sustainability

MDPI

Article

Early Identification of Significant Patents Using Heterogeneous Applicant-Citation Networks Based on the Chinese Green Patent Data

Xipeng Liu * and Xinmiao Li

School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
* Correspondence: xp_liu@163.sufe.edu.cn

Abstract: With the deterioration of the environment and the acceleration of resource consumption, green patent innovation focusing on environmental protection fields has become a research hot-spot around the world. Previous researchers constructed homogeneous information networks to analyze the influence of patents based on citation ranking algorithms. However, a patent information network is a complex network containing multiple pieces of information (e.g., citation, applicant, inventor), and the use of a single information network will result in incomplete information or information loss, and the obtained results are biased. In addition, scholars constructed centrality indicators to assess the importance of patents with less consideration of the age bias problem of algorithms and models, and the results obtained are inaccurate. In this paper, based on the Chinese green patent (*CNGP*) dataset from 1985 to 2020, a *CNGP* heterogeneous applicant-citation network is constructed, and the rescaling method and normalization procedure are used to solve the age bias. The results illustrate that the method proposed in this paper is able to identify significant patents earlier, and the performance of the rescaled indegree (*R_ID*) works best such as the *IR* score is 17.32% in the top 5% of the rankings, and it is the best in the constructed dynamic heterogeneous networks as well. In addition, the constructed heterogeneous information network has better results compared with the traditional homogeneous information network, such as the *NIR* score of *R_ID* metrics can be improved by 2% under the same condition. Therefore, the analysis method proposed in this paper can reasonably evaluate the quality of patents and identify significant patents earlier, thus providing a new method for scientists to measure the quality of patents.

Keywords: green innovation; Chinese green patents; heterogeneous network; age bias

check for updates

Citation: Liu, X.; Li, X. Early Identification of Significant Patents Using Heterogeneous Applicant-Citation Networks Based on the Chinese Green Patent Data. *Sustainability* **2022**, *14*, 13870. https://doi.org/10.3390/su142113870

Academic Editor: Luigi Aldieri

Received: 22 September 2022
Accepted: 22 October 2022
Published: 25 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, with the aggravation of global environmental pollution and resource shortages, it has been proposed to use green technology and other measures to solve this problem [1]. The Chinese government has issued relevant policies to encourage and support the development of green industries, thus promoting the rapid development of green technology and environmental protection industries [2,3]. Existing studies use patents as an important indicator of innovation, and China has become the world's largest patent application country since 2011, but the quality of patents has become uneven extremely, and how to identify significant patents from them earlier is a topic worthy of research [4]. Scholars have built patent citation networks based on citation data and created a series of centrality metrics to identify the impact of patents [5]. However, those networks are homogeneous information networks, which only contain patent nodes while ignoring the impact of other information of patents. More seriously, most scholars assessed the impact or significance of patents using citation-based ranking metrics, and fewer analyzed the age bias problem that exists in ranking metrics and the impact on the performance evaluation indicators [6]. Therefore, this paper needs to address the following two issues:

first, using the multidimensional information of patents, and constructing a heterogeneous information network to evaluate the importance of patents more comprehensively. Second, when using the citation-based ranking metrics, it needs to analyze the age bias, so as to identify significant patents earlier.

It is found that if only analyzing the homogeneous information network, it generally misses important information that is useful for further exploring the nature and laws of the research target. Most of the information networks existing, in reality, belong to the heterogeneous network, which contains richer information, hence building a heterogeneous information network can be more comprehensive and closer to the real information [7]. Since patent information contains applicants and inventors, among which applicants are subjects who can apply for and obtain patent rights, they are of great significance to patents. For example, in a certain technological domain, if an authoritative applicant has the strength of technical monopoly or has a certain influence in the domain, then the quality of his patent application is generally higher than that of other applicants' patents [8]. Therefore, this paper combines the applicant and citation information of patents to build a *CNGP* heterogeneous information network, which not only can alleviate the sparsity problem of homogeneous information network, but also can reflect the actual situation of patents more truly.

When we use citation-based ranking metrics to evaluate the node importance of a network, we inevitably discuss the age bias induced by the ranking metrics [9,10]. Since the constructed datasets are truncated and the number of citations accumulates over time, old patents have the advantage of a long time span to obtain more citations compared to young patents [10,11]. Then, the use of classic ranking metrics (e.g., the citation count, PageRank) suffers from age bias [9,12]. In order to suppress age bias, scholars have proposed methods such as CiteRank [13], Time Weighted PageRank [14], and rescaling method [10], among which the rescaling method proposed by Mariani et al. [10] was shown to be effective in suppressing age bias of ranking metrics, it was used to adjust the ranking metrics to achieve relative fairness in ranking old and young patents.

In this paper, based on the Chinese green patent (*CNGP*) dataset from 1985 to 2020, a *CNGP* heterogeneous applicant-citation network is constructed and uses the Chinese Patent Award (*CPA*) data as the expert-selected significant patents to identify. In the model analysis, the rescaling method and normalization procedure are used to solve the problem of age bias. The results illustrate that the proposed analysis method can not only identify the significant patents earlier, but also the constructed heterogeneous information network with better performances. Therefore, the analysis method proposed in this paper can reasonably evaluate the quality of patents and provides a new method to measure the quality of patents.

It is an extraordinary significance to evaluate the quality of patents by building heterogeneous information networks. This study has the following innovations. (1) To the best of our knowledge, this paper is the first study to build a dataset of Chinese green patents and conduct patent importance analysis, laying a solid foundation for the research of green innovation in China. (2) Combining the applicant information of patents with the citation information is a way to analyze patents using multidimensional information, which provides scholars with new perspectives to study patent quality. (3) In the heterogeneous information network, we consider the effect of age bias, and the rescaling method and normalization procedure are used to solve the age bias. Thus, constructing a complete analysis method and providing a new approach for scholars to study the importance of patents.

The paper is organized as follows. In Section 2, we present the work related to our research. In Section 3, we describe and analyze the dataset in this study, including the build steps of the *CNGP* dataset and the expert-selected significant patents, then in-depth analyze the obtained data. In Section 4, we introduce the heterogeneous applicant-citation networks, and present the considered ranking algorithms and the evaluation indicators.

In Section 5, we evaluate and analyze the results. In Section 6, we offer some discussion. Finally, in Section 7, we draw the conclusions of this paper.

2. Related Works

2.1. Patent Quality Analysis

The scientific research on patent analysis is of tremendous interest to scientists and practitioners because of its importance in, for instance, strategic planning [15], analyses of the competitiveness of companies [16], research and development (R&D) planning [17], and technology forecasting [18]. In general, the quality of patents cannot be directly measured, most scholars [19,20] evaluated the quality of patents by estimating the value of patents. The value of a patent includes the following three parts: commercial value, technological value, and legal value. Kogan et al. [21] evaluated the patent's commercial value based on the market reaction caused by its granted announcement. Yuan and Li [22] believed that patents with more scientific knowledge have higher technological value, and Mezzanotti [23] indicated that those patents with legal proceedings or disputes generally have higher legal value than other patents. Therefore, according to these three values of patents, researchers have constructed many reasonable and novelty indicators to evaluate the value of patents.

As a phased achievement of technological innovation, scientists and scholars use the inherent attributes and document information of patents to put forward dozens of quantitative indicators and then evaluate the technological value of patents. For example, Harhoff et al. [19] identified forward citations in patents as a reliable indicator to detect the value of a patent. Other patent indicators include the quality of claims, family size, year of the grant, the validity of the patent, and science intensity, etc. [24,25].

With the development of network scientific research, state-of-the-art models and algorithms are gradually applied in patent information analysis. Lin et al. [26] established a patent citation network to evaluate the value of patents, and Chung and Sohn [27] applied a deep learning framework to mine the textual information of patents to predict the number of forward citations. Although the repeated evidence suggests a positive relationship between citations received and different measures of value, it is generally acknowledged that the relationship is noisy, so using more elaborated indicators may be preferable to simple citations in identifying the value of patents. Here, we should discuss two basic questions, how to build a patent information network to evaluate the value of patents more comprehensively, and how to identify those significant patents earlier, although they did not have enough time to accumulate a high number of citations.

2.2. Patent Information Network

The patent dataset covers the basic information of patents, such as applicant and inventor, citation information, classification number, date and text, etc. Recently, social network analysis methods have been introduced into patent information analysis, and various types of patent information network frameworks have been proposed. (1) Using applicant or inventor information to establish a patent collaboration network which is useful for collaborative innovation, and scholars have explored the factors influencing patenting activity and the motivation for cooperation [28,29]. (2) Using patent citation information to establish a patent citation network is conducive to analyzing core patents and emerging technologies, and exploring the evolution path of innovative technologies [30,31]. (3) Using the classification number of patents to establish a subject word network or a subject similarity network, it is beneficial to discover the centrality and similarity of the technology [32,33].

Bibliometric methods have been used to measure and analyze citation networks in various scenarios, such as the scientific impact of papers [7], individual researchers [34], journals [35], etc. Patent citation analysis gained traction relatively late (in the 1990s) compared to their scholarly articles' counterparts. Karki [36] proposed several technological indicators based on citations among patents. Mariani et al. [31] use the US patent citation

network to early identify a list of expert-selected historically significant patents through citation network analysis. Most of the information networks established in literature belong to homogeneous information networks [10,30]. However, the data in the real world generally contains heterogeneous information of different types of nodes and edges. For example, the entities included in the patent information include the patent itself, applicants, and inventors. The edges of the patent include the citation relationships and the affiliation relationships. Zhou et al. [7] proposed an interactive model of author-paper bipartite networks and an iterative algorithm to obtain a better ranking for scientists and their publications. Du et al. [37] presented an iteration algorithm called inventor-ranking, to sort the influences of patent inventors in heterogeneous networks constructed based on their patent data. Zhao et al. [38] utilized heterogeneous author-citation networks to measure authors' academic influence. In addition, scholars mainly used patent citation data to construct a homogeneous network to analyze the importance of patents, and less considered the impact of other information on the importance of patents.

Table 1 summarizes the literature related to the patent information network, including the type of network, the category of information, the purpose of research, and the reference sources. From Table 1 we can see that information contained in patents such as citation information, inventor information, or classification information are used to analyze the patents. In addition, when constructing the heterogeneous information network analysis, some scholars mainly discuss the influence of scholars in the literature network of scholars, or the influence of inventors in the inventor-citation network, and fewer analyze the impact or importance of patents by constructing the applicant-citation network, where the significance of the applicant to the patent has been described above. Therefore, this paper uses the patent's applicant and citation information to construct a heterogeneous applicant-citation network, so as to analyze the importance of patents more comprehensively.

Table 1. Summary of patent information network.

Network Type	Information	Purpose	References
Homogeneous information network	Inventors of patents	Explore the importance of the mobility of knowledge workers for the formation of collaborative patents across different regional contexts.	Miguelez [28]; Liu et al. [29]
	Citations of patents	Identify clusters of patents and prediction them;identify significant patent.	Érdi et al. [30]; Miriani et al. [31]
	Classifications of patents	Using the classification number of patents to establish a subject word network or a subject similarity network.	Tseng et al. [32]; Yoon et al. [33]
Heterogeneous information network	Authors and citations of papers	Build a heterogeneous author-citation academic network for ranking authors and their papers.	Zhou et al. [7]; Zhao et al. [34]
	Inventors and citations of patents	Sort the influences of patent inventors in heterogeneous networks constructed based on their patent data.	Du et al. [38]

2.3. Citation-Based Ranking Metrics and Bias

Ranking metrics are pervasive in our increasingly digitized society, with important real-world applications including recommender systems, search engines, and influencer marketing practices [6]. From a network science perspective, citation-based ranking metrics constitute a key tool in scientometrics and play an increasingly important role in research evaluation [5]. On a patent citation network, patents are connected by citation relationships, and the value of patents can be simply evaluated by calculating citations (referred to as in-degree centrality in the network science) for each patent. However, this metric considers that the importance of each citation is the same and ignores the differences between citation relationships. Hence, Brin and Page [39] proposed the popular PageRank metric by using the global information of the network. Its core idea is that: "a node is important if it is linked by other important nodes" [12]. Due to the originality of this metric, it has been widely ap-

plied in real systems ranging from information to biological and infrastructure networks [9]. However, the PageRank metric is not suitable for all specific problems, so variants of PageRank have been proposed. For example, LeaderRank has shown good performance in both social networks and citation networks [40]. In addition, Namtirtha et al. [41] proposed a ranking metric named the network global structure-based centrality, which has good performance in identifying important nodes in complex networks as well.

