4-2017

Predicting the impact of software engineering topics: An empirical study

Santonu SARKAR

Rumana LAKDAWALA

Subhajit DATTA
Singapore Management University, subhajitd@smu.edu.sg

Predicting the Impact of Software Engineering Topics An Empirical Study

Santonu Sarkar
BITS Pilani K K Birla Goa Campus, Goa, India
santonus@acm.org

Rumana Lakdawala
BITS Pilani K K Birla Goa Campus, Goa, India
rumanalakdawala@gmail.com

Subhajit Datta
Singapore University of Technology and Design, Singapore
subhajit.datta@acm.org

ABSTRACT

Predicting the future is hard, more so in active research areas. In this paper, we customize an established model for citation prediction of research papers and apply it on research topics. We argue that research topics, rather than individual publications, have wider relevance in the research ecosystem, for individuals as well as organizations. In this study, topics are extracted from a corpus of software engineering publications covering 55,000+ papers written by more than 70,000 authors across 56 publication venues, over a span of 38 years, using natural language processing techniques. We demonstrate how critical aspects of the original paper-based prediction model are valid for a topic-based approach. Our results indicate the customized model is able to predict citations for many of the topics considered in our study with reasonably high accuracy. Insights from these results indicate the promise of citation of prediction of research topics, and its utility for individual researchers, as well as research groups.

Keywords

software engineering publication, topic model, citation prediction

1. INTRODUCTION

1.1 Background

Across domains, researchers are deeply interested in knowing how long their results will continue to attract attention in the community. Being cited by peer researchers reflects an important aspects of such attention. Since citations are gathered over long periods of time *after* a result is published, it is challenging to predict how important the result will remain in future. This importance prevails across different levels, from the individual to the organizational. Individual researches need to make informed decisions about the research threads they choose to pursue, while research organizations - both academic and industrial - require objective assessments to guide allocation of resources. In this context, Wang et al. have developed a mechanistic model for citation prediction of research papers [18].

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053051

1.2 Context and Approach

While papers are widely regarded as research units, we posit that *topics* represent more reliable quanta of research. As we discuss further in the subsequent sections of this paper, a topic is a collection of papers that are thematically linked; given a corpus of papers, topics are discovered by natural language processing algorithms such as Latent Dirichlet Allocation (LDA). Across the development life cycle of a discipline, individuals as well as institutions are more concerned with how the importance of research topics - rather than individual papers - vary over time. From this perspective, we have customized the paper-based model presented in [18] to be applicable in a topic-based context. We examine key components of the methodology of [18] when applied to topics vis-a-vis papers - growth of topic volume, preferential attachment, and temporal decay to establish the validity of the model for topics. We then apply the model on a large corpus of software engineering publications to extract insights on the predictability of topic citations.

1.3 Research contribution

The research contribution from this study can be summarized as:

- To the best of our knowledge, this is the first study on citation prediction of topics.
- We demonstrate the effective application of a mechanistic model for topics.
- We apply the model on a large corpus of research publications and derive insights with implications for individuals as well as organizations.

1.4 Organization of the paper

The paper has been organized as follows. In Section 2, we summarize the related work in this area. Next, we provide a brief description of the software publication data, as well as an approach to derive a set of topics from the dataset in Section 3. Next, we describe the prediction model construction approach in Section 4. We demonstrate the prediction accuracy of the model in Section 5. We briefly describe the threats to validity of this approach in Section 6. Finally we conclude our paper.

2. RELATED WORK

There has been a growing research interest in modeling and analyzing the dynamics behind scientific research collaboration, publication and its impact based on citation. The pioneering work by [11, 9, 10] on the science of science observed the exponential nature of the publication volume and recognized the importance of citations in the world of scientific publications. A significant number of research publications on citation based analysis of published papers appeared in [13, 14, 4, 1, 15, 18, 19]. The work by [14] is an early report on modeling the aging phenomenon of a published paper. The work by [12, 15] discussed the phenomenon of preferential attachment in scientific papers. One common characteristics of these contribution is that the analysis was mostly done on the publication data from journals related to theoretical sciences. In contrast to the existing approaches we have taken the discipline of software engineering, which is one of the earliest, recognized disciple in computer science. The publication data we have considered comprises of not only journals but conferences, whose longevity is arguably less compared to an archival publication. In one of our earlier work we have shown that the preferential attachment is present for papers the domain of software engineering [6]. However, what differentiates our current work from the existing body of literature is that we have attempted to define a prediction model on a set of topics, which is at a higher level of abstraction than papers. The modeling effort is nontrivial as the topic model derived from papers is a probabilistic one based on the similarity of content of papers. In our earlier work [7, 8], we have shown that the topic model based on the information content of papers work reasonably well in terms of predicting the longevity of the topic in the domain of software engineering. In the current approach, as described later in this paper, we have shown that a simple adaptation of the citation prediction model on papers, proposed by Wang et al. [18] also can be useful in predicting citations that a topic will attract in the near future.

