# Community-based Endogamy as an Influence Indicator

Thiago H. P. Silva, Mirella M. Moro, Ana Paula C. Silva
Wagner Meira Jr., Alberto H. F. Laender
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{thps,mirella,ana.coutosilva,meira,laender}@dcc.ufmg.br

## ABSTRACT

Evaluating researchers (individually or in groups) usually depends on qualifying their publications and influence. Here, we aid such crucial task by introducing two new metrics (*C-Endo* and *Comb*) that rely on the concept of endogamy for communities of authors who publish in conferences and journals, and produce patents. Endogamy here measures how tightly structured the groups of authors are within a community. We validate and evaluate the metrics by using real datasets, two ground-truth rankings and citation count. We also perform random sampling analysis to account for any unbalance from the ground-truth rankings. Overall, such a thorough evaluation shows that our metrics are successful in defining community-based endogamy as an influence indicator.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## Keywords

Research performance; Bibliometric indicators

## 1. INTRODUCTION

The study of coauthorship networks and scientific collaborations informs on the evolution of science. Mining such networks enables to identify communities, their members, experts, influential people and more [4, 15, 16, 17]. Such networks may also expose patterns of cooperation among researchers and their publications impact [2, 18]. Evaluating a group of researchers (an ongoing problem for institutions and funding agencies) is often based on metrics for assessing publications impact. Indeed, analyzing networks may aid the evaluation of, for example, research groups and graduate programs [11, 13, 14].

In this paper we propose to measure influence by the concept of endogamy. Overall, people who potentially make a network stronger by building social capital or bringing new ideas to a different group have been called *weak ties* by Granovetter [5], *brokers* by Burt [3], the *newcomer on the networks* by Guimerà et al. [6], whereas Montolio et al. [15] define such influential people by proposing the concept of *research endogamy*. Within Social sciences, *endogamy* is the tendency of marrying to members of one's own social group. Then, *research endogamy* reveals the *degree of cooperation among researchers*: groups with members that always cooperate among themselves have a high endogamy value, whereas groups that tend to make new relationships have low endogamy. All the aforementioned work have one insight in common: evaluating the structural properties of the networks may be employed for assessing the quality of the work produced or the influence of the people involved.

Indeed, Montolio et al. [15] propose an *endogamy indicator* for assessing the quality of publication venues, focusing on Computer Science conferences and journals. The results show a strong correlation between low endogamy and the quality of conferences. However, although there was some correlation for journals, the conclusion was that their indicator is not suitable for journals.

We propose a novel community-based endogamy metric called *C-Endo*. Here, a community is intuitively defined by a set of people who share common research interests (e.g., a scientific conference usually has a list of topics of interest that define its scope). In this context, *C-Endo* measures how frequently an author publishes with members from the same community. Furthermore, differently from [15], *C-Endo* emphasizes the role of weak ties (brokers, newcomers, etc) in each community, i.e., people who are influential by bringing new ideas to a community. For example, a group with many publications in Information Retrieval is likely to bring new ideas when publishing in Digital Libraries; *C-Endo* accounts for such behavior better than [15]. At the end, we show that *C-Endo* can be successfully applied as an indicator for evaluating the influence of conferences, journals and patents.

We validate our approach on a real dataset of scientific publications using two ground-truths. Our results surpass the baseline ([15]), and the new indicator is able to measure the quality of *both* conferences and journals. We also propose an aggregated indicator (called *Comb*) that combines different approaches to compute endogamy. Furthermore, we analyze the correlation between endogamy and another well-known bibliometric indicator, the citation count [7]. The results show a strong correlation between them: publications with small endogamy have more citations. We further analyze our new indicators by applying them to patents – we define the categories of the patents as communities. The

results consider the number of citations received and show a strong correlation as well. Therefore, publishing patents in varied categories is an indicative of influence.

Next, we present fundamental concepts and discuss related work (Section 2). Then, we go over our main contributions, which are summarized as follows:

- We introduce two metrics for qualifying conferences, journals and patents by computing the endogamy based on authors and their communities (Section 3);
- After presenting an experimental methodology (Section 4), we validate the metrics as influence indicators for conferences and journals, followed by random sampling analysis (Section 5);
- We also analyze the distribution of endogamy by tier and its sensibility to the number of authors and extend the concept of community to work on patents as well. We experimentally analyze the metrics in such contexts by comparing results to state-of-the-art rankings and citation indexes (Section 6).

# 2. BACKGROUND

This section introduces the background to understand our contributions. Specifically, Section 2.1 goes over the fundamental concepts used through out this paper and Section 2.2 emphasizes our contributions over the related work.

## 2.1 Fundamental Concepts

The research endogamy of a set of **authors** $A$ is defined by Montolio et al. [15] as Equation 1.

$$Endo(A) = \frac{|d(A)|}{|\cup_{a \in A} d(\{a\})|} \quad (1)$$

where $d(A)$ is the co-authored papers by $A$. For example, let $A = \{author_1, author_2\}$, $d(author_1) = \{p_1, p_2, p_3\}$, and $d(author_2) = \{p_2, p_3, p_4\}$. Then, $Endo(A) = 0.5$, because authors $author_1$ and $author_2$ have coauthored two ($p_2$ and $p_3$) out of four papers ($p_1$ to $p_4$).

Having defined the endogamy of a set of authors, the next steps are to define the *collective* endogamy of a publication and then of a whole venue. First, the research endogamy of a **publication** $P$ (work, article, paper, patent, etc.) is defined as the average of endogamy values of the power set of its authors, i.e., the average of all subsets formed by more than one author. Given $A(P)$ as the set of authors of the publication $P$, and $L_i(p) = P_i(A(p))$ the subsets formed by the set of authors with size equal to $i$. Then, the set of all subsets with more than one author of $P$ is given by $L(P) = \bigcup_{i=2}^{i=|A|} L_i$. Finally, the endogamy of $P$ is the average of its endogamies $L(P)$ given by Equation 2.

$$Endo(P) = \frac{\sum_{x \epsilon L} Endo(x)}{|L|} \quad (2)$$

The research endogamy of a **venue** $V$ is the average of the endogamies of its set of publications $P$ as Equation 3.

$$Endo(V) = \frac{1}{|V|} \sum_{p \epsilon V} Endo(p) \quad (3)$$

Note that such equation considers only the collaborations made before the date of the publication.

## 2.2 Related Work

People who potentially make a network stronger (or build social capital) have been subject of multiple studies under different names. For instance, in Social Sciences, Granovetter [5] claims that *weak ties* bring real strength to a social network because they are likely to link people from different groups, building bridges in the network. Burt [3] proves the concept of *brokers* as the people who work across the structural holes between groups, providing visions otherwise unseen and building social capital. The same concept is regarded as the *newcomer on the networks* by Guimerà et al. [6], who found that the performance of research teams in social psychology and ecology can be significantly better when such newcomers enter the group. Montolio et al [15] define *research endogamy* as a metric for evaluating coauthorship networks and assessing the quality of conferences based on their networks. Likewise, Kato and Ando [9] show that international collaboration (i.e., people from outside local groups) improves the overall performance of researchers in Chemistry. No matter the name, all aforementioned publications emphasize the importance of bridging actors (weak ties, etc.) when evaluating networks of researchers in order to identify influential people, quantify the quality of a network or even measure the overall performance of a network.

In a different way, Yan et al [19] propose a set of features for evaluating and predicting the influence of a publication. Two claims are close to ours: the authors' network structure and the entropy (diversity) of the author's topic distribution are relevant for assessing quality. They also assess citation count as a metric of influence. The conclusions corroborate the already cited work: social relationships and the author's ability to work with different groups (in distinguished topics) impact on the influence and quality of a work.