However, a patent information network is a growing network in which the number of nodes gradually increase over time, old nodes have more time to acquire citations than young nodes [42]. The average citations of those patents with a fixed age will gradually increase [6]. Therefore, when we use ranking metrics such as citations or PageRank to measure network centrality, we inevitably need to consider the impact of age bias included in these metrics. Mariani et al. [10] argued that the age bias of these metrics can be rescaled by using a transformation that ensures that the average score of a node and its standard deviation are independent of the age of the node. After adopting this transformation, the resulting "rescaled" score can identify important nodes earlier, which is significant for the early detection of milestone papers [40], patents [31], etc.

The main hope motivating the use of metrics for ranking and prediction tasks is that they might provide a relatively objective evaluation of the value of an agent (e.g., the quality of a patent), whereas human or expert judgment might be subjective and influenced by biases and social factors [6]. However, most of the constructed datasets are obtained through manual processing such as patents selected by experts being used as significant patents. If our task is to identify such patents, the age distribution of the selected significant patents would greatly impact the performance of the ranking metric as well. If the expert-selected significant patents are older, then performance evaluation metrics that ignore this bias will favor ranking metrics that favor old patents. Hence, 'corrected' performance evaluation metrics that penalized those biased metrics are not affected by this confounding effect [40]. Therefore, this paper not only explores the age bias in the ranking metrics on the patent information network, but also discusses the interplay between the bias of the evaluated ranking metrics and the bias of the significant patents.

3. Data and Analysis

We collected the Chinese green patents (*CNGP*) dataset, which spans the period covering the years 1985 to 2020. This dataset contains patent citations and applicant information, where patents gradually appear with time. Nodes include patents and applicants, directed links represent patent citation relationships, and undirected links represent patent-applicant relationships. There are a set of corresponding expert-selected patents of high impact that are referred to as significant patents. Table 2 summarizes the analyzed dataset's basic characteristics, including the time span, patent nodes, applicant nodes, patent citation edges, patent-applicant edges, and the corresponding sets of significant patents.

Table 2. Basic characteristics of the Chinese green patent dataset.

Data Field	Data Value
Time span	1985–2020
Patent nodes	878,007
Applicant nodes	202,764
Patent-patent edges	1,676,458
Patent-applicant edges	516,201
Significant patents	839

This section includes the following three parts: first, we describe how to build the *CNGP* applicant-citation dataset. Then, we explain the source of expert-selected significant patents and match them to the *CNGP* dataset. Finally, we have an in-depth analysis of the obtained data, thereby laying the foundation for the research of this study.

3.1. CNGP Applicant-Citation Dataset

As one of the important means of environmentally sustainable development, green innovation has attracted more attention from the society and government. However, as the world's largest patent applicant country since 2011, China has not existed a corresponding dataset to study green innovation yet. As the phased achievements of technological innovation, patents are of extraordinary significance to analyze. Hence, the establishment of a Chinese green patent dataset is beneficial to the in-depth exploration of the Chinese green innovation process and to discover significant patents earlier. The flowchart for building the *CNGP* applicant-citation dataset is shown in Figure 1.

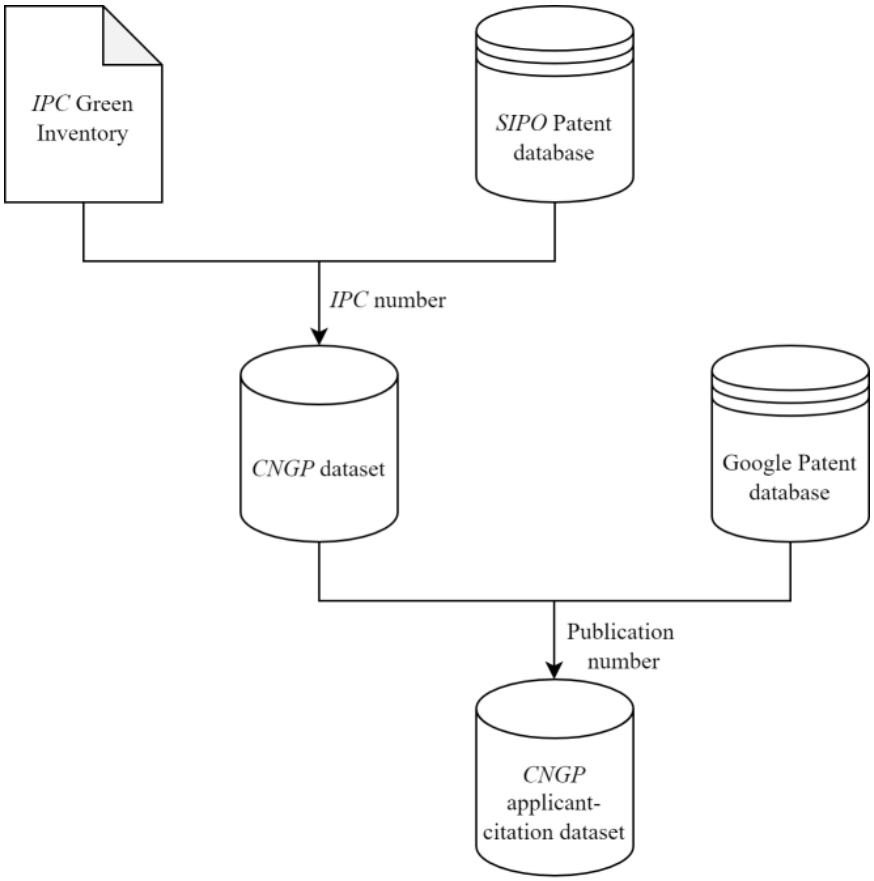


Figure 1. The flowchart for building the *CNGP* applicant-citation dataset.

The procedure of building the *CNGP* applicant-citation dataset as follows:

(1) Using crawler technology to collect the patent invention applications from the "China National Intellectual Property Administration" (*CNIPA*) website (http://epub.cnipa.gov.cn/ (accessed on 1 July 2022)). Some scholars studied show that the data from this website is valuable for studying patent quality and can be used as a proxy variable for studying innovation [43,44]. After collecting and sorting, the total number of invention patents was 12,814,946, obtained from 1985 to 2020.

(2) In order to identify Chinese green patents, we are using the patent's International Patent Classification (*IPC*) number to match the *IPC* green inventory published by the World Intellectual Property Organization (WIPO) (https://www.wipo.int/classifications/ipc/green-inventory/home (accessed on 1 July 2022)), which left a total of 1,670,450 green patents in China.

(3) Since the *CNIPA* database does not include patents' citation information. Google Patent (https://patents.google.com/ (accessed on 1 July 2022)), fortunately, provides related citation data for all Chinese patent applications. Moreover, Google Patent

updates the forward citation data of each patent according to the information on the backward citation. We link those data with the *CNGP* dataset through the patent's publication number [45,46].

(4) Finally, we further process the obtained data. Referring to the operation of Kogan et al. [21], we retain the citation relationships from the year of 1985 to 2020 and ensure that the cited patents belong to the *CNGP* dataset. In addition, we keep information about each corresponding applicant and link it to the patent. Therefore, the *CNGP* applicant-citation dataset is composed of the number of patent nodes which are 878,007, and the number of applicant nodes are 202,764, the number of directed citation edges are 1,676,458, and the number of undirected applicant-patent edges are 516,201 in the end.

3.2. Expert-Selected Significant Patents

The Chinese Patent Award (*CPA*) is the highest award in the field of Chinese patents, which is jointly issued by the *CNIPA* and the *WIPO* since 1989. This award is the only government award in China that specifically award authorized patents, and has a certain international influence. The evaluation criteria for the patent award not only grab the legal, technical, and market dimensions of the patent, but also care about the social benefits and development prospects of the patent. Therefore, patents that have won this award are scientific and feasible as high-quality patents.

In addition, Moser and Nichalas's [47] studies of US patents found that the use of incentives attracts more innovators and has a positive impact on more patents and better innovation at a later stage. The *CPA* is a government department of award set up by the Chinese government to find and reward high-quality Chinese patents, which reflects the technological quality and economic benefits of Chinese patents, focuses on the value of Chinese patents, and plays a role in leading innovation [48]. Some scholars believed that the award-winning patents selected by the *CPA* are characterized by high creation quality, strong patent protection, and good patent application, and thus the intellectual property rights and innovations represented by the *CPA* have a great impact and contribution to society [49,50]. Other scholars also directly take the Chinese Patent Gold Award as a high-value patent and analyze the differences in patent quality among different types of patent owners and different regions [51,52]. Based on the data of *CPA*, some researchers construct a patent quality assessment indicator framework, and the studies find that the award results of *CPA* are relatively fair, and the awarded patents have a higher value than the non-awarded patents [48,53,54]. Therefore, it is reasonable to adopt the *CPA* as the label of significant patents in this paper.

The steps for collecting the award-winning patent dataset are as follows. First, by visiting the website of the *CNIPA* (https://www.cnipa.gov.cn/col/col41/index.html (accessed on 1 July 2022)), collecting and sorting out all the award-winning information of the patent invention applications, and standardizing the processing of the dataset, we obtain the number of all the award-winning patents which is 6169. Then, restricting our analysis to those patents that were issued within our dataset's temporal span, and matching the green patents in the *CNGP* applicant citation dataset, and the number of the green patents is 839 in the end. From 1989 to 2020, there are a total of 22 sessions of the Chinese Patent Award. The number of green patents awarded in the *CPA* by session is shown in Figure 2. Before 2007, this award was held biennially, and after that, it was held annually. Figure 2 shows that the number of invention patents awarded is not fixed in each session, but the number of awards for green patents has increased obviously since 2015. It can be seen that to achieve sustainable development, the research of green patents has gradually received attention from society and the government.

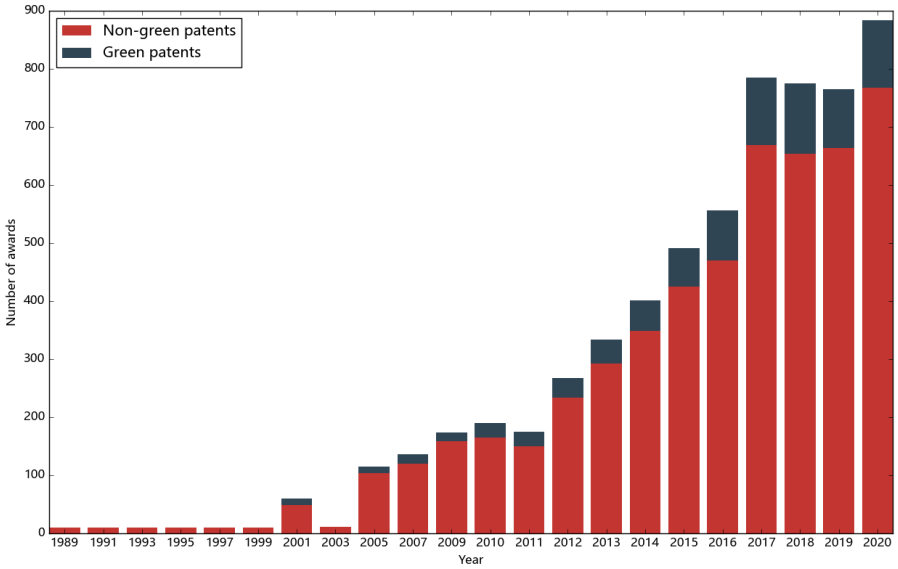


Figure 2. Number of green patents awarded in the Chinese Patent Award by session.

