Papers in Evolutionary Economic Geography

# 20.43

Improvement on the association strength: implementing a probabilistic measure based on combinations without repetition

Mathieu P.A. Steijn

Improvement on the association strength: implementing a probabilistic measure based on combinations without repetition.[*]

Mathieu P.A. Steijn[†]

September 2020

Abstract — The use of co-occurrence data is common in various domains. Co-occurrence data often needs to be normalised to correct for the size-effect. To this end, van Eck and Waltman (2009) recommend a probabilistic measure known as the association strength. However, this formula is based on combinations with repetition, even though in most uses self-co-occurrences are non-existent or irrelevant. A more accurate measure based on combinations without repetition is introduced here and compared to the original formula in mathematical derivations, simulations, and patent data, which shows that the original formula overestimates the relation between a pair and that some pairs are disproportionally more overestimated than others. The new measure is available in the EconGeo package for R by Balland (2016).

Keywords — co-occurrence, network analysis, similarity measure, probabilistic measures.

[*] I thank Ludo Waltman, Nees Jan van Eck, Pierre-Alexandre Balland and Ron Boschma for useful comments on this paper. This work has benefited from grant 438-13-406 from JPI Urban Europe.(**contact:** m.p.a.steijn@uu.nl)

[†] Department of Human Geography and Planning, Utrecht University, Princetonlaan 8a, 3584CB Utrecht and Department of Spatial Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam.

1 Introduction

The use co-occurrence data is popular in numerous scientific domains like scientometrics (see for example van Eck and Waltman, 2009), computational linguistics (see for example Schutze, 1998), community ecology (see for example Peres-Neto, 2004), development economics (see for example Hidalgo et al., 2007), molecular biology (see for example Maslov and Sneppen, 2002 and evolutionary economic geography (see for example Boschma et al., 2015). Its use is widespread and in close relation with the popularity of network analysis across disciplines.

Co-occurrence data is used to infer the relation, referred to as relatedness here following Hidalgo et al. (2007), between entities, which can be species of fish, authors or technological classes, by observing how each of these co-occur with others in places, like streams, articles or patents. However, the total number of co-occurrences between a pair of entities cannot be used straightforwardly to reflect the relatedness between them because entities with more observations are more likely to co-occur than entities with fewer observations. To correct for this size-effect a normalisation measure is applied to the data.[1] van Eck and Waltman (2009) review the most popular normalisation measures and make a convincing case for the use of a probability-based measure known as the association strength. This measure is based on dividing the *observed* number of co-occurrences over the *expected* numbers of co-occurrences when assuming observations are randomly distributed over co-occurrences.[2]

In this paper, it is shown that the probability formula of the association strength, as proposed by van Eck and Waltman (2009), is not optimized to calculate the expected

[1] Note that it depends on the goal of the research if it is necessary to correct for the size-effect or that absolute counts are more relevant. In this paper and the research cited here and in van Eck and Waltman (2009) normalisation are assumed to be necessary.

[2] As such, a value of one indicates that exactly the same amount of co-occurrences are observed as expected. While a value above one or below one indicates respectively a stronger relation or a weaker relation between the two entities.

number of co-occurrences. The formula of van Eck and Waltman (2009) is proportional to probability calculations based on combinations with repetition, which means that when estimating the probability that two entities co-occur an observation drawn in the first draw is assumed to be available for drawing again when drawing the second observation. However, in the use of co-occurrence data the co-occurrence of observations from the same entity is disregarded.[3] This makes the possibility of drawing the same observation or any other observation from the same entity impossible in the second draw once an observation from this entity has been drawn in the first draw.

Therefore, an improved formula for the association strength is introduced derived from, but not equal to, probability measures based on combinations without repetition. Furthermore, two refinements are made regarding the inputs to the formula, which in the current definition do not properly take into account how the number of observed co-occurrences are calculated.

The improved formula is compared to the original formula in a theoretical setting, a number of simulations, and a real world application using patent data. It is shown that: firstly, the original formula overestimates the relatedness between a pair, when these co-occur at least once. This indicates that the original formula can wrongly identify two entities as related whereas in fact they are not; and, secondly, the original formula overestimates the relatedness between some pairs more than other. This indicates that the overestimation is not proportional and that the differences between the relatedness values for each pair are also distorted.

In the theoretical analysis, the improved formula is subtracted from the original formula, to obtain a formula for the difference. By considering the domain of each variable, it is shown that the original formula underestimates the number of expected occurrences in all cases and therefore overestimates the relationship between two entities when there is at

[3] This holds for the work referred to in this paper and those by van Eck and Waltman (2009).

least one observed co-occurrence. Continuing the theoretical exploration, the first order partial derivatives of the difference with respect to each variable is taken, which shows that the overestimation is not equal across all possible types of co-occurrence matrices.

Just taking the partial derivatives is not sufficient to show the size of the difference for each case, as the values of the variables are interconnected in ways that do not allow for analytical solving. Therefore, simulations are ran in which four different exemplary cases are taken to the extreme to demonstrate the effect on the difference. The simulations show that the overestimation by the original formula can be close to 0% but also close to 100% of the relatedness value given by the improved formula depending on the specificities of the co-occurrence matrix.

To measure to what extent these theoretical simulations are representative of real world applications of research on co-occurrence data, a number of patent samples, containing data on the technology classes per document, is treated to compare the results of both formulas. In these samples the overestimation of relatedness values for individual pairs varies between close to 0% to up to 3.234% of the value given by the improved formula and therefore does not attain the most extreme values obtained in the simulation. Nonetheless, it clearly confirms that some pairs are more overestimated than others. The results also show that some pairs are misidentified as being related by the original formula but that this is only the case for a rather small share of the pairs up to about 0.29% of the number of pairs identified by the original formula.

All in all, it is advisable to use the improved formula when working with co-occurrence data, where self co-occurrences are non-existent or irrelevant. The reformulation of the probability measure does not in any way alter the conclusion by van Eck and Waltman (2009) that probability based measures outperform so-called set-theoretic measures in normalising co-occurrence data. The improved measure, including the recommended method of implementation, is available in the EconGeo package for R by Balland (2016).

This paper is organised as follows: Section 2 gives a short overview of the use of co-occurrence data and the association strength; Section 3 discusses the refinements; Sections 4 to 6 explore the overestimation by the original formula respectively in a theoretical setting, simulations, and in a real world example using patent data; and Section 7 concludes.

