Empirical comparative study of similarity indexes in scientometrics co-authorship analysis

Hamid Bouabid
Faculty of Science, Mohammed V University in Rabat,
Agdal, Rabat, MOROCCO
e-mail: h.bouabid@um5r.ac.ma
ORCID ID: H.Bouabid: 0000-0001-7965-7870

ABSTRACT

Similarity indexes are widely used in the field of scientometrics either in co-words, co-citations, bibliographic coupling, or co-authorship, and very recently in link prediction and system recommender. Despite the rich literature on the comparison of various indexes very rarely a consensus is being reached on the appropriateness of a specific one. This paper aims to enhance empirical understanding of similarity indexes within the context of co-authorship networks, which are widely used and highly relevant in scientometrics. The objective is to assist scientometricians in better analyzing co-authorship networks and selecting the most suitable similarity index for their studies. The research examines two types of co-authorship networks - one with low density at the individual level and another with high density at the country level - using five commonly applied similarity indexes: Jaccard, Salton, Dice-Sorenson, Pearson, and Association Strength. The study confirms that, as theoretically expected, the Salton index follows a concave increasing function of the Jaccard index, with Jaccard values consistently lower, regardless of network density. The concave shape of the curve is more pronounced in the case of low dense network. A linear function is found between Dice-Sorenson and Salton. Additionally, Pearson is observed to be 'orthogonal' to Jaccard, Salton, and Dice-Sorenson, indicating a lack of direct correlation. In contrast, Association Strength behaves differently: in a high-density network, it is 'orthogonal' to Jaccard, Salton, and Dice-Sorenson and shows no correlation with Pearson. However, in a low-density network, Association Strength displays the opposite behavior.

Keywords: Co-authorship networks; Scientometrics; Similarity indexes; Network analysis: Scientific collaboration.

INTRODUCTION

The ‘similarity’ concept, is widely used as a mathematical approach to quantitatively assess similarity or proximity between entities - in a group, system or cluster - according to specific characteristics. It has early found its way to the field of scientometrics and has been enhanced by various tools for science mapping (Sternitzke & Bergmann, 2009), system recommender (Linyuan et al., 2012) and link prediction in co-authorship networks (Lü et al., 2012; Chuan et al., 2018). Similarity metrics are fundamental to compute link predictors based on either nodes or vertices (or both) features. Chuan et al. (2018) stated that a similarity index used as a metric for link prediction in co-authorship may adapt better for a specific case, but may be less accurate in another one.

Several studies have examined and compared various similarity indexes, yielding differing results and recommendations. However, there is no clear consensus on which index is most suitable for determining similarity or proximity between entities within a group (Hamers et al., 1989; Ahlgren et al., 2003, White, 2003; White, 2004; Bensman, 2004; Egghe, 2009; Van Eck & Waltman, 2009). Furthermore, for the purpose of science mapping, different indexes reveal varying perspectives on the research landscape (Sternitzke & Bergmann, 2009), leading to diverse interpretations, conclusions, implications, and recommendations. These findings are crucial for decision-making based on co-authorship network analysis, whether viewed from an individual (Bouabid & Achachai, 2021), institutional, or international perspective (Wei et al., 2017).

A recent work by Adnani et al. (2020) conducted a comparative analysis of the five most used similarity indexes, namely, Jaccard, Dice-Sorensson, Salton's Cosine, Pearson, and Association Strength, for the three common scientometric analysis types: co-word, co-citation and co-authorship. Among other findings, this work showed that (a) there is still no common agreement on the appropriateness of an index for co-authorship analysis (beside co-word analysis), (b) and that the Association Strength is the less covered and compared to other indexes for the analysis, in both theoretical and empirical levels in all analyses, particularly co-authorship.

The aim of this paper is to enhance the empirical understanding of similarity indexes in co-authorship analysis, particularly in co-publication, a key aspect of network-based studies in scientometrics. By assisting scientometricians, this paper emphasises the suitability of major similarity indexes in analyzing co-authorship networks, offering insights that lead to a more accurate and unbiased interpretation of scientific collaboration patterns, which are crucial for evaluating and mapping scholarly relationships.

MATERIALS AND METHODS

Corpus of analysis

This empirical study considered two distinct cases of co-authorship networks. The networks were selected to represent diversity in terms of scope - micro-level (individual) and macro-level (country) - as well as consistency, with one being low-density and the other high-density. Both networks were constructed based on co-authored papers. Indeed, the difference between the two networks is that the first one is at the 'individual level' (nodes refer to authors), while the second one is at the 'aggregated level', i.e. country (nodes refer to countries). The first network was chosen due to its composition of highly prolific authors from various countries, but with relatively few co-authored papers. The co-authorship network among scientometricians and informetricans was selected as a case study because this community tends to form a "small-world" topology (Erfanmanesh et al., 2012; Abrizah et al. 2014), characterised by a small number of co-authored papers (few vertices), making it an ideal index-sensitive network for analysis.