3. DATA DESCRIPTION

In this paper we examine a body of 55,000+ papers collected from over 56 venues authored by $\sim$ 73000 researchers between the years 1975-2013. Our primary source of data is from Microsoft Academic Graph Data[16]. Paper abstracts were extracted from the AMiner[17] database. While the AMiner data contains publication data in computer science, the microsoft academic graph contains data from other branches as well. The data was filtered to match 56 prominent venues in software engineering which account for the papers with appropriate citation data used in this paper. After filtering the papers, we built our corpus comprising of paper published in software engineering venues since its inception (1968) till 2013. The corpus comprises of i) papers ii) authors iii) publication venues iv) paper abstract v) citation data restricted to software engineering domain only. Merging the two massive data sets, was a challenging effort where we linked the publication data from one data set to the other using approximate string matching approach over the paper title. We skip the nitty gritties of the corpus creation process in this paper.

3.1 Topic Modeling

Though ACM classification of computer science topics[1] is being enforced in most of the recent publications, the classification framework has evolved over time. Most of the publications that we are dealing with since 1975 certainly do not follow this classification framework. In view of this we have used the Latent Dirichlet Allocation (LDA) model to identify topics from the collection of 55,000+ papers. LDA has been widely used to model topics from text corpora particularly in the context of research publications. The work by Datta et. al [7, 8] described the process of extracting and finalizing topics from the corpora of papers $\mathcal{P}$, where each paper p is treated as a document comprising of the title and the abstract. In this paper we briefly mention this process for brevity.

The LDA algorithm, based on Blei's original implementation [3], considers each paper $p \in \mathcal{P}$ to be a probability distribution θ_p over a limited number of topics $\Gamma = \{\tau_1 \cdots \tau_k\}$. LDA also generates another distribution of keywords (collected from all the papers in the corpus) over topics. We do not consider this distribution for our analysis in this paper. We only use $\Theta = \{\theta_p | p \in \mathcal{P}\}$ topic-paper mixture model. Next, we use a threshold ψ to discard a paper p belonging to a topic τ when $\theta_p(\tau) < \psi$. Subsequently, for each topic, we denote $\mathcal{P}_\tau$ to be the set of papers that belongs to this topic. Identifying the number of topics is an important part of this process. In our case, we have used a log-likelihood measure to decide the number of topics. In our study, we have generated 60 topics from the software engineering publication data for our work. More details about the process of topic generation may be found in [8].

4. CITATION PREDICTION MODEL

In this paper we have adapted the citation prediction model for papers proposed by Wang et al. [18]. The authors have built the prediction model from the publication data from a few journals. The original prediction model comprises of the following key components:

1. Growth of papers over time (for the journal under consideration): The authors have observed that the publication volume, denoted by $N(t)$ grows exponentially with time.
2. Authors bring the notion of temporal decay, described as the aging factor which captures the notion that novel ideas proposed in a paper are assimilated in subsequent publications as this paper is read and cited over time. This in a sense, "fades" the novelty factor of the original paper and eventually it's citation count decreases. They have also observed that the decay phenomenon can be best modeled by a log-normal survival probability $P(\Delta t_p)$, which essentially computes the probability that a paper p will be cited again in the future after time $\Delta t_p = t - t^p_{pub}$, where t^p_{pub} is the publication year of p. This essentially indicate the survival rate of p's novelty factor to the research community.
3. Alongwith the decay, there is a notion of "preferential attachment" which captures the idea that highly cited papers are more visible and likely to be cited again than less cited paper. The probability that a paper p's will be cited again is proportional to the total number of citations c^t_p that p received previously.

