Old is Not Always Gold: Early Identification of Milestone Patents Employing Network Flow Metrics

Manajit Chakraborty
Faculty of Informatics
Università della Svizzera italiana
Lugano, Switzerland
chakrm@usi.ch

Fabio Crestani
Faculty of Informatics
Università della Svizzera italiana
Lugano, Switzerland
fabio.crestani@usi.ch

Abstract

Bibliometrics has been employed previously with patents for technological forecasting. The primary challenge that technological forecasting faces is early-stage identification of technologies with the potential to have a significant impact on the socio-economic landscape. With this aim, we carry out an exploratory study using various network-based metrics on patent citation network to identify patents which are possible candidates for major influence in the immediate future. To effectively uncover these patents shortly after they are issued, we need to go beyond raw citation counts and take into account both the citation network topology and temporal information. We posit that, as with scholarly citations, not all patent citations carry equal importance with age. This information is captured by dynamic network flow metrics that take the effect of time on the citations into account. Identifying top patents can aid in re-ranking of search results in a patent search. We carry out our experiments on two standard collections of patents and present some insightful results and observations based on rigorous analysis.

1 Introduction

Patent citations, namely references to prior patent documents and the state-of-the-art included therein, and their frequency are also often used as indicators for the technological and commercial value of a patent and to identify "key" patents, which often varies depending on the nature of the technology. Previous research has already endorsed technological forecasting[1] as an integral element to stay ahead of the curve for corporations and governments (Campbell, 1983). Acs et al. (2002) suggested that patents provide a fairly reliable measure of innovative activity. Identifying important patents, observing their change of importance as captured by the variation of citation measures and analyzing them can lend us new insights as to how innovation evolves over a period of time. This could be beneficial for innovators and companies who are actively involved in producing patents. It would facilitate them to take stock of the innovation quotient of a particular technological area and help measure its growth and potential over a certain period of time.

In this paper, we aim to identify influential patents from different technological areas from patent citation network using network flow algorithms. Identifying top patents from any particular category can help companies interested in patenting to glean an overview of the important innovations in their field of concern. It can also benefit governments in deciding various policies such as funding to technological areas that have shown promise over the last few years. We argue that while citation count may help us identify important patents, it tends to favour patents that have been filed or granted long ago, thus providing it a longer citation accumulation period. While PageRank helps to mitigate the situation to a certain extent by considering the whole network instead of simple citation count, PageRank too has been known to be biased against recent network nodes. CiteRank (Walker et al., 2007) introduces exponential penalization of old nodes, thus modelling the node score such that it captures the future citation count gain. However,

[1] https://hbr.org/1967/03/technological-forecasting

due to CiteRank's known limitations, we propose a new model called Time-Attentive Ranking, which helps to capture the temporal changes and their effect on certain nodes. We carry out our experiments on two different datasets to determine the efficacy and effectiveness of our method against baselines both qualitatively and quantitatively. We then carry out a comparison of the top-N ranked list of patents provided by three algorithms using Rank Biased Overlap (Webber et al., 2010) and against a list of significant patents by Strumsky and Lobo (2015), to point out the relative changes. We posit that top-ranked patents or the ranking criteria for the same could be employed for a ranking based patent retrieval method as have been exploited by Xue and Croft (2009) and Liao and Veeramachaneni (2010). Our experiments are two-pronged – first, we study the effect of the network metrics on European patents from the MAREC dataset and secondly, we employ an adapted version of a deep learning model that infuses both textual content and network flow metrics on USPTO patents in order to spot influential patents and validate our hypothesis.

2 Related Works

The notion of quantitative evaluation of scientific and technological impact builds on the basic idea that the scientific merits of papers (Radicchi et al., 2008), scholars (Egghe, 2006), journals (Bollen et al., 2006), universities (Molinari and Molinari, 2008) and countries (Cimini et al., 2014) can be gauged by metrics based on the received citations. Bibliometrics has been employed in a variety of scenarios to measure and analyze citations since they provide a rich source of information. Scientific papers and scholarly articles have been investigated using various bibliometric tools, especially citations for a long period (Narin et al., 1976; Bakkalbasi et al., 2006). One of the early studies to measure the technological impact based on patent citations was done by Karki (1997). He proposed a host of technological indicators based on citations among patents.