The second case represents a dense network, largely attributed to the successive and consistent European Framework Programmes for Research and Innovation. These programmes, which have been one of the world's primary instruments for implementing common scientific and innovation policies, have operated in four-year cycles since the first framework in 1984. The latest, "Horizon Europe" (2021–2027), has fostered strong and close collaborative scientific partnerships between European countries (Wagner and Leydesdorff, 2003; Balland et al., 2019).

The 14 scientometrics and informetrics scientists (Table 1, Case 1) were selected based on their prolific publication records in relevant journals and ISSI conference proceedings, all indexed in the Web of Science, for the period from 2010 to 2019.

- *Journal of Informetrics;*
- *Scientometrics;*
- *Journal of the Association for Information Science and Technology;*
- *Research Evaluation;*
- *Research Policy;*
- *Proceedings of the International Society of Scientometrics and Informetrics (ISSI) Conference*

On the other hand, the second network is based on co-authored papers (co-publications) among 28 European countries [including the UK, which was part of the EU during the study period] (Table 1, Case 2). These papers were retrieved from Web of Science for the period from 2016 to 2018.

Table 1: List of the Entities (Nodes) in the Networks under Study

a) Case 1 : List of top 14 scientometrics and informetrics' scientists (in alphabetical order and in capital letters)	
ABRAMO G	GUAN JC
BORNMANN L	HUANG MH
CHEN DZ	LARIVIERE V
D'ANGELO CA	LEYDESDORFF L
DING Y	PRATHAP G
EGGHE L	ROUSSEAU R
GLANZEL W	THELWALL M

b) Case 2 : List of 28 EU countries (in alphabetical order)			
Austria	Estonia	Italy	Portugal
Belgium	Finland	Latvia	Romania
Bulgaria	France	Lithuania	Slovakia
Croatia	Germany	Luxembourg	Slovenia
Cyprus	Greece	Malta	Spain
Czech	Hungary	Netherlands	Sweden
Denmark	Ireland	Poland	United Kingdom

The matrices of co-authorship for the two networks are given in Appendices 1 and 2. Table 2 shows the characteristics of the two networks with regard to number of entities, number of co-occurences (co-authored papers) and number of zeros co-occurences in the raw matrix. The raw matrix of co-authored papers represents shared papers during the decade from 2010 to 2019 extracted from the Web of Science Core Collection database by Clarivate Analytics.

Table 2: Characteristics of the Two Cases under Study

	Case 1: Individual co-authorship	Case 2: Country co-authorship
Number of entities	14	28
Number of co-occurences	(14*13) = 182	(28*27) = 756
Zeros co-occurences in the matrix: number & (%)	166 (91.2%)	0 (0%)

The selection of these two distinct co-authorship network formats aims to enhance understanding of the suitability of different similarity indexes across varied cases. This approach highlights the characteristics of the data for which each index is most effective. Previous research has demonstrated that the effectiveness of a similarity index depends on the specific characteristics of the data to which it is applied (Schneider & Borlund, 2007).

Similarity indexes

The five indexes considered in this study are Jaccard, Dice-Sorenson, Salton's Cosine, Pearson and Association Strength. These five indexes have been addressed in the comparative study by Adnani et al. (2020). Jaccard, Salton's Cosine, Dice-Sorenson, and Association Strength (known also as Proximity Index) were chosen because they are the most widely used indexes in the field of scientometrics (Van Eck & Waltman, 2009). Dice-Sorenson, somewhat similar to Jaccard index, was chosen as it is one of the oldest similarity measures (1945) that is still widely used in the scientometrics field. Pearson was considered because it is an indirect similarity assessment index, a vector-space index and is the only index to describe both similarity and dissimilarity using a -1 to +1 range (Pearson, 1895). Association Strength is the index that has no vector-variant formula and was given a probabilistic conceptualization by Van Eck & Waltman (2009).

For the analysis, non-vector formulas (Euclidean formulas) of the indexes were utilised. Note that Association Strength lacks a vector form, unlike the other four indexes. The mathematical formulations of these similarity indexes are detailed below.

For all equations, ***n*** represent the number of entities within a group (words, authors, institutions, countries, fields, journals, citations, etc), ***i*** and ***j*** are two entities for which similarity is to be assessed; X_{ij} is the number of co-occurrences of both entity ***i*** and entity ***j***. Y_{it} and Y_{tj} are the total numbers of co-occurrences of entities ***i*** and ***j*** respectively, where the indice ***t*** refers to the total number of co-occurences for a given entity:

$$Y_{it} = \sum_{j=1}^{n} X_{ij} \qquad\qquad Y_{tj} = \sum_{i=1}^{n} X_{ij} \tag{1}$$

The matrix $\lfloor X_{ij} \rfloor$ may be symmetric or not. For each index, the vector and non vector formulas are presented in Adnani et al. (2020).

