On Modeling and Predicting individual Paper Citation Count over Time

Shuai Xiao[1], **Junchi Yan**[23*], **Changsheng Li**[3], **Bo Jin**[2]
Xiangfeng Wang[2], **Xiaokang Yang**[1], **Stephen M. Chu**[3], **Hongyuan Zha**[2]
[1] Shanghai Jiao Tong University [2] East China Normal University [3] IBM Research - China
{benjaminforever,xkyang}@sjtu.edu.cn
{jcyan,bjin,xfwang,zha}@sei.ecnu.edu.cn, {lcsheng,schu}@cn.ibm.com

Abstract

Evaluating a scientist's past and future potential impact is key in decision making concerning with recruitment and funding, and is increasingly linked to publication citation count. Meanwhile, timely identifying those valuable work with great potential before they receive wide recognition and become highly cited papers is both useful for readers and authors in many regards. We propose a method for predicting the citation counts of individual publications, over an arbitrary time period. Our approach explores paper-specific covariates, and a point process model to account for the aging effect and triggering role of recent citations, through which papers lose and gain their popularity, respectively. Empirical results on the Microsoft Academic Graph data suggests that our model can be useful for both prediction and interpretability.

1 Introduction and related work

Integral to the success of scientific research is the impact of works. Paper citation and its derivatives e.g. g-index [Egghe, 2006], H-index [Hirsch, 2005; Acuna *et al.*, 2012] have become popular measures to gauge the journals, scholars, labs, departments, and institutes [Fuyuno and Cyranoski, 2006], despite their well-known lack of predictive power to future impact [Wang *et al.*, 2013]: current citations and the derived metrics can only capture past accomplishments.

A candidate's potential *future* impact e.g. his/her citation count often plays a more important role for policy/decision making concerning with recruitment, promotion and funding, because the ultimate question is: Who will be the most successful in this position, with this fellowship? When an early-career candidate is selected for a tenure-track position, it is an investment. In those institutions with low tenure rate, this can amount to an outright bet on one scientist who acquire a start-up package up to millions of dollars [Stephan, 2012].

The scientific literature is turning into an unbounded collection such that it becomes intimidating to have a thorough comprehension on relevant papers even in one area. To find frontier research materials, there is also a need for identifying the pertinent and influential work in a setting where a plenty of papers emerge each day, before they become widely recognized. Also, researchers may re-think if their research is on an exciting path or a dead end that will end the careers prematurely. A (reliable) prediction model can serve as a self-evaluation tool to streamline their research agenda.

We agree that the main way of predicting a researcher's future impact is peer-assessment, but also think that algorithmic approaches could be valuable complementary ways, especially for junior scientists representing a group closer to the typical case in which algorithmic approaches will be applied in real academic hiring decisions, under an appropriate mechanism. It is felt that 'pipeline' leaks in the later career decision points, especially confounded with the subjective gender bias in academic career [Ginther and Kahn, 2004].

As a widely recognized metric [Wang *et al.*, 2013] to scientific impact, however, predicting an individual paper's citation count over time is (arguably) very difficult. For instance, a seminal work may start-up by a small number of follow-up papers that builds up to a pioneering work within a field, and it takes a long time before they generate greater impact. Or a researcher may work on a hot topic, and publish a novel method related to this topic, which immediately draw the community's wide attention. Or simply different papers by the same author can have significant citation variation due to various reasons such as the topic, timing, fields, etc. Such heterogeneous citation curves call for advanced models.

We give an overview about the general problem – scientific impact analysis and prediction, and then focus on the literature on (long-term) individual-level paper citation prediction, which is (arguably) more challenging and has become an emerging applied research topic [Wang *et al.*, 2013]. **Scientific impact analysis and prediction** Since predicting individual paper's citation count looks very challenging, and the skewed distribution of citations often obeys a power-law [Dong *et al.*, 2015] or log-normal form [Radicchi *et al.*, 2008], many researchers resort to other more accessible impact analysis and prediction problems. For instance, [Petersen *et al.*, 2014] perform a longitudinal analysis to measure the effect of the central author's reputation on the paper citation rate. [Pan and Fortunato, 2014] give a formal definition concerning the author-wise impact metric *Author Impact*

[*] Corresponding author is Junchi Yan. The work is partially supported by STCSM (15JC1401700, 14XD1402100), NSFC (61527804, 61221001, 61129001), the 111 Program (B07022).

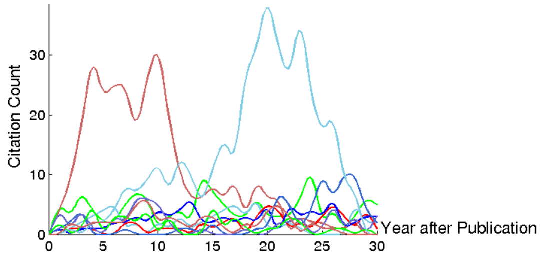

Figure 1: Citation of 30 randomly selected papers over time.

Factor (AIF) and perform empirical study to verify its capability to capture the trends and variations of the impact of the scientific output of scholars in time. However, they are prescriptive methods with no capability of prediction.

