Gatekeeper: **Analyzing G-Indexes and Improving Service Quantification**

Jingtao Han*
Peking University
Haidian District, Beijing, China
hanjt@pku.edu.cn

Spyke Krepshaw
Penn State University
University Park, PA, USA
spyke@psu.edu

Dongwon Lee
Penn State University
University Park, PA, USA
dongwon@psu.edu

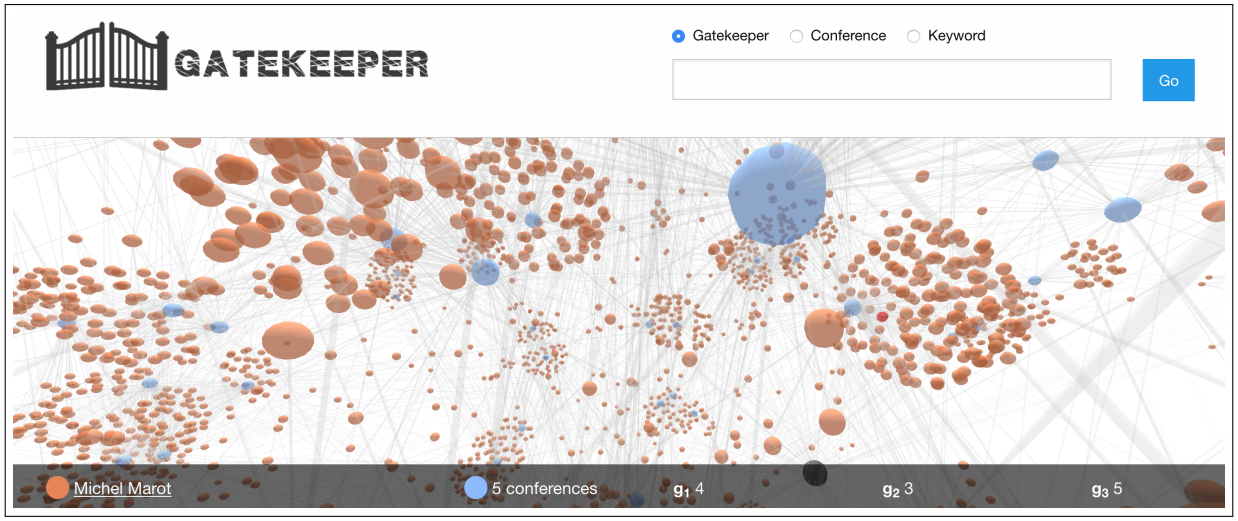


Figure 1: Screenshot of the Gatekeeper **system.**

ABSTRACT

While it has been extensively studied on how to model and measure a scholar's research impact (e.g., citation analysis), there have been very few studies that systematically collect and quantify a scholar's *service impact* to scientific communities. To address this lack of studies, we have developed a prototype digital library, named as Gatekeeper, that crawls, extracts, and quantifies scholars' service impacts based on their roles as "gatekeepers" in Computer Science conferences. Continuing this effort, in this work, we further theoretically analyze and improve the understanding on the expected behavior of three quantification measures (i.e., G-indexes) being used in Gatekeeper. In addition, we demonstrate that the stretched-exponential model fits significantly better than three other heavy-tail models (i.e., power-law, log-normal, and parabolic-fractal) in capturing scholars' service impacts via three quantification measures. Finally, using the analyzed quantification measures, we present leading scholars and conferences with respect to their service impacts. Our prototype is available at: https://gatekeeper.ist.psu.edu.

CCS CONCEPTS

• **Information systems** → *Digital libraries and archives*; *Web crawling*; • **Applied computing** → *Digital libraries and archives.*

KEYWORDS

Service Impact, Citation Analysis, *h*-Index, Stretched-Exponential Model

ACM Reference Format:
Jingtao Han, Spyke Krepshaw, and Dongwon Lee. 2020. Gatekeeper: Analyzing G-Indexes and Improving Service Quantification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3383583.3398528

1 INTRODUCTION

Being able to model and quantify the impacts of a scholar's *service* to research community and society has many utilities in applications– e.g., hiring and promoting scholars, or finding experts for service based committees. Yet, it is inherently subjective and ambiguous to quantify one's service impact. To start with, the precise definition

*Part of the work was done while visiting Penn State as a summer intern in 2019.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06…$15.00
https://doi.org/10.1145/3383583.3398528

of "service" is not straightforward. For instance, a scholar's service may include diverse activities such as participating in conference organizational or technical committees, serving in the editorial boards of journals, delivering talks in events, reviewing books, serving in funding-related panels, or interviewing with press and media, just to name a few.

Toward this ambiguity and complexity of the challenge, in [10], therefore, we limited our focus of service as the participation in the organizational or technical committees in Computer Science conferences, proposed the idea of measuring one's overall service impact by means of the *quantity* and *quality* of committees where one has served, and developed a prototype digital library, named as Gatekeeper, that implemented these ideas. We named it as Gatekeeper since scholars play an important role in the spread of research findings as "gatekeepers" by serving in conference committees. Continuing this effort, in this work, we make several significant improvements in Gatekeeper.

First, we theoretically analyze and improve the understanding on the expected behavior of three quantification measures (i.e., G-indexes) being used in Gatekeeper. We examine four alternatives (e.g., stretched-exponential, power-law, log-normal, and parabolic-fractal) to model the citation distributions of conferences against real data and demonstrate the superiority of the stretched-exponential model in capturing scholars' service impacts under three quantification measures.

Second, we significantly improve Gatekeeper in several aspects: (1) we increased the numbers of conferences, gatekeepers, and gatekeeper-conference service records in Gatekeeper by 4-5 times, respectively; (2) we developed and deployed machine learning based models to classify gatekeeper webpages, extract gatekeeper entities, and discern gatekeeper roles in a great detail.

Third, in comparison with other existing quantification measures for research impacts, we compare G-indexes against h-index and identify leading scholars and conferences whose service impacts are among the highest.

2 RELATED WORK

2.1 Quantifying Individual Research Outputs

One of the most popular methods to quantify the research impacts of a scholar is the h-index [7]. A scientist has index h if h of his or her papers have at least h citations each and the rest of papers have less or equal than h citations each. The h-index measures both the productivity (i.e., how many articles) and impact (i.e., how many citations) of one's research articles. While capturing both the productivity and impact of one's research well, the h-index method fails to recognize scholars who have made seminar findings with a small number of publications as they will have a low number of h. To improve on this shortcoming, the G-index method [3] further modifies the h-index such that a scholar receives a G-index score of G if she has published at least G articles that have been cited "collectively" at least G^2 times. This change has the effect of allowing highly-cited articles to effectively assist the low-cited articles in the calculation.