The Chinese Patent Award has three categories of awards: gold, silver, and excellent award. The reasons why we do not distinguish the different types of awards are as follows. (1) The selected award-winning patents in this paper are the invention patents, not the utility model patents. Due to the value of the invention, patents are better than that of utility model patents and are more reflective of innovation. (2) We know that different category of awards means different values of patents. However, it is well-known that the category of award and the total number is not fixed in each session. For example, the silver award was awarded in 2018, hence the total volume is minimal. While the number of gold awards has increased from 10 to 30 per session. The number of patents awarded excellent awards is far more than the number of gold and silver awards. (3) The number of green patents studied in this paper is 878,007, and the number of green patents awarded with gold, silver, and excellent awards are 42, 27, and 770, respectively. The total number of awarded green patents and its percentage of the total number of green patents is 0.95‰, moreover, the percentage of both gold and silver awards is less than one ten-thousandth. Therefore, it is not meaningful to distinguish different categories of awarded patents. In addition, the purpose of this paper is to construct a patent heterogeneous network to identify significant patents earlier, so the effect of different types of *CPA* on the identification effect is not considered.

3.3. Data Analysis

The *CNGP* applicant-citation dataset has been constructed through the above-mentioned section, and it is necessary to analyze the obtained dataset in-depth. This dataset not only contains the citation information of patents but also the applicant information, it is beneficial to identify significant patents by using the characteristics of the heterogeneous information dataset.

Foremost, analyzing the distribution characteristics of patents as shown in Figure 3. From Figure 3a, it is found that the log-scale number of green patents almost increases linearly as time goes by (the point corresponding to the last two years can be ignored, since the data might be incomplete caused by data lag). The result shows that there is a continued focus on green technology innovation and development in China and that more patented products are being used in the environment and resource areas. In addition, we analyze the distribution of the total number of patents with the number of citations, the result as Figure 3b shown. This result indicates that most "normal" patents have few citations, while a few "seminal" patents have large citations. Such a network is called a scale-free network. Moreover, we calculated the average path length of this dataset, which

value is equal to 4.34, representing that it belongs to the small-world network (Six Degrees of Separation). Hence, those metric applicable to complex networks could be applied to our dataset as well.

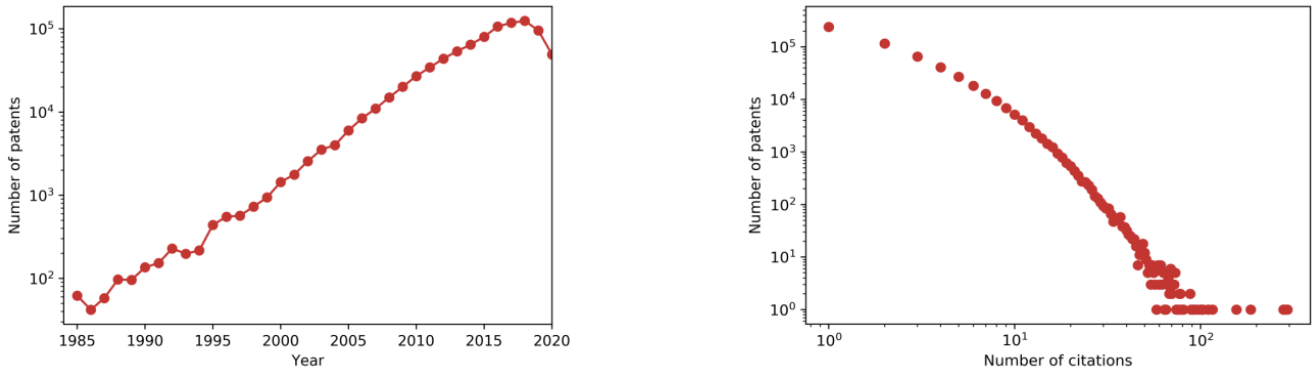


Figure 3. Distribution characteristics of the patents in the *CNGP* dataset. (**a**) The number of patents with time; (**b**) The number of patents with citations.

Many scholars applied the number of citation counts to assess the quality of patents [19,23]. We use the same method to analyze our dataset, the result of the top 10 patents ranked by the number of citations count as shown in Table 3. It includes the patent's rank, application number, title, application year, applicant, and the "count" refers to the citation count of the patent. The patent title clearly indicates that it belongs to the green patent product, and we can see from the year of application of patents, the top 10 patents are relatively old. It illustrates that, in the *CNGP* dataset, old patents have a longer time span to obtain citation relationships, and consequently old patents have more advantage over other young peers to obtain more citations. The more citation counts, the higher the value of the patent. This conclusion is consistent with the results of other literature [10,11].

Table 3. Top 10 patents ranked by the number of citation count in the *CNGP* dataset.

Rank	Number	Patent Title	Year	Applicant	Count
1	CN101665311A	Catalysis and micro-electrolysis combined technology for high-concentration refractory organic wastewater	2009	Central South University	293
2	CN1470327A	Metal nitride catalyst preparing method and catalyst	2002	Beijing Institute of Petrochemical Technology	285
3	CN105251527A	Composite molecular sieve and hydro-desulfurization catalyst prepared with composite molecular sieve as carrier	2015	China University of Petroleum Beijing	280
4	CN1262969A	Catalyst using TiO_2 as carrier to load metal nitride Mo_2N	2000	Nankai University	279
5	CN104229957A	Compound flocculant taking natural minerals as main components	2013	Zhang Jia-ling	186
6	CN101327976A	Efficient water treatment flocculant	2008	Nantong Liyuan Water Treatment Technology Co., Ltd.	156
7	CN101724474A	Mode for preparing M15-M100 complexing methanol petrol and gas normal-temperature liquefaction mixed combustion technology	2008	Chen Ruo-xin	116

Table 3. *Cont.*

Rank	Number	Patent Title	Year	Applicant	Count
8	CN105956923A	Asset transaction platform and digital certification and transaction method for assets	2016	Shanghai Dove Investment Co., Ltd.	110
9	CN102599161A	Hormone hydrogel rooting film	2012	Puyang Academy of Forest	103
10	CN101816305A	Bactericide pesticide composite containing oligochitosan	2010	Hainan Zhengye Zhongnong Hi Tech Co., Ltd.	101

We divided the applicant for patents into four types: enterprise, university, research institute, and individual. Figure 4 shows the proportion of different applicant types in our dataset, it illustrates that enterprises constitute 65.03% of the Chinese green patent applicants and are the largest component. The second-largest component is individuals, which account for 31.82% of applicants. Meanwhile, research institutes and universities constitute 2.01% and 1.15% of applicants, respectively. These findings show that in Chinese green patents, enterprises are dominant, followed by individuals, while research institutes and universities' participation is relatively small. The high proportion of enterprises and individuals illustrates that they are more inclined to transfer technological innovation into patents and could benefit from their own patents, such as by enhancing their core competitiveness. On the contrary, the proportion of research institutes and universities is smaller, because the number of them is limited, but their R&D capabilities cannot be ignored, so the analysis of applicants is particularly indispensable.

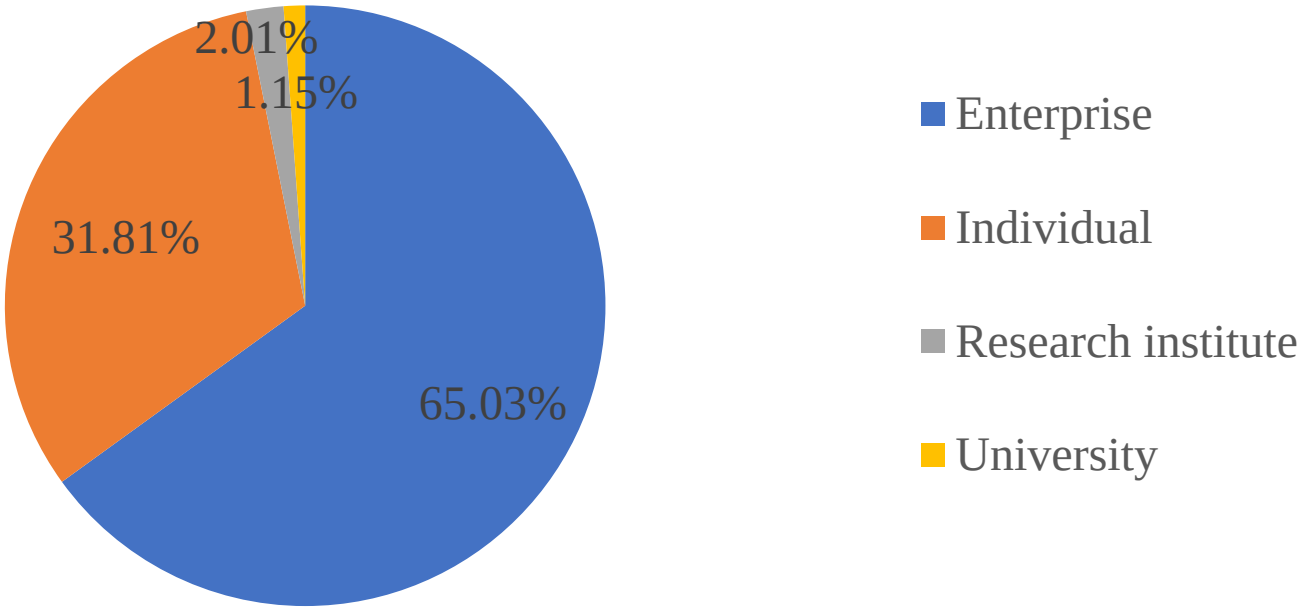


Figure 4. Proportion of different applicant types in the *CNGP* dataset.

Many studies rank applicants according to the number of patents the applicants own [33,55,56]. Therefore, we simply analyze the number of Chinese green patents owned by different applicants and found differences in the status of technological innovation, we do not consider whether it is a current name or a former name. The result is shown in Table 4, the "number" refers to the number of patents owned by a certain applicant, and the "average citations" indicates the average number of citations of all the patents invented by the applicant in this dataset.

Table 4. Top 10 applicant ranked by the number of Chinese green patents.

Rank	Applicant	Number	Average Citations
1	State Grid Corporation of China (former name)	21,266	2.58
2	China Petroleum & Chemical Corporation	10,181	2.05
3	Tsinghua University	5848	2.97
4	Zhejiang University	5438	2.69
5	State Grid Corporation of China	4873	0.69
6	Southeast University	4406	3.08
7	South China University of Technology	4117	2.71
8	Shanghai Jiaotong University	3712	2.88
9	SINOPEC Research Institute of Petroleum Processing	3431	2.17
10	China Electric Power Research Institute	3416	3.94

Table 4 illustrates that, in the Chinese green patent dataset, the State Grid Corporation of China has the largest number of patents, indicating it has a high level of R&D and a strong awareness of intellectual property protection. China Petroleum & Chemical Corporation follows and also has an important status in Chinese green patent innovation. Additionally, there are 3 enterprises, 5 universities, and 2 research institutes among the top 10 applicants. Individuals are not ranked in the top 10, implying that, although individual applicants comprise the proportion of 31.82%, the influence of individuals is weaker than that of organizations. Among the five universities, all belong to "Project 211 (a National Key Universities)", indicating that most green technological innovation activities are in well-known universities. Only larger state-owned enterprises will actively participate in green innovation, perhaps because green innovation belongs to a new field, and green patents are difficult to convert into market value, hence other enterprises are not paying enough attention to green patents. By applying for patents, universities and research institutes can not only improve their innovation capabilities, but also obtain considerable economic income through the transfer of patents. Therefore, universities and research institutes have become the main force of green invention patents. Moreover, we calculated the average number of patent citations for the applicants in this dataset as the value of 1.62, however, only the State Grid Corporation of China of the above Top 10 applications is smaller than the overall average, because it is a company that changed its name only in 2017 and has not enough time to obtain citations. In general, patent applicants with influence or authority in the field, such as the China Electric Power Research Institute and Southeast University, generally have the strength of technological monopoly, and their average patent citations are much higher than those of common applicants. Thus, the patents invented by these applicants may all contain high technological innovation and value.

In the past, lots of literature generally used the method of citation networks to analyze the importance of patents and did not consider the influence of applicant information on it. On the one hand, the homogeneous information network does not conform to the actual situation, and on the other hand, the obtained results are biased. Therefore, this paper intends to build a heterogeneous information network, which contains abundant information (e.g., applicants and citations) and is more feasible to study the importance of patents.

4. Methods

In this section, we build a heterogeneous applicant-citation network to represent the patent innovation data. Then we use four distinct citation-based ranking metrics that are described below and their rescaled variants to suppress the problem of age bias. Furthermore, we introduce the evaluation indicators of this study.

