Journal of Informetrics 9 (2015) 359–372

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Who is collaborating with whom? Part I. Mathematical model and methods for empirical testing

Hildrun Kretschmer[a,*], Donald deB. Beaver[b], Bulent Ozel[c], Theo Kretschmer[d]

[a] COLLNET Center, Germany
[b] Williams College, 117 Bronfman Science Center, 18 Hoxsey St., Williamstown, MA 01267, USA
[c] University of Jaume I, Department of Economy, Castellon De La Plana, Spain
[d] COLLNET Center, Borgsdorfer Str. 5, D-16540 Hohen Neuendorf, Germany

A R T I C L E I N F O

Article history:
Received 5 February 2013
Received in revised form
30 December 2014
Accepted 5 January 2015
Available online 19 February 2015

Keywords:
Social network analysis
Self-organization
Mathematical model
Co-authorship
3-D computer graphs
Animation

A B S T R A C T

There are two versions in the literature of counting co-author pairs. Whereas the first version leads to a two-dimensional (2-D) power function distribution; the other version shows three-dimensional (3-D) graphs, totally rotatable around and their shapes are visible in space from all possible points of view. As a result, these new 3-D computer graphs, called "Social Gestalts" deliver more comprehensive information about social network structures than simple 2-D power function distributions. The mathematical model of Social Gestalts and the corresponding methods for the 3-D visualization and animation of collaboration networks are presented in Part I of this paper. Fundamental findings in psychology/sociology and physics are used as a basis for the development of this model.

The application of these new methods to male and to female networks is shown in Part II. After regression analysis the visualized Social Gestalts are rather identical with the corresponding empirical distributions ($R^2 > 0.99$). The structures of female co-authorship networks differ markedly from the structures of the male co-authorship networks. For female co-author pairs' networks, accentuation of productivity dissimilarities of the pairs is becoming visible but on the contrary, for male co-author pairs' networks, accentuation of productivity similarities of the pairs is expressed.

1. Introduction

The rise in collaboration in science and technology experienced world-wide at both national and international levels, has assumed such overriding importance that there is now a perceptibly urgent need to study such processes with a view to acquiring fundamental knowledge for organizing future research and its application to science and technology policies. New concepts have emerged in order to understand pattern formation in interactional processes of collaboration (Yin, Kretschmer, Hanneman, & Liu, 2006). Some of these concepts are self-similarity, self-organization, power laws, complex networks of interactions and others.

Two different bibliometric analysis techniques are usually used for gender and for collaboration studies:

* Corresponding author. Permanent address: Borgsdorfer Str. 5, D-16540 Hohen Neuendorf, Germany. Tel.: +49 3303 500 866; fax: +49 3303 50 4838.
E-mail addresses: kretschmer.h@onlinehome.de (H. Kretschmer), dbeaver@williams.edu (D.deB. Beaver), ozel@uji.es (B. Ozel), kretschmer.h@onlinehome.de (T. Kretschmer).

http://dx.doi.org/10.1016/j.joi.2015.01.004

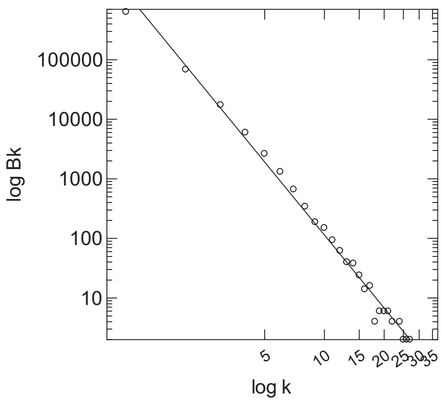

Fig. 1. Power law distributions of co-author pairs' frequencies B_k. Journal of Biochemistry, $B_k = c/k^b$, with c = constant $n = 26$; $R = 0.987$, $R^2 = 0.975$; F = Ratio = 508.65; $B_k = 5.39–4.539 \log k$.

- Descriptive analysis methods (standard bibliometric indicators, social network analysis, etc.), for example Naldi, Luzi, Valente, and Parenti (2004), Melin (2000), Carr, Pololi, Knight, and Conrad (2009), Kyvik and Teigen (1996), Pepe and Rodriguez (2009), etc.
- Parametric models or laws, for example Lotka (1926), Bradford (1934), Price (1963), Egghe (2008), Newman (2005), and Morris and Goldstein (2007).

The Social Gestalt model is a new parametric model with four parameters. First we refer to the version of counting co-author pairs leading to power laws and second to the new version of counting co-author pairs leading to 3-D graphs. In contrast to a single power function distribution (2-D graphs), the mathematical model of "Social Gestalts" (Kretschmer, 1999, 2002; Kretschmer & Kretschmer, 2007) visualizes the 3-D graphs, using animation in the form of rotating these graphs.

McGrath (2002, p. 314) has added the following remarks to the model of "Social Gestalts":

> "The social Gestalt can be defined as a holistic configuration of the parts of a relationship. Each Gestalt can be graphed as a 3-dimensional array of co-authorships. Though the interrelationships may vary, they can always be represented in a single holistic graph that, when stable exemplifies the conciseness principle."

The term "Social Gestalt" is selected in honor of both Wolfgang Metzger's deliberations in 1967 about the formation of social groups on the basis of the conciseness principle and in honor of the famous Berlin "Gestalt Psychology" at the beginning of the 20th century Metzger (1986).

1.1. First version of counting co-author pairs leading to power laws

Counting the number of publications of co-author pairs: The pairs are counted as units (P,Q) in analogy to single authors P in Lotka's Law, where k is the number of joint publications of the pair P, Q (For example: Smith & Miller). We assume there is a regularity for the distribution of coauthor pairs' frequencies B_k with k publications per co-author pair (P,Q) in form of a power law distribution:

Morris and Goldstein (2007) have already shown this kind of regularity in one of their empirical studies. In this connection Egghe (2008) has presented a theoretical model for the size-frequency function of co-author pairs.

We have studied the regularities for distributions if one counts the number of publications of the co-author pairs in the Journal of Biochemistry. We found a power law distribution, cf. Fig. 1.

Special collaboration structures, for example scale-free network models, self-similarity, power laws and others (Egghe, 2008; Morris & Goldstein, 2007; Newman, 2005) could be found in many larger networks. However, the investigation in large networks often rely on a wealth of bibliographic data, but very little or no other information about the individuals in the network (Pepe & Rodriguez, 2009).