[1] http://www.acm.org/about/class/1998

4. Finally, the model uses a constant called a "fitness" value, for the lack of a better term, λ_p for a paper p that essentially denotes several other intangible factors that can influence the novelty and hence the citation of the paper p, relative to another paper.

The prediction model based on the paper citation combines these components and solves the equations to come up with a final prediction model.

While it is intuitively plausible that such a model is adaptable for topics, the adaptation is not trivial. While citations of a paper is a ground truth, citations for a topic has to be derived. While calculating the aging factor of a paper, there is a clear starting point, which is the year of publication of the paper. However, a topic does not have such a birth date. Furthermore, the original model was applied only for a handful set of journals, whereas we are attempting to fit the model for a domain of software engineering comprising of 50+ journals and conferences, covering almost the entire breadth of the domain.

A research topic in reality can be impacted by various sociological factors, unlike a paper. A research topics can be attractive or "saturated" due to factors that are beyond the realm of any meaningful modeling, such as technological developments or influences from the industry, an increase or the lack of funding, a scientific revolution and so on. These factors can influence researchers to publish or cite papers in a particular research topic. Our current model does not consider such factors.

4.1 Topic Citation

In order to define the prediction model, we first define the notion of topic citations c_τ^t uptil the year t as the weighted sum of the citations of the component papers, the weight being the probability of paper touching upon a given topic. Thus,

$$c_\tau^t = \sum_{p \in \mathcal{P}_\tau} \theta_p(\tau) \times c_p^t \tag{1}$$

where c_p^t is the cumulative citation of the paper p uptil the year t since its publication.

4.2 Growth of topic volume

The growth of scientific publications is exponential is well documented (first by Price et. al [9]) but various groups have also shown the growth of publication volume is exponential within each discipline [4]. We find this to be consistent within our dataset. While [18] observed the exponential growth $N(t) \sim e^{\beta t}$ of paper within a set of journals, we observed the same phenomenon to be true for the entire corpus of software engineering publication where where $\beta = 16^{-1}$ for the software engineering publications.

In order to extend the growth model for topics, we define the growth of volume for a topic τ, to be

$$N_\tau(t) = |\{\ p \mid t_{pub}^p = t \wedge p \in \mathcal{P}_\tau\ \}| \tag{2}$$

1. For each topic, we plot $N_\tau(t)$ vs. t and observed an exponential growth.
2. For each topic, we fit the publication count data on an exponential curve like $ae^{bt} + c$ and obtain parameters a, b, c using least squares.
3. We then use the R^2 as well as Kolmogrov Smirnov (KS) D statistic measure which is used to evaluate the maximum deviation for the fitted data to the empirical citation data.

We thus verify the hypothesis that the growth of volume of papers in each topic is $N_\tau(t) \sim e^{t\beta_\tau}$. We have shown the best, average and the worst fit in Figure 1.