For scientific impact prediction, [Dong *et al.*, 2015] formulates and addresses the classification problem of whether a paper will influence an author's h-index *within a pre-defined timeframe*, whereby various factors/features are fed into the classification model including publication topic, venue, author's affiliation etc. Similarly, [Acuna *et al.*, 2012] present an approximate formula obtained from linear regression with elastic net regularization, to estimate the future h-index of life scientists of which the factors relate to current h-index, number of written papers of authors, years since first published article etc. [Penner *et al.*, 2013] identify the flaws of the above linear regression model and empirically show this model tend to overestimate the scientist's future impact. Specifically, they suggest that h-index, or other cumulative metrics are inappropriate targets of *regression* based predictive models in that they contain trivial correlation by construction (see 'Discussion' in [Penner *et al.*, 2013]). In contrast, in this paper, our *point process* based model does not suffer from this limitation, and directly estimates the transient citation counts at any future time point or period. [Stern, 2014] empirically study the social science top-ranked journals and discover that '*half of the papers in the top 20% in 2012 were already in the top 20% in the year of publication (2006)*'.

Paper citation prediction As shown in Fig.1, individual level paper citation prediction is challenging. A line of methods formulate the paper citation prediction problem into a regression task, and examine which covariates are effective input features. [Yan *et al.*, 2011; Chakraborty *et al.*, 2014] extract author-centric attributes (e.g. productivity, co-author network, influence), paper-specific features (e.g. team-size, reference count), and venue-centric features (e.g. venue rank) to build a supervised regression model. [Yu *et al.*, 2012] study the problem of predicting the linkage i.e. citation between a pair of papers. However, their approach is based on link prediction and cannot predict the dynamic citation count at any time point. In the seminal *Science* paper [Wang *et al.*, 2013], the authors propose a point process based behavioral model which tries to capture the dynamics of the individual paper citations. Its intensity function is a multiplication of three factors: i) fitness term, which is paper's intrinsic value being different from paper to paper; ii) the aging effect over time; iii) the reinforcement term indicating the well-documented fact that highly cited papers are more visible and are more likely to be cited again than less-cited contributions. They employ maximum likelihood estimation for individual paper to infer its set of parameters, which is at the risk of over-fitting as also observed in [Wang *et al.*, 2014b]. To mitigate this issue, [Shen *et al.*, 2014] adopt a Bayesian treatment by using a conjugate prior for the fitness parameter μ_d. This prior is not paper-specific nor flexible to capture paper's arbitrary profile.

Despite the recent advances in scientific impact prediction and more specifically, paper citation prediction, it is still unclear and even controversial on the reliability and bound of prediction accuracy of a long-term citation prediction model – see the comments [Wang *et al.*, 2014b] and response [Wang *et al.*, 2014a] published in the *Science communication* papers (http://www.science.com/) after the pioneering work [Wang *et al.*, 2013] and improvement [Shen *et al.*, 2014].

Contribution of the paper This paper is aimed to provide in-depth findings on a recently released real-world dataset.

Specifically, we propose a novel point process model for long-term paper citation prediction, which is also quite general in applicability. Our approach captures the Matthew effect [Merton, 1968] (or accumulated advantage, richer get richer and the poor get poorer) and the recency effect of past citations. In particular, it can help better address the common but unresolved 'second-acts' scenarios in [Wang *et al.*, 2013][1] (a.k.a. 'Sleep Beauty' [Ke *et al.*, 2015]). The covariates w.r.t the author and paper are also incorporated in the intensity function to improve interpretability and mitigates overfitting.

We also provide an empirical analysis of the predictive power and interpretability of the learned point process model on the public Microsoft Academic Graph [Sinha *et al.*, 2015]. Our method consistently outperforms the state-of-the-arts approaches [Wang *et al.*, 2013; Shen *et al.*, 2014].

We think more importantly, this work provides a new investigation on how effective algorithmic citation prediction can be devised, regarding the recent arguments appear on *Science communication* [Wang *et al.*, 2014b; 2014a].

2 Model and algorithm

2.1 Model formulation

The received citation count of an individual paper d during time period $[0, T]$ is characterized by a time-stamped sequence $\{t_i^d\}_{i=0}^n$ when a citation occurs – which we dub it as event in the setting of point process in this paper: $0 = t_0^d \leq t_1^d \leq ... \leq t_i^d \leq ... \leq t_n^d \leq T$. The goal is to model and predict the future citation count over an arbitrary time window given the historical citations and other available covariates.

It is clear that papers having been cited frequently tend to accumulate more citations, especially for recent citations. It is also clear that, with time, even the most novel paper loses its popularity. Some papers, however, seem to have an inherent 'quality' that can be interpreted as a community's recog-

[1] The 'second-acts' e.g. the citation burst for superconductivity papers after the discovery of high-temperature superconductivity in the 1980s, or delayed impact, like the citation explosion to Erdős and Rényi's work 40 years after their publication [Barabási and Albert, 1999], following the emergence of network science [Redner, 2005].

nition to the work. Building on a foundation of the above observations, we derive our prediction model in three regards.

