EARLY IDENTIFICATION OF IMPORTANT PATENTS THROUGH NETWORK CENTRALITY

Manuel Sebastian Mariani, Matúš Medo, François Lafond

Early identification of important patents through network centrality

Manuel Sebastian Mariani[a,b], Matúš Medo[a,c,b], François Lafond[d,e]

[a]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, PR China*
[b]*Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland*
[c]*Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, 3010 Bern, Switzerland*
[d]*Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX2 6ED, UK*
[e]*Smith School for Enterprise and the Environment, University of Oxford, Oxford OX1 3QY*

Abstract

One of the most challenging problems in technological forecasting is to identify as early as possible those technologies that have the potential to lead to radical changes in our society. In this paper, we use the US patent citation network (1926-2010) to test our ability to early identify a list of historically significant patents through citation network analysis. We show that in order 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. In particular, an age-normalized measure of patent centrality, called rescaled PageRank, allows us to identify the significant patents earlier than citation count and PageRank score. In addition, we find that while high-impact patents tend to rely on other high-impact patents in a similar way as scientific papers, the patents' citation dynamics is significantly slower than that of papers, which makes the early identification of significant patents more challenging than that of significant papers.

Keywords: Patent analysis, Citation networks, Significant patents, Technological forecasting, PageRank

1. Introduction

While many inventions are granted a patent, only a small fraction of them represent "important" technological advances or will have a significant impact on the market. As a result, a key problem in technological forecasting is to detect which patents are important as early as possible. The literature has designed various indicators of patent importance based on patent data analysis, and it has been found quite consistently (see Section 2) that at least on average, important patents tend to receive more citations. However, this relationship is typically noisy, which suggests that more sophisticated metrics could outperform simple citation count in identifying important patents. Importantly, it takes time for a patent to accumulate citations, which implies that simply counting the number of citations received by a patent may be effective for uncovering old important patents, but not to detect important patents shortly after they are granted.

In this paper, we propose a network-based metric that identifies important patents better and earlier than citation count. Our metric, time-rescaled PageRank, was introduced by Mariani et al (2016) to identify expert-selected important papers in physics. It is built on Google's PageRank algorithm (Brin and Page, 1998) by requiring that node score is not biased by node age. This metric is computationally efficient and thus can be applied on very large datasets (Vaccario et al, 2017). Here we validate this metric on the US patent citation network (1926-2010),

by evaluating its ability to detect the expert-selected "important" patents from Strumsky and Lobo (2015).

We find that Google's PageRank outperforms raw citation count in identifying the important patents, which supports the idea that important patents tend to be cited by other important patents. This idea is further supported by the strong assortative degree-degree correlations observed in the network (Fig. 1 below): highly-cited patents are typically cited by other highly-cited patents significantly more than what we would expect by chance; at the same time, highly-cited patents tend to cite other highly-cited patents. However, both PageRank and citation count are biased towards old patents; removing this bias is crucial to compare young and old patents on the same scale (Mariani et al, 2016).

To demonstrate the usefulness of removing the age bias in the context of technological forecasting, we evaluate the metrics' performance in identifying the important patents shortly after they are issued, and find that time-rescaled PageRank significantly outperforms citation count and original PageRank in the first 10 years after issuing, approximately. Finally, we use a time-respecting network null model (Ren et al, 2017) to generate randomized networks where the individual patents' citation count dynamics is the same as in the original network. We find that both the observed degree-degree correlations and the performance advantage of PageRank-related metrics over citation count cannot be found in the randomized networks, which indicates that these properties emerge as a result of network effects that go beyond patents' neighborhood.

Our findings demonstrate that in order to timely identify the significant patents, both network topology and time informa-

Email address: `manuel.mariani@unifr.ch` (Manuel Sebastian Mariani)

tion play an essential role. In more general terms, the advantage of PageRank-related metrics over citation counting metrics, together with the strong degree-degree correlations of the patent citation network, support the hypothesis that significant technological advances "lean on the shoulders of giants" in a similar way as scientific advances (Bornmann et al, 2010). Yet, we find that the citation dynamics of scientific papers and patents are characterized by substantially different timescales. As a result, because patents (on average) take more time than papers to accumulate citations, the early identification of significant patents is more challenging than that of significant papers.

2. Related work

Broadly speaking, our work is related to those studying the relation between popularity metrics and significance in creative works such as scientific papers (Mariani et al, 2016; Comins and Leydesdorff, 2017), movies (Spitz and Horvát, 2014; Wasserman et al, 2015; Ren et al, 2017), or music albums (Monechi et al, 2017).

In the context of patent analysis, it is well known that patents are of extremely different quality (Silverberg and Verspagen, 2007). While a direct measure of patent value is unavailable, patent data are very rich and there have been many attempts at providing indicators of patent value or novelty based on data contained in patent documents, such as the number of claims, the number and type of technology categories, the size of the patent family, and renewal fees, to give just major examples. By far the most widely used patent impact indicator is the number of citations received, and many studies have established a correlation between patent citations and patent value. For instance, Trajtenberg (1990) found that to understand the evolution of the social value generated by the CT scan industry, it was better to count citations received by patents rather than simply counting patents. Albert et al (1991) asked experts to rate the technical impact of patents in the area of silver halide technology, and found that highly cited patents received higher ratings. Harhoff et al (1999), Jaffe et al (2000a) and Harhoff et al (2003), using survey data, found that citations were correlated with the value reported by the inventors. Lanjouw and Schankerman (2004) collected several indicators of patent quality and concluded that citations and the number of claims were the most important indicators of quality. Recently, Zhang et al (2017) proposed to weight 11 indicators of patent value using the Shannon entropy, and selected forward citations as one of the most important indicators for technological value. Hall et al (2005) found that firm market value (Tobin's Q ratio) was correlated to the citation-weighted patent portfolio of the firms. 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).

But in spite of the repeated evidence of the positive relationship between citations received and different indicators of value or quality, it is often acknowledged that this relationship is very noisy (Harhoff et al, 1999), thus leaving open the possibility

that more elaborated indicators could outperform simple citations count in predicting patent value. Here we address two basic (and well-known) problems of evaluation by simply counting citations: when evaluating a given patent's score, it fails to take into account the importance of the citing patents (Narin et al, 1976); and it fails to correct for the fact that young but potentially important patents did not have the time to accumulate a high number of citations.

The basic motivation for using citations received as an indicator of quality is that citations indicate some form of knowledge spillovers. As argued by Jaffe et al (2000b), 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 the patents. But not all measures of centrality are appropriate. For instance, in the case of patents, we want to value positively how many citations are received, but not necessarily how many citations are made.

Whether the references made by a given patent can be used to infer the patent's importance is a delicate issue. In principle, one could argue that a patent with many references has high potential, because it draws from many existing inventions. But an opposite argument could be made as well, because a patent with many references makes it also (legally) clear that its claims are somewhat limited by the claims of the cited patents – in that sense, references indicate a limitation of novelty. It is not yet well-understood which of these two arguments is the most appropriate, and the empirical evidence so far is inconclusive (Jaffe and de Rassenfosse, 2017); here, we will consider that citations received are a weaker signal of importance when they come from patents that make a lot of references.

Based upon the aforementioned considerations, Google's PageRank centrality (Brin and Page, 1998) is especially suited for identifying important patents for three reasons: (i) It takes into account how many citations are received by a patent, (ii) It takes into account how many citations are received by the citing patents, and (iii) it takes into account that citations from patents that have many references are less indicative of the cited patent's quality.

We are not the first ones to suggest that PageRank (Lukach and Lukach, 2007; Bedau et al, 2011; Shaffer, 2011; Dechezleprêtre et al, 2014; Bruck et al, 2016) and similar eigenvector-based metrics (Corredoira and Banerjee, 2015) can be computed on patent citation networks to identify important patents. However, robust evidence that PageRank is more effective than citation count in identifying the key patents is still lacking. In addition, both citation counts and PageRank fail to take into account the dynamic, evolving nature of the citation network. Because the patent system grows with time, older patents tend to have more citations simply because they have been there for a longer time and, on top of that, the preferential attachment mechanism (Valverde et al, 2007) further magnified their advantage. This problem has been long acknowledged and the usual solution is either to limit citation counts to a fixed citation time span, such as the first five years after issuing (e.g., Lanjouw and Schankerman (2004)), or to control for the grant year

in regressions (e.g., Kogan et al (2017)).