1.2. Second version of counting co-author pairs leading to 3-D graphs

Because of the former missing information about the individuals in the network, the present paper is focused on social network analysis (SNA) applied to collaboration structures in co-authorship networks with special focus on the topic "Who is collaborating with whom". A developed procedure for visualizing a bivariate distribution of co-author pairs' frequencies hence producing 3-D graphs is presented. This distribution is described by a mathematical model, presented in Section 2. The detailed methods of counting the co-author pairs are shown in Section 3.

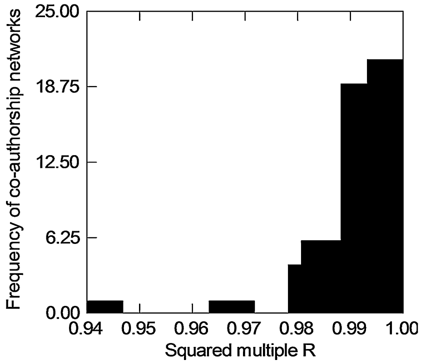

Fig. 2. Copy of Fig. 13, Page 110, published in Kretschmer & Kretschmer (2013): Frequency of co-authorship networks (ordinate) depend on the squared multiple R (abscissa) after regression analysis (empirical distribution of co-author pairs' frequencies ($\log N_{ij}$) and social Gestalt). Note: The squared multiple R ranges between 0.944 and 1.000 and the median is equal to 0.993. For 96% of the co-authorship networks the squared multiple R is larger than 0.98.

1.3. Validity of the mathematical model of Social Gestalts and reliability of the empirical studies

The study of the co-authorship networks presented in this paper is a part of all of our studies on these networks by the new mathematical model of Social Gestalts (3-D graphs). This model has already been applied to 52 large co-authorship networks (Kretschmer & Kretschmer, 2013, open access, http://dx.doi.org/10.4236/sn.2013.23011). The visualized Social Gestalts in form of 3-D computer graphs are rather identical with the corresponding empirical distributions. After regression analysis, for 96% of them the squared multiple R is larger than 0.98% and for 77% of the 52 networks even equal or larger than 0.99 (cf. Fig. 2). The corresponding 40 Social Gestalts (with $R^2 \geq 0.99$) in combination with empirical data are presented in the Appendix of the open access paper mentioned above, cf. pages 117–137. These 3-D graphs are rotated patterns as shown in the present paper, Section 3.4.2, Fig. 6: The rightmost pattern at the first row but without any vertical spikes.

After publication of the open access paper we have continued these studies (Kretschmer, Kretschmer, & Stegmann, 2014; Ozel, Kretschmer, & Kretschmer, 2014) resulting together with the other 52 networks in 62 Social Gestalts in total. For 80% of these networks the squared multiple R is larger than 0.99 and for 97% larger than 0.98. The median is equal to 0.994. Question: Can we expect a general validity of this mathematical model for research of co-authorship networks?

The reliability is given by both using official sources of data (for example Web of Science database) and by application of the methods/analysis presented by Kretschmer and Kretschmer (2013) and application of the extended version presented in Section 3 of this paper.

Both the theoretical approach of the mathematical model of Social Gestalts presented in Section 2 and the methods (analysis) in Section 3 of this paper, are an essentially improved description and interpretation of the original model and analysis published in the open access paper by Kretschmer and Kretschmer (2013): http://dx.doi.org/10.4236/sn.2013.23011.

The Social Gestalt model adds a new dimension to studies on *interactions* in complex social networks. It allows researchers to identify and to examine special regularities of network structures based on *interpersonal attraction* and *characteristic features of the subjects*. Fundamental findings in psychology/sociology and physics are used as a basis for the development of the intensity function of interpersonal attraction.

The new objectives of this paper are the application of the methods based on the model of the Social Gestalts to male and to female networks for the proof of the hypothesis that the shapes of male and female Social Gestalts are different from each other.

The paper "Who is Collaborating with Whom?" is presented in two parts:

- Part I: Mathematical Model and Methods for Empirical Testing.
- Part II: Application of the Methods to Male and to Female Networks.

2. Theoretical approach of the mathematical model of Social Gestalts

There are two parts of the theoretical approach.

Sections 2.1–2.4: Theoretical approach in general.

Section 2.5: Intensity Function of Interpersonal Attraction: Special Model for the Distribution of Co-Author Pairs' Frequencies.

2.1. Fundamental findings in psychology/sociology and physics used as basis for the development of the intensity function of interpersonal attraction to describe the Social Gestalts

Fundamental findings in psychology/sociology and physics are used as a basis for the development of the intensity function of interpersonal attraction. This intensity structure is described by a special power function's combination based on two crucial determinants of interpersonal attraction (psychology) and the model of complementarities (physics). For better understanding the visible varying shapes of the 3-D graphs in this paper, the development of this intensity function of interpersonal attraction is explained in the Sections 2.2–2.4. The model is applied on co-authorship pair frequencies' distributions (Z_{XY}) which are particularly conditioned by the productivity level of each author in a dyad (X and Y).

Eqs. (1) and (2) are derived from findings in psychology/sociology but Eqs. (3) and (4) from findings in physics. The Eqs. (1)–(4) are the basis for the Eqs. (5)–(8). The Eqs. (1)–(8) are parts of the Intensity Function of Interpersonal Attraction.

2.2. Interpersonal attraction (psychology/sociology)

Interpersonal attraction is a major area of study in social psychology (Wolf, 1996). Whereas in physics, attraction may refer to gravity or to the electro-magnetic force, interpersonal attraction can be thought of *force* acting between two people (attractor/attracted) tending to draw them together, i.e. creating *social interactions* (as for example co-authorship):

- When measuring the *intensity of interpersonal attraction between two individuals* in a social network, one must refer to the qualities of the attracted (Variable X) as well as the qualities of the attractor (Variable Y). That means one must refer to their personal characteristics.
- This *intensity* is mirrored in the magnitude or frequency of social interactions (Variable Z_{XY}) as dependent third variable.

Examples (types) of qualities or personal characteristics of these individuals are age, labor productivity, education, professional status, gender, etc., while examples (types) of social interactions are collaboration, friendships, marriages, etc. In the present paper we study especially co-authorship networks (Z_{XY}) depending on productivity of the individuals (measured by X and Y with $X_{\min} = Y_{\min}$ and $X_{\max} = Y_{\max}$).

Interpersonal attraction reveals two pivotal aspects (Wolf, 1996):

- "Similarity" and "dissimilarity" of personal characteristics:

"*Similarity*": The intensity of interpersonal attraction between individuals who have similar values of X and Y is stronger than between individuals having dissimilar values of X and Y.