Finally, the work by Montolio et al. [15] is the most related to ours. The authors argue that cooperating with new researchers is very likely to introduce new ideas to a research community. To measure it, they propose calculating the research endogamy of authors, publications and venues (as presented in Section 2.1) and validate their metric for Computer Science conferences and journals. Their most important conclusions are: ($i$) for conferences, there is a strong negative correlation between endogamy and quality; and ($ii$) for journals, although there is some correlation, endogamy is *not* suitable for qualifying them.

Here, we take such studies one step forward and propose two new metrics for qualifying the influence of conferences, journals and patents based on their authors community endogamy. We experimentally validate such claim by considering real datasets and performing random sampling analysis. Note that, as explained next, the concept of community is versatile enough to be applied to conferences and journals as well as patents. All studies consider state-of-the-art rankings and metrics as ground-truths and baselines.

# 3. COMMUNITY-BASED ENDOGAMY

This section presents our endogamy metrics based on the cooperation among communities (Section 3.1). It also discusses the combination of such metrics (Section 3.2).

## 3.1 C-Endo

As discussed in Section 2.2, the existing endogamy computation (defined by [15] and here called *Endo*) explores the

degree of new collaborations, but not the *importance* of relationship between authors and their publications venue. For instance, researchers who tend to publish in few venues with a selected set of authors are considered very important to the community because, often, such authors act as hub due to high expertise, thus making the community stronger as a whole by following the aforementioned concepts of weak ties, bridges and so on. However, *Endo* fails in capturing such important relationships. Specifically, the bridging-authors attain *less* weight when compared to the others. Then, even when researchers bring new ideas to a group that potentially produce high quality work, *Endo* penalizes the little interactivity with new researchers, so generally they have high endogamy values (i.e., are less influent).

In a different perspective, we propose a new metric called *C-Endo* (community-based endogamy for conferences, journals and patents) for considering the importance of influent groups in the communities. A community is given by a set of people who share common research interests and work in groups or individually towards developing them and publishing relevant results. For example, journals and conferences define their scope through a list of topics of interest; likewise, patents are produced within categories and subcategories. *C-Endo* measures the degree of participation of the set of authors in the communities. Specifically, given a set of authors $A$, a community $c$ and a subset of works in that community $l_c(A)$, *C-Endo* is defined by Equation 4.

$$C\text{-}Endo(A,c) = \frac{|l_c(A)|}{|\cup_{a\in A}\, l_c(\{a\})|} \quad (4)$$

For instance, let $a$ and $b$ be authors with publications (or patents) $d(\{a\}) = \{v_1, x_1, y_1, z_2\}$ and $d(\{b\}) = \{y_1, z_2, w_2\}$, where for each publication $p_c$, $c$ indicates that $p$ belongs to community $c$. So, $C\text{-}Endo(\{a,b\},1) = 1/3$ and $C\text{-}Endo(\{a,b\},2) = 1/2$.

Following Equations 1 and 3, *C-Endo* of a work is also defined by the average of the endogamy values of the power set of authors with more than one author. Then, Equation 5 adapts Equation 3 for the new *C-Endo* endogamy.

$$C\text{-}Endo(V,c) = \frac{1}{|V|}\sum_{p\epsilon V} C\text{-}Endo(p,c) \quad (5)$$

*C-Endo* follows the idea that researchers who publish in other communities are more likely to introduce new ideas from their competence in such contexts. Such authors are highly influent for connecting parts of a network, then tightening and making a whole community stronger.

## 3.2 Combined Metrics

Bibliometrics is very restrict for *individually* assessing the quality or influence of a specific publication. For example, take citation count as a sole quality metric: we all know that an article receiving 10 citations does not indicate that it is better than others with less citations. The same is also true for other 39 impact indicators [1]. Hence, previous work (e.g., [12]) have considered *combining* bibliometric indicators for improving such assessment by considering a wider range of aspects, then making the results less unfair and biased. Following such a trend, we also propose combining *Endo* and *C-Endo* as a new indicator called *Comb*. In summary, *Comb* uses the average of the two metrics normalized by the overall endogamy average for all venues[1].

[1]Similar results were obtained using the average of all articles in the normalization instead of the average of venues.

**Table 1: Publication venues of ERA in DBLP.**

| | Conferences | Journals |
|---|---|---|
| A | 144 | 105 |
| B | 98 | 76 |
| C | 89 | 69 |
| Total | 331 | 250 |

**Table 2: Publication venues of Qualis in DBLP.**

| | Conferences | Journals |
|---|---|---|
| A1 | 86 | 56 |
| A2 | 77 | 42 |
| B1 | 116 | 59 |
| B2 | 55 | 46 |
| B3 | 26 | 23 |
| B4 | 18 | 14 |
| Total | 378 | 240 |

# 4. EXPERIMENTAL METHODOLOGY

This section presents the methodology underlying the experiments described in Sections 5 and 6. In particular, we aim to investigate the endogamy in a real publication dataset against two ground-truths and in a real patent dataset.

## 4.1 Venue Ranking Ground-truths

Evaluating a group of researchers is often based on bibliometrics and needs qualified committees to interpret the results and provide a fair comparative analysis. For instance, consider the discrepancy between the publication patterns of areas (Physics and Computer Science) and subareas (Databases and Theory), where some focus on publishing innovative results in conferences and others in journals. Moreover, one key process is to rank the venues in which the researchers publish for qualifying their publications. Here, we consider the existing ranks from two government agencies as ground-truth for ranking publication venues.

**ERA Ranking.** Excellence in Research for Australia (ERA)[2] is a project administered by the Australian Research Council (ARC). It aims at promoting excellence of research activities in Australia's higher education institutions. For this purpose, ERA ranks the publication venues in three categories: A, B and C, where A is the best one and B is better than C. This classification is determined by committees of distinguished researchers from Australia and overseas. We take the ERA evaluation of 2010. This baseline is the same used by Montolio et al. [15] in their experiments.

**Qualis Ranking.** Qualis is an initiative of CAPES[3] (the Brazilian Ministry of Education agency that assesses and funds graduate programs) for rating publication venues. Similar to ERA, highly qualified committees from different areas classify the publication venues in eight categories: A1, A2, B1 to B5 and C, where A1 is better than A2, A2 is better than B1, and so on. The C category does not have weight and comprises, for example, new publications with no history. We consider the latest Qualis from 2012.

## 4.2 Datasets

Our evaluation considers three different datasets. The first one contains data from the DBLP[4] digital library, lim-

[2]ERA: `http://www.arc.gov.au/era`
[3]WebQualis: `http://qualis.capes.gov.br`
[4]DBLP: `http://www.informatik.uni-trier.de/~ley/db`

**Table 3: Agreements for conferences and journals for *All* tiers and tier *A* according to ERA.**

| | Conferences | | Journals | |
|---|---|---|---|---|
| **Metric** | **All** | **A** | **All** | **A** |
| **Endo** | 75.32 | 77.75 | 60.01 | 62.88 |
| **C-Endo** | 70.65 | 74.40 | 71.02 | 75.07 |
| **Comb** | 76.08 | 79.49 | 68.21 | 71.61 |

**Table 4: Agreements for conferences and journals for All tiers, $G_1$ ($\{A_1, A_2\}$, $\{B_1, B_2\}$, $\{B_3, B_4\}$) and $G_2$ ($\{A_1, A_2, B_1\}$, $\{B_2, B_3, B_4\}$) according to Qualis.**

| | Conferences | | | Journals | | |
|---|---|---|---|---|---|---|
| **Metric** | **All** | **$G_1$** | **$G_2$** | **All** | **$G_1$** | **$G_2$** |
| **Endo** | 73.71 | 78.24 | 79.42 | 58.13 | 58.66 | 60.39 |
| **C-Endo** | 76.03 | 78.91 | 78.89 | 72.07 | 73.77 | 72.24 |
| **Comb** | 79.92 | 83.45 | 83.16 | 67.06 | 68.75 | 69.08 |

ited to the same venues used by ERA and Qualis (by matching title, acronym or ISSN): conference venues with more than 500 papers and journal venues with more than 100 papers. These limitations are based on Montolio et al. [15], who showed that such cleanup does not affect the conclusions. Tables 1 and 2 show the number of venues retrieved by category for each baseline. Note that categories B5 and C from Qualis are not included because they have no or little participation within DBLP.