Here, we propose an alternative approach, put forward recently by Mariani et al (2016) in the context of scientific publications, which can be applied equally well to citation counts and other centrality metrics, and produces a single score without (or with dramatically reduced) age bias.

Our work complements other efforts to identify important items using citation networks. For instance, Comins and Leydesdorff (2017) report that Reference Publication Year Spectroscopy, a method that looks at the temporal distribution of cited references, is able to identify the biomedical research milestones listed by experts. In the patent literature, Castaldi et al (2015) proposed to identify "superstar" patents as those in the extreme right tail of the citation count distribution, where a power law behavior was observed. Another popular approach, main path analysis, was introduced in the bibliometric literature by Hummon and Dereian (1989), and further developed and applied to patents by Verspagen (2007). In the spirit of the betweenness centrality, it seeks to extract important nodes and edges based on how often geodesic paths pass through them, thus revealing continuity or disruption in technological trajectories. This aspect was exemplified by Martinelli (2012) for the telecommunication switching industry, and by Epicoco (2013) for the green chemistry sector. Triulzi et al (2017) measured patent centrality using a normalized version of centrality metrics, and found that technological domains with central patents also tend to have faster technological improvement rates (a separately measured indicator of progress in technological performance). Finally, as a last example of this rich literature, Martinelli and Nomaler (2014) proposed to measure knowledge persistence by giving higher value to patents that are cited by patents that dont cite many patents - an idea that we will use here too, as PageRank normalizes the received citations by the outdegree of the citing nodes.

In this work, we focus on comparing PageRank with citation counts, and age-rescaled metrics with non-rescaled metrics. This allows us to evaluate whether network-based metrics outperform raw citations counts, and to determine over which range of time the rescaling procedure allows us to better identify the significant patents. In addition, because our analysis follows closely the study of milestone physics papers by Mariani et al (2016), we are able to evaluate the similarities and differences between the scholarly and patent citation data. We find that patents take much longer than papers to receive citations, which makes it harder to identify important patents early on.

3. Data

3.1. The U.S. patent citation network

We analyzed the US patents dataset collected by Kogan et al (2017) that spans the period between 01-01-1926 and 11-02-2010. As compared to the well-known NBER patent data, this dataset has a vastly improved coverage. We pre-processed the data to only keep the citations between patents that were issued within this temporal period, removing thereby the citations to patents issued before 01-01-1926. The resulting citation network analyzed in this paper is composed of $N = 6,237,625$ nodes (patents) and $E = 45,962,301$ directed edges (citations) between them.

In this dataset, the in- and out-degree distribution of the US patent citation network are in agreement with previous findings (Valverde et al, 2007; Csárdi et al, 2007; Silverberg and Verspagen, 2007): the two distributions are relatively broad and span more than three orders of magnitude.

In previous works, Mariani et al (2016) found that PageRank-related metrics outperform citation-counting metrics in identifying significant nodes in a scientific paper citation network, whereas the same does not happen in a movie citation network (Ren et al, 2017). Additionally, Ren et al (2017) found remarkably different degree-degree correlations for the two networks: the papers' citation network is strongly assortative, whereas the movies' citation network is basically uncorrelated. This observation led Ren et al (2017) to suggest that the relative performance of PageRank-related and citation-counting metrics may be related to the network correlation patterns: when the network is uncorrelated, PageRank and indegree bring basically the same information (Fortunato et al, 2008); when there are significant structural correlations, PageRank brings additional information that may be valuable to improve ranking performance.

Fig. 1 shows that the US patent network exhibits strong degree-degree correlations[1]: highly-cited patents tend to be cited by other highly-cited patents, and to cite other highly-cited patents. This assortative pattern cannot be explained by a null model that preserves the temporal evolution of node degree ((Ren et al, 2017), see Section 5.4 for details), which suggests that it is a genuine network effect.

In agreement with similar findings for scientific papers (Ren et al, 2017; Bornmann et al, 2010), Fig. 1A suggests that high-impact patents are able to inspire other high-impact patents more than expected by chance, whereas low-impact patents tend to be cited by other low-impact patents; at the same time (Fig. 1B), high-impact patents rely on other high-impact patents more heavily than expected by chance. Following Ren et al (2017), the strong degree-degree correlations in the patent citation network opens the door to the possibility that metrics that take higher-order network effects into account outperform simple citation counts.

3.2. Expert-selected historically significant patents

In a recent work, 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,

---

[1]The assortativity plot used here is arguably the simplest method to visualize network structural correlations, as simply represents the average (in- or out-)degree of nodes' neighbors as a function of node (in- or out-)degree. Since node centrality is related to incoming connections, we focus here on the average indegree of citing and cited nodes as a function of node indegree.
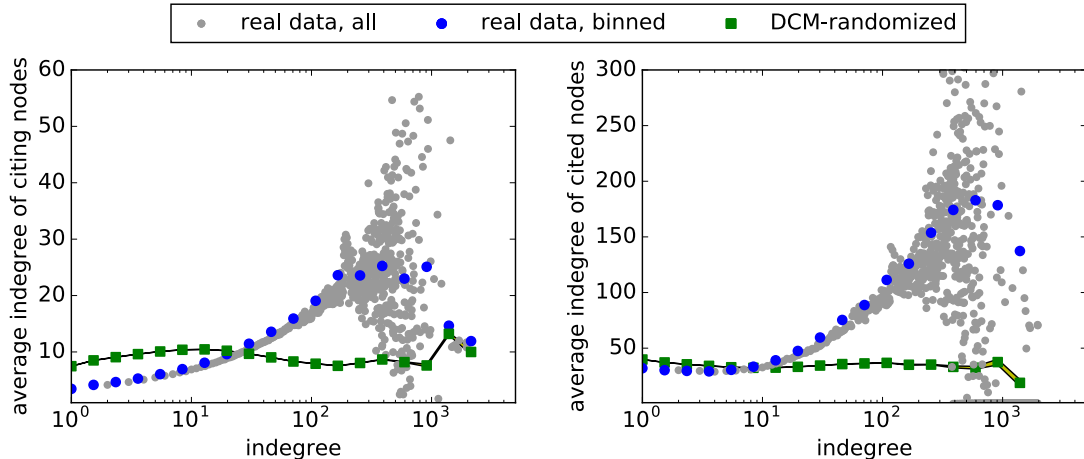
Figure 1: Degree-degree correlations in the US patents' citation network. The gray circles represent the observed average neighbors' indegree for all the indegree values; the blue circles represent the same information in a histogram with equal bin length on logarithmic scale; the green squares represent the mean behavior observed within ten realizations of the dynamic configuration model (see Section 5.4 for a description of the model); the shadows around the line that connect the green squares represent one standard deviation around the mean.

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 the significant patents. The complete list of significant patents can be found in Appendix C of Strumsky and Lobo (2015); the list is quite heterogeneous and comprises patents ranging from simple objects that are part of our everyday life (like the toothbrush and the post-it note) to more sophisticated inventions (like the Game Boy and the Desk Jet printer).

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 main goals 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 the design patents which are absent in our dataset. This leaves us with $M_0 = 112$ significant patents.