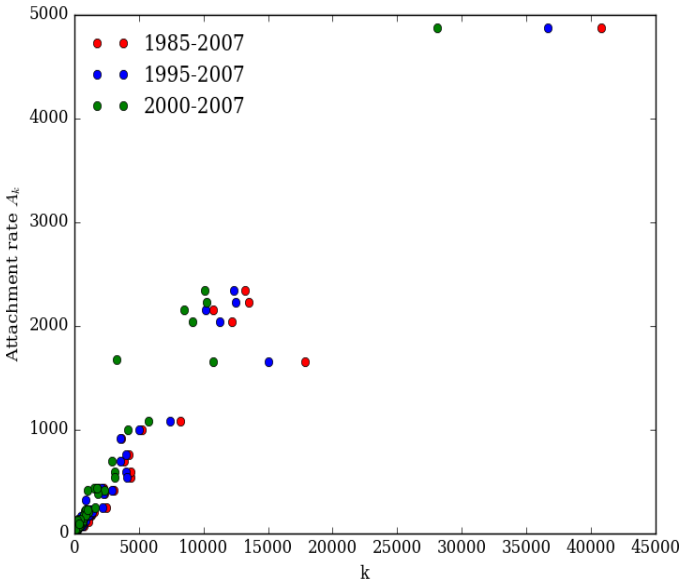


Figure 2: Empirical validation of preferential attachment

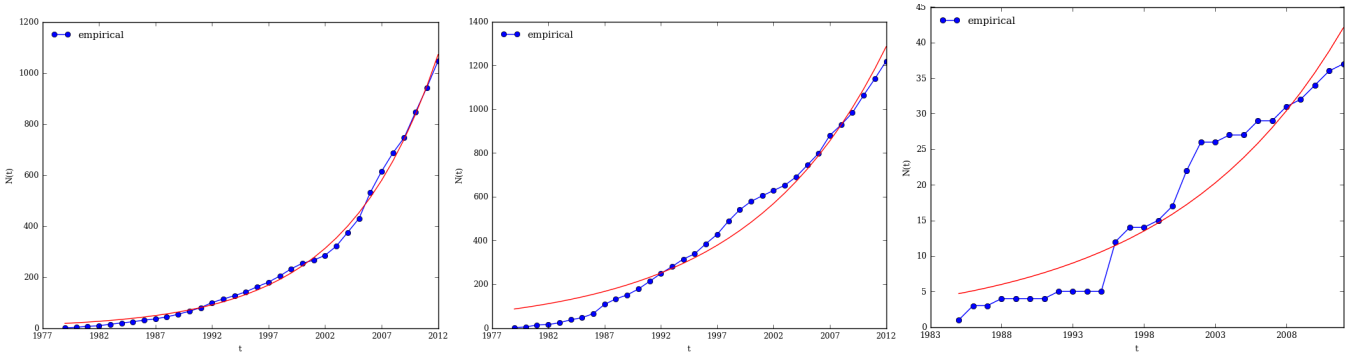
4.3 Preferential Attachment

While the paper citation count has been a standard metric to evaluate the impact (and hence the longevity) of a paper, such a measure is useful for a topic as well, even though there is a subtle difference between the longevity of a paper and a topic. A topic remains prevalent in the research community when researchers not only publish more papers in the topic, but refer to published papers from this topic as well. In the case of a paper, it is quite possible that a seminal paper remains highly cited after a long time since its publication even though the field or a topic does not remain "attractive" to the next generation of researcher. As a result, the papers in the topic do not attract citations over time, but the individual paper remains visible due to its seminal contribution.

For a topic τ, we define the notion its preferential attachment as the probability that a topic τ is cited again in year t is proportional to c_τ^t, where $c_{\tau,t}$ is the citations received by τ *in year* t, and c_τ^t is the cumulative citations received by τ *until year* t as defined earlier.

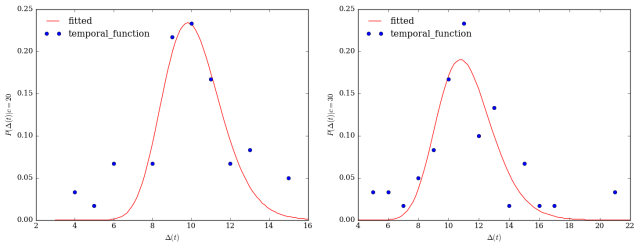
In Figure 2 we document the prevalence of preferential attachment. To verify the preferential attachment, for each topic, we plot a distribution of attachment rate A vs c_τ^t. The notion of attachment rate in the case of a topic is essentially the evolution of citations for that topic. Formally, it gives the likelihood that a topic τ with c_τ^t citations will be cited again in the near future. To compute the attachment rate, we perform the following:

1. We take the slices $T = \{[t_1 \cdot \cdot t_2], [(t_1 + \delta) \cdot \cdot t_2], [(t_1 + 2\delta) \cdot \cdot t_2], \}$
2. For each time slice T_i, for a topic τ, we compute the cumulative citations c^{T_i} obtained by τ during this time. We perform the computation for all the topics.



(a) Best fit (b) Average fit (c) Worst fit

Figure 1: Exponential growth of volume of topics



(a) $c_a = 20$ (b) $c_a = 30$

Figure 3: Temporal Decay graph for topics indicating log-normal distribution

3. We now cluster the topics based on the c^{T_i} ($\pm 5\%$) values. Let this cluster be $G_{k,T_i} = \{\tau_i \cdots \tau_j\}$, where for notational convenience we denote the range c^{T_i} ($\pm 5\%$) to be k.
4. For each k we now compute

$$A_{k,T_i} = \frac{1}{|G_{k,T_i}|} \sum_{\tau \in G_{k,T_i}} c_\tau^{t_2+1} \tag{3}$$

We plot k vs average A_{k,T_i} for each time slice. As shown in Figure 2, the relationship appears linear.