2.2 Measuring Conference Quality

Same as metrics for scholars, there are also a variety of methods to measure venue qualities. Some of the work use scores of authors and institutions of the papers admitted to the publishing venues [1][6].

There are also network-based methods for ranking venues have included citation information [22], and are able to produce temporal models of quality. Previous work [21] has developed methods for ranking by scores with seed-based measure that does not use citation analysis, and a realistic browsing-based measure that takes an article reader's behavior into account. However, this approach required manually labeled seeds indicating what a good work is. One work[4] used the average number of committee members, the average number of published articles by committee members, and the average closeness centrality of committee members as criteria to measure whether conferences are of the same quality. Recent development proposed a model that ranks scholars and venues based solely on individuals' status as faculty members, National Science Foundation grants and University of California salary data [9]. In this work, we generate scores for committee members with conference citation data and get conference metric from committee members' G-indexes.

2.3 Individual Paper Citation Distributions

The citation distribution of academic papers was first studied by Derek J. de Solla Price [14], whose work has indicated a power-law distribution. The author also proposed the so-called Cumulative Advantage Processes to understand the dynamics of citation , where a statement of "a paper which has been cited many times is more likely to be cited again than one which has been little cited" was presented. The mechanism could also be understood as preferential attachment in the framework of evolving networks [8]. More recently, Laherrère and Sornette [11] studied the citation record of the 1120 most cited physicists over the period between 1981 and 1997 as evidence for a stretched exponential distribution. Redner [15] suggested a power-law decay for the large citation tail by studying the citation distribution of 783,339 papers published in 1981, and the corresponding 6,716,198 citations to these papers between 1981 and 1997. Tsallis and de Albuquerque [20] proposed a continuous distribution from the non-extensive thermostatistical formalism. However, the citation distribution of conferences (total citation of papers accepted by a conference) have not been studied. We observe a similar pattern in the citation distribution of conference in real life data. Therefore, we use power-law and stretched exponential and other heavy-tailed functions as assumptions to analyze the 3 versions of G-indexes theoretically and numerically.

2.4 "Gatekeeper" System Prototype

In the previous work [10], to quantify the service impact of scientists, a prototype called Gatekeeper was proposed that crawled and kept records of information of computer science conferences and their program committee members. In this work, we aim to improve the system in terms of accuracy and extensiveness by new entity extraction algorithms and a webpage detection classifier. Inspired by the h-index that quantifies the impact of research outputs, the

previous work [10] proposed 3 versions of G-indexes with different emphasis using the citation data of conferences that a scholar served as program committee members. In this work, we further explore the idea by analyzing the 3 versions of G-indexes both theoretically and numerically, validating their quality and discussing their differences with existing metrics in evaluating scholars and conference.

3 ANALYSING G-INDEXES

3.1 Preliminaries

To quantify one's service impact, in [10], we leveraged on the concept of h-index and proposed 3 versions "Gatekeeper"-index methods. The Gatekeeper-index methods aim to capture both service productivity (i.e., how many service roles a scholar has served) and impact (i.e., how good a conference is where a scholar serves) of service. Intuitively, a scholar who has served more committees of conferences of higher qualities tends to have a higher Gatekeeper-index score. The first Gatekeeper-index uses the citations of conferences as a quality metric as follows:

Definition 3.1 (G_1-index). A scholar has the G_1-index score of N if he or she has served in N conferences and each conference has accrued at least a total of $f(N)$ citations, where $f(x)$ is a normalization function.

As the aggregated citation count of conferences are usually much larger than those of individual scholars, we use a normalization function $f(x)$ to suppress this inflation (e.g., x^2 and x^3). For instance, a scholar with the G_1-index score of 10 has to serve in the program committee of 10 different conferences with more than 10^2 or 10^3 citations each. Table 1, for instance, shows a list of conferences where a scholar "Lise Scholar" has served as a technical program committee member and their corresponding aggregated citation counts, fetched from the Microsoft Academic Graph (MAG) [16] digital library. Note that due to the collection delay, information of latest years is missing, and there is a wide fluctuation of the number of citations across conferences.

While intuitive to implement, the G_1-index disproportionately favors a large conference as they tend to have more articles, thus more aggregated total citations. To mitigate this problem, next, we propose the G_2-index that uses the *average* number of citations per article in a conference instead of the aggregated total of citations of a conference, as the quality metric of conferences.

Definition 3.2 (G_2-index). A scholar has the G_2-index score of N if she has served in N conferences and an article of each conference has accrued on average at least N citations.

Under the G_2-index, a scholar who has served in more impactful conferences, regardless of the size of conferences (thus, with higher citation counts per paper) tends to have a higher score than otherwise. Therefore, this G_2-index does not penalize scholars who served in many small conferences much as long as they are good ones. Finally, we propose the third version of G-index inspired by the notion of g-index [3] as follows:

Definition 3.3 (G_3-index). A scholar has the G_3-index score of N if she has served in top-N conferences (sorted in descending order of the citations of conferences) that have collectively received at least N^2 citations: i.e., $\sum_{N \geq i} C_i \geq N^2$, where C_i is the citation count of a conference among top-N conferences.

Table 1: A list of conferences served by a scholar "Lise Getoor" and their aggregated citation counts.

