Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

On Predictive Patent Valuation: Forecasting Patent Citations and Their Types

Xin Liu,[1*] **Junchi Yan,**[12*] **Shuai Xiao,**[3] **Xiangfeng Wang,**[1] **Hongyuan Zha,**[1] **Stephen M. Chu**[2]
[1]East China Normal University [2]IBM Research – China [3]Shanghai Jiao Tong University
xinchrome@gmail.com, {jcyan,xfwang,zha}@sei.ecnu.edu.cn, benjaminforever@sjtu.edu.cn, schu@us.ibm.com

Abstract

Patents are widely regarded as a proxy for inventive output which is valuable and can be commercialized by various means. Individual patent information such as technology field, classification, claims, application jurisdictions are increasingly available as released by different venues. This work has relied on a long-standing hypothesis that the citation received by a patent is a proxy for knowledge flows or impacts of the patent thus is directly related to patent value. This paper does not fall into the line of intensive existing work that test or apply this hypothesis, rather we aim to address the limitation of using so-far received citations for patent valuation. By devising a point process based patent citation type aware (self-citation and non-self-citation) prediction model which incorporates the various information of a patent, we open up the possibility for performing predictive patent valuation which can be especially useful for newly granted patents with emerging technology. Study on real-world data corroborates the efficacy of our approach. Our initiative may also have policy implications for technology markets, patent systems and all other stakeholders. The code and curated data will be available to the research community.

Introduction

Questions involving patent valuation have intrigued scholars for decades. Assessing the value of a patent is crucial both at the licensing stage and during the resolution of a patent infringement lawsuit. The demand for Intellectual Property (IP) valuation is increasing, with IP now being of great importance or collateral in financial investment and loan decisions. Many companies, especially high-tech firms, apply and receive thousands of granted patents each year[1]. IP valuation is also growing more important so as to make fair and reliable valuation information available to investors.

However, patents often show less explanatory power than other R&D measures. (Griliches, Pakes, and Hall 1986) find that patents are an extremely noisy measure of the underlying economic value of the innovations with which they are associated. By their nature, patented technologies differ greatly in quality, and the distribution of patent values is skewed (Pakes 1984; Griliches 1990). Hence estimating the patent value via a principled approach is a central issue to technology markets, patent systems and other stakeholders.

In current practices, patents without established market values (e.g., no negotiated royalty rates) are often valued by comparing the number of citations the patent has received to the numbers received by other patents whose market values are established. Specifically, the citations that a patent receives from subsequent patents are called forward citations. Forward citation analysis has become increasingly dominant for patent valuation in litigation, transfer pricing, and other purposes; see reference cases e.g., Oracle v. Google, Finjan v. Blue Coat Systems, Realtek Semiconductor v. LSI and Agere[2]. The standard method is to compare the patent being valued against patents, or portfolios of patents, with established values, such as licensing fees or sales prices.

However, patent valuation that relies on the count of *observed* and *received* citations has fundamental limitations. In an extreme case, a patent that has received no citation by the time of appraisal can not be valued or is trivially valued as zero, though the patent might receive numerous citations in future. For most of recently-issued patents, if the patent has received only a few citations, then comparisons are subject to the noisiness of small numbers. Even for relatively old patents that have had time to accumulate many citations, two patents currently with the same citation counts in general will not have the same number of lifetime citations.

This paper initiates an alternative way by adopting the prediction of the accumulated citation count that a patent will receive in a relatively long prospect time window, e.g. 5 or 10 years. This is a natural generalization (but a very early exploration in this direction (Falk and Train 2016)) towards the current prevalent citation based patent valuation approaches, whereby the future citations are also accounted.

*Correspondence author is Junchi Yan. The first two authors contribute equally to this paper. This research was supported by National Natural Science Foundation of China (61602176, 61672231), Science and Technology Commission of Shanghai Municipality (15JC1401700, 14XD1402100), China Postdoctoral Science Foundation Funded Project (2016M590337).

[1]According to http://www.ificlaims.com/, in 2015, IBM received 7,355 granted U.S. patents, followed by Samsung (5,072), Canon (4,134), Qualcomm (2,900), Google (2,835), Toshiba (2,627), Sony (2,455), LG Electronics (2,242), Intel (2,048).

[2]Oracle v. Google Inc., 3:10-cv-03561-WHA (8/24/2012, N.D. Cal.); Finjan, Inc. v. Blue Coat Systems, Inc., 13-cv-03999-BLF (7/14/2015, N.D.Cal.); Realtek Semiconductor Corp. v. LSI Corp. and Agere Systems LLC, C-12-03451-RMW (1/6/2014, N.D.Cal.).

On top of conquering the limitation of using current citations discussed above, our new methodology is also suited in practical settings. In fact, there are more and more patent infringement litigations along with massive young patents being granted. The so-far received citation cannot reflect the future potential and this calls for a principled approach.

One shall note without combining with effective citation based patent valuation methods e.g. (Harhoff, Scherer, and Vopel 2003) and additional financial value as reference cases (Hall, Jaffe, and Trajtenberg 2000), our citation prediction method alone has no direct output for patent valuation. In fact, we believe citation is a key proxy but not a precise delegator for patent value. We leave for immediate future work to integrate the predictive citations and other factors into a real-world patent valuation web system allowing people to valuate their patents by uploading a patent number or a whole patent text. Indeed, predictive valuation can be useful especially for emerging technologies owner, like startups whose valuation may largely depend on their (newly) patented cutting-edge technologies.

Related Work and Overview

Citation for patent valuation Though value constructs are not always precisely defined in the patent literature, it is in general (arguably) accepted that patent citation counts are positively and significantly related to patent value or patent quality (Carpenter, Narin, and Woolf 1981; Trajtenberg 1990; Harhoff et al. 1999; Anthony and Mogee 2002; Harhoff, Scherer, and Vopel 2003; Martinez-Ruiz and Aluja-Banet 2009; Sterzi 2013). Meanwhile, citation data is also often used for patent retrieval (Fujii 2007). Therefore, patent citation has been used as a proxy for patent value regarding with innovation (Ahuja and Lampert 2001) and knowledge flow (Rosenkopf and Almeida 2003). In (Hall, Jaffe, and Trajtenberg 2001), the authors define a 'Generality Index' that describes the variety of fields of a patent's forward citations, which is used by (Layne-Farrar and Lerner 2011) for the examination of patent pools. (Hall, Jaffe, and Trajtenberg 2001) also define an 'Originality Index' for a patent's backward citations which refer to the citation to previous work by the studied patent, which is followed by (Gompers, Lerner, and Scharfstein 2005) to study the creation of startups.

Implicit factors for patent valuation Besides citation count, there are other factors discussed by previous works e.g. number of patent claiming points and the content, number of inventor, classification, applicant type and patent family. These factors in fact are implicitly reflected by the citation count and hence a common practice is directly using citation rather than drilling down to the behind variables.

For instance, more claiming points tend to increase the possibility of patent renew (Moore 2004; Liu et al. 2008), which signifies the patent value. In (Han and Sohn 2015), text mining is performed to extract keywords from patent claims and to measure the Euclidian distance between a patent and its backward or forward cited patents regarding with their claims. In fact, the similarity among the claims of relevant patents will affect the risk of patent infringement. The inventor list also relate to the patent quality and (Sapsalis, La Potterie, and Navon 2006) find the number of inventors can usually reflect the importance and investment having been made by the patent applicant. The number of assigned international patent classifications (IPCs) indicates the number of application areas for a patent (Fischer and Henkel 2011). Therefore, the factors that are important in affecting patent value in terms of novelty are large numbers of claims, inventors, and IPCs (Martinez-Ruiz and Aluja-Banet 2009). Another important factor is the number of patent families (Harhoff, Scherer, and Vopel 2003).

Citation prediction Citation prediction has drawn attentions in the field of scientific impact analysis, and specifically paper citation prediction. A body of literatures formulate the paper citation prediction problem as a regression task. (Yan et al. 2011; Chakraborty et al. 2014) extract author-wise attributes, paper-specific and venue-centric features to build regression model where the supervision is based on the citation counts in a predefined time window. (Yu et al. 2012) study the problem of predicting the citation between a pair of papers. Similar to regression based methods, their approach is based on link prediction rather than predicting the citation count at an arbitrary time point.

Apart from the standard Poisson process model (Falk and Train 2016), for a related task of paper citation prediction, many methods adopt advanced point process models (Zhou, Zha, and Song 2013b; Yang and Zha 2013; Wang, Song, and Barabási 2013; Shen et al. 2014; Xiao et al. 2016; Yan et al. 2016; Xiao et al. 2017). One technical difference among them is how prior is used (Shen et al. 2014), and whether the learning problem is solved in a closed-form by differential equation (Wang, Song, and Barabási 2013) or iteratively, e.g. by gradient descent (Xiao et al. 2016).

Main idea and highlight We treat the citation sequence received by a patent as a point process, and model its conditional intensity function by a mixture of its intrinsic features and the conditioned history citations. The model is learned from observed history and used for future citation prediction over time. Our study contributes to the growing body of work on patent valuation, while in an orthogonal and relatively new direction i.e. predictive citation modeling. Specifically, the highlights of the paper are:

1) We propose to predict the future citations of a patent to push forward the frontier of patent valuation. The novelty pops up as most existing work focus on valuation by so-far received citations and citation itself has rich implications for different stakeholders. **2)** Guided by clear physical meaning, the proposed point process model is designed such that the time-varying features and a time-delayed effect kernel can be readily encoded with a tailored learning algorithm. Meanwhile, the sample-specific and sample-agnostic parameters are carefully balanced to avoid model underfitting and overfitting. **3)** To our best knowledge, this is the first work for predicting citations in a self/non-self citation type aware fashion. The model thus can potentially be applied to other scenarios e.g. impact prediction for papers, researchers or affiliations as non-self citation often out-weights self-citation. **4)** We verify the efficacy of our method on a dataset collected from multiple open venues including National Bureau of Economic Research (NBER) (http://www.nber.org/patents/),

Table 1: Patent features used in our method. The definition for 'Measure of Generality' and 'Measure of Originality' can be found in (Trajtenberg, Henderson, and Jaffe 1997).