Intrinsic popularity The quality of the paper is its intrinsic factor contributing to its popularity. To some extent, the quality can be measured by its paper/author-specific covariates, such as the H-Index [Hirsch, 2005] of the author, the field where the paper is published etc. In line with [Yan *et al.*, 2011; 2012; Chakraborty *et al.*, 2014], we extract a set of covariates for each paper as listed in Table 1. We also plot three of their scatters regarding with the citation count in Fig.2 and some of them exhibit strong correlations. Therefore, we can use these covariates to regress the associated coefficients of the intrinsic popularity. Here the Lasso (ℓ_1 norm) can be used to induce sparse coefficients to mitigate overfitting.

Impact decaying over time A general and common trend is that the paper's attractiveness fades away over time. This can be explained by the fact that a topic goes through its lifecycle and ends up with an out-of-dated status, or still being a hot topic, but the novelty is incorporated in subsequent work that dilutes its impact and relevance to other work.

Recency-sensitive citation triggering Previous methods [Wang *et al.*, 2013; Shen *et al.*, 2014] ignore the time-stamp, and aggregate all past citations to model the intensity, which might be less effective to capture the citation dynamics – see our experiments. We propose to use a self-triggering process a.k.a. Hawkes processes [Hawkes, 1971; Hawkes and Oakes, 1974] which favor more on recent citations and the effect time-decaying window can be controlled by, for computational effectiveness, a Laplace kernel [Yan *et al.*, 2013; 2015]. The 'recency-sensitive' model can naturally address those papers with spiking citation curve being still remain unresolved in [Wang *et al.*, 2013; Shen *et al.*, 2014].

Hence, we define the citation count intensity of paper d:

$$\lambda_d(t) = \underbrace{\boldsymbol{\beta}\boldsymbol{s}_d(t)}_{\text{Quality}}\underbrace{\mathrm{e}^{-w_{1d}t}}_{\text{Aging}} + \underbrace{\alpha_d \sum_{i,t_i<t} \mathrm{e}^{-w_{2d}(t-t_i)}}_{\text{Triggering weighted by recency}} \tag{1}$$

where $\boldsymbol{s}_d(t) = (1, s_{d1}, s_{d2}, ..., s_{dK})$ is a *row* vector encoding the K paper-specific covariates for paper d and $\boldsymbol{\beta}$ is a *column* vector for the coefficients. $\mathrm{e}^{-w_{1d}t}$ is the aging function accounting for attractiveness decrease since its publication. α_d is the triggering strength of each citation before current time point t, with the decaying effect $\mathrm{e}^{-w_{2d}(t-t_i)}$. For point process, the estimated citation count can in general be computed by integrating Eq.1 over a specified future time period.

2.2 Discussion on peer methods

As an emerging problem, the most relevant work to ours is the Reinforced Poisson Process (RPP) model as presented in [Wang *et al.*, 2013] and [Shen *et al.*, 2014], whereby the latter adds a conjugate prior on the fitness of an individual to the former work [Wang *et al.*, 2013] published in *Science* studying the problem of long-term *individual* citation dynamics. We also mention the form of Hawkes processes [Hawkes, 1971; Hawkes and Oakes, 1974] as our model is partly originated from this type of point process. We also solve the learning problem via a tailored ADMM based algorithm.

Table 1: Paper/author/venue-centric covariates in our model.

Type	Covariates	Description	Rank
Author-wise	hindex	H-index of anthor	1
	authorrank	rank of author	2
	noca	number of co-authors	5
	insitrank	rank of author's institute	6
	producibility	publications by author	7
	authordiv	diversity of author's topic	8
	authorcen	centrality of authors	10
	teamsize	number of authors of papers	14
	insitnum	number of institute	17
Venue-wise	venuerank	rank of venues	3
	venuecen	degree of centrality of venues	13
	venuepub	number of publications of venues	19
	venuediv	topic diversity of venue	21
	venueaut	number of authors of venues	22
Paper-wise	firstPA	first order of preferential attachment	4
	secPA	second order of preferential attachment	9
	topdiv	topic diversity of the paper	11
	filedhot	topic hotness of the paper	12
	refdiv	topic diversity of reference	15
	firstRef	first order of reference	16
	secRef	seconde order of rerference	18
	keydiv	keywork diversity of the paper	20

Reinforced Poisson Process – RPP The seminal work published in *Science* [Wang *et al.*, 2013] begins to study the fundamental problem for the predictability of long-term citation.

Their point process model involves three individual paper-specific parameters: the relative fitness λ_i capturing a paper's importance relative to to other papers; immediacy μ_i governing the time for a paper to reach its citation peak; and a longevity parameter σ_i accounting for the decay rate of its popularity. By solving a master equation associated with the intensity function, they directly give the equation for the number of citations $c_i^t(\lambda_i, \mu_i, \sigma_i)$ at time t (see Eq.2 and Eq.3 in that paper). The parameters are then estimated by least-square-fitting, given historical citation data for paper i.

To overcome the problem that maximum likelihood parameter estimation suffers from overfitting, especially for relatively small sample size as we need to train one model for each paper by its citations, [Shen *et al.*, 2014] adopt a Bayesian treatment by adding a conjugate prior on the relative fitness parameter μ_d via a gamma distribution, for each paper and showcase superior results compared with [Wang *et al.*, 2013] for prediction accuracy. While it incurs some doubts from more recent study [Wang *et al.*, 2014b].