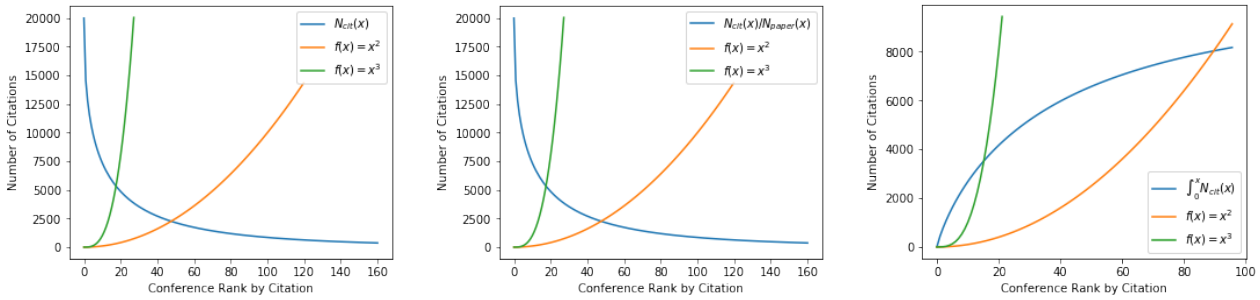
Conference	Citation #	Conference	Citation #
SIMBig 2018	83	ISWC 2009	4,093
IJCAI 2017	1,279	ILP 2009	1,324
CoDS 2017	462	KDD 2009	1,046
NeurIPS 2016	9,601	ICML 2008	4,150
MLG 2016	12	AAAI 2007	1,811
WACCK 2014	1,071	ICML 2007	9,725
ICML 2014	11,672	ECML-PKDD 2007	878
KDD 2014	7,345	SDM 2007	9,644
ICML 2013	11,559	SUM 2007	838
BIOKDD 2013	27,880	UAI 2006	4,951
MLG 2013	2,804	AAAI 2006	17,680
KDD 2012	9,987	KDD 2006	16,023
ICML 2011	15,687	ICML 2005	13,284
SemSearch 2010	1,699	AISTATS 2005	2,963
ICWSM 2010	7,845	MSW 2004	1,648
DyNaK 2010	2,441	SIGMOD 2004	15,137
SOMA 2010	1,235	SIGKDD 2003	16,414
CNIKM 2009	3,213	ICML 2003	13,940

Therefore, a scholar who has served in many highly-cited influential conferences (e.g., flagship conferences of sub-disciplines) is likely to have a higher G_3-index score. The schematic curves of three versions of G-indexes are shown in Figure 2.

3.2 Distribution of Conference Citations

All three G-indexes use some notions of quality of conferences by means of the total citation counts of conferences. As such, it is useful to understand how the overall citation counts of conferences behave and how we can model them. Previous works have found that both power-law and stretched-exponential models can accurately describe citation distributions of individual scholars. For instance, [14] showed that the citation distribution of papers indicates a power-law distribution. More recently, [11] reported that the citations of the 1,120 most-cited physicists from 1981 to 1997 are better modeled by the stretched-exponential distribution.

However, unlike these prior studies, the citation distribution of "conferences" (i.e., total aggregated citation count of all articles published in a conference) have not been well studied. After observing some data samples, we found that the citation distribution of conferences that a scholar has served as a program committee member shows some similarity with individual scholar's citation distribution. Both the conference citation distribution of a scholar and individual scholar's citation distribution are heavy-tailed. As shown in Figure 3, which illustrates a power-law like distribution (in a log-log plot) between the number of citations and the rank of conferences, a small fraction of conferences accrue a large number of citations while a large number of conferences receive only a small number of citations.



(a) G_1-index (b) G_2-index (c) G_3-index

Figure 2: Illustrations of 3 versions of G-index

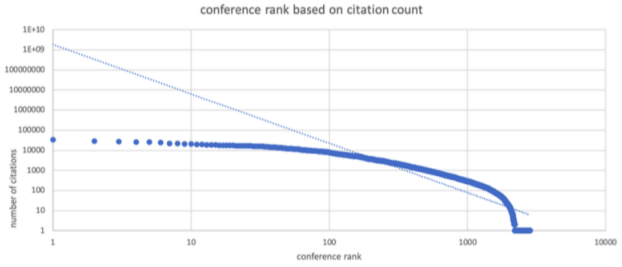


Figure 3: A power-law like distribution of conference citations.

Next, under the stretched-exponential model, we theoretically analyze the three versions of G-indexes, exploring their relations with the total citations of conferences, total number of conferences, and the maximum number of citations of conferences where a gatekeeper served in. Similar analysis for the power-law model is available in Appendix.

3.3 The Stretched-Exponential Model

Under this assumption of the Stretched-Exponential model, the total citation number of the k-th most cited conference where a scholar served as gatekeeper, $N_{cit}(k)$, follows:

$$N_{cit}(k) = N_0 e^{-k_0 k^\beta} \tag{1}$$

where $N_{cit,total}$ is the total number of citations of a conference where a scholar served in as gatekeeper, and N_{conf} is the total number of conferences where a scholar served in as gatekeeper. From this, we get $N_{cit,total}$ as follows:

$$N_{cit,total} \approx \int_0^\infty N_{cit}(k)dk = \int_0^\infty N_0 e^{-k_0 k^\beta} dk$$

In this case, we assume that the total number of conferences where a scholar served in as gatekeeper is large enough. Therefore, it is admissible to extend the upper limit of the integral to infinity, at the cost of slight overestimation of the total number of citations. Under this approximation, the integral can be analytically re-written as follows:

$$\int_0^\infty e^{-k_0 k^\beta} dk = \frac{\Gamma(1+\frac{1}{\beta})}{k_0^{\frac{1}{\beta}}}$$

where Γ is the usual Gamma function [17]. Hence, the distribution function can be re-written as:

$$N_{cit}(k) = N_{cit,total} \frac{k_0^{\frac{1}{\beta}}}{\Gamma(1+\frac{1}{\beta})} e^{-k_0 k^\beta}$$

The distribution function given by Equation 1 gives zero citation only when the rank is infinity. We assume that the least cited conference has the rank of pN_{conf} so that $N_{cit}(pN_{conf}) = 1$. Then, we get:

$$1 = \overline{N_{cit}} N_{conf} \frac{k_0^{\frac{1}{\beta}}}{\Gamma(1+\frac{1}{\beta})} e^{-k_0 p^\beta N_{conf}^\beta}$$

where p is the fraction of conferences which have been cited at least once, and $\overline{N_{cit}}$ is the average number of citations for those conferences. This expression can be treated as a transcendental equation in k_0. In fact, we get the following:

$$\beta[\frac{p\Gamma(1+\frac{1}{\beta})}{\overline{N_{cit}}}]^\beta = \beta k_0 p^\beta N_{conf}^\beta e^{-\beta k_0 p^\beta N_{conf}^\beta} = x e^{-x} \leq e^{-1} \tag{2}$$

where $x = \beta k_0 p^\beta N_{conf}^\beta$. The solutions of this equation for x can be obtained numerically. However, because the maximum of the function xe^{-x} is e^{-1}, Equation 2 only has a solution if:

$$\overline{N_{cit}} > p\Gamma(1+\frac{1}{\beta}) e^{\frac{1}{\beta}} \beta^{\frac{1}{\beta}}$$

According to the definition of G_1 index, then:

$$f(G_1) = N_{cit}(G_1)$$

where f is the normalization function. Hence,

$$f(G_1) = N_{cit,total} \frac{k_0^{\frac{1}{\beta}}}{\Gamma(1+\frac{1}{\beta})} e^{-k_0 {G_1}^\beta}$$

$$\frac{\Gamma^\beta(1+\frac{1}{\beta})}{\beta^2 k_0^3 N_{cit,total}^\beta} = \frac{e^{-\beta k_0 G_1^\beta}}{(\beta k_0 G_1^\beta)^2} = \frac{e^{-x}}{x^2} \quad when \quad f(x) = x^2$$

Table 2: Heavy-tailed functions.