"*Dissimilarity:* The reverse phenomenon is emergent.

- "Edge effect" and "medium effect":

"*Edge effect*": The intensity of interpersonal attraction between individuals who have either high values of X and Y or low values of X and Y is stronger than between individuals having medium level of X and Y.

"*Medium effect*": The reverse phenomenon is emergent.

The *crucial determinant of interpersonal attraction (similarity or dissimilarity)* suggests considering the *distance A* between the personal characteristics for the derivation of the intensity function of interpersonal attraction:

$$A = \left|X - Y\right| \tag{1}$$

As the *opposite* of subtraction A, the other *crucial determinant of interpersonal attraction (edge or medium effects)* suggests considering the *addition B* of the personal characteristics:

$$B = X + Y \tag{2}$$

Important remarks for the development of the "Intensity Function of Interpersonal Attraction": A and B can vary independently from each other.

For example, X is equal to Y: $A = |X - Y| = 0$ is constant but the value of B can vary.

Examples of varying B values: $B = 2$ with $X = Y = 1$, $B = 4$ with $X = Y = 2$, $B = 6$ with $X = Y = 3$, etc.

Because of possible independent variations of A and B both Eqs. (1) and (2) are selected as parts for the intensity function.

2.3. Complementarities (physics)

The notion of "birds of a feather flock together" points out that similarity is a crucial determinant of interpersonal attraction. Do birds of a feather flock together, or the opposites attract? The "edge effect" is another crucial determinant of interpersonal attraction. Is the "edge effect" dominant or the "medium effect"?

Both opposing proverbs and both the edge and the medium effects give rise to reflection about the notion of complementarity. Capra (1996, p. 160) wrote that the term complementarity (e.g. particle/wave) introduced by the physicist Niels

Table 1
Example: $A_{min} = 0$, $A_{max} = 3$.

A	$A_{complement}$
0	3
1	2
2	1
3	0

Bohr has become a firm integral part of the conceptual framework within which physicists attentively weigh the problems of nature and that Bohr had repeatedly indicated that this idea could also be beneficial outside of physics. Niels Bohr recognized that the modern notion of complementarity had existed already in a clear-cut manner in old Chinese thought, in the Yin/Yang teaching.

Yin and Yang have to be seen as polar forces of only one whole, as complementary tendencies interacting dynamically with each other, so that the entire system is kept flexible and open to *change* because there is *not only Yin or only Yang.*

In other words, there is not either "Birds of a feather flock together" or "Opposites attract"; but these complementary tendencies interact dynamically with each other, so *the entire system is open to change from accentuation of "Birds of a feather flock together" to the accentuation of "Opposites attract" and* vice versa. *The same is valid for the edge and medium effects*. This phenomenon is explained in detail in Section 2.4 and shown in Fig. 4. Three-dimensional patterns in social networks owe their shapes and the change of their shapes to the balancing interaction of these forces but these complementary tendencies are interacting dynamically with each other, producing an infinite number of shapes (Section 2.4, Fig. 5).

For the derivation of the intensity function, the model of *complementarities* leads to the conclusion:

- to use in addition to A the "*complement of A": $A_{complement}$* and
- to use in addition to B the "*complement of B": $B_{complement}$*.

$A_{complement}$:

As mentioned above, there is a *complementary variation of similarity and dissimilarity*. As dissimilarity increases between persons, similarity decreases, and vice versa. Similarity is greatest at the minimum of A and least at the maximum and vice versa, dissimilarity is greatest at the maximum and least at the minimum.

A is a variable with the two opposite poles A_{min} and A_{max}. The sum of A_{min} and A_{max} is a constant. Thus,

$$A_{complement} = A_{min} + A_{max} - A \quad (3)$$

That means, the variable $A_{complement}$ increases by the same amount as the variable A decreases and vice versa, cf. Table 1. In continuation, the final measurement of the expressions of the complementarities (similarities, dissimilarities and edge effect, medium effect) is given in Section 2.4.

$B_{complement}$:

There is a complementary variation of edge effect and medium effect.

We obtain in analogy to A and $A_{complement}$:

$$B_{complement} = B_{min} + B_{max} - B \quad (4)$$

$$A_{min} = (|X - Y|)_{min} \quad (5)$$

$$A_{max} = (|X - Y|)_{max} \quad (6)$$

$$B_{min} = (X + Y)_{min} \quad (7)$$

$$B_{max} = (X + Y)_{max} \quad (8)$$

Measurement of the variables X and Y including $X_{min} = Y_{min}$ and $X_{max} = Y_{max}$ depends on the subject being studied.

Examples (types) of characteristics or of qualities of these individual persons (X or Y) are age, labor productivity, education, professional status, degree of a node in a network, etc.

2.4. Intensity function of interpersonal attraction based on fundamental findings in psychology/sociology and physics

Eqs. (1) and (2) derived from findings in psychology/sociology as well as Eqs. (3) and (4) derived from findings in physics and following the Eqs. (5)–(8) are parts of the *Intensity Function of Interpersonal Attraction.*

We assume the *intensity structure of mutual* attraction Z_{XY} can be described by a function of a special power functions' combination (X is the value of a special personality characteristic (quality) of an attracted and Y is the value of the same personality characteristic (quality) of the attractor and in case of mutual attraction also vice versa).

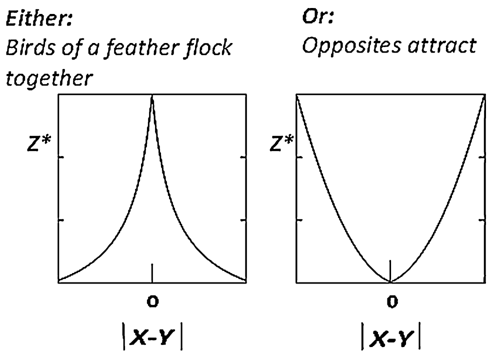

Fig. 3. Power functions with different values of parameter α (non-log presentation). In both patterns $|X-Y|$ is the abscissa with $|X-Y|=0$ in the middle (Similarity is highest) and Z^* is the ordinate. On the left pattern, the parameter α is negative: "Birds of a feather flock together", i.e. decrease of interpersonal relations with increasing dissimilarity. On the right pattern, the parameter α is positive: "Opposites attract", i.e. increase of interpersonal relations with increasing dissimilarity (This figure is a copy of Fig. 12 in Kretschmer & Kretschmer, 2007).