2 Normalising co-occurrence data through probabilistic similarity measures

Co-occurrence data is generally derived from a binary occurrence matrix O of some order $m \times n$. The rows of O correspond to the places in which the observations occur and the columns to the entities to which they belong. There is a large variety of what these places and entities can be.[4] The example in Matrix 1 shows three patents that contain a reference to, respectively, only class c; class c & class d; and all classes a to d.

Matrix 1

$$\left(\begin{array}{ccccc} & \mathit{Class}\,a & \mathit{Class}\,\mathrm{b} & \mathit{Class}\,\mathrm{c} & \mathit{Class}\,\mathrm{d} \\ \mathit{Patent}\,1 & 0 & 0 & 1 & 0 \\ \mathit{Patent}\,2 & 0 & 0 & 1 & 1 \\ \mathit{Patent}\,3 & 1 & 1 & 1 & 1 \end{array}\right)$$

By multiplying the transpose of O by O itself the co-occurrence matrix C is obtained[5]. In which both the rows and the columns represent the entities and the matrix gives how often they co-occur with the other.

In the case of our example, this would yield the co-occurrence matrix C given in Matrix 2. Where class a co-occurs once with b, c, and d; class b co-occurs once with a, c and

[4] There are for example occurrence matrices of: scientific publications by research institutions (e.g. Hoekman et al., 2010); countries by industries (e.g. Hidalgo et al., 2007); streams by fish species (e.g. Peres-Neto, 2004); and patent documents by technology classes (e.g. Boschma et al., 2015).

[5] If the rows of O indicate the entities and the columns indicate the places where they co-occur then it is the other way around and O should be multiplied by its transpose.

d; class *c* co-occurs once with *a* and *b*, and twice with *d*; and class *d* co-occurs once with *a* and *b*, and twice with *c*. The diagonal is set to zero as the reference to a certain class does not entail a co-occurrence between that class and itself. This has important implications down the line.

Matrix 2

$$\left(\begin{array}{ccccc} & \mathit{Class}\,\mathrm{a} & \mathit{Class}\,\mathrm{b} & \mathit{Class}\,\mathrm{c} & \mathit{Class}\,\mathrm{d} \\ \mathit{Class}\,\mathrm{a} & 0 & 1 & 1 & 1 \\ \mathit{Class}\,\mathrm{b} & 1 & 0 & 1 & 1 \\ \mathit{Class}\,\mathrm{c} & 1 & 1 & 0 & 2 \\ \mathit{Class}\,\mathrm{d} & 1 & 1 & 2 & 0 \end{array}\right)$$

In many applications of co-occurrence data, such as the concept of relatedness, the raw numbers of co-occurrences between entities cannot straightforwardly be interpreted as giving the strength of the relation between each pair of entities. There is a so-called size-effect, as some classes co-occur more often with others for the simple reason that these classes have more occurrences in the first place. Like in our example, where *d* has more co-occurrences with *c* than with *a* or *b* but *c* also has more occurrences in total and therefore is more likely to co-occur with any class.

To correct the absolute number of co-occurrences for the size-effect data is normalised (van Eck and Waltman, 2009).[6] Correcting co-occurrence data for the size-effect to derive relationships between entities is done through direct similarity measures.[7] van Eck and Waltman (2009) wrote an extensive review on the most popular direct similarity measures, being: the cosine, the Jaccard index, the inclusion index and the association strength. Of these the last is a probabilistic measure, while the others are set-theoretic

[6] In some cases, more normalisation measures are deemed necessary. For example, Neffke et al. (2011) who look at the co-occurrence of products in the production process of the same plant also correct for the profitability of the respective products.

[7] Another option to derive similarities or relationships between entities is by comparing co-occurrence profiles of the entities, which are known as indirect similarity measures (see van Eck and Waltman, 2009).

measures. The authors show that set-theoretic measures do not properly correct for the size effect and argue in favour of the association strength.

The usability of their formula exceeds the domain of scientometrics. Hidalgo et al. (2007) developed an influential network analysis tool to derive the what they call relatedness between entities on the basis of co-occurrences. Although they use a different probabilistic direct similarity measure than the ones covered by van Eck and Waltman (2009), other authors (*e.g.* Balland et al., 2015) building on the framework of Hidalgo et al. (2007) do opt for the association strength.

Albeit influential, refinements to the work of van Eck and Waltman (2009) are in place. The probabilistic formula should be based on combinations *without* repetition instead of *with* repetition. Furthermore, the definitions of the inputs for the formula are imprecise. These points will be treated in the following section. It should be noted that the refinements to the measure do not undermine in anyway the statement of van Eck and Waltman (2009) that probabilistic measures outperform set-theoretic measures in normalising co-occurrence data to control for the size-effect.

3 Refinement to the association strength

The objective of the association strength is to estimate the number of expected co-occurrences for each pair assuming that these are randomly distributed and compare this to the number of observed co-occurrences to give an indication of the relation between a pair of entities when corrected for the size-effect. The challenge therefore is to correctly estimate the number of expected co-occurrences per combination.

As an intuitive example Matrix 3 gives a co-occurrence matrix C in which three classes (a, b, and c) exist and co-occur exactly once with each other:[8]

[8] This C would result from our example O in Matrix 1 if one would remove class d and its observations.

Matrix 3

$$\left(\begin{array}{cccc} & Class\text{ a} & Class\text{ b} & Class\text{ c} \\ Class\text{ a} & 0 & 1 & 1 \\ Class\text{ b} & 1 & 0 & 1 \\ Class\text{ c} & 1 & 1 & 0 \end{array}\right)$$

As each class has two observations and two possible other classes to co-occur with the expected number of co-occurrences is logically $\frac{2}{2} = 1$ for each combination (a & b, a & c, and b & c).

In this case, the matrix of expected co-occurrences is exactly the same as the matrix of observed co-occurrences given in Matrix 3. Therefore, we observe as many co-occurrences as expected and $\frac{Observed}{Expected}$ should be equal to one for each combination.