4.1. Heterogeneous Applicant-Citation Networks

The heterogeneous applicant-citation networks consist of two types of nodes, i.e., applicants and patents. There are two types of links, including the citation link between patents, and the applicant link between a patent and an applicant. We define the importance of patents is judged by analyzing the role they play in heterogeneous networks.

Given a set of applicants $\boldsymbol{a} = \{a_1, a_2, \ldots, a_m\}$ and a set of patents $\boldsymbol{p} = \{p_1, p_2, \ldots, p_m\}$. Let E_{PP} denote the citation links between patents; E_{PA} denote the applicant relation between a patent and an applicant. The heterogeneous applicant-citation network is a graph $G = (\boldsymbol{a} \cup \boldsymbol{p}, E_{PP} \cup E_{PA})$. A network contains m applicants and n patents, the graph can be represented by a binary $(m+n) \times (m+n)$ adjacency matric A:

$$A = \begin{pmatrix} A_{PP} & A_{PA} \\ A_{AP} & 0 \end{pmatrix} \tag{1}$$

where A_{PP} is the citation matrix between patents, A_{PA} and A_{AP} represent patent-applicant relations where $A_{PA} = A_{AP}^T$, and $A_{ij} = 1$ if node i points to node j. Our goal is to obtain a vector r for the network G, where r can reflect the importance of patents p. The proposed networks is shown in Figure 5.

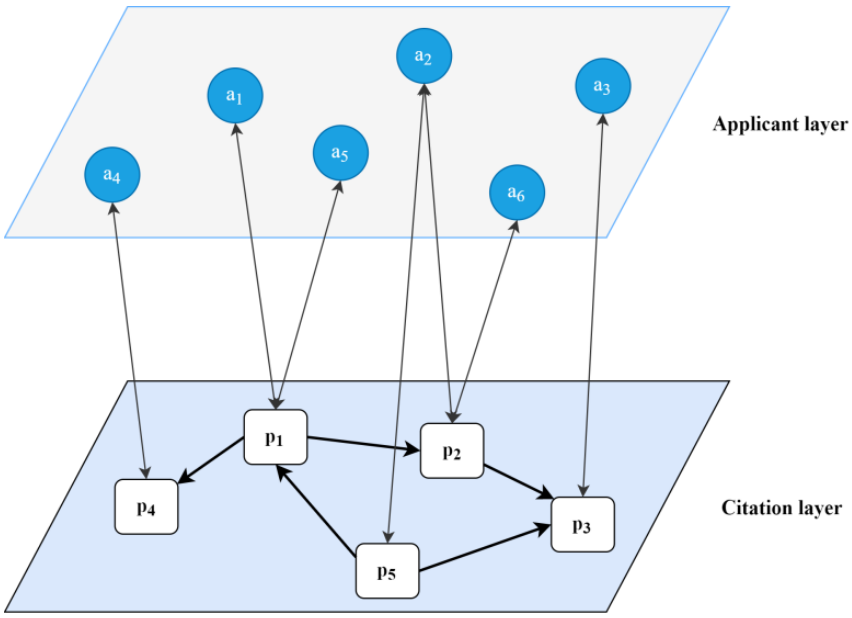


Figure 5. Heterogeneous applicant-citation networks structure.

As can be seen from Figure 5, the heterogeneous applicant-citation network structure consists of two layers, the applicant layer and the citation layer. The applicant layer includes all applicants in the dataset, the citation layer is the patents' citation network, where node denotes patent and edge denotes citation relationship. Linking applicant layer and citation layer by patent's application number. There are no edges between applicants, which differs from other paper/author networks such as the one proposed by Zhou et al. [57], Sun et al. [58], and West et al. [59]. As the links between applicants may make applicant social networks dominate the ranking system, in our study the ranking should be directed mainly by patents, rather than applicants' information. Therefore, we exclude the links between co-applicants in the applicant layer, and the undirected link of the applicant-patent is represented by a bidirectional link as Figure 5 shown. We treat the applicant-patent link as a citation link and use the citation-based ranking metrics to analyze the significance of patents on this network.

4.2. Citation-Based Ranking Metrics

From a network science perspective, citation-based ranking metrics constitute a key tool in scientometrics and play an increasingly important role in research evaluation [5]. In this section, we use four distinct network centrality metrics, and their variants where the age bias of metrics has been removed by the rescaling procedure introduced in Mariani et al. [10].

4.2.1. Citation Count (ID)

Citation count is one of the most commonly used metrics for evaluating a patent's impact, a patent with more citation count is considered to have a higher impact. For patent i, citation count is defined as $ID_i = \sum_j A_{ji}$, ID_i is referred to as the node i's in-degree in the network science language [11]. The aim of this metric is to mirror the impact and quality of patents. Therefore, ranking the patents by citation count assume that a patent is important if it is pointed by many other nodes [60].

4.2.2. PageRank (PR)

PageRank [39] is an algorithm used by Google Search to rank web pages in their search engine. It is a way of measuring the importance of website pages, and is later applied to evaluate the significance of publications [12]. In a directed network composed of N nodes, the vector of the PageRank score $\{PR_i\}$ can be found as the stationary solution of the following set of recursive linear equations:

$$PR_i^{(t+1)} = d \sum_{j:k^{out}>0} \frac{a_{ji}}{k_j^{out}} PR_j^{(t)} + d \sum_{j:k^{out}=0} \frac{PR_j^{(t)}}{N} + \frac{(1-d)}{N} \tag{2}$$

where k_j^{out} is the out-degree of node j, d is the teleportation parameter, and t is the iteration number. Equation (2) represents node i by a random walker who with probability d follows the network's links and with probability 1-d teleports to a random node. The iterative process starts from the uniform score vector $PR_i^{(0)} = 1/N$ and stop iterating when its average score change is small enough, in which $\sum_i \left| PR_i^{(t)} - PR_i^{(t-1)} \right| < \varepsilon$ where we set $\varepsilon = 10^{-8}$ (the same stopping condition is used in LeaderRank metric). At the same time, we set $d = 0.5$ which is the common choice in citation networks [12,61].

4.2.3. LeaderRank (LR)

LeaderRank was introduced by Lü et al. [62] to identify influential users in networks. To rank the users, it adds a ground node which connects to every node through bidirectional links, we compute each node score by the iterative equation as:

$$LR_i^{(t+1)} = \sum_{j=1}^{N+1} \frac{a_{ji}}{k_j^{out}} LR_j^{(t)} \tag{3}$$

where $a_{ji} = 1$ if node j points to node i and 0 otherwise, k_j^{out} denotes the out-degree of node j. The initial scores are given by $LR_i^{(0)} = 1$ for all node i (while the ground node is equal to 0). The final score of node i to be the LeaderRank score LR defined as:

$$LR_i = LR_i^{(t_c)} + \frac{LR_g^{(t_c)}}{N} \tag{4}$$

where t_c is the convergence time and $LR_g^{(t_c)}$ is the score of the ground node at steady states. After calculating the LR scores of all nodes, sort the LR scores in descending order. The larger the LR score, the greater the importance of the nodes, thus the update order of the nodes is obtained.

4.2.4. Network Global Structure-Based Centrality (NGSC)

Network global structure-based centrality was introduced by Namtirtha et al. [41] to search the crucial nodes in complex networks. *NGSC* intelligently combines existing k-shell and the sum of neighbors' degree methods with knowledge of the network's global structured-based centrality. The *NGSC* score for node i as:

$$NGSC_i = \sum_{j \in N(i)} (tune_1 \times ks(i) + tune_2 \times k(i)) + (tune_1 \times ks(j) + tune_2 \times k(j)) \quad (5)$$

where $N(i)$ is the set of neighbor of node i, $ks(i)$ and $ks(j)$ are the k-shell indices of node i and node j, $k(i)$ and $k(j)$ are the degree of node i and node j, respectively. The range of the tunable parameters $tune_1$ and $tune_2$ lie between $[0, 1]$. In our directed networks, we focus on the in-neighbor and in-degree of patent i, and we set $tune_1 = tune_2 = 1$ in this study.

4.2.5. Rescaled Metric Variants

The strong age bias of the centrality metrics implies that nodes that appeared in some time periods are much more likely to rank than other nodes, independently of their properties such as novelty and significance. Mariani et al. [10] proposed the rescaling procedure to suppress the age bias of the ranking metrics, and we use this method in this study. The rescaled score $R(m_i)$ for metric m and node i is calculated by the z-score of metric m score for a group of nodes applied in a similar time as m_i:

$$R(m_i) = \frac{m_i - \mu_i(m)}{\sigma_i(m)} \quad (6)$$

where m_i is the original score of node i as produced by metric m, $\mu_i(m)$ is the metric mean value and $\sigma_i(m)$ is the metric standard deviation, they computed over node i's reference set. In our dataset, by labeling the patents in order of decreasing age, patent i's reference set is the set of $\triangle$ patents j such that $j \in [\max\{i - \triangle/2, 0\}, \min\{i + \triangle/2, n\}]$, we set $\Delta = 2000$ in this study.

4.3. Evaluation of the Metrics' Performance in Identifying the Significant Patents

To make quantitative statements on the ability of the metrics to identify the significant patents of different ages, we introduce two evaluation indicators: the identification rate and the average precision.

4.3.1. Identification Rate (IR)

The identification rate is an estimate of the probability that a subject is identified correctly at least at rank-N. We defined it as $f_z(m)$, which means that of a given metric m is defined as the fraction of significant patents that are ranked among the top z N patents by metric m. This quantity is commonly referred to as *recall* in the information filtering community. It is worth noting that $z \in (0, 1)$ is an evaluation parameter, and to reflect our goal of evaluating the ranking metrics by whether they rank the significant patents "highly", we set a small number $z = 5\%$ for all experiments.

First, we evaluate the identification rate on the complete heterogeneous networks. Then, we assess the ranking metrics' performance as a function of the age of significant patents. In this way, we could untangle the role of patent age in determining the metrics' performance, dissect the network evolution by constructing network snapshots at the end of each calendar year and rank all the nodes on each network snapshot. At each network snapshot computation time $t^{(c)}$, ignore all patents and citations that appear afterward, only preserve the patents applied before $t^{(c)}$ and corresponding citations and the corresponding applicant data. Individually for each significant patent i (applied at time t_i), we measure its age $\Delta t = t^{(c)} - t_i$ at each network snapshot analysis. 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 z N patents by metric m when they were Δt years old. For example,

$f_z(m;\Delta t=3)$ is the fraction of the significant patents that are in the top z fraction of the ranking when they are three years old.

The identification rate $f_z(m;\Delta t)$ is computed as:

$$f_z(m;k\Delta d)=\frac{1}{M(t)}\sum_{t^{(c)}}\sum_{t\in M}\delta\left(\left\lfloor\left(t^{(c)}-t(i)\right)/\Delta d\right\rfloor,k\right)\times\chi\left(\gamma\left(z,i;t^{(c)}\right)\leq z\right) \quad (7)$$

where $k=0,1,2,\ldots,20$, $\triangle d=365$ *days*, $M(t)$ denotes the distribution of the number of significant patents with age at least t years old at the end of the dataset, $\lfloor x\rfloor$ denotes rounding down to the nearest integer, $\delta(x,y)$ denotes the Kronecker delta function of x and y, $\chi\left(\gamma\left(z,i;t^{(c)}\right)\leq z\right)$ is equal to one if patent i is among the top $z\,N\left(t^{(c)}\right)$ patents in the ranking by metric m at time $t^{(c)}$, equal to zero otherwise.

4.3.2. Average Precision (AP)

The average precision is a value obtained by computing the average of the non-interpolated precision scores at each rank where a relevant entity is retrieved and therefore factors in precision at all *recall* levels [63]. However, for many applications only the top results are valuable, then we focus on the common indicator named $AP@n$, which is defined as follows:

$$AP@n=\frac{1}{min(|Rel|,n)}\cdot\sum_{k=1}^{n}P@k\cdot isrel(k) \quad (8)$$

where $|Rel|$ is the set of relevant significant patents, $min(|Rel|,n)$ refers to the total number of ground truth positives, n refers to the total number of patents you are interested in, $P@k$ refers to the precision at k, $isrel(k)$ is a binary function that returns one if the significant patents at rank k are relevant or equals to zero otherwise. The score of $AP@n$ measures how many relevant significant patents are returned by a query at the top n ranked patents and their average precision. For instance, if we consider five significant patents, according to their citation counts, are ranked at positions 1, 3, 6, 8, and 12 in a list of otherwise non-significant patents. The AP@10 for citation counts using the 5 significant patents as the test set would be $(1/1+2/3+3/6+4/8)/5=0.53$. Note that score of AP is biased toward the top of the rankings.