Jaccard (J):
$$J_{ij} = \frac{X_{ij}}{Y_{it} + Y_{tj} - X_{ij}} \tag{2}$$

Dice-Sorenson (D):
$$D_{ij} = \frac{2.X_{ij}}{Y_{it} + Y_{tj}} \tag{3}$$

Salton's Cosine (COS):
$$Cos_{ij} = \frac{X_{ij}}{\sqrt{Y_{it}.Y_{tj}}} \tag{4}$$

Pearson (r):
$$r_{ij} = \frac{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}}).(X_{kj} - \overline{X_{.j}})}{\sqrt{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}})^2} . \sqrt{\sum_{k=1}^{n} (X_{kj} - \overline{X_{.j}})^2}} \tag{5}$$

where $\overline{X_{i.}} = \frac{1}{n}\sum_{k=1}^{n} X_{ik}$ and $\overline{X_{.j}} = \frac{1}{n}\sum_{k=1}^{n} X_{kj}$

Association Strength (AS): $$AS_{ij} = n\frac{X_{ij}}{Y_{it}Y_{tj}} \quad (6)$$

To apply each index to the raw matrix in both cases, the Similarity Index Computation Programme (SICoP), developed by Adnani et al. (2020), was used. This programme automatically generates normalised matrices for each index, facilitating comparisons. The SICoP outputs, in .net format, were then exported to the open-source software Gephi to create the corresponding network maps.

RESULTS AND DISCUSSION

The results of the comparison of each pair of indexes for the two cases 'Individual level' and 'aggregate level' are reported in Figures 1 and 2. The two figures show that Jaccard is always smaller than Salton, regardless of the network size. In the context of the countries (aggregate level) co-authorship, Jaccard produced smaller similarities than Salton. This is also supported by the theoretical findings of Egghe (2009), when comparing eight similarity indexes (including Jaccard, Salton and Dice-Sorenson).

Figures 1 and 2 also reveal that Salton follows a concave (the curve is entirely located below its tangents) increasing function of Jaccard for both individual and country co-authorship cases. A result that has been pointed out also by Egghe (2009) for theorethical calculation and Van Eck and Waltman (2009) for an empirical case.

Considering two vectors $\vec{X}$ and $\vec{Y}$, the relationship between Salton and Jaccard, as established by Egghe (2009), can be rewritten as follows:

$$Cos = \frac{J}{1+J}(\lambda - \frac{1}{\lambda}) \text{ where } \lambda = \sqrt{\frac{\sum_{k=1}^{n}(X_i)^2}{\sum_{k=1}^{n}(Y_i)^2}} = \frac{\|\vec{X}\|}{\|\vec{Y}\|}$$

Figure 2 shows that the concave shape of the curve is more pronounced in the case of low dense network, with few co-authored papers. Obviously, when applying a similarity index to networks, such as a country co-authorship one, underestimation or overestimation of some network links depends on the characteristic of the index being used and the nature of the network itself, unlike the findings of Luukkonen et al. (1993) who reported that Jaccard tends to underestimate the collaboration of smaller countries with larger ones, while Salton tends to underestimates the collaboration among smaller countries.

The other finding that can be highlited from the two figures is the linear function of Dice-Sorenson with Salton. Since Jaccard and Dice-Sorenson are somewhat correlated (Eqs 2 and 3), this aligns with Egghe's (2009) theoretical findings. However, and even if the curve appears to be a linear function, one could see that Dice-Sorenson is always smaller, utmost equal, to Salton (see Figures 1 and 2). This is contrary to the results by Lü and Zhou (2011) when comparing nine similarity indexes (including Salton, Jaccard, Dice-Sorenson) in co-authorship, stating that Jaccard and Dice-Sorenson experienced higher score, while Salton showed a lesser, but a very closer, score to these two indexes.

For both studied co-authorship cases, high and low dense networks, Pearson is 'orthogonal' to Jaccard, Salton and Dice-Sorenson. That is, when Pearson' score is small (almost zero), the three indexes' score is high and vice-versa (see Figures 1 and 2).

Considering two vectors $\vec{X}$ and $\vec{Y}$, Egghe and Leydesdorff (2008) establish the relation between Salton and Pearson as follows :

$$r = \frac{n}{\sqrt{n-a^2}.\sqrt{n-b^2}}(Cos - \frac{ab}{n}) \text{ where } a = \frac{\sum_{k=1}^{n} X_i}{\sqrt{\sum_{k=1}^{n}(X_i)^2}} \text{ and } b = \frac{\sum_{k=1}^{n} Y_i}{\sqrt{\sum_{k=1}^{n}(Y_i)^2}}$$

Since, ***a*** and ***b*** vary, the relation between ***r*** and ***Cos*** is not a functional relation. Egghe and Leydesdorff (2009) have found the relation as an increasing cloud of points (sheaf). The empirical results in Figure 1 and 2 show that this relation is somewhat orthogonal. This finding is observed in the maps shown in Figure 3. As an example, bold links as expressed by Jaccard, such as that for Glanzel-Guan and Thelwall-Ding in Case 1 or for Bulgaria-Croatia and Bulgaria-Slovenia in Case 2, become weak links in favor of a completely other distinct links expressed by Pearson such as Bornemann-Prathap and Guan-Egghe in Case 1 or Poland-Hungary and Denmark-Finland in Case 2.