For the association strength, van Eck and Waltman (2009) use a simplified formula in the main text but describe formula 1 on p.1636:[9,10]

$$S_{Original}(C_{ij}, S_i, S_j, T, m) = \frac{C_{ij}}{(\frac{S_i}{T}\frac{S_j}{T} + \frac{S_j}{T}\frac{S_i}{T})m}, i \neq j, \tag{1}$$

In which S_i and S_j are the number of occurrences of entity i respectively j involved in co-occurrences where $i \neq j$. To calculate S_i one can use the row sum or the column sum of row i, respectively, column i of the C when the diagonal is set to zero. This slightly

[9] I argue that it is more advantageous to use the full formula, which entails exactly dividing the number of observed co-occurrences over the number of expected co-occurrences as it gives a clear threshold of one when $Observed = Expected$. As such, values below one indicate that less co-occurrences are observed than could be expected given a random distribution, whereas values above indicate the opposite. This threshold holds in all cases, even when matrices with different numbers of occurrences are compared. In contrast, the simplified formula would have a different value indicating that the number of observed co-occurrences equals expected depending on the matrices, even though it is proportional to the more detailed formula by a factor of $2m$.

[10] This formula is also presented in rewritten form in equation 1 in Waltman et al. (2010).

diverges from the explanation of van Eck and Waltman (2009).[11] T is the total number of occurrences and equal to $\sum_{i=1}^{n} S_i$ with n being the total number of entities, and m is the total number of co-occurrences and therefore equal to $\frac{\sum_{i=1}^{n} S_i}{2}$, which is half of T as each co-occurrence involves 2 occurrences. This definition also diverges from van Eck and Waltman (2009).[12] C_{ij} is the number of *observed* co-occurrences between i and j.

In essence, the denominator gives that the chance of encountering a co-occurrence between an observation of class i and an observation of class j is equal to the probability of first drawing one of the observations of class i out of the total number of occurrences times the chance of drawing an observation belonging to class j out of the total number of occurrences plus the probability of first drawing j and then i times the total number of co-occurrences.

Calculating this formula for our example C in Matrix 3 would yield Relatedness Matrix R given in Matrix 4 below:

Matrix 4

$$\begin{pmatrix} & Class\,\text{a} & Class\,\text{b} & Class\,\text{c} \\ Class\,\text{a} & 0 & 1.5 & 1.5 \\ Class\,\text{b} & 1.5 & 0 & 1.5 \\ Class\,\text{c} & 1.5 & 1.5 & 0 \end{pmatrix}$$

[11] van Eck and Waltman (2009, p. 1636) state that for S_i both the number of occurrences of entity i can be used or the number of co-occurrences in which i is involved. However, it is important to emphasize that single occurrences, as in Patent 1 of the example O in Matrix 1, should be ignored as these do not lead to co-occurrences. This also holds for self co-occurrences of i with i as both of these cannot be part of C_{ij} where $i \neq j$. Setting the diagonal to zero resolves both these issues.

[12] van Eck and Waltman (2009, p. 1648) state that m should be equal to "the number of documents". However, this only holds when the number of documents is equal to the number of co-occurrences. In the example O in Matrix 1 patent 1 is one document but only refers to one class so it does not involve any co-occurrences and is therefore not equal to one co-occurrence. Patent 3, on the other hand, refers to all classes a to d and therefore leads to 6 unique co-occurrences (a&b, a&c, a&d, b&c, b&d, c&d). All together the example consists of three documents and seven unique co-occurrences. As a result, in this case using the number of documents would underestimate the expected number of co-occurrences as the probability of encountering a co-occurrence is multiplied by a too small number of co-occurrences than are actually possible. This explanation is the same as in Waltman et al. (2010).

It is clear that the formula does not provide the intuitive answer of 1 but actually overestimates the relationship by returning that each pair co-occurs more often than could be expected given a random distribution.

The flaw cannot lie in the numerator, which is equal to the number of observed co-occurrences. Therefore the problem lies in the denominator. The formula to calculate the expected number of co-occurrences includes the possibility that when an occurrence of a certain entity is drawn the same occurrence or another occurrence of the same entity (if present) can be drawn in the next draw to complete the co-occurrence. This is known as combinations with repetition. However, as self co-occurrences are non-existent one knows that one cannot redraw the same occurrence, but also none of the other occurrences of that class.

In the case of our example, the denominator of formula 1 yields an expected number of $\frac{2}{3}$ co-occurrences. This is because the formula observes 2 occurrences for each class and 3 possible partners to co-occur with even though there are only 2 possible partners. Class a can co-occur with class b and class c but not with itself.[13]

In the case of co-occurrence data in which none of the observations belonging to the previously drawn entity can be drawn in the second draw the correct probabilistic measure would be formula 2:

$$S_{Improved}(C_{ij}, S_i, S_j, T, m) = \frac{C_{ij}}{(\frac{S_i}{T}\frac{S_j}{T-S_i} + \frac{S_j}{T}\frac{S_i}{T-S_j})m}, i \neq j, \tag{2}$$

Here, the denominator gives that the chance of encountering a co-occurrence between an observation of class i and an observation of class j is equal to the probability of first drawing one of the observations of class i times the chance of drawing an observation

[13] To be exact the denominator of formula 1 would be equal to $(\frac{2}{6}\frac{2}{6} + \frac{2}{6}\frac{2}{6})3$ for each pair outside of the diagonal in the matrix of this example.

belonging to class j knowing that none of the observations of class i can be drawn plus the chance of first drawing one of the observations of class j times the chance of drawing an observation belonging to class i knowing that any other observations of class j cannot be drawn.

The implications of using formula 1 instead of formula 2 are that the relatedness between a pair is overestimated when at least one co-occurrence is observed and that the overestimation is larger for certain pairs than others. These implications are demonstrated and further explored in the following parts. First in a theoretic setting, then by running simulations and concluding with the analysis of a real world example using patent data.

4 Theoretical exploration of the overestimation.

An obvious first notion from observing formula 1 and formula 2 is that there is no difference in outcome when the number of observed co-occurrences is zero, as the numerator C_{ij} will then be zero.

Furthermore, it can be assumed that formula 1 overestimates the relation between two entities when there is at least one co-occurrence. The assumption in the probabilistic measure of formula 1 is that the same observation and other observations from the same entity can be drawn again while this is not possible. This enlarges the total pool from which observations can be drawn and therefore decreases the likelihood that a certain co-occurrence can be drawn. This leads to the denominator, which contains the expected number of co-occurrences, in formula 1 being smaller than the one in formula 2 in all cases. As was the case for the example Matrix 3, where the denominator indicated a co-occurrence probability of $\frac{2}{3}$ for each pair where actually only two options instead of three existed and therefore $\frac{2}{2}$ should have been the answer.

