Learning from Citation: Interpretable Patent Valuation via Conditional Variational Autoencoder

Chuan-Wei Kuo[1*], Bo-Yu Chen[1†], Chih-Chieh Hung[2†], Wen-Chih Peng[1†] and Hsin-Ning Su[3†]

[1*]Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University.
[2]Department of Management Information Systems, National Chung Hsing University.
[3]Institute of Management of Technology, National Yang Ming Chiao Tung University.

*Corresponding author(s). E-mail(s):
auberonkuo.cs07@nycu.edu.tw;
Contributing authors: mowd.cs07@nycu.edu.tw;
smalloshin@nchu.edu.tw; wcpeng@g2.nctu.edu.tw;
ningsu@nycu.edu.tw;
[†]These authors contributed equally to this work.

Abstract

Patent value prediction is one of the most fundamental components in many applications for knowledge value retrieval. It is important in medical and high-tech fields and crucial in commercial legal proceedings. Since the patent citation relationship has been proven to be one of the most critical factors affecting the value of a patent. Hence, extracting such a citation relationship becomes an essential process. However, extracting the citation relationship between patent documents is still challenging because patent relationship updates frequently. This paper embeds the citation relationship from the patent's Cross Relationship Network (CRN) via the Conditional Variational Autoencoder (CVAE) into prediction tasks. This is the first attempt to leverage the deep generative model to extract a patent's citation relationship into a patent valuation prediction task. Moreover, we also

design an interpretable mechanism to judge the goodness. Our approach shows a significant positive effect on refining the prediction of patent value and can enormously assist the Intellectual Property (IP) valuation. We also demonstrate the importance of utilizing patent CRN for extracting the citation relationship. After verifying real-world data, our proposed method shows outstanding results in valuing the newly granted patents by prediction models. In addition, the code and dataset employed in this research will be available to the research community.

1 Introduction

Patents have been used to instantiate knowledge to protect and commercialize the technology. In addition, researchers have continued to study the questions involving patent valuation for decades because of the real-world crucial commercialization issues of knowledge. Especially in the patent, they were licensing stage and when facing an infringement lawsuit process. The demand for Intellectual Property (IP) valuation has been increasing for many years, and IP occupies a significant position in financial investments and loan decisions. Many companies have received thousands of granted patents, especially in high-tech fields. Assessing the value of IP also becomes a critical issue [1–3] in building a fair and reliable investing environment.

Therefore, generating a principled approach to estimate the patent value is a central issue for the technology market, IP system, and other stakeholders. In general, we can measure the patent value from two aspects: 1) Domain experts can provide a professional review for patents by investigating the depth of technology. 2) Utilizing machine learning that conducts predictive results for patents by employing its characteristics (i.e., bibliographic features and other extra information) as inputs. This paper focuses on leveraging machine learning methods because they can quickly and efficiently analyze patent value by various learning methods. In addition, citation received counts have been seen as critical for determining a patent's value [4]. Formally to the definition, the citations that a patent receives from subsequent patents are called forward citations. Moreover, the citations that a patent citing existing patents are called backward citations. Specifically, forward citation analysis has become progressively dominant for evaluating patents value in Mergers and Acquisitions, infringement disputes, and other purposes. In contrast, backward citation analysis is a basis for responding to technological completeness. In current practices, the standard operating process compares those valued patents or portfolios (i.e., licensing fee or market price) to the patent wait to valuation. However, it only relies on the total received citations to do patent valuation has fundamental limitations.
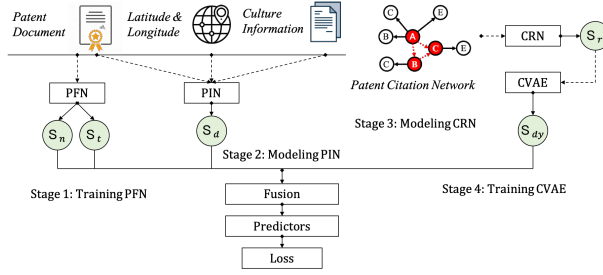
Fig. 1 Overview of the framework of our proposed approach, from the feature process and leverage CVAE to build dynamic interaction intensity.

A severe problem arises when the total number of citations is used as a prediction label for the learning models. The earlier the patent is issued, the higher the number of citations, unfair to the newly issued patents. To obtain a fairer and more useful predictive label, limiting the time window of received citations (i.e., calculating the number of citations received after N years from the issuance) is required to fairly compare different patents, shown significantly in several empirical studies.

Another essential issue in predicting patent value by machine learning is the quality of input features, and high-quality features are more likely to represent the patent sufficiently. Moreover, five vital bibliographic features have been highly relevant to patent value: 1) the number of foreign patent citations. 2) the number of forward and backward citations. 3) the number of non-patent references. 4) the number of claims. 5) the number of Cooperative Patent Classification (CPC). Differ from the bibliographic features of the patent itself. Many previous studies show that external information is also essential for evaluating the patent value. They extracted the external information as new patent characteristics through their defined formulas and proved their usefulness for patent value analysis. For instance, [5] designed an algorithm to calculate the patent's geographical intensity by the latitude and longitude the inventors live as a new feature. [6] has proposed the concept of cultural distance to combine the external scores into a new patent feature.

However, utilizing this external information only focuses on the patent's characteristics individually without taking the cross-relationship between the patents in the citation network. Extracting such cross-relationships from citation networks is still challenging. In recent practices, generative model Generative Adversarial Networks (GANs) and CVAE have achieved outperforms in many tasks. Such as image generation, speech generation, and data generation. They learn the structure of inputs very well through the different kinds of the f-divergency loss functions. We believe that the citation relationship between patents can also be seen as a particular structure intercepted by some restrictions. Hence, we use the generative model to extract cross-relationship between patents through its capability of learning data structure in this paper. In addition, we also propose a mechanism to interpret how this method can help patent valuation.

We propose a Cross Relationship Network (CRN) to simulate the dynamic interaction between patents. Besides, to handle that high-dynamic transition change between patents, we utilized the CVAE model to compress and extract the dynamic interaction between patents. In addition, Patent Feature Network (PFN) provides the patent's bibliographic information and textual information from the patent documents. The Patent Index Network (PIN) calculates the external patent information to model the patent value's geographical and cultural effect. The ability of generated features from CVAE will measure in four cases, which observe the predictive performance. In summary, our contributions can be summarized as follows:

- We design a mechanism to interpret the importance of the cross-relationship for patent valuation. By giving the complete process of extracting citation relationship, and improving the performance of prediction models.
- For cross-relationship extracting, we design a method to utilize the generative model to capture the dynamic interaction between patents.
- We conduct extensive experiments on real-world datasets to demonstrate the effectiveness of our proposed method on the patent value prediction task. In addition, our approach significant enhances the performance on our baseline models.

2 Related Work

2.1 Citation for patent valuation

The value of technology can be determined and estimated by machine learning methods, which have been explored in many papers. Hence, an accepted consensus is that the patent citation counts can represent the patent value, significantly related to patent value. In addition, two crucial patent metrics related to citation are widely used in patent value prediction tasks: [7] define the "Generality Index" that describes the variety of fields of a patent's forward citations, which is used by [8] for the examination of patent pools. They also define the "Originality Index" for a patent's backward citations, which is used to study the creation of startups.

2.2 Bibliographic factors for patent valuation

Apart from the citation, patent bibliographic factors have been used in many previous empirical studies, for instance, the reference, non-patent reference, claim, cooperative patent classification, patent family size, litigation status, etc. In current standard practices, they directly use these instead of comprehensively considering the other factors because they believe the bibliographic features can represent the patent well. For instance, "claim" in the patent document is used to mainly protect and vehemently declare the right of patent, which significantly reflects the patent value. Litigation status can also indirectly indicate the value of a patent, and generally, more valuable patents are

more likely to be prosecuted. The former can be analyzed using text mining methods. In addition, similarity among the claims of relevant patents will also affect the risk of patent infringement. However, only considering the bibliographic factors will extensively erase the explanation power.

2.3 External information for patent valuation

Questions involving how external information affect the valuation of the patent have attracted researchers in recent years [5]. Differ from bibliographic factors representing the patent information directly, external data is used to activate the outside influence of patents. Recent empirical studies show that geography and cultural factors have strongly influenced patent quality. The geographic factor is calculated by the living city's longitude and latitude of the relevant inventors, which indicates that the combination of researchers from different regions affects the performance of the patent. In addition, the cultural factor will be scored for different cultural degrees such as social norms, ethnicities, and beliefs, and each patent will have an overall score.

2.4 Prediction Citations

Prediction citations have drawn the attention of researchers, especially in paper value prediction tasks. However, even though the relevant research involves various analytical methods (i.e., from traditional statistical analysis to deep learning methods). Research has only focused on leveraging the bibliographic factors and few external information for predicting the patent value: [9] focus on the problem of predicting the citation between a pair of papers, and [10, 11] extract author-wise attributes, paper-specific, and venue-centric features to generate a regression model. The supervision restriction in a predefined time window is based on the citation counts. Most of the task of paper citation prediction used advanced point process models [12–15]. [16] applied the standard Poisson process model, and [12] predicting the citation count at an arbitrary time point through the EM algorithm. The difference technical among them is how prior is used [13], and whether the learning problem is solved in closed-form by differential equation iteratively by gradient descent [14], or maximal likelihood estimate [12].

2.5 Network Embedding

Network embedding is designed to learn the low-dimensional potential representation of nodes in the network and has been used for many empirical kinds of research [17–23]. The learned feature representations can be used as features of various tasks based on graphs. DeepWalk [24] bridges the gap between network embedding and word embedding by treating nodes as words and generating short random walks as sentences. A neural language model such as skip-gram [25] can then be applied to these random walks to obtain network embedding. LINE [26] uses a breadth-first search strategy to generate context nodes: only nodes that are up to two hops from a given node are considered to

be neighbors. In addition, it uses negative sampling to optimize the skip-gram model compared to the layered softmax used in DeepWalk.

Recent years have witnessed an increasing interest in deep generative models that generate observable data based on hidden parameters [27–30]. Unlike Generative Adversarial Networks (GANs)[28] which generate data based on arbitrary noises, the Conditional Variational Auto-encoder (CVAE) [31] setting we adopted is more expressive since it tries to model the underlying probability distribution of the data by latent variables so that we can sample from that distribution to generate new data accordingly. An increasing number of models and applications are proposed which consider data in different modalities, such as generating images or natural language. Works on generative relation discovery with a probabilistic graphic model that requires handcrafted relation-level features. As far as we know, the knowledge value discovers through embedding the knowledge relational intensity into the patent entity, which is suitable for deep generative modeling, has not been studied in a generative perspective with restricted data requirement.

2.6 Overview of our Approach

Fig. 1. presents the framework of our proposed approach. It consists of four components: 1) Patent Feature Network (PFN) for word proximity and bibliographic feature extraction of patent documents; 2) Patent Index Network (PIN) for external information calculating; 3) Cross Relationship Network (CRN) for generating dynamic patent correlation knowledge flow; 4) CVAE model for generating the dynamic interaction intensity between patents from CRN. For more specific, Fig. 2, 5, 6 shows the detailed introduction of our proposed method, and the notations used in this paper are summarized in Table 1. Section 3 and 4 will illustrate the mathematical formula inference and how the algorithm works in our model.

3 Bibliographic and External Information

Patent Feature Network: Under our proposed framework, PFN (see Fig. 2) will obtain the patent's bibliographic factors in two ways: 1) The normalization method will process numerical features. 2) the textual data (e.g., title abstracts and claims) will be embedded using the BERT model, which proved to be one of the most effective methods for embedding textual data in several real-world studies.

Patent Index Network: PIN integrates external data and calculates it through its algorithm to gather external data representation. In addition, considering that the use of external data requires the advice of domain experts, we also took the advice from domain experts. This task used three distance embedding methodologies as patent's external information. We considered the three state-of-the-art ideas referring to the technology management field: 1) Researching for geography indicators affecting the patent value is currently an import issue [5]. 2) The originality and generality intensity of a patent can
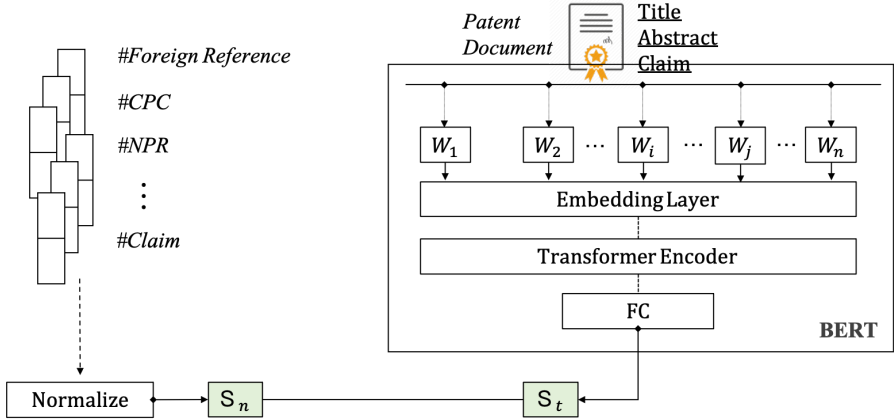
Fig. 2 Patent Feature Network (PFN). Pre-processes bibliographic and text features. S_n and S_t represent the output sequence after the two processes.

strongly reflect the value of the specific patent entity. 3) The cultural factor is the crucial implicit factor for evaluating the value of the patent.

3.1 Collaboration Distance

This patent indicator uses the natural log of the average geographic distance (in kilometers) between all inventors in a given firm and their partners. Partnerships are determined using patent data (i.e., the inventor and assignee of one specific patent entity). The geographic distance between a pair of inventors involved in a patented invention is defined as the spatial distance between their cities, which is based on the latitudes and longitudes of the city. The definition of collaboration distance [5] as below:

$$Collaboration_{ij} = r\{\arccos[\sin(lat_i)\sin(lat_j) \\ +\cos(lat_i)\cos(lat_j)\cos(|long_i - long_j|)]\} \tag{1}$$

3.2 Generality and Originality Distance

The originality of a patent indicates the diversity of cited patents and is based on the distribution (by ratio) of cited patents over specialized classes. The originality index can also be applied at the country, industry, and assignee levels for illustrating the distribution (by ratio) of cited patents over different assignees, industries, and governments. The generality index is used to measure the externality of knowledge spillovers. The generality index indicates the diversity of citing patents and is based on the distribution (by ratio) of citing patents over specialized classes. The generality index can also be applied at the assignee, industry, and country levels for illustrating the distribution
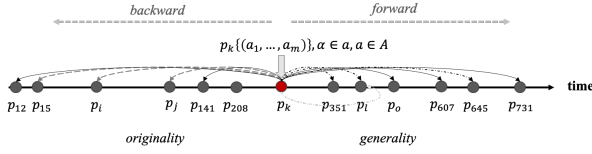
Fig. 3 Generality and originality distance. p_k is the anchor node, and the "forward" means received citation called generality, and the "backward" means citing another patent called originality.

(by ratio) of citing patents over different organizations, industries, and countries. Fig. 3. shows the structure of generality and originality distance. The definition of originality and generality distance [32] as below:

$$\forall R(p_k, p), \exists \alpha, \{p_k(\alpha_\theta) = p(\alpha_\theta) | \theta = 1, 2, ..., \gamma\} \tag{2}$$

$$Originality = 1 - \sum_{\theta}^{n_k} s_{k\theta}^2 \tag{3}$$

where $s_{k\theta}$ denotes the percentage of citations made by p_k that belong to patent class θ out of n_k patent classes.

$$Generality = 1 - \sum_{\theta}^{n_k} s_{k\theta}^2 \tag{4}$$

where $s_{k\theta}$ denotes the percentage of citations received by p_k that belong to patent class θ out of n_k patent classes.

3.3 Cultural Distance

Cultural distance [6] indicates differences between national cultures, such as social norms, ethnicities, and beliefs, and is usually used in international collaboration and business research. It computes six national cultural dimensions using country scores from the Hofstede Center survey: (1) Power Distance, (2) Uncertainty Avoidance, (3) Individualism, (4) Masculinity, (5) Long term orientation (6) Indulgence. Updated national scores on these dimensions have been relatively stable over time, which are often used in previous research. Each pair of countries constructs a measure of cultural distance as the Euclidean space distance between two countries' scores on six dimensions. The definition of cultural distance is as below:

$$Cultural_{ij} = \sqrt{\sum_{6}^{k=1} \frac{(I_{ik} - I_{jk})^2}{V_k}} \tag{5}$$

where i and j represent country i and j; k denotes the cultural dimension $k(k = 1, \ldots, 6)$; I_{ik} and I_{jk} are cultural scores on the dimension k; and V_k shows variance of all cultural scores on the dimension k. We embedding all of these three important patent distances into our dataset to enhance the analysis ability and comparing to the original one.

Table 1 Notations in this paper.

Symbols	Description
p	The granted patent p from the U.S. patent office
a	Attribute of the p
$lat_a, long_a$	Latitude and longitude information of the attribute a
d	Collaboration distance of patent p
r^*	The radius of the earth, set to $3,963$ miles
θ	Patent class, for instance, chemical and electronic
S_θ	The percentage of citations received by p
I_k	Culture scores of dimension $k, k = 1, \ldots, 6$
V_k	Variance of all culture scores
P_{citing}, P_{cited}	Citing and received citation of p
$Q_\phi(\cdot)$	Simple distribution (i.e., a gaussian distribution)
$P_\theta(\cdot)$	The generated distribution from our model
V_k	Variance of all culture scores
R	Citation relationship between two patent
V	A set of n granted p
w_{ij}	Denote the adjacency matrix $(4,681 * 4,681)$

4 Generative Embedding for CRN

To obtain the strength of dynamic interaction between patents, we define a firm-citation-based patent Cross Relationship Network (CRN), which consists of those firms that issued more than 100 patents in 2013 (a total have 4,681 firms, see Fig. 4 and Table 2). Specifically, this paper aims to find the implied association strength of each firm, which can be regarded as a representation of the firm's strength. In addition, we use the CVAE model to learn the characteristics of data distribution by KL-divergence under the restriction of specific conditions and use the "cited company" itself as the restriction to learn the distribution of the "cited company" cited by other firms. Moreover, the CVAE model can generate the cross-citation strength of the "cited company" through the generator. Finally, each company can have its dynamic interaction intensity, which means the generator will generate 4,681 dynamic interaction intensity.

A three-phase module was designed in our CVAE model: encoder, decoder, and generator. Encoder module takes the citation relation patent pair entity (p_{citing}, p_{cited}) and a conditional label c (i.e., "cited company") as the inputs,
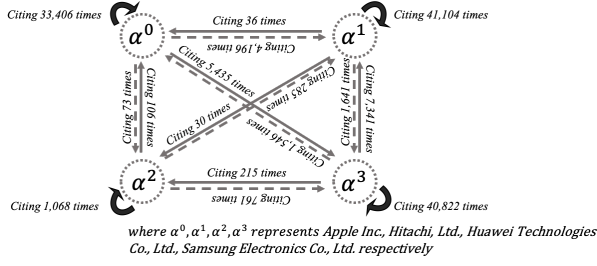
Fig. 4 There are three types of firm citation relationships: self-citation, citing others, and citing by others. Four nodes will generate up to 16 results; inference to N nodes will maximum have N^2 results.

Table 2 The representation of CRN. We choose the companies that have granted more than 100 patents in 2013 from the USPTO. 4,681 companies have been selected.

Company	$Apple_{cited}$	$Google_{cited}$	$Huawei_{cited}$	$Sony_{cited}$	$Samsung_{cited}$
$Apple_{citing}$	33,406	2,268	106	472	1,546
$Google_{citing}$	3,444	18,274	75	298	952
$Huawei_{citing}$	73	51	1,068	56	215
$Sony_{citing}$	10,687	4,369	273	40,740	10,630
$Samsung_{citing}$	5,435	3,933	761	3,887	40,822

then trained it to enhance the patent pair entity representations. Thus, encode the diversely expressed patent entity pair for each patent related to a latent space with simple distribution Q_ϕ. The decoder is jointly trained to reconstruct the patent relational entity as P_θ. The generator module shares the same learned structure as the decoder module.
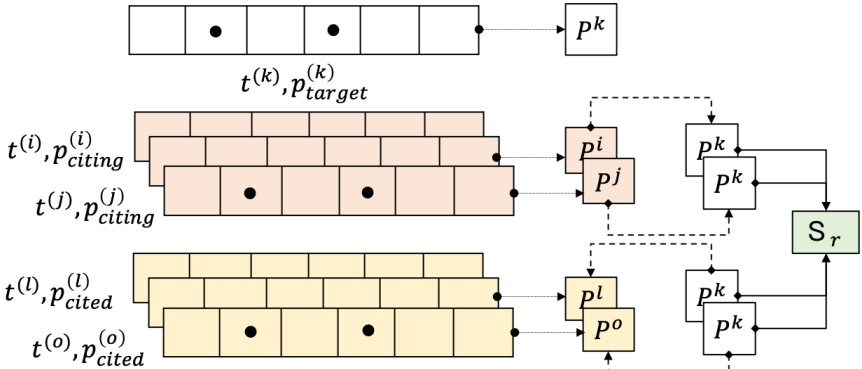
Fig. 5 Detail of Cross Relationship Network (CRN). Three types of citation relationships will be generated in this module as an adjacency matrix. p^k represents the target node, and S_r is the output sequence.
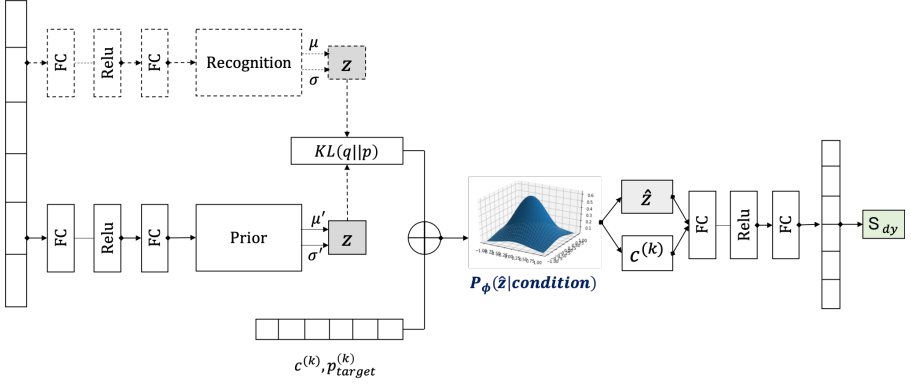
Fig. 6 Detail of CVAE. The inputs from CRN and the condition label are the "cited company," and the output sequence is S_{dy}. Random variable z from Gaussian distribution $N(0,1)$. KL-divergence uses to estimate the similarity of distribution. The activation function uses ReLu.

4.1 Cross Relationship Network

Cross relationship networks (CRN) as an essential first step. This network interpreted the company citation process. Two formal definitions of a CRN as below:
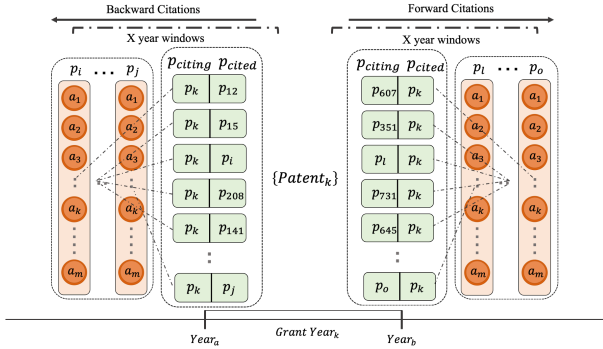
Definition 1: A patent citation network (see Fig. 7) consists a network $G = (P, A)$, the patent P and attributes $A = \{a_1, a_k, \ldots, a_m\}$ where m is the maximum number of attributes of patent p, where $p \in P$. Also include the citation records set $R = \{p_{citing}, p_{cited}\}$, where p_{citing} and p_{cited} represents a citation pair r as citation relation between two patents, where $r \in R$, which means a patent p_i citing another patent p_j, where $p_i \neq p_j$.

Definition 2: Given a patent citation network. The objective function is $\delta(\cdot)$ and have a restriction setting ξ. The CRN is to select a set of target patents P, which has same restriction ξ under $\delta(\cdot)$ for each patent p, where $p \in P$. The restriction ξ in this paper is for those companies that granted more than 100 patents in 2013.

Fig. 8. shows that the structure of patent citation (i.e., if p_k role the p_{citing} and p_i role the p_{cited}, means $Patent_k$ citing the $Patent_i$, each pair of block represent one citation relationship). The $GrantYear_k$ represents that the patent was given legal rights from patent offices between $Year_a$ and $Year_b$. Generally, the patent owner has the patent right of about 20 years. Forward citations show all the citations from other patents, and backward citations show all the citations it cites. X year time windows are the citation time windows, and we can limit the received citation counts year within X years. In addition, the 5-year citation windows are used in this paper.
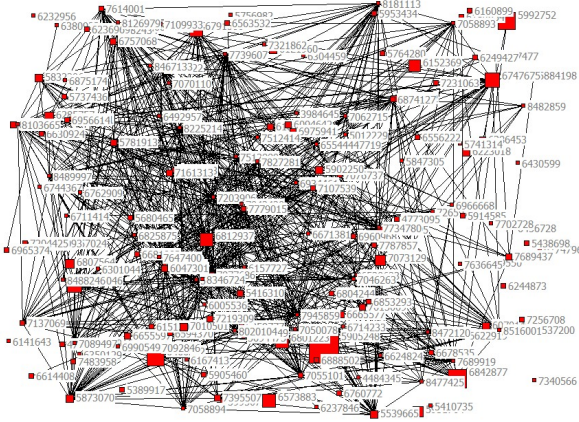
Fig. 7 Real-world patent citation network. The red node represents patents, and the edge represents a relationship (one-to-many). Patents used are from the U.S. Patent Office (USPTO).

Fig. 8 Detail of the patent citation network. The patent is chosen under the same grant year and limited to an X-year time window of citations.

4.2 Problem Definition

Let $G = \{V, A, W\}$ denote as an input knowledge graph. $V = \{v_i\}_{i=1}^n$ is a set of n patent nodes (i.e., n granted patents), $W = \{w_{ij}\} \in \mathbb{R}^{n \times n}$ denotes the adjacency matrix (CRN). $w_{ij} = t$ denotes there exists t times citations between v_i and v_j. $A = \{a_{ij}\} \in \mathbb{R}^{n \times d}$ represents the attribute matrix. The i-th row $a_i \in \mathbb{R}^d$ denotes the attributes of node v_i. V and A used to build the CRN.

We focus on embedding the dynamic interaction intensity between patents in this paper and how learning its representation from CRN is a critical contribution to our approach. To achieve this, we must use a network embedding method that is adequate for our needs. Besides, a suitable network embedding method should preserve the proximity between nodes to get the low-dimensional representation. In other words, an appropriate embedding

method should put the similar nodes together and push the different nodes away in the vector space.

For instance, the proximity between different nodes can represent as relational linkage block $\{v_i|v_{pos}, v_{neg}\}$ where v_{pos} denote the positive nodes which is similar with v_i, and v_{neg} denote the negative nodes which is dissimilar with v_i. In our task, CRN's node similarity is to learn a map $\phi : \{W\} \to E$, where $E \in \mathbb{R}$ denotes the low-dimensional representation, and the low-dimensional representation should satisfy:

$$sim(E_{i\cdot}, E_{pos\cdot}) > sim(E_{i\cdot}, E_{neg\cdot}) \tag{6}$$

where $sim(\cdot)$ denotes the proximity between two patent nodes, $E_{i\cdot}$ represents the low-dimensional embedding of i-th node v_i.

A suitable network embedding method can deconstruct the relationship between targets, which means discovering the proximity of each node from the original variable space. Moreover, it will preserve the relationship in the low-dimensional space. In other words, only the well-discovered proximity can be stored in the low-dimensional space as new representations. Learning a good node representation is the goal of different embedding methodologies. However, the edges in the network are usually very sparse and not good enough to discover the proximity between other nodes.

Furthermore, we need an embedding method that can learn the "dynamic interaction," not only the need for the similarity between nodes alone but also the need to ensure the randomness. Current network embedding proposed various methods for finding underlying proximities, for instance, first-order proximity (i.e., the probability of co-occurrence of two nodes), second-order proximity (i.e., the conditional probability of another node appearing when given one node). However, most of them lack to consider the randomness of generation. Based on the above. This paper uses the CVAE generative model as the network embedding method. Besides learning the similarity between nodes, the generator allows specific conditions and uses random Gaussian sampling to ensure the randomness of the result.

Definition of CVAE: The most architecture of CVAE model inherits from variational autoencoder. The roots in Bayesian inference, and assumes the input data x can be encoded into a set of latent variables z with certain distributions, for instance, Gaussian distributions. The latent variables z are generated by the generative distribution $P_\theta(z), x^{'}$ generated with a Bayesian model by a conditional distribution on z (i.e., $P_\theta(x^{'}|z)$. The latent distribution $P(z)$ infer from $P_\theta(z|x)$, which can consider as some mapping from x to z. In addition, $P_\theta(z|x)$ is usually inferred using a simple distribution $Q_\phi(z|x)$, for instance, Gaussian distribution. Moreover, its extension with a particular label c, both x and latent variables z are conditioned on the label c. The objective of the model is to optimize it's variational lower bound with condition c as below:

$$\mathcal{L}_{CVAE}(x, c;\ \theta, \phi) = -KL[Q_\phi(z|x,c)||P_\theta(z|x)]$$
$$+log\left(P_\theta(x|c)\right) \tag{7}$$

where the first term uses the *KL-divergence* to minimize the difference between simple distribution $Q_\phi(z|x,c)$ and the true distribution $P_\theta(z|x)$, maximizes the $log(P_\theta(x|c))$. This model can directly sample from the learned latent variable distribution to generate knowledge relational intensity for one particular patent assignee entity. Here, we will introduce how to generate dynamic interaction intensity from CRN using the CVAE model. A tuple (w_{ij}, w_{ik}) and a conditional label c as the inputs, where w_{ij} and w_{ik} are represent the firm-citation relationship. Conditional label c represent the current "cited company".

4.3 Encoder, Decoder and Generator in our model

The inputs of this model consist of the initial firm-citation relationship entity pair (w_{ij}, w_{ik}) and the conditional label c. Encoder module first extracts and take the logarithm of the vector of the firm-citation relationship entity pair. Then maps the $log(w_{ij}, w_{ik})$ to a latent space $Q_\phi(z|w_{ij}, w_{ik}, c)$ by multiply fully connected layers (i.e., we will obtain the latent variable l that preserve the relational information as well as the relationship of the interaction from two firm-citation entities through applying five consecutive linear fully connected layers on the $log(w_{ij}, w_{ik})$). We assume $Q_\phi(z|w_{ij}, w_{ik}, c)$ follow simple Gaussian distribution for the firm-citation relationship pair (w_{ij}, w_{ik}) and the conditional label c. Moreover, for each entity pair (w_{ij}, w_{ik}) and the conditional label c, a mean vector μ and a variance vector σ^2 can be learned as latent variables to model $Q_\phi(z|w_{ij}, w_{ik}, c)$, the detail of μ and σ^2 shows as below:

$$\mu = [l, c] \cdot W_\mu + bias_\mu \tag{8}$$

$$\sigma^2 = [l, c] \cdot W_\sigma + bias_\sigma \tag{9}$$

where the one-hot conditional label $c \in \mathbb{R}$ is used for the patent conditional label c and $|R|$ is the number of all conditions ($|R|$ is $4,681$, which means the $4,681$ companies in this paper). $W_\mu, W_\sigma \in \mathbb{R}^{(Dim_l + |R|) \times Dim_L}$ are the weight terms, and the $bias_\mu, bias_\sigma \in \mathbb{R}^{1 \times Dim_L}$ are bias. Thus, Dim_L is the dimension for latent variables and Dim_l is the dimension for l. The decoder module uses the latent variables μ and σ^2 and the conditional label c to reconstruct the firm-citation entity pair, and implement the $P_\theta(w_{ij}, w_{ik}|z, c)$.

This module first uses the μ, σ^2, and the latent value z directly getting from the $N(\mu, \sigma^2)$, where $z = \mu + \sigma \cdot \epsilon, \epsilon$ is the error that random pick from Gaussian distribution. Similarly, as we use the of multiple linear fully connected layers for the mapping in the encoder module, multiple linear fully connected layers are used for an inverse mapping in the decoder module, and obtain $log(w_{ij}, w_{ik})'$ after the decoder process.

The generator tries to sample the $\hat{z}$ directly from $P_\theta(\hat{z}|c)$. After obtained the $\hat{z}$, the same structure from the decoder module is used to decode the patent

relational entity pair. The matrix consists of $w_{ij} \in W$, where W denotes the adjacency matrix, and the definition of this adjacency matrix is as below:

$$w_{ij} = \sum_{1}^{j} (firm_i^{citing}, firm_j^{cited}) \tag{10}$$

where j is number of companies. Specifically, running process has shown in two algorithms. The problem of generating the dynamic interaction intensity can be presented and solved as follows:

Algorithm 1 Citation Relationship Embedding

Input: Patent p, $\forall p \in P$, $P \in \mathbb{R}^{N \times D}$; Φ_{SR} for storing the reliable entity pair (p_i, p_k); Φ_{RV} for recording the verified entity pair (p_i, p_k); citation entity pair $r(p_i, p_j)$, where $r \in R$, and i, j $\in \mathbb{R}^{|R|}$

Output: Adjacency matrix A, $A \in \mathbb{R}^{|R| \times |R|}$ ▷ Generated by Φ_{SR}, Φ_{RV}

  1: Initialization $\Leftarrow \rho(\cdot)$;   ▷ ρ is the threshold of Φ_{RV}
  2: $A \Leftarrow \varnothing$, $\Phi_{SR} \Leftarrow \varnothing$, $\Phi_{RV} \Leftarrow \varnothing$;
  3: **while** $A = \varnothing$ **do**
  4:     **for** $p \in P$ **do**
  5:         **if** p pass $\rho(p)$ **then**
  6:             Φ_{RV}.append(p);
  7:         **else**
  8:             Continue;
  9:         **end if**
 10:     **end for**
 11:     **for** $p_i \in \Phi_{RV}$ **do**
 12:         Φ_{SR}.append($r(p_i, p_j)$), $\forall i \neq j, p_j \in \Phi_{RV}$;
 13:     **end for**
 14:     update A by Φ_{SR};
 15: **end while**
 16: **return** A ▷ Adjacency matrix of R within threshold ρ

(Algorithm 1) Given a list of N tuples of p_i^*, p_k^* with d dimension attributes:

$$\forall p_i^*, p_k^* \in p^*, p^* = f(p|p_1, ..., p_N \leq \rho, \in \mathbb{R}) \tag{11}$$

Φ_{SR} storing the reliable patent entity pair (p_i^*, p_k^*), and Φ_{RV} recording the verified patent entity pair. V must make sure that the total entity is not changeable.

Given a list of n tuples of (p_i^*, p_k^*), where $n \leq N$ with d dimension attributes from equation 11.

$$\forall p_i^*, p_k^* \in p^*, p^* \subseteq \xi(p) \tag{12}$$

(Algorithm 2) The mathematical inference for extracting dynamic inter-action intensity using CVAE from CRN is follows:

$$\mathcal{L}_{CVAE}(p, c;\ \theta, \phi) = -KL[Q_\phi(z|w_{ij}, w_{ik}, c)||P_\theta(z|w_{ij}, w_{ik}, c)]$$
$$+log\,(P_\theta(w_{ij}, w_{ik}|c)) \tag{13}$$

(Reformulated)

$$\mathcal{L}_{CVAE}(p, c;\ \theta, \phi) = -KL[Q_\phi(z|w_{ij}, w_{ik}, c)||P_\theta(z|c)]$$
$$+E[log(P_\theta(w_{ij}, w_{ik}|z, c))] \tag{14}$$

$P_\theta(z|c)\ N(0,1)$ and $Q_\phi(z|w_{ij}, w_{ik}, c)\ N(\mu, \sigma^2)$, then a close-form solution for the (15) can be derived as:

$$-KL[Q_\phi(z|w_{ij}, w_{ik}, c)||P_\theta(z|c)] = -KL[N(\mu, \sigma^2)||N(0,1)]$$
$$= -\frac{1}{2}(tr(\sigma^2) + \mu^T\mu - D_L - logdet(\sigma^2)) \tag{15}$$
$$= -\frac{1}{2}\sum_{l}^{D_L}(\sigma_l^2 + \mu_l^2 - 1 - log(\sigma^2))$$

then, modeled the $log(\sigma^2)$ as σ^2, final close-form of (16) is:

$$-\frac{1}{2}\sum_{l}^{D_L}(exp(\sigma^2)_l + \mu_l^2 - 1 - \sigma_l^2) \tag{16}$$

The second term in (15) use the maximum likelihood, the mean square error is adopted to calculate the difference:

$$E[log(P_\theta(w_{ij}, w_{ik}|z, c))] =$$
$$\frac{1}{2D_p}(\|w_{ij} - w_{ij}{}'\|_2^2 + \|w_{ik} - w_{ik}{}'\|_2^2) \tag{17}$$

The final loss function used for training is formulated as:

$$\mathcal{L}_{CVAE}(p, c;\ \theta, \phi) = -\frac{1}{2}\sum_{l}^{D_L}(exp(\sigma^2)_l + \mu_l^2 - 1 - \sigma_l^2)$$
$$+\frac{1}{2D_p}(\|w_{ij} - w_{ij}{}'\|_2^2 + \|w_{ik} - w_{ik}{}'\|_2^2) \tag{18}$$

5 Result and Discussion

5.1 Experiment Settings

In this paper, we choose the patent dataset granted in 2013 related to $4,681$ companies that have more than 100 patents; the total has $195,045$ patents. The 5-years received citations count as prediction labels. Bibliographic features are

---

Algorithm 2 Generative Model on Citation Relationship

Input: Firm citation entity pair (w_{ij}, w_{ik}), where $r \in R$, and i, j $\in \mathbb{R}^{|R|}$;
     Adjacent network A, $R \in A$, $A \in \mathbb{R}^{|R| \times |R|}$; Label c, $c \in C$, $C \in \mathbb{R}^{|R|}$
Output: Generative feature f_g^* correspond to label c

  1: **Initialization**;   ▷ build model
  2: $W \Leftarrow Random(\cdot)$;   ▷ W: initial weights
  3: $x \Leftarrow Variable(\cdot)$;
  4: $c \Leftarrow Embed(c)$;   ▷ Embed: one hot
  5: CVAE(W, x, c);   ▷ Layer: 5, Activation: Sigmoid, ReLu
  6: **for** epochs k **do**
  7:      $l, W^{Encode} \Leftarrow Encode(x)$;   ▷ l: latent variable
  8:      $W^{Decode} \Leftarrow Decode(z, c)$;   ▷ W^{Decode}: weight of decode model
  9:      $\hat{z} \Leftarrow Gaussian(z) \sim N(0, 1)$;   ▷ $\hat{z}$: random pick
10:      $f_g \Leftarrow Generate(\hat{z}, c | W^{Decode})$;
11:      update $W \Leftarrow Loss(\cdot)$;
12: **end for**
13: **return** f_g^*

---

chosen by the previous studies that have proven high related to patent value. External information was selected from suggestions by the domain expert. The ratio of our sample data for training/testing/validation is 70%/20%/10% respectively, and the batch size is 128, optimizer use Adam and the loss rate is $1e - 3$. Everage running time for training is $7hrs$, for evaluating is $48hrs$, the complexity is $\mathcal{O}(n^2)$. Our device use Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, CPU cores is 6, RAM is 32G, GPU is GeForce RTX 2070 Rev. A.

5.2 Compared methods and metrics

Up to this point, however, there were few empirical studies of directly predicted patent value by learning ways. To highlight the advantages of dynamic interaction intensity, we chose the machine learning models that keep their explanatory power for the prediction task:

- **lightGBM** is a framework for gradient boosted machines. Default light-GBM will train a gradient boosted decision tree and also supports random forests, dropouts meet multiple additive regression trees, and gradient-based one-side sampling [33].
- **XGBoost** is a scalable implementation of Gradient Boosting Decision Tree algorithm. We set the parameter of *objective* in XGBoost as $reg : linear$ and "$random_state = 42$" to build a regression model for patent valuation prediction [34].
- **Linear Regression** is a classical regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.
- **Ridge** is a variant of a linear regression model with L2 regularization.

- **Decision Tree** is a classical algorithm that falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- **Gradient Boosting** is a regression and classification model, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

In line with [12, 35], we use three metrics to verify results. The first is the Mean Absolute Percentage Error (MAPE), which measures the average deviation between predicted and ground truth. The second metric is Mean Squared Error (MSE), which measures the quality of an estimator. The last one is Accuracy (ACC), which measures the fraction of correctly predicted for a given error tolerance ϵ. Let N denote the number of nodes, c^i denote the predicted number of citations for patent i and r^i is the real number of citations.

MAPE in this paper:

$$\frac{1}{N}\sum_{i=1}^{N}|\frac{r^i - c^i}{r^i}| \times 100 \tag{19}$$

MSE in this paper:

$$\frac{1}{N}\sum_{i=1}^{N}(r^i - c^i)^2 \tag{20}$$

ACC in this paper:

$$\frac{1}{N}\sum_{i=1}^{N}|i : \frac{r^i - c^i}{r^i} \leq \epsilon|, \epsilon = 0.1 \tag{21}$$

We designed four cases to explore the usefulness of dynamic interaction intensity for patent value prediction. Thus through the prediction model to examine the validity of the generated dynamic interaction intensity, the four cases are illustrated as follows.

- **Case I:** Compare whether using dynamic interaction intensity. Other feature choosing: 1) Bibliographic fields (foreign references, references, non-patent references, number of claims, number of CPC). 2) External information (Collaboration Distance).
- **Case II:** Compare whether using dynamic interaction intensity. Other feature choosing: 1) Bibliographic fields (foreign references, references, non-patent references, number of claims, number of CPC). 2) External information (Originality Distance)
- **Case III:** Compare whether using dynamic interaction intensity. Other feature choosing: 1) Bibliographic fields (foreign references, references, non-patent references, number of claims, number of CPC). 2) External information (Cultural Distance)

Table 3 Performance evaluation of our method (generating f_g^*) on predicting task. Measure the magnitude of MAE increase or decrease for different standard machine learning methods by adding f_g^*.

	MAE Changes (f_g^* on different characteristics)			
	Case I	Case II	Case III	Case IV
LightGBM	↓ (0.41%)	↓ (25.67%)	↓ (0.31%)	↓ (0.43%)
XGBoost	↓ (0.60%)	↓ (25.81%)	↓ (0.50%)	↓ (0.61%)
Linear Regression	↑ (2.40%)	↓ (78.84%)	↑ (1.69%)	↓ (18.82%)
Ridge	↑ (1.73%)	↓ (69.88%)	↑ (1.16%)	↓ (18.28%)
Decision Tree	↑ (6.94%)	↓ (40.33%)	↑ (5.78%)	↑ (7.77%)
Gradient Boosting	↓ (0.35%)	↓ (23.66%)	↓ (0.35%)	↓ (0.39%)

Table 4 Performance evaluation of our method (generating f_g^*) on predicting task. Measure the magnitude of MAPE increase or decrease for different standard machine learning methods by adding f_g^*.

	MAPE Changes (f_g^* on different characteristics)			
	Case I	Case II	Case III	Case IV
LightGBM	↓ (1.75%)	↓ (56.92%)	↓ (1.42%)	↓ (1.54%)
XGBoost	↓ (1.54%)	↓ (56.88%)	↓ (1.47%)	↓ (1.39%)
Linear Regression	↓ (0.44%)	↓ (46.89%)	↓ (0.39%)	↓ (3.21%)
Ridge	↓ (0.42%)	↓ (43.42%)	↓ (0.37%)	↓ (3.11%)
Decision Tree	↑ (11.47%)	↓ (52.37%)	↑ (7.07%)	↑ (13.11%)
Gradient Boosting	↓ (1.64%)	↓ (56.87%)	↓ (1.47%)	↓ (1.47%)

- **Case IV:** Compare whether using dynamic interaction intensity. Other feature choosing: Bibliographic fields (foreign references, references, non-patent references, number of claims, number of CPC).

5.3 Discussion

Observed in Table 3, 4, and 5, Case II and Case IV significantly reduce MSE and MAPE. Especially in Case II, where MSE and MAPE are reduced considerably, the ACC is greatly improved. However, the performance of Case I and Case III was relatively poor. Although, the external data used in Case I and III are Collaboration Distance and Cultural Distance, which are considered in the management field as the critical factors for judging the quality of patents. But we think it is not suitable for prediction tasks because the calculation process involves geographic distance and the defined cultural scores. The calculation process takes too many external factors that can harm the analysis, and it's easy to destroy data analyzability. Differ from Case I and Case III. The feature used in Case II and Case IV is calculated by simply using citation-related information, which guarantees the analyzability. The dynamic

Table 5 Performance evaluation of our method (generating f_g^*) on predicting task. Measure the magnitude of ACC increase or decrease for different standard machine learning methods by adding f_g^*.