Name	Function
Stretched-Exponential	$e^{-x_0 x^\beta}$
Power-Law	$x^{-\alpha}$
Log-Normal	$\frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}$
Parabolic-Fractal	$x^{-b}e^{-c(logx)^2}$

$$\frac{\Gamma^\beta(1+\frac{1}{\beta})}{\beta^3 k_0^4 N_{cit,total}^\beta} = \frac{e^{-\beta k_0 G_1^\beta}}{(\beta k_0 G_1^\beta)^3} = \frac{e^{-x}}{x^3} \quad when \quad f(x) = x^3$$

Therefore, the solutions of G_1-index can be obtained numerically. The analysis for G_2-index is similar using average citation counts.

We can also get relations between the G-index, the maximum citation of conferences where a gatekeeper served, and the total number of conferences where a gatekeeper served.

- Let $k = G_1$, then, we get:

$$max(N_{cit}(k)) = N_0 = G_1^2 e^{k_0 G_1^\beta} \quad when \quad f(x) = x^2$$

$$max(N_{cit}(k)) = N_0 = G_1^3 e^{k_0 G_1^\beta} \quad when \quad f(x) = x^3$$

Let $N_{cit}(k) = 1$, we get

$$N_{conf} = (2lnG_1/k_0 + G_1^\beta)^{\frac{1}{\beta}} \quad when \quad f(x) = x^2$$

$$N_{conf} = (3lnG_1/k_0 + G_1^\beta)^{\frac{1}{\beta}} \quad when \quad f(x) = x^3$$

According to the definition of the G_3 index,

$$f(G_3) = \int_1^{G_3} N_{cit}(k)dk \approx \int_0^\infty N_{cit}(k)dk$$

$$\int_0^\infty N_{cit}(k)dk = \int_0^\infty N_0 e^{-k_0 k^\beta} dk = N_0 \frac{\Gamma(1+\frac{1}{\beta})}{k_0^{\frac{1}{\beta}}}$$

where f is the normalization function. Hence,

$$f(G_3) \approx N_{cit,total}$$

$$G_{3a} = (N_{cit,total})^{\frac{1}{2}} \quad when \quad f(x) = x^2$$

$$G_{3b} = (N_{cit,total})^{\frac{1}{3}} \quad when \quad f(x) = x^3$$

- The case of $k = G_2$ is the same as that of $k = G_1$, thus omitted.
- Let $k = G_3$, we get,

$$max(N_{cit}(k)) = N_0 = \frac{G_3^2}{\int_1^{G_3} e^{-k_0 k^\beta} dk} \quad when \quad f(x) = x^2$$

$$max(N_{cit}(k)) = N_0 = \frac{G_3^3}{\int_1^{G_3} e^{-k_0 k^\beta} dk} \quad when \quad f(x) = x^3$$

Let $N_{cit}(k) = 1$, we get

$$N_{conf} = (\frac{lnN_0}{k_0})^{\frac{1}{\beta}}$$

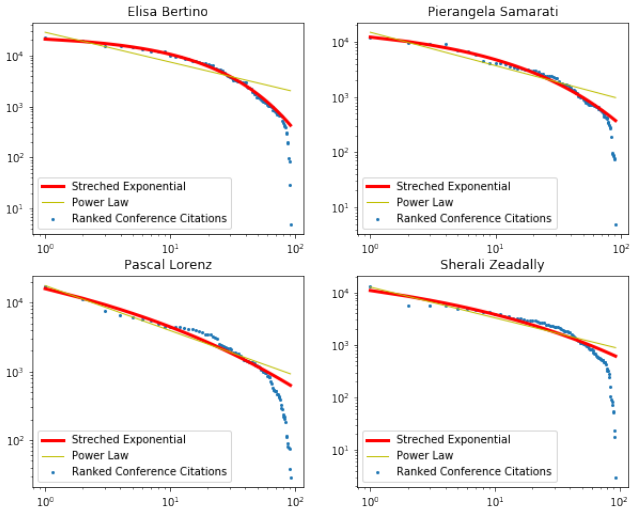


Figure 4: Most distributions, including four examples here, are best described by the Stretched-Exponential function.

4 FITTING REAL DATA

As we analyzed the stretched-exponential model with respect to three G-indexes in Section 3.3, in this section, we test if the stretched-exponential model is indeed a good one using real data. We use the `optimize` function in the SciPy [1] package to fit four alternative heavy-tailed functions listed in Table 2 against conference citation distributions of 6,813 scholars who have served in at least 10 conference committees. We use the Residual Sum of Squares (RSS) as the measure to evaluate the fitness:

$$RSS = \sum_{k=1}^{N_{conf}} (N_{cit}(k) - \hat{N_{cit}}(k))^2$$

To contrast models and get the overall quality of the fitness, we calculate the Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \frac{RMSE}{max(N_{cit}) - min(N_{cit})}$$

where $RMSE = \sqrt{\frac{RSS}{N_{conf}}}$.

This way, we normalize the errors by each gatekeeper's conference counts and conference citation scales. We can see in Table 3 that the stretched-exponential function performs the best. That is, 5,909 out of 6,417 data points are best described by the stretched-exponential, while 501 of 6,417 data points are best described by the power-law, 7 of 6,417 data points are best described by the log-normal, and none of the data is best described by the parabolic-fractal. Some examples of fitting are shown in Figure 4.

Next, by calculating the Spearman's ρ correlation measure [18] among paper counts, citation counts, and average citation per paper of a conference, we can see in Table 4 that the correlation between those metrics are weak. This suggests that additional fitting for the average citation statistics is necessary for the modeling of the G_2-index.