Due to the smaller expected probability, formula 1 divides the number of observed co-occurrences over a too small number of expected co-occurrences and therefore the relatedness between these two entities is overestimated, when at least one co-occurrence is observed.

That the denominator of formula 1 underestimates the expected number of co-occurrences can also be proven analytically. The original probabilistic measure of van Eck and Waltman (2009) in the denominator of formula 1 is rewritten and given in formula 3, while the improved probabilistic measure used in the denominator of formula 2 is rewritten and given in formula 4:

$$E(C_{ij})_{Original}(S_i, S_j, T) = \frac{S_i S_j}{T}, i \neq j, \tag{3}$$

$$E(C_{ij})_{Improved}(S_i, S_j, T) = \frac{S_i S_j (2T - S_i - S_j)}{2(T - S_i)(T - S_j)}, i \neq j, \tag{4}$$

Let $D_{probability}$ be equal to $E(C_{ij})_{Improved} - E(C_{ij})_{Original}$. It can be shown that this difference $D_{probability}$ is equal to formula 5.

$$D_{probability}(S_i, S_j, T) = \frac{S_i S_j (S_i T + S_j T - S_i S_j)}{2T(T - S_i)(T - S_j)}, i \neq j, \tag{5}$$

For $E(C_{ij})_{Improved}$ to be larger than $E(C_{ij})_{Original}$ formula 5 gives that $S_i T + S_j T$ must be larger than $S_i S_j$. As $S_i \geq 1$, $S_j \geq 1$, and $T = S_i + S_j + S_k + ... + S_n$ it is clear that $T > S_i$ and $T > S_j$ and therefore $S_i T + S_j T > S_i S_j$ must hold.[14]

This means that $D_{probability}$ is positive in all circumstances, which indicates that the improved formula predicts in all cases that more co-occurrences can be expected between i and j. Which makes sense as the improved formula excludes the possibility of drawing a combination of i and i making it more likely to draw a combination with j.

[14] If entities can partially occur in a place then the values for S_i and S_j can be below one but in any case not below or equal to zero and the same statements hold.

Because the number of observed co-occurrences, C_{ij}, is divided over the number of expected co-occurrences, the original formula 1 leads to larger results than the improved formula 2 in all possible cases, when $C_{ij} > 0$. This can also be shown mathematically: Let $D_{Formula}$ be equal to $S_{Original}(C_{ij}, S_i, S_j, T) - S_{Improved}(C_{ij}, S_i, S_j, T)$.[15] It can be shown that the difference $D_{Formula}$ is equal to formula 8 after rewriting formula 1 to formula 6 and formula 2 to formula 7.

$$S_{Original}(C_{ij}, S_i, S_j, T) = \frac{TC_{ij}}{S_i S_j}, i \neq j, \tag{6}$$

$$S_{Improved}(C_{ij}, S_i, S_j, T) = \frac{2(T - S_i)(T - S_j)C_{ij}}{S_i S_j (2T - S_i - S_j)}, i \neq j, \tag{7}$$

$$D_{Formula}(C_{ij}, S_i, S_j, T) = \frac{(S_i T + S_j T - 2S_i S_j)C_{ij}}{S_i S_j (2T - S_i - S_j)}, i \neq j, \tag{8}$$

Three important notions can be derived from formula 8. First, it is confirmed that when there are no observed co-occurrences, i.e. $C_{ij} = 0$, the difference is zero. Second, if and only if $C_{ij} > 0$ then $S_i \geq S_j \geq 1$ and $T \geq S_i + S_j$ and therefore $(S_i T + S_j T > 2S_i S_j$. This indicates that formula 1 yields larger outcomes than formula 2 in all possible cases, with at least one observed co-occurrence. Effectively overestimating the relation between entity i and j. Third, for different values of S_i, S_j, C_{ij} and T the difference between formula 1 and formula 2 will also vary. This means that the difference between the formulas is not proportional for each pair but the relatedness between certain pairs is more strongly overestimated than for other pairs.

To explore the difference due to different values of S_i, S_j, C_{ij} and T the partial derivatives are taken of $D_{Formula}$ with respect to each. Because T is a function of S_i, S_j, and all

[15] Note that the order of the original formula and the improved formula has been altered compared to the previous calculation of the difference of the respective probabilistic measures.

other co-occurrences, $\sum_{k \neq i,j}^{n} S_k$. T is replaced by $S_i + S_j + L$ in formula 10 in which $L = \sum_{k \neq i,j}^{n} S_k$ and its range is equal to or larger than zero.

The partial derivatives $\frac{\delta D_{Formula}}{\delta C_{ij}}$, $\frac{\delta D_{Formula}}{\delta S_i}$, and $\frac{\delta D_{Formula}}{\delta L}$ are respectively given in formulas 9, 10, and 11. [16]

$$\frac{\delta D_{Formula}}{\delta C_{ij}} = \frac{(S_i^2 + S_j^2 + S_i L + S_j L)}{S_i S_j (S_i + S_j + 2L)}, i \neq j, \tag{9}$$

$$\frac{\delta D_{Formula}}{\delta S_i} = \frac{C_{ij}(S_i^2 S_j + S_i^2 L - 2S_i S_j^2 - 3S_i S_j L - S_j^3 - 3S_j^2 L - 2S_j L^2)}{S_i^2 S_j (S_i + S_j + 2L)^2}, i \neq j, \tag{10}$$

$$\frac{\delta D_{Formula}}{\delta L} = \frac{-C_{ij}(S_i - S_j)^2}{S_i S_j (S_i + S_j + 2L)^2}, i \neq j, \tag{11}$$

Given the domain of each formula, formula 9 is always positive, and, when at least one co-occurrence exists, formula 10 can be positive or negative depending on the respective inputs and formula **??** is always negative.

This last statement suggests that a relationship between two entities will be more overestimated by formula 1 when there is a smaller amount of other possibilities to co-occur with.