The score of AP is a statistical index in bibliometrics, which is an approximation of the area under the precision-recall curve. In machine learning, $AP@n$ seems to be the most stable under varying cut-off thresholds n. In order to be consistent with the identification rate, we rank the network patents by their score according to a given metric m and compute the score of AP that are among the top $z\,N$ patents, then we focus on the ranking positions of the significant patents, and these are reflected by $AP_z(m)$ and $AP_z(m;\Delta t)$ as similarity description above.

In this study, all the computations were performed in the PYTHON programming environment, version 3.7.3, with a processor of Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz (6 CPUs) and 16 GB RAM on a Windows 10 environment.

5. Results

5.1. Metrics' Performance on the Complete Heterogeneous Networks

We start by measuring the identification rate and the average precision of all the metrics on the complete heterogeneous applicant-citation networks, where uses $z=5\%$ to single out the significant patents. The colors of the bars are used to distinguish the original ranking metrics (red) and their age-rescaled counterparts (dark).

Figure 6 shows that all the original metrics have a similar performance in identifying the significant patents on the complete heterogeneous networks, besides their rescaled metrics as well. However, the scores of *IR* and *AP* are not very high. The reason may be shown that significant patents are difficult to separate from the other patents in our dataset. In addition, from the identification rate in Figure 6a, we can notice that: (1) *ID* is the best

performing metric with a small over the others original metrics and a large margin over all rescaled metrics. (2) The ratio between the best and the worst metric's *IR* score is 2.05. (3) All rescaled metrics perform significantly worse than their non-rescaled counterparts. In Figure 6b, all the metrics' performance of the score of *AP* are quite small. The reason is that the Top *z N* data contained a small proportion of significant patents (as shown in Figure 3b), and its ranking is also a very important factor, so the resulting score appears to be very small. In addition, *PR* and *R_ID* perform better than other metrics on the scores of *AP*, respectively.

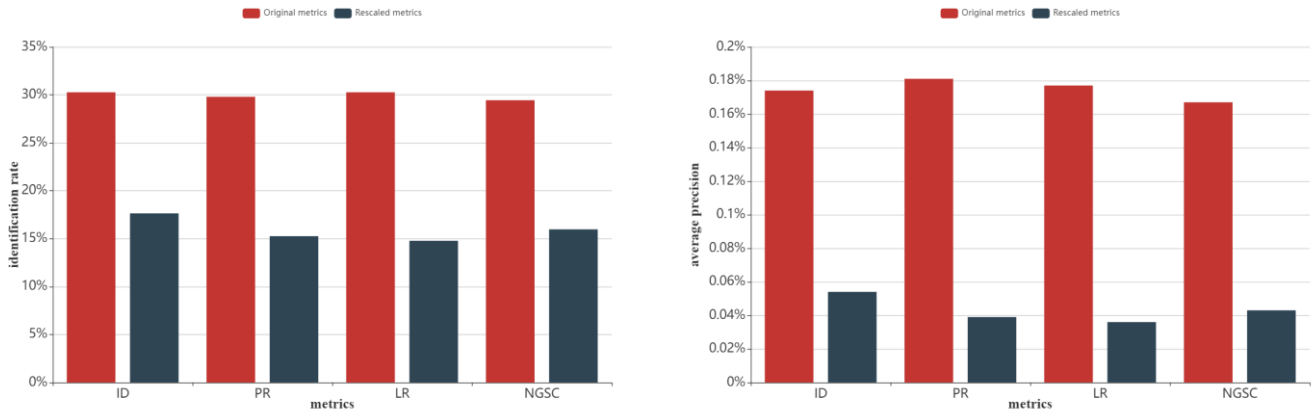


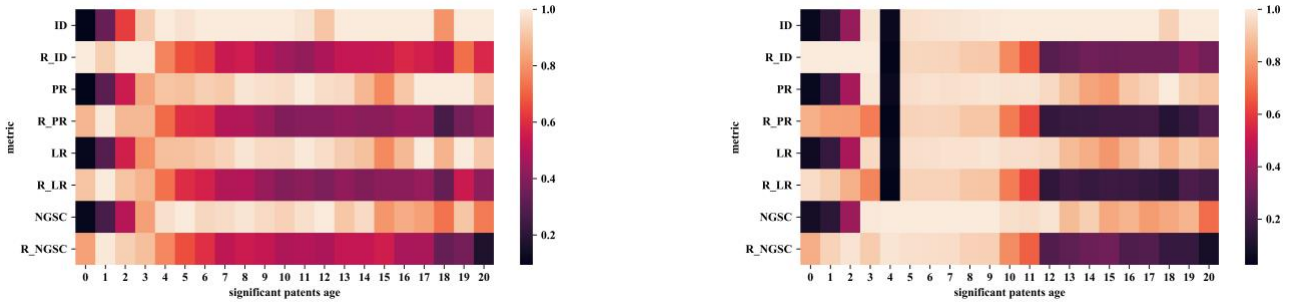
Figure 6. Metrics' performance in identifying the significant patents on the complete networks. (**a**) Identification rate; (**b**) Average precision.

From the above analysis, we know that a better metric for the analysis of complete heterogeneous networks does not exist. The original metrics get better performance over the rescaled metrics, while it contains age-biased ranking metrics. Therefore, it is crucial to analyze the effect of the patents' age by dissecting the complete networks.

5.2. Metrics' Performance with Patents Age

Although the analysis in the previous section reveals the important differences among all the ranking metrics, the main objective of this paper is to reveal the dependence of the ranking metrics' performance as a function of patent age and evaluate the ability of metrics to identify the significant patents earlier. Therefore, we dissect the network evolution by constructing network snapshots at the end of each calendar year and ranking all the nodes on each network snapshot, the detailed steps refer to Section 4.2.1. The results, as Figure 7 shows, show the metrics' performance as a function of the age of the significant patent. This method is used to reveal the time evolution of the metrics' performance. To facilitate the comparison between the obtained results, the performance of each metric was normalized to the best metric in each age bin. Specifically, it means that there is the best metric that receives the best score in each age bin. For instance, a metric with zero *IR* then obtains a zero score, while a metric that achieves the best *IR* for given significant patents age obtains one score. In addition, the same process is used for *AP* as well.

As shown in Figure 7a, the relative performance of metrics changes dramatically with the age of the significant patent. All the rescaled metrics that work well shortly after the date of application lose their advantage as the significant patents become older, then the original metrics perform better. In our study, there is no single metric that performs well for most age values. Rescaled indegree (*R_ID*) is better until age 3, then indegree (*ID*) is better until age 10, LeaderRank (*LR*) is the best until age 13, and then *ID* and *PR* are better until age 20 take place by turn. However, Figure 7b indicates that the *R_ID* metric is the best until age 2, *NGSC* is the best until age 9, and *ID* is better from then until age 20. The rescaled metrics perform worse if the significant patent's age is more than 12 years old.



(**a**) (**b**)

Figure 7. Metrics' relative performance as a function of the significant patents age. (**a**) Identification rate; (**b**) Average precision.

The above analysis shows that the rescaled metrics can earlier identify the significant green patents in the dataset. With the patent's age increasing, the rescaled metrics' scores are far less than the original metrics. This proves the validity of the rescaling procedure method, and analyzes the influence of the existence of the patent's age bias on the ranking centrality results.

5.3. Further Explanation about the Results

To further understand the differences between the ranking metrics, we used the Spearman ranking correlation of all patents' rankings to assess their pairwise similarity. In statistics, Spearman's ρ is a non-parametric measure of ranking correlation. This method is based on the *L1* distance of the ranks of patents in two ranked lists and provides a quantitative measure to compare how similar these lists are. The value is $\rho \in [-1, 1]$, the higher the absolute value, the better the similarity between the two rankings, and the value is 0 denoting there is no correlation. It can be computed using the popular formula like that:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} (R(X_i) - R(Y_i))^2}{N(N^2 - 1)} \tag{9}$$

where N denotes the number of patents, $R(X_i)$ and $R(Y_i)$ are the position of patent i in different ranked lists.

In addition, we use hierarchical clustering analysis to cluster the results. It is an algorithm that groups similar objects into groups. The metrics' hierarchical clustering is obtained by the unweighted pair-group method with the arithmetic means (*UPGMA*) method. The result is shown in Figure 8 together with metric clustering based on the received correlation metrics.

From Figure 8, the following points should be noted. (1) The clustering of metrics is extremely stable in the *CNGP* dataset. (2) The hierarchical clustering revealed that there are two groups of metrics that were ranked similarly to each other. The larger group includes four ranking metrics: *PR*, *LR*, *ID*, and *NGSC*. The smaller group includes some of their rescaled variants: *R_ID*, *R_PR*, and *R_LR*. (3) However, *R_NGSC* is not clustered with other rescaled metrics, probably because the rescaling procedure has no effect on the *NGSC* metric. (4) Within each of the two mentioned clusters, the pairwise Spearman's rank correlation coefficients are rather higher (above 0.78 in our dataset), which indicates a high degree of similarity among the respective metrics.

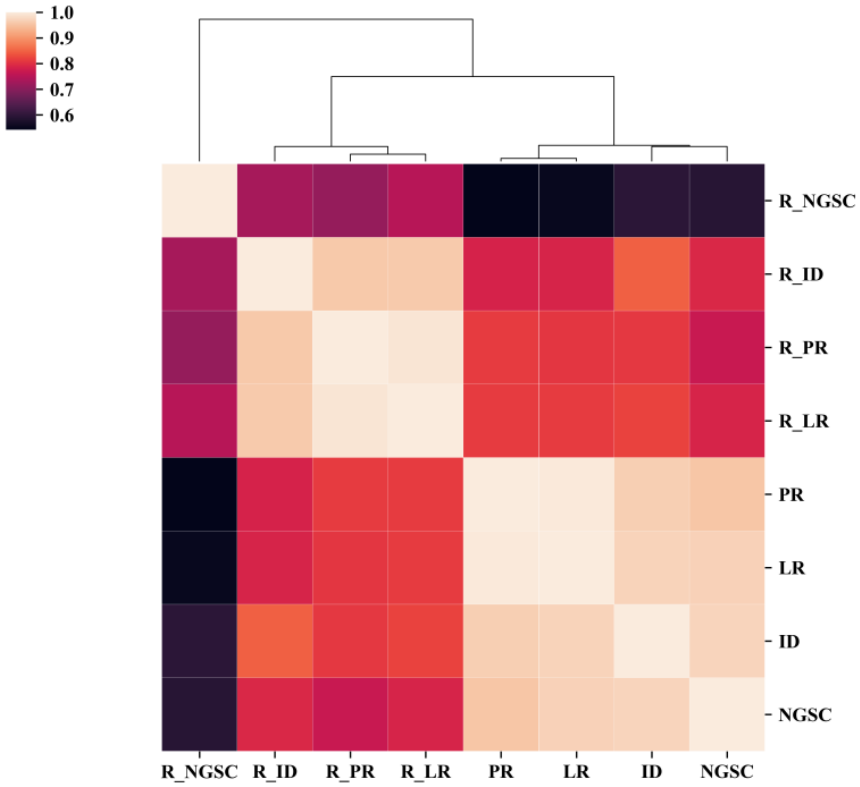


Figure 8. Similarity of the evaluated metrics as measured by the Spearman's rank correlation.

5.4. Caveats of the Evaluation Indicators

As shown in Figure 6 above, the ranking metrics are not good enough in the evaluation indicators of *IR* and *AP* assessment performance. In order to illustrate where they come from, we explore the age distribution of the significant patents in our dataset. First, check the distribution of the significant green patents according to their application date. Second, sort all green patents according to the application date and divide them into 40 equally-sized age groups, and count the distribution number of the significant patents according to groups, the results as Figure 9 shown.