The *crucial determinant of interpersonal attraction (similarity or dissimilarity)* suggests considering the *distance A* between the qualities of persons (Eq. (1)) as the independent variable of a power function:

$$Z^* = c_1 \cdot (A+1)^{\alpha} \tag{9}$$

with c_1 = constant; the 1 is added to A because $\log A$ is not possible in case $A=0$. We see that *as A increases, dissimilarity increases.*

A power function with only one parameter (unequal to zero) is either only monotonically decreasing or only monotonically increasing; when referred to both proverbs we obtain: *either "Birds of a feather flock together" or" the "Opposites attract"*, cf. Fig. 3.

In order to fulfill the inherent requirement that both proverbs with their extensions can be included in the representation, the next step of approximation follows.

The model of *complementarities* leads to the conclusion to use additionally to the distance A in Eq. (9) the *complement of the distance A: "$A_{\text{complement}}$"* as the independent variable of a second power function (Eq. (10)).

$$Z** = c_2 \cdot (A_{\text{complement}} + 1)^{\beta} \tag{10}$$

Both power functions are combined in Eq. (11):

$$Z_A = \text{constant}_A \cdot (A+1)^{\alpha} \cdot (A_{\text{complenet}} + 1)^{\beta} \tag{11}$$

The relationships of the two parameters α and β to each other (cf. Eq. (11)) determine the expressions of the complementarities (similarities, dissimilarities) in each of the 8 shapes in Fig. 4, left side. In correspondence with changing relationships of the two parameters α and β to each other a systematic variation is possible from "Birds of a feather flock together" to "Opposites attract" and vice versa. Let us repeat what we said in Section 2.3: There is not either "Birds of a feather flock together" or "Opposites attract"; but these complementary tendencies interact dynamically with each other, producing an infinite number of shapes. That means, the 8 shapes in Fig. 4, left side box, are only special selected patterns for demonstration but in reality an infinite number of shapes exist between each two neighboring shapes. The same is valid for the edge and medium effects, Fig. 4, right side box.

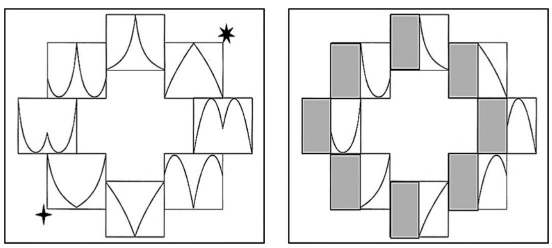

Fig. 4. Left side box: Symmetrical patterns with varying relationships of the two parameters α and β to each other (non-log presentation, but the same kind of patterns can also be produced by log–log presentation). In all of the 8 patterns $|X-Y|$ is the abscissa with $|X-Y|=0$ in the middle and Z_A (Eq. (11)) is the ordinate (this figure is a copy of Fig. 7 in Kretschmer & Kretschmer, 2007). Both, one example of the original patterns and its mirror image are marked by stars. Right side box: Non-symmetrical patterns with varying relationships of the two parameters γ and δ to each other (non-log presentation, but the same kind of patterns can also be produced by log–log presentation). In all of the 8 patterns $(X+Y)$ is the abscissa and Z_B (Eq. (14)) is the ordinate. Note: the 8 shapes both left side and right side boxes, are special selected patterns for demonstration but in reality an infinite number of shapes exist between each two neighboring shapes.

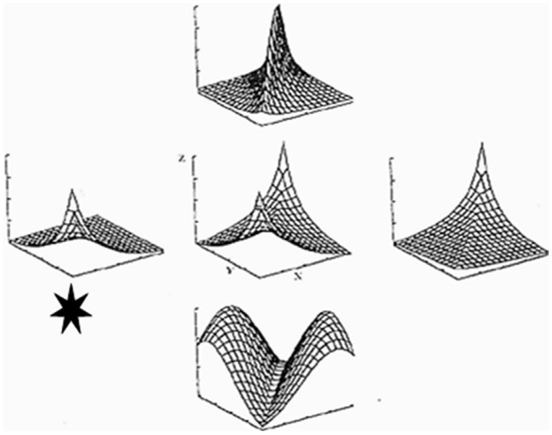

Fig. 5. Prototypes of social Gestalts (non-logarithmic presentation). The 3-D graphs owe their shapes to the balancing interaction of the complementary poles. The distribution of co-author pairs' frequencies is similar to the left prototype (marked by a star) in Fig. 5. However, in this paper we are showing the corresponding log–log–log 3-D presentation ($\log N_{ij}$ with $\log i$ and $\log j$). Several empirical patterns matching the 5 Prototypes in Fig. 5 were already taken out and presented in Fig. 4, Kretschmer (2002) and in Fig. 6, Kretschmer and Kretschmer (2007).

Left side box: While in the right neighboring pattern (marked by a star) of the upmost pattern, "Birds of the feather flock together" is more likely to be in the foreground, the left neighboring pattern (marked by another star) of the bottom pattern reveals that "Opposites attract" is more likely to be salient. Starting pattern by pattern counter clockwise from the right neighboring pattern of the upmost pattern towards the left neighboring pattern of the bottom pattern, "Birds of a feather flock together" diminishes as "Opposites attract" emerges. Vice versa, starting pattern by pattern counter clockwise from the left neighboring pattern of the bottom pattern toward the right neighboring pattern of the upper pattern, "Opposites attract" diminishes as "Birds of the feather flock together" emerges.

For the *purpose of completion*,

- Let the addition B (Eq. (2)) as the *opposite* of subtraction A (Eq. (1)), be the independent variable of the third power function.

$$Z***=c_3\cdot(\mathrm{B}+1)^{\gamma} \tag{12}$$

- and the complement, Eq. (4) be the independent variable of the fourth power function.

$$Z****=c_4\cdot(B_{\text{complement}}+1)^{\delta} \tag{13}$$

Both power functions are combined in Eq. (14):

$$Z_B=\text{constant}_B\cdot(B+1)^{\gamma}\cdot(B_{\text{complement}}+1)^{\delta} \tag{14}$$

The relationships of the two parameters γ and δ to each other determine the expressions of the complementarities (edge effect, medium effect) in each of the non-symmetrical 8 shapes; cf. Fig. 4, right side box. In correspondence with changing relationships of the two parameters γ and δ to each other a systematic variation is possible from "Edge effect" to "Medium effect" and vice versa.