The second dataset (called *SHINE$^{+J}$*) comes from Lima et al. [12] and is based on SHINE[5] (Simple HINdex Estimator). It contains 1,018,410 publications from 1,934 conferences with 8,780,752 citations, plus 201,593 articles from 188 journals with 4,499,049 citations. The data collected ranges from 2000 to 2012.

The third dataset is the NBER U.S. Patent Citation [7] provided by the National Bureau of Economic Research. For each patent, this dataset includes: the inventors, location of the first inventor, citations received and other variables as the technological subcategory. The dataset comprises almost 3 million U.S. patents (1963-1999) with over 16 million of citations made between 1975 and 1999.

## 4.3 Evaluation Metric

The experimental validation considers the ranking of all venues from the datasets by endogamy values. The degree of similarity between two rankings is given by *agreement*. Specifically, for any pair ($p_1$, $p_2$) of conferences or journals, if $p_1$ is from a category better than $p_2$, then the pair is concordant; otherwise, it is discordant. Ties are not computed. The agreement between two rankings is given by Equation 6.

$$\rho = 100\frac{p}{p+f} \tag{6}$$

where $p$ and $f$ are the number of concordant and discordant pairs. We also test the statistical significance[6] of the rank correlation by means of the Kendall Tau coefficient [10] – the same used by Montolio et al. [15] in their experiments.

# 5. EXPERIMENTAL VALIDATION

[5]SHINE: `http://shine.icomp.ufam.edu.br`

[6]All statistical tests performed with confidence level $\alpha = 0.05$.

**Table 5: Means of agreements for conferences and journals with random samples of 50 venues per tier by ERA ranking.**

| | Conferences | | Journals | |
|---|---|---|---|---|
| **Metric** | **All** | **A** | **All** | **A** |
| **Endo** | 73.16 | 76.20 | 59.60 | 63.43 |
| **C-Endo** | 69.57 | 74.75 | 69.16 | 74.73 |
| **Comb** | 75.43 | 80.27 | 68.72 | 73.70 |

**Table 6: Means of agreements for conferences and journals with random samples of 10 venues per tier for All tiers, $G_1$ ($\{A_1, A_2\}$, $\{B_1, B_2\}$, $\{B_3, B_4\}$) and $G_2$ ($\{A_1, A_2, B_1\}$, $\{B_2, B_3, B_4\}$), by Qualis ranking.**

| | Conferences | | | Journals | | |
|---|---|---|---|---|---|---|
| **Metric** | **All** | **$G_1$** | **$G_2$** | **All** | **$G_1$** | **$G_2$** |
| **Endo** | 76.71 | 79.66 | 82.31 | 59.23 | 58.92 | 62.74 |
| **C-Endo** | 77.97 | 81.17 | 84.26 | 77.30 | 79.34 | 78.39 |
| **Comb** | 81.31 | 84.62 | 87.38 | 69.98 | 70.74 | 71.09 |

We introduced two new ways for calculating community-based endogamy. After presenting the ground truths, datasets and evaluation metric, we now proceed by detailing how the approaches are validated against the state-of-the-art. First, we validate the proposed approaches (*C-Endo* and *Comb*) against ERA and Qualis (Section 5.1). The number of venues per tier in both ERA and Qualis is not equal, which may introduce bias. Hence, we confirm our results by a random sampling analysis (Section 5.2).

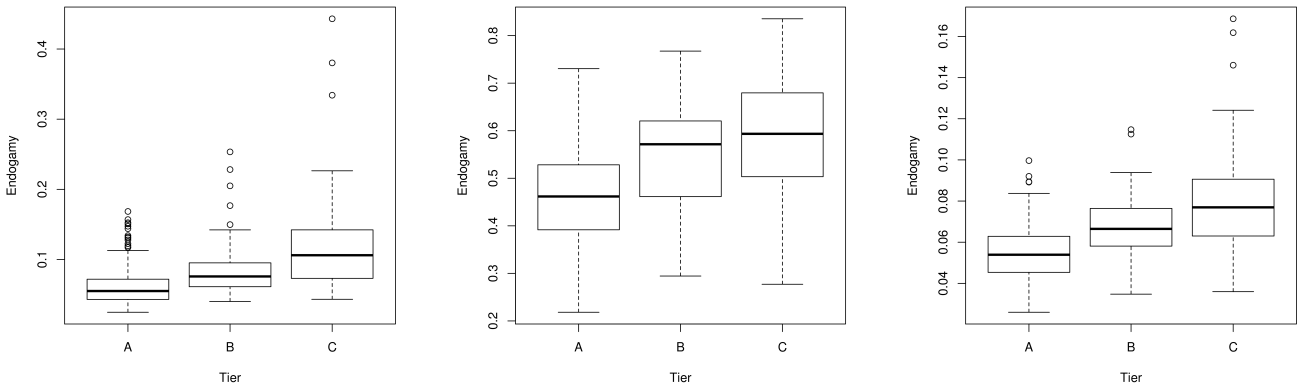
## 5.1 Initial Results

Table 3 shows the agreement results for ERA, where *All* considers all categories (ERA's A, B and C), and *A* emphasizes the best category, leaving B and C as having the same weight. For conferences, all results show very strong rank correlation, with most values between 75 and 80%. *Comb* provides the best results for both scenarios (*All* and *A*), but with a little difference from the baseline (*Endo*). *C-Endo* suffers a little with a difference of about 5% (*All*) and 1.7% (*A*) from the best result. For journals, *C-Endo* is the best and has a large difference from the others: 11% (*All*) and 12.1% (*A*) from *Endo*, and 2.1% (*All*) and 3.4% (*A*) from *Comb*. The combined version was negatively affected by the *Endo*, but still has a strong correlation.

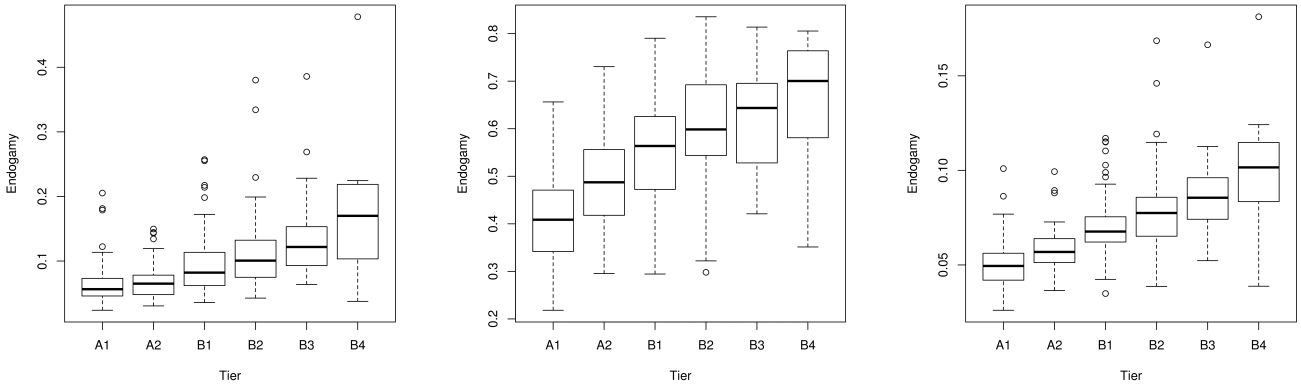
Table 4 shows the results for Qualis, where: *All* is for all categories (A1 to C), $G_1$ for grouping categories A1 and A2 as the top tier, B1 and B2 as the intermediate tier, and B3 and B4 as inferior tier; and $G_2$ for grouping A1, A2 and B1 as the top tier, and B2 to B4 as the inferior tier. For both conferences and journals, the best results are the same of ERA: *Comb* and *C-Endo*, respectively.