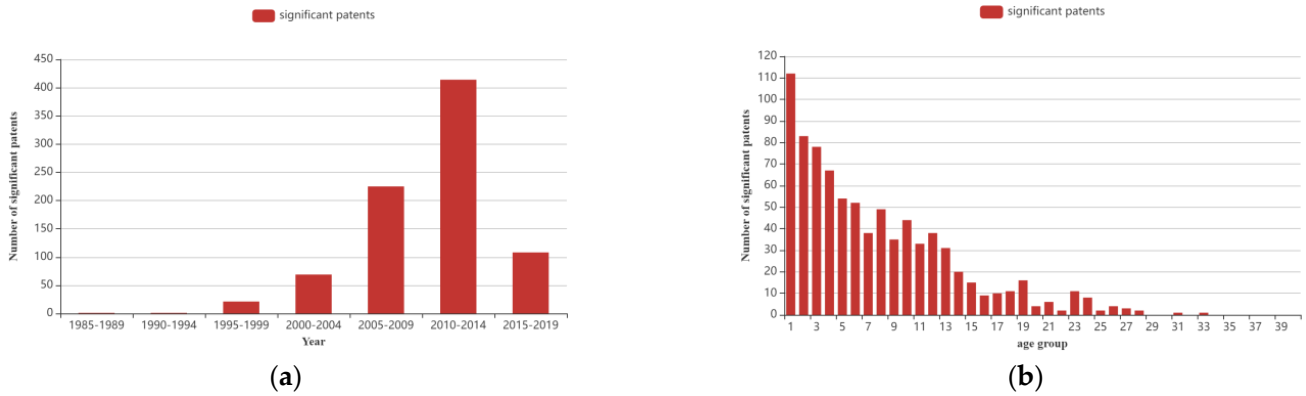


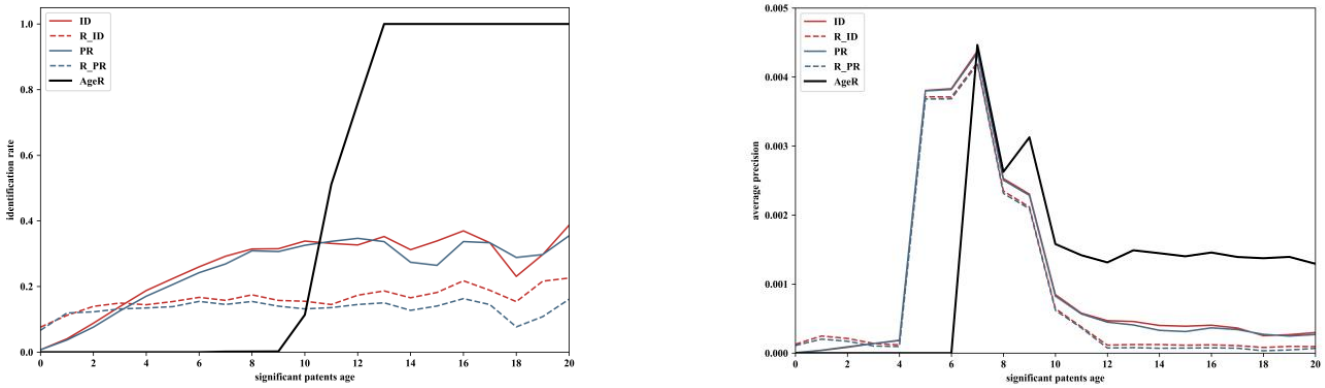
Figure 9. Distributions of the significant patents in the *CNGP* dataset. (**a**) Patents application date; (**b**) 40 equally-sized age groups.

In Figure 9a, we count the distribution of the significant patents in a span of five years, which shows the significant patents' application dates mostly in the year scope of 2010–2014. Although before 2004, the time span was as long as 20 years, the proportion of the significant patents was only 10.96%, and the analysis found that the average application

year is 2010. In addition, we sorted the complete dataset by patents' application date, and split them into 40 equally-sized age groups (with groups 1 and 40 containing the oldest and the most recent patents, respectively). As Figure 9b applies that the significant patents are distributed unevenly among the age groups. Most of them were distributed in the older age groups (about 73% in age group 1 to age group 10), and even less than 10 significant patents in the age group 24 to age group 40. Compare with Figure 9a,b, the reason for the difference between them is the exponential increase in the number of new patents application each year (as shown in Figure 3a). The number of recent new patents is so much larger that they "push" the significant patents to the earlier age groups, resulting in an uneven distribution, as shown in Figure 9b.

According to the above analysis, the strong temporal non-uniformity of the significant patents can have decisive consequences. First, it is not beneficial to age-rescaled metrics, because the rescaled procedure method strives for a uniform representation of all age groups among the top-ranked patents. In addition, for the dataset in this paper, patents from age groups 20 to 40 can contribute only marginally to the evaluation scores of *IR* and *AP*, where there are only a few significant patents among them. By contrast, original non-rescaled metrics are generally biased towards older patents, and such metrics are more advantageous when a given set of significant patents has the same bias towards older patents.

The age bias of the significant patents in the *CNGP* dataset is so strong that a simple ranking of nodes by age is achievable (we refer to this metric as Xu et al. [40] named *AgeR*, when ranking, old nodes are at the top) to outperform all other metrics. In this paper, we choose indegree (*ID*), PageRank (*PR*), and their rescaled metrics with *AgeR* in identifying the significant patents, and the results of metrics' performance as a function of the significant patents age as Figure 10 shown.



(**a**) (**b**)

Figure 10. Selected metrics' performance with the significant patents age. (**a**) Identification rate; (**b**) Average precision.

Figure 10a clearly illustrates the selected metrics' performance with the significant patents age in the evaluation indicator of *IR*, the rescaled metrics (*R_ID* and *R_PR*) get better performance while the significant patents age is younger than 3. When the significant patents' age is from 4 to 10, *ID* and *PR* have a better performance. The *AgeR* metric receives a score of *IR* zero when the significant patents are young, the reason for this is determined by the mechanism of this metric, which simply puts older patents at the top of the ranking. For example, as Figure 9a shows that before 2000, the number of significant patents is particularly rare, at only 23. However, as the size of the network increased, the advantages of adopting this metric continued to emerge, significant patents begin to be identified continuously at the age of 9, and the *AgeR* algorithm becomes the best metric after age 11. Moreover, if the age of the significant patents is more than 13 years, the *AgeR* metric can identify 100% of them. This suggests that evaluating ranking metrics by their ranking of

the significant patents is of limited relevance. For instance, *AgeR* which metric completely ignores the actual influence of the patents, is finally able to outperform the other ranking metrics. In addition, Figure 10b shows the results of the *AP* evaluation indicator, which shows that the rescaled metrics exhibit better performance when the age of the significant patents is younger than 3. The original metrics' were better than the rescaled metrics' after that. After the significant patents age more than 9, *AgeR* becomes the best metric, and the effect is absolutely significant as well.

5.5. Penalizing Age-Biased Metrics

From the above analysis, the age bias is implicit in the selected set of significant patents by experts. In this section, we need to apply an additional penalty for biased metrics when we use the indicators of *IR* and *AP* to evaluate the performance of ranking metrics. We adopt the normalized method as Mariani et al. [10] applied which imposes a penalty on metrics that are age-biased. The specific description process could refer to the literature written by Xu et al. [40]. In this paper, we're measuring the evaluation indicators of *NIR* and *NAP* for the metrics' performance on the complete dataset and the metrics' performance as a function of the age of the significant patents as well.

To define the *NIR* $\widetilde{f}_z(m;\Delta t)$ of a metric, at each computation time $t^{(c)}$ we divide the $N\left(t^{(c)}\right)$ patents into 40 groups according to their age. The *NIR* of metric m is defined as:

$$\widetilde{f}_z(m,k\triangle d)=\frac{1}{M(t)}\sum_{t^{(c)}}\sum_{t\in M}\delta\left(\left\lfloor\left(t^{(c)}-t(i)\right)/\triangle d\right\rfloor,k\right)\times\chi\left(\gamma\left(z,i;t^{(c)}\right)\leq z\right)y\left(n\left(m,i;t^{(c)}\right)\right) \tag{10}$$

where the above variables are denoted the same as those described in Equation (7), and $y\left(n\left(m,i;t^{(c)}\right)\right)$ is a decreasing function of the fraction $n\left(m,i;t^{(c)}\right)$ of patents that belong to the same age group of patent i and are ranked among the top z $N\left(t^{(c)}\right)$ by metrics. Denoting by $n_0\left(i;t^{(c)}\right)=1/40$ the expected value of $n(\cdot,i;t^{(c)})$ for an unbiased ranking, which is defined as follows:

$$y\left(n\left(m,i;t^{(c)}\right)\right)=\begin{cases}\left(n\left(m,i;t^{(c)}\right)/n_0\left(i;t^{(c)}\right)\right)^{-1} & ,if\ n\left(m,i;t^{(c)}\right)>n_0\left(i;t^{(c)}\right)\\ 1 & ,\ otherwise\end{cases} \tag{11}$$

Besides, the similar operation is used for *NAP* as well.

Figure 11 shows the performance of using the normalized method for evaluating selected metrics with the significant patents age, the results show that using this method indeed solves the problem encountered when using the ordinary evaluation metrics, such as in Figure 10 *AgeR* metrics ultimately outperform others selected metrics. However, whilst penalizing age-biased metrics, *AgeR* becomes the worst metric regardless of the significant patents age, which applies to ranking metrics that actually ignore the impact of patents in the network. With the function of the significant patents age, the rescaled metrics mostly get better performance than the original metrics. This suggests that the use of the normalized procedure weakens the mutually reinforcing link between the age-biased ranking metric and the age-biased sets of significant patents.

The reason for applying normalized processes, which is a significant patent, is just the tip of the iceberg of high-quality patents, it is inevitable to ignore other significant patents that are not identified by experts. In summary, the normalized method corresponds to the task of ranking the best patents for each age group, where the given significant patents are a potentially biased sample. Then, we assess the performance of the metrics using the normalized process on the complete heterogeneous networks as Figure 12 shown.

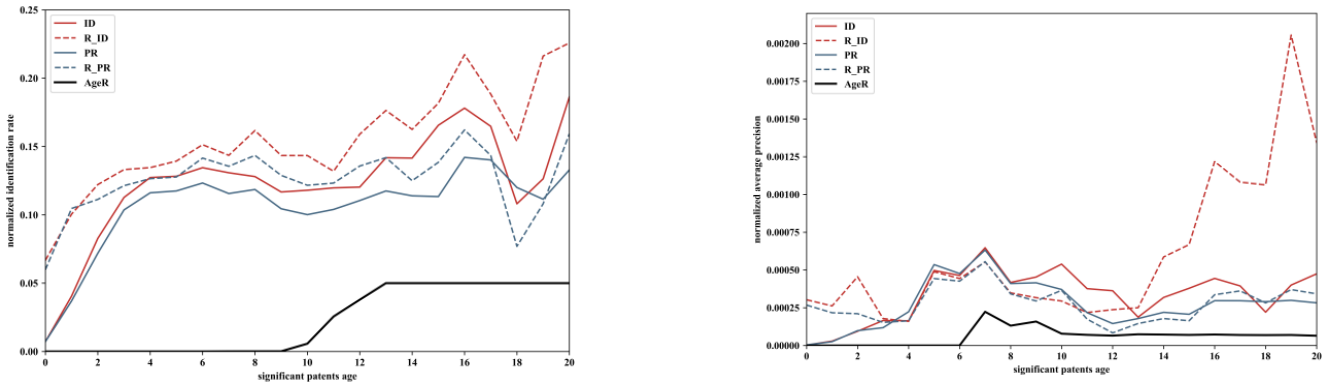


Figure 11. Normalized method to penalize age-biased metrics with the significant patents age. (**a**) Normalized identification rate; (**b**) Normalized average precision.