Furthermore, one could see that in low dense network, Pearson amplifies dissimilarity. Indeed, Table 1 shows that Pearson experiences negative scores, thus amplifying dissimilarities within the network, contrary to the second case of the dense network, recalling that it is the only index, out of the five studied here, for which the value ranges in the interval [-1, +1]. In this latter case of dense network, the average clustering coefficient is the highest for Pearson index (with a diameter value of 1), which means that this index favors in the global network tightly knit groups, which fits with the model of 'small world' (Watts & Strogatz, 1998, Erfanmanesh et al., 2012).

Table 3: Network Parameters for Cases 1 and 2

Case 1: Authors					
	Average weighted degree	Diameter	Density	Modularity	Average Clustering Coefficient
Jaccard	0.187	4	0.231	0.392	0.700
Salton	0.977	4	0.231	0.514	0.700
Pearson	-0.264	1	1.000	-5.409	1.000
Dice-Sorenson	3.710	4	0.231	0.324	0.700
Association Strength	0.052	4	0.231	0.278	0.700
Case 2: Countries					
Jaccard	4.882	1	1.000	0.206	1.000
Salton	11.686	1	1.000	0.000	1.000
Pearson	9.958	1	1.000	0.005	1.000
Dice-Sorenson	33.829	1	1.000	0.158	1.000
Association Strength	0.000	1	1.000	0.000	1.000

Regarding Association Strength, Van Eck & Waltman (2009), when comparing theoretically and empirically four similarity indexes (Association Strength, Salton, Inclusion index and Jaccard) found that Association Strength was considered the best index, compared to other indexes such as Jaccard or Salton. Association Strength was later reported by Egghe (2010a) to be a function of Jaccard and also of Salton, which evolves into two stages: first convex and then concave.

To check whether an index qualify as a similarity one, Egghe (2010b) introduced two properties to be satisfied by any similarity measure. He considered Jaccard, Dice-Sorenson, Salton, Overlap and Association Strength for his theoretical analysis. The two properties are namely: (a) if adding a constant vector to both vectors, then the similarity must increase, and (b) if one of the two vectors is added to both vectors, then the similarity must also increase. Egghe's findings indicate that Dice and Jaccard meet the first property, while Salton and Association Strength do not. For the second property, Dice, Jaccard, and Salton are compliant, whereas Association Strength is not. While limited to the two scientometric analysis types, co-word and co-citation, Egghe (2010b, p. 29) concluded in his paper that '*we do not have … a similarity measure that satisfies all properties*'. This conclusion extends to other scientometric analysis types, such as co-authorship, highlighting that no single similarity measure is universally applicable across all types of scientometric analysis.

The next step is to evaluate the suitability of Association Strength for co-authorship analysis. The results reveal two key characteristics of Association Strength in a dense network (Case 2) compared to Jaccard, Salton, and Dice-Sorenson. First, Association Strength moderates relative scores within the network rather than absolute ones, due to its different scale compared to the 0-1 range of the other indexes. Second, in a dense network (Case 2), Association Strength is 'orthogonal' to Jaccard, Salton, and Dice-Sorenson, with no clear relationship to Pearson. This is illustrated by the maps in Figure 3.

Conversely, in a low dense network (Case 1), Association Strength appears 'orthogonal' to Pearson, but shows no clear relationship with Jaccard, Salton, and Dice-Sorenson. Figure 3 shows that the links with high scores using Pearson (for example: Prathap-Bornemman, Guan-Egghe, Egghe-Glanzel, Huang-Ding) disappeared when using Association Strength in favour of other links (for example: Guan-Glanzel mostly).

Figure 1: Comparison of Similarity Indexes in the Context of Individual Co-authorship (The Matrix is (14x14), hence each graph depicts (14*13)/2 points)