Variable	Description	Static?
claims	number of claims	✓
inventors	number of inventors	✓
backwards	number of backward citation	✓
classification	patent classification	✓
assignee type	organizations, governments or individuals	✓
backward lag	mean backward citation lag	✓
self-citations	percentage of self-citations	✓
backward similarity	mean document similarity with backward citations	✓
originality	measure of originality*	✓
forward lag	mean forward citation lag	×
forward similarity	mean document similarity with forward citations	×
legal status	current legal status, alive or expired	×
generality	measure of generality*	×

United States Patent and Trademark Office (USPTO) (https://www.uspto.gov/), Google Patent Search.

Finally, it is important to note that the presented model can go beyond patent citation prediction. Methodologically it can be applied in other scenarios when similar event dynamics exist akin to patent citation data, especially if event type need be considered. It shall also be noted that this paper is not an effort for verifying the hypothesis that citation is positively related to patent value, which has been discussed extensively in literature and is out of the scope of this paper.

Data Collection and Preprocessing

Our dataset originates from patent information in NBER (Hall, Jaffe, and Trajtenberg 2001), USPTO and Google Patent Search. The NBER dataset comprises abundant information on almost 3 million U.S. patents, we add more patent information, e.g. patents profiles, patent citations, patent full text, legal status. In fact, in the US system, when a new patent is filed, the inventor references the existing prior art, and demonstrates how the new invention represents an advance over this prior art. We consider important patent features and Table 1 presents an overview of the used features. In this study, we use 10,000 live patents collected by joining from multiple data sources.

Intrinsic (static) features For a granted patent, some of its profile information can be immediately obtained, e.g. patent claim, inventor, patent classification and assignee types, these features are termed as 'intrinsic features' in this paper. To some extent, they can reflect the intrinsic quality of inventions. In general, one can regress the coefficients of the features by applying regression models for a specific task.

Time-varying features Besides the various static features discussed above, there are also rich features evolving with time, e.g. citation lag, technological category, measure of generality, and also derived complex measurements like exponential moving average of text similarity of forward citations, mean forward citation lag, some of which are described in detail in (Hall, Jaffe, and Trajtenberg 2001). Among these features, one important indicator is the legal status of a patent. Patent holders must periodically pay a renewal fee to keep their patent rights, and renewal fee increases with the lapse of time after patent registration (Han and Sohn 2015). If a patent has high market valuable and/or indeed includes core technology, the holder will afford the increased renew fee and keep the patent alive. In our model, legal status is considered as an important time-varying feature, which may be either 'alive' or 'expired'. The associated date information for the status is obtained from USPTO.

Two-type (forward) citations For each patent, we construct a forward citation sequence i.e. the citation received since its grant. Meanwhile, we compute the number of backward citation i.e. the older patents cited by the current patent. In this paper, if not otherwise explicitly specified, 'citation' refers to the forward citations based on which we build our model. Compared with the intrinsic features, the (forward) citation sequence can reflect the dynamic behavior of patents. As discussed earlier in the paper, citation count is widely used as an important proxy for patent valuation.

In line with the definition by (Hall, Jaffe, and Trajtenberg 2001), we further divide the citation into two types: i) self-citation, which refers to the citing patent and cited patent are from the same assignee (not necessarily the same inventors); ii) the non-self citations, i.e. the involved patents in the citation are from different assignees. The distinction for self-citation from others has important implications, inter alia, for the study of spillovers (Hall, Jaffe, and Trajtenberg 2001): presumably citations to patents belonging to the same assignee indicate the knowledge transfers are mostly internalized. Comparatively citations to patents of others are closer to the true notion of spillovers.

Document similarity Similarity between patents' full texts is meaningful information. If a patented invention is considered as an essential technology, it tends to receive forward citations from similar inventions with overlapping or related content. To some extent, information about semantic similarity reflects the originality of an invention. In order to measure the semantic similarity between patent documents, we vectorize the patent claims and descriptions with Term Frequency Inverse Document Frequency (TF-IDF) (Ramos 2003). We also further transform the high-dimensional document vectors from the TF-IDF output into topic representations by applying the Latent Semantic Index (LSI) (Deerwester, Dumais, and Harshman 1990) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) respectively and concatenate as a whole as our text feature vector. The Euclidian distance distribution measured in the vectorized feature space is shown in Fig. 1 which shows the LSI and LDA feature have better discriminative capability.

Model and Algorithm

In this section, we introduce the proposed point process based citation prediction model, discussing our motivations and differences compared to related previous works.

Point processes are effective and compact mathematical tool to model the occurrences of discrete events. The conditional intensity function $\lambda(t)$ represents the expected instantaneous rate of future events at time t conditioned on the history. One basic intensity function is the constant $\lambda(t) = \lambda_0$

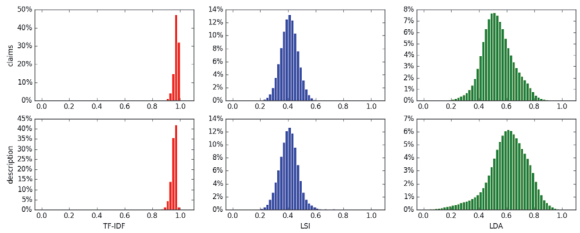


Figure 1: Euclidian distance distribution by applying different features. LSI and LDA lead to scattered distance distribution and we use their concatenation as our textual feature.

over time, as used in the homogeneous Poisson process. Another popular conditional intensity function is the one used by the Hawkes process: $\lambda(t) = \gamma_0 + \alpha \sum_{t \in \tau} \gamma(t, t_i)$ where τ denotes the event history and $\gamma(t, t_i) \geq 0$ is a triggering kernel capturing the temporal dependency. We build our intensity function similar to the Hawkes process as it involves two terms that model the intrinsic intensity and the temporal dependency respectively. The readers are referred to (Aalen, Borgan, and Gjessing 2008) for a textbook treatment for different forms of point processes and their intensity functions.

In our case, we have two interdependent intensity functions for the self-citation behavior ($m = 1$) and non-self-citation behavior ($m = 2$) for patent i, respectively. Note we only consider the relation over citations of different types associated with the single patent that receives these citations.

Conditional intensity function modeling

In principle, we aim to derive a compact and effective intensity function formulation, which ideally can decouple the effect of the intrinsic features (e.g. inventors, claims) from those time-varying features (e.g. legal status) and dynamic citations. Specifically, our conditional intensity function comprises of two ingredients as follows.

Intrinsic component The first term encodes the intrinsic time-constant features to account for the intrinsic quality of the patent. We define the feature vector for patent i, which is given by $\boldsymbol{x_i} = [x_{i1}, x_{i2}, ..., x_{iP}]$ for $x_{i1} = 1$, and $[x_{i2}, x_{i3}, ..., x_{iK}]$ are intrinsic features, while $[x_{i,K+1}, x_{i,K+2}, ..., x_{i,P}]$ are time-varying features representing the patent's transient state at time t. According to (Pakes and Schankerman 1984; Bessen 2008), a patent's quality will normally depreciate over time due to the technological obsolescence or because competitors are able to 'invent around' the patent. Specifically, we follow the constant depreciation rate assumption suggested in (Bessen 2008), also for its computational tractability and popularity in the statistics literature. The intrinsic part is modeled as follows:

$$\lambda_i^m(t)_{int} = \sum_{p=1}^{K} \beta_p^m x_{ip} e^{-\theta_p^m t},$$

where $\boldsymbol{\beta}^m \in \mathbb{R}^K$ is the encoding coefficients associated with the intrinsic feature vector $\boldsymbol{x}_i \in \mathbb{R}^K$. Note here we allow the depreciation kernel $e^{-\theta_p^m t}$ have its separate parameters θ_p^m for each intrinsic feature value x_{ip} to improve its modeling capability. It is believed some features have long-term impact, while some tend to lose their effect rapidly over time. Our model allows for this heterogeneity. From the parameter learning perspective, the model is still manageable since the parameter θ_p^m are shared by all patent samples. Note we write the formula in scalar form for all the parameters as the involved exponential operator disallows a compact form by vector multiplication.

Triggering component Then we consider the dynamic effect. We first define the received citation sequence of patent i as $S_i = \{(t_j, m_j) | j = 1, ..., N_i\}$ where m is the type of forward citation events, and we consider two types of citations: self-citation and non-self-citation in this paper. Like many real-world phenomenon e.g, paper citation (Wang, Song, and Barabási 2013; Xiao et al. 2016), patent citation events tend to exhibit temporal clustering behavior which sometimes is also interpreted as the Matthew Effect – richer get richer. Thus we build our triggering term as a recency-weighted effect accumulation of the received citations, whereby each past citation for $t_j < t$ imposes a positive impact on the intensity function. We write out the devised triggering part as follows:

$$\lambda_i^m(t)_{tri} = \sum_{j, t_j < t} \underbrace{\left(\gamma_i^m e^{-w_i^m |t - t_j - b_i^m|}\right)}_{\text{sample-specific term}} \underbrace{\left(\sum_{p=1}^{P} \alpha_p^{mm_j} x_{ip}(t_j)\right)}_{\text{sample-agnostic term}},$$

The above formulation adds up all effects from each past citation and the parameters can be divided into two categories: i) individual patent specific parameter γ_i^m and w_i^m, b_i^m. The former is the weight coefficient of the triggering effect from the history citations, and the latter measures the scale and shape of the recency kernel. This is because we believe the effect of its received citations γ_i^m can be heterogeneous among individual patents and so for the specific triggering kernel shape controlled by w_i^m, b_i^m. In particular, here we adopt an exponential decaying model with a time delay b_i^m to enhance its flexibility; ii) individual patent agnostic (i.e. cross-sample shared) parameter $\alpha_p^{mm_j}$ which encodes the mutual-effect by the two citation types for patents i.e. self-citation and non-self-citation. The rationale is that we try to avoid involving too many parameters in our model which may incur overfitting especially when event data is sparse. In fact, consider the size of the feature vector $\boldsymbol{x}_i(t)$ i.e. $K > 10$ in our case, with multiplication by the number of citation type can be further intimidating if they were independently estimated for each patent.

Based on the above discussion and in line with (Xiao et al. 2016), we write out the proposed conditional intensity function as the summation of the intrinsic and triggering terms:

$$\lambda_i^m(t) = \lambda_i^m(t)_{int} + \lambda_i^m(t)_{tri} \quad (1)$$

For notational clarity, we further rewrite the parameters in their matrix form, i.e. individual patent agnostic parameters: $\boldsymbol{\beta} \in \mathbb{R}^{M \times K}$, $\boldsymbol{\theta} \in \mathbb{R}^{M \times K}$, $\boldsymbol{\alpha} \in \mathbb{R}^{M^2 \times P}$; and individual patent specific parameters $\boldsymbol{w} \in \mathbb{R}^{M \times I}$, $\boldsymbol{b} \in \mathbb{R}^{M \times I}$, $\boldsymbol{\gamma} \in \mathbb{R}^{M \times I}$. Here $M = 2$ is the number of citation types, and I is the total number of studied patents.

Model learning and prediction

For each observed forward citation cascade $\{t_j, m_j\}_{j=1}^{N_i}$, let $t_0 = 0$, $t_{N_i+1} = T_i$, $g_i^m(t) = \exp(-w_i^m |t - b_i^m|)$,

and its integral $G_i^m(t) = \int_0^t g_i^m(s)ds$, $\Gamma_i^{m_1 m_2}(t) = \gamma_i^{m_1} \sum_{p=1}^P \alpha_p^{m_1 m_2} x_{ip}(t)$, $\mu_i^m(t) = \sum_{p=1}^K \beta_p^m x_{ip} e^{-\theta_p^m t}$ and its integral $U_i^m(t) = \int_0^t \mu_i^m(s)ds$. Then by combining the parameters into $\mathbf{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{b}\}$, and slightly abusing G for G^m, the log likelihood of our model is given by:

$$\begin{aligned}\mathcal{L}_i(\mathbf{\Theta}) &= \sum_{j=1}^{N_i} \log \lambda_i^{m_j}(t_j) - \sum_{m=1}^{M} \int_0^{T_i} \lambda_i^m(t)dt \\ &= \sum_{j=1}^{N_i} \log \left(\mu_i^{m_j}(t_j) + \sum_{k<j} \Gamma_i^{m_j m_k}(t_k) g_i^{m_j}(t_j - t_k) \right) - \\ &\sum_{m=1}^{M} U_i^m(T_i) - \sum_{m=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{j} \Gamma_i^{m m_j}(t_j) \Big(G(t_{j+1} - t_k) - G(t_j - t_k) \Big),\end{aligned}$$

Model parameters are estimated by minimizing negative log-likelihood function over all training samples. Note that train time window $[0, T_i]$ for each patent can be different.

$$\min_{\mathbf{\Theta}} \mathcal{L}_\sigma(\mathbf{\Theta}) = -\sum_{i=1}^{I} \mathcal{L}_i(\mathbf{\Theta}) + \frac{1}{2}\sigma(||\boldsymbol{\alpha}||_F^2 + ||\boldsymbol{\beta}||_F^2) \quad (2)$$

We use squared Frobenius norm $||\cdot||_F^2$ to avoid overfitting. We employ the Majorization-Minimization principle (Lange, Hunter, and Yang 2000) to update the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ by minimizing a tight upper-bound for $\mathcal{L}_\sigma(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ based on the Jensen's inequality (Yan et al. 2016) as follows:

$$\begin{aligned}Q_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} = -\sum_{i=1}^{I} \Bigg[\sum_{j=1}^{N_i} \Bigg(\sum_{p=1}^{K} \psi_{jp}^i \log \frac{\beta_p^{m_j} x_{ip} \exp(-\theta_p^{m_j} t_j)}{\psi_{jp}^i} + \\ \sum_{k<j} \sum_{p=1}^{P} \phi_{jkp}^i \log \frac{\gamma_i^{m_j} \alpha_p^{m_j m_k} x_{ip}(t_k) g_i^{m_j}(t_j - t_k)}{\phi_{jkp}^i} \Bigg) - \\ \sum_{m=1}^{M} U_i^m(T_i) - \sum_{m=1}^{M} \sum_{j=1}^{N_i} \sum_{k=1}^{j} \Gamma_i^{m m_j}(t_j) \Big(G(t_{j+1} - t_k) - \\ G(t_j - t_k) \Big) \Bigg] + \frac{1}{2}\sigma(||\boldsymbol{\alpha}||_F^2 + ||\boldsymbol{\beta}||_F^2) \quad (3)\end{aligned}$$

where ψ_{jp}^i can be interpreted as the effect that feature p triggers citation t_j, ϕ_{jkp}^i can be interpreted as the effect that citation t_k triggers citation t_j. Expectation step and minimization step are performed iteratively until convergence.

Expectation step:

$$\psi_{jp}^i := \frac{\beta_p^{m_j} x_{ip} \exp(-\theta_p^{m_j} t_j)}{\mu_i^{m_j}(t_j) + \sum_{k<j} \Gamma_i^{m_j m_k}(t_k) g_i^{m_j}(t_j - t_k)} \quad (4)$$

$$\phi_{jkp}^i := \frac{\gamma_i^{m_j} \alpha_p^{m_j m_k} x_{ip}(t_k) g_i^{m_j}(t_j - t_k)}{\mu_i^{m_j}(t_j) + \sum_{k<j} \Gamma_i^{m_j m_k}(t_k) g_i^{m_j}(t_j - t_k)} \quad (5)$$

Minimization step By zeroing partial derivatives w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, i.e. $\frac{\partial Q}{\partial \alpha} = \frac{\partial Q}{\partial \beta} = \frac{\partial Q}{\partial \gamma} = 0$, Q is minimized, thus:

$$\alpha_p^{m_1 m_2} := \frac{-C_p^{m_1 m_2} + \sqrt{(C_p^{m_1 m_2})^2 + 4\sigma B_p^{m_1 m_2}}}{2\sigma} \quad (6)$$

$$\beta_p^m := \frac{-E_p^m + \sqrt{(E_p^m)^2 + 4\sigma D_p^m}}{2\sigma} \quad (7)$$

$$\gamma_i^m := \frac{\sum_{j:1 \le j \le N_i, m_j = m} \sum_{k=1}^{j-1} \phi_{jkp}^i}{\sum_{j=1}^{N_i} \sum_{k=1}^{j} \sum_{p=1}^{P} A_{ijkp}^{t_j} \alpha_p^{m m_j}} \quad (8)$$

Algorithm 1 Type-aware prediction for patent citations

```
Input: event sequence S_i and features x_i for each patent i
Initialize Θ = {α, β, γ, θ, w, b} randomly
while L_σ(Θ) not converged do
    while Q_{α,β,γ} not converged do
        Calculate ψ, φ via Eq. 4, Eq. 5
        Update α, β, γ via Eq. 6, Eq. 7, Eq. 8
    end while
    Update θ, w, b via gradient descent
end while
```

for $A_{ijkp}^{t_n} = (G(t_{j+1} - t_k) - G(t_j - t_k)) x_{ip}(t_n)$, $B_p^{m_1 m_2} = \sum_{i=1}^{I} \sum_{j:j \le N_i, m_j = m_1} \sum_{k:k<j, m_k = m_2} \phi_{jkp}^i$, $C_p^{m_1 m_2} = \sum_{i=1}^{I} \sum_{j:j \le N_i, m_j = m_1} \sum_{k:k \le j, m_k = m_2} A_{ijkp}^{t_k} \gamma_i^{m_j}$, $D_p^m = \sum_{i=1}^{I} \sum_{j:j \le N_i, m_j = m} \psi_{jp}^i$, $E_p^m = \sum_{i=1}^{I} \frac{x_{ip}}{\theta_p^m}(1 - e^{-\theta_p^m T_i})$. To minimize $\mathcal{L}_\sigma(\mathbf{\Theta})$, we update $\boldsymbol{\theta}$, $\boldsymbol{w}$ and $\boldsymbol{b}$ by gradient descent. The overall algorithm is shown in Alg.1.

The learned model is applied to predict the two types of citations on a yearly basis: the model trained using window $[0, T-1]$ is used to predict citations for year T. Then based on citation in $[0, T]$, model outputs citations for year $T+1$.

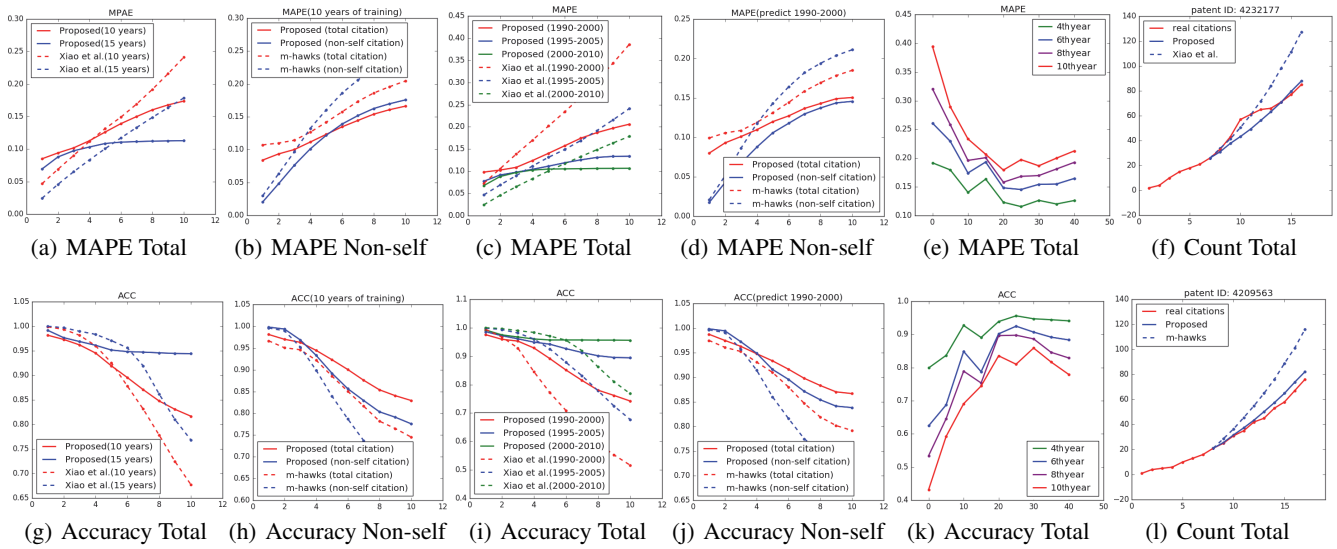
Experiments and Discussion

The experiments are performed on a machine installing Microsoft Windows 10, equipped with 4 cores: Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, and 8GB RAM.

Compared methods and metrics Though there is few direct methods for patent citation prediction, the idea for event forecasting has been studied in literature. Here we compare one state-of-the-art method (Xiao et al. 2016) which has shown superior performance over other popularity prediction models (Shen et al. 2014; Wang, Song, and Barabási 2013). However, the above three methods have the fundamental limitation for being unable to handle citation (i.e. event) type. Thus we also compare with the multi-dimensional Hawkes process (Zhou, Zha, and Song 2013a) that can be used to predict future event at the type level. Note here we remove its low-rank and sparse regularization term for the mutual effect matrix as we only have two dimensions.

In line with (Shen et al. 2014; Xiao et al. 2016), we use two metrics for prediction result evaluation. The first is the Mean Absolute Percentage Error (MAPE). It measures the average deviation between predicted and ground truth over N patents. Let $c^i(t)$ denote the predicted number of citations for patent i up to time t and with $r^i(t)$ its real number of citations, MAPE is given by $\frac{1}{N} \sum_{i=1}^{N} \left| \frac{c^i(t) - r^i(t)}{r^i(t)} \right|$. The other metric is Accuracy. It measures the fraction of papers correctly predicted for a given error tolerance ϵ. Hence the accuracy of popularity prediction on N patents is $\frac{1}{N} \sum_{i=1}^{N} \left| i : \left| \frac{c^i(t) - r^i(t)}{r^i(t)} \right| \le \epsilon \right|$ for $\epsilon = 0.3$ in our experiments. The above two definitions are also applied to the self-citation and non-self citation in a routine fashion.

Results and discussion We choose patents which a) are assigned to U.S. companies or U.S. government and b) are assigned between 1975 and 1985 and c) have at least 5 for-



(a) MAPE Total (b) MAPE Non-self (c) MAPE Total (d) MAPE Non-self (e) MAPE Total (f) Count Total

(g) Accuracy Total (h) Accuracy Non-self (i) Accuracy Total (j) Accuracy Non-self (k) Accuracy Total (l) Count Total

Figure 2: MAPE and accuracy on citation prediction over years. Column 1-2: fix training period to 10 and 15 years, and for total citation and non-self citation respectively. Column 3-4: fix the prediction period to 1990-2000, 1995-2005, 2000-2010, and for total citation and non-self citation respectively. The x-axis denote the prediction year. Column 5: MAPE and accuracy on patents whose observation window is fixed to 10 years and x axis denotes the received citation count within the window. The performance is measured on the (4,6,8,10)-th year for prediction. Column 6: predicted total citation curves on two examples.

ward citations within 5 years after grant. By applying these filter rules, the resulting data size is 11,878, whose average percentage of self-citations is around 24.7%.

The evaluation involves two cases. **In the first case**, we fix the training period to 10 or 15 years for each patent and verify the prediction performance in a period after the training periods. The corresponding performance curves over the prediction years are displayed in Fig.2(a)2(g) (for total citation), Fig.2(b)2(h) (for non-self citation). **In the second and more realistic case**, as shown in Fig.2(c)2(i) (for total citation), Fig.2(d)2(j) (for non-self citation), we fix the prediction period to three periods: 1990-2000, 1995-2005, 2000-2010 for all granted patents in these periods respectively, and use the time period before 1990, 1995, 2000 back to the grant year as the (varying) training period for each patent. As a result, each patent has its own observation window for training. In parallel, we further plot the curves for predicting the non-self and total citations and compare with the multi-dimensional Hawkes model. For all the cases, our method outperforms by a notable margin as shown in Fig.2. We also experimentally find our method can converge at comparable speed with peer methods i.e. multi-dimensional Hawkes model (Zhou, Zha, and Song 2013a) and (Xiao et al. 2016).

We show in Fig. 2(e)2(k) that prediction becomes more difficult when there are fewer citations (in a fixed 10-year observation window), the trend for MAPE and accuracy is similar and when the received citation count of a patent is larger than 20, the prediction performance becomes relatively flat. Note the x-axis in the figure is the received total citation count within the first 10 years since a patent's grant. Different curves denote the prediction curve at a given future year (4, 6, 8, 10). Finally we plot the prediction on two individual patents regarding with its citations in Fig.2(f)2(l). One can find our method can better track the ground truth compared with (Xiao et al. 2016) and the multi-dimensional Hawkes model (Zhou, Zha, and Song 2013a).

Conceptually, we attribute the superior performance of our approach against the peer method (Xiao et al. 2016) and others (Shen et al. 2014; Wang, Song, and Barabási 2013) to the following main differentiating characteristics:

1) Our model allows for encoding time-varying feature and uses them to parameterize both background rate and history effects. Hence the information can be shared among patents through sharing the sample-agnostic parameters. Compared to training different models for each patent, this can help avoid under fitting (when too few parameters are used in a model) or overfitting (when too many parameters are used for an individual patent's model). **2)** Our model accounts for the heterogeneity of citation types and outputs the type-specific probability. Indeed, the directed effects between self and non-self citation can be very different. **3)** We design a more flexible and realistic recency kernel to model the possible delay of the immediate triggering effect i.e. $\exp(|t - t_j - b_i^m|)$. **4)** We devise effective iterative algorithm to learn the model with the above characters.

Conclusion

We have developed a patent citation prediction model which serves as a building block for patent valuation. Our model can cover the dynamics of both self and non-self citations. Results on collected U.S. patent data show the efficacy of our approach. One direction is to extend to the valuation of a patent portfolio which are technologically related or serve

for a strategic impact. Another challenging topic is predicting the maintenance/renew action of patents.

References

Aalen, O.; Borgan, O.; and Gjessing, H. 2008. *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Ahuja, G., and Lampert, C. M. 2001. Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal* 22:521–543.

Anthony, B., and Mogee, M. E. 2002. The many applications of patent analysis. *Journal of Information Science* 28(3):187–205.

Bessen, J. 2008. The value of us patents by owner and patent characteristics. 37(5):932–945.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. 993–1022.

Carpenter, M. P.; Narin, F.; and Woolf, P. 1981. Citation rates to technologically important patents. *World Patent Information* 3(4):160–163.

Chakraborty, T.; Kumar, S.; Goyal, P.; Ganguly, S.; and Mukherjee, A. 2014. Towards a stratified learning approach to predict future citation counts. In *JCDL*.

Deerwester, S.; Dumais, S.; and Harshman, R. 1990. Indexing by latent semantic analysis. 41(6):391.

Falk, N., and Train, K. 2016. Patent valuation with forecasts of forward citations. *Journal of Business Valuation and Economic Loss Analysis*.

Fischer, T., and Henkel, J. 2011. Patent trolls on markets for technology - an empirical analysis of trolls' patent acquisitions. *Research Policy* 41(9):1519–1533.

Fujii, A. 2007. Enhancing patent retrieval by citation analysis. In *SIGIR*, 793–794.

Gompers, P.; Lerner, J.; and Scharfstein, D. 2005. Entrepreneurial spawning: Public corporations and the genesis of new ventures, 1986 to 1999. *The Journal of Finance* 60(2):577–614.

Griliches, Z.; Pakes, A.; and Hall, B. H. 1986. The value of patents as indicators of inventive activity.

Griliches, Z. 1990. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* 28(4):1661–1707.

Hall, B. H.; Jaffe, A. B.; and Trajtenberg, M. 2000. Market value and patent citations: A first look. *No. w7741, National Bureau of Economic Research*.

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

Han, E., and Sohn, S. 2015. Patent valuation based on text mining and survival analysis. *The Journal of Technology Transfer* 40(5):821–839.

Harhoff, D.; Narin, F.; Scherer, F. M.; and Vopel, K. 1999. Citation frequency and the value of patented inventions. *Review of Economics and statistics* 81(3):511–515.

Harhoff, D.; Scherer, F.; and Vopel, K. 2003. Citations, family size, opposition and the value of patent rights. *Research Policy* 32(8):1343–1363.

Lange, K.; Hunter, D.; and Yang, I. 2000. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* 9(1):1–20.

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

Liu, K.; Arthurs, J. D.; Cullen, J. B.; and Alexander, R. T. 2008. Internal sequential innovations: How does interrelatedness affect patent renewal? *Research Policy* 37(5):946–953.

Martinez-Ruiz, A., and Aluja-Banet, T. 2009. Toward the definition of a structural equation model of patent value: Pls path modelling with formative constructs. *REVSTAT–Statistical Journal* 7(3):265–290.

Moore, K. A. 2004. Worthless patents. *George Mason Law & Economics Research Paper* (04-29).

Pakes, A., and Schankerman, M. 1984. The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. 73–88.

Pakes, A. 1984. Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica* 54(4):755–84.

Ramos, J. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first Instructional Conference on Machine Learning*.

Rosenkopf, L., and Almeida, P. 2003. Overcoming local search through alliances and mobility. *Management Science* 49(6):751–766.

Sapsalis, E.; La Potterie, B. V. P. D.; and Navon, R. 2006. Academic versus industry patenting: An in-depth analysis of what determines patent value. *Research Policy* 35(10):1631–1645.

Shen, H.; Wang, D.; Song, C.; and Barabási, A. 2014. Modeling and predicting popularity dynamics via reinforced poisson processes. In *AAAI*.

Sterzi, V. 2013. Patent quality and ownership: An analysis of uk faculty patenting. *Research Policy* 42(2):564–576.

Trajtenberg, M.; Henderson, R.; and Jaffe, A. 1997. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology* 5(1):19–50.

Trajtenberg, M. 1990. A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics* 172–187.

Wang, D.; Song, C.; and Barabási, A. 2013. Quantifying long-term scientific impact. *Science* 342(6154):127–132.

Xiao, S.; Yan, J.; Li, C.; Jin, B.; Wang, X.; Yang, X.; Chu, S.; and Zha, H. 2016. On modeling and predicting individual paper citation count over time. In *IJCAI*.

Xiao, S.; Yan, J.; Yang, X.; Zha, H.; and Chu, S. 2017. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*.

Yan, R.; Tang, J.; Liu, X.; Shan, D.; and Li, X. 2011. Citation count prediction: Learning to estimate future citations for literature. In *CIKM*, 1247–1252.

Yan, J.; Xiao, S.; Li, C.; Jin, B.; Wang, X.; Ke, B.; Yang, X.; and Zha, H. 2016. Modeling contagious merger and acquisition via point processes with a profile regression prior. In *IJCAI*.

Yang, S., and Zha, H. 2013. Mixture of mutually exciting processes for viral diffusion. In *ICML*.

Yu, X.; Gu, Q.; Zhou, M.; and Han, J. 2012. Citation prediction in heterogeneous bibliographic networks. In *SDM*.

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*.