	ACC Changes (f_g^* on different characteristics)			
	Case I	Case II	Case III	Case IV
LightGBM	↑ (0.01%)	↑ (22.99%)	↑ (0.02%)	↑ (0.01%)
XGBoost	↑ (0.22%)	↑ (21.00%)	↑ (0.13%)	↑ (0.18%)
Linear Regression	↓ (0.23%)	↓ (2.38%)	↓ (0.12%)	↓ (1.67%)
Ridge	↓ (0.27%)	↓ (4.46%)	↓ (0.23%)	↓ (2.12%)
Decision Tree	↑ (2.01%)	↑ (29.30%)	↑ (1.62%)	↑ (1.77%)
Gradient Boosting	↑ (0.25%)	↑ (21.54%)	↑ (0.20%)	↑ (0.25%)

interaction intensity we generated can significantly reduce MAPE and MSE and improve model accuracy under this situation.

6 Conclusion

We have proposed an approach for extracting a patent's dynamic interaction intensity, which can be a crucial component for patent valuation. The results show that our method has a positive impact on the task of patent valuation. The generative CVAE model can learn the dynamic structure well, and real-world data collected from the U.S. patent office can reveal the efficacy of our model. The future works are toward the challenging tasks for patent valuation by reassembling the distribution and multi-restriction of a dataset by a multi-deep probability learning framework.

References

[1] Fu, T.-y., Lei, Z., Lee, W.-C.: Patent citation recommendation for examiners. In: 2015 IEEE International Conference on Data Mining, pp. 751–756 (2015). IEEE

[2] Jin, B., Ge, Y., Zhu, H., Guo, L., Xiong, H., Zhang, C.: Technology prospecting for high tech companies through patent mining. In: 2014 IEEE International Conference on Data Mining, pp. 220–229 (2014). IEEE

[3] Vitt, C.A., Xiong, H.: The impact of patent activities on stock dynamics in the high-tech sector. In: 2015 IEEE International Conference on Data Mining, pp. 399–408 (2015). IEEE

[4] Lee, C., Cho, Y., Seol, H., Park, Y.: A stochastic patent citation analysis approach to assessing future technological impacts. Technological Forecasting and Social Change **79**(1), 16–29 (2012)

[5] Su, H.-N., Moaniba, I.M.: Does geographic distance to partners affect firm r&d spending? the moderating roles of individuals, firms, and countries. Journal of Business Research **106**, 12–23 (2020)

[6] Guan, J.C., Yan, Y.: Technological proximity and recombinative innovation in the alternative energy field. Research Policy **45**(7), 1460–1473 (2016)

[7] Hall, B.H., Jaffe, A.B., Trajtenberg, M.: The nber patent citation data file: Lessons, insights and methodological tools. Technical report, National Bureau of Economic Research (2001)

[8] Layne-Farrar, A., Lerner, J.: To join or not to join: Examining patent pool participation and rent sharing rules. International Journal of Industrial Organization **29**(2), 294–303 (2011)

[9] Yu, X., Gu, Q., Zhou, M., Han, J.: Citation prediction in heterogeneous bibliographic networks. In: Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 1119–1130 (2012). SIAM

[10] Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: Towards a stratified learning approach to predict future citation counts. In: IEEE/ACM Joint Conference on Digital Libraries, pp. 351–360 (2014). IEEE

[11] Yan, R., Tang, J., Liu, X., Shan, D., Li, X.: Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1247–1252 (2011)

[12] Liu, X., Yan, J., Xiao, S., Wang, X., Zha, H., Chu, S.: On predictive patent valuation: Forecasting patent citations and their types. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)

[13] Shen, H., Wang, D., Song, C., Barabási, A.-L.: Modeling and predicting popularity dynamics via reinforced poisson processes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)

[14] Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., Chu, S.M., Zha, H.: On modeling and predicting individual paper citation count over time. In: IJCAI, pp. 2676–2682 (2016)

[15] Yan, J., Xiao, S., Li, C., Jin, B., Wang, X., Ke, B., Yang, X., Zha, H.: Modeling contagious merger and acquisition via point processes with a profile regression prior. In: IJCAI, pp. 2690–2696 (2016)

[16] Falk, N., Train, K.: Patent valuation with forecasts of forward citations.

Journal of Business Valuation and Economic Loss Analysis **12**(1), 101–121 (2017)

[17] Fei, H., Tan, S., Li, P.: Hierarchical multi-task word embedding learning for synonym prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 834–842 (2019)

[18] Hao, J., Chen, M., Yu, W., Sun, Y., Wang, W.: Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1709–1719 (2019)

[19] Kumar, S., Zhang, X., Leskovec, J.: Predicting dynamic embedding trajectory in temporal interaction networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1269–1278 (2019)

[20] Lu, Y., Shi, C., Hu, L., Liu, Z.: Relation structure-aware heterogeneous information network embedding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4456–4463 (2019)

[21] Wang, H., Wang, J., Wang, J., Zhao, M., Zhang, W., Zhang, F., Xie, X., Guo, M.: Graphgan: Graph representation learning with generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[22] Yuan, J., Gao, N., Xiang, J.: Transgate: knowledge graph embedding with shared gate structure. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3100–3107 (2019)

[23] Zhang, Y., Yao, Q., Shao, Y., Chen, L.: Nscaching: simple and efficient negative sampling for knowledge graph embedding. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 614–625 (2019). IEEE

[24] Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710 (2014)

[25] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[26] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: Proceedings of the 24th International

Conference on World Wide Web, pp. 1067–1077 (2015)

[27] Gao, H., Pei, J., Huang, H.: Progan: Network embedding via proximity generative adversarial network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1308–1316 (2019)

[28] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)

[29] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[30] Zhang, C., Li, Y., Du, N., Fan, W., Yu, P.S.: On the generative discovery of structured medical knowledge. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2720–2728 (2018)

[31] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)

[32] Lee, P.-C.: Investigating the knowledge spillover and externality of technology standards based on patent data. IEEE Transactions on Engineering Management (2019)

[33] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30**, 3146–3154 (2017)

[34] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[35] Ponta, L., Puliga, G., Oneto, L., Manzini, R.: Identifying the determinants of innovation capability with machine learning and patents. IEEE transactions on engineering management (2020)