Hawkes Process The intensity of self-exciting Hawkes process [Hawkes, 1971; Hawkes and Oakes, 1974] is given by:

$$\lambda_d(t) = \mu_d + \alpha_d \sum_{j,t_j<t} \mathrm{e}^{-w(t-t_j)}.$$

Here a paper-specific parameter μ_d for paper d is used instead of the parameterized linear regression term $\boldsymbol{\beta}\boldsymbol{s}_d$ by our approach, and other parameters have the similar meaning compared to Eq.1. There are some recent studies/applications on Hawkes processes, [Zhao *et al.*, 2015] use this model to predict the popularity of Twitter, where popularity intensity is determined by a stochastic infectiousness process (Cox process). [Zhou *et al.*, 2013] introduce the low-rank sparsity on the infect matrix formed by the mutually-exciting Hawkes model. [Luo *et al.*, 2015] propose a multi-task learning variation for the mutually-exciting Hawkes model and [Yan *et al.*, 2015] adopt it for sales pipeline modeling.

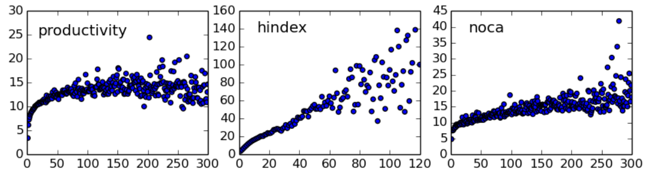

Figure 2: Covariates (x axis) – citation (y axis) scatter.

Compared with the above methods, especially [Wang *et al.*, 2013; Shen *et al.*, 2014], the bullets of our approach are:

i) **Introducing paper-specific covariates** The paper-specific covariates are involved via parameterizing the quality term (see Eq.1). In RPP and Hawkes process, for each paper, their quality term μ_d is modeled by one their own parameter respectively, which we think is the main reason for over-fitting. One shall note that in their methods, each of these parameters is learned from its past citations of each paper independently. Meanwhile, the covariates are not used which otherwise can play a bridge to cross-distributing the information over training samples i.e. papers. For instance, for those papers with very short observation window and with few citations, it is difficult to interpret and predict the behavior of such papers without exploring the covariates used in Table 1. In this sense, involving the informative covariates will mitigate the over-fitting problem and meanwhile help improve the interpretability of our model. Note in the improved RPP [Shen *et al.*, 2014], they only impose a prior on the global distribution of μ_d, without using the covariates to parameterize the prior and such valuable information is ignored.

ii) **Modeling citation recency by self-exciting kernel** We account for citation recency by modeling triggering effect in continuous time space. This feature is inspired from the Hawkes process. To our best knowledge, this is the first time for adapting this component in paper citation prediction. More importantly, we find this recency-weighted triggering model is more appropriate for the citation dynamics, especially for those 'second-acts' and 'delayed-impact' phenomena that once appear in citation history [Redner, 2005].

iii) **Additive intensity model** We model the relation of the first two components with the third one by an *additive* composition rather than multiplication used in [Wang *et al.*, 2013; Shen *et al.*, 2014]. In general these two forms have their respective strengths (see more details in Chapter 4 in [Aalen *et al.*, 2008]), in the analogous context of the multiplicative Cox and the additive Aalen functions [Aalen *et al.*, 2008].

Specifically, the additive model decouples the temporal aging from the triggering effect while the multiplicative couples each other. We simplify the temporal aging term by a general decaying kernel for the difficulty to capture various and unknown citation life-cycle patterns. Our additive mechanism can isolate the adverse effect by this coarse design. Moreover, it is mathematically easier and more efficient to learn the additive model than a multiplicative one [Vu *et al.*, 2011][2].

2.3 Model learning and prediction

The length of time interval between two consecutive citations follows an inhomogeneous Poisson process. Therefore, given that the $(i-1)$th citation arrives at t_{i-1}, the probability that the ith citation arrives at t_i follows

$$p(t_i|t_{i-1}) = \exp\left(-\int_{t_{i-1}}^{t_i} \lambda(t)dt\right)\lambda(t_i)$$

Then the log-likelihood of time-stamped sequence is:

$$\log\prod_{i=1}^{n}\lambda(t_i)\text{exp}\left(-\int_0^T \lambda(s)ds\right) = \sum_{i=1}^{n}\log\lambda(t_i) - \int_0^T \lambda(t)dt,$$

