair (i, j) with added JOB and EDU information, respectively:

$$MU^{JOB}(i, j) = Rank^{BASE+JOB}(i, j) - Rank^{BASE}(i, j),$$

$$MU^{EDU}(i, j) = Rank^{BASE+EDU}(i, j) - Rank^{BASE}(i, j). \quad (3)$$

where $Rank^X(i, j)$ denotes the rank order of competitive scores for the firm pair (i, j) computed by the model trained on the basis of the heterogeneous network X . MU is the additional utility gained from the model by adding a specific piece of information (JOB or EDU) into the network. To align the scales, we further propose *Normalized Marginal Utility (NormMU)* as:

$$NormMU^{JOB}(i, j) = (MarginalUtility^{JOB}(i, j))/N,$$

$$NormMU^{EDU}(i, j) = (MarginalUtility^{EDU}(i, j))/N, \quad (4)$$

where N is the total number of firm pairs. We then scatter all firm pairs given their *NormMU* in Figure 2. The more the points are scattered towards the upper-right corner (the first quadrant), the higher the marginal utility (JOB and EDU) would be. The

shape of the distribution reveals moderate correlation between the two variables, which also signifies that occupation and education information tend to offer similar efficacy in assessing inter-firm competition. We highlight some representative firm pairs in the plot and find that *Computer Hardware* firms are more likely to reside in the first quadrant while *Computer Software* firms tend to be in the third quadrant. This pattern aligns with our earlier observations.

6. Conclusion

Summary. In this paper, we attempt to address inter-firm competition assessment problem and adopted the *Information Systems Design Theory* (ISDT) framework (Walls et al., 1992) to guide the design of our IT artifact. We start by constructing a *Heterogeneous Occupation Network* (HON) among firms and employees using employees' occupation and education information, and then leverage a *Metapath2Vec*-based heterogeneous network embedding algorithm to learn firms' latent representations (embeddings). Then, the inter-firm competition is assessed in two ways: 1) by computing the similarities of firm embeddings and 2) by building a broad set of supervised classifiers on the concatenated firm embeddings. Following the principle of *design as a search process* (Hevner et al., 2004), we performed extensive model evaluations and comparisons with a focus on *IT industry* using the data sourced from *LinkedIn* and *Owler*.

Limitations and future work. We further discuss some limitations of our work, which aim to foster more future research. First, our study focuses solely on IT industry. It is worth investigating other industries to see whether similar conclusions can be drawn. Second, our main objective is to develop an IT artifact using a heterogeneous network embedding model by leveraging its capability of integrating multiple types of information. One follow-up work is to investigate and renovate other heterogeneous network embedding models for better model performance. Third, we can systematically fuse the two components: firm embeddings and competition assessment, to yield an end-to-end solution. Currently, putting aside this idea is owing to one main objective: to examine the potential of firm embeddings that are learned exclusively from employees' occupation and education information in an unsupervised manner. Meanwhile, we aim to further increase the interpretability of our enhanced model.

References

Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In

- Social network data analytics* (pp. 243–275). Springer.
- Bao, S., Li, R., Yu, Y., & Cao, Y. (2008). Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(10), 1297–1310.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management*, 17(1), 99–120.
- Barney, J. B. (2014). Gaining and sustaining competitive advantage. Harlow: Pearson.
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. Columbia University Press, New York.
- Bergen, M., & Peteraf, M. A. (2002). Competitor identification and competitor analysis: A broad-based managerial approach. *Managerial and decision economics*, 23(4-5), 157–169.
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Info. Proc. Systems*, 26.
- Chen, M.-J. (1996). Competitor analysis and interfirm rivalry: Toward a theoretical integration. *Academy of management review*, 21(1), 100–134.
- Clark, B. H., & Montgomery, D. B. (1999). Managerial identification of competitors. *Journal of Marketing*, 63(3), 67–83.
- Cohen, M. D., & Bacdayan, P. (1994). Organizational routines are stored as procedural memory: Evidence from a laboratory study. *Organization science*, 5(4), 554–568.
- Conner, K. R., & Prahalad, C. K. (1996). A resource-based theory of the firm: Knowledge versus opportunism. *Organization science*, 7(5), 477–501.
- Daft, R. L. (1983). *Organization theory and design*, west pub. St. Paul, Minnesota.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, 1–30.
- Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2d knowledge graph embeddings, In *Thirty-second aaai conference on artificial intelligence*.
- Dong, Y., Chawla, N. V., & Swami, A. (2017). Metapath2vec: Scalable representation learning for heterogeneous networks, In *Proceedings of acm sigkdd international conference on knowledge discovery and data mining*.