Also, for Qualis, the differences between the best results and the baselines are more significant: 6.2% (*All*), 5.2% ($G_1$) and 3.7% ($G_2$) for conferences, and 13.9% (*All*), 15.1% ($G_1$) and 11.8% ($G_2$) for journals. Again, all rankings have a strong correlation for conferences, and the worse results are *Endo* for journals. Note that *C-Endo* has better results for *All* and $G_1$ than *Endo* for conferences (opposite to ERA).

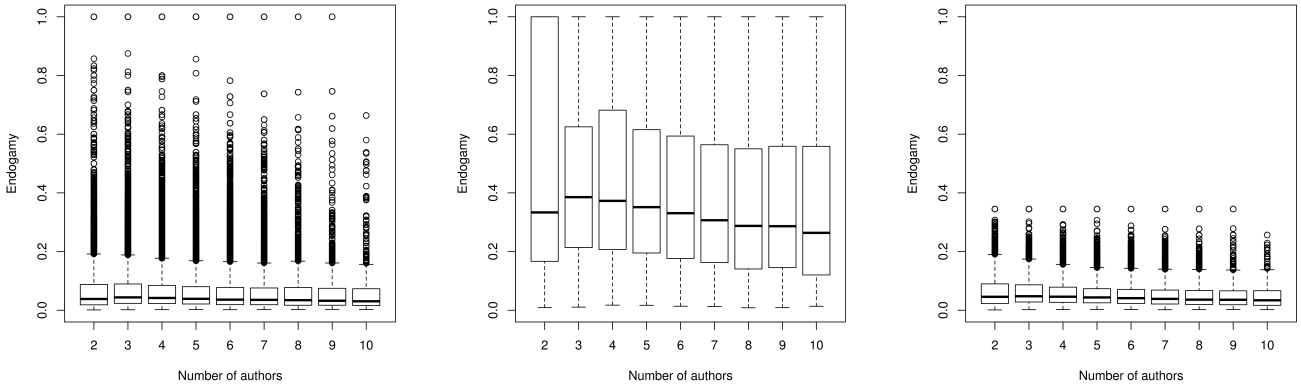
The weak results of *Endo* for journals were reported by Montolio et al. [15], which concluded that *Endo* is not suitable to be used as indicator of quality for journal venues.




(a) Endo per ERA tier (b) C-Endo per ERA tier (c) Comb per ERA tier




(d) Endo per Qualis tier (e) C-Endo per Qualis tier (f) Comb per Qualis tier




(g) Endo by number of authors (h) C-Endo by number of authors (i) Comb by number of authors

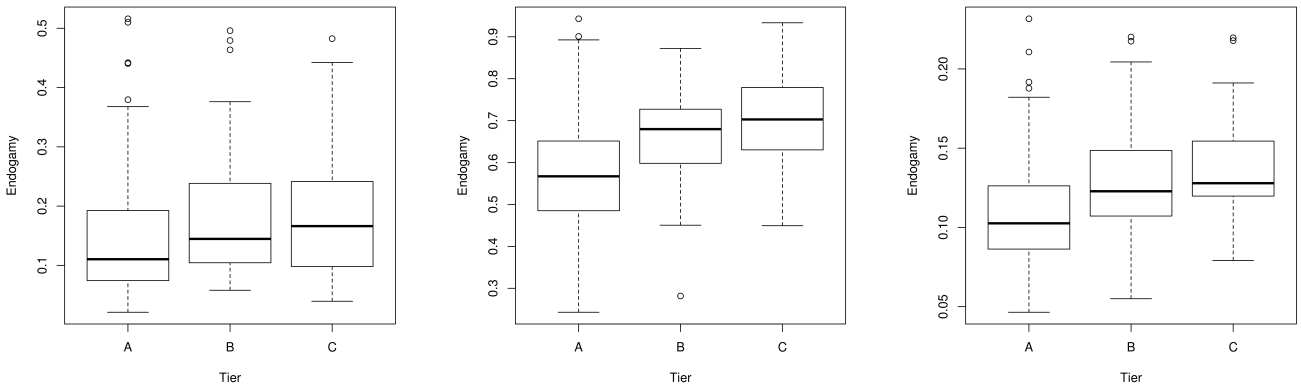
**Figure 1: Endogamy for ERA (a to c) and Qualis (d to f), and by number of authors in conferences (g to i).**

Furthermore, Montolio et al. [15] discuss the reasons of such bad results and point out that most innovative ideas are presented in conferences, whereas journals serve for archival. Then, the conclusion was that many journals focus on deeper analysis of previous ideas benefiting, therefore, the same group of authors. In addition, conferences allow researchers to travel and have direct contact with other researchers (in presentations, workshops, coffee breaks, etc.), thus resulting in a greater likelihood of new cooperation.

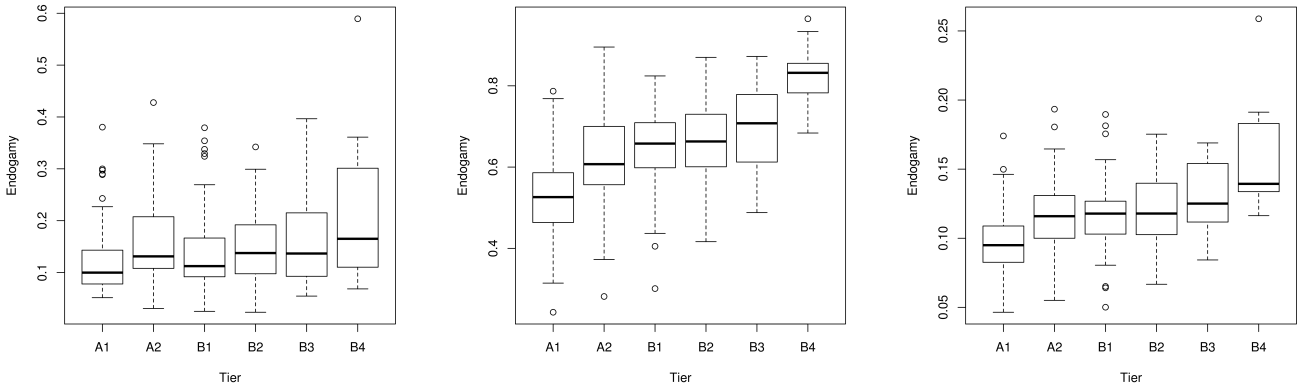
Comparing the results for conferences and journals, the presence of influential groups of researchers from different venues are high and very similar. Then, *C-Endo* is suitable for being used as indicator in both scenarios, as well as the combined version. All the correlations showed are statistically significant by means of the Kendall Tau coefficient.