4. Methods

In this section, we define the metrics used to identify important patents, and the indicators of performance that we use to evaluate them. Many network centrality metrics (Lü et al, 2016; Liao et al, 2017) and bibliometric indicators (Waltman, 2016) have been devised in the literature. Here, we narrow our focus to four metrics (see Table 1 for a summary): citation count c, PageRank score p, (age-)rescaled citation count $R(c)$ and (age-)rescaled PageRank $R(p)$. Differently from citation count, PageRank score takes the whole network structure into account and weights citations differently according to the centrality of the citing nodes. Rescaled citation count and rescaled PageRank score are obtained from citation count and PageRank score, by explicitly requiring that node score is not biased
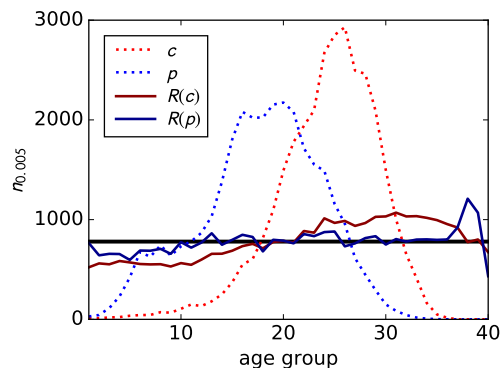

Figure 2: Bias by node age of the rankings by the metrics. Patents are divided by their age into 40 equally-sized groups; the bias by patent age is represented by the number $n_{0.005}$ of patents from each age group in the top $f = 0.5\%$ of the overall patent ranking. The black horizontal line represents the expected unbiased value $n_{0.005}^{(0)} = 0.005\,N/40$. Results for different (small) values of f are qualitatively similar.

by node age (see details below).

4.1. Static patent-level metrics

Citation count, c. The citation count of a given patent is simply the total number of citations the patent has received so far. In terms of the patent citation network's adjacency matrix A (A_{ij} is equal to one if patent j cites patent i, zero otherwise), the citation count c_i of patent i is defined as $c_i = \sum_j A_{ij}$; c_i is referred to as the node i's indegree in the network science language (Newman, 2010). Ranking the patents by citation count assumes that *a patent is important if it is cited by many other patents.*

The ranking by citation count is strongly biased by node age in our dataset. To visualize and quantify this bias, we divide the N patents into 40 equally-sized age-groups based on age. We

4

	Static	Age-rescaled
Citation-counting	**Citation count**, c A patent is important if it is cited by many other patents	**Rescaled citation count**, $R(c)$ Built on citation count by requiring that patent score is not biased by node age
Network-based	**PageRank score**, p A patent is important if it is cited by other important patents	**Rescaled PageRank score**, $R(p)$ Built on PageRank score by requiring that patent score is not biased by node age

Table 1: Metrics considered in this paper together and their main assumptions.

then count how many patents from each age group are in the top-f fraction of the patent ranking by c. For an ideal unbiased ranking, for each age group, we would expect $n_f^{(0)} = f N/40$ patents in the top-f fraction, with small deviations. The result is strikingly different for citation count (see Fig. 2) which underestimates both the oldest and the most recent patents in the dataset. While the bias against recent patents is expected (they have had less time to accumulate citations), the bias against older patents is more surprising, and it can be due to a variety of factors such a higher propensity to cite patents available electronically, a prevalence of patents in technological domains for which less citations tend to be made, patent citation patterns changing with time – for a discussion of these and other reasons for citation bias, see Jaffe and de Rassenfosse (2017). To counteract this bias, we use a simple normalization procedure, described in paragraph 4.2.

PageRank, p. Google's PageRank is a node ranking algorithm introduced by Brin and Page (1998) with the original goal to rank websites in the World Wide Web. Since then, the algorithm has found applications in a broad range of real systems (Gleich, 2015; Liao et al, 2017). The PageRank score p_i of node i is defined through the equation (Berkhin, 2005)

$$p_i = \alpha \sum_{j:k^{out}>0} \frac{A_{ij}}{k_j^{out}} p_j + \alpha \sum_{j:k^{out}=0} \frac{p_j}{N} + \frac{1-\alpha}{N} \quad (1)$$

where $k_j^{out} = \sum_l A_{lj}$ is the number of references made by patent j (k_i^{out} referred to as the node i's *outdegree* in the network science language) and the term $(1 - \alpha)/N$ represents the so-called "teleportation term" (Berkhin, 2005; Gleich, 2015). The algorithm is built on the thesis that *a node is important if it is cited by other important nodes* (Franceschet, 2011): the score of a given patent i depends linearly on the scores of the patents that cited patent i. We set $\alpha = 0.5$ which is the common choice in citation networks (Chen et al, 2007; Walker et al, 2007; Bruck et al, 2016).

In practice, the vector of PageRank scores can be obtained from Eq. (1) by the power iteration method. Starting from a uniform score vector $p_i^{(0)} = 1/N \,\forall i$, we iteratively update the scores according to the equation (Berkhin, 2005)

$$p_i^{(n+1)} = \alpha \sum_{j:k^{out}>0} \frac{A_{ij}}{k_j^{out}} p_j^{(n)} + \alpha \sum_{j:k^{out}=0} \frac{p_j^{(n)}}{N} + \frac{1-\alpha}{N}. \quad (2)$$

Note that the previous equation is the master equation of a two-fold stochastic process on the network where at each step n, a random walker either performs a jump along the network edges (with probability α), or "teleports" to a randomly chosen node in the network (with probability $1 - \alpha$). The PageRank vector of scores $\mathbf{p} = \{p_i\}$ can therefore be interpreted as the stationary state of this Markov process. We halt the iterations when

$$\sum_i \left| p_i^{(n)} - p_i^{(n-1)} \right| < \epsilon, \quad (3)$$

where we set $\epsilon = 10^{-9}$. This procedure guarantees convergence after a number of iterations smaller than $\log \alpha / \log \epsilon$, independently of N (Berkhin, 2005).

While PageRank's basic premise is plausible, the algorithm is *static*, whereas real networks evolve in time. This causes the ranking by the algorithm to be severely biased by node age in growing networks (Chen et al, 2007; Mariani et al, 2015, 2016; Vaccario et al, 2017). The ranking by PageRank is strongly biased by node age also in our dataset (see Fig. 2). PageRank's bias in the patent citation network has different features with respect to its bias in papers' citation network reported by Mariani et al (2016). While in both datasets recent nodes are strongly disadvantaged by the algorithm, the oldest patents are not the most overvalued by the PageRank algorithm as opposed to what has been observed for papers (Mariani et al, 2016). This is a direct consequence of the significantly smaller citation count of the oldest patents. The peak of $n_{0.005}(p)$ is shifted to the left from the peak of $n_{0.005}(c)$, which means that PageRank nevertheless tends to favor older nodes with respect to citation count.

4.2. Time rescaled metrics $R(p)$ and $R(c)$

The strong age bias of the rankings by citation count and PageRank score implies that patents that appeared in some time periods are much more likely to rank well than other patents, independently of their properties such as novelty and significance. In bibliometrics (Radicchi et al, 2008; Waltman, 2016) and patent analysis (Triulzi et al, 2017), it is common to attempt to suppress this bias by age through various normalization procedures.

Here, we apply the rescaling procedure proposed by Mariani et al (2016) to citation count and PageRank. The rescaling procedure consists of comparing the score s_i of a given patent i with scores of the patents that belong to a reference-set of

patents of similar age[2] as patent i. By labeling the patents in order of decreasing age[3], the reference set is the set of Δ patents j such that $|i - j| < \Delta/2$.[4] Constructing the set of comparable patents based on a continuously moving window centered on a focal patent is advantageous with respect to grouping the patents by year as the latter results in imposing a sharp distinction between patents granted very closely in time but on either side of the January 1st boundary.

Denoting with $\mu_i(s)$ and $\sigma_i(s)$ the mean value and the standard deviation, respectively, of score s over the patent i's reference set, the rescaled score $R_i(s)$ of patent i is given by

$$R_i(s) = \frac{s_i - \mu_i(s)}{\sigma_i(s)}. \tag{4}$$

In this work, we set $\Delta = 15,000$, yet our results are robust with respect to other choices of Δ (not shown here).

As shown in Fig. 2, the rescaled scores $R(c)$ and $R(p)$ are much less biased by node age than the original scores c and p: $n_{0.005}(R(c))$ and $n_{0.005}(R(p))$ are remarkably stable across different age groups, and their value is always close to the expected unbiased value $n_{0.005}^{(0)}$. In agreement with Mariani et al (2016); Vaccario et al (2017); Liao et al (2017), this shows that the proposed rescaling procedure is effective in suppressing the temporal biases of the static metrics. By giving to old and recent patents the same chance of appearing to the top of the ranking, we expect the rescaled metrics to bring a substantial advantage in identifying valuable patents shortly after issuing. As the rankings by static metrics are biased toward old patents, we also expect the rescaled metrics' advantage in identifying significant patents to shrink (and eventually vanish) as we consider older significant patents. These hypotheses are validated in the next Section.