Despite being informative, partial derivatives give an incomplete picture of the discrepancy between the two formulas as these give the direction of a function with respect to an infinitesimal increase in one of the variables while keeping the others equal, even though it is in reality impossible to keep the other variables equal as the inputs are all related to

[16] The partial derivatives $\frac{\delta D_{Formula}}{\delta S_i}$ and $\frac{\delta D_{Formula}}{\delta S_j}$ are very similar in the sense that one can interchange the S_i and S_j to obtain the same formula, therefore $\frac{\delta D_{Formula}}{\delta S_j}$ is not shown.

each other. Necessarily C_{ij} consists of S_i and S_j, and if not all S_i co-occur with S_j then L must at least have enough occurrences to co-occur with the remaining i and js. In other words, the following logical conditions hold: $C_{ij} \leq min\{S_i, S_j\}$; and $L \geq |S_i - S_j|$. In the next section theoretical simulations are run in which these conditions can be met.

5 Simulational exploration of the overestimation

For the theoretical simulations a simple co-occurrence matrix C depicted in Matrix 5 is used. Albeit it simple, this matrix allows for some exploration of the numerical difference between formula 1 or formula 2 for different values of S_i, S_j, C_{ij}, and L. In four different simulations, hypothetical and rather extreme situations are simulated to get insight on the effects of increasing the values of each of the variables S_i, S_j, C_{ij}, and L, while meeting the conditions $C_{ij} \leq min\{S_i, S_j\}$; and $L \geq |S_i - S_j|$.

Matrix 5

$$\begin{pmatrix} Classes & \text{a} & \text{b} & \text{c} & \text{d} \\ \text{a} & 0 & 1 & 1 & 1 \\ \text{b} & 1 & 0 & 1 & 1 \\ \text{c} & 1 & 1 & 0 & 1 \\ \text{d} & 1 & 1 & 1 & 0 \end{pmatrix}$$

In the first simulation, Matrix 5 is taken and the number of co-occurrences between c & d is increased by 1 in each step k, ceteris paribus. Matrix 6 gives this simulation:

Matrix 6

$$\begin{pmatrix} Classes & \text{a} & \text{b} & \text{c} & \text{d} \\ \text{a} & 0 & 1 & 1 & 1 \\ \text{b} & 1 & 0 & 1 & 1 \\ \text{c} & 1 & 1 & 0 & \mathbf{1+k} \\ \text{d} & 1 & 1 & 1+k & 0 \end{pmatrix}$$

In each step k the resulting relatedness matrix using formula 1 is subtracted from the resulting relatedness matrix using formula 2 and divided over the value of formula 2 to express the difference in percentages. The relatedness values for the pairs a & b, and c & d are then plotted for each step. Each of these two changing relationships represent a different scenario:

- ***a* & *b*.** The changing difference in relatedness for the pair a & b simulates a steady increase in L, keeping $C_{ij} = 1$ and $S_i = S_j = 3$. This result is depicted in Figure 1.
- ***c* & *d*.** The changing difference in relatedness between classes c & d simulates a steady increase in C_{ij} but also in S_i and S_j, keeping $L = 6$. To increase C_{ij} beyond the maximum value of S_i and S_j S_i and S_j also have to increase. From the partial derivatives can be derived that an increasing C_{ij} would increase the difference whereas an increase in S_i and S_j can both increase or decrease the difference. The result of the simulation is depicted in Figure 2.

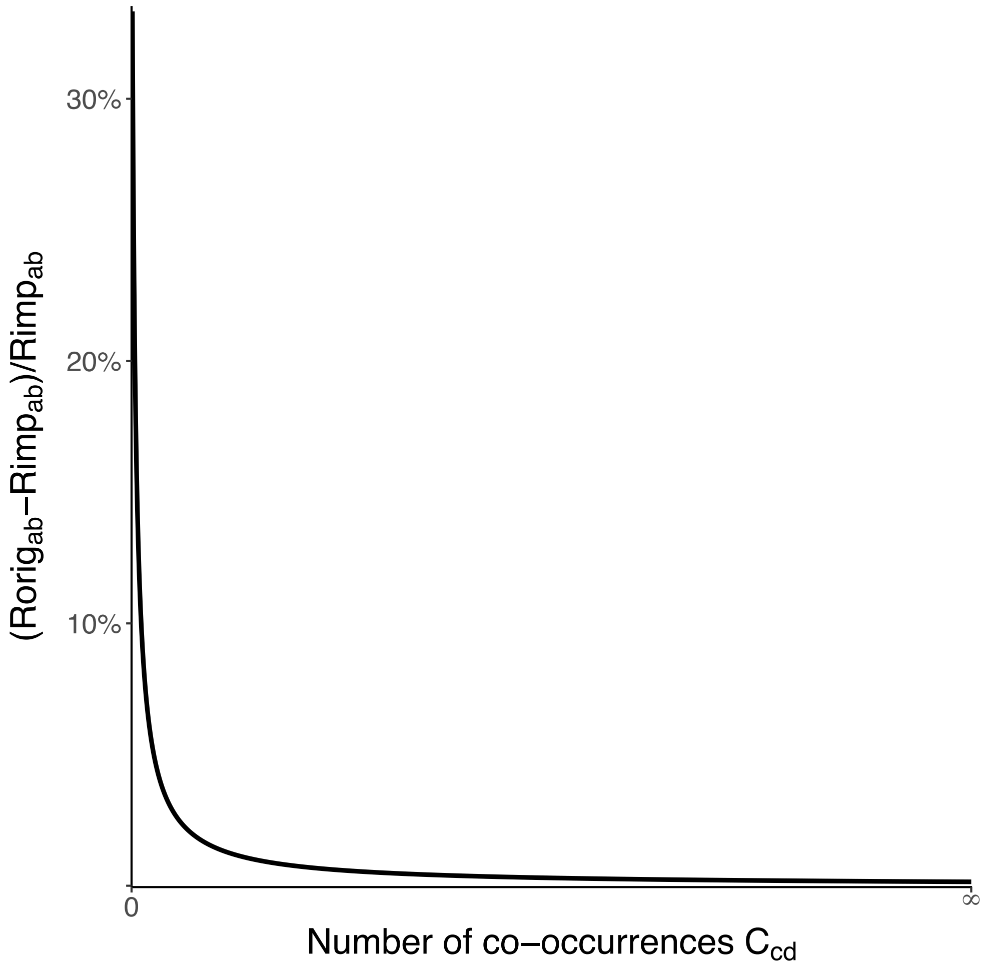

Figure 1 – The difference in relatedness between the original formula and the improved formula for class a & b when L increases.