## 5.2 Random Sampling Analysis

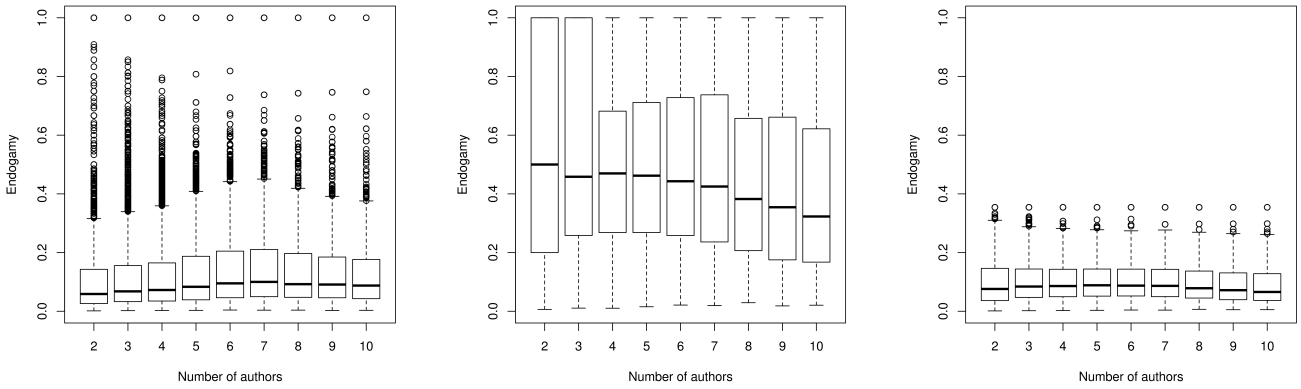
The previous evaluations were performed on real data and provided strong results that validate *C-Endo* and *Comb*. However, the ground truth rankings have unbalanced number of venues per tier which may affect the results (see Tables 1 and 2). For instance, tier A in ERA has 144 venues, whereas B has 98 and C has 89. With a good classification only for tier A, there will be a lot of concordant pairs and,




(a) Endo per ERA tier (b) C-Endo per ERA tier (c) Comb per ERA tier




(d) Endo per Qualis tier (e) C-Endo per Qualis tier (f) Comb per Qualis tier




(g) Endo by Number of authors (h) C-Endo by number of authors (i) Comb by number of authors

**Figure 2: Endogamy for ERA (a to c) and Qualis (d to f), and by number of authors in journals (g to i).**

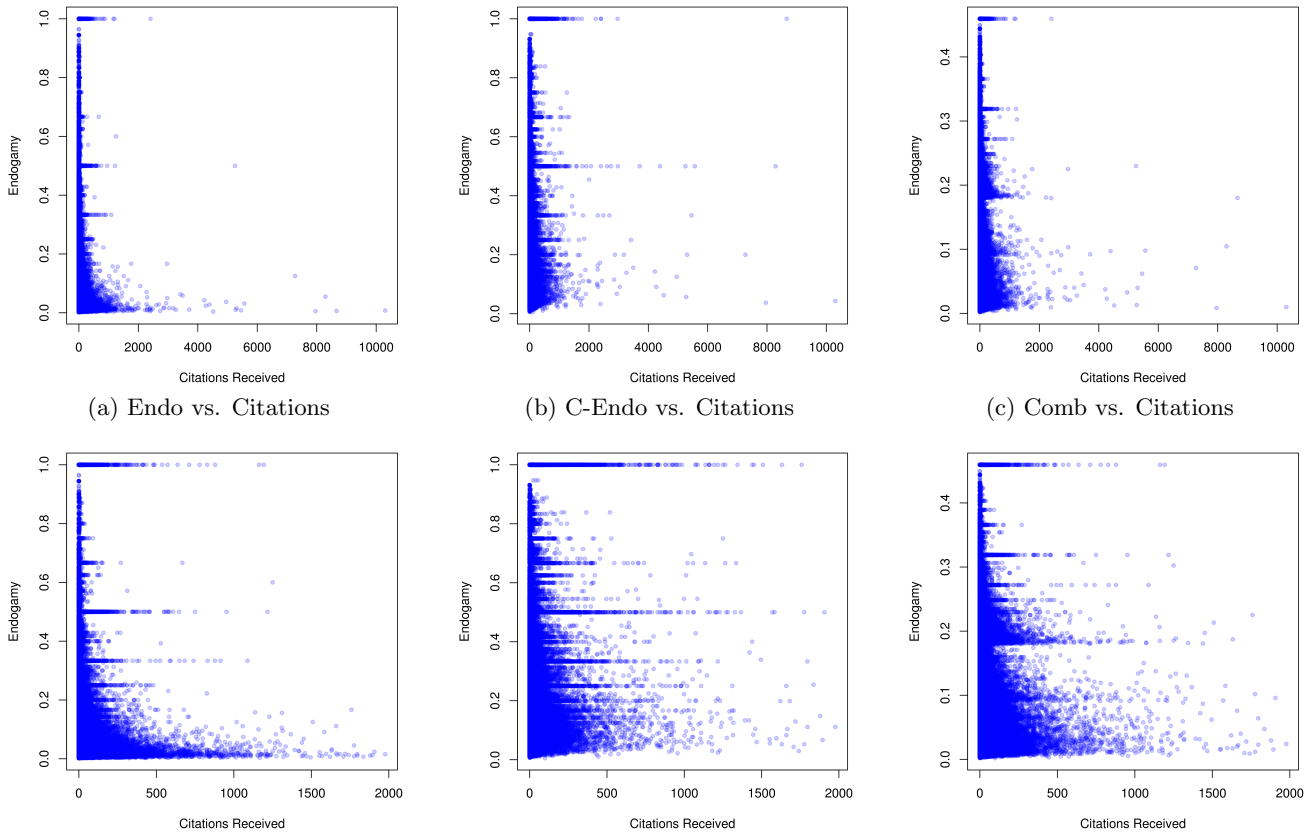
hence, the agreement shall have a high correlation. But, for a bad classification, an opposite effect will appear. Hence, we now select random samples of venues and compute their rank correlation.

Table 5 shows the agreement average of ten series performed with random samples of 50 venues per tier for ERA. Even with agreement values smaller than the real dataset (Table 4), the rank correlation is still strong. The best results for conferences come from *Comb*, and for journals from *C-Endo*. It is very important to notice that *C-Endo* for journals is way better than *Endo*, emphasizing *C-Endo*'s great performance for journals.

Table 6 shows the agreement average of ten series performed with random samples of 10 venues per tier for Qualis: *Comb* provides the best results for conferences and *C-Endo* for journals. Note that such values are superior in comparison with the real dataset (Table 4).

By the Kendall Tau coefficient, the correlations were not statistically significant for *Endo* in journals (agreeing with [15]). Still, this is an exception, for all the rank correlations are statistically significant by means of Kendall Tau coefficient.

Overall, the whole evaluation in random samples confirmed the results obtained for the real datasets. *C-Endo*




(a) Endo vs. Citations (b) C-Endo vs. Citations (c) Comb vs. Citations

(d) Zoom of Endo vs. Citations (e) Zoom of C-Endo vs. Citations (f) Zoom of Comb vs. Citations

**Figure 3: Endogamy values versus citations received for *SHINE*$^{+J}$ dataset of (a) Endo, (b) C-Endo and (c) Comb. The second row shows the zoom ones.**

and *Comb* have a strong rank correlation and are, then, suitable to be used as bibliometric indicators. However, *Endo* is still not suitable for journals (again, agreeing with [15]).

# 6. ENDOGAMY DISTRIBUTION

Now, we check if the metrics distinguish venue tiers through endogamy and their bias on the number of authors. We analyze the distribution of endogamy by tier and its sensibility to the number of authors for conferences (Section 6.1) and journals (Section 6.2), and to the number of citations received by publications and patents (Section 6.3).

## 6.1 Conference Analysis

Figures 1(a) to 1(c) show the endogamy values per tiers for ERA. All endogamy values have medians increasing from tier A to C. Also, *Endo* has a large number of outliers that is reduced in the combined version. *C-Endo* does not have outliers, but the minimum value (inferior extremity) of tier B is bigger than C, and the median of tier B is bigger than the first quartile and very close to the median of C. Figures 1(d) to 1(f) show the endogamy values per tiers for Qualis. Again, all versions of endogamy have the medians increasing from tier A1 to B4. Note that the minimum and maximum values are not well defined (extremities of the boxes).

