Research Endogamy as an Indicator of Conference Quality

Sergio Lopez Montolio, David Dominguez-Sal, Josep Lluis Larriba-Pey
DAMA-UPC
Universitat Politecnica de Catalunya, Barcelona Tech
Barcelona
{slopez,ddomings,larri}@ac.upc.edu

ABSTRACT

Endogamy in scientific publications is a measure of the degree of collaboration between researchers. In this paper, we analyze the endogamy of a large set of computer science conferences and journals. We observe a strong correlation between the quality of those conferences and the endogamy of their authors: conferences where researchers collaborate with new peers have significantly more quality than conferences where researchers work in groups that are stable along time.

1. INTRODUCTION

Social sciences define endogamy as "the custom of marrying only within the limits of a local community, clan, or tribe"[1]. We can extend this concept to measure the degree of collaboration between persons. In the context of scientific publications, we consider endogamy as the inclination of a person or a group to usually collaborate (i.e., publish papers) within a small group of selected people.

Coauthorship networks represent authors as nodes in a graph and edges linking people who coauthor a paper. They provide information about how the researchers cooperate to produce new ideas [11]. It is known that not all collaborations have an equal impact, and some of them produce higher research impact [2]. Furthermore, Guimerà et al. studied a small set of journals and found that endogamy is a significant factor in the performance of research teams in some research fields such as social psychology or ecology [6]. The collaborations with new researchers open new streams of ideas, and hence are a positive indicator of good research.

In this paper, we go further in the study of the endogamy in computer science collaborations. We apply this endogamy to calculate the endogamy of a broad spectrum of computer science conferences (926) and journals (317). We observe that there is a strong influence of the endogamy of the research teams publishing in a conference on the quality of such conference (up to 80% agreement with the ERA conference ranking[2]). This shows the social importance of conferences for computer scientists, where they are able to meet new peers that in turn lead to better publications. In particular, reputed conferences such as PODS, ICDT, SIGMOD, VLDB or ICDE stand out among database conferences as having particularly low endogamy. Although this collaborative strategy works well for conferences, it is not universal, because we found that computer science journals are not affected by endogamy alike.

The correlation found between the endogamy and the quality of conferences opens the possibility to consider having metrics to evaluate the quality of a conference that are based on the social aspects of research. Currently, the evaluation of conferences relies mostly on measures based on the citations: h-index, cites per paper, pagerank, etc. [1, 5] and in few occasions (e.g. program committee relations [14]) personal relations are analyzed. But, the extraction of cites is not an easy task [3] and error free citation collection requires a large manual effort. Furthermore, the median age of citation is several years (e.g. the median age for TODS is over 10 years [13]), which delays the release of reliable qualifications for conferences and journals. In contrast, coauthor networks are easy to obtain and they describe the current information without delay. Although social metrics cannot be used to evaluate the content of an article because scientific excellence is determined by article's content and not by authors' profiles, social metrics can be computed to obtain early estimates of the quality of recent conferences.

We define the endogamy in Section 2. Then, we describe the experimental environment in Section 3. After computing the endogamy for all the available

[1] `http://oxforddictionaries.com/definition/endogamy`

[2] Previously known as CORE. Available at `http://www.arc.gov.au/era/era_2010/archive/default.htm`

journals and conferences in our dataset, we evaluate the results for conferences in general in Section 4, and for database conferences in Section 5. Finally, we evaluate analyze the endogamy of journals in Section 6.

2. ENDOGAMY COMPUTATION

Research is based on the proposal and study of new ideas. The collaboration with researchers external to the usual research team is a very good means to introduce such new ideas and allow merging the expertise from multiple fields. In this paper, we quantify this degree of new collaborations by means of a new indicator called endogamy.