4.4 Temporal Decay

Adapting the approach in [18] proposed for papers, we define the temporal decay of a topic τ to be:

$$P(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma_\tau \Delta t} \exp\left(-\frac{(\ln \Delta t - \mu_\tau)^2}{2\sigma_\tau^2}\right) \tag{4}$$

where μ_τ denotes the time for the topic τ to reach its citation peak and σ_τ denotes the decay rate for the topic τ. To understand our adaption of this decay model, we first briefly describe how this aging function was derived for papers.

1. The authors considered papers for a given journal, published during a fixed time frame of 1950-1960.
2. For a fixed cumulative citation $c_a = 10$, for each paper p whose $t^p_{pub} \in \{1950 - 1960\}$, they computed Δt_p for the year when the paper's citation reaches c_a.
3. The Δt values thus collected for each paper were plotted against the $P(\Delta t \mid c_a)$ distribution to observe a lognormal distribution.

Since our data set of software engineering publications shows a strong prevalence of temporal decay for the papers, we were motivated to adapt and extend the decay model for topics as well. In order to now adapt the model for the topic, it is important to define a publication year for a topic. In the current model we have considered the following approach:

1. We create three time slices $T = \{1975 - 1989, 1989 - 1998, 1999 - 2008\}$ and for each time slice T_i, for each topic τ we calculate a cumulative citation $c_\tau^{T_i}$ as:

$$c_\tau^{T_i} = \sum_{p \in P_\tau \wedge t^p_{pub} \in T_i} \theta_p \times c_p^t \tag{5}$$

2. We compute the $\Delta t_{\tau,T_i}$ for each topic τ in each time slice T_i by taking $\Delta t_{\tau,T_i} = t - t_{T_i[0]}$ where $T_i[0]$ is the start year of each time slice.
3. The $\Delta t_{\tau,T_i}$ are collected for each topic in each time slice and their distribution is plotted ($P(\Delta t_{\tau,T_i} \mid c_a)$ vs $\Delta t_{\tau,T_i}$) to observe a lognormal distribution.
4. We get three plots for each c_a value corresponding to the distribution for each time slice in $T = \{T_1, T_2, T_3\}$
5. For the model construction we repeated this for $c_a = 15, 20, 30$

With the above modification, we observed that even the topics follow a similar decay model. We haves shown the decay model for $c_a = 20$ and $c_a = 30$ in Figure 3.

4.5 Topic's Citation Prediction Model

Following the approach proposed by [18], we combined the temporal decay, preferential attachment and the fitness to obtain the probability that a topic τ is cited again at time t as

$$\Pi_\tau(t) \sim \eta_\tau c_\tau^t P_\tau(t) \tag{6}$$

One can solve the above equation alongwith the equation for growth in volume, to get the following master equation that allows us to predict the cumulative number of citations a topic receives in year t:

$$c_\tau^t = m\left[e^{(\lambda_\tau)\Phi\left(\frac{\ln t - \mu_\tau}{\sigma_\tau}\right)} - 1\right] \text{ where}$$
$$\lambda_\tau = \frac{\beta_\tau \eta_\tau}{D_\tau} \text{ and } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-y^2}{2}} dy \tag{7}$$

The Φ function is the cumulative normal distribution, whereas m, is a constant that indicates the average number of references each new topic contains. The β_τ parameter cap-

Table 1: Summary of Results

	R^2	NRMSE	RMSE	MAPE	KS-stat
Best	1	0.02	188.39	0.01	0.2
Average	0.708	0.151	188.94	0.03	0.25
Median	0.89	0.12	84.135	0.03	0.2
Worst	-1.23	0.53	1211.38	0.09	0.4