Carpenter et al. (1981) and Fontana et al. (2013) compared patents associated with inventions that received a prize and patents from a control group, finding again evidence that "important" patents are more cited (the mean number of citations received was found to be about 50% higher for important patents). As argued by (Jaffe et al., 2000), citations reflect the fact that either a new technology builds on an existing one or that they serve a similar purpose. As a result, chains of citations allow us to trace the technological evolution, and hence patent centrality in the citation network can be used to score patents. In our preliminary citation analysis, we have adopted a couple of PageRank based approaches along with other citation metrics. PageRank (Bedau et al., 2011; Bruck et al., 2016) and similar eigenvector-based metrics (Doira and Banerjee, 2015) has been computed on patent citation networks earlier. Mariani et al. (2016) argued on similar lines in the case of scholarly articles and proposed a re-scaled version of PageRank that discounts citations for old papers based on age. We build upon this notion and perform a thorough analysis of patent citation network in sub-categories and sectors and in the presence (or absence) of patent content by employing a proposed network flow algorithm.

3 Methodology

We employ three different network-based patent-level metrics for comparison: PageRank scores P, CiteRank score C and our proposed Time-Attentive Rank score T.

3.1 PageRank

PageRank (Brin and Page, 1998) is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the PageRank of E. PageRank normalizes the number of links on a document by not counting each of them as equal. PageRank can be defined as follows (Equation 1):

$$P_i^{n+1} = \alpha . \sum_{j:k_j^{out}>0} A_{ij} \frac{p_j^n}{k_j^{out}} + \alpha . \sum_{j:k_j^{out}=0} \frac{p_j^n}{N} + \frac{1-\alpha}{N} \tag{1}$$

where $k^o_j ut = \sum_l A_{lj}$ is the outdegree of node j, α is the teleportation parameter, and n is the iteration number. The PageRank score P_i of node i can be interpreted as the average fraction of time spent on node i by a random walker who with probability α follows the network's links and with probability $1 - \alpha$ teleports to a random node. We consider $\alpha = 0.5$ throughout this paper since it is the accepted choice for citation networks (Chen et al., 2007).

3.2 CiteRank

CiteRank (Walker et al., 2007) was designed specifically for ranking papers in a citation network. CiteRank performs a random walk on an aggregated citation graph but initiates the walk from a recent paper chosen with the probability that depends on its age. Authors estimated parameters of the random walk by fitting papers' CiteRank score to the number of citations accrued by papers over some time period. Let us suppose M is a transfer matrix with elements $M_{ij} = 1/L_j$ if paper j cites i and 0 otherwise. The probability that a researcher follows the citation links to encounter a paper is defined as in Equation 2:

$$\vec{C} = I_0.\vec{\rho} + (1-\alpha)M.\vec{\rho} + (1-\alpha)^2 M^2.\vec{\rho} + \ldots \quad (2)$$

where I_0 is an identity matrix, ρ_i =exp$^{-age_i/\tau}$ is the probability of initially selecting paper i, age_i is the age of the paper and τ is the characteristic decay time. In this paper, we consider $\alpha = 0.5$ and $\tau = 2.6$ years, as specified by Walker et al. (2007).

3.3 Time-Attentive Rank

Our proposed model Time-Attentive Ranking is based on the notion that 'An inventor or patentee can find patents by following citations links back in time from a particular patent'. The number of paths that can be attenuated between patent p_i and p_j can be expressed as a *contagion matrix* M given by Equation 3:

$$M_{N,\alpha} = \sum_{i=1}^{N-1} \alpha^i A^i{}_N \quad (3)$$

where A_n is the adjacency matrix of patents citing each other for a particular year t_n and α is the probability of following a citation link. The more paths there are from patent p_i to p_j, the higher the likelihood that an inventor will find p_j by following citation chains from p_i, which is similar to α-Centrality (Bonacich, 1987) and Katz centrality (Katz, 1953) metrics. Since the existing contagion matrix does not account for time and hence weights each edge equally, the authors propose a *retained adjacency matrix* which is given by Equation 4:

$$R_{n,\gamma}(i,j) = \begin{cases} \gamma^{N-n_i}, & \text{if } p_i \text{cites } p_j \text{and} \\ & t(p_i) = t_{n_i} \leq t_n \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $\gamma < 1$ is the retention probability given to attach more weight to a recent patent and decrease the weight as the patent keeps ageing. The contagion matrix can then be written as Equation 5 (using Equation 3 and 4):

$$EM_{N,\alpha,\gamma} = \sum_{i=1}^{N-1} \alpha^i R^i{}_N \quad (5)$$

and hence the score of a patent p_j at the end of a time period $[t_i, t_N]$ is given by $EM_N(j) = \sum_i EM_N(i,j)$. For our experiments we consider the best possible settings by empirically setting $\alpha = 0.1$ and $\gamma = 0.3$. Elsewhere in the paper we refer to the EM score as $\mathcal{T}$ for uniformity and ease of comprehension.