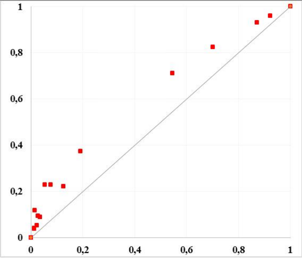
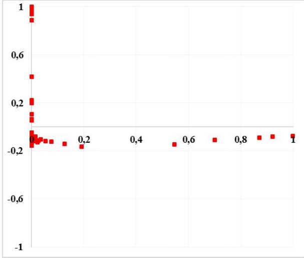
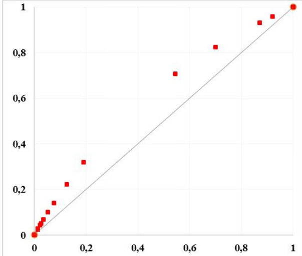
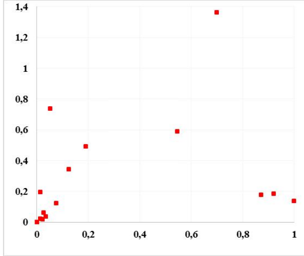
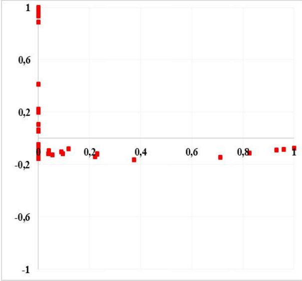
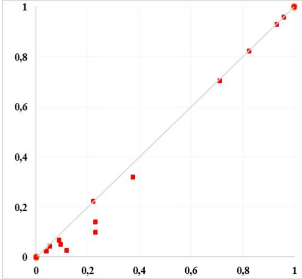
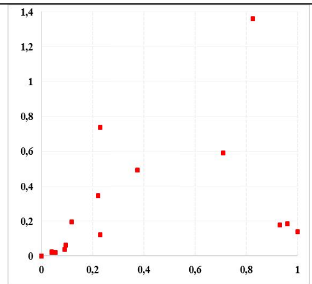
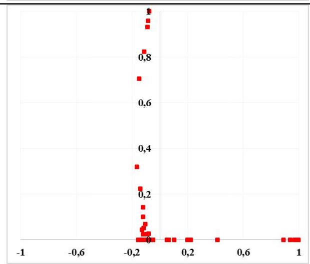
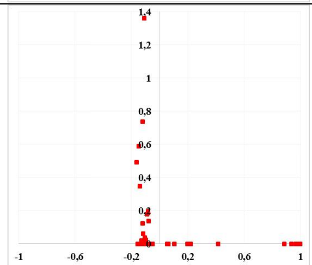
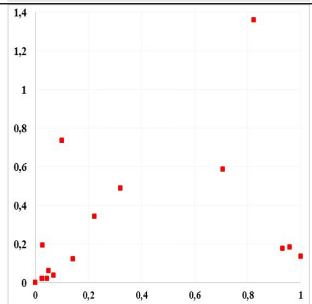
	Salton	**Pearson**	**Dice-Sorenson**	**Association Strength**
Jaccard				
Salton				
Pearson				
Dice-Sorenson				

Figure 2: Comparison of Similarity Indexes used for Country Co-authorship (The Matrix is (28x28), hence each graph depicts (28*27)/2 points)