The absolute difference between the calculated relatedness of formula 1 and formula 2 for the pair a & b is equal to 1/3 across the entire simulation. However, as the number of other co-occurrences L increases potential co-occurrence candidates increase as well and therefore the expected number of co-occurrences for a & b decreases. As a result, relatedness values are higher as L increases and the relative difference decreases, as can be seen in Graph 1.

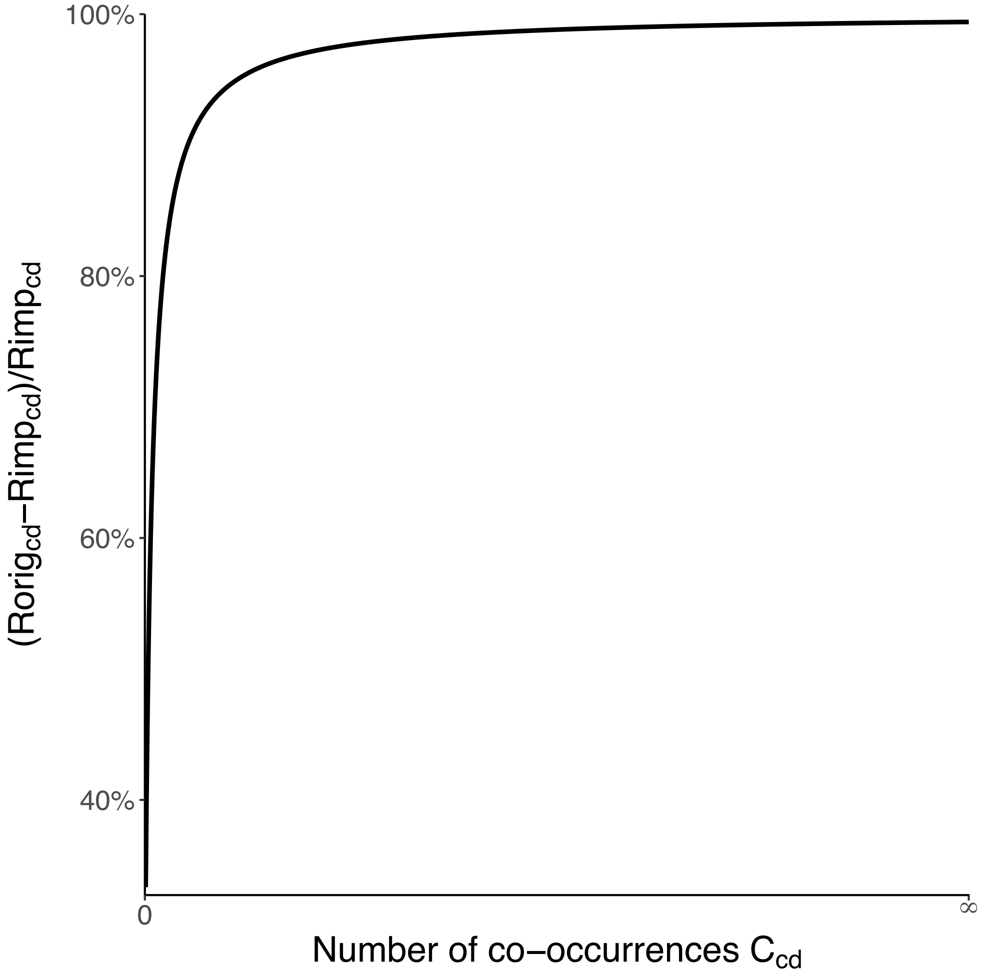

Figure 2 – The difference in relatedness between the original formula and the improved formula for class c & d when C_{cd}, S_c and S_d increase.

For pair c & d L remains equal to 6 but C_{cd}, S_c and S_d increase. Figure 2 depicts how the difference in the estimated relatedness increases asymptotically converging from 33.3% to the value of 100%. As the $\frac{Observed}{Expected}$ should be close to one when two entities are close to having 100% of the occurrences in the sample but the values of the original formula 1 converges to two the difference is close to 100% of the correct value.

To simulate an increase in C_{ij} while keeping S_i, S_j, and L equal, ceteris paribus, another simulation is needed: matrix 1 is altered by replacing the number of co-occurrences between entities a & b and c & d by a large amount of co-occurrences x.

Then in each step k of the simulation a co-occurrence is subtracted from this amount x and added to the co-occurrences between entities a & d and b & c. See matrix 6. This keeps S_i, S_j and L equal but increases C_{ij} for the relatedness between a & d. Note that the result is insensitive to the exact value of x as the resulting change in the denominator and numerator cancel each other out.

Matrix 7

$$\begin{pmatrix} Classes & \text{a} & \text{b} & \text{c} & \text{d} \\ \text{a} & 0 & x-k & 1 & 1+k \\ \text{b} & x-k & 0 & 1+k & 1 \\ \text{c} & 1 & 1+k & 0 & x-k \\ \text{d} & \mathbf{1+k} & 1 & x-k & 0 \end{pmatrix}$$

The result is a stable overestimation of 33.3% for all values of k. When a & d co-occur more often but the total number of co-occurrences in the sample stays the same the relatedness between a & d naturally increases. Nonetheless, the increase in relatedness is proportional for the two formulas and therefore the difference remains 33.3%.

Lastly, an increase in S_i and S_j while keeping C_{ij} equal is simulated. The simulation is very similar to the first simulation except that next to increasing the co-occurrences between c & d also those between b & c is increased in each step k, see matrix 4. As a result, S_b and S_c increases while C_{bd} is kept at one. L increases necessarily as well in the form of S_c to match the added co-occurrences of S_b and S_d.

Matrix 8

$$\begin{pmatrix} Classes & \text{a} & \text{b} & \text{c} & \text{d} \\ \text{a} & 0 & 1 & 1 & 1 \\ \text{b} & 1 & 0 & 1+k & \mathbf{1} \\ \text{c} & 1 & 1+k & 0 & 1+k \\ \text{d} & 1 & 1 & 1+k & 0 \end{pmatrix}$$

Once again the percentual difference between calculating the level of relatedness for the pair b & d using formula 1 and formula 2 is stable at 33.3% for all values k. This time the relatedness between b & d decreases as k increases because their total number of occurrences S_b and S_d increase but their number of co-occurrences remains 1.

The simulations in this section show that the difference can range between close to 100% and close to 0. In real world applications of co-occurrence data the bias introduced by using formula 1 instead of formula 2 will be somewhere in between the extreme scenarios simulated here. In which each respective value in the relatedness matrix will be closer to a specific scenario than others.