We compute the endogamy of a set of authors as the inclination of a person or a group to usually collaborate (i.e., publish papers) within a small group of selected people as:

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

where A is a set of authors, and $d(A)$ is the set of papers that were published by the *full* set of authors, in other words, papers coauthored by all the members of A. For example, consider the endogamy of a group formed by authors x and y, who have individually published three papers ($d(\{x\}) = \{a, b, c\}$ and $d(\{y\}) = \{b, c, d\}$). Since they have collaborated in half of their publications their endogamy, $Endo(\{x, y\})$, is: $2/4 = 0.5$

Endogamy of a paper: Let $A(p)$ be the set of authors of a paper p and $L_i(p) = \mathcal{P}_i(A(p))$ be the power set of authors of size i (the set of all subsets with size i within $A(p)$). Then, $L(p) = \bigcup_{i=2}^{i=|A|} L_i$ is the set of all the subsets with more than one author. We compute the endogamy of a paper p, as the aggregation of the endogamies of $L(p)$. We test several endogamy aggregations:

- `Max`: Maximum of the endogamies of all groups:

$$Endo(p) = max_{x \in L(p)}(Endo(x))$$

- `Min`: Minimum of the endogamies of all groups:

$$Endo(p) = min_{x \in L(p)}(Endo(x))$$

- `Med`: Median of the endogamies:

$$Endo(p) = med_{x \in L(p)}(Endo(L_i))$$

- `Avg`: Arithmetic mean of the endogamies:

$$Endo(p) = \frac{\sum_{x \in L} Endo(x)}{|L|}$$

	Conferences	Journals
A/A*	223	122
B	308	87
C	395	108
Total	926	317

Table 1: Conferences and journals by tier.

- `Harm`: Harmonic mean of the endogamies within $L(p)$:

$$Endo(p) = harm(\{Endo(x) | x \in L(p)\}),$$
$$where\ harm(X) = \frac{|X|}{\sum_{x \in X} \frac{1}{x}}$$

- `Avg size`: Arithmetic mean of the endogamies of the subsets of authors grouped by size:

$$Endo(p) = \frac{1}{|A| - 1} \cdot \sum_{i=2}^{i=|A|} \frac{\sum_{x \in L_i(p)} Endo(x)}{|L_i(p)|}$$

- `Harm size`: Harmonic mean of the endogamies of the subsets of authors grouped by size:

$$Endo(p) = harm(\{harm(L_i(p)) | 2 \le i \le |A|\})$$

Endogamy of a conference/journal: Let C be the set of articles published in a conference or a journal. We compute the endogamy as the average endogamy of its papers:

$$Endo(C) = \frac{1}{|C|} \sum_{p \in C} Endo(p) \quad (2)$$

Endo must not be seen as an absolute value of the research quality of a group of people. Indeed, the quality of an individual paper cannot be computed by simply stating the persons who wrote it. High quality research relies on good scientific content, which can be potentially written by any person. *Endo* should be seen as a probability distribution of the quality of a paper. The *Endo* value associated to a group is a number between 0 and 1. An *Endo* value close to 1 indicates that the paper is not likely to bring new ideas because the authors are not working with other members of the community. Values close to 0 show that the researchers constantly collaborate with new researchers, and thus they are more likely to introduce new ideas.

3. EXPERIMENTAL ENVIRONMENT

In order to study the influence of the endogamy of authors on the quality of conferences and journals, we rank the computer science conferences and journals available in the DBLP database[3] by their *Endo* value[4]. In order to verify the quality of the ranking, we take the quality indicators published by the project Excellence in Research for Australia (ERA) as reference. We take the ERA evaluation performed in 2010, which ranks conferences and journals in three categories: A, B and C. In this classification, publications in category A are better than publications in category B, and publications in category B are better than publications in category C. Since the titles in DBLP and ERA are not normalized, we only select those conferences and journals that appear in both datasets with exactly the same title or acronym. After this process, we retrieve 926 conferences and 317 journals that belong to all the three ranks of ERA as shown in Table 1.