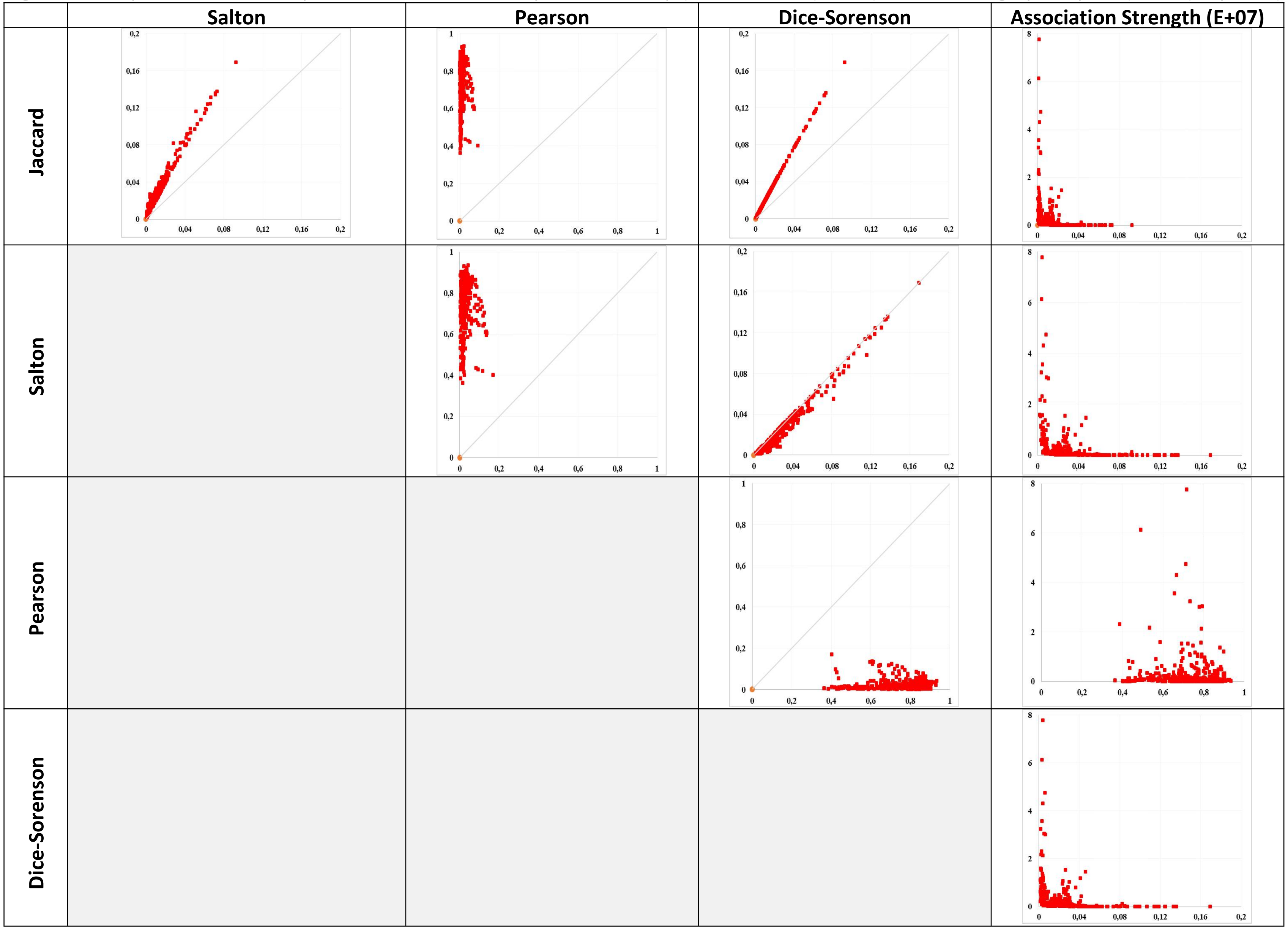
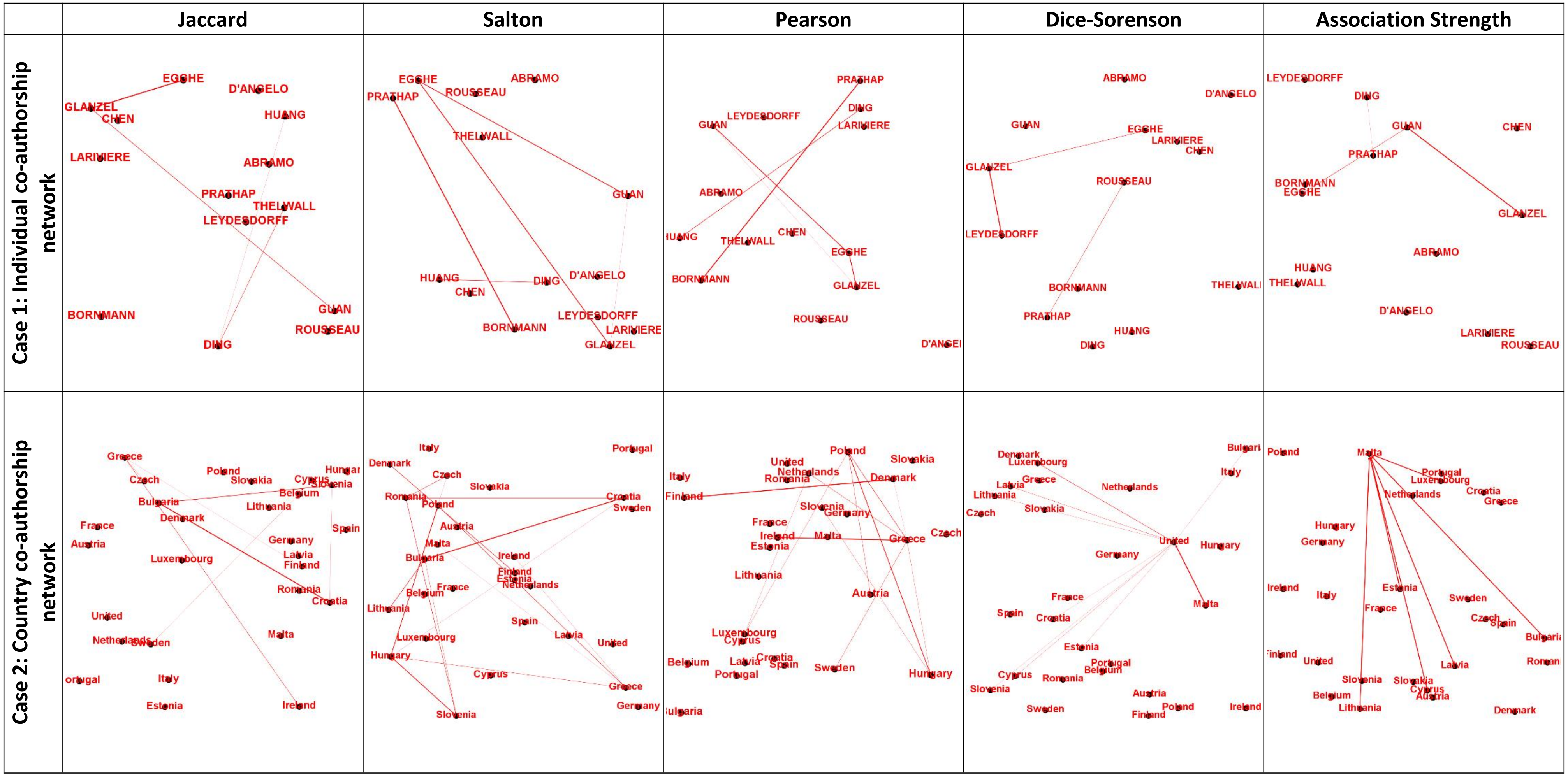


Figure 3: Network Mapping using each of the Similarity Indexes



CONCLUSIONS

Co-authorship is an increasingly crucial component in scientometrics to assess scientific collaboration networks and extentively analyses science teams. To draw the network of co-authorship, similarity indexes are often used to detect, analyse and/or predict collaborative links in terms of co-authored papers and localize core and peripheral entities within a network. Regardless of these levels, the question that remains is how to draw the co-authorship network as accurately as possible. This empirical study aims to improve understanding of which index is best suited for co-authorship analysis. To achieve this, two distinct cases of low and high-density co-authorship networks were examined: the first is a micro-level network composed of the top 14 authors in scientometrics and informetrics, while the second is a macro-level network representing co-authorship across European countries. In both cases, five most well-known similarity indexes were used: Jaccard, Salton, Dice-Sorensen, Pearson, and Association Strength.