6 Real world data-based exploration of the overestimation

The theoretical and simulational explorations demonstrate that formula 1 overestimates the relatedness between entities compared to formula 2 in a way that disproportionally affects certain pairs more than other pairs. However, the question remains how close these examples are to real world applications.

Therefore, the outcomes of formula 1 and formula 2 are compared using USPTO technology class data from utility patents in periods of 5 years from 1855 to 2014.[17]

In the occurrence matrix O of each time period the rows indicate patent numbers and the columns technology classes, like the example in Matrix 1. By multiplying the transpose

[17] A period of 5 years is also used by Boschma et al. (2015).

of O by O itself a technology classes by technology classes co-occurrence matrix C is obtained. As before, the diagonal of C is set to zero and S_i can then be calculated as the column sum of column i or the row sum of row i.[18] Next formula 1 and formula 2 are calculated using the C of each time period and the results are compared in Table 1.

Table 1 gives a number of statistics for each time period mentioned in the respective header. The first row gives the number of different technology classes (n) referred to on the patents. This number is equal to the number of columns/rows in C. The second line gives the number of pairs that have a value higher than 1 according to formula 1 by van Eck and Waltman (2009), these relatedness pairs have more or just as much observed co-occurrences as expected and are therefore seen as related in research within this domain (see for example Balland et al., 2015). The third line gives the same statistic but employs the improved formula 2. On line four the difference between the number of related pairs according to each formula is given.[19]. Difference (%) expresses this difference as a percentage of the number of related pairs according to the improved formula 2.

Focussing on these first five statistics it can be seen that in 1855 to 1859 patents made references to 327 different technology classes and that according to formula 1 5154 pairs of technology classes can be seen as related, while formula 2 identifies 5150 related pairs. As a result, formula 1 identifies 4 pairs or $\frac{4}{5150} \times 100 = 0.07\%$ more as related than formula 2.

In later time periods the differences increase both in absolute terms as in relative terms with a maximum in relative terms of 0.29% in 1885-1889 and a maximum in absolute terms with 62 pairs wrongly seen as related in 1955-1959.

Next to the overestimation another problem of using formula 1 instead of formula 2 is

[18] Note that the relatedness function in the EconGeo package for R (see Balland, 2016) sets the diagonal of the input co-occurrence matrix to zero automatically.

[19] Note that there are no pairs identified as related by formula 2 that are identified as unrelated by formula 1, as formula 1 > formula 2, when $C_{ij} > 0$. See also Section 4.

that the relatedness between some pairs is more overestimated than between other pairs. The last four statistics explore this disproportionality. The largest difference in value gives the largest difference in the relatedness value of a single pair between formula 1 and formula 2, while its percentage counterpart gives the largest overestimation relative to the value given by formula 2. In relative terms the highest over estimation is 3.23% and occurs in 2000-2004, this percentage is way below some of the extreme scenarios simulated in Section 5. The largest absolute difference is 0.837 in 1860-1864.

The last two statistics are similar but give the smallest difference, when $C_{ij} > 0$.[20] When at least one co-occurrence exists between a pair its relation is overestimated as already shown mathematically in Section 4. The values are close to zero both in absolute terms as in relative terms and therefore in strong contrast to the highest values, showing that some pairs get more overestimated than others.

The results also show that there is not necessarily a direct connection between the number of technology classes and the number of related pairs or the overestimation. In 2000-2004, there is the second highest number of different technology classes, while the number of related pairs is lower than in 1950-1954 when fewer technology classes were in use.

When comparing these specific time periods, 2000-2004 turns out to have a much more concentrated co-occurrence matrix C than the one in 1950-1954. In 2000-2004 each row or column i contains a few pairs with a lot of observations while others have relatively few observations. This contrasts with the more even spread of observations across C in 1950-1954. The average Gini coefficient per row of C in 2000-2004 is 0.936 versus 0.909 in 1950-1954.

Very much like the simulation based on matrix 7, where S_i and S_j was increased while keeping C_{ij} equal, the pairs with little co-occurrences are less overestimated when there

[20] When $C_{ij} = 0$ both formulas return 0 and the difference is therefore also zero and obviously the smallest.

are more occurrences of the same technology class with other classes, as is more the case in 2000-2004. The pairs with relatively high numbers of co-occurrences have a larger share of the sample in 2000-2004 compared to 1950-1954, like in matrix 6, where C_{ij} is increased while S_i and S_j are kept equal, these pairs are more overestimated in 2000-2004. The pairs with relatively many co-occurrences are likely to pass the threshold of 1 using either formula, the stronger overestimation for these pairs in 2000-2004 does not lead to much change with respect to passing this threshold. This is not the case for the pairs with relatively fewer co-occurrences, which are less overestimated in 2000-2004 than in 1950-1954. Therefore in 2000-2004, these are less likely to pass the threshold irrespective of whether formula 1 or formula 2 is used. While in 1950-1954 these pairs are more likely to pass the threshold using formula 1 but not when using formula 2. As a result, 2000-2004 has larger overestimations of individual relatedness values but less pairs that are wrongly identified as related.

The comparison shows that using formula 1 instead of formula 2 in research can lead to non-negligible differences and that some pairs and matrices are affected disproportionally. Note that with an incorrect specification of S_i, S_j and m formula 1 becomes even more inaccurate, see Section 4. It is unlikely that papers employing formula 1 instead of formula 2 would have reached fundamentally different conclusions but a risk is more present in some cases than others. It is recommended to use formula 2 in future research.