[1] https://www.scipy.org

Table 3: Fitting statistics for citations.

	NRMSE (Mean)	NRMSE (Median)	Best-Fit (%)
Stretched-Expo.	0.048	0.046	92.08%
Power-Law	0.103	0.100	7.81%
Log-Normal	0.314	0.307	0.11%
Parabolic-Frac.	0.430	0.418	0.00%

Table 4: Spearman's ρ correlation between different metrics.

	Paper #	Citation #	Average Citation #
Paper #	1.000	0.371	-0.155
Citation #	0.371	1.000	0.377
Average Citation #	-0.155	0.377	1.000

Table 5: Fitting statistics for average citations.

	NRMSE (Mean)	NRMSE (Median)	Best-Fit (%)
Stretched-Expo.	0.051	0.049	88.83%
Power-Law	0.100	0.094	10.26%
Log-Normal	0.202	0.208	0.00%
Parabolic-Frac.	0.267	0.267	0.91%

As listed in Table 5, the fitting statistics shows that the Stretched-Exponential function also performs the best in describing the distribution of average citation counts per paper of conferences. That is, 5,700 out of the 6,417 data points are best described by the stretched-exponential, while 658 of the 6,417 data points are best described by the power-law, none of the data is best described by the log-normal, and 58 of the 6,417 data points are best described by the parabolic-fractal.

By fitting the distribution of citation counts of conferences in which gatekeepers served as committee members, we can get the distributions of parameters in the Stretched-Exponential model. Here, we only present a realistic model for the G_1-index with the normalization function $f(x) = x^2$. The modeling for the other G_2-index and the G_3-index can be acquired in a similar manner.

Combined with our analysis, then, a given scholar's conference citation distribution can be modeled by choosing appropriate fitting parameters. This way, we get the relationship between the G_1-index and the highest citation counts N_0, the total conference number, N_{conf}, of conferences where a scholar served as committee members. For instance, when the normalization function $f(x) = x^2$, for $\beta = 1$, if $k_0^{-\frac{1}{\beta}} = 5$, we have $N_0 = G_1^2 e^{\frac{G_1}{5}}$, $N_{conf} = 10lnG_1 + G_1$. If $k_0^{-\frac{1}{\beta}} = 7$, we have $N_0 = G_1^2 e^{\frac{G_1}{7}}$ and $N_{conf} = 14lnG_1 + G_1$.

As scholars tend to gain more experience and be assigned more important roles (e.g., program chair vs. technical program committee member) through their academic career, we can expect the G-index scores of any scholars would increase with time. If a scholar stops her service in conference committees, their G-index are expected to increase for a while and then stay constant due to the limited number of influential conferences where they served.

Table 6: Data statistics in the prototype.

	Conferences	Gatekeepers	Relationships	Years
[10]	2,825	56,187	87,368	27
Updated	7,409	134,689	392,276	27

5 IMPROVING GATEKEEPER SYSTEM

5.1 Expanding and Elaborating Data

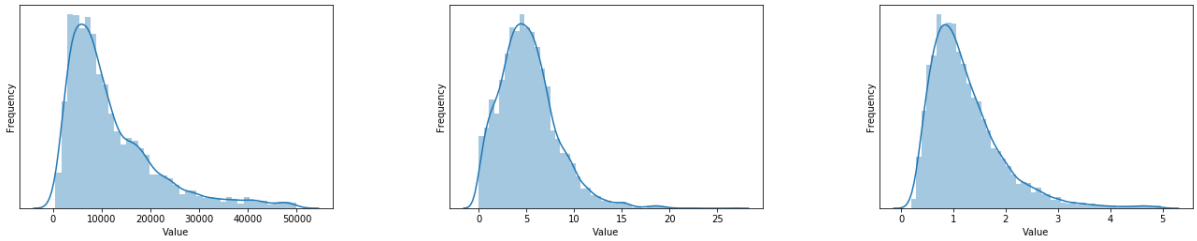
The previous prototype in [10] did not fully address the problem of acquiring citation data. To get extensive and detailed records of paper and citation counts for each conferences, we used the 2019-05-05 version of ArnetMiner Citation Network Dataset [19], which is extracted from the DBLP Computer Science Bibliography [12], the ACM Digital Library, and the Microsoft Academic Graph [16]. This dataset contains 4,107,340 papers with 36,624,464 citation relationships. By mining and merging the data by venue name and publishing year, then, we obtain paper and citation counts of 130,705 unique venues from 1970 to 2019. Finally, we match the data with our own 13,491 conference records, which results in 12,668 matches.

The previous method used in [10] to extract the names of gatekeepers simply cleaned, tokenized and tagged webpage texts as a whole using the Stanford NER (Named Entity Recognizer) package [5]. It obtained the names by combining consecutive tokens tagged as person names. We improved this method further by extracting a list of webpage text entries, splitting them with punctuation and tagging them separately to get results. As shown in Figure 6, the names on committee pages are usually separated with punctuation in different entries. In practice, the new method performed a lot better than the previous one. We obtained 134,689 gatekeepers from 7,409 conference records, improving from 56,187 gatekeepers from 2,825 conference records extracted by the previous method [10].

To improve the quality and robustness of our data, we got a list of top 425 computer science conferences ranked by the $h5$-index provided by Google Scholar Metrics from Guide2Research [2]. We further get committee pages links for top 100 conferences of different years manually, resulting in 2,970 records. In order to get paper and citation counts of existing conferences in the database, we matched the record with the aggregated version of the ArnetMiner Citation Network Dataset [19], which resulted in 2,195 matches. We extracted names and roles, added conference and gatekeeper data to our database and filtered out duplicates already existed in our database. The statistics of the improvement in current version of Gatekeeper is summarized in Table 6.

5.2 Extracting Service Role Titles

Gatekeepers often have different roles in program committees, we addressed the role extraction problem by a simple but efficient and accurate approach with the help of the Stanford NER (Name Entity Recognizer) package [5]. The cleaning and chucking approach is the same as the name extraction method, then we tagged role titles with keywords "committee", "chair", "member", "board", NS "advisory" in the token and names with the Stanford NER (Named Entity Recognizer) package. We attached person names with the role title



(a) N_0 (b) $k_0^{-\frac{1}{\beta}}$ (c) β

Figure 5: Parameter distributions of the Stretched-Exponential model of the G_1-index.