All of the patterns in Fig. 4 (both on the left and on the right sides boxes) are arranged that the shapes of the opposite placed patterns show the mirror images of the original shapes (For example the shapes of the upmost and bottom patterns) in connection with opposite meanings regarding the relationships of the two proverbs (or of the other complementarities: edge effect, medium effect) to each other.whereas the shapes of the neighboring patterns look similar to each other these similarities decrease toward the oppositely placed patterns up to the mirror images of the original shapes.

The strength of visible similarities between two shapes is mirrored in the strength of similarities of their parameter values of the mathematical model (α, β, γ and δ), i.e. the differences between the parameter values of two shapes increase with the increasing dissimilarities of their shapes. The differences are highest between an original pattern and its mirror image (For example, the two shapes marked by stars).

Because the function Z_A can vary independently from the function Z_B we assume the intensity of mutual attraction Z_{XY} is proportional to the product of the two functions Z_A and Z_B:

$$Z_{XY}\sim Z_A\cdot Z_B \tag{15}$$

Whereas Z_A and Z_B each alone produce two-dimensional patterns (2-D patterns), the bivariate function Z_{XY} (cf. Eq. (16)) shows three-dimensional patterns (3-D patterns), cf. Fig. 5 with non-logarithmic presentation. The distribution of co-author pairs' frequencies N_{ij} is similar to the left prototype in Fig. 5. However, in this paper we are showing the corresponding log–log–log 3-D presentation (log N_{ij} with log i and log j). As already mentioned in Section 2.3 "Complementarities (Physics)" the 3-D graphs owe their shapes to the balancing interaction of the complementary poles.

In continuation of Eq. (15) the Intensity Function of Interpersonal Attraction (Social Gestalt) can be formalized as follows:

$$Z_{XY} = \text{constant} \cdot (A+1)^{\alpha} \cdot (A_{\text{complement}}+1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{\text{complement}}+1)^{\delta} \tag{16}$$

with $A = \left|X - Y\right|$ and $B = X + Y$.

Measurement of the variables X, Y and Z_{XY} including $X_{\min} = Y_{\min}$ and $X_{\max} = Y_{\max}$ depends on the subject being studied.

As already mentioned above, examples (types) of social interactions (Z_{XY}) are collaboration, co-authorships, friendships, marriages, etc., while examples (types) of characteristics or of qualities of these individual persons (X or Y) are age, labor productivity, education, professional status, degree of a node in a network, etc.

2.5. *Intensity function of interpersonal attraction: special model for the distribution of co-author pairs' frequencies*

In the present paper we describe the example of how to measure the variables X and Y in relation to the function of the distribution of co-author pairs' frequencies with $Z'_{XY} = N_{ij}$. The physicist and historian of science de Solla Price (1963) conjectured that the logarithm of the number of publications has greater importance than the number of publications per se.

Thus, using the logarithm of the number of publications: log i or log j respectively with $(\log i)_{\min} = 0$ and $(\log j)_{\min} = 0$ as an indicator of the personal characteristic 'productivity', we define:

$$X = \log\ i \tag{17}$$

$$Y = \log\ i \tag{18}$$

$$A = \left|\log\ i - \log\ j\right| \tag{19}$$

$$B = \log\ i + \log\ j \tag{20}$$

Consequently:

$$\begin{aligned} A_{\min} &= \left|X - Y\right|_{\min} = 0 \\ &\text{with } \log\ i = \log\ j \end{aligned} \tag{21}$$

$$A_{\max} = \left|X - Y\right|_{\max} = \left|(\log\ i)_{\max} - \log\ 1\right| = \left|\log\ 1 - (\log\ j)_{\max}\right| = (\log\ i)_{\max} = (\log\ j)_{\max} \tag{22}$$

$$B_{\min} = (X + Y)_{\min} = \log\ 1 + \log\ 1 = 0 \tag{23}$$

$$B_{\max} = (X + Y)_{\max} = (\log\ i)_{\max} + (\log\ j)_{\max} = 2(\log\ i)_{\max} = 2(\log\ j)_{\max} \tag{24}$$

Let us assume a specific value for the maximum possible number of publications i (or j respectively) of an author as a standard for such studies, which does not vary depending upon the given sample. We assume that the maximum possible number of publications of an author is equal to 1000, i.e.

$$A_{\max} = \log\ 1000 = 3 \tag{25}$$

$$B_{\max} = 2A_{\max} = 6 \tag{26}$$

Thus it follows that:

$$A_{\text{complement}} = 3 - \left|\log\ i - \log\ j\right| \tag{27}$$

$$B_{\text{complement}} = 6 - (\log\ i + \log\ j) \tag{28}$$

Thus, the theoretical mathematical function for describing the social Gestalts of the distribution of co-author pairs' frequencies results in the logarithmic version

$$\log\ N_{ij} = c + \alpha \cdot \log(\left|X - Y\right| + 1) + \beta \cdot \log(4 - \left|X - Y\right|) + \gamma \cdot \log(X + Y + 1) + \delta \cdot \log\ \ (7 - X - Y) \tag{29}$$

with $X = \log i$ and $Y = \log j$ and with c = constant.

As mentioned above, whereas Z_A (Eq. (11)) and Z_B (Eq. (14)) each alone produce two-dimensional patterns, the bivariate function Z_{XY} (Eq. (16)) shows three-dimensional patterns. In analogy:

Whereas log N_{ij} (Eq. (29)) produces 3-D patterns,

$$\log N_A = \text{constant}_{NA} + \alpha \cdot \log(\left|X - Y\right| + 1) + \beta \cdot \log(4 - \left|X - Y\right|) \tag{30}$$

with $X = \log i$ and $Y = \log j$

Table 2
Artificial table of co-author pairs N_{ij}.

i/j	1	2	3	N_i
1	30	20	10	60
2	20	25	5	50
3	10	5	2	17
N_j	60	50	17	$N = 127$

Note: $Ni = \Sigma jNij$ is the sum of co-authors of all authors with i publications per author.
$Nj = \Sigma iNij$ is the sum of co-authors of all authors with j publications per author.
N = Total sum of degrees of all nodes in a network, equal to the total sum of pairs including F_x each, with x ($x = 1, 2, \ldots, n$).

and

$$\log N_B = \text{constant}_{NB} + \gamma \cdot \log(X + Y + 1) + \delta \log(7 - X - Y) \quad (31)$$

with $X = \log i$ and $Y = \log j$

each alone produce two-dimensional patterns independently from each other.