Table 2: Prediction Accuracy: Metric Ranges

Range%	R^2	NRMSE	MAPE
75-100	$[0.95-1]$ 17	$[0.02-0.07]$ 17	$[0.01-0.02]$ 14
50-75	$[0.95-0.89]$ 14	$[0.08-0.12]$ 14	$[0.03-0.04]$ 14
25-50	$[0.88-0.72]$ 15	$[0.13-0.19]$ 15	$[0.04-0.09]$ 17
0-25	$[<0.71]$ 11	$[0.19-0.53]$ 11	$[<0.9]$ 15

tures the growth rate of a topic, and D_τ is a normalization constant for a topic. Since we have adapted the approach in [18], we skip the detailed derivation of c_τ^t here. Furthermore, analogous to a paper's fitness, λ_τ represents a topic's fitness which we empirically derive from the papers for each topic τ. In the original model however, the authors considered all the papers published in a particular journal to be their corpus, and hence they treated β and D to be a system-wide constant.

5. EXPERIMENTAL RESULTS

To test our model, we have split each topic's citation history into test and training sets. Therefore for a topic τ since we have computed the c_τ^t for $t = \{1975, \cdots, 2013\}$, we define the training period for each topic as 1975- 2008 (containing 67% papers) and the test period as 2009-2013 (33% papers).

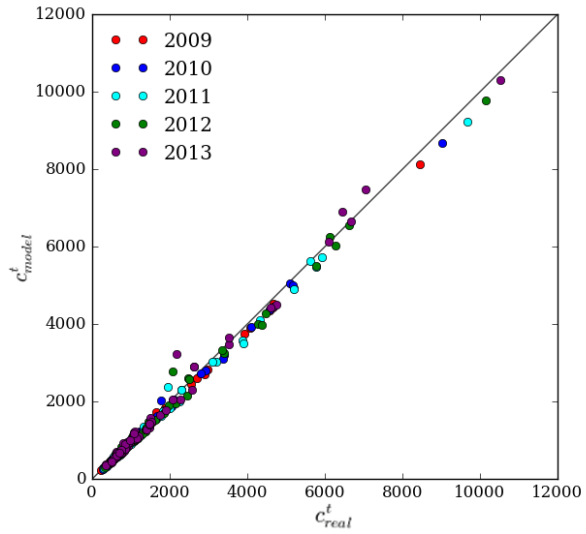
We apply the model for each topic as follows:

1. Using citation training data for topic τ, we obtain unique $\mu_\tau, \sigma_\tau, \lambda_\tau$ by collapsing the citation histories into the equation for c_τ^t by applying non linear least squares fit to the equation over the training citation data points.
2. For each year t in the test phase we obtain c_τ^t as the predicted cumulative citation by plugging parameters $\mu_\tau, \sigma_\tau, \lambda_\tau$ into the equation.
3. We test these predicted citation on various metrics to see that the model indeed works and provides a good prediction.

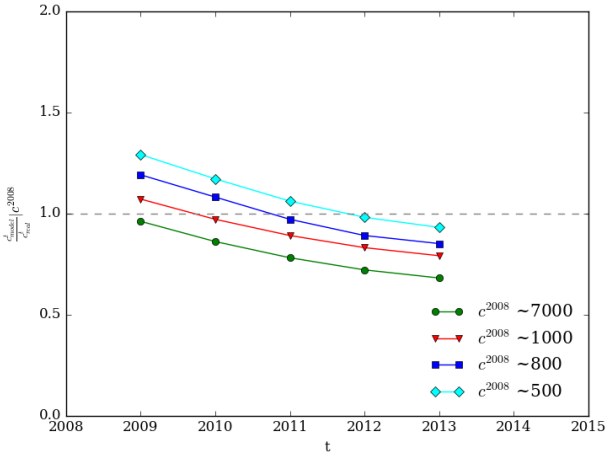
To quantify the model's predictive accuracy we tested its output on various statistical metrics like Normalized Root Mean Square, Coefficient of determination (goodness of fit), MAE and KS test. The summary of the results are shown in Table 1.

Next we show the prediction accuracy over the topics in Table 2. As shown in this table, there are 17 topics, for which the prediction accuracy with respect to R^2 value, lies within $[0.95-1]$ and for 80% of the topics, the R^2 value is > 0.7. However, for a small 5% of the topics, we have observed a negative R^2 value, which is quite possible for a non-linear model [5] like our approach.

For NRMSE, we observe in Table 2 that 17 topics have NRMSE value within $[0.02-0.07]$ and for 80% of the topics, the NRMSE value is < 0.19.