Figure 6: Structure of JCDL 2020 committee webpage with gatekeeper information.

Figure 7: A word-cloud of all role names before merge.

appeared first ahead of them. In this way, we are able to differentiate roles of gatekeepers in conference committees. As shown in Figure 6, service role titles highlighted in blue are followed by committee members serving the role highlighted in yellow. Simple as it sounds, the method performed well in practice, with 68% of the 394,208 gatekeeper-conference-title relation records fall into top 39 title names after aggregation of similar titles(Figure 7). Some of the top title names that appeared the most are listed in Table 7. Considering the vast varieties of formats of committee pages, this is a satisfying result.

5.3 Machine Learning Based Gatekeeper Page Detection

After examining 2,970 conference websites in the top conference data from Guide2Research [2], we have found out that a large fraction of the conference websites contain several committee pages, sometimes up to 8. For example, the website of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2019 have lists of Research Track Program Committee and Applied Data Science Track Program Committee on different webpages. It is also worth mentioning that about half of the links of committee pages do not contain any of the keywords above and some are formatted as PDF

Table 7: Top-15 most frequently occurring roles of gatekeepers after merge.

Rank	Title	Count
1	Technical Program Committee	179,681
2	Technical Program Chair	28,614
3	Area Chair	7,016
4	Steering Committee	6,596
5	Publicity Chair	4,603
6	Advisory Committee	3,149
7	Workshop Chair	2,971
8	Local Organization Chair	2,326
9	Organizing Committee	2,170
10	Publication Chair	1,958
11	Finance Chair	1,445
12	Tutorial Chair	1,167
13	Web Chair	645
14	Registration Chair	494
15	Poster Chair	365

or TXT files. As shown in Figure 6, webpages containing committee information usually have a large portion of person names and contains certain keywords like program, committee, chair, organization, chair, university and institute. Therefore, to solve those issues,

Table 8: 10-fold Cross Validation Scores

1	2	3	4	5	6	7	8	9	10
0.86	0.82	0.83	0.90	0.82	0.84	0.79	0.85	0.83	0.82
0.95	0.96	0.96	0.96	0.94	0.93	0.89	0.95	0.92	0.95
0.96	0.96	0.96	0.96	0.96	0.96	0.93	0.94	0.97	0.96

we built machine learning classifiers by fitting different popular models such as Naïve Bayes, Random Forest and Support Vector Machines on our existing dataset, using keywords, person name counts, total webpage word counts as features. We can train and test the models in Python using popular machine learning software packages [13]. Below are the results:

The 95% confidence interval for 10-fold cross validation scores of Multinomial Naïve Bayes, Random Forest and Support Vector Machine are 0.84 ± 0.06, 0.94 ± 0.04 and 0.96 ± 0.02 respectively.

A support vector machine constructs hyper-planes that can be used for binary classification. Intuitively, a good separation is achieved by the hyper-plane that has the largest functional margin. In the case, we use the radial basis function as the kernel function.

$$K(x, x') = e^{-\gamma \|x - x'\|^2} \tag{3}$$

After tuning parameters, the support vector machine model with the regularization parameter $C = 10$ and the kernel coefficient $\gamma = 0.0001$ performed the best, with a 95% confidence interval of 0.96 ± 0.02 for accuracy in a 10-fold cross-validation on 17,112 records.

We also implemented file reading functions and added more columns in our database to allow for multiple committee pages for a single conference. We got 10,543 gatekeeper pages for 13,491 conference records, comparing to 8,340 gatekeeper pages from 11,174 conference records identified by the previous method [10].

5.4 The Gatekeeper Prototype

The previous work [10] already built a Gatekeeper prototype which allowed users to query scholars, conferences, research keywords and got a node-based graph representing the connections of the conferences with their gatekeepers through a front-end interface. The steps constructing the database are: (1) finding conference websites, (2) finding committee Webpages (among hundreds of webpages in a conference website), (3) scraping gatekeeper-related information (e.g., name, affiliation, email, etc), and (4) scraping research keywords of conferences.

We implement three improvements into the prototype in Step 1 by adding top conferences records, in Step 2 by building the keyword-based machine learning model to detect committee pages with improved precision, in Step 3 by incorporating the role extraction algorithm along with the new name extraction method. Figure 8 is a screenshot of the main query page for gatekeepers, conferences and keywords, along with top scholars and top conferences ranked by the G-indexes. Figure 9 is a screenshot of the scholar profile in the interface with conferences they served, the role they served in conferences and a node graph of scholars and conferences for relation analysis (conferences of the same series or domain are usually clustered together).

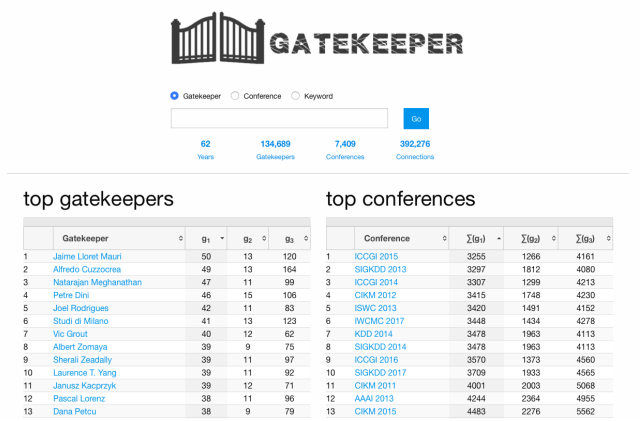


Figure 8: Screenshot of the main query page in Gatekeeper**.**

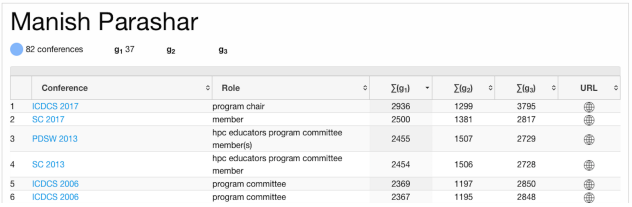


Figure 9: Screenshot of a scholar page in Gatekeeper**.**

6 RELATIONSHIP BETWEEN SERVICE AND RESEARCH IMPACT

The method of quantifying the impacts of a scholar's service can be used in various applications. A high G-index score in Gatekeeper implies that the scholar has richer experience in serving computer science conferences and playing the role of gatekeepers (often with respect to both quantity and quality of services).