4 Experimental Setup

4.1 Datasets

For this study, we used two datasets: (1) European Patent (EP) collection from the MAtrixware REsearch Collection (MAREC)[2] and (2) US patents dataset collected by Kogan et al. (2017) that spans the period between 01-01-1926 and 11-02-2010. To the best of our knowledge, there exists no study of a similar kind on the European Patents, which is why we chose to work with the former collection from MAREC. However, this presents a unique challenge of finding a respective gold-standard list of "milestone" patents, which is not available. Hence for this collection, we resort to a qualitative evaluation as described in later sections. To compare our proposed approach's performance against the state-of-the-art and perform a quantitative evaluation, we repeated our experiments on the USPTO dataset as well. Additionally, we also performed validation of our model's performance by a deep learning technique as suggested in Chung and Sohn (2020) to identify a patent's grade in determining in value.

4.2 Preprocessing

MAREC collection We only considered granted patents from the 'EP' collection. For uniformity,

[2] http://www.ifs.tuwien.ac.at/imp/marec.shtml

we removed patents that had some metadata missing, such as classification codes or patent citations. We also did not consider the non-patent citations since they are out of the scope of our study. We pre-processed the data to only keep the citations between patents that were issued within 1976-2008, removing thereby the citations to patents issued before 05-1976. Hence, we were left with a network consisting of only EP-EP patent citations formed out of 251,664 patents having 350,164 citations.

USPTO collection Unlike the well-known NBER patent data, the dataset provided by Kogan et al. (2017) has a vastly improved coverage. We pre-processed the data to only keep the citations between patents that were issued within the time period of 01-01-1926 and 11-02-2010, removing thereby the citations to patents issued before 01-01-1926. The resulting citation network analyzed in this paper is composed of 6,237,625 patents and 45,962,301 citations between them.

5 Results and Analysis

5.1 MAREC Patents

In this section, we present a qualitative of the results followed by a comprehensive analysis of the same. We ran the experiments on networks that were spliced in time (yearly) over a ten year period (1998-2008). The same set of experiments were carried out on networks comprising of all patents, patents belonging to a certain sector and patents belonging to a certain category. On studying the in-degree and out-degree of patents, we observed that the degrees are very skewed, *i.e.*, only a handful of patents gets a large number of citations while most of the patents in the network have less than ten citations. Hence, this network follows a similar pattern to that of a scholarly article citation network (albeit with more skewness). Hence, the network-flow algorithms that can be employed with paper citation networks can also be adapted here. Moreover, there is a strong correlation between the in-degree and out-degree of the patents in both collections, which implies that highly-cited patents tend to be cited by other highly-cited patents and to cite other highly-cited patents (Ren et al., 2018).

Qualitative Comparison of Top Patents : For an intuitive understanding of how the different network-flow metric scores affect the rank, it is important to observe the top-ranked patents according to the PageRank score P, CiteRank score C and Time-Attentive Rank score T. As mentioned earlier, each patent is endorsed with several classification codes that classify them into sectors, categories, sub-categories *etc*. The highest level of classification is according to sectors (A-H). Each sector consists of several categories (A61K, A61P, ...), each category consists of several sub-categories and so on. A single patent can belong to several sectors and categories.

Table 1: Top-5 Patents by citation count

PatentID	No. of Citations
EP0037691	125
EP0272189	121
EP1049021	121
EP0364618	121
EP0527247	121

5.1.1 Complete Network

From Table 2, we can observe significant changes in the ranking of the patents. The tables reveal that there are more recent patents granted after the year 2000 in the top-10 list ranked by Time-Attentive Ranking than that produced by either PageRank or CiteRank. To be precise, the Time-Attentive ranking method includes five patents granted after the year 2000 in the top-10 list, while for PageRank and CiteRank, it is four out of ten for both. Of course, the difference is even more pronounced as we go deeper in the lists, say, top-15, top-20 and so on, which we could not present here due to space constraints.