- Gathmann, C., & Schönberg, U. (2010). How general is human capital? a task-based approach. *Journal of Labor Economics*, 28(1), 1–49.
- Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks, In *Proceedings of acm sigkdd int'l conf. on knowledge discovery & data mining*.
- Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28, 75–105.
- Kunegis, J., Blattner, M., & Moser, C. (2013). Preferential attachment in online networks: Measurement and explanations, In *Acm web science conference*.
- Lado, A. A., & Wilson, M. C. (1994). Human resource systems and sustained competitive advantage: A competency-based perspective. *Academy of management review*, 19(4), 699–727.
- Li, B., & Pi, D. (2020). Network representation learning: A systematic literature review. *Neural Computing and Applications*, 1–33.
- Li, N. (2017). Labor Market Peer Firms.
- Li, R., Bao, S., Wang, J., Yu, Y., & Cao, Y. (2006). CoMiner: An Effective Algorithm for mining competitors from the web, In *Proceedings of ieee international conference on data mining*, IEEE.
- Liu, J., Guo, C., Ng, Y. C., Wood, K. L., & Lim, K. H. (2019). Ipod: Corpus of 190,000 industrial occupations. *arXiv preprint arXiv:1910.10495*.
- Liu, Y., Pant, G., & Sheng, O. R. (2020). Predicting labor market competition: Leveraging interfirm network and employee skills. *Information Systems Research*, 31(4), 1443–1466.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space, In *Proceedings of int'l conf. on learning representations workshop*.
- Pant, G., & Sheng, O. R. (2009). Avoiding the blind spots: Competitor identification using web text and linkage structure, In *Proceedings of the international conference on information systems*.
- Pant, G., & Sheng, O. R. (2015). Web footprints of firms: Using online isomorphism for competitor identification. *Information Systems Research*, 26(1), 188–209.
- Pennings, J. M., & Wezel, F. C. (2007). *Human capital, inter-firm mobility and organizational evolution*. Edward Elgar Publishing.
- Shaw, K. L. (1987). Occupational change, employer change, and the transferability of skills. *Southern Economic Journal*, 702–719.
- Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2016). A survey of heterogeneous information network analysis. *IEEE Trans. on Knowledge and Data Engineering*, 29(1), 17–37.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Intro. to data mining*. Pearson Education India.
- Tang, J., Qu, M., & Mei, Q. (2015). Pte: Predictive text embedding through large-scale heterogeneous text networks, In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*.
- Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant eis. *Information Systems Research*, 3(1), 36–59.
- Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network, In *The world wide web conference*.
- Wright, P. M., McMahan, G. C., & McWilliams, A. (1994). Human resources and sustained competitive advantage: A resource-based perspective. *International journal of human resource management*, 5(2), 301–326.
- Yang, C., Xiao, Y., Zhang, Y., Sun, Y., & Han, J. (2020). Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.
- Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network, In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*.
- Zhang, L., Xu, T., Zhu, H., Qin, C., Meng, Q., Xiong, H., & Chen, E. (2020). Large-scale talent flow embedding for company competitive analysis, In *Proceedings of the web conference*.
- Zhang, W., Cai, Y., Lau, R. Y., Liao, S. S., & Kwok, R. C.-W. (2012a). Semi-supervised text mining for dynamic business network discovery., In *Pacis*.
- Zhang, W., Lau, R. Y., Liao, S. S., & Kwok, R. C.-W. (2012b). A probabilistic generative model for latent business networks mining, In *Proceedings of the international conference on information systems*.
- Zhang, Z., Guo, C., & Goes, P. (2013). Product comparison networks for Competitive Analysis of Online Word-of-Mouth. *ACM Transactions on Management Information Systems*, 3(4), 1–22.