In analogy to the functions Z_A (Eq. (11)) and Z_B (Eq. (14)) the function $\log N_A$ can vary independently from the function $\log N_B$. Following:

$$\log\ N_{ij} = \text{constant}_{Nij} + \log\ N_A + \log\ N_B \quad (32)$$

3. Methods (partly presented in previous studies of Social Gestalts)

3.1. Remarks

Some of the methods are partly presented in previous studies: Kretschmer and Kretschmer (2009), Kretschmer, Kundra, Beaver, and Kretschmer, (2012), Ozel et al. (2014).

The method of counting co-author pairs based on social network analysis (SNA), the logarithmic binning procedure and the method of visualizing the 3-D collaboration patterns are presented. PWQ is used as an example.

3.2. Method of counting co-author pairs, based on social network analysis (SNA)

For the purposes of analysis, a social network can be considered as consisting of two sets, a set of n nodes (individuals) and a set of m edges (undirected relations) between pairs of the nodes. The degree of a node F_x with x ($x = 1, 2, \ldots, n$) is equal to the number of nodes (or edges) that are attached to the node F_x. In co-authorship networks between two authors (nodes) F_x and F_y, there exists an edge if both were acting as co-authors one times at least.

"Example:

The authors, F_1, F_2 and F_3 have published a paper together. Thus three edges are existing producing three co-author pairs:

- One edge between the pair F_1 and F_2.
- One edge between the pair F_1 and F_3.
- One edge between the pair F_2 and F_3".

An author's productivity is measured by his number of publications. The number of publications i per author F_x or j per possible co-author F_y respectively are determined by using the 'normal count procedure'. Each time the name of an author appears, it is counted.

The n authors F_x are grouped according to their productivities i or j respectively. The co-author pairs of authors F_{xi} (who have the number of publications i) in co-authorship with authors F_{yj} (who have the number of publications j), are counted. The resulting sum of co-author pairs N_{ij} is equal to the sum of degrees of the authors F_{xi} to the co-authors F_{yj}. Therefore, the matrix of N_{ij} is symmetrical (cf. Table 2).

In other words: N_{ij} is equal to the sum of co-author pairs of authors who have the number of publications i in co-authorship with authors who have the number of publications j.

N is equal to the total sum of degrees of all n nodes (all authors F_x) in a network, equal to the total sum of pairs.

3.3. Logarithmic binning procedure

Distributions of this kind of co-author pairs' frequencies (N_{ij}) have already been published (Guo, Kretschmer, & Liu, 2008; Kretschmer & Kretschmer, 2007; Kundra, Beaver, Kretschmer, & Kretschmer, 2008). However, these distributions were restricted to $i_{\max} = 31$.

Usually the stochastic noise increases with higher productivity because of the decreasing number of authors. We intend to overcome this problem in this paper with help of the *logarithmic binning procedure*. Newman has already proposed in

Table 3
Matrix of $N^S{}_{ij}$ in dependence on i' (bin) and j' (bin) (PWQ) with $N = 4324$. Copy of Table 3 from Kretschmer et al. (2012).

i' (bin)/j' (bin)	1	2	4	8	Sum
1	2228	499	186	143	3056
2	499	154	53	52	758
4	186	53	34	24	297
8	143	52	24	12	231
Sum	3056	758	297	231	$N = 4342$

Table 4
Matrix of the width of the bin: $\Delta i' \cdot \Delta j'$ (as example from 1–8 only).

$\Delta i'/\Delta j'$	1	2	4	8
1	1	2	4	8
2	2	4	8	16
4	4	8	16	32
8	8	16	32	64

2005 using the logarithmic binning procedure for the log–log scale plot of power functions. To get a good fit of a straight line (log–log scale plot of power functions, for example Lotka's distribution), we need to bin the data i into exponentially wider bins. Each bin is a fixed multiple wider than the one before it. For example, choosing the multiplier of 2 we receive the intervals 1 to 2, 2 to 4, 4 to 8, 8 to 16, etc. For each bin we have ordered the corresponding first value of i (or j) to this bin. Thus, the sequence of bins i' or j' is:i' ($i' = 1, 2, 4, 8, 16, 32, 64, 128, 256, \ldots$). The same holds for the bins j'. The sizes or widths of the bins ($\Delta i'$) are: 1, 2, 4, 8, 16, etc. The same holds for ($\Delta j'$).

However, because of the bivariate presentation the width of a bin ($\text{cell}_{i'j'}$) in the matrix is the product of $\Delta i'$ and $\Delta j' = (\Delta i' \cdot \Delta j')$. The sum of co-author pairs in a bin ($\text{cell}_{i'j'}$) is called $N^S{}_{ij}$, cf. Table 3. The total sum of $N^S{}_{ij} = \sum_{ij} N^S{}_{ij}$ is equal to the total number of co-author pairs N of a co-authorship network:

$$N = \sum_{ij} N_{ij}^S$$

3.4. *Method of visualizing the 3-D collaboration patterns (PWQ is used as an example)*

3.4.1. *Remarks*

The following methods will be presented (PWQ is used as an example):

- Visualizing empirical patterns (Distributions of Co-Author Pairs' Frequencies).
- Visualizing theoretical patterns and overlay of empirical and theoretical patterns into a single frame.
- Standardization of theoretical patterns.

3.4.2. *Visualizing empirical patterns (distributions of co-author pairs' frequencies)*

For visualizing the original data we use the sum of co-author pairs in a bin ($\text{cell}_{i'j'}$), i.e. $N^S{}_{ij}$ directly in dependence on i' (bin) and j' (bin) (cf. Table 3). Because log 0 is not given, we are using the value "0" for presentation of $N^S{}_{ij}$ in the tables but not for regression analysis. One case is presented in Part II: Section 3.1, Table 2.

The matrix of $N^S{}_{ij}$ (Distribution of co-author pairs' frequencies), Table 3, is used as an example for explaining the following steps of the methods.

As the next step in the logarithmic binning procedure: $N^S{}_{ij}$ of a cell ($\text{cell}_{i'j'}$) has to be divided by the width of the bin: ($\Delta i' \cdot \Delta j'$), matrix of the width, cf. Table 4. In other words, the new value in a bin (Example, cf. Table 5 for PWQ) is simply the arithmetic average of all the points in the bin. This new value, i.e. the ratio, is called the average co-author pairs' frequency N'_{ij}.