For comparison, the list of top five patents on the basis of citation count is presented in Table 1. One can observe that none of the three metrics ranks the patents from Table 1 in their top five list. In fact, within the top fifteen results, only patent *EP0272189* and *EP0364618* feature in the lists compiled according to PageRank and CiteRank scores, while patent *EP0527247* is listed by PageRank only. The rest do not find a place in the top-10 of any network-based score list. This corroborates our initial hypothesis that is simply acquiring a high citation count does not indicate the importance of a patent.

5.1.2 Network for a particular sector

Next, we perform the same set of experiments over individual sectors of patents. Similar to the trend shown by the complete network, for sector B which has the highest number of patents, we observe from Table 3 that Time-Attentive Ranking features the

Table 2: Top-10 patents for 2008 ranked by scores

	Rank	PatID	Title	Date	#Citations
PageRank score P	1	EP0251752	Aluminum-stabilized ceria catalyst compositions and method of making same.	29-06-1987	111
	2	EP1304455	Particulate filter for purifying exhaust gases of internal combustion engines	17-10-2002	103
	3	EP0728435	Cyclone dust extractor	20-02-1996	87
	4	EP1031939	COMPOSITE IC CARD	16-11-1998	69
	5	EP1261147	A method and system for simultaneous bi-directional wireless communication between a user station and first and second base stations	21-05-2001	78
	6	EP0776864	Process for the aerobic biological purification of water	10-07-1996	34
	7	EP1466940	Carbon fiber composite material and process for producing the same	13-04-2004	57
	8	EP1400858	PHOTORESIST STRIPPER COMPOSITION	21-06-2002	28
	9	EP1059092	Use of complexes among cationic liposomes and polydeoxyribonucleotides as medicaments	08-06-1999	19
	10	EP0534904	Imidazolylmethyl-pyridines.	21-09-1992	23
CiteRank score C	1	EP0251752	Aluminum-stabilized ceria catalyst compositions and method of making same.	29-06-1987	111
	2	EP1304455	Particulate filter for purifying exhaust gases of internal combustion engines	17-10-2002	103
	3	EP0728435	Cyclone dust extractor	20-02-1996	87
	4	EP1031939	COMPOSITE IC CARD	16-11-1998	69
	5	EP1261147	A method and system for simultaneous bi-directional wireless communication between a user station and first and second base stations	21-05-2001	78
	6	EP0776864	Process for the aerobic biological purification of water	10-07-1996	34
	7	EP1466940	Carbon fiber composite material and process for producing the same	13-04-2004	57
	8	EP1400858	PHOTORESIST STRIPPER COMPOSITION	21-06-2002	28
	9	EP1059092	Use of complexes among cationic liposomes and polydeoxyribonucleotides as medicaments	08-06-1999	19
	10	EP0534904	Imidazolylmethyl-pyridines.	21-09-1992	23
TimeAttentiveRank score T	1	EP0251752	Aluminum-stabilized ceria catalyst compositions and method of making same.	29-06-1987	111
	2	EP1304455	Particulate filter for purifying exhaust gases of internal combustion engines	17-10-2002	103
	3	EP0728435	Cyclone dust extractor	20-02-1996	87
	4	EP1031939	COMPOSITE IC CARD	16-11-1998	69
	5	EP1835243	Evaporator with electronic circuit printed on a first side plate	26-02-2007	21
	6	EP1261147	A method and system for simultaneous bi-directional wireless communication between a user station and first and second base stations	21-05-2001	78
	7	EP1466940	Carbon fiber composite material and process for producing the same	13-04-2004	57
	8	EP0364618	Multiple signal transmission device.	18-10-1988	121
	9	EP0776864	Process for the aerobic biological purification of water	10-07-1996	57
	10	EP1400858	PHOTORESIST STRIPPER COMPOSITION	21-06-2002	28

Table 3: Sector B patents of 2008 ranked

	Rank	Patent ID	Date
PageRank P	1	EP0728435	20-02-1996
	2	EP0008860	20-07-1979
	3	EP0095603	07-05-1983
	4	EP1142619	26-09-2000
	5	EP0466535	18-06-1991
CiteRank C	1	EP0728435	20-02-1996
	2	EP1304455	17-10-2002
	3	EP1142619	26-09-2000
	4	EP0534904	21-09-1992
	5	EP1731327	10-06-2005
TimeAttentiveRank T	1	EP0728435	20-02-1996
	2	EP1329412	10-10-2000
	3	EP1489033	05-06-2004
	4	EP1306147	23-10-2002
	5	EP1674419	21-12-2005