By plugging Eq.1 into the above function and adding sparsity regularization $||\boldsymbol{\beta}||_1$, for $G_d(t) = \int_0^t g_d(t)dt$ we reach:

$$\mathcal{L}_\beta = -\sum_{d=1}^{N}\Bigg\{\sum_{i=1}^{n}\log\Big(\boldsymbol{\beta}\boldsymbol{s}_d\mathrm{e}^{-w_{1d}t} + \sum_{t_j<t_i}\alpha_d g_d(t_i-t_j)\Big) \quad (2)$$
$$-\boldsymbol{\beta}\boldsymbol{s}_d G_d(T) - \sum_{j=1}^{n}\alpha_d G_d(T-t_j)\Bigg\} + \lambda||\boldsymbol{\beta}||_1$$

where $g_d(t) = \mathrm{e}^{-w(t-t_j)}$ is the triggering kernel in Eq.1. Adding ℓ_1 norm renders Eq.2 non-differentiable. We apply the idea of Alternating Direction Method of Multipliers (ADMM) [Boyd *et al.*, 2011] to convert the optimization problem to several sub-problems that are easier to solve. The optimization problem in Eq.2 can be rewritten as the following equivalent form by introducing an auxiliary variable z:

$$\min\mathcal{L} + \lambda||\boldsymbol{z}||_1 \quad s.t.\boldsymbol{\beta} = z. \quad (3)$$

The corresponding augmented Lagrangian of the problem is:

$$\mathcal{L}_\rho = \mathcal{L} + \lambda||\boldsymbol{z}||_1 + \rho\boldsymbol{u}(\boldsymbol{\beta}-\boldsymbol{z}) + \frac{\rho}{2}||\boldsymbol{\beta}-\boldsymbol{z}||_2^2, \quad (4)$$

where $\boldsymbol{u}$ is the scaled dual variables corresponding to the constraint $\boldsymbol{\beta} = \boldsymbol{z}$, and ρ is the penalty parameter, which is usually used as the step size in updating the dual variable. Solving the above augmented Lagrangian problem using ADMM algorithm involves the following sub-problem:

$$\boldsymbol{\beta}^{l+1}, \boldsymbol{\alpha}^{l+1} = \text{argmin}_{\boldsymbol{\beta}\geq 0, \boldsymbol{\alpha}\geq 0}\mathcal{L}_\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z}^l, \boldsymbol{u}^l), \quad (5)$$
$$\boldsymbol{z}^{l+1} = S_{\lambda/\rho}(\boldsymbol{\beta}^{l+1} + \boldsymbol{u}^l), \quad \boldsymbol{u}^{l+1} = \boldsymbol{u}^l + \boldsymbol{\beta}^{l+1} - \boldsymbol{z}^{l+1}$$

$S_{\lambda/\rho}$ is soft thresholding [Donoho and Johnstone, 1995].

To update $\boldsymbol{\beta}$ and α in Eq.5 efficiently, we adopt EM framework to solve this convex problem. The EM step is as follows. Let p_{ki} denotes probability that feature k triggers event t_i and the p_{ij} denotes the probability that event t_i triggers event t_j.

We empirically iterate the expectation step (Eq.6, 7) maximization step (Eq.8, 9) until convergence:

$$p_{ki}^{d\ (l+1)} = \frac{\boldsymbol{\beta}_k\boldsymbol{s}_{dk}\mathrm{e}^{-w_{1d}t_i}}{\boldsymbol{\beta}\boldsymbol{s}_d\mathrm{e}^{-w_{1d}t} + \sum_{t_j<t_i}\alpha_d g_d(t_i-t_j)} \quad (6)$$

$$p_{ij}^{d\ (l+1)} = \frac{\alpha_d g_d(t_i-t_j)}{\boldsymbol{\beta}\boldsymbol{s}_d\mathrm{e}^{-w_{1d}t} + \sum_{t_j<t_i}\alpha_d g_d(t_i-t_j)} \quad (7)$$

$$\boldsymbol{\beta}_k^{(l+1)} = \frac{-B + \sqrt{B^2 + 4\rho\sum_{d=1}^{N}\sum_{i=1}^{n}p_{ki}^d}}{2\rho} \quad (8)$$

$$\alpha_d^{(l+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}p_{ij}^d}{\sum_{i=1}^{n}G_d(T-t_i)} \quad (9)$$

[2] One technical issue is our additive model does not automatically guarantee the non-negativeness of the first term in Eq.1. Thus we normalize the covariates $\mathbf{s}_d$ to [0,1] and make them almost always greater than zero, and β are ensured to be positive according to Eq.8.