Empirical results revealed that some indexes produced similar predictions for similarity scores based on theoretical calculations, while others showed no clear relationship. Jaccard, Salton, and Dice-Sorensen were found to be correlated. Specifically, Salton followed a concave increasing function relative to Jaccard, though Jaccard scores were consistently lower. Dice-Sorensen, which is quite similar to Jaccard, displayed a linear relationship with Salton. Pearson, as an indirect similarity index that accounts for both similarity and dissimilarity, was found to be orthogonal to Jaccard, Salton, and Dice-Sorensen - meaning that when the Pearson score was low, the scores of the other three indexes were high, and vice versa. Furthermore, Association Strength, proposed as a probabilistic measure with no vector formula - unlike the other indexes and with an unlimited scale - offers a distinct measure for co-authorship analysis. Empirically, Association Strength is found to be orthogonal to Jaccard, Salton, and Dice-Sorensen, but shows no clear correlation with Pearson in high-density networks. However, in low-density networks, Association Strength is orthogonal to Pearson, while lacking a clear correlation with Jaccard, Salton, and Dice-Sorensen.

ACKNOWLEDGEMENT

This study receives no funding from any source, and no financial or non-financial interests that could be perceived as influencing the study or its outcomes. The author would like to thank Prof. Rachid Bouabid for his valuable support in reviewing this paper and Prof. Dr Vincent Larivière for his kind help to provide data on the co-publications of the 28 EU countries. Finally, the author acknowledges the anonymous referee for his/her insightful comments and substantial suggestions on an earlier draft of this paper.

CONFLICT OF INTERESTS

The author declares no conflict of interest.

APPENDIX

Appendix 1 : Raw Matrix of Individual Co-authorship of Top 14 Scientometricians in the Web of Science for the Time Period 2010-2019

	BORNMANN L	GLANZEL W	LEYDESDORFF L	ABRAMO G	D'ANGELO CA	THELWALL M	PRATHAP G	HUANG MH	ROUSSEAU R	DING Y	LARIVIERE V	EGGHE L	GUAN JC	CHEN DZ
BORNMANN L		2	68	0	0	1	0	0	0	0	0	3	0	0
GLANZEL W			0	0	0	0	0	0	4	0	0	0	0	0
LEYDESDORFF L				0	0	0	1	0	2	1	0	0	0	0
ABRAMO G					101	0	0	0	0	0	0	0	0	0
D'ANGELO CA						0	0	0	0	0	0	0	0	0
THELWALL M							0	0	0	0	7	0	0	0
PRATHAP G								0	0	0	0	0	0	0
HUANG MH									0	0	0	0	0	70
ROUSSEAU R										0	0	12	1	0
DING Y											2	0	0	6
LARIVIERE V												0	0	0
EGGHE L													0	0
GUAN JC														0
CHEN DZ														

Appendix 2 : Raw Matrix of Country Co-authored Papers of UE-28 Countries in the Web of Science for the Time Period 2016-2018