Using the log–log–log presentation after the logarithmic binning procedure, the sequence of $\log i'$ (rows) is as follows: $\log i'$ ($\log i' = 0, 0.301, 0.602, 0.903, \ldots$); the same holds for $\log j'$ (columns) resulting in a square matrix. An example of the

Table 5
Matrix of the average co-author pairs' frequencies N'_{ij} in dependence on i' (bin) and j' (bin) (PWQ).

i' (bin)/j' (bin)	1	2	4	8
1	2228	249.5	46.5	17.875
2	249.5	38.5	6.625	3.25
4	46.5	6.625	2.125	0.75
8	17.875	3.25	0.75	0.1875

Table 6
Matrix of $\log N'_{ij}$ in dependence on $\log i'$ and $\log j'$ (PWQ).

$\log i'/\log j'$	0	0.30103	0.60205999	0.90308999
0	3.34791519	2.39707055	1.66745295	1.25224605
0.30103	2.39707055	1.58546073	0.82118588	0.51188336
0.60205999	1.66745295	0.82118588	0.32735893	−0.12493874
0.90308999	1.25224605	0.51188336	−0.12493874	−0.72699873

matrix of the logarithm of the average co-author pairs' frequencies $\log N'_{ij}$ is shown in Table 6 (distribution of co-author pairs' frequencies in logarithmic version: $\log N_{ij}$). The values are obtained from PWQ.

In 3-D presentations $\log i'$ is placed on the X-axis, $\log j'$ on the Y-axis and $\log N'_{ij}$ on the Z-axis, cf. Fig. 6 as an example.

The view at the three patterns on the first row and at the rightmost pattern at the second row is given from the bottom right corner of the matrix, Table 6, to the top left corner (i.e. along the main diagonal).

One can follow the process of making these patterns visible starting with the leftmost pattern on the first row of Fig. 6 followed by the other two patterns. The empirical values ($\log N'_{ij}$) are presented as dots on the top of the corresponding vertical spikes (But empirical values (dots) can also be used separately without any vertical spikes).

On the leftmost pattern (first row) one can see the dots on the main diagonal for

- $\log N'_{ij} = -0.72699873$ with $\log i' = \log j' = 0.90308999$ in front,
- in the middle: $\log N'_{ij} = 0.32735893$ with $\log i' = \log j' = 0.60205999$ and $\log N'_{ij} = 1.58546073$ with $\log i' = \log j' = 0.30103$ and
- $\log N'_{ij} = 3.34791519$ with $\log i' = \log j' = 0$ in the background.

On the second pattern (first row) all of the 16 empirical values (dots on the top of the corresponding vertical spikes) are plotted (But empirical values (dots) can also be separately used without any vertical spikes).

- The curvilinear line from left to right of the second pattern (first row) presents the transferred data from the secondary diagonal (Table 6) to the corresponding places of the Z-axis.
- The curvilinear line from top to bottom presents the transferred data from the main diagonal (Table 6) to the corresponding places of the Z-axis.

3.4.3. Visualizing the theoretical pattern and overlay of empirical and theoretical patterns into a single frame

For better understanding; as the first step we show *examples* after overlay of empirical and theoretical patterns into a single frame presented in Fig. 6. Explanation about visualizing the theoretical pattern and the method of overlay of empirical and theoretical patterns into a single frame follow afterwards.

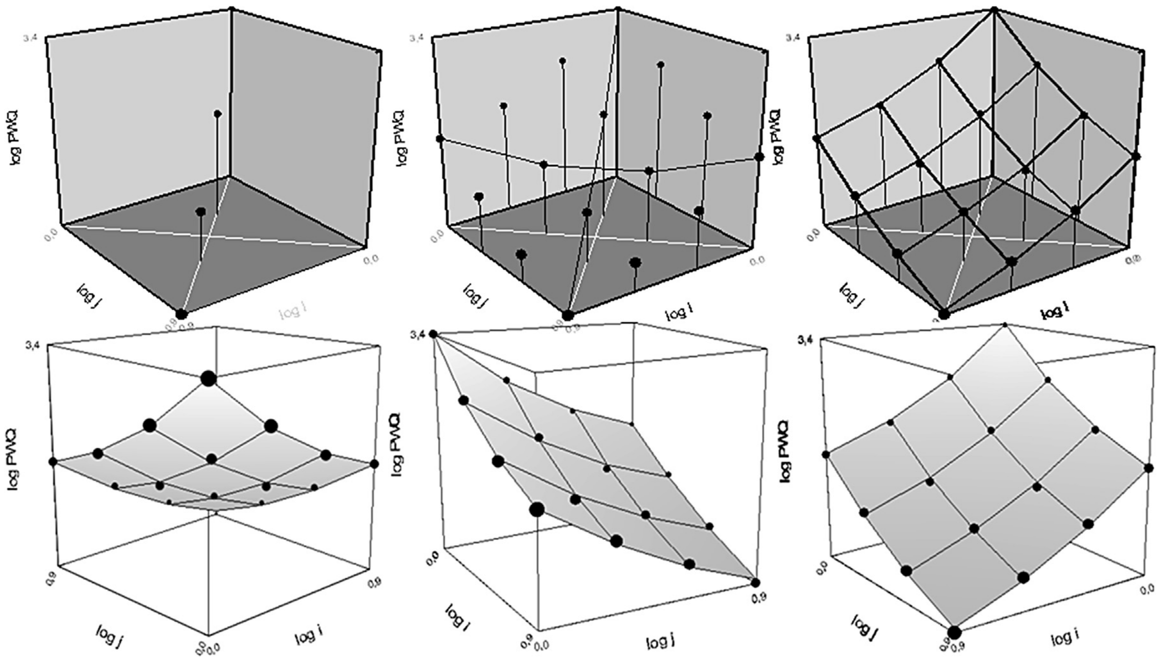

Fig. 6. Visualizing 3-D collaboration patterns (PWQ is used as an example).

Examples:

The rightmost pattern on the first row of Fig. 6 presents the overlay of the empirical data (dots) taken from the second pattern of the first row and the corresponding theoretical pattern (lines). But the overlay of empirical dots and theoretical lines in combination with the appearance of the corresponding white colored 3-D surface can be found on the three patterns, second row.

As mentioned above, the Social Gestalt shows well-ordered three-dimensional bodies, totally rotatable around and their manifold shapes are visible in the space from all possible points of view. Thus two examples, i.e. the rotation twice in succession of the rightmost pattern, second row, are selected. The view at the pattern in the middle is given from the lower left entry of the matrix (Table 6) to the upper right entry (i.e. along the secondary diagonal). The view at the leftmost pattern is given from the top left corner of the matrix, Table 6, to the bottom right corner (i.e. along the main diagonal).