Table 4: Category A61K patents of 2008 ranked

	Rank	Patent ID	Date
PageRank P	1	EP0776864	10-07-1996
	2	EP0728435	20-02-1996
	3	EP0071564	19-07-1982
	4	EP0002210	17-11-1978
	5	EP0447285	27-02-1991
CiteRank C	1	EP0776864	10-07-1996
	2	EP0728435	20-02-1996
	3	EP1835243	26-02-2007
	4	EP1568666	22-02-2005
	5	EP0770375	13-09-1996
TimeAttentiveRank T	1	EP0776864	10-07-1996
	2	EP0527247	08-08-1991
	3	EP0364618	18-10-1988
	4	EP0272189	17-12-1987
	5	EP0728435	20-02-1996

more recent patents in their top five as compared to their counterparts.

5.1.3 Network for a particular category

While it is interesting to study the complete network and find the most influential patents as identified by Time-Attentive Rank, it does not deliver us

Table 5: RBO@20 for 2008

	P	**C**	**T**
P	–		
C	0.4981	–	
T	0.3921	0.5741	–

Table 6: RBO@20 for A61K

	P	**C**	**T**
P	–		
C	0.3430	–	
T	0.2568	0.3963	–

Table 7: RBO among the full ranked lists

	P	**C**	**T**
P	–		
C	0.8548	–	
T	0.6307	0.7270	–

a lot of information. On the other hand if we limit the patent citation network by categories, it could provide us some insights as to which technologies have been gaining momentum in the last few years of the patent data. The total number of categories in the patent database exceeds hundred. Not surprisingly, the distribution of patents against categories is also skewed. For brevity, we present only the results for the most popular category A61K.

From Table 4, we observe a certain peculiarity. None of the top five patents ranked by the Time-Attentive Ranking mechanism is a post-2000 patent. This is interesting because it implies that while Time-Attentive rank gives more weightage to recent citations, it does not bias towards recent patents, thus maintaining a balance between older and newer patents. So, the top-ranked patent in all three cases is the same indicating that *EP0776864* is indeed the most important patent in category A61K.

Metric for comparison of ranked lists: Since our hypothesis hinges on the ranking of patents over a network metric based score, it is imperative that the lists generated by PageRank and CiteRank and TimeAttentiveRank will be different in their ordering of elements (ranks). As the lists are quite long, their scores are not directly comparable, and for a given depth d the two lists may not even have the same set of elements, we will have to resort to indefinite ranking (Webber et al., 2010). To this end, we employ *rank-biased overlap* (RBO) to measure the similarity and agreement between the two lists. The RBO values for the year 2008 compared over the complete list of ranked results is presented in Table 7. The Rank-Biased Overlap is defined as in Equation 6.

$$\mathrm{RBO}(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} . A^{d} \qquad (6)$$

where S and T are two indefinite ranked lists. p stands for *user's persistence*, which determines how steep is the decline in weights: the smaller p, the more top-weighted is the metric. A_d, *agreeement* can be defined as the proportion of S and T that are overlapped at depth d. Rank-biased Overlap falls in the range $[0, 1]$, where 0 means disjoint, and 1 means identical. While RBO is the agreement score between two indefinite lists, we are more concerned with the top-k elements in the lists and hence RBO@k provides us a better measure to compare the top-ranked elements. It is imperative to note that RBO $>$ RBO@k. For our case, we empirically consider $k = 20$ and $p = 0.9$.

Tables 5, 6 and 7 present the RBO confusion matrix. We can clearly observe a pattern here. The overlap between CiteRank and TimeAttentive ranked lists are certainly more than the overlap (agreement) between PageRank and TimeAttentive Rank, which confirms our intuition that recent patents receive more preference in the weighted citation measures rather than unweighted citations of PageRank.

5.2 USPTO Patents

For this collection, we adopt a different approach for carrying out our experiments. The experiments on the MAREC patents were solely based on network flow metrics, which we could not assess quantitatively due to the lack of a standard baseline. Instead, for the US patents collections, we compare our approach against the state-of-the-art Re-scaled PageRank method proposed by Mariani et al. (2019) to identify milestone patents. As a second objective, we wanted to determine the value added by textual content in determining a patent's worth. This objective stems from similar studies on patents where it was shown that exploiting the multimodal nature of patents yields better prediction performance (Chakraborty et al., 2020). For this purpose, we adapt the deep learning approach proposed by Chung and Sohn (2020). Due to the incompatibility of NLP based approach proposed by Chung and Sohn (2020) and network flow metrics-based approaches such as the one by Mariani et al. (2019) and ourselves in this paper, we only adopt the deep learning approach (DEP-net) to determine a patent's grade, which is another measure of patent's importance. As per Chung and Sohn (2020), a patent's quality is assigned one of three grades (A, B, or C) based on the average number of forward citations per year. The deep learning approach is briefly summarised below:

- A patent grade (A, B, or C) is assigned based on a threshold determined by the average forward citations accrued per year by the patent.
- Textual content (abstract and claim) from the patent data is extracted along with several other indices such as number of claims, number of inventors, number of backward citations, number of IPCs, *etc*.
- Abstract and claims are transformed (vectorized) into word embeddings as matrices.
- A deep neural network composed of Bi-LSTM layer is added to the CNN structure using multiple filters that fuses the four components (abstract, claims, indices, network-metric score) as input to train and evaluate a patent's quality. Finally, we also evaluate the patent quality for test data.

It is to be noted that we add an extra component to the original model proposed by Chung and Sohn (2020), *i.e.*, the network-metric score. Both the rescaled PageRank score (R) and the Time-Attentive Rank score are fed separately as inputs to the deep neural model. To simplify things, we retained the parametric setting of the neural model as proposed by Chung and Sohn (2020). Finally, the features from the abstracts, claims, indices and network-flow metrics are fused and used as inputs to the fully connected layer. The loss function was cross-entropy, and the activation function was softmax. We label this model as DEP-netPlus (as we add value to the DEP-net model).

5.2.1 Expert-selected historically significant patents

Strumsky and Lobo (2015) listed 175 patents carefully selected "on the basis of consultation with engineers, scientists, historians, textbooks, magazine articles, and internet searches". The patents in the list "all started technological pathways which affected society, individuals and the economy in a historically significant manner" (Strumsky and Lobo, 2015). These significant patents thus provide a good "ground-truth" set of patents that can be used to discern the ability of different metrics to uncover these significant patents. The complete list of these patents can be found in Appendix C of (Strumsky and Lobo, 2015). Presence in the list of significant patents by Strumsky and Lobo is a binary variable: a patent is either in the list or not. We can therefore study the ability of the metrics to rank these outstanding patents as high as possible, in agreement with the objectives of this paper. While there are 175 significant patents in the Strumsky-Lobo list, we restrict our analysis to those patents that were issued within our dataset's temporal span and remove those that are absent in our dataset. This leaves us with M = 112 significant patents.

5.2.2 Comparison against baselines

In this section we inspect the top-ranked patents. For simplicity, we focus on the top-10 patents as ranked by PageRank P and Re-scaled PageRank R and our Time-AttentiveRank T scores (Table 8). From Table 8, we can observe that the top-10 patents by Re-scaled PageRank span a wider temporal range (1942–2010) than the top-10 by PageRank (1942–1996), which is a direct consequence of the age-bias removal. The same temporal span is retained by the Time-Attentive Rank as well. However, it is noteworthy that our proposed method can pick more (3) patents from Strumsky-Lobo's list of significant patents. Among the ten top-ranked patents, two are from 2010 (the last year in the dataset) and received only one citation. This happens because only a few among the most recent patents received citations, which results in temporal windows with a large fraction of patents with zero citations. Thus, within such a temporal window, a patent can achieve large T score thanks to one single citation. A possible solution for this issue is to only include the patents whose temporal windows contain a certain minimal number of incoming citations. Another observation is that both the Re-scaled PageRank and Time-Attentive rank do not necessarily rank patent with grade A in a higher position, so the ranking is not solely dependent on the citation count but also on the network structure.

5.2.3 Performance comparison against DEP-net

To illustrate the importance of including network-flow based metric as a component, we performed the patent grade classification as described in (Chung and Sohn, 2020). We used the same dataset

Table 8: Top-10 patents ranked by Network-metric scores. Asterisks mark the Strumsky-Lobo significant patents.