(a) Scatter plot all topics



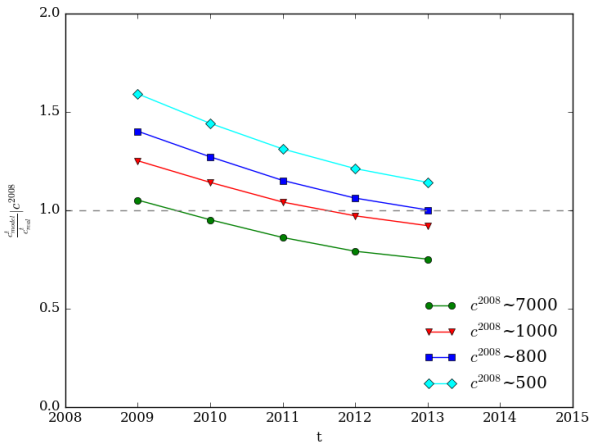
(b) Accuracy as a ratio for group of topics

Figure 4: Actual citations vs predicted citations for all topics

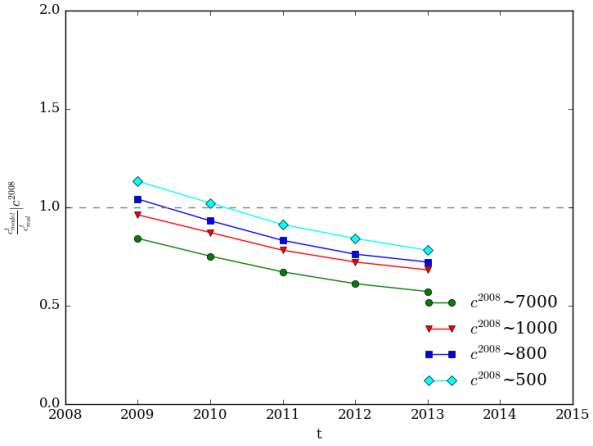
We also applied KS test as shown in the Table 1. We find that 78% topics have a p value > 0.7 confirming our Null Hypothesis that the empirical topic citations follow the same distribution as the citations predicted from Eq 3.t.

We group the topics into high (7000), good (1000), moderate(800) and low(500) categories of citations (the numbers are chosen so as make the group sizes to be roughly equal), where the cumulative citations are computed till the training period of 2008. We then plot a ratio $\frac{c_{model}}{c_{real}}$ of the average citation per category for the prediction years $[2009-2013]$. Figure 4 shows that the trends are close to best prediction accuracy (close to 1).

We also highlight that the prediction accuracy does not depend on the value of m we choose. We have varied the value of m starting with 10 to values close to 100 and observed that the prediction accuracy of the model remains unaffected.



(a) Pref Attachment only



(b) Temporal decay only

Figure 5: Accuracy of alternate models

5.1 Power of the combined model

We also compared the prediction accuracy of the combined model (that combines the growth, the temporal decay and the preferential attachment) with the individual components of the model, namely the preferential attachment and the temporal decay and found that the combine model indeed has the best performance.

5.1.1 *Preferential Attachment Only*

Bianconi et al. [2] proposed a citation prediction model for papers based on preferential attachment only. This model has the same conceptual basis of Wang et al. [18] and our approach. Hence, we considered the preferential attachment based prediction model [2] with suitable modifications for topics as:
$\Pi_\tau(t) \sim \lambda_\tau \times c_\tau^t$ where c_τ^t models the preferential attachment and $c_\tau^t \sim N_\tau^{\lambda_\tau}$ where N is a growth term, computed for each topic.

We observed that this model being exponential, over estimates the citation dynamics compared to the empirical observations as shown in Figure 5a. Though for high citation topics, we observed that this model performs shows better performance than topics with lesser citations.

5.1.2 *Temporal decay Only*

To illustrate that only the temporal decay based model is also not good enough, we tested our data on the lognormal decay based prediction model $c_\tau^t \sim \Phi\left(\frac{\ln t - \mu_\tau}{\sigma_\tau}\right)$

We observed that this model underestimates the citation prediction for topics in the absence of the preferential attachment term as shown in Figure 5b. Specifically, this model underestimates significantly for topics with high citations.