	Austria	**Belgium**	**Bulgaria**	**Croatia**	**Cyprus**	**Czech**	**Estonia**	**Finland**	**France**	**Germany**	**Greece**	**Hungary**	**Ireland**	**Italy**	**Latvia**	**Lithuania**	**Luxem-bourg**	**Malta**	**Netherla nds**	**Poland**	**Portugal**	**Romania**	**Slovakia**	**Slovenia**	**Spain**	**Sweden**	**United Kingdom**
Austria		2714	880	1027	472	2672	652	1704	5440	14869	1752	2055	1072	6159	399	587	199	54	4033	2534	1663	1094	1080	1135	4291	3166	7296
Belgium			782	796	561	1848	743	1977	11588	10781	1880	1521	1498	8065	410	650	538	96	10355	2494	2235	830	373	618	6386	3460	12085
Bulgaria				583	392	946	452	600	1223	1640	907	868	507	1344	346	472	32	23	640	1172	823	605	420	397	1188	530	1325
Croatia					428	910	504	772	1370	1845	837	863	569	1743	357	476	53	75	665	1022	676	406	307	861	1223	567	1637
Cyprus						467	424	495	683	857	1260	445	461	857	292	405	17	46	331	516	504	122	71	85	744	273	1189
Czech							649	1377	4120	6062	1491	1773	792	3839	418	632	64	35	2127	3399	1478	977	2899	970	3227	2075	4480
Denmark							437	2570	5364	9268	1483	1120	1002	5242	157	263	188	74	5594	1943	1591	881	675	732	4626	6797	10523
Estonia								1366	938	1354	551	587	540	994	545	573	48	29	519	803	563	180	141	144	989	888	1358
Finland									3887	6084	1366	1122	1093	3780	484	654	124	50	3114	1876	1280	501	356	333	3332	5580	6645
France										27176	3923	2842	2801	21655	515	852	762	158	12044	6015	4630	2763	1154	1326	17247	7865	28400
Germany											4837	4253	3552	23455	661	1070	984	163	21450	8752	4850	2661	1566	1797	18860	12828	40021
Greece												1254	937	4681	373	562	102	111	2567	1972	1555	1052	666	719	3626	1899	6531
Hungary													679	2796	383	526	48	45	1748	1833	1155	1318	892	688	2431	1626	3581
Ireland														2778	354	482	101	71	2175	1282	1011	451	165	230	2567	1341	7643
Italy															529	929	455	326	11996	5978	4543	2528	1139	1860	18038	7592	27517
Latvia																553	17	18	178	544	385	115	88	83	501	238	560
Lithuania																	24	33	342	880	622	174	142	160	875	476	1008
Luxembourg																		16	355	145	136	55	47	59	336	228	621
Malta																			144	78	93	56	30	60	160	97	347
Netherlands																				3564	2810	1445	866	1041	9279	6960	23169
Poland																					1907	1579	1419	984	4902	2879	7136
Portugal																						943	582	714	8055	1823	6496
Romania																							716	666	1925	991	2156
Slovakia																								603	1054	746	1192
Slovenia																									1280	919	1527
Spain																										6256	22339
Sweden																											14543
United Kingdom																											

REFERENCES

Abrizah, A., Erfanmanesh, M., Rohani, V. A., Thelwall, M., Levitt, J. M., & Didegah, F. (2014). Sixty-four years of informetrics research: productivity, impact and collaboration. *Scientometrics, 101*(1), 569-585.

Adnani, H., Cherraj, M. & Bouabid, H. (2020). Similarity indexes for scientometric research: a comparative analysis. *Malaysian Journal of Library & Information Science, 25*(3), 29-46.

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550-560.

Balland, P.A., Boschma, R., & Ravet, J. (2019). Network dynamics in collaborative research in the EU, 2003-2017. *European Planning Studies, 27*(9), 1811-1837.

Bensman, S. J. (2004). Pearson's *r* and author co-citation analysis: a commentary on the controversy. *Journal of the American Society for Information Science and Technology, 55*(10), 935.

Bouabid, H., & Achachi, H. (2022). Size of science team at university and internal co-publications: Science policy implications. *Scientometrics, 127*(12), 6993-7013.

Chuan, P. M., Son, L. H., Ali, M, Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence, 48*(8), 2470-2486.

Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology, 60*(2), 232-239.

Egghe, L., & Leydesdorff, L. (2009). Relation between Pearson's correlation coefficient *r* and Salton's cosine measure. *Journal of the American Society for Information Science and Technology, 60*(5), 1027-1036.

Egghe, L. (2010a). On the relation between the Association Strength and other similarity measures. *Journal of the American Society for Information Science and Technology, 61(*7), 1502-1504.

Egghe, L. (2010b). Good properties of similarity measures and their complementarity. *Journal of the American Society for Information Science and Technology, 61*(10), 2151-2160.

Erfanmanesh, M., Rohani, V. A., & Abrizah, A. (2012). Co-authorship network of scientometrics research collaboration. *Malaysian Journal of Library and Information Science, 17*(3), 73-93.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing & Management, 25*(3), 315-318.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*(1), 1-49.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: a survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150-1170.

Luukkonen, T., Tijssen, R. J., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28*(1), 15-36.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*, 240-242.

Schneider, J. W., & Borlund, P. (2007). Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. *Journal of the American Society for Information Science and Technology, 58*(11), 1596-1609.

Sternitzke, C., & Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics, 78*(1), 113-130.

Van Eck, N. J. & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of The American Society for Information Science and Technology, 60*(8), 1635-1651.

Watts D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Wagner, C. S., & Leydesdorff, L. (2003). Mapping global science using international co-authorships: a comparison of 1990 and 2000. In G. Jiang, R. Rousseau, & Y. Wu (Eds.), *Proceedings of the Ninth International Conference on Scientometrics and Informetrics - ISSI 2003*, (pp. 330-340).

Wei, F., Zhang, G., Feng, Y., Liu, L., & Shao, Z. (2017). A co-authorship network-based method for understanding the evolution of a research area: A case of information systems research. *Malaysian Journal of Library & Information Science, 22*(2), 1-14.

White, H. D. (2003). Author cocitation analysis and Pearson's *r*. *Journal of the American Society for Information Science and Technology, 54*(13), 1250-1259.

White, H. D. (2004). Replies and a correction. *Journal of the American Society for Information Science and Technology, 55*(9), 843-844.