4.3. Evaluation of the metrics' performance in identifying the significant patents

To make quantitative statements on the ability of the metrics to single out the significant patents of different age, we introduce two evaluation metrics: the average ranking ratio and the identification rate.

Average ranking ratio. A straightforward way to assess the metrics' performance in identifying the significant patents would consist in calculating the average ranking position of the significant patents t years after they are issued. However, this simple measure is highly sensitive to the ranking position of the lowest-ranked items (Mariani et al, 2016).

To prevent this shortcoming, it is preferable to measure the *average ranking ratio* of the target items by the different metrics (Mariani et al, 2016), which is defined as follows. Denoting the rank of patent i by metric m as $r_i(m)$, the ranking ratio of metric m is defined as $\hat{r}_i(m) = r_i(m) / \min_{m'}\{r_i(m')\}$. The metric achieves the best-possible ranking ratio of one if it ranks a given significant patent best of all metrics; the lower the value, the better. The *average ranking ratio* $\langle\hat{r}\rangle(m)$ of metric m is the average of the ranking ratios $\hat{r}_i(m)$ of all significant patents, and it quantifies how much the metric underperforms, on average, with respect to the best-performing metric. A metric that outperforms all the other metrics for all the target nodes achieves an average ranking ratio $\langle\hat{r}\rangle = 1$; larger values of $\langle\hat{r}\rangle$ indicate a worse performance.

Identification rate. The identification rate $f_z(m)$ – commonly referred to as *recall* in the information filtering community (Lü et al, 2012) – of a given metric m is defined as the fraction of significant patents that are ranked among the top zN patents by metric m. Hence, while the average ranking ratio takes all significant patents and their ranking into account, the identification rate measure focuses on the top-items by each ranking.

4.4. Evaluating the evolution of the metrics' performance with patent age

To uncover the metrics' ability to early identify the significant patents, we evaluate the metrics' average ranking ratio and identification rate as a function of patent age. In this way, we are able to untangle the role of patent age in determining the metrics' performance; for example, a metric that is biased toward old patents only performs well in detecting *old* important patents, whereas we expect good early-identification metrics to perform well in detecting *recent* important patents.

To untangle the role of patent age in determining the metrics' performance, we dissect the network evolution by computing the rankings by the metrics every six months. At each ranking computation time $t^{(c)}$, only the patents issued before $t^{(c)}$ are included in the analysis. For significant patent i (issued at time t_i), we measure the age $\Delta t = t^{(c)} - t_i$ of the significant patent at time $t^{(c)}$. Then, we determine its ranking ratio values $\hat{r}_i(m; \Delta t)$ for all considered metrics m. Patent i's ranking ratio $\hat{r}_i(m; \Delta t)$ contributes to metric m's average ranking ratio $\langle\hat{r}\rangle(m; \Delta t)$ for Δt year-old patents. After having analyzed the whole network history, we can thus determine the average ranking ratio $\langle\hat{r}_i\rangle(m; \Delta t)$ of metric m for Δt years old patents as the average of $\hat{r}_i(m; \Delta t)$ over all the significant patents included in the analysis.

In the same way, we define the identification rate $f_z(m; \Delta t)$ of metric m for Δt years old patents as the fraction of significant patents that were ranked among the top zN patents by metric m when they were Δt years old.

---

[2] A potential limitation of this approach is that by comparing each patent's score with only the scores of patents of similar age, it may underestimate the importance of patents that happened to be issued in periods during which many breakthrough inventions took place. However, despite the well-known theory of Kondratiev waves and innovation clustering, robust empirical evidence for the existence of such periods is weak and debated. For instance, Silverberg and Verspagen (2003) found no evidence for innovation clustering in a list of basic inventions,whereas Korotayev et al (2011) found an evidence of Kondratiev cycles in the world-level patent output per inhabitant.

[3] We order by increasing ID those patents that are issued on the same day.

[4] The temporal window is defined in a slightly different way for patents close to the beginning and the end of the dataset. For the $\Delta/2$ patents closest to the beginning (end) of the dataset, the temporal window is given by the Δ oldest (most recent) patents in the dataset.

5. Results

5.1. Metrics' performance on the time-aggregate network

We start by assessing the average ranking ratio (Fig. 3A) and the identification rate (Fig. 3B) of the metrics on the whole dataset. The results show a clear advantage of the network-based metrics, p and $R(p)$, over the citation-counting metrics. According to the average ranking ratio (Fig. 3A), rescaled PageRank is the best-performing metric with a small margin over PageRank and a large margin over raw and rescaled citation count. Rescaled PageRank and PageRank also achieve the highest identification rates (Fig. 3B).

To understand where the gaps between the metrics stem from, we inspect the patents that give the largest contribution to c and $R(c)$'s ranking ratio – i.e., the patents that are ranked much better by p and $R(p)$ than by c and $R(c)$. We find a significant contribution coming from patent $4,237,224$ ("Process for producing biologically functional molecular chimeras", $c = 285$), which is ranked 2nd by $R(p)$ ($\hat{r} = 1$), 3rd by p ($\hat{r} = 1.5$), 1079th by $R(c)$ ($\hat{r} = 539.5$), and 1181st by c ($\hat{r} = 590.5$). Importantly, this patent gives the same contribution (equal to one) to all metrics' identification rate as all the metrics rank it among the top-0.5% patents. This example shows well that patents that are ranked at the top of the ranking by all metrics can have very different ranking ratio values. The second largest contribution to c's and $R(c)$'s average ranking ratio comes from patent $4,438,032$ ("Unique T-lymphocyte line and products derived therefrom", $c = 73$), which is ranked 253rd by p ($\hat{r} = 1$), 562nd by $R(p)$ ($\hat{r} = 2.2$), $48,742$nd by c ($\hat{r} = 192.7$), 66,014th by $R(c)$ ($\hat{r} = 260.9$). To check that the advantage of network-based metrics was not entirely due to these two patents, we have excluded them from the analysis and recalculated the metrics' average ranking ratio. PageRank and rescaled PageRank remain the two best-performing metrics ($\langle\hat{r}\rangle(p) = 4.2$, $\langle\hat{r}\rangle(R(p)) = 6.1$), yet their edge over the link-counting metrics ($\langle\hat{r}\rangle(c) = 8.1$, $\langle\hat{r}\rangle(R(c)) = 10.0$) significantly decreased.

5.2. Age-rescaling matters most for young patents

While the analysis of the previous Section reveals important differences among the metrics, the main goal of this manuscript is to reveal the dependence of the metrics' performance as a function of patent age, and to assess the metrics' ability to early-identify the significant patents. To this end, by following the procedure described in Section 4.4, we consider the ranking positions[5] of the group of expert-selected significant patent by Strumsky and Lobo (2015) one (Figs. 4A,D), five (Figs. 4B,E) and ten (Figs. 4C,F) years after issuing. Due to their lack of time bias, the rescaled metrics rank the significant patents much better than the corresponding static metrics one year after issuing (see Figs. 4A,D). On the other hand, the ranking positions by rescaled and static metrics are comparable ten years after issuing (see Figs. 4C,F).

The evolution of the ranking position of the significant patents as evaluated by p and $R(p)$ is shown in Supplementary

Movie M1; the same for c and $R(p)$ is shown in Supplementary Movie M2. The moving dots in these movies represent the significant patents, and the displacements of the dots represent the change in the significant patents' ranking position as they get older[6]. Movies M1 and M2 show that short after issuing, all significant patents are ranked higher by rescaled PageRank than by PageRank and citation count, respectively, consistently with Figs. 4A,D. In movie M1 that compares the rankings by p and $R(p)$, as the significant patents get older, the entity of their displacements in the ranking plane diminish, and they gradually drift toward the diagonal of the plane, which means that the gap between their ranking position by p and $R(p)$ shrinks. After ten years, most of the significant patents lie close to the diagonal, which indicates that the rankings of the significant patents by p and $R(p)$ are comparable.