We report the degree of similarity between the ERA and *Endo* rankings by means of the agreement between both series. Given the two rankings, a pair of conferences c_1 and c_2 is concordant if $c_1 > c_2$ for both rankings (and by symmetry $c_1 < c_2$ for both rankings). Otherwise, the pair is discordant. We compute for all pairs of conferences (or journals) in the dataset, the number of concordant pairs p, and the number of discordant ones f (ties are not considered). The following percentage ratio computes the *agreement* between both rankings:

$$\rho = 100 \cdot \frac{p}{p+f} \tag{3}$$

We verify the statistical significance of our results by means of the Kendall tau [12], which is a non parametric test that measures the rank correlation between two lists without making assumptions of the sorting method, and ANOVA, which is suited for comparing different configurations of our metric using the R statistical package[5].

4. CONFERENCE ANALYSIS

We ranked the conferences using the six described variants of *Endo*. In this first experiment, we removed entities with low activity: those conferences with less than 500 papers in all their history. With this, we ended up with a total of 241 conferences to be used for the first experiment. We show later that the conclusions are the same if no cleanup is performed. The dark series of Figure 1 shows the agreement for each aggregation technique. We observe that the ranking of conferences performed by *Endo* has a very strong agreement with those of ERA independently of the aggregation performed. By means of the Kendall Tau coefficient test, we found that such correlations are statistically significant for all the aggregation techniques. Among them, `Max` and `Avg` are the best aggregation techniques. This corresponds to selecting the most endogamous group of authors, or average endogamy of all subsets of authors, respectively.

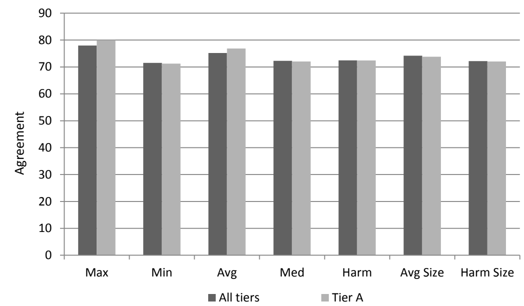

Figure 1: Agreement ρ for conferences with more than 500 papers.

We also consider the case of deciding whether a conference is a top tier (A) or a non top tier conference (B and C) according to ERA. We depict the agreement with this binary decision in the light series of Figure 1 showing that it also correlates well, being the influence statistically significant considering the Kendall coefficient.

We observed that depending on the conference tier, the distribution of *Endo* changes. We illustrate this change as a boxplot in Figure 2 for the conferences in the previous experiment, where we depict *Endo* using `Avg` with respect to the ERA tier. Note that the median *Endo* increases as we lower the conference quality, and the median *Endo* of a tier is lower than the first quartile of the next ranked tier consistently.

We verify the significance of the differences by means of an ANOVA test. We first performed a random sample of 50 conferences of each tier, adding up 150 conferences in total and compared their *Endo* in logarithmic scale. The ANOVA allows us to conclude that there exists statistically significant differences between the three tiers considered with respect to *Endo*. In order to improve the confidence of our statistical analysis, we applied resampling. We selected ten new samples, where each sample contains 50 conferences in each tier, and recomputed the ANOVA procedure. In all the cases, the results showed significant differences between tiers,

[3]`http://www.informatik.uni-trier.de/~ley/db`

[4]When we compute of a paper p using Equation 1, we consider only collaborations performed before the publication date of p. So, we do not introduce unavailable information about subsequent collaborations after p was published.

[5]All statistical test in the paper are performed with confidence level $\alpha = 0.05$

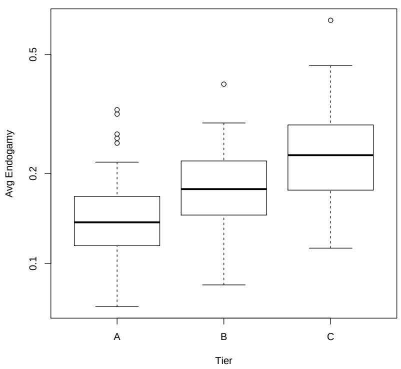

Figure 2: *Endo* per conference tier using `Avg`.