	Rank	PatID	Title	Date	#Citations	Grade
PageRank score P	1	4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	28-07-1987	1956	A
	2	4683202	Process for amplifying nucleic acid sequences	28-07-1987	2169	A
	3	4237224	(*) Process for producing biologically functional molecular chimeras	02-12-1980	285	B
	4	4395486	Method for the direct analysis of sickle cell anemia	26-07-1983	71	B
	5	4723129	Bubble jet recording method and apparatus in which a heating element generates bubbles in a liquid flow path to project droplets	02-02-1988	1962	A
	6	3813316	Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof	28-05-1974	16	C
	7	5536637	Method of screening for cDNA encoding novel secreted mammalian proteins in yeast	16-06-1996	422	A
	8	4558413	Software version management system	10-12-1985	1956	A
	9	4358535	Specific DNA probes in diagnostic microbiology	09-11-1982	436	A
	10	2297691	SElectrophotography	06-10-1942	588	B
Re-scaled PageRank score R	1	7764447	Optical element holding device, lens barrel, exposing device, and device producing method	27-7-2010	1	C
	2	4237224	(*) Process for producing biologically functional molecular chimeras	02-12-1980	285	B
	3	2297691	Electrophotography	06-10-1942	588	B
	4	7749477	Carbon nanotube arrays	06-07-2010	1	C
	5	7784029	Network service for modularly constructing a software defined radio	24-08-2010	1	C
	6	5536637	Method of screening for cDNA encoding novel secreted mammalian proteins in yeast	16-07-1996	422	A
	7	4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	28-07-1987	1956	A
	8	5523520	Mutant dwarfism gene of petunia	04-06-1996	1139	A
	9	4395486	Method for the direct analysis of sickle cell anaemia	26-07-1983	71	B
	10	4683202	Process for amplifying nucleic acid sequences	28-07-1987	2169	A
TimeAttentiveRank score T	1	4683202	Process for amplifying nucleic acid sequences	28-07-1987	2169	A
	2	4237224	(*) Process for producing biologically functional molecular chimeras	02-12-1980	285	B
	3	2297691	Electrophotography	06-10-1942	588	B
	4	D268584	(*) Personal computer	12-04-1983	3	C
	5	7749477	Carbon nanotube arrays	06-07-2010	1	C
	6	7784029	Network service for modularly constructing a software defined radio	24-08-2010	1	C
	7	5536637	Method of screening for cDNA encoding novel secreted mammalian proteins in yeast	16-07-1996	422	A
	8	5225539	(*) Using recombinant DNA to produce an altered antibody	06-07-1993	549	A
	9	4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	28-07-1987	1956	A
	10	4395486	Method for the direct analysis of sickle cell anemia	26-07-1983	71	B

Table 9: Performance matrix for DEP-netPlus. Best results are marked in bold.

	DEP-net			DEP-netPlus		
Measure	A grade (%)	B grade (%)	C grade (%)	A grade (%)	B grade (%)	C grade (%)
Precision	78.00	51.48	74.85	**79.03**	**52.14**	**75.01**
Recall	75.53	46.65	73.22	74.67	45.98	**73.30**
F-measure	76.74	48.95	74.03	**76.84**	**49.06**	**74.15**

of 296,933 USPTO patents pertaining to "semiconductor" technology collected within the temporal span of 2000 to 2015. We carried out the same preprocessing steps along with down-sampling of the data or certain classes to maintain uniformity. The results of experiments performed with an additional component, *i.e.*, our proposed TimeAttentiveRank score to the deep learning model which we refer to as DEP-netPlus are presented in Table 9.

From the table, we can clearly observe that the classification model is enhanced by the inclusion of a network flow metric that account for the network effect due to citations. This also confirms the superiority of our model in capturing not only the "importance" of a patent but also in evaluating the patent's grade.

6 Conclusion and Future Work

In this paper, we proposed a method to proactively identify milestone patents that have been granted in recent years. We compared the performance of three network-flow algorithms for this purpose on two different datasets. On the second dataset, we used a deep-learning-based approach to fuse patent content along with network flow metrics, to compare against state-of-the-art and discovered that our proposed approach results in better performance both in identifying "milestone" patents as well as improving the patent grade prediction. From the experimental results, we summarily concluded that raw citation count is not enough to capture the *importance* of a patent since it does not take into account the age of citations. When accounted for the same using a balanced metric like Time-Attentive ranking, we are guaranteed to identify potential patents that are likely to spur technological growth in the near future. We also identified top patents per category and sector, which can help in the identification of niche areas for innovation. Although patent retrieval is a recall-oriented task, these criteria may also help in re-ranking the results against a keyword search for patents.

As part of our future work, we would like to study the importance of geographical location on influential patents, such as the country they originated from, the citations received from other countries and so on. We also plan to experiment with the various granularity of time such as a month, year, 5-year period, and so on.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was partially supported by The Global Structure for Knowledge Networks project grant (No. 167326) under the SNSF National Research Programme 75 "Big Data" (NRP 75).