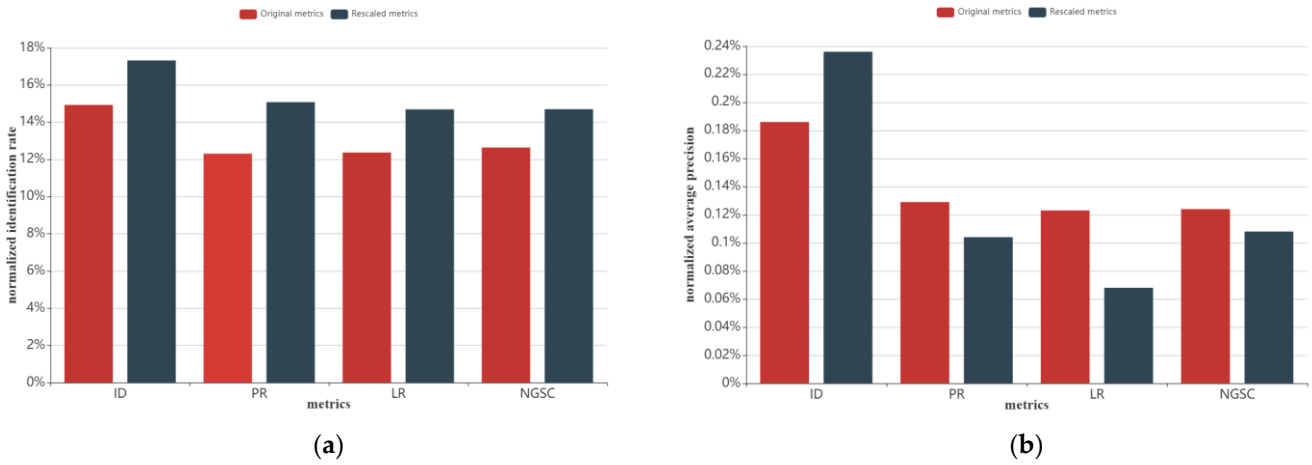


Figure 12. Metrics' performance by using normalized method on the complete heterogeneous networks. (**a**) Normalized identification rate; (**b**) Normalized average precision.

Figure 12 shows the ranking metrics evaluated by their *NIR* and *NAP* on the complete heterogeneous networks. We observe something from Figure 12a as follows. (1) The rescaled metrics generally perform better than their original counterparts here. (2) The *NIR* scores are much lower than the previously reported *IR* scores (From Figure 6a). This is a direct effect of the penalty approach introduced by *NIR* that severely penalizes biased ranking metrics and, unbiased ranking metrics are not good at identifying the biased significant patents. Figure 12b shows that *R_ID* has the best performance than other metrics. However, the *NAP* scores of *PR*, *LR*, and *NGSC* are not better than their rescaled metrics. Through comparison with their original performance, we find out the original metrics of *NAP* are generally lower than *AP* values, while the rescaled metrics of *NAP* are much better than *AP* values. The reason for this result may be the rescaled metrics can reduce the age-biased influence in our dataset.

It can be seen from the above analysis that the combination of rescaled metrics and normalized processing methods can effectively suppress the age deviation of patents and the age-biased cause by expert-selected significant patents, which is conducive to identifying significant patents at an earlier stage. From Figure 12 we know that *R_ID* has better performance both in *IR* and *AP* evaluation indicators, hence the results of the Top 10 patents as ranked by *R_ID* score, as Table 5 shows.

Table 5. Top 10 patents as ranked by rescaled in-degree (*R_ID*) score.

Rank	Number	Patent Title	Year	*ID*	*R_ID*
1	CN105251527A	Composite molecular sieve and hydro-desulfurization catalyst prepared with composite molecular sieve as carrier	2015	280	52.29
2	CN1470327A	Metal nitride catalyst preparing method and catalyst	2002	285	49.14
3	CN1262969A	Catalyst using TiO_2 as carrier to load metal nitride Mo_2N	2000	279	48.79
4	CN101665311A	Catalysis and micro-electrolysis combined technology for high-concentration refractory organic wastewater	2009	293	44.1
5	CN104229957A	Compound flocculant taking natural minerals as main components	2013	186	39.08
6	CN105956923A	Asset transaction platform and digital certification and transaction method for assets	2016	110	33.04
7	CN101327976A	Efficient water treatment flocculants	2008	156	28.49
8	CN109754605A	A kind of traffic forecast method based on attention temporal diagram convolutional network	2019	32	26.94
9	CN108010360A	A kind of automatic Pilot context aware systems based on bus or train route collaboration	2017	49	26.38
10	CN104992364A	Unattended electric car rental system and rental method	2015	89	26.07

Table 5 shows that the top 10 patents by *R_ID* span a wider temporal range (2000–2019) than the top 10 by *ID* (2000–2016) in Table 3, which is a direct result of the age-bias removal. In the meantime, this method can also identify those significant patents earlier, thereby contributing to sustainable development.

6. Discussion

To verify the validity of the heterogeneous applicant-citation networks built in this paper, we analyze the homogeneous patent citation network as well. This homogeneous information network only contains citation relationships of patents, not including applicant information. We use the same ranking metrics as described in Section 4.2 to evaluate the complete network. The results of the comparison of the metrics' performance in identifying the significant patents on the heterogeneous and homogeneous information network as Table 6 shown.

Table 6. Comparison of the metrics' performance on different information network.

	Heterogeneous Network (%)				Homogeneous Network (%)			
Metric	***IR***	***AP***	***NIR***	***NAP***	***IR***	***AP***	***NIR***	***NAP***
ID	30.27	0.174	14.93	0.186	26.34	0.157	11.95	0.149
R_ID	17.64	0.054	17.32	0.236	15.49	0.048	15.36	0.198
PR	29.80	0.181	12.31	0.129	26.58	0.166	9.91	0.113
R_PR	15.26	0.039	15.08	0.104	11.68	0.028	11.57	0.183
LR	30.27	0.177	12.37	0.123	26.94	0.160	10.22	0.112
R_LR	14.78	0.036	14.69	0.068	10.61	0.026	10.56	0.039
NGSC	29.44	0.167	12.64	0.124	25.39	0.149	9.90	0.112
R_NGSC	15.97	0.043	14.70	0.108	12.99	0.035	12.76	0.219

From Table 6 we can conclude that the heterogeneous network has better performance than the homogeneous networks under the same conditions. For instance, the *IR* scores of the heterogeneous network are at least 12% larger than the *IR* scores of the homogeneous network. In addition, after using the normalized method to penalize age-biased, the conclusion remains the same. Therefore, it can be seen from the above that the performance of the heterogeneous information network obtained by adding the applicant's information to the patent citation network is better than that of the homogeneous information network. It is because those influential applicants attracted more attention in the net-

work, and those significant patents can be better identified by analyzing the heterogeneous information network.

Since most of the previous literature identified significant patents by constructing patent citation network, for example, Mariani et al. [31] and Xu et al. [40] analyzed US patents and found that the *NIR* scores of rescaled_PageRank and rescaled_LeaderRank in the static network in the top 1% rankings were about 38%, while our results have the best *NIR* score is 15% in the top 5% rankings, which is surprisingly a huge difference. We found two possible reasons through analysis: (1) the citation relationship of Chinese patents is non-compulsory disclosure, so the constructed patent network is sparser and harder to identify by the centrality metrics than the US patent network; (2) the number of expert-selected significant patents we used is more than those literature used, and the method of selecting those significant patents is also different. Since their dataset does not contain information on patent applicants, the model proposed in this paper cannot be used to construct a heterogeneous information network to compare the performance of the Chinese and the US patent datasets horizontally. In conclusion, when we analyze a problem, we not only need to consider the importance of the algorithm but also need to deeply analyze and explore the impact caused by the original data on the results.

All the abbreviations and variables in this paper are shown in below.

7. Conclusions

In this study, we based on the Chinese green patent dataset from 1985 to 2020, and construct a *CNGP* heterogeneous applicant-citation network for identifying expert-based significant patents earlier. We use the rescaled method to suppress the age bias in citation-based ranking metrics, and construct static and dynamic citation networks to more comprehensively analyze the impact of patent age. To analyze the reasons for the poor model performance, we deeply analyze the source data and find that there is a strong age distribution bias in the expert-selected significant patents, so we use the normalized method to penalize the age bias of the evaluation indicator and obtain a reasonable evaluation performance. The experimental results show that the *R_ID* metric has the best performance and identifies significant patents earlier. In addition, compared to the patent citation network, the heterogeneous information network constructed by combining patent applicants is beneficial to improve the performance of identifying significant patents. Therefore, the analysis method in this paper not only evaluates the patent quality reasonably but also identifies significant patents earlier, which provides scientists with new methods to measure the importance of patents.

There are three major directions for extending this research. (1) When building heterogeneous applicant-citation networks, the citation layer excludes those patents that have no citation relationship. These patents may be old patents or newly applied patents belonging to isolated nodes. Other literature generally denotes that those patents have not been cited, indicating that the quality of those patents is very low. The above operations may cause bias in network analysis. Therefore, when building the heterogeneous information network, these patents need to be taken into account for further research. (2) In the heterogeneous application-citation networks analysis, we transform each applicant-citation relationship into bidirectional links and treat them as simple citation links. We use the unweighted links method to analyze this network. However, the actual situation should be more complicated than that, we can design a set of weight distribution principles to calculate the weight of applicant-citation links and patent citation links. In addition, we can try to use other ranking metrics to analyze the network. (3) As we know, *WIPO* divides green patents into seven categories, we should consider both the age bias and category bias of the analyzed ranking metrics as done by Vaccario et al. [64]. Since the number of patents in different categories varies greatly, for example, the category of waste management accounts for 25%, and nuclear power generation only for 1%, so it is extremely significant to add the green patent category bias to the analysis.

Author Contributions: Conceptualization, X.L. (Xipeng Liu) and X.L. (Xinmiao Li); methodology, X.L. (Xipeng Liu); soft-ware, X.L. (Xipeng Liu); validation, X.L. (Xinmiao Li) and X.L. (Xipeng Liu); formal analysis, X.L. (Xinmiao Li); investigation, X.L. (Xinmiao Li); resources, X.L. (Xipeng Liu); data curation, X.L. (Xipeng Liu); writing—original draft preparation, X.L. (Xinmiao Li); writing—review and editing, X.L. (Xipeng Liu); visualization, X.L. (Xinmiao Li); supervision, X.L. (Xinmiao Li); project administration, X.L. (Xipeng Liu); funding acquisition, X.L. (Xipeng Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shanghai University of Finance and Economics Postgraduate Innovation Fund Project (CXJJ-2021-396).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: Thanks to everyone who contributed to this paper. We would also like to thank the reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations and Variables

Abbreviation	Specific Name
AgeR	Ranking by age
AP	Average Precision
CNGP	Chinese Green Patents
CNIPA	China National Intellectual Property Administration
CPA	Chinese Patent Award
ID	Citation Count (In-degree)
IPC	International Patent Classification
IR	Identification Rate
LR	LeaderRank
NAP	Normalized Average Precision
NGSC	Network Global Structure-based Centrality
NIR	Normalized Identification Rate
PR	PageRank
R_ID	Rescaled indegree
R_LR	Rescaled LeaderRank
R_NGSC	Rescaled Network Global Structure-based Centrality
R_PR	Rescaled PageRank
R&D	Research and Development
UPGMA	Unweighted Pair-Group Method with Arithmetic means
WIPO	World Intellectual Property Organization
Variable	**Definition**
$AP@n$	The average precision score of the significant patents rankings that are in the top n
ID_i	The number of in-degree of node i
LR_i	The LeaderRank score of node i
$NGSC_i$	The NGSC score of node i
PR_i	The PageRank score of node i
$R(m_i)$	The rescaled score for metric m and node i
$f_z(m)$	The fraction of the significant patents that are in the top zN nodes by a given metric m ranking score.

$f_z(m;\Delta t)$	The fraction of the significant patents that are in the top zN patents by a given metric m ranking score when they are Δt years old.
$\widetilde{f}_z(m;\Delta t)$	The fraction of the significant patents that are in the top zN patents by a given metric m ranking score when they are Δt years old, and which adopt the normalized method to penalize the age-biased metric.