Table 1 – Patent comparison results

	1855-9	1860-4	1865-9	1870-4	1875-9	1880-4	1885-9	1890-4
Number of technology classes	327	335	343	356	361	372	379	385
Number of related pairs (Original formula)	5154	4902	7910	8954	10100	12396	13438	13484
Number of related pairs (Improved formula)	5150	4898	7892	8934	10080	12370	13398	13464
Difference	4	4	18	20	20	26	40	20
Difference (%)	0.07	0.08	0.22	0.22	0.19	0.21	0.29	0.14
Largest difference in value	0.827	0.837	0.788	0.786	0.822	0.662	0.593	0.63
Largest difference (%) in value	2.643	2.177	2.009	1.961	2.258	2.333	2.425	2.36
Smallest difference in value	0.00599	0.00505	0.00234	0.00169	0.0011	0.00107	0.00084	0.00082
Smallest difference (%) in value	0.0294	0.0268	0.01	0.0075	0.0085	0.0037	0.004	0.0032

	1895-9	1900-4	1905-9	1910-4	1915-9	1920-4	1925-9	1930-4
Number of technology classes	385	387	390	394	403	404	405	415
Number of related pairs (Original formula)	14196	15866	16372	16742	17784	18036	19560	21432
Number of related pairs (Improved formula)	14160	15842	16338	16694	17754	17990	19528	21396
Difference	36	24	34	48	30	46	32	36
Difference (%)	0.25	0.15	0.20	0.28	0.16	0.25	0.16	0.16
Largest difference in value	0.625	0.515	0.586	0.666	0.645	0.753	0.711	0.677
Largest difference (%) in value	2.568	2.341	2.303	2.441	2.536	2.173	1.933	1.872
Smallest difference in value	0.00063	0.00056	0.00042	0.00051	0.00038	0.00039	0.00023	0.00018
Smallest difference (%) in value	0.0026	0.0054	0.0055	0.0071	0.0052	0.0023	0.0036	0.0071

	1935-9	1940-4	1945-9	1950-4	1955-9	1960-4	1965-9	1970-4
Number of technology classes	414	417	413	423	427	430	432	434
Number of related pairs (Original formula)	22852	23430	23336	25104	24422	25326	25932	25590
Number of related pairs (Improved formula)	22814	23388	23280	25060	24360	25280	25902	25544
Difference	38	42	56	44	62	46	30	46
Difference (%)	0.16	0.17	0.24	0.17	0.25	0.18	0.11	0.18
Largest difference in value	0.557	0.56	0.525	0.492	0.557	0.529	0.579	0.661
Largest difference (%) in value	1.641	1.76	1.772	1.726	1.51	1.561	1.602	1.892
Smallest difference in value	0.00015	0.00015	0.00022	0.00014	0.00014	0.00011	0.00009	0.00008
Smallest difference (%) in value	0.003	0.0034	0.0063	0.0019	0.0029	0.0018	0.0015	0.0006

	1975-9	1980-4	1985-9	1990-4	1995-9	2000-4	2005-9	2010-4
Number of technology classes	436	435	435	435	431	437	436	438
Number of related pairs (Original formula)	25350	25012	24712	23982	24120	24422	24356	26382
Number of related pairs (Improved formula)	25324	24980	24676	23928	24084	24388	24310	26348
Difference	26	32	36	54	36	34	46	34
Difference (%)	0.10	0.12	0.14	0.22	0.14	0.13	0.18	0.12
Largest difference in value	0.684	0.694	0.69	0.524	0.501	0.581	0.592	0.64
Largest difference (%) in value	2.29	2.52	2.192	2.293	2.404	3.234	3.176	2.834
Smallest difference in value	0.00008	0.00008	0.00006	0.00005	0.00005	0.00004	0.00003	0.00002
Smallest difference (%) in value	0.0018	0.0033	0.004	0.0028	0.0033	0.0036	0.0013	0.0012

Notes: A pair is seen as related when the respective formula returns a value of 1 or higher for a certain pair. The statistics expressed in percentages are taken with respect to the value returned by the improved formula 2.

7 Conclusion

Co-occurrence data is commonly used in various domains. Researchers generally apply normalisation measures to correct for the size-effect. To this end, van Eck and Waltman (2009) make a convincing case to use a probability-based measure known as the association strength. In which the number of observed co-occurences is divided over the number of expected co-occurrences, assuming that observations are randomly distributed over co-occurences.

However, the probability formula to calculate the expected number of co-occurrences is not suited for the co-occurrence analysis it is recommended for. In this line of research self-co-occurrences are non-existent or irrelevant, whereas the probability formula assumes that an observation from an entity can be drawn again after been picked in the first draw.

This paper introduces a formula that is based on, but not equal to, combinations *without* repetition in which the probability of drawing entity i and j together is calculated as the probability of drawing i first and then j, knowing that none of the observations pertaining to i can be drawn plus the the probability of drawing j and then j, knowing that none of the observations pertaining to i can be drawn. This formula gives the correct results in an intuitive example.

Furthermore, it is shown that the original formula overestimates the relatedness between a pair of entities compared to the improved formula introduced here, when there is at least one observed co-occurrence, and that the overestimation is not proportional across pairs. Simulations show that the over estimation of the relatedness can range between virtually 0% and almost 100% of the correct value given by the improved formula. In a real world example, a number of patent samples showed that the overestimation of individual values was between virtually 0% and 3.234%, while the difference in the number of pairs that can be seen as related can be 0.29% more than the number of pairs identified as related by the improved formula.

All in all, it is evident that the formula presented here is better equipped for the analysis of co-occurrence data. The formula, including all recommendations for inputs and treatment, is available in the EconGeo package for R by Balland (2016).

References

Balland, P.-A. (2016). EconGeo: Computing Key Indicators of the Spatial Distribution of Economic Activities.

Balland, P.-A., Rigby, D. L., and Boschma, R. (2015). The technological resilience of US cities. *Cambridge Journal of Regions, Economy and Society*, 8(2):167–184.

Boschma, R., Balland, P.-A., and Kogler, D. F. (2015). Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, 24(1):223–250.

Hidalgo, C. A., Kilinger, B., Barabási, A.-L., and Hausmann, R. (2007). The Product Space Conditons the Develpment of Nations. *Science*, 317(July):482–487.

Hoekman, J., Frenken, K., and Tijssen, R. J. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5):662–673.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, 296(5569):910–3.

Neffke, F., Henning, M., Boschma, R., Lundquist, K. J., and Olander, L. O. (2011). The Dynamics of Agglomeration Externalities along the Life Cycle of Industries. *Regional Studies*, 45(1):49–65.

Peres-Neto, P. R. (2004). Patterns in the co-occurrence of fish species in streams: the role of site suitability, morphology and phylogeny versus species interactions. *Oecologia*, 140(2):352–360.

Schutze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1).

van Eck, N. J. and Waltman, L. (2009). How to Normalize Cooccurrence Data? An Analysis of SomeWell-Known Similarity Measures. *Journal of the American Society for Information Science*, 60(8):1635–1651.

Waltman, L., van Eck, N. J., and Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4):629–635.