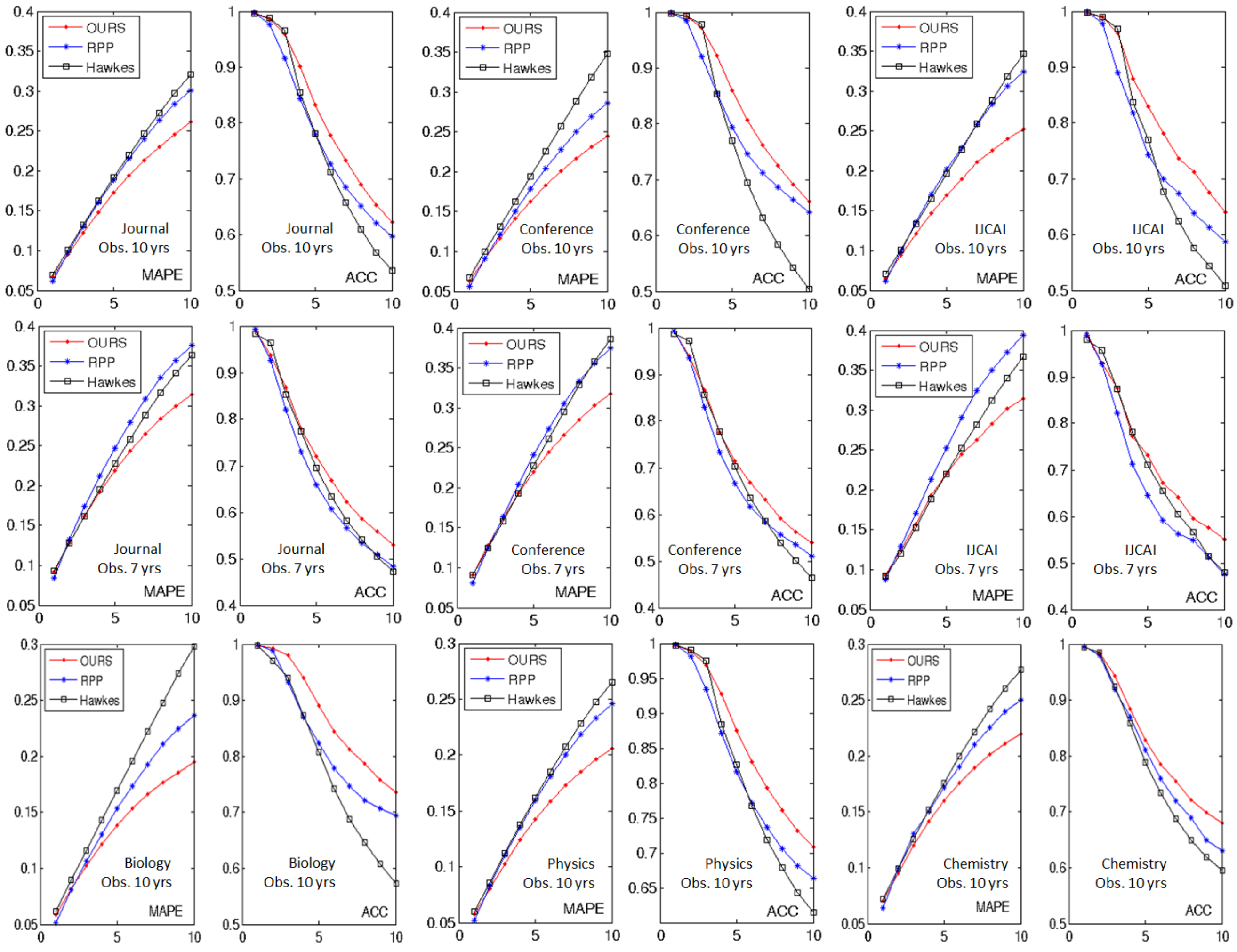

Figure 3: MAPE and accuracy comparison by different observation time windows for training. **Row 1-2**: *Computer Science* each column denotes results of papers in *journal*, *conference*, and *IJCAI* proceedings from 1969 to 1989 and the observation window is 10 years and 7 years in each row respectively. **Row 3**: *Biology*, *Physics*, *Chemistry* with 10-year time window.

where $B = \sum_{d=1}^{N} \boldsymbol{s}_{dk} G_d(T) + \rho(\boldsymbol{u}_k - \boldsymbol{z}_k)$. In fact, in our tests, we always find our method converges to a stationary point though its convergence property is not theoretically proved in the paper.

We update w_1, w_2 for paper d (below d is omitted for notational simplicity) by gradient descent:

$$\frac{\partial \mathcal{L}_\rho}{\partial w_1} = \sum_{i=1}^{n} \frac{\boldsymbol{\beta s} e^{-w_1 t_i}(-t_i)}{\lambda(t_i)} - \frac{\boldsymbol{\beta s} e^{-w_1 T} T w_1 - (1 - e^{-w_1 T})}{{w_1}^2} \tag{10}$$

$$\frac{\partial \mathcal{L}_\rho}{\partial w_2} = \sum_{i=1}^{n} \frac{\sum_{t_j < t_i} \alpha e^{-w_2(t_i - t_j)}(t_j - t_i)}{\lambda(t_i)} \tag{11}$$

$$- \frac{\sum_{i=1}^{n} \alpha e^{-w_2(T - t_i)}(T - t_i) w_2 - (1 - e^{-w_2(T - t_i)})}{{w_2}^2}$$

After learning the parameters, we simulate the Hawkes process by Ogata's thinning algorithm [Ogata, 1981] and estimate the predicted citations before time t, denoted by $c^d(t)$.

3 Experiments and discussion

3.1 Experimental settings

Dataset and compared methods We perform citation count prediction on the real-world dataset: Microsoft Academic Graph [Sinha *et al.*, 2015] of which the papers are well collected, complete and authorized. We select publications in *Computer Science*, which consists of 3,539,403 papers authored by 1,598,575 researchers. Two networks are constructed: the co-author collaboration network with 1,598,575 vertices and the other is citation network with time-stamped directed link, indicating when the citation is received.