Method of Visualizing Theoretical Patterns and Overlay

Theoretical patterns are obtained by regression analysis based on the mathematical model for the intensity function of interpersonal attraction (Eq. (29)). For visualizing the theoretical patterns (lines and/or the 3-D surfaces as in Fig. 6) in combination with the empirical values (dots) we use the Function Plot of SYSTAT for the theoretical and the Scatterplot for the empirical patterns.

After regression analysis using the Eq. (29) with $\log N_{ij} = \log N'_{ij}$ and after logarithmic binning, we obtain 4 parameters α, β, γ, and δ plus a constant c which are entered into the Function Plot (Z is the dependent variable and X and Y are the independent):

$$Z = c + \alpha \cdot \log(\left|X - Y\right| + 1)\beta \cdot \log(4 - \left|X - Y\right|) + \gamma \cdot \log(X + Y + 1) + \delta \cdot \log(7 - X - Y) \tag{33}$$

The parameter values and the constant for PWQ can be found in Part II: Section 3.2, Table 7.

Scale Range: The maximum and minimum values appearing on the axis are specified, i.e. both all of the empirical and corresponding theoretical data have to be presented. Any data values outside these limits will not appear on the display. The minimum for the X-axis is in principle specified as 0 ($(\log i')_{\min} = 0$) and the maximum is equal to $(\log i')_{\max}$ of the empirical data. For example, in Table 6: $(\log i')_{\max} = \log 8$. The same holds for the Y-axis $(\log j')_{\max} = \log 8$.

The minimum and maximum values for the Z-axis are selected according to the minimum and maximum values of the whole Gestalt produced by the function. In case there are empirical values greater or less than these two theoretical values, the minimum or maximum of the Z-axis has to be extended accordingly so that all of the empirical values become visible.

The Surface and Line Style dialog box is used to customize the appearance of lines or surfaces. The used XY Cut Lines are in two directions. The number of cuts in the grid has to be specified by the number of bins i' (or j' respectively) minus 1 in the data set. For example, a special data set has 4 bins i' as in Table 6 (PWQ); the number of cuts in the grid is specified by $4-1=3$. The resulting number of lines of the theoretical pattern (Gestalt) is equal to the double of the number of bins i' ($2\cdot 4 = 8$, cf. Fig. 6). The number of points where two of the lines intersect, is equal to the square of the number of bins i' ($4^2 = 16$). The Scale Range of the empirical pattern has to be equal (or slightly less) to the theoretical Gestalt (cf. Fig. 6).

After the overlay of the empirical distribution and the theoretical pattern into a single frame as in Fig. 6 the goodness-of-fit is highest in the case where the empirical values (dots) are directly placed on the points where two of the theoretical lines intersect. In the case the distance between the intersection points and the dots increases, the goodness-of-fit decreases.

For simplification hereafter we use the dots for presenting the empirical values but not the vertical spikes.

Analogously to the rotated 3-D collaboration pattern (PWQ), shown at the second row of Fig. 6, the results after overlay of empirical and theoretical patterns into single frames are presented for the six bibliographies (including PWQ) in Part II: Section 3.4, Figs. 5 and 6.

Standardization of theoretical patterns for pattern comparison

For comparing the theoretical patterns of the six bibliographies with each other the theoretical patterns are standardized as follows:

- The number of cuts in the grid is specified by 6 for all of the theoretical patterns. Thus, the resulting number of lines in a theoretical pattern equals 14.
- There are no axes to display.
- There are no scales to display.
- All of the theoretical patterns are zoomed and presented at the same size.

Standardized theoretical patterns presented in Fig. 7:

- After upwards rotation once of the leftmost pattern on the second row of Fig. 6, this pattern is selected as the basis for the standardized patterns in Fig. 7.

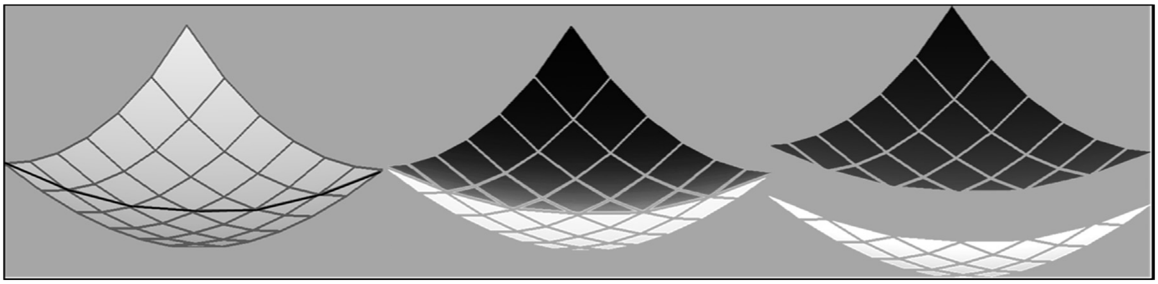

Fig. 7. Three versions of standardized theoretical 3-D patterns based on the data from PWQ.

- (6-PWQ): The curvilinear line from left to right of the leftmost pattern (Fig. 7) presents the data from the secondary diagonal (Table 6) transferred to the corresponding places of the *Z*-axis.
- The difference between several standardized theoretical patterns can become pronounced as shown in Fig. 7 by using dark color for the part of a pattern obtained from the data left on the secondary diagonal of the matrix of log N'_{ij} and light color for the other part (cf. pattern in the middle of Fig. 7). This can become reinforced by additional separation of the two parts (cf. rightmost pattern of Fig. 7).

Whereas in Part I examples of using the methods is given with help of one of the bibliographies especially; in Part II both the theoretical and empirical patterns of the six bibliographies are compared and the differences between the female and male groups will be specially discussed.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2015.01.004.

References

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *23*(3), 85–88.

Capra, F. (1996). *Wendezeit. Bausteine für ein neues Weltbild*. München: Deutscher Taschenbuch Verlag GmbH & Co. KG.

Carr, P. L., Pololi, L., Knight, S., & Conrad, P. (2009). Collaboration in academic medicine: Reflections on gender and advancement. *Academic Medicine*, *84*(10), 1447–1453.

Egghe, L. (2008). A model for the size-frequency function of co-author pairs. *Journal of the American Society for Information Science and Technology*, *59*(13), 2133–2137. November 2008

Guo, H., Kretschmer, H., & Liu, Z. (2008). Distribution of co-author pairs' frequencies of the journal of information technology. *COