Dec 12th, 12:00 AM

Firm Profiling and Competition Assessment via Heterogeneous Occupation Network

Hao Zhong

Chuanren Liu

Firm Profiling and Competition Assessment via Heterogeneous Occupation Network

Completed Research Paper

Hao Zhong
Information & Operations Management
ESCP Business School
Paris, France
hzhong@escp.eu

Chuanren Liu
Business Analytics and Statistics
The University of Tennessee, Knoxville
Knoxville, TN, United States
cliu89@utk.edu

Abstract

Extensive efforts have been made by both academics and practitioners to understand inter-firm competitive relationship. However, it has never been an easy task to fully characterize firms and assess their competitive relationship owing 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 firm embeddings as input, we train multiple classifiers to assess the competitive relationship among the firms. 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, competition assessment, heterogeneous occupation network.

Introduction

Competitor analysis, as an essential component of corporate strategy, has long been a critical yet challenging task to business. Competitor identification (CI) is a necessary precursor to the task of competitor analysis and serves as the starting point of forming competitive strategies (Bergen and Peteraf 2002). Timely and precise identification of potential competitors has considerable impacts on several key tactical and strategic business goals, e.g., company benchmarking (Knuf 2000; Leem et al. 2008), marketing strategy planning (Zinkhan and Pereira 1994), talent acquisition and retention (Lewis and Heckman 2006).

From managerial perspectives, two approaches have been proposed for the identification of competitors, i.e., *supply-based* and *demand-based* (Clark and 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; Zheng et al. 2012), the *supply-based* approach is favored for the identification of competing firms in the literature (Bao et al. 2008; Li et al. 2006; Pant and Sheng 2009; Pant and Sheng 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) articulated 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 and Wilson 1994). Evidently, 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 and 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; Peffers et al. 2007; 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 and 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*[1] and *Owler*[2], and have the following findings. First, our proposed HON-based system is proven effective in assessing *general* inter-firm competitions. Second, some unique information (e.g., employees' occupational details and education attainments) pertains essential substances in revealing competitive relationships between firms. Third, our cross-industry studies reveal some managerial insights on varying effects of employees' occupation information on inter-firm competition assessment.

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 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*.

The rest of our paper is organized as follows. Section *Related Work* reviews the relevant literature and Section *Kernel Theory-based Design* elaborates the details of our kernel theory-based IT artifact design. Section *Design Evaluation* discusses data collection, experimental setup, and evaluation results. Section *Discussions and Implications* shows empirical findings and discusses major implications. And Section *Conclusion* concludes our paper in the end.

Related Work

This section reviews three streams of literature. We start with the managerial studies on inter-firm competition and human resources, which closely relates to our kernel theory. Then we review multiple data-driven inter-firm competition analyses as they are the closest to this study. At last, we summarize the state-of-the-art studies on heterogeneous network embedding models.

[1]https://www.linkedin.com/
[2]https://corp.owler.com/

Inter-firm competition and human resources

Understanding competitors and developing an advantage against them is critical to business, which draws considerable interests of management, economics, and marketing literature (Clark and Montgomery 1999). As an earlier work in managerial identification of competitors, Clark and Montgomery (1999) reviewed some prior work and summarized two approaches in identifying competitors: *supply-based* and *demand-based*. The *supply-based* approach identifies competitors based on the similarity of the competing firms' attributes (e.g., business strategy, technology, and products offered) while the *demand-based* approach relies on the attributes of customers (e.g., customer attitudes and behaviors). Clark and Montgomery (1999)'s *supply-based* approach is conceptually linked to one of the two views in understanding inter-firm rivalry in (Chen 1996), i.e., through *resource similarity*. Their common proposition is to study inter-firm competitions by regarding firms as bundles of resources and capabilities.

There has been a large body of work 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 and Prahalad 1996). The concept of *firm resources* is adopted from Daft (2006): 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) discussed why human resource systems can have a profound impact on firm's sustained competitive advantage. They argue 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, a great deal of work on human resource/capital and inter-firm competition emerged in the literature. 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 *Kernel Theory-based Design*. 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*. In her paper, she 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. In a nutshell, it is of sufficient evidence that human resource/capital can serve as a credible proxy to understanding inter-firm competitions.

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. Unlike *CoMiner*, they refined the labels using competitors' data from a more credible source, Hoover's API[3]. Besides online web metrics, they incorporated offline metrics (e.g., SIC codes and market values of firms) to enhance the model's predictive capability (Pant and Sheng 2015). 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

[3]http://developer.hoovers.com/page/overview

their relationships (Zhang et al. 2012a).

Most recently, studies on the use of human resource information for firm competitor identification have emerged. As an example, 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. Despite that our study shares some common theoretical foundations with theirs (e.g., the notion of human resource bundles from RBV), there are still several notable distinctions. One major difference lies on 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 (given that our source of competitor labels are indistinctive and general). 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, thanks to 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 *Design Evaluation*.

As another relevant work, Zhang et al. (2020) also utilized human capital flow (*talent flow* in their paper) to assess competitions between companies. Despite that they developed a Talent Flow Embedding (TFE) model to measure pairwise competitive relationship of firms, the underlying network still relies on the aggregated flows, similar to 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 cannot reflect the *genuine* inter-firm competitive relationship. Instead, our model's capability is examined using real and credible competitor labels, thanks to *Owler*'s data.

Heterogeneous network embedding models

Graph embedding models have enjoyed rising popularity in various scholarly and practical domains in recent years. Depending on the types of nodes and/or edges, the models are categorized into *homogeneous network-based* (only a single type of node and edge involved) and *heterogeneous network-based* (with multiple types of nodes and/or edges). 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. They can be further divided into two groups: random walk approaches (e.g., *Metapath2Vec* (Dong et al. 2017) and *HIN2Vec* (Fu et al. 2017)) and first/second-order proximity-based approaches (e.g., *PTE* (Tang et al. 2015) and *HEER* (Shi et al. 2018)). In contrast, message-passing methods aim to learn node embeddings by aggregating the information from neighbors, e.g, *HAN* (Wang et al. 2019), *HetGNN* (Zhang et al. 2019) and *HGT* (Hu et al. 2020). 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), DistMult (Yang et al. 2015) 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.

Kernel Theory-based Design

The design of our system is essentially guided by *Information System Design Theory (ISDT)*, which was originally developed by Walls et al. (1992). It consists of four components: *kernel theories*, *meta-requirement*, *meta-design*, and *testable hypotheses*.

Kernel theory: inter-firm personnel mobility as a form of transformation of firms' competitive surroundings

The *kernel theory* for our IT artifact design originates from (Pennings and 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 and 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 and Wezel 2007). According to the Resource-based View (RBV), human capital is a special type of resource, which combines skills and knowledge gained from prior work experiences and education. Firms competing for talents with comparable work experience or education background tend to be more homogeneous and therefore have a higher chance to compete. Thus, we aim to investigate the extent to which employees' occupation and education details can distill inter-firm competitions.

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 the environment of business competition.

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 and 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 design details in the following.

The design of Heterogeneous occupation network (HON)

We first need to construct a heterogeneous information network to incorporate various information (e.g., occupational records and educational attainments) in a flexible manner. 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 (HON)) *A Heterogeneous Occupation Network (HON) is defined as* $\boldsymbol{G}_O = (\boldsymbol{V}_O, \boldsymbol{E}_O, \boldsymbol{A}_O, \boldsymbol{C}_O)$. $\boldsymbol{A}_O = \{F, P, J, S\}$ *where each denotes, respectively, Firm (F), Employee (P), Job (J) and School (S) and* $\boldsymbol{C}_O = \{offer, hire, work, study\}$ *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 Figure 1, we present a schematic illustration of our HON network with some typical scenarios. For

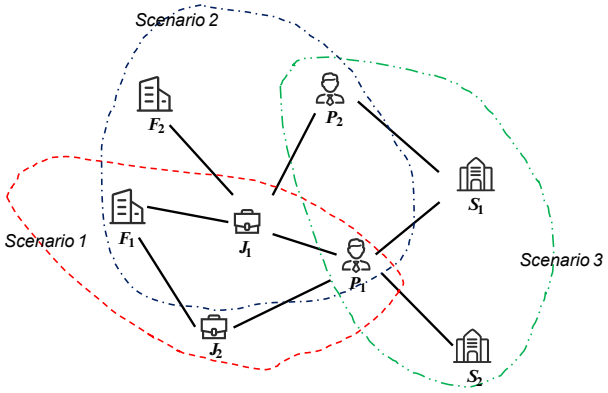


Figure 1. A conceptual presentation of our HON network

example, *Scenario 1* shows that employee P_1 has worked at firm F_1 with two different jobs J_1 and J_2. In *Scenario 2*, we can observe an interesting case where firm F_1 and F_2 offer the same job J_1 which is nonetheless taken by two different employees P_1 and P_2. *Scenario 3* presents a possible case where one individual P_1 has studied at two different schools S_1 and S_2 while s/he and P_2 are both alumni of school S_1. 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 $a_1 \overset{c_1}{\longleftrightarrow} a_2 \overset{c_2}{\longleftrightarrow} \cdots \overset{c_L}{\longleftrightarrow} a_{L+1}$, where $a_l \in \boldsymbol{A}$ are node types and $c_l \in \boldsymbol{C}$ are link types, respectively, and $1 \leq l \leq L$.*

For simplicity, we can also use a sequence of node types to denote the meta path if there is only one relation type between the same pair of link types, e.g., $\mathcal{P} = (a_1 a_2 \cdots a_{L+1})$. As such, in our proposed HON network, one typical example of *metapath* is $F \overset{offer}{\longleftrightarrow} J \overset{worked\_by}{\longleftrightarrow} P \overset{study}{\longleftrightarrow} S$, which can be denoted $(FJPS)$. Note that the links are *undirected* in our HON network.

Firm representation learning

Definition 3 (Heterogeneous Network Embedding) *For a given heterogeneous network $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{A}, \boldsymbol{C})$, heterogeneous network embedding is a mapping function $\mathcal{F}: \boldsymbol{V} \mapsto \mathbb{R}^{|V| \times d} (d \ll |\boldsymbol{V}|)$, which defines the latent representation of each node $v \in \boldsymbol{V}$ and captures network topological information in $\boldsymbol{E}$.*

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 and 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 $\mathcal{P}: a_1 \overset{c_1}{\longleftrightarrow} a_2 \overset{c_2}{\longleftrightarrow} \cdots \overset{c_L}{\longleftrightarrow} a_{L+1}$, the transition probability at step l is defined as:

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

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

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

where $\boldsymbol{q}_u, \boldsymbol{q}_v$ 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 $(F_1P_1F_2J_1P_2...)$ 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. But later in Section *Design Evaluation*, we will observe that the results are statistically stable.

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

Note that we will perform inter-firm competition assessment in two ways. First, we can compute the similarity of pairwise firm embeddings as the measure of competition. Or we can leverage supervised classifiers to further distill more granular competitive dynamics between firms.

Testable hypotheses

As for the *testable hypotheses*, we mainly focus on the *feasibility* and the *effectiveness* of the IT artifact (Walls et al. 1992). The *feasibility* is demonstrated by the final instantiation of our developed system, which will be detailed in Section *Design instantiation*. The *effectiveness* is evaluated through extensive experimental studies and tested on our proposed hypotheses. Prior to that, we first articulate our 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 and Zaki 2011). There have been numerous studies of network link prediction methods, mostly in the field of social network analysis (Al Hasan and Zaki 2011; Liben-Nowell and Kleinberg 2007). These traditional methods of extracting structural information from network data usually depend on topological statistical information, aggregation coefficients or the limitations of handcrafted features (Li and Pi 2020). 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. Network representation learning has proven useful in many tasks of data mining and machine learning such as link prediction, node classification, and network reconstruction (Li and 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 and 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 and 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.*

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)*. To examine the effects of heterogeneous data, we construct a number of network variants using varying types of nodes and edges. Four HON variants are investigated, whose metapath sets are summarized here: *BASE*: $\{FPF\}$; *BASE+JOB*: $\{FPF, FJPJF\}$; *BASE+EDU*: $\{FPF, FPSPF\}$; and *OVERALL*: $\{FPF, FJPJF, FPSPF\}$. As an example, $\{FPF, FJPJF\}$ means the network is constructued using two types of metapaths: 1) $F \xleftrightarrow{hire} P \xleftrightarrow{hired\_by} F$ and 2) $F \xleftrightarrow{offer} J \xleftrightarrow{worked\_by} P \xleftrightarrow{work} J \xleftrightarrow{offered\_by} F$. 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.

Design Evaluation

Data Collection

To our knowledge, there is no single publicly-accessible database storing all the required information for our study (e.g., employees' occupation, educational background and inter-firm competition). We therefore retrieved and processed the original data from two data sources and integrated them into one unique dataset for our analysis.

LinkedIn data. Launched in 2003, *LinkedIn* is a professional networking platform which allows members (both employees and employers) to develop profiles and build "connections".[4] It reaches over 700 million registered members from more than 200 countries and territories worldwide as of November 2020.[5] LinkedIn users generally have commitments to assure the credibility of their self-reported personal 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. Our data was collected by crawling individual's profile, following a well-selected list of IT-related industry sectors. Specifically, we searched individual profiles using the industry as the query keyword and tried to extract all matched profiles for each industry. Figure 2 showcases the distribution of employee profiles across all industry categories in our data, in which *Information Technology and Services* and *Computer Software* are the most dominant groups. Recall that our HON is constructed with four types of node (*Firm, Employee, Job*, and *School*) and multiple metapaths. We compute and present the key statistics of our HON as well as its node-degree distribution in Figure 3. Meanwhile, Figure 4a presents the longitudinal distribution of the employment and education records. We can see most employment records are between 1990 and 2018 and the majority of education records are situated in the

[4]https://en.wikipedia.org/wiki/LinkedIn
[5]https://about.linkedin.com/

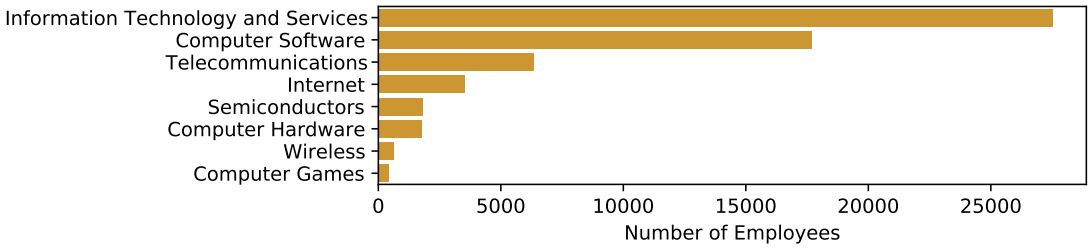


Figure 2. Distribution of employee profiles across all industry sectors

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

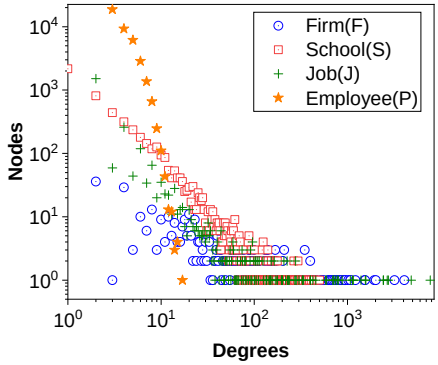


Figure 3. HON network statistics and its node-degree distribution

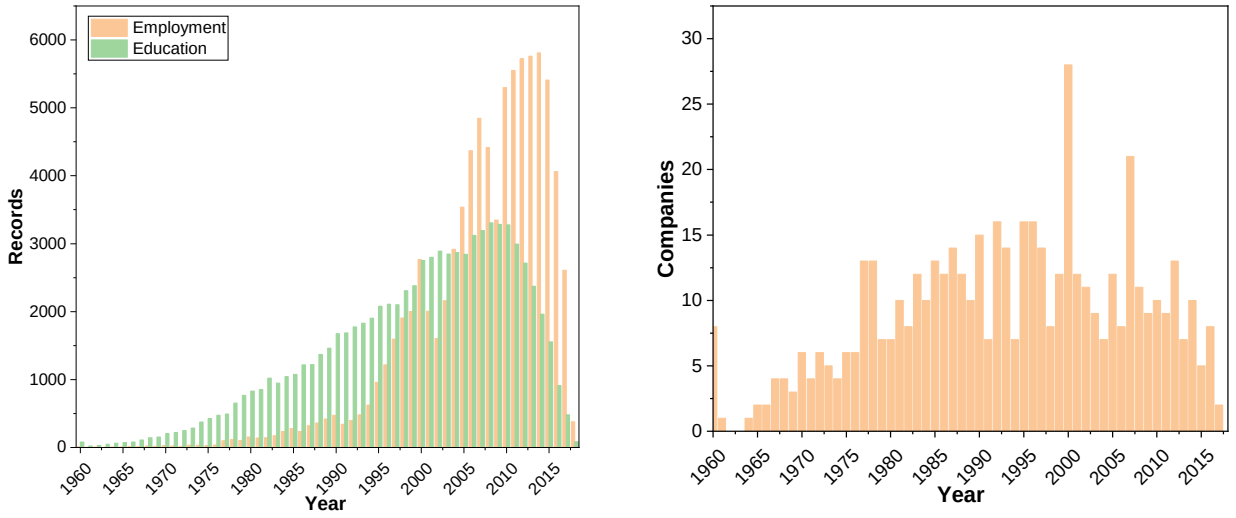
1970-2018 timeframe. On the other hand, Figure 4b displays the longitudinal distribution of firms by year of their first recruitment record, ranging from 1960 to 2018 in our dataset. In fact, our database contains a broad set of companies, from *old giants* (IBM, Apple and Microsoft), *new giants* (Google, Amazon and LinkedIn) to *new upstarts* (Zynga, Pinterest and Instagram).

Original *job titles* in our dataset are self-reported and relatively cluttered. It is crucial to normalize job titles and retain commonly-shared terms pertaining to job responsibilities and functions. In this regard, we leverage the Industrial and Professional Dataset (IPOD) released by Liu et al. (2019). 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*. Owler is specialized in business intelligence with its data mined and used by a community of over 5M reliable and knowledgeable business professionals[6]. Meanwhile, it is subscribed by FACTSET[7], a globally-recognized financial data and service provider who works with over 7000 financial service firms worldwide. In Owler, for each focal company, top 10 competitors are presented in its 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*. Table 1 lists a curated set of firms and their top three competitors retrieved from *Owler*.

These two databases are linked using a *fuzzy entity matching* of firm names through measuring their *Levenshtein distances*. We retain only the firm pairs with similarity scores of 95% or higher as candidates and then confirm the 520 matched companies via manual verification. More statistics are shown in Figure 3.

[6]https://corp.owler.com/data-licensing
[7]https://www.factset.com/marketplace/catalog/product/owler-private-company-data



(a) Employment and education records by year **(b) Firms by year of the first recruitment record**

Figure 4. Data distribution in chronological order

	Dell	Microsoft	Amazon	LinkedIn	Salesforce	Best Buy	Uber
#1	HP	Apple	Walmart	New Work	Oracle	Walmart	Lyft
#2	Lenovo	Cisco	ebay	indeed	SAP	Target	Ola
#3	IBM	Google	Alibaba	ZipRecruiter	Adobe	Costco	Didi

Table 1. Examples of firms and their top-3 competitors (from Owler)

Evaluation Setup

This section details our evaluation setup. Note that 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 start by downgrading 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**), which simply counts the number of neighbors that two nodes have in common: $score(u, v) = |\Gamma(u) \cap \Gamma(v)|$, where $\Gamma(\cdot)$ denotes the set of neighbors. The second is *Jaccard's Coefficient* (**JC**), measuring the probability that both u and v have common features, where "features" are node neighbors: $score(u, v) = |\Gamma(u) \cap \Gamma(v)|/|\Gamma(u) \cup \Gamma(v)|$. The third, *Preferential Attachment* (**PA**), refers to the observation that in networks that grow over time, the probability that an edge is "attached" to a node with d neighbors is *proportional* to d (Kunegis et al. 2013), with a score function is as $score(u, v) = |\Gamma(u)| \cdot |\Gamma(v)|$ (Liben-Nowell and Kleinberg 2007).

HMO embedding-based proximity model. We also employ a HMO node embedding model, *Node2Vec* (Grover and 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

	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

Table 2. Comparison of the link prediction models and embeddings-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

Table 3. Comparison of the HON embedding-based models

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. Note that the node embeddings are randomly initialized with the dimensionality of 64. We denote this model as **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.

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 heteroge-

neous 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*.

	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****

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

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

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 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 signed-rank test results in Table 5 further affirms this argument, which therefore supports *Hypothesis 3*.

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

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

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

Discussions and Implications

Lift measure. We start by examining the effectiveness of firm latent representations on assessing inter-firm competition. In specific, we use the *lift* measure, which is interpreted as the extent to which it "pushes up" the positive instances over the negative instances in an ordered list and thus quantifies the comparative advantages over random guessing (Provost and Fawcett 2013). The calculation process is as follows. Given any firm i in our data, we first compute the *cosine similarity* between its embedding and that of all other firms learned by our proposed HON-based embedding models. Then with any given percentage p, we choose the top p% of firms based on their calculated similarity scores w.r.t. firm i. We label $I^X(p,i) = 1$ if *any* selected firm pair has true competitive relationships, verified using our ground-truth labels from *Owler*; otherwise 0. Note that X denotes any of these four variations of HON, i.e., *BASE*, *BASE+JOB*, *BASE+EDU*, or *OVERALL*. As a baseline, we repeat such calculation by randomly selecting p% of firms and calculating $I^R(p,i)$. Our *lift* measure ends up as

$$lift^X(p) = \frac{\sum_{i=1}^{N} I^X(p,i)}{\sum_{i=1}^{N} I^R(p,i)}. \tag{3}$$

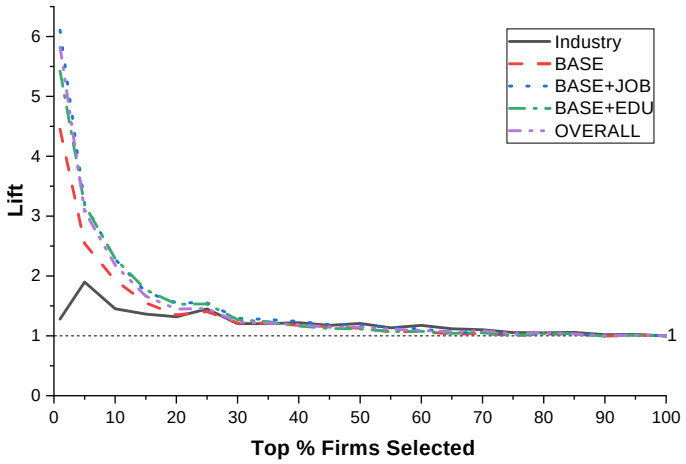


Figure 5. Lift curves of the industry-based model and HON-based models

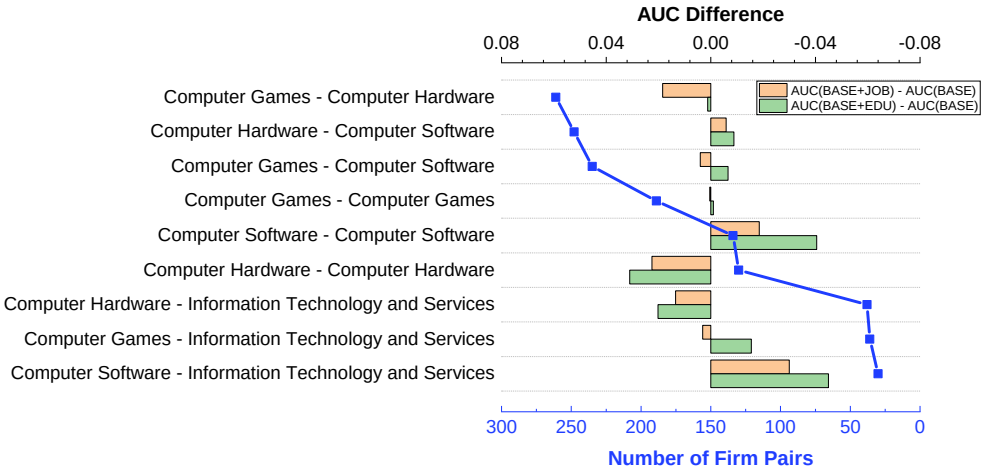


Figure 6. AUC differences by different industry pairs

By altering p, we can generate the *lift* curves given different variation of HON, as shown in Figure 5. As a benchmark, we also include the *lift* curve calculated using an industry-based approach: viewing firms within the same industry as competitors. The plots clearly show that our HON-based models have significant advantages over the industry-based baseline. Note that our best model has over three times of the chance, as opposed to random guessing, in correctly predicting firm i's competitors if taking top 5% other firms as competitor candidates.

Added value of JOB and EDU. To better understand the added-value of occupation and education information in inter-firm competition assessment, we group the firm pairs based on their industry sectors and examine the model's performance separately. Given each industry pair, we calculate the 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 6. 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 and Schönberg 2010; Kambourov and Manovskii 2009; Parent 2000; Shaw 1987). It is broadly recognized that human capital is partially transferable across occupations (Gathmann and 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

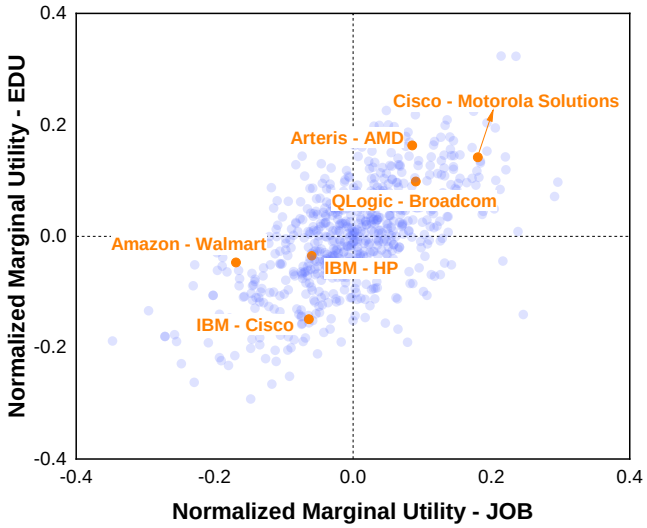


Figure 7. Firm-pair distributions based on normalized marginal utilities

interesting pattern in Figure 6 is that, for *Computer Software*-related pairs, the added JOB information does not enhance the model's differentiability in terms of inter-firm competition, while the improvement is more prominent 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 and software development) 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. Furthermore, we attempt to understand the marginal utility of JOB and EDU information to inter-firm competition assessment. 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* for a given competitive firm pair (i, j) with added JOB and EDU information, respectively:

$$\begin{aligned} Marginal\_Utility^{JOB}(i,j) &= Rank^{BASE+JOB}(i,j) - Rank^{BASE}(i,j), \\ Marginal\_Utility^{EDU}(i,j) &= Rank^{BASE+EDU}(i,j) - Rank^{BASE}(i,j). \end{aligned} \tag{4}$$

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. *Marginal Utility* 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* as:

$$\begin{aligned} Normalized\_Marginal\_Utility^{JOB}(i,j) &= \frac{Marginal\_Utility^{JOB}(i,j)}{N}, \\ Normalized\_Marginal\_Utility^{EDU}(i,j) &= \frac{Marginal\_Utility^{EDU}(i,j)}{N}, \end{aligned} \tag{5}$$

where N is the total number of firm pairs. We then scatter all firm pairs given their *Normalized Marginal Utilities* in Figure 7. 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.

Conclusion

Summary. In this paper, we have attempted to address the important yet challenging inter-firm competition assessment problem. We adopted the *Information Systems Design Theory* (ISDT) framework (Walls et al. 1992) to guide the design of our IT artifact. Our proposed system is built on the kernel theory 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. In specific, 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, including 1) assessing the effectiveness of our model on four different HON variants and 2) comparing our model with other selected baseline models. Our analysis focuses on *IT industry* using the data sourced from *LinkedIn* and *Owler*. To summarize, our study has the following findings. The design process and experimental analysis demonstrate the effectiveness of our IT artifact in assessing general inter-firm competitions. Meanwhile, the rigorous statistical testings reveal that employees' occupation and education details can further improve the model's capability for our task.

Limitations and future work. We further discuss some limitations of our work, which aims to foster more future research. First, our study focuses solely on IT industry. It would be worth investigating other industry sectors in the future 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. We have not exerted ourselves to completely re-build the *Metapath2Vec*-based model or examine other network embedding models. One follow-up work is to investigate and renovate other heterogeneous network embedding models for better firm representations learning. Third, individual employment data are dynamic and inter-firm competitive relationships also change over time. It is worth studying the variation of inter-firm competition over time using dynamic heterogeneous occupation networks. Lastly, we can later develop an end-to-end solution by systematically fusing the two components: firm embeddings learning and inter-firm competition assessment. The reason to put aside this idea is owing to one main objective in our study: to examine the potential of firm embeddings that are exclusively learned from employees' occupation and education information in an unsupervised manner. But an integral end-to-end system is surely worth more in-depth research.

References

Al Hasan, M. and Zaki, M. J. 2011. "A survey of link prediction in social networks". In: *Social network data analytics*. Springer, pp. 243–275.

Bao, S., Li, R., Yu, Y., and Cao, Y. 2008. "Competitor mining with the web". *IEEE Transactions on Knowledge and Data Engineering* (20:10), pp. 1297–1310.

Barney, J. 1991. "Firm resources and sustained competitive advantage". *Journal of Management* (17:1), pp. 99–120.

Barney, J. B. 2014. *Gaining and sustaining competitive advantage. Harlow*.

Becker, G. S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Columbia University Press, New York.

Bergen, M. and Peteraf, M. A. 2002. "Competitor identification and competitor analysis: a broad-based managerial approach". *Managerial and Decision Economics* (23:4-5), pp. 157–169.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. 2013. "Translating embeddings for modeling multi-relational data". *Advances in Neural Information Processing Systems* (26).

Chen, M.-J. 1996. "Competitor analysis and interfirm rivalry: Toward a theoretical integration". *Academy of Management Review* (21:1), pp. 100–134.

Clark, B. H. and Montgomery, D. B. 1999. "Managerial identification of competitors". *Journal of Marketing* (63:3), pp. 67–83.

Cohen, M. D. and Bacdayan, P. 1994. "Organizational routines are stored as procedural memory: Evidence from a laboratory study". *Organization Science* (5:4), pp. 554–568.
Conner, K. R. and Prahalad, C. K. 1996. "A resource-based theory of the firm: Knowledge versus opportunism". *Organization Science* (7:5), pp. 477–501.
Daft, R. 2006. *Organization Theory and Design*. Available Titles CengageNOW Series. Cengage Learning.
Demšar, J. 2006. "Statistical comparisons of classifiers over multiple data sets". *The Journal of Machine Learning Research* (7), pp. 1–30.
Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. 2018. "Convolutional 2d knowledge graph embeddings". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.
Dong, Y., Chawla, N. V., and Swami, A. 2017. "metapath2vec: Scalable representation learning for heterogeneous networks". In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144.
Fu, T.-y., Lee, W.-C., and Lei, Z. 2017. "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning". In: *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 1797–1806.
Gathmann, C. and Schönberg, U. 2010. "How general is human capital? A task-based approach". *Journal of Labor Economics* (28:1), pp. 1–49.
Grover, A. and Leskovec, J. 2016. "node2vec: Scalable feature learning for networks". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864.
Hevner, A., March, S., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research". *MIS Quarterly* (28), pp. 75–105.
Hu, Z., Dong, Y., Wang, K., and Sun, Y. 2020. "Heterogeneous graph transformer". In: *Proceedings of The Web Conference 2020*, pp. 2704–2710.
Kambourov, G. and Manovskii, I. 2009. "Occupational specificity of human capital". *International Economic Review* (50:1), pp. 63–115.
Knuf, J. 2000. "Benchmarking the lean enterprise: organizational learning at work". *Journal of Management in Engineering* (16:4), pp. 58–71.
Kunegis, J., Blattner, M., and Moser, C. 2013. "Preferential attachment in online networks: Measurement and explanations". In: *Proceedings of the ACM Web Science Conference*, pp. 205–214.
Lado, A. A. and Wilson, M. C. 1994. "Human resource systems and sustained competitive advantage: A competency-based perspective". *Academy of Management Review* (19:4), pp. 699–727.
Leem, C. S., Kim, B. W., Yu, E. J., and Paek, M. H. 2008. "Information technology maturity stages and enterprise benchmarking: an empirical study". *Industrial Management & Data Systems* (108:9), pp. 1200–1218.
Lewis, R. E. and Heckman, R. J. 2006. "Talent management: A critical review". *Human Resource Management Review* (16:2), pp. 139–154.
Li, B. and Pi, D. 2020. "Network representation learning: a systematic literature review". *Neural Computing and Application* (32:21), pp. 16647–16679.
Li, N. 2017. "Labor Market Peer Firms". Available at SSRN: https://ssrn.com/abstract=2558271.
Li, R., Bao, S., Wang, J., Yu, Y., and Cao, Y. 2006. "CoMiner: An Effective Algorithm for Mining Competitors from the Web". In: *Proceedings of IEEE International Conference on Data Mining*. IEEE, pp. 948–952.
Liben-Nowell, D. and Kleinberg, J. 2007. "The link prediction problem for social networks". *Journal of the American Society for Information Science and Technology* (58:7), pp. 1019–1031.
Liu, J., Guo, C., Ng, Y. C., Wood, K. L., and Lim, K. H. 2019. "IPOD: Corpus of 190, 000 Industrial Occupations". *CoRR* (abs/1910.10495). arXiv: 1910.10495.
Liu, Y., Pant, G., and Sheng, O. R. 2020. "Predicting Labor Market Competition: Leveraging Interfirm Network and Employee Skills". *Information Systems Research* (31:4), pp. 1443–1466.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. "Efficient estimation of word representations in vector space". In: *Proceedings of International Conference on Learning Representations Workshop*.
Pant, G. and Sheng, O. R. 2009. "Avoiding the blind spots: Competitor identification using web text and linkage structure". In: *Proceedings of the Int'l Conf. on Information Systems*, p. 57.
Pant, G. and Sheng, O. R. 2015. "Web footprints of firms: Using online isomorphism for competitor identification". *Information Systems Research* (26:1), pp. 188–209.