We further use papers published during 1969-1989 from the so-called 'main' *Computer Science* venues (refer to http://libra.msra.cn/), including 1,240 journals and 547 conference series, resulting in a total of 47,293 papers. Similar to the protocol in [Wang *et al.*, 2013; Shen *et al.*, 2014], we use papers with more than 5 citations during the first 5 years after publication as training data and predict their citations in the

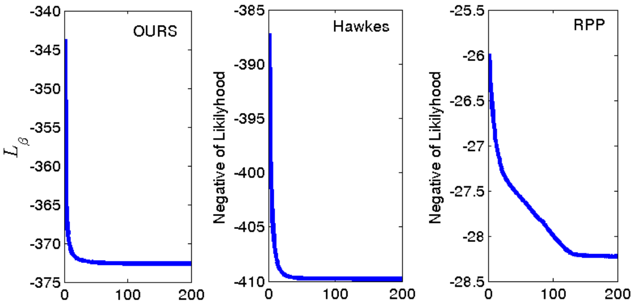

Figure 4: Objective function by iteration.

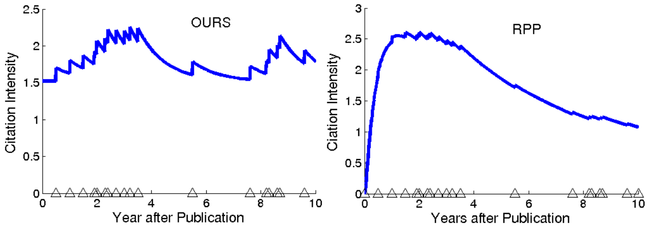

Figure 5: The learned intensity function over time by RPP and ours. Black triangles denote citation events.

next 10 years. Other fields, *Physics*, *Biology* and *Chemistry* are also evaluated. The improved RPP [Shen *et al.*, 2014] based on [Wang *et al.*, 2013], the Hawkes model [Hawkes, 1971] are implemented and tuned to their best performance.

Two metrics used in [Shen *et al.*, 2014] are also used:

Mean Absolute Percentage Error (MAPE) It measures the average deviation between predicted and true popularity over N papers. Denoting with $c^d(t)$ the predicted number of citations for a paper d up to time t and with $r^d(t)$ its real number of citations, MAPE is given by $\frac{1}{N}\sum_{d=1}^{N}\left|\frac{c^d(t)-r^d(t)}{r^d(t)}\right|$.

Accuracy It measures the fraction of papers correctly predicted for a given error tolerance ϵ. Hence the accuracy of popularity prediction on N papers is $\frac{1}{N}\sum_{d=1}^{N}\left|d : \left|\frac{c^d(t)-r^d(t)}{r^d(t)}\right| \leq \epsilon\right|$. [Shen *et al.*, 2014] set $\epsilon = 0.1$ on their dataset. We find in our test, our methods always outperforms regardless ϵ and we set $\epsilon = 0.3$.

3.2 Results and further discussion

MAPE and accuracy They are given in Fig.3 where each column for the first two rows shows the results for *Computer Science* papers published in journal, conference, and IJCAI respectively. Our method (denoted by *OURS*) consistently outperforms across different observation time window (7, 10 years). The third row on *Biology*, *Physics* and *Chemistry* reveal that our method performs robust across fields.

Time complexity The time cost for RPP, Hawkes and our method is 0.355, 0.805, 0.860 seconds per iteration. However, our efficient EM framework renders *OURS* converges with less iterations as illustrated in Fig.4. The total consumption time for RPP, Hawkes and our method is 42.6, 24.1 and 17.2 minutes in average for training the *Computer Science* models. This suggests our method is efficient and accurate.

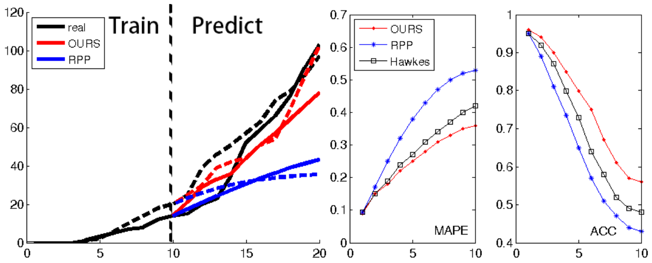

Figure 6: Performance on the 'second-acts' papers. Two papers with solid [Kahn and Roth, 1971] and dashed [Eklundh, 1986] lines are used to exemplify the real and predicted cumulative citation count over years – the first 10 years are used as training window, and the next 10 years are for prediction.

Visual comparison We also compare the learned intensities of RPP and our method. Fig.5 shows the learned intensity distribution and citation events. Event intensity of RPP relies strongly on the time decaying function while ours can flexibly capture the triggering effect of recent citations.

Study on 'second-acts' papers We conduct experiments on the so-called 'second-acts' phenomenon that papers receive plenty of citations in their late-stage of life cycle. This type of papers is also called in *Sleep Beauty* by [Ke *et al.*, 2015]. There is a continuous spectrum of delayed recognition where both the hibernation period and the awakening intensity are taken into account. Fig.6 compares both the real and cumulative predicted citations for sleep beauties type of papers, where the observation window is 10 years. 249 papers are chosen by setting i) less than 20 citations in the first 10 years since publication, and ii) larger than 70 citations in the next 10 years. Note the MAPE is worse than the results in Fig.3 while the Hawkes model and our method performs relatively better. We visualize real and predicted citations of two concrete examples: solid line indicates paper [Kahn and Roth, 1971], and dash line [Eklundh, 1986]. RPP increases linearly regardless of the booming citations in the late stage (around the end of the 10 year observation time window) while our method follows the trend more timely and closely.