References

Zoltan J Acs, Luc Anselin, and Attila Varga. 2002. Patents and innovation counts as measures of regional production of new knowledge. *Research Policy*, 31(7):1069 – 1085.

Nisa Bakkalbasi, Kathleen Bauer, Janis Glover, and Lei Wang. 2006. Three options for citation tracking: Google scholar, scopus and web of science. *Biomedical Digital Libraries*, 3(1):7.

Mark A Bedau, Andrew Buchanan, Devin W Chalmers, C Cooper Francis, Norman H Packard, and Noah M Pepper. 2011. Evidence in the patent record for the evolution of technology using citation and pagerank statistics. In *ECAL*, pages 77–84. Citeseer.

Johan Bollen, Marko A. Rodriquez, and Herbert Van de Sompel. 2006. Journal status. *Scientometrics*, 69(3):669–687.

Phillip Bonacich. 1987. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107 – 117. Proceedings of the Seventh International World Wide Web Conference.

Péter Bruck, István Réthy, Judit Szente, Jan Tobochnik, and Péter Érdi. 2016. Recognition of emerging technology trends: class-selective study of citations in the u.s. patent citation network. *Scientometrics*, 107(3):1465–1475.

Richard S. Campbell. 1983. Patent trends as a technological forecasting tool. *World Patent Information*, 5(3):137 – 143.

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

Manajit Chakraborty, Seyed Ali Bahrainian, and Fabio Crestani. 2020. Forecasting patent growth by combining time-series signals using covariance patterns. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020*, volume 2621 of *CEUR Workshop Proceedings*.

P. Chen, H. Xie, S. Maslov, and S. Redner. 2007. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8 – 15.

Park Chung and So Young Sohn. 2020. Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change*, 158:120146.

Giulio Cimini, Andrea Gabrielli, and Francesco Sylos Labini. 2014. The scientific competitiveness of nations. *PloS one*, 9(12):e113470.

Rafael A. Corre Doira and Preeta M. Banerjee. 2015. Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity. *Research Policy*, 44(2):508 – 521.

Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics*, 69:131–152.

Roberto Fontana, Alessandro Nuvolari, Hiroshi Shimizu, and Andrea Vezzulli. 2013. Reassessing patent propensity: Evidence from a dataset of r&d awards, 1977-2004. *Research Policy*, 42(10):1780 – 1792.

Adam B Jaffe, Manuel Trajtenberg, and Michael S Fogarty. 2000. The meaning of patent citations: Report on the nber/case-western reserve survey of patentees. Technical report, National bureau of economic research.

M.M.S. Karki. 1997. Patent citation analysis: A policy analysis tool. *World Patent Information*, 19(4):269 – 272.

Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. 2017. Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics*, 132(2):665–712.

Wenhui Liao and Sriharsha Veeramachaneni. 2010. Unsupervised learning for reranking-based patent retrieval. In *Proceedings of the 3rd International Workshop on Patent Information Retrieval*, PaIR '10, pages 23–26, New York, NY, USA. ACM.

Manuel Sebastian Mariani, Matúš Medo, and François Lafond. 2019. Early identification of important patents: Design and validation of citation network metrics. *Technological Forecasting and Social Change*, 146:644–654.

Manuel Sebastian Mariani, Matúš Medo, and Yi-Cheng Zhang. 2016. Identification of milestone papers through time-balanced network centrality. *J. Informetrics*, 10(4):1207–1223.

Jean-Francois Molinari and Alain Molinari. 2008. A new methodology for ranking scientific institutions. *Scientometrics*, 75(1):163–174.

Francis Narin et al. 1976. *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Computer Horizons Cherry Hill, NJ.

Filippo Radicchi, Santo Fortunato, and Claudio Castellano. 2008. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272.

Zhuo-Ming Ren, Manuel Sebastian Mariani, Yi-Cheng Zhang, and Matúš Medo. 2018. Randomizing growing networks with a time-respecting null model. *Physical Review E*, 97(5):052311.

Deborah Strumsky and José Lobo. 2015. Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8):1445–1461.

Dylan Walker, Huafeng Xie, Koon-Kiu Yan, and Sergei Maslov. 2007. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38.

Xiaobing Xue and W. Bruce Croft. 2009. Automatic query generation for patent search. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 2037–2040, New York, NY, USA. ACM.