We first make a list of scholars with top G_1-index scores and find out that scholars with high G-index scores usually also have decent[2] h-index scores [7]. The means of both G_1-index (with $f() = x^3$ as the normalization function) and h-index among all scholars in Gatekeeper are 2.5 and 22.2, respectively. When we fetch a list of all 1,163 ACM fellows form the official website, 602 fellows exist as gatekeepers in Gatekeeper. The mean G_1-index of these ACM fellows in Gatekeeper is 9.6, which is significantly higher than the average among scholars in Gatekeeper, indicating that ACM fellows (who are in general regarded to have made high research impacts) tend to have made high service impacts as well.

To compare our service metrics with the widely used h-index, which captures the impact of a scholar's research output, we use Spearman's ρ correlation [18] between our G-indexes and the h-index obtained from Google Scholar. For comparison, we collected a randomly selected subset (n = 9,447) from our Gatekeeper database.

We can see in Table 9 that our G-indexes are highly correlated since they all measure service impacts. However, the correlation

[2]For instance, the h-index score of 40 usually characterizes an outstanding scientist, likely to be found among scholars at top universities or major research laboratories [7].

Table 9: Spearman's ρ correlation between different impact measures.

	h-index	G_1-index	G_2-index	G_3-index
h-index	1.00	0.40	0.47	0.38
G_1-index	0.40	1.00	0.89	0.84
G_2-index	0.47	0.89	1.00	0.81
G_3-index	0.38	0.84	0.81	1.00

between research impact and service impact are not the strongest. However, as Table 10 lists, top scholars with respect to their research impacts (having high h-index) tend to have reasonably high G-index scores as well. Therefore, we can say that not all high-achieving scholars have high-achieving service impacts, but most do.

Another intriguing application of individual service metric such as G-indexes is to quantify the aggregated conference impact scores. One may assume that important and selective conferences in fields are usually organized by a group of experienced senior scholars who often have made significant research contributions in a community. Therefore, by adding or averaging out their service impacts, one can measure the impact scores of conferences. For instance, Table 11 shows that conferences with high aggregated G-index scores are all influential ones judging by their h5-index, which is the h-index of articles published in the last 5 years of a conference. We fetched the h5-index data from Google Scholar via Guide2Research [2].

6.1 Limitation and Future Work

While we tried to improve our data collection and cleaning processes as much as possible, there are still non-negligible number of data points for scholars and conferences missing. These errors may influence the calculation of G-index scores, introducing errors. Due to the name disambiguation problem, it is also possible that two different scholars' G-index calculation is merged incorrectly. Finally, many websites of older conferences are no longer maintained, and thus introducing null data points in Gatekeeper. In future, we plan to attempt to find such missing websites using web tools (e.g., Wayback Machine of Internet Archive). Finally, as major publication outlets of other disciplines than computer science are often journals, we plan to crawl and extract the editorial board information of journals.

In addition, also, note that we do not differentiate different types of "service" in the definitions of G-indexes. That is, serving as a program committee member vs. as a program chair is counted equally. Therefore, it is also plausible to define a more fine-grained definitions of G-index, by giving disproportionate weights to more senior role types of service (e.g., serving as a program chair is viewed as twice more service than serving as a program committee member). We leave this exploration as future work.

Table 10: Examples of scholars with top G-index scores (sorted in descending order in G_1-index scores and $f() = x^3$ is used for G_1-index).

Name	G_1-index	G_2-index	G_3-index	h-index
Lise Getoor	19	20	42	65
Christos Faloutsos	18	19	48	126
Andrew McCallum	18	22	26	98
Elisa Bertino	18	22	101	98
Raghu Ramakrishnan	18	22	34	87
Ricardo Baeza-Yates	18	19	71	78
Wolfgang Nejdl	18	16	80	70
Thorsten Joachims	18	19	29	69
Dale Schuurmans	18	20	31	47
Jiawei Han	17	20	70	168
Gerhard Weikum	17	18	54	86
Jian Pei	17	20	92	85
Johannes Gehrke	17	21	37	76
Jieping Ye	17	16	41	70
Ming-Syan Chen	17	20	44	64
Hang Li	17	22	58	63
Irwin King	17	16	54	59
Yannis Ioannidis	17	16	36	59
Wei Wang	17	16	93	53
Rich Caruana	17	17	19	47
Sunita Sarawagi	17	17	29	45
Masaru Kitsuregawa	17	19	40	43
Alfredo Cuzzocrea	17	17	190	39
Cordelia Schmid	16	16	21	112
Jure Leskovec	16	19	54	95
Bing Liu	16	19	52	87
Ajith Abraham	16	13	78	85
Dimitrios Gunopulos	16	16	53	72
Andrew Tomkins	16	19	33	64
Haixun Wang	16	17	42	64
Raymond Ng	16	17	27	63
Minos Garofalakis	16	16	51	63
Nick Koudas	16	19	53	62
Xindong Wu	16	16	50	60
Charles Elkan	16	14	28	55
Carla Brodley	16	15	18	55
Naren Ramakrishnan	16	11	46	46
Jaime Lloret Mauri	16	13	120	43
Olfa Nasraoui	16	11	42	39
Giovanni Semeraro	16	10	61	38
Tiziana Catarci	16	13	54	37
Witold Pedrycz	15	11	42	109
Francesco Ricci	15	7	36	107
Sushil Jajodia	15	14	44	104
Victor Bahl	15	21	45	91

7 CONCLUSION

To enable to track and quantify service impacts of scholars, we have proposed three measures of G-indexes and implemented them in the prototype digital library, Gatekeeper. In this work, further, we have analyzed the alternative models to capture the distributions of conference citations, that are used in the definition of G-indexes, and found that the stretched-exponential model fits the best.

8 ACKNOWLEDGEMENT

This work was in part supported by NSF award #1934782. We appreciate anonymous reviewers for their constructive comments.

Table 11: Conferences with top aggregated G-index scores.