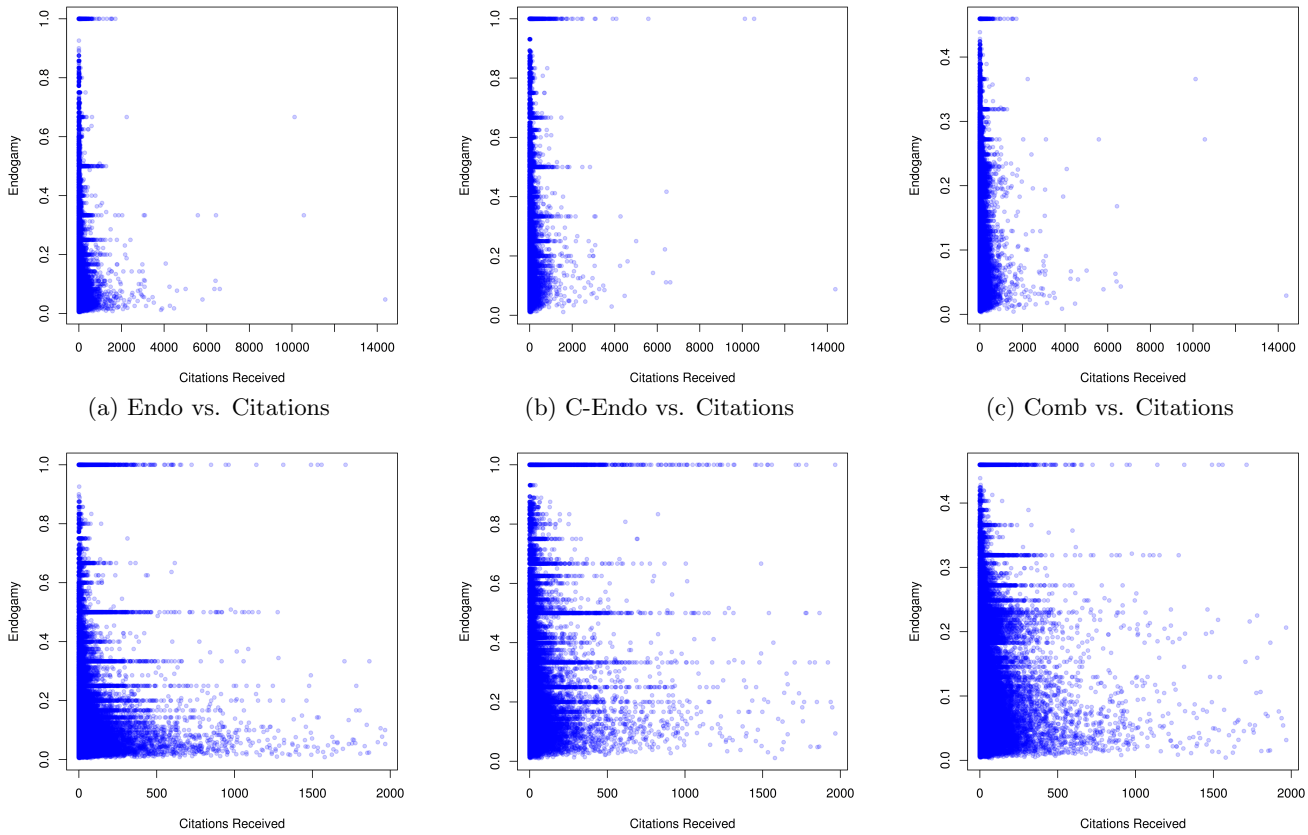
Figures 1(g) to 1(i) show the influence of the number of authors. *C-Endo* shows a little variation among the medians, as they decrease slightly for more than six authors. *Endo* and *Comb* show an homogeneous distribution.

Overall, the results show that all indicators have high values for rank correlations and are *not* sensitive to the number of authors. Hence, all forms of computing endogamy are suitable for conferences. Among them, *Comb* provides the best results due to the more concentrated values in accordance with the tiers and few number of outliers.

## 6.2 Journal Analysis

Figures 2(a) to 2(c) show the endogamy per tiers for ERA. Again, all versions of endogamy have the medians increasing from tier A to C. For *Endo*, the first quartile (bottom of box) of tier B is superior to tier C, and the third quartile (top of box) is nearly equal. For *C-Endo*, the range of value of tier A is very expressive, as well as the minimum and maximum values against the other tiers. Despite the maximum value of tier B and the proximity between the medians of tiers B and C, *Comb* is the best defined.

Figures 2(d) to 2(f) show the endogamy values per tiers for Qualis. For *Endo*, there is no distribution well-defined between the values of endogamy and the tiers. For *C-Endo*, there is a better distribution, but tiers B1 and B2 are very similar because they belong to the critical transition zone. The bad result in *Endo* affects *Comb*, where the tiers A2, B1 and B2 are very similar.




(a) Endo vs. Citations (b) C-Endo vs. Citations (c) Comb vs. Citations

(d) Zoom of Endo vs. Citations (e) Zoom of C-Endo vs. Citations (f) Zoom of Comb vs. Citations

**Figure 4: Endogamy values versus citations received for journals of (a) Endo, (b) C-Endo and (c) Comb. The second row shows the zoom ones.**

Figures 2(g) to 2(i) show the influence of the number of authors in the endogamy for journals, with similar results to those of conferences. Despite *C-Endo* having a little variation in the medians between 5 and 10 authors, it has a homogeneous distribution as well as *Comb*.

*Endo* does not perform well for Qualis, with weak agreement values, which corroborates that *Endo* is not suitable to assess the quality of journals. *C-Endo* provides the best results and can be used to assess the quality of journals, although it cannot distinguish well tiers B1 and B2 for Qualis (these two tiers are in the borderline of high quality venues on A1 and A2 and low quality venues on B3 and B4).

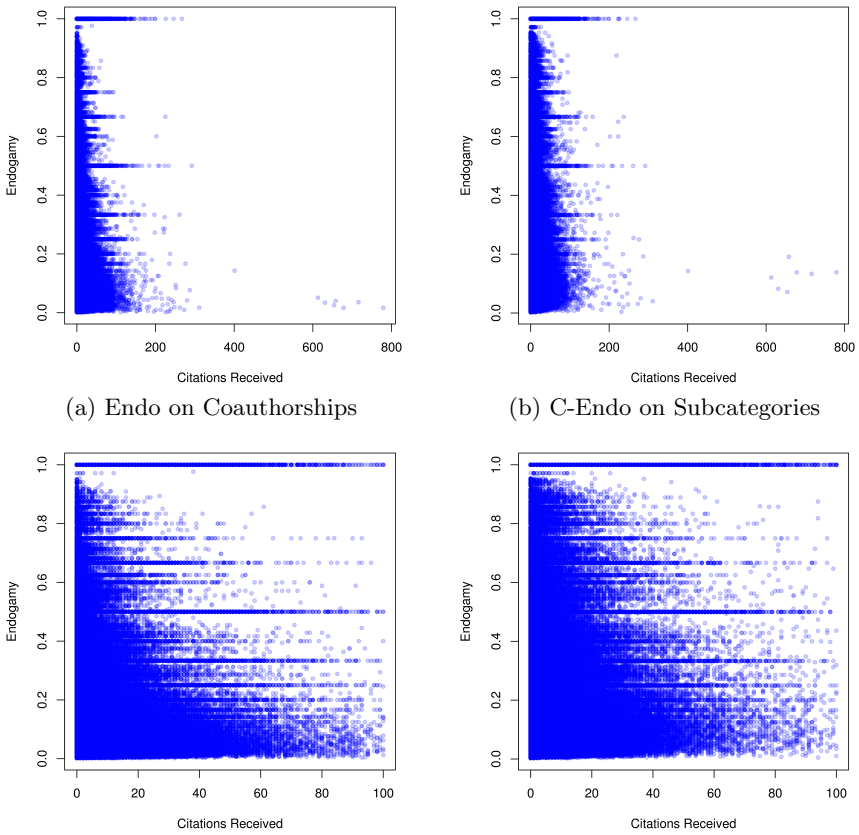
## 6.3 Endogamy versus Citations

We now compare endogamy and the number of citations. We assume that the lower the endogamy, the greater the likelihood of new ideas being introduced for producing high quality work, then the more citations a work receives.