5.3. Comparison of the four metrics' performance for different patents' age

The above-discussed Fig. 4 and Supplementary Movies M1–M2 show that the age of the significant patents has a large impact on the ability of the metrics to identify them. The goal of this section is to quantify the magnitude and the duration of the advantage of rescaled metrics in early identifying the significant patents, and to compare the obtained results with known results for scientific papers (Mariani et al, 2016).

To quantify how well the different metrics recognize the significant patents short after their issuing, we focus on the $M_{20} = 77$ patents that are at least 20 years old at the end of the dataset. By performing the evaluation procedure described in Section 4.4, we study how their average ranking ratio and identification rate depend on their age up to 20 years after issuing. We focus thus on a fixed group of target patents, which allows us to gauge the impact of time on the metrics' performance[7].

5.3.1. Average ranking ratio

In qualitative agreement with Fig. 4, Fig. 5A shows striking differences between the metrics' performance. Shortly after issuing, the rescaled metrics achieve an average ranking ratio much lower than that of the non-rescaled metrics. For example, one year after publication, PageRank's and rescaled PageRank's average ranking ratio are equal to 20.8 and 1.6, respectively, which indicates a performance advantage of one order of magnitude in favor of $R(p)$. The gap between rescaled PageRank and PageRank (rescaled citation count and citation count) closes 12 (7) years after issuing. There is therefore a medium-term temporal window over which the rescaled metrics rank the significant patents remarkably better than the non-rescaled metrics.

---

[5]The ranking positions considered in this paper are always normalized by the size of the system at the time when the ranking is computed.

[6]We only represent the significant patents after they have received their first citation. This is the reason why during the dynamics, some dots appear on the plane out of nowhere.

[7]Patents less than are 10 years old, for example, could not contribute to the age bins from 10 to 20 years after issuing. Were we including also them in the control group of significant patents, we would have ended up with a control group with different composition for different age bins, which would have confounded the temporal effects that we focus on here.
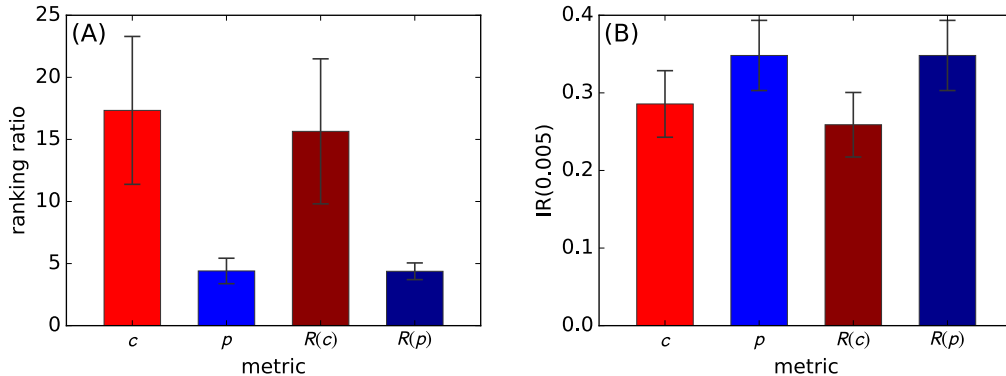
Figure 3: Performance of the metrics in identifying the significant patents from the list by Strumsky and Lobo, as measured by the metrics' average ranking ratio (panel A, the lower the better) and their identification rate (panels B, the higher the better) evaluated on the complete patent citation dataset

Importantly, once we have suppressed the age bias of c and p, we are able to reveal the advantage of using (higher-order) network information to rank the significant patents instead of simply counting citations, which manifests itself in the performance advantage of $R(p)$ over $R(c)$.

5.3.2. Identification rate

Fig. 5B shows the dependence of the metrics' identification rate $f_{0.005}(\Delta t)$ as a function of patent age. This evaluation measure quantifies the fraction of significant patents ranked in the top 0.5% by the metrics when they were Δt years old. The rescaled metrics outperform the non-rescaled metrics shortly after publication; the gap between rescaled and non-rescaled metrics closes eventually: p's performance reaches $R(p)$'s performance 12 years after issuing, and c's performance reaches $R(c)$'s performance 6 years after issuing. These two timescales are consistent with those observed for the average ranking ratio, and they define a temporal window over which the rescaled metrics achieve an improved identification of the significant patents.

5.4. The role of the network structure

In this Section, we address the following question: to which extent can the improved performance of PageRank-related metrics be explained by citation count dynamics alone? In other words, once we control for the effect of citation count dynamics and randomize the rest, can we reproduce the results in Fig. 5?

To address these questions, we use the the Dynamic Configuration Model (DCM) introduced by Ren et al (2017) to generate random networks that preserve the individual nodes' citation count time-series observed in the original network. Differently from the widely-used configuration model (Molloy and Reed, 1995), the DCM preserves the original network's temporal linking patterns (Ren et al, 2017). In the DCM, the total system time span T is divided into L layers of equal duration $\Delta t = T/L$. The randomized networks are thus generated by rewiring the existing connections, within each layer, by preserving each node's indegree and outdegree variation in that layer (see Ren et al

(2017) for the details). The expected number of edges $E_{ij}(n)$ from node j to node i at layer n is given by

$$E_{ij}(n) = \frac{\Delta k_i^{in}(n)\,\Delta k_j^{out}(n)}{E(n)},\qquad(5)$$

where Δk_i^{in} ($\Delta k_j^{out}(n)$) denotes the indegree (outdegree) increase of node i (j) in layer n, and $E(n)$ denotes the total number of edges in layer n. In our work, we set $L = 100$ which results in $\Delta t = 310$ days.

We compare the metrics' performance in the thus-generated random networks with the performance observed in the real data. By construction, the model preserves the indegree time-series of the original network; as a consequence, the performance of the citation count and rescaled citation count is the same as in the real data (Fig. 6). The model allows us to assess whether the advantage of network-based metrics (Fig. 6) is a genuine network effect or if it can be explained by random fluctuations.

In the randomized networks, the network-based metrics have no advantage with respect to citation-counting metrics in identifying the significant patents (Fig. 6A). In fact, $R(p)$ falls slightly below $R(c)$ for almost every patent age. Fig. 6B shows that the performance difference between the performance of PageRank-related metrics in real and randomized networks is significantly positive. We conclude that controlling for the individual nodes' citation count dynamics is not sufficient to explain our findings. Therefore, (higher-order) network structure plays a significant role for the advantage of network-based metrics with respect to citation-counting metrics in identifying the significant patents.

5.5. Top patents

In this section we inspect the top-ranked patents. For simplicity, we focus on the top-15 patents as ranked by PageRank (Table 3) and rescaled PageRank score (Table 4).