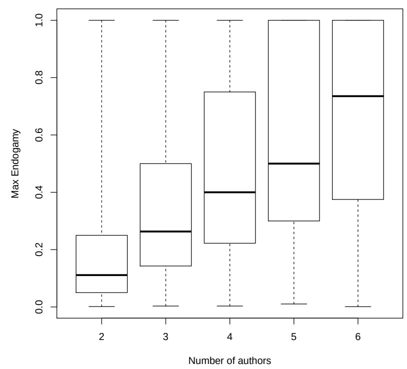

Figure 3: *Endo* using `Max` vs. authors of a paper.

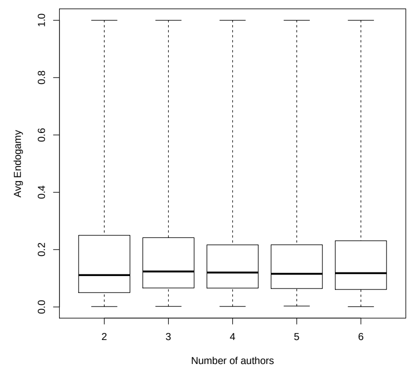

Figure 4: *Endo* using `Avg` vs. authors of a paper.

and thus, we conclude that each conference tier has a characteristic *Endo*. The different pairs of tiers have been compared using a Tukey's test, concluding that for any pair of tiers their *Endo* is statistically different.

Impact of parameters in *Endo*: We observed that `Max` and `Avg` are the best candidates to be considered as quality indicators of conferences. After verifying the significance of their predictions (we showed in the previous section the results for `Avg` and for space reasons we do not report those for `Max`), we proceed to analyze with more detail the impact of the variables involved in the computation of *Endo*.

First, we analyze the impact of the number of authors in the computation of the endogamy of a paper. We separate the papers in groups by the number of authors and plot *Endo* for each paper in the group as a boxplot in Figures 3 and 4. We expected that the number of authors would not be relevant for the quality of the paper. We found that despite the higher precision of `Max`, the value of *Endo* obtained with it depends on the number of authors of a paper: more authors imply larger *Endo*. `Max` takes into account only the most endogamic group, and with more authors there are more subgroups that may have large endogamy. On the other hand, Figure 4 shows an homogeneous distribution of endogamies for `Avg` no matter the number of authors. We conclude that `Max` gives biased results between conferences with different distributions of authors but this is not the case for `Avg`. Therefore, in the following experiments we focus on `Avg`.

In our next experiment, we study if the number of papers of a conference and the number of papers per author have an impact in the accuracy of *Endo* as a predictor. We set five levels for each variable:

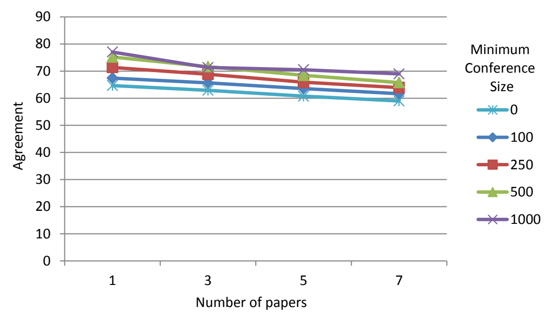

Figure 5: Agreement ρ for *Endo* using `Avg`. Series are the minimum count of papers of a conference. The X-axis is the minimum number of papers for a group of authors.

we study conferences with any number of papers and a minimum of 100, 250, 500 and 1.000 papers; and filter groups of authors with at least 1, 3, 5 and 7 papers. This produces twenty configurations in a full factorial design, which are plotted in Figure 5. We observe a defined trend for each variable. First, we observe that considering authors with few papers (novel authors) improves the accuracy of *Endo*. This result suggests that the impact of non experienced researchers in research teams is not negligible. Since people who publish for the first time reduce the endogamy of the research team, these results suggest that the inexperience of new researchers is overcome by the novelty of ideas that they can provide. With respect to conference size, we see that for conferences with a large number of papers the agreement is larger.