**Endogamy of Conference Papers versus Citations Received.** Figures 3(a) to (c) show the endogamy values versus the number of citations received for conferences papers; whereas Figures 3(d) to (f) zoom in such results (up to 2000 citations received) for better visualization. We expected all graphs to be filled with more points in the *low endogamy/high citations received* quadrant. The results are better than expected and show a well defined behavior with low endogamy papers receiving more citations – clarified by the zoom graphs. *Endo* has a cleaner drawing, but all graphs are very similar. There are many outliers with maximum values for endogamy in all graphs. Also, *Comb* provides two triangle aspects (bottom and endogamy nearly to 0.2) due to the average used in combination.

Note that the number of points close to few citations was expected, since the new publications have had less time to be cited. For example, the citation average of conference papers in 2000 is 23.6, whereas in 2012 is 0.41 (citations collected in 2013). However, the density close to the origin is not very clear in the visualization. Hence, to clarify the results, the papers are grouped by endogamy and quality venues in Table 7(a). The column *All* shows the citations distribution of all papers grouped in *C-Endo*[7] intervals varying in 0.1, showing that the average citations is higher for lower values of endogamy. The remaining columns group the papers in tiers and enable to analyze the behavior of endogamy according to the quality of the venue (following the Qualis ranking that is more estratified than ERA). For instance, A1 has the highest average citations for the lowest endogamy interval $[0.0, 0.1)$, whereas B4 has the worst. Furthermore, for the same range both the averages and the medians decrease from A1 to B4. Overall, these results emphasize the use of endogamy to assess the influence of conferences.

[7] *Endo* and *Comb* have similar results, but were not shown due to limited space.




(a) Endo on Coauthorships (b) C-Endo on Subcategories

(c) Zoom on Endo (d) Zoom on C-Endo

**Figure 5: Endogamy versus Citations Received for NBER Patents based on (a) Coauthorships and (b) Subcategories. The second row shows the zoom ones.**

**Endogamy of Journal Articles versus Citations Received.** Figures 4(a) to (c) show the endogamy values versus the number of citations received for journal articles, with zoom in Figures 4(d) to (f). Again, the results are better than expected for all graphs, emphasizing a well defined behavior with low endogamy articles receiving more citations. The citation average of journal articles is 58 in 2000 and 1.2 in 2012, so again the large number of points close to few citations was expected. However, the results are not as clean as for conferences because there are outliers with too many citations that affect the visualization. Then, Table 7(b) summarizes the citation distribution. The column *All* shows that the average citations is greater for the low endogamy intervals (with few exceptions, e.g., at interval $(0.5, 0.4]$ that is bigger than $(0.4, 0.5]$). Endogamy is best defined (low endogamy/high average citation) for the top tiers A1 and B1, whereas other tiers present noise on some intervals. Tier A2 does not have good citation average in the lowest endogamy interval, but its remaining values are suitable. Tier B1 has the highest average at the lowest endogamy interval and overcomes A2 in other intervals; however, it also has a large standard deviation and bad results at interval $(0.8, 0.9)$. Then, the tier with lowest quality (B4) is the only one that does not have papers in the best endogamy interval. All results confirm those for conferences.

**Endogamy of Patent versus Citations.** Now, we verify the endogamy behavior for patents. The goal of this experiment is to show that our definition of community-based endogamy is versatile because it may be applied in contexts beyond the usual analysis of publication venues (conferences and journals). The NBER patents dataset has a classification for patents that considers 6 categories and 36 subcategories. Note that *Endo* focuses on the patents' coauthorships. On the other hand, we apply *C-Endo* considering the subcategories as *communities*, i.e., *C-Endo* measures the presence of influential groups of inventors in different subcategories. Also, following [7] and [8], we consider that influential patents present bigger citation count, which is the baseline metric in this evaluation.

Figure 5 shows the results for *Endo* on coauthorships (a) and *C-Endo* on subcategories (b). Note that unlike publications, patents (1963-1999) have a very large interval to be cited (1975-1999). Again, we expected that all graphs were filled with more points in the *low endogamy/high citations received* quadrant. However, all show an aspect similar to the previous results. Also, although the triangle aspects remain, there are more points with intermediate and high endogamy values with more citations than for the publications graphs. Nonetheless, these results confirm the use of endogamy as an indicator of influence using citation as baseline. They reinforce that the endogamy computation

**Table 7: Statistics (average, standard deviation, and median) of distributions of citations for (a) conferences and (b) journals, with publications grouped by tiers (Qualis) and *C-Endo*'s intervals varying in 0.1.**

| (a) | ***All*** | | | **A1** | | | **A2** | | | **B1** | | | **B2** | | | **B3** | | | **B4** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Interval | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | |
| Endogamy | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. |
| [0.0, 0.1] | 18.2 | 88.9 | 2.0 | 29.8 | 121.0 | 7.0 | 16.2 | 44.9 | 5.0 | 7.4 | 20.2 | 2.0 | 3.7 | 8.5 | 1.0 | 3.7 | 7.3 | 1.0 | 1.2 | 4.6 | 0.0 |
| (0.1, 0.2] | 13.8 | 66.1 | 2.0 | 27.1 | 103.9 | 6.0 | 16.3 | 38.2 | 5.0 | 7.8 | 19.1 | 3.0 | 5.0 | 17.7 | 1.0 | 3.5 | 8.9 | 1.0 | 1.4 | 3.5 | 0.0 |
| (0.2, 0.3] | 10.8 | 41.6 | 1.0 | 22.5 | 67.2 | 6.0 | 15.9 | 37.4 | 5.0 | 8.0 | 18.9 | 3.0 | 4.9 | 10.1 | 1.0 | 3.8 | 10.2 | 1.0 | 1.6 | 5.3 | 0.0 |
| (0.3, 0.4] | 9.4 | 43.5 | 1.0 | 22.6 | 81.3 | 5.0 | 14.9 | 33.6 | 5.0 | 7.9 | 20.6 | 3.0 | 4.7 | 11.6 | 1.0 | 3.3 | 6.8 | 1.0 | 1.6 | 4.6 | 0.0 |
| (0.4, 0.5] | 8.5 | 56.9 | 1.0 | 22.8 | 105.0 | 5.0 | 14.0 | 85.2 | 4.0 | 8.0 | 25.4 | 3.0 | 4.4 | 10.1 | 1.0 | 3.6 | 9.5 | 1.0 | 1.7 | 4.8 | 0.0 |
| (0.5, 0.6] | 7.7 | 31.5 | 1.0 | 19.5 | 59.9 | 5.0 | 13.3 | 35.5 | 4.0 | 7.1 | 16.7 | 2.0 | 4.4 | 8.3 | 2.0 | 3.7 | 8.6 | 1.0 | 1.8 | 5.2 | 0.0 |
| (0.6, 0.7] | 7.5 | 33.3 | 1.0 | 20.6 | 67.2 | 5.0 | 12.3 | 27.5 | 5.0 | 7.1 | 19.8 | 3.0 | 4.3 | 8.8 | 1.0 | 3.4 | 6.4 | 1.0 | 1.6 | 4.0 | 0.0 |
| (0.7, 0.8] | 7.5 | 29.7 | 1.0 | 16.8 | 54.2 | 5.0 | 16.1 | 38.8 | 5.0 | 7.1 | 14.5 | 2.0 | 4.3 | 8.4 | 1.0 | 4.0 | 7.6 | 1.0 | 3.2 | 20.4 | 0.0 |
| (0.8, 0.9] | 7.4 | 28.2 | 0.0 | 17.8 | 50.9 | 4.0 | 12.3 | 22.4 | 4.0 | 5.9 | 9.1 | 3.0 | 3.9 | 6.4 | 2.0 | 2.8 | 5.5 | 1.0 | 0.8 | 1.8 | 0.0 |
| (0.9, 1.0] | 6.0 | 32.8 | 1.0 | 18.7 | 78.0 | 5.0 | 11.3 | 35.3 | 4.0 | 6.8 | 17.3 | 2.0 | 4.3 | 11.7 | 1.0 | 3.5 | 8.9 | 1.0 | 1.8 | 7.0 | 0.0 |