Table 3 shows that also patents with relatively few citations can reach the top of the ranking by PageRank score, which confirms the idea that in citation networks, the PageRank algorithm
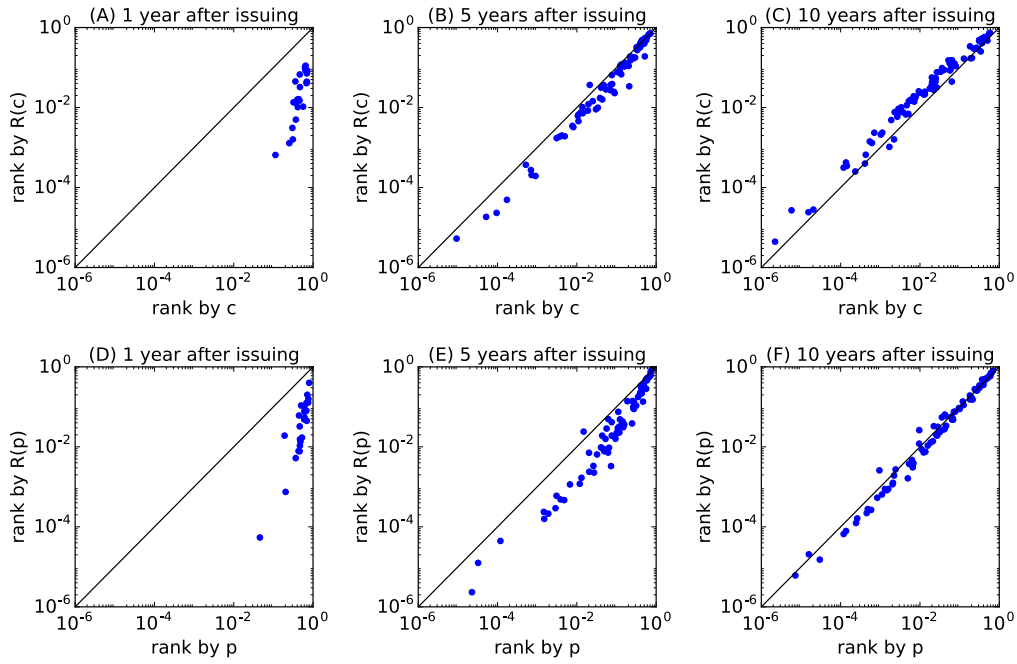
Figure 4: A comparison of the relative rankings (the lower, the better) of the significant patents by c and $R(c)$ one (panel A), five (panel B) and ten (panel C) years after issuing. Only the patents that received at least one citation at a given age are included. The same comparison between p and $R(p)$ is shown in panels D–F.
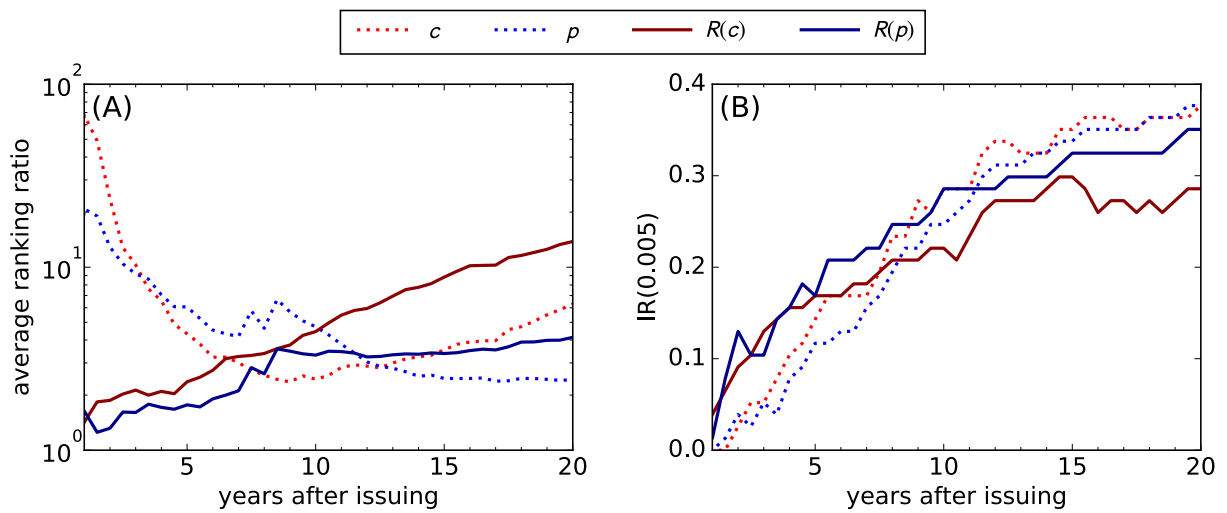

Figure 5: Performance of the metrics in identifying the significant patents from the list by Strumsky and Lobo over a 20-year time window after their issuing, as explained in the main text. *(A)* Average ranking ratio as a function of patent age. *(B)* Identification rate as a function of patent age.
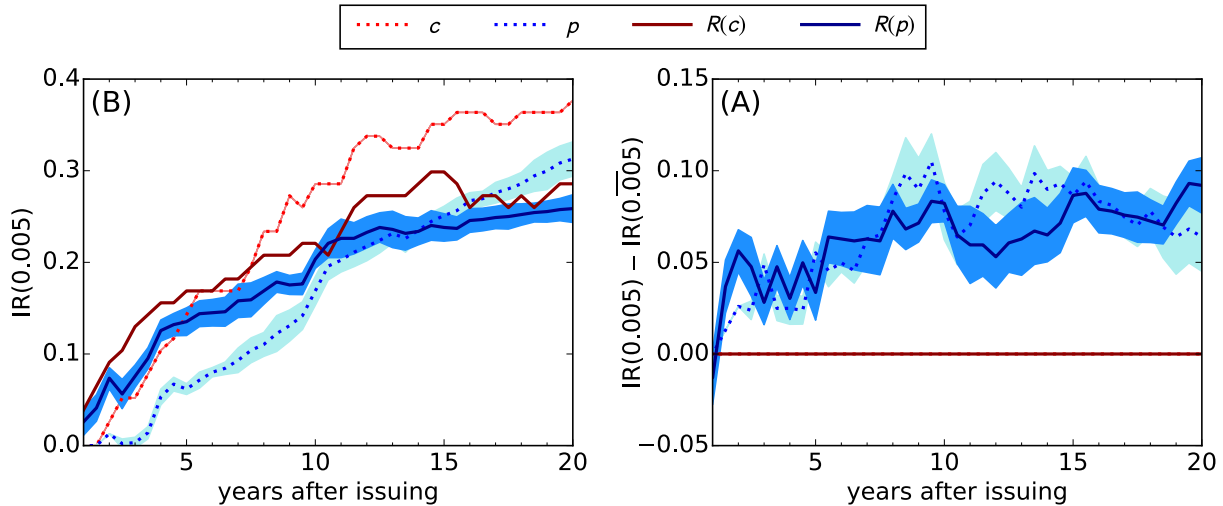
Figure 6: *(A)* Metrics' performance in identifying the significant patents in a random network generated with the Dynamic Configuration Model. The lines and the shaded areas around the lines represent the mean and the standard error, respectively, measured over 12 realizations of the randomization process. *(B)* Difference between the performance observed in the real data and that observed in the randomized networks generated with the Dynamic Configuration Model.

can identify "hidden gems" (Chen et al, 2007) that are underestimated by citation count. A paradigmatic example in this sense is patent 3813316 ("Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof"). The patent is ranked 6th by PageRank despite having been cited only 16 times. By inspecting the patent's neighborhood, it emerges that the reason for this is that the patent has been cited by patents with relatively large citation count and, additionally, small outdegree. For example, patent 3813316 is the only patent cited by patent 4237224 ("Process for producing biologically functional molecular chimeras", $c = 285$, included in the Strumsky-Lobo list of significant patents) which is ranked 3rd by PageRank. Highly scoring patent 3813316 refers only to patent 3723248 ("Method for producing ketoglucaric acid") which is consequently ranked 38th by PageRank despite having received only one citation. Small outdegree of the citing patents is crucial because it implies that a large portion of the citing patents' score will contribute to the cited patent's score in the score redistribution process defined by Eq. (1).

Table 4 shows that the top-15 patents by rescaled PageRank span a wider temporal range (1934-2010) than the top-15 by PageRank (1942-1996), which is a direct consequence of the age-bias removal. On the other hand, Table 4 also points out a potential limitation of the rescaling procedure. Among the 15 top-ranked patents, four are indeed 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. Within such a temporal window, a patent can achieve large rescaled 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. However, we prefer to show the

scores of all the patents in order to highlight the subtleties associated with the evaluation of very recent patents.

6. A comparison of the APS papers' and the US patents' citation network dynamics

Section 5 validates the rescaled metrics as better indicators of significance of recent patents than the non-rescaled metrics. Yet, there is a remarkable difference between the behavior of the identification rate observed in our analysis of the US patent dataset (Fig. 5B) and that reported by Mariani et al (2016) in their analysis of the American Physical Society (APS) paper citation network: Mariani et al (2016) found that $R(p)$ ranks more than 30% of the Physical Review Letters milestone letters in the top 0.5% already one year after publication, whereas it only ranks 1% of the Strumsky-Lobo significant patents in the top 1% one year after issuing.