6. THREATS TO VALIDITY

We now highlight the threats to validity of our results. Customizing a model developed for research papers to be effective for research topics has required assumptions and workarounds that we have explained in previous sections. In many cases, we have validated the assumptions on the basis of our data-set. However, inherent difference between papers and topics can contribute to some divergence in the results. So far as the construct validity is concerned, we have adapted established bibliometric measures and shown that the individual components of the model follows the observations of the established techniques. Therefore, our study is free from threats to construct validity. As we have used bibliographic information available in the public domain, threats to internal validity are limited by the veracity of these data-sets. Though we have used a well-known topic modeling technique, the granularity of the topics, can have an impact on the prediction, specifically, the preferential attachment of a topic may be impacted by the topic granularity. Threats to external validity is defined by how much our results can be generalized. Since we have only considered one research domain, our results are not fully generalizable yet. We plan to validate the model with data from other domains in future. Given access to the data, our results can be easily replicated; hence this study is free from threats to reliability.

7. CONCLUSIONS

In this paper we have proposed a model for predicting citations that research topics would receive in the near future, based on a citation prediction model by Wang et al. [18] for published papers. For a topic, we defined the topic citations from its constituent papers using the topic distribution. While the original model was defined with a specific journal in mind, and for papers, we have shown that a simple adaption of the same works well for topics which can be defined as a probability distribution over a set of papers. In order to show that the model works for topics, we have shown that all of the trends observed, i.e. i) aging of a paper follows a lognormal distribution, ii) pref attachment exists in paper citations,and iii) paper volume grows exponentially for papers, also apply to the topic distribution.

We have built and verified that topic citations prediction model from a corpus of academic publications in the software engineering domain since its inception till 2013, having 55000+ papers, spanning across 56 venues. Since our empirical model is based on topics, which itself is a derived data, as opposed to a paper, the generation of topics from papers can have an impact on the prediction model. As a part of the future work, we intend to study whether varying the number of topics can have any impact on the accuracy of the prediction model. As a part of the future work, we would like to extend the model for other disciplines of computer science, and observe how the longevity of the topics

can be predicted from the citation prediction approach for other disciplines of computer science. It would be interesting to compare and contrast the accuracy of predictions of various topics extracted from publications in other computer science disciplines.

8. REFERENCES

[1] U. B, M. S, S. M, and J. B. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[2] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters (EPL)*, 54(4):436–442, 2001.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, J.-P. Bianconi, Ginestra andBouchaud, and A.-L. Barabási. Dynamics of ranking processes in complex systems. *Physical Review Letters*, 109(12), 2012.

[5] A. C. Cameron and F. A. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, apr 1997.

[6] S. Datta, N. Kumar, and S. Sarkar. The social network of software engineering research. In *ISEC Š12: Proceedings of the 5th India Software Engineering Conference*, pages 61–70. ACM, 2012.

[7] S. Datta, S. Sarkar, A. Sajeev, and N. Kumar. Discovering the rise and fall of software engineering ideas from scholarly publication data. In *Proceedings of the 24th International Conference on World Wide Web - WWW'15 Companion*. Association for Computing Machinery (ACM), 2015.

[8] S. Datta, S. Sarkar, and A. S. M. Sajeev. How long will this live? discovering the lifespans of software engineering ideas. *IEEE Transactions on Big Data*, 2(2):124–137, jun 2016.

[9] D. de Solla Price. *Little Science, Big Science... and Beyond*. Columbia Univ Pr, 1963.

[10] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[11] G. E. Citation indexes for science: A new dimension in documentation through association of ideas. *Sci*, 122(3159):108–111, 1955.

[12] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PLoS ONE*, 6(9):e24926, sep 2011.

[13] W. Glanzel. Towards a model for diachronous and synchronous citation analyses. *Scientometrics*, 60(3):511–522, 2004.

[14] H. KB and S. P. Modelling aging characteristics in citation networks. *Physica A*, 368(2):575–582, 2006.

[15] W. M, Y. G, and Y. D. Measuring the preferential attachment mechanism in citation network. *Physica A*, 387(18):4692–4698, 2008.

[16] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web - WWW'15 Companion*, pages 243–246. ACM, 2015.

[17] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

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

[19] J. Wang. Citation time window choice for research impact evaluation. *Scientometrics*, 94(3):851–872, 2013.