| (b) | ***All*** | | | **A1** | | | **A2** | | | **B1** | | | **B2** | | | **B3** | | | **B4** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Interval | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | | Citations | | |
| Endogamy | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. | Avg. | SD. | Med. |
| [0.0, 0.1] | 41.3 | 196.2 | 9.0 | 44.5 | 148.4 | 12.0 | 12.8 | 43.4 | 3.0 | 91.9 | 526.4 | 9.0 | 13.5 | 40.9 | 2.0 | 4.9 | 18.5 | 0.0 | - | - | - |
| (0.1, 0.2] | 33.2 | 134.3 | 8.0 | 39.2 | 155.0 | 11.0 | 16.2 | 65.1 | 4.0 | 38.2 | 121.6 | 6.0 | 15.8 | 49.2 | 4.0 | 6.9 | 33.5 | 1.0 | 3.1 | 6.1 | 0.5 |
| (0.2, 0.3] | 27.6 | 109.2 | 7.0 | 32.3 | 122.3 | 10.0 | 16.7 | 77.8 | 5.0 | 31.7 | 113.0 | 6.0 | 14.0 | 33.8 | 3.0 | 6.3 | 15.2 | 1.0 | 1.4 | 1.9 | 0.0 |
| (0.3, 0.4] | 24.0 | 83.3 | 6.0 | 29.4 | 99.4 | 9.0 | 16.0 | 45.9 | 4.0 | 20.1 | 58.2 | 4.0 | 16.3 | 69.9 | 3.0 | 8.1 | 20.7 | 1.0 | 2.0 | 4.9 | 0.0 |
| (0.4, 0.5] | 21.3 | 74.6 | 6.0 | 27.0 | 90.7 | 9.0 | 14.8 | 40.0 | 4.0 | 16.7 | 57.8 | 4.0 | 14.5 | 54.7 | 3.0 | 8.3 | 30.3 | 1.0 | 4.7 | 16.9 | 0.0 |
| (0.5, 0.6] | 22.1 | 60.0 | 6.0 | 25.6 | 56.2 | 8.0 | 13.4 | 23.0 | 5.0 | 26.3 | 87.3 | 5.0 | 13.8 | 68.4 | 2.0 | 7.9 | 21.2 | 1.5 | 2.7 | 4.1 | 1.0 |
| (0.6, 0.7] | 20.2 | 59.7 | 5.0 | 26.2 | 69.3 | 8.0 | 11.7 | 28.8 | 4.0 | 17.9 | 66.3 | 4.0 | 9.2 | 21.5 | 2.0 | 5.0 | 10.2 | 0.5 | 4.4 | 9.5 | 0.0 |
| (0.7, 0.8] | 18.1 | 42.0 | 6.0 | 23.2 | 50.7 | 8.0 | 11.2 | 22.4 | 4.0 | 14.1 | 32.4 | 3.0 | 11.2 | 23.7 | 3.0 | 1.6 | 3.1 | 0.0 | 4.3 | 6.0 | 2.0 |
| (0.8, 0.9] | 22.7 | 69.2 | 7.0 | 23.2 | 35.0 | 11.0 | 7.0 | 12.0 | 2.0 | 88.5 | 199.0 | 11.0 | 4.3 | 6.6 | 2.0 | - | - | - | - | - | - |
| (0.9, 1.0] | 17.6 | 90.6 | 4.0 | 23.9 | 121.2 | 7.0 | 13.6 | 42.2 | 4.0 | 14.2 | 70.0 | 3.0 | 10.7 | 35.4 | 2.0 | 7.3 | 23.4 | 1.0 | 3.3 | 9.0 | 1.0 |

for publications can be extended to consider the subareas of knowledge given here by patent subcategories.

# 7. CONCLUDING REMARKS

We have presented two novel metrics called *C-Endo* and *Comb* for qualifying publications and patents by computing the endogamy based on authors and their communities. We have validated the metrics and explored them as influence indicators through an experimental analysis which also included random sampling. Although the good results, we understand that the metrics cannot individually assess quality or influence, since quality is defined by the work content. Nonetheless, both *C-Endo* and *Comb* are one step closer to more robust production evaluation mechanisms. As future work, we plan to expand our approaches by considering other indicators from the researchers' areas as communities.

**Acknowledgements.** This work has been partially funded by CNPq and FAPEMIG, Brazil.

# References

[1] J. Bollen et al. A Principal Component Analysis of 39 Scientific Impact Measures. *CoRR*, abs/0902.2183, 2009.

[2] K. Börner et al. Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*, 10(4):57–67, 2005.

[3] R. S. Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[4] H. Deng et al. Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In *JCDL*, pages 71–80, Washington, DC, USA, 2012.

[5] M. S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[6] R. Guimerà et al. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performanced. *Science*, pages 697–702, 2005.

[7] B. H. Hall, A. B. Jaffe, and M. Trajtenberg. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Technical Report NBER Working Paper No. 8498, National Bureau of Economic Research, 2001.

[8] B. H. Hall, A. B. Jaffe, and M. Trajtenberg. Market Value and Patent Citations. *The RAND Journal of Economics*, 36(1):16–38, 2005.

[9] M. Kato and A. Ando. The relationship between research performance and international collaboration in chemistry. *Scientometrics*, 97(3):535–553, 2013.

[10] M. Kendall. A New Measure of Rank Correlation. *Biometrika*, 1938.

[11] A. H. F. Laender et al. Building a Research Social Network from an Individual Perspective. In *JCDL*, Ottawa, Canada, 2011.

[12] H. Lima et al. Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas. In *JCDL*, pages 97–106, Indianapolis, USA, 2013.

[13] G. R. Lopes et al. Ranking Strategy for Graduate Programs Evaluation. In *ICITA*, pages 253–260, Sydney, Australia, 2011.

[14] G. R. Lopes et al. Scientific Collaboration in Research Networks: A Quantification Method by Using Gini Coefficient. *IJCSA*, 9(2):15–31, 2012.

[15] S. L. Montolio, D. Dominguez-Sal, and J. L. Larriba-Pey. Research Endogamy as an Indicator of Conference Quality. *SIGMOD Rec.*, 42(1):11–16, July 2013.

[16] M. E. Newman. Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks. In *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 337–370. Springer, 2004.

[17] M. E. J. Newman. Coauthorship Networks and Patterns of Scientific Collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5200–5205, 2004.

[18] L. Soulier et al. BibRank: A Language-based Model for Co-ranking Entities in Bibliographic Networks. In *JCDL*, pages 61–70, Washington, DC, USA, 2012.

[19] R. Yan et al. To Better Stand on the Shoulder of Giants. In *JCDL*, pages 51–60, Washington, DC, USA, 2012.