The qualitative difference between Fig. 5B and Fig. 3B in Mariani et al (2016) for significant papers motivated us to explore the differences between the dynamics of (significant) patents and that of (significant) papers. To this end, we analyzed an extension of the dataset[8] used by Mariani et al (2016), and compared the obtained results with those obtained for the US patent citation network. The results of our analysis are summarized in Tables 2. The table shows that both the significant papers and the significant patents: (1) tend to be cited more than ordinary papers and ordinary patents, respectively, in the respective datasets; (2) tend to accrue citations faster than ordinary papers and ordinary patents, respectively. Like patents of high economic value (Lee and Sohn, 2017), the Strumsky-Lobo

---

[8]In particular, we analyzed the APS citation network from 1893 to 2016, which comprises 593,443 papers and 7,031,030 citations between them.

Dataset	Group of nodes	citations	τ_3	τ_5
US patents	Significant patents	105.6	9.6 y	12.0 y
	All patents	7.4	24.9 y	31.4 y
APS papers	Significant papers	457.0	1.0 y	1.4 y
	All papers	11.8	3.6 y	4.8 y

Table 2: A comparison between the average properties of all nodes and the average properties of the significant nodes in the APS paper and US patent citation network. The significant nodes are the milestone letters and the Strumsky-Lobo significant patents in the two datasets, respectively.

significant patents tend to receive the first few citations quicker than ordinary patents.

However, there is a striking difference between the dynamics of the two datasets: the APS papers tend to accrue citations much quicker than US patents. For example, the time needed for papers that received at least three total citations to receive their first three citations is much smaller for papers (3.6 years on average) than for patents (24.9 years). The same is true if we restrict the analysis to the significant papers (1.0 years) and patents (9.6 years), respectively. These results point out that the smaller identification rate for patents shortly after issuing is partly a manifestation of the slower citation dynamics of patents with respect to the citation dynamics of papers.[9]

7. Conclusions

Our paper has two main messages.

First, we find that using the whole network topology instead of only counting citations brings a substantial advantage in identifying the significant patents. Both the observed degree-degree correlations (Fig. 1) and the performance edge of PageRank-related metrics over citation-counting metrics (Fig. 5) suggest that important patents build on other other important patents. This supports the hypothesis that high-impact patents "stand on the shoulders of giants", in a similar way as scientific papers (Bornmann et al, 2010), although the high prevalence of examiner-added citations in patents makes the analogy imperfect.

Second, we show that removing the time bias of static centrality metrics allows one to identify significant patents much earlier than it is possible with conventional static metrics. The rescaling procedure which we use to remove the time bias is efficient and thus applicable even to large-scale datasets (Vaccario et al, 2017).

There are some limitations to our work that deserve to be discussed. First of all, we have pointed out that the early-identification of significant patents is more difficult than that

of significant papers, because patents take more time to accumulate citations (Section 6). Second, the time-rescaled metrics are based on the assumption that a good ranking of the patents should give the patents from different age periods the same chance to get to the top of the ranking. While this assumption is customary in paper citation analysis (Waltman, 2016), it creates a bias against patents that appear in periods of intensive breakthrough inventive activity, if they exist. Third, the rescaled metrics evaluate the most recent patents on the basis of citations received in a relatively short time period. While this may be justified by the finding that patents in rapidly growing domains are highly cited shortly after issuing (Benson and Magee, 2015), it potentially misses out "sleeping-beauty" (Ke et al, 2015) patents that received a substantial amount of citations only many years after issuing.

We see three major directions for extending this research. The most obvious is to acknowledge that there are different citation practices across technological fields, just as different scientific fields exhibit different citation patterns (Waltman, 2016). Based on the results by Vaccario et al (2017), we know that the rescaling procedure can in principle be extended to suppress the bias by technological field as well. However, while it is natural to suppress biases by scientific field of paper-level metrics due to their use in research evaluation, it remains unclear whether a similar approach would be the most effective strategy to rank patents. Besides, using patent classification information is problematic when the goal is to rigorously test predictive ability, because the classification system is changing often and many patents are reclassified (Lafond and Kim, 2017). Second, while there exist theoretical explanations for how the broad citation count distribution and the bias of citation-based metric by node age emerge as a result of the dynamics of the system (Valverde et al, 2007; Newman, 2009; Medo et al, 2011; Mariani et al, 2015), a model-based explanation of the strong degree-degree correlations and the improved PageRank performance observed in our dataset is still lacking. Third, while we studied PageRank as a paradigmatic network-based metric because of its plausible assumption ("a node is important if it is cited by other important nodes"), other network-based metrics (Liao et al, 2017) can be analyzed in a similar way to improve our understanding of which metrics best identify important patents.

Albert MB, Avery D, Narin F, McAllister P (1991) Direct validation of citation counts as indicators of industrially important patents. Research policy 20(3):251–259

Bedau MA, Buchanan AJ, Chalmers FCC D W, Packard NH, Pepper NM (2011) Evidence in the patent record for the evolution of technology using citation and pagerank statistics. In: Lenaerts T, Giacobini M, Bersini H, Bourgine P, Dorigo M, Doursat R (eds) Advances in artificial life, ECAL 2011: Proceedings of the eleventh European conference on the synthesis and simulation of living system, MIT Press, p 7784

Benson CL, Magee CL (2015) Quantitative determination of technological improvement from patent data. PloS one 10(4):e0121,635

---

[9] We emphasize that the APS papers dataset contains only citations among the APS papers. By considering the citations to APS papers from non-APS papers, τ_3 and τ_5 would further decrease and the difference from the patent dataset (which, by contrast to the APS dataset, is comparably complete) would further magnify.

Berkhin P (2005) A survey on pagerank computing. Internet Mathematics 2(1):73–120

Bornmann L, de Moya Anegón F, Leydesdorff L (2010) Do scientific advancements lean on the shoulders of giants? a bibliometric investigation of the ortega hypothesis. PLoS One 5(10):e13,327

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30(1):107–117

Bruck P, Réthy I, Szente J, Tobochnik J, Érdi P (2016) Recognition of emerging technology trends: class-selective study of citations in the us patent citation network. Scientometrics 107(3):1465–1475

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

Castaldi C, Frenken K, Los B (2015) Related variety, unrelated variety and technological breakthroughs: an analysis of us state-level patenting. Regional studies 49(5):767–781

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

Comins JA, Leydesdorff L (2017) Citation algorithms for identifying research milestones driving biomedical innovation. Scientometrics 110(3):1495–1504

Corredoira RA, Banerjee PM (2015) Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity. Research Policy 44(2):508–521

Csárdi G, Strandburg KJ, Zalányi L, Tobochnik J, Érdi P (2007) Modeling innovation by a kinetic description of the patent citation system. Physica A: Statistical Mechanics and its Applications 374(2):783–793

Dechezleprêtre A, Martin R, Mohnen M (2014) Knowledge spillovers from clean and dirty technologies. Centre for Economic Performance WP

Epicoco M (2013) Knowledge patterns and sources of leadership: Mapping the semiconductor miniaturization trajectory. Research Policy 42(1):180–195

Fontana R, Nuvolari A, Shimizu H, Vezzulli A (2013) Reassessing patent propensity: Evidence from a dataset of r&d awards, 1977–2004. Research Policy 42(10):1780–1792

Fortunato S, Boguñá M, Flammini A, Menczer F (2008) Approximating pagerank from in-degree. In: Algorithms and Models for the Web-Graph, Springer, pp 59–71

Franceschet M (2011) Pagerank: Standing on the shoulders of giants. Communications of the ACM 54(6):92–101

Gleich DF (2015) Pagerank beyond the web. SIAM Rev 57(3):321–363

Hall BH, Jaffe A, Trajtenberg M (2005) Market value and patent citations. RAND Journal of economics pp 16–38

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

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

Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of dna theory. Social networks 11(1):39–63