Both trends indicate that the more observations are taken into account (and thus the endogamy of more papers and more authors), the better the pre-

Conference	Tier	Avg. Endo
PODS	A	0.083
ISIT	B	0.085
EDBT	A	0.095
ICDE	A	0.108
PKDD	A	0.120
SIGMOD	A	0.122
MDM	C	0.126
ASIACRYPT	A	0.132
DASFAA	A	0.132
PAKDD	A	0.133

All history

Conference	Tier	Avg. Endo
PODS	A	0.058
CRYPTO	A	0.062
ICDT	A	0.065
DBLP	B	0.065
SIGMOD	A	0.073
EUROCRYPT	A	0.073
VLDB	A	0.077
EDBT	A	0.079
ASIACRYPT	A	0.080
ICDE	A	0.081

Years 2003-2012

Figure 6: "Data Format" conferences in ERA with the lowest *Endo*.

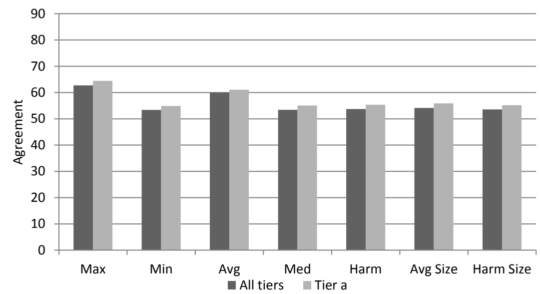

Figure 7: Agreement ρ for journals with more than 100 papers

diction power of *Endo*. As more papers are aggregated, the trends for *Endo* are stronger as a consequence of the law of large numbers.

5. DATABASE CONFERENCES

For this section, we focus on the set of conferences marked as "Data Format" in the ERA list. We computed the *Endo* value of all these conferences and ranked them. In Figure 6, we report the top 10 conferences in terms of *Endo*. We computed two result sets: the one on the left considers all the editions performed by the conferences, and the one on the right only accounts for the last ten years. We find that both lists contain a majority of conferences of excellence: on the left and on the right, 8 and 9 out of the 10 conferences classified belong to tier A, respectively. *Endo* is able to distinguish the most relevant conferences in the area: PODS, ICDE, SIGMOD, EDBT, VLDB, ICDT... Most of them appear in both lists showing the correlation of *Endo* and the quality of database conferences.

We found that time is a relevant factor in computing the endogamy of database conferences, as can be seen comparing both lists. In absolute terms, the endogamy of the latest years is considerably smaller than twenty or thirty years ago. The reason is that the database field has been a popular one and the number of authors has grown in the latest years, which provides a potentially larger number of collaborations. For example, the number of different authors that have published in SIGMOD in the last decade (2003-2012) is 2,349 compared to 1,465 in the previous decade (1993-2002).

On the other hand, conferences (and in particular those considered as the best) tend to have a worse *Endo* in the first editions and reduce their *Endo* along time. One example is VLDB that in the first five editions had endogamies above 0.4, which is significantly larger to the average of the latest ten years, 0.077. For this reason, VLDB is classified in the 11th position in the left list and does not appear in the list. We detected similar patterns for SIGMOD, ICDE, EDBT or DASFAA, just to mention a few. This pattern seems more correlated to the longevity of the conference rather than the exact year because conferences starting in 70's, 80's and 90's show such a lowering trend, as discussed with more detail in [8]. According to these results, the evaluation of a recent window of years provides more accurate tier predictions by *Endo*.

6. JOURNAL ANALYSIS

For the journals included in DBLP and ERA lists, we initially expected a similar influence of endogamy. However, after performing the same procedures we observed that the endogamy is not strongly influenced by the quality of the journal. In fact, with a study similar to that for the conference analysis, we obtained a maximum agreement of 62% (Figure 7). Although this number indicates some correlation, there is a big difference between the agreement of journals and conferences.