Parent, D. 2000. "Industry-specific capital and the wage profile: Evidence from the national longitudinal survey of youth and the panel study of income dynamics". *Journal of Labor Economics* (18:2), pp. 306–323.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research". *Journal of Management Information Systems* (24:3), pp. 45–77.

Pennings, J. M. and Wezel, F. C. 2007. *Human capital, inter-firm mobility and organizational evolution*. Edward Elgar Publishing.

Provost, F. and Fawcett, T. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Shaw, K. L. 1987. "Occupational change, employer change, and the transferability of skills". *Southern Economic Journal* (), pp. 702–719.

Shi, C., Li, Y., Zhang, J., Sun, Y., and Philip, S. Y. 2016. "A survey of heterogeneous information network analysis". *IEEE Transaction on Knowledge and Data Engineering* (29:1), pp. 17–37.

Shi, Y., Zhu, Q., Guo, F., Zhang, C., and Han, J. 2018. "Easing embedding learning by comprehensive transcription of heterogeneous information networks". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2190–2199.

Tan, P.-N., Steinbach, M., and Kumar, V. 2016. *Introduction to data mining*. Pearson Education India.

Tang, J., Qu, M., and Mei, Q. 2015. "Pte: Predictive text embedding through large-scale heterogeneous text networks". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an information system design theory for vigilant EIS". *Information Systems Research* (3:1), pp. 36–59.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., and Yu, P. S. 2019. "Heterogeneous graph attention network". In: *The World Wide Web Conference*, pp. 2022–2032.

Wright, P. M., McMahan, G. C., and McWilliams, A. 1994. "Human resources and sustained competitive advantage: a resource-based perspective". *International Journal of Human Resource Management* (5:2), pp. 301–326.

Yang, B., Yih, W., He, X., Gao, J., and Deng, L. 2015. "Embedding Entities and Relations for Learning and Inference in Knowledge Bases". In: *Proceedings of International Conference on Learning Representations*.

Yang, C., Xiao, Y., Zhang, Y., Sun, Y., and Han, J. 2020. "Heterogeneous Network Representation Learning: Survey, Benchmark, Evaluation, and Beyond". *CoRR* (abs/2004.00216). arXiv: 2004.00216.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. 2019. "Heterogeneous graph neural network". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 793–803.

Zhang, L., Xu, T., Zhu, H., Qin, C., Meng, Q., Xiong, H., and Chen, E. 2020. "Large-Scale Talent Flow Embedding for Company Competitive Analysis". In: *Proceedings of The Web Conference*, pp. 2354–2364.

Zhang, W., Cai, Y., Lau, R. Y., Liao, S. S., and Kwok, R. C.-W. 2012a. "Semi-Supervised Text Mining For Dynamic Business Network Discovery." In: *PACIS*, p. 138.

Zhang, W., Lau, R. Y., Liao, S. S., and Kwok, R. C.-W. 2012b. "A probabilistic generative model for latent business networks mining". In: *ICIS*. Citeseer.

Zhang, Z., Guo, C., and Goes, P. 2013. "Product Comparison Networks for Competitive Analysis of Online Word-of-Mouth". *ACM Transactions on Management Information Systems* (3:4), pp. 1–22.

Zheng, Z. E., Fader, P., and Padmanabhan, B. 2012. "From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data". *Information Systems Research* (23:3-part-1), pp. 698–720.

Zinkhan, G. M. and Pereira, A. 1994. "An overview of marketing strategy and planning". *International Journal of Research in Marketing* (11:3), pp. 185–218.