Interpretability of covariates By using the sparsity regularization (set $\lambda = 2$ in Eq.1), we can select the most important and interpretable features. Table.1 ranks the covariates by the amplitude of coefficients. The most important factors are author's authority, such as H-index, author rank, and venue's rank, which relate to the novelty of the scientific works.

4 Conclusion

We present an individual paper citation prediction model. Empirical results suggest that its utility for prediction and interpretability. It also gives an independent study on the argument for how effective algorithmic citation prediction approaches can be devised among *Science communication* [Wang *et al.*, 2014b; 2014a]. We empirically find robust methods is achievable for individual paper citation prediction by appropriate modeling in line with [Wang *et al.*, 2014a].

References

[Aalen *et al.*, 2008] O. Aalen, O. Borgan, and H. Gjessing. Survival and event history analysis: A process point of view. In *Springer*, 2008.

[Acuna *et al.*, 2012] D. Acuna, E. Daniel, S. Allesina, and K. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[Barabási and Albert, 1999] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[Boyd *et al.*, 2011] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

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

[Dong *et al.*, 2015] Y. Dong, R. Johnson, and N. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *WSDM*, 2015.

[Donoho and Johnstone, 1995] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

[Egghe, 2006] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[Eklundh, 1986] B. Eklundh. Channel utilization and blocking probability in a cellular mobile telephone system with directed retry. *IEEE Transactions on Communications*, 34(4):329–337, 1986.

[Fuyuno and Cyranoski, 2006] I. Fuyuno and D. Cyranoski. Cash for papers: putting a premium on publication. *Nature*, 441(7095):792–792, 2006.

[Ginther and Kahn, 2004] D. Ginther and S. Kahn. Women in economics: moving up or falling off the academic career ladder? *Journal of Economic perspectives*, 18:193–214, 2004.

[Hawkes and Oakes, 1974] A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 1974.

[Hawkes, 1971] A. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971.

[Hirsch, 2005] J. Hirsch. An index to quantify an individuals scientific research output. *PNAS*, 2005.

[Kahn and Roth, 1971] M. Kahn and B. Roth. The near-minimum-time control of open-loop articulated kinematic chains. *Journal of Dynamic Systems, Measurement, and Control*, 93(3):164–172, 1971.

[Ke *et al.*, 2015] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. Defining and identifying sleeping beauties in science. *PNAS*, 2015.

[Luo *et al.*, 2015] D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *IJCAI*, 2015.

[Merton, 1968] R. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[Ogata, 1981] Y. Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[Pan and Fortunato, 2014] R. Pan and S. Fortunato. Author impact factor: tracking the dynamics of individual scientific impact. *Scientific Reports*, 4, 2014.

[Penner *et al.*, 2013] O. Penner, R. Pan, A. Petersen, K. Kaski, and S. Fortunato. On the predictability of future impact in science. *Scientific Reports*, 3, 2013.

[Petersen *et al.*, 2014] A. Petersen, S. Fortunatoo, R. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. Stanley, and F. Pammolli. Reputation and impact in academic careers. *PNAS*, 111(43):15316–15321, 2014.

[Radicchi *et al.*, 2008] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, 105(45):17268–17272, 2008.

[Redner, 2005] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 2005.

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

[Sinha *et al.*, 2015] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *WWW*, pages 243–246, 2015.

[Stephan, 2012] P. Stephan. *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA, 2012.

[Stern, 2014] D. Stern. High-ranked social science journal articles can be identified from early citation information. *PLoS ONE*, 9(11):0112520, 2014.

[Vu *et al.*, 2011] D. Vu, A. Asuncion, D. Hunter, and P. Smyth. Continuous-time regression models for longitudinal networks. In *NIPS*, 2011.

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

[Wang *et al.*, 2014a] D. Wang, C. Song, H. Shen, and A. Barabási. Response to comment on quantifying long-term scientific impact. *Science*, 345(6193):149–149, 2014.

[Wang *et al.*, 2014b] J. Wang, Y. Mei, and D. Hicks. Comment on "quantifying long-term scientific impact". *Science*, 345(6193):149–149, 2014.

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

[Yan *et al.*, 2012] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *JCDL*, 2012.

[Yan *et al.*, 2013] J. Yan, Y. Wang, K. Zhou, J. Huang, C. Tian, H. Zha, and W. Dong. Towards effective prioritizing water pipe replacement and rehabilitation. In *IJCAI*, 2013.

[Yan *et al.*, 2015] J. Yan, C. Zhang, H. Zha, M. Gong, C. Sun, J. Huang, S. Chu, and X. Yang. On machine learning towards predictive sles pipeline analytics. In *AAAI*, 2015.

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

[Zhao *et al.*, 2015] Q. Zhao, M. Erdogdu, H. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *SIGKDD*, 2015.

[Zhou *et al.*, 2013] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processe. In *AISTATS*, 2013.