References

1. Chai, K.C.; Yang, Y.; Sui, Z.; Chang, K.C. Determinants of highly-cited green patents: The perspective of network characteristics. *PLoS ONE* **2020**, *15*, e0240679. [CrossRef]
2. Zhang, S.; Andrews, S.; Zhao, X.; He, Y. Interactions between renewable energy policy and renewable energy industrial policy: A critical analysis of China's policy approach to renewable energies. *Energy Policy* **2013**, *62*, 342–353. [CrossRef]
3. Hu, Y.; Jiang, H.; Zhong, Z. Impact of green credit on industrial structure in China: Theoretical mechanism and empirical analysis. *Environ. Sci. Pollut. Res.* **2020**, *27*, 10506–10519. [CrossRef] [PubMed]
4. Qi, S.; Lin, S.; Cui, J. Do environmental rights trading schemes induce green innovation? Evidence from listed firms in China. *China Econ. Res.* **2018**, *53*, 129–143.
5. Bornmann, L.; Williams, R. Can the journal impact factor be used as a criterion for the selection of junior researchers? A large-scale empirical study based on ResearcherID data. *J. Informetr.* **2017**, *11*, 788–799. [CrossRef]
6. Mariani, M.S.; Lü, L. Network-based ranking in social systems: Three challenges. *J. Phys. Complex.* **2020**, *1*, 011001. [CrossRef]
7. Zhou, Y.B.; Lü, L.; Li, M. Quantifying the influence of scientists and their publications: Distinguishing between prestige and popularity. *New J. Phys.* **2012**, *14*, 033033. [CrossRef]
8. Sun, Y.; Han, J. Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Sci. Technol.* **2013**, *18*, 329–338. [CrossRef]
9. Mariani, M.S.; Medo, M.; Zhang, Y.C. Ranking nodes in growing networks: When PageRank fails. *Sci. Rep.* **2015**, *5*, 16181. [CrossRef]
10. Mariani, M.S.; Medo, M.; Zhang, Y.C. Identification of milestone papers through time-balanced network centrality. *J. Informetr.* **2016**, *10*, 1207–1223. [CrossRef]
11. Newman, M. *Networks: An Introduction*; Oxford University Press: New York, NY, USA, 2010.
12. Chen, P.; Xie, H.; Maslov, S.; Redner, S. Finding scientific gems with Google's PageRank algorithm. *J. Informetr.* **2007**, *1*, 8–15. [CrossRef]
13. Walker, D.; Xie, H.; Yan, K.K.; Maslov, S. Ranking scientific publications using a model of network traffic. *J. Stat. Mech. Theory Exp.* **2007**, *6*, P06010. [CrossRef]
14. Luo, D.; Gong, C.; Hu, R.; Duan, L.; Ma, S. Ensemble enabled weighted PageRank. *arXiv* **2016**, arXiv:1604.05462. [CrossRef]
15. Lee, C.; Cho, Y.; Seol, H.; Park, Y. A stochastic patent citation analysis approach to assessing future technological impacts. *Technol. Forecast. Soc. Chang.* **2012**, *79*, 16–29. [CrossRef]
16. Lee, W.S.; Han, E.J.; Sohn, J.D. Predicting the pattern of technology convergence using bigdata technology on large-scale triadic patents. *Technol. Forecast. Soc. Chang.* **2015**, *100*, 317–329. [CrossRef]
17. Yoon, J.; Park, H.; Kim, K. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. *Scientometrics* **2013**, *94*, 313–331. [CrossRef]
18. Hsieh, H.C. Patent value assessment and commercialization strategy. *Technol. Forecast. Soc. Chang.* **2013**, *80*, 307–319. [CrossRef]
19. Harhoff, D.; Scherer, F.M.; Vopel, K. Citations, family size, opposition and the value of patent rights. *Res. Policy* **2003**, *32*, 1343–1363. [CrossRef]
20. Abbas, A.; Zhang, L.; Khan, S.U. A literature review on the state-of-the-art in patent analysis. *World Pat. Inf.* **2014**, *37*, 3–13. [CrossRef]
21. Kogan, L.; Papanikolaou, D.; Seru, A.; Stoffman, N. Technological innovation, resource allocation and growth. *Q. J. Econ.* **2017**, *132*, 665–712. [CrossRef]
22. Yuan, X.; Li, X. The evolution of the industrial value chain in China's high-speed rail driven by innovation policies: A patent analysis. *Technol. Forecast. Soc. Chang.* **2021**, *172*, 121054. [CrossRef]
23. Mezzanotti, F. Roadblock to innovation: The role of patent litigation in corporate R&D. *Manag. Sci.* **2021**, *67*, 7362–7390. [CrossRef]
24. Krestel, R.; Chikkamath, R.; Hewei, C.; Risch, J. A survey on deep learning for patent analysis. *World Pat. Inf.* **2021**, *65*, 102035. [CrossRef]
25. Zhou, Y.; Dong, F.; Liu, Y.; Ran, L. A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool. *Scientometrics* **2021**, *126*, 969–994. [CrossRef]
26. Lin, H.; Wang, H.; Du, D.; Wu, H.; Chang, B.; Chen, E. Patent quality valuation with deep learning models. In *International Conference on Database Systems for Advanced Applications*; Springer: Cham, Switzerland, 2018; pp. 474–490. [CrossRef]
27. Chung, P.; Sohn, S.Y. Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technol. Forecast. Soc. Chang.* **2020**, *158*, 120–146. [CrossRef]
28. Miguelez, E. Collaborative patents and the mobility of knowledge workers. *Technovation* **2019**, *86*, 62–74. [CrossRef]

29. Liu, W.; Song, Y.; Bi, K. Exploring the patent collaboration network of China's wind energy industry: A study based on patent data from CNIPA. *Renew. Sust. Energ. Rev.* **2021**, *144*, 110989. [CrossRef]
30. Érdi, P.; Makovi, K.; Somogyvári, Z.; Tobochnik, J.; Volf, P.; Zalányi, L. Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics* **2013**, *95*, 225–242. [CrossRef]
31. Mariani, M.S.; Medo, M.; Lafond, F. Early identification of important patents: Design and validation of citation network metrics. *Technol. Forecast. Soc. Chang.* **2019**, *146*, 644–654. [CrossRef]
32. Tseng, Y.H.; Lin, C.J.; Lin, Y.I. Text mining techniques for patent analysis. *Inform. Process. Manag.* **2007**, *43*, 1216–1247. [CrossRef]
33. Yoon, J. The evolution of South Korea's innovation system: Moving towards the triple helix model? *Scientometrics* **2015**, *104*, 265–293. [CrossRef]
34. Medo, M.; Cimini, G. Model-based evaluation of scientific impact indicators. *Phys. Rev. E* **2016**, *94*, 032312. [CrossRef]
35. Harzing, A.W.; Van Der Wal, R. A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *J. Am. Soc. For. Inf. Sci. Tec.* **2009**, *60*, 41–46. [CrossRef]
36. Karki, M.S. Patent citation analysis: A policy analysis tool. *World Pat. Inf.* **1997**, *19*, 269–272. [CrossRef]
37. Du, Y.P.; Yao, C.Q.; Li, N. Using heterogeneous patent network features to rank and discover influential inventors. *Front. Inf. Technol. Electron. Eng.* **2015**, *16*, 568–578. [CrossRef]
38. Zhao, F.; Zhang, Y.; Lu, J.; Shai, O. Measuring academic influence using heterogeneous author-citation networks. *Scientometrics* **2019**, *118*, 1119–1140. [CrossRef]
39. Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. [CrossRef]
40. Xu, S.; Mariani, M.S.; Lü, L.; Medo, M. Unbiased evaluation of ranking metrics reveals consistent performance in science and technology citation data. *J. Informetr.* **2020**, *14*, 101005. [CrossRef]
41. Namtirtha, A.; Dutta, A.; Dutta, B.; Sundararajan, A.; Simmhan, Y. Best influential spreaders identification using network global structural properties. *Sci. Rep.* **2021**, *11*, 2254. [CrossRef]
42. Newman, M. The first-mover advantage in scientific publication. *Europhys. Lett.* **2009**, *86*, 68001. [CrossRef]
43. Tan, Y.; Tian, X.; Zhang, X.; Zhao, H. The real effect of partial privatization on corporate innovation: Evidence from China's split share structure reform. *J. Corp. Financ.* **2020**, *64*, 101661. [CrossRef]
44. He, H.; Zhang, T.; Cai, Q. Import liberalization and Chinese firms' innovation—Evidence from patent quality and quantity. *China Econ. Q.* **2021**, *21*, 597–616. [CrossRef]
45. Sun, Z.; Lei, Z.; Wright, B.; Cohen, M.; Liu, T. Government targets, end-of-year patenting rush and innovative performance in China. *Nat. Biotechnol.* **2021**, *39*, 1068–1075. [CrossRef]
46. Lin, J.; Wu, H.M.; Wu, H. Could government lead the way? Evaluation of China's patent subsidy policy on patent quality. *China Econ. Rev.* **2021**, *69*, 101663. [CrossRef]
47. Moser, P.; Nichalos, T. Prizes, publicity and patents: Non-monetary awards as a mechanism to encourage innovation. *J. Ind. Econ.* **2013**, *61*, 763–788. [CrossRef]
48. Chen, K.Z. Research on the Value of Chinese Awarded Patent Based on Patent Index. Master's Thesis, Huazhong University of Science & Technology, Wuhan, China, 2019.
49. Yuan, H.X.; Ding, W.; Wu, Z.; Hu, Y.H. Analysis of Chinese patent award. *China Invent. Pat.* **2018**, *15*, 55–59. [CrossRef]
50. Jiang, K.L. Analysis and research of previous China's top patent awards. *China Invent. Pat.* **2018**, *15*, 49–54. [CrossRef]
51. Qiao, Y.Z.; Wang, Z.L. Research on the quality of invention patents winning the China patent gold award. *J. Intell.* **2018**, *37*, 120–125. [CrossRef]
52. Deng, S.M. Information analysis on regional layout of China's high value patents based on patent gold award. *J. Libr. Inf. Sci.* **2018**, *3*, 55–61. [CrossRef]
53. Liu, C.Y. Patent Quality Evaluation System Construction and Empirical Analysis: Based on the Comparative Collection of Chinese Patent Awards. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2020.
54. Zhen, S.; Li, H. Study on the construction of high-value patent evaluation index system of Chinese patent medicine. *China Pharm.* **2020**, *12*, 2152–2157. [CrossRef]
55. Zhu, R.; Liu, S. Identifying tacit university-industry collaborations in Chinese patents based on inventor-author analysis. *World Pat. Inf.* **2020**, *62*, 101986. [CrossRef]
56. Chang, S. Patent analysis of the critical technology network of semiconductor optical amplifiers. *Appl. Sci.* **2020**, *10*, 1552–1562. [CrossRef]
57. Zhou, D.; Orshanskiy, S.A.; Zha, H.; Giles, C.L. Co-ranking authors and documents in a heterogeneous network. In Proceedings of the Seventh IEEE International Conference on Data Mining, Omaha, NE, USA, 28–31 October 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 739–744. [CrossRef]
58. Sun, Y.; Han, J.; Zhao, P.; Yin, Z.; Cheng, H.; Wu, T. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology, Saint Petersburg, Russia, 24–26 March 2009; ACM Press: New York, NY, USA; pp. 565–576. [CrossRef]
59. West, J.D.; Jensen, M.C.; Dandrea, R.J.; Gordon, G.J.; Bergstrom, C.T. Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *J. Am. Soc. Inf. Sci. Tec.* **2013**, *64*, 787–801. [CrossRef]

60. Trajtenberg, M. A penny for you quotes: Patent citations and the value of innovations. *Rand. J. Econ.* **1990**, *21*, 172–187. [CrossRef]
61. Bruck, P.; Rethy, I.; Szente, J.; Tobochnik, J.; Erdi, P. Recognition of emerging technology trends: Class-selective study of citations in the US patent citation network. *Scientometrics* **2016**, *107*, 1465–1475. [CrossRef]
62. Lü, L.; Zhang, Y.C.; Yeung, C.H.; Zhou, T. Leaders in social networks, the delicious case. *PLoS ONE* **2011**, *6*, e201202. [CrossRef]
63. Dunaiski, M.; Geldenhuys, J.; Visser, W. On the interplay between normalization, bias, and performance of paper impact metrics. *J. Informetr.* **2019**, *13*, 270–290. [CrossRef]
64. Vaccario, G.; Medo, M.; Wider, N.; Mariani, M.S. Quantifying and suppressing ranking bias in a large citation network. *J. Informetr.* **2017**, *11*, 766–782. [CrossRef]