Title	$\sum G_1$-index	$\sum G_2$-index	$\sum G_3$-index	$h5$-index
AAAI	6,087	4,643	12,575	95
KDD	5,451	4,596	12,510	86
ICCV	4,108	3,670	5,721	129
CIKM	4,015	3,223	8,646	48
OOPSLA	3,860	2,359	7,284	34
ISWC	3,431	2,450	8,472	21
ICML	3,261	2,891	5,078	135
SIGIR	2,997	2,523	5,468	55
ECCV	2,839	2,652	3,587	137
ICDM	2,792	2,171	8,094	44
WWW	2,763	2,518	5,609	70
EMNLP	2,679	2,364	4,393	88
SC	2,640	2,046	5,609	43
WSDM	2,620	2,278	5,507	51
LREC	2,483	2,093	3,775	45
IWCMC	2,432	1,462	4,243	21
ICDE	2,408	2,141	5,286	14
CVPR	2,369	2,191	3,077	240

REFERENCES

[1] Emery Berger. 2020. *CSRankings: Computer Science Rankings.* http://csrankings.org/.
[2] Imed Bouchrika. 2019. Guide2Research: Top Computer Science Conferences. http://www.guide2research.com/topconf/
[3] Leo Egghe. 2006. Theory and Practise of the G-index. *Scientometrics* (2006).
[4] Ergin Elmacioglu and Dongwon Lee. 2009. Oracle, Where Shall I Submit My Papers? *Commun. ACM* (2009).
[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*.
[6] Robert Geist, Madhu Chetuparambil, Stephen Hedetniemi, and A. Joe Turner. 1996. Computing Research Programs in the U.S. *Commun. ACM* (1996).
[7] J. E. Hirsch. 2005. An Index to Quantify an Individual's Scientific Research Output. *PNAS* (2005).
[8] H Jeong, Z Néda, and A. L Barabási. 2003. Measuring Preferential Attachment in Evolving Networks. *Europhysics Letters* (2003).
[9] Leonid Keselman. 2019. Venue Analytics: A Simple Alternative to Citation-Based Metrics. In *JCDL*.
[10] Spyke Krepshaw and Dongwon Lee. 2019. Gatekeeper: Quantifying the Impacts of Service to the Scientific Community. In *TPDL*.
[11] J. Laherrère and D. Sornette. 1998. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal* (1998).
[12] Michael Ley. 2002. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *SPIRE*.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011).
[14] Derek De Solla Price. 1976. A General Theory of Bibliometric and other Cumulative Advantage Processes. *Journal of the American Society for Information Science* (1976).
[15] S. Redner. 1998. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *The European Physical Journal B: Condensed Matter and Complex Systems* (1998).
[16] Boris Schauerte. 2008. Microsoft Academic: Conference Field Ratings. http://www.conferenceranks.com/visualization/msar2014.html
[17] Jerome Spanier and Keith B. Oldham. 1987. *An Atlas of Functions.*
[18] C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* (1904).
[19] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD*.
[20] C. Tsallis and M F de Albuquerque. 2000. Are Citations of Scientific Papers a Case of Nonextensivity? *The European Physical Journal B - Condensed Matter and Complex Systems* (2000).
[21] Su Yan and Dongwon Lee. 2007. Toward Alternative Measures for Ranking Venues: A Case of Database Research Community. In *JCDL*.
[22] Tong Zhange. 2004. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In *ICML*.

A APPENDIX

A.1 Power-law to Model Conference Citations

Under this assumption, the total citation number of the k-th most cited conference where a scholar served as gatekeeper is: $N_{cit}(k)$ follows $N_{cit}(k) = N_0 k^{-\alpha}$ Then,

$$N_{cit,total} = \int_1^{N_{conf}} N_{cit}(k)dk = \int_1^{N_{conf}} N_0 k^{-\alpha} dk$$

where $N_{cit,total}$ is the total number of citations a scholar served in as a gatekeeper, N_{conf} is the total number of conferences a scholar served in as a gatekeeper. From this,

$$N_0 = (1-\alpha)N_{cit,total}N_{conf}^{\alpha-1} = (1-\alpha)\overline{N_{cit}}N_{conf}^{\alpha}$$

where

$$\overline{N_{cit}} = N_{cit,total}N_{conf}^{-1}$$

According to the definition of the G_1-index, $f(G_1) = N_{cit}(G_1)$, where f is the normalization function. Hence,

$$G_1 = [(1-\alpha)\overline{N_{cit}}N_{conf}^{\alpha}]^{\frac{1}{(\alpha+2)}}, \quad when \quad f(x) = x^2$$

$$G_1 = [(1-\alpha)\overline{N_{cit}}N_{conf}^{\alpha}]^{\frac{1}{(\alpha+3)}}, \quad when \quad f(x) = x^3$$

We can also get relations between the G-index, the maximum citation of conferences a gatekeeper served in and the total number of conferences a gatekeeper served in. Let $k = G_1$, we get

$$max(N_{cit}(k)) = N_0 = G_1^{\alpha+2} \quad when \quad f(x) = x^2$$

$$max(N_{cit}(k)) = N_0 = G_1^{\alpha+3} \quad when \quad f(x) = x^3$$

Let $N_{cit}(k) = 1$, we get

$$N_{conf} = G_1^{1+\frac{2}{\alpha}} \quad when \quad f(x) = x^2$$

$$N_{conf} = G_1^{1+\frac{3}{\alpha}} \quad when \quad f(x) = x^3$$

According to the definition of the G_3 index,

$$f(G_3) = \int_1^{G_3} N_{cit}(k)dk$$

where f is the normalization function. Hence,

$$G_3 = [\overline{N_{cit}}N_{conf}^{\alpha}]^{\frac{1}{(\alpha+1)}}, \quad when \quad f(x) = x^2$$

$$G_3 = [\overline{N_{cit}}N_{conf}^{\alpha}]^{\frac{1}{(\alpha+2)}}, \quad when \quad f(x) = x^3$$

The case of $k = G_2$ is the same as that of $k = G_1$, thus omitted. Let $k = G_3$, we get

$$max(N_{cit}(k)) = N_0 = (1-\alpha)G_3^{3-\alpha} \quad when \quad f(x) = x^2$$

$$max(N_{cit}(k)) = N_0 = (1-\alpha)G_3^{4-\alpha} \quad when \quad f(x) = x^3$$

Let $N_{cit}(k) = 1$, we get

$$N_{conf} = [(1-\alpha)G_3^{3-\alpha}]^{\frac{1}{\alpha}} \quad when \quad f(x) = x^2$$

$$N_{conf} = [(1-\alpha)G_3^{4-\alpha}]^{\frac{1}{\alpha}} \quad when \quad f(x) = x^3$$