Jaffe AB, de Rassenfosse G (2017) Patent citation data in social science research: Overview and best practices. Journal of the Association for Information Science and Technology 68(6):1360–1374

Jaffe AB, Trajtenberg M, Fogarty MS (2000a) Knowledge spillovers and patent citations: Evidence from a survey of inventors. American Economic Review Papers and Proceedings 90(2):215–218

Jaffe AB, Trajtenberg M, Fogarty MS (2000b) The meaning of patent citations: Report on the nber/case-western reserve survey of patentees. Tech. rep., National bureau of economic research

Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences 112(24):7426–7431

Kogan L, Papanikolaou D, Seru A, Stoffman N (2017) Technological innovation, resource allocation, and growth. Quarterly Journal of Economics

Korotayev A, Zinkina J, Bogevolnov J (2011) Kondratieff waves in global invention activity (1900–2008). Technological forecasting and social change 78(7):1280–1284

Lafond F, Kim D (2017) Long-run dynamics of the US patent classification system. arXiv 1703.02104

Lanjouw JO, Schankerman M (2004) Patent quality and research productivity: Measuring innovation with multiple indicators. The Economic Journal 114(495):441–465

Lee J, Sohn SY (2017) What makes the first forward citation of a patent occur

earlier? Scientometrics pp 1–20

Liao H, Mariani MS, Medo M, Zhang YC, Zhou MY (2017) Ranking in evolving complex networks. Physics Reports 689:1–54

Lü L, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T (2012) Recommender systems. Physics Reports 519(1):1–49

Lü L, Chen D, Ren XL, Zhang QM, Zhang YC, Zhou T (2016) Vital nodes identification in complex networks. Physics Reports 650:1–63

Lukach R, Lukach M (2007) Ranking uspto patent documents by importance using random surfer method (pagerank). Available at SSRN: https://papersssrncom/sol3/paperscfm?abstract_id=996595

Mariani MS, Medo M, Zhang YC (2015) Ranking nodes in growing networks: When pagerank fails. Scientific Reports 5

Mariani MS, Medo M, Zhang YC (2016) Identification of milestone papers through time-balanced network centrality. Journal of Informetrics 10(4):1207–1223

Martinelli A (2012) An emerging paradigm or just another trajectory? understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. Research Policy 41(2):414–429

Martinelli A, Nomaler Ö (2014) Measuring knowledge persistence: a genetic approach to patent citation networks. Journal of Evolutionary Economics 24(3):623–652

Medo M, Cimini G, Gualdi S (2011) Temporal effects in the growth of networks. Physical Review Letters 107(23):238,701

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random structures & algorithms 6(2-3):161–180

Monechi B, Gravino P, Servedio VD, Tria F, Loreto V (2017) Significance and popularity in music production. Royal Society Open Science 4(7):170,433

Narin F, et al (1976) Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. Computer Horizons Washington, D. C

Newman M (2009) The first-mover advantage in scientific publication. Europhys Lett 86(6):68,001

Newman M (2010) Networks: an introduction. Oxford University Press

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences 105(45):17,268–17,272

Ren ZM, Mariani MS, Zhang YC, Medo M (2017) A time-respecting null model for growing networks. arXiv preprint arXiv:170307656

Shaffer MJ (2011) Entrepreneurial innovation: Patent rank and marketing science. Washington State University

Silverberg G, Verspagen B (2003) Breaking the waves: a poisson regression approach to schumpeterian clustering of basic innovations. Cambridge Journal of Economics 27(5):671–693

Silverberg G, Verspagen B (2007) The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance. Journal of Econometrics 139(2):318–339

Spitz A, Horvát EÁ (2014) Measuring long-term impact based on network centrality: Unraveling cinematic citations. PLoS One 9(10):e108,857

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

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

Triulzi G, Alstott J, Magee CL (2017) Predicting technology performance improvement rates by mining patent data. SSRN 2987588

Vaccario G, Medo M, Wider N, Mariani MS (2017) Quantifying and suppressing ranking bias in a large citation network. Journal of Informetrics 11(3):766–782

Valverde S, Solé RV, Bedau MA, Packard N (2007) Topology and evolution of technology innovation networks. Physical Review E 76(5):056,118

Verspagen B (2007) Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems 10(01):93–115

Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment 2007(06):P06,010

Waltman L (2016) A review of the literature on citation impact indicators. Journal of Informetrics 10(2):365–391

Wasserman M, Zeng XHT, Amaral LAN (2015) Cross-evaluation of metrics to estimate the significance of creative works. Proceedings of the National Academy of Sciences 112(5):1281–1286

Zhang Y, Qian Y, Huang Y, Guo Y, Zhang G, Lu J (2017) An entropy-based

indicator system for measuring the potential of patents in technological innovation: rejecting moderation. Scientometrics 111(3):1925–1946

Table 3: Top-15 patents as ranked by PageRank score p (asterisks mark the Strumsky-Lobo significant patents).

Rank	Patent ID	Patent title	Issuing date	c	$p \cdot 10^5$
1	4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	28-7-1987	1956	2.824
2	4683202	Process for amplifying nucleic acid sequences	28-7-1987	2169	2.691
3	4237224	(*) Process for producing biologically functional molecular chimeras	2-12-1980	285	2.687
4	4395486	Method for the direct analysis of sickle cell anemia	26-7-1983	71	1.731
5	4723129	Bubble jet recording method and apparatus in which a heating element generates bubbles in a liquid flow path to project droplets	2-2-1988	1962	1.416
6	3813316	Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof	28-5-1974	16	1.399
7	5536637	Method of screening for cDNA encoding novel secreted mammalian proteins in yeast	16-6-1996	422	1.344
8	4558413	Software version management system	10-12-1985	1956	1.326
9	4358535	Specific DNA probes in diagnostic microbiology	9-11-1982	436	1.324
10	2297691	Electrophotography	6-10-1942	588	1.312
11	4463359	Droplet generating method and apparatus thereof	31-7-1984	1694	1.263
12	5523520	Mutant dwarfism gene of petunia	4-6-1996	1139	1.221
13	4812599	Inbred corn line PHV78	14-3-1989	179	1.187
14	4740796	Bubble jet recording method and apparatus in which a heating element generates bubbles in multiple liquid flow paths to project droplets	26-4-1988	1663	1.093
15	5103459	System and method for generating signal waveforms in a CDMA cellular telephone system	7-4-1992	1208	1.034

Table 4: Top-15 patents as ranked by rescaled PageRank score $R(p)$ (with $\Delta_p = 15,000$). Asterisks mark the Strumsky-Lobo significant patents.

Rank	Patent ID	Patent title	Issuing date	c	$R(p)$
1	7764447	Optical element holding device, lens barrel, exposing device, and device producing method	27-7-2010	1	104.9
2	4237224	(*) Process for producing biologically functional molecular chimeras	2-12-1980	285	99.1
3	2297691	Electrophotography	6-10-1942	588	91.8
4	7749477	Carbon nanotube arrays	6-7-2010	1	84.8
5	7784029	Network service for modularly constructing a software defined radio	24-8-2010	1	78.9
6	5536637	Method of screening for cDNA encoding novel secreted mammalian proteins in yeast	16-7-1996	422	78.9
7	4683195	Process for amplifying, detecting, and/or-cloning nucleic acid sequences	28-7-1987	1956	78.2
8	5523520	Mutant dwarfism gene of petunia	4-6-1996	1139	76.9
9	4395486	Method for the direct analysis of sickle cell anemia	26-7-1983	71	75.0
10	4683202	Process for amplifying nucleic acid sequences	28-7-1987	2169	74.6
11	7779788	Animal training system with multiple configurable correction settings	24-8-2010	1	73.3
12	1970578	Assistants for the textile and related industries	21-8-1934	241	73.3
13	3813316	Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof	28-5-1974	16	72.7
14	5572643	Web browser with dynamic display of information objects during linking	5-11-1996	1120	71.8
15	4723129	Bubble jet recording method and apparatus in which a heating element generates bubbles in a liquid flow path to project droplets	2-2-1988	1962	68.9