These results show that there is a behavior change in the way that people collaborate for publishing in journals, which can be explained in terms of previous studies. A recent survey among 22 editors from major software engineering journals report a general agreement that many publications in journals have archival intention and not innovative objectives [10]. Laender et al. [7] indicate that most journal papers have a conference prelude. Furthermore, the fraction of papers in journals that extend previous conference works has been estimated around 30% on average [4, 10], and for some journals it has been observed above 50% [10]. Since many works in journals focus on deeper analysis of previous ideas, journal publications benefit from groups of authors that have already collaborated. Therefore, we believe that the lower influence of endogamy in the case of journals is explained by a large set of journal papers from authors that collaborate again to extend ideas already presented in conference papers. For those journal papers, the endogamy approach is not indicative and alters the results.

7. CONCLUSIONS

The analysis introduced in this paper suggests that endogamy is a fundamental factor in understanding the generation of new scientific knowledge. The impact of social behavior in science is still a relatively unexplored topic, whose deeper understanding could be used to improve the efficiency in research innovation and effective team formation.

We observe that papers published in highly reputed conferences are published by groups of authors with low endogamy. On the other hand, low quality conferences tend to publish articles where authors have collaborated in many occasions. This stresses the importance of social contact in research and the opportunity that conferences offer to exchange new ideas and start collaborations.

We have also observed that high impact research in computer science does not have a unique strategy. Journal impact is not affected by endogamy in contrast to results in other research areas [6]. Although this seems a peculiar consequence of the extended versioning and archival focus of many computer science journals, we believe that it will be interesting to analyze the factors that determine the impact in computer science journal papers.

Our results show that endogamy could be used as a feature for determining the quality of conferences and, in particular, this applies to database conferences [9]. The endogamy of a group of authors can be computed when the paper is just published, in contrast to the number of citations to a paper, which may require years to be collected. Since an evaluation metric relying only on endogamy could be easily abused by dishonest conferences (by simply accepting papers that have small endogamy) we believe that endogamy should be taken as a complement to other metrics to obtain fast evaluation of conferences. An interesting research topic could be whether it is possible to design metrics based on endogamy which are difficult to flaw.

Acknowledgements

The authors thank the Ministry of Science and Innovation of Spain for grants TIN2009-14560-C03-03, PTQ-11-04970; and Generalitat de Catalunya for grant GRC-1087.

8. REFERENCES

[1] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009.

[2] K. Borner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):57–67, 2005.

[3] E. Cortez, A. da Silva, and Gonçalves et al. FLUX-CIM: Flexible unsupervised extraction of citation metadata. In *Proc. JCDL*, pages 215–224, 2007.

[4] M. Eckmann, A. Rocha, and J. Wainer. Relationship between high-quality journals and conferences in computer vision. *Scientometrics*, 90(2):617–630, 2012.

[5] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108, 1955.

[6] R. Guimerà, B. Uzzi, J. Spiro, and L. Nunes. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308:697–702, 2005.

[7] A. Laender, C. de Lucena, et al. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40(2):135–145, 2008.

[8] S. Lopez-Montolio. Research endogamy as an indicator of conference quality. *UPC Master Thesis*, 2013.

[9] W. Martins, M. Gonçalves, et al. Learning to assess the quality of scientific conferences: a case study in computer science. In *Proc. JCDL*, pages 193–202, 2009.

[10] M. Montesi and J. Owen. From conference to journal publication: How conference papers in software engineering are extended for publication in journals. *J. Am. Soc. Inf. Sci. Technol.*, 59(5):816–829, 2008.

[11] M. Newman. Coauthorship networks and patterns of scientific collaboration. *Nat. Ac. Sc. USA*, 101(1):5200–5205, 2004.

[12] R. Porkess. *Statistics defined and explained*, page 64. Collins, 2005.

[13] E. Rahm and A. Thor. Citation analysis of database publications. *SIGMOD Record*, 34(4):48–53, 2005.

[14] Z. Zhuang, E. Elmacioglu, D. Lee, and C. Giles. Measuring conference quality by mining program committee characteristics. In *Proc. JCDL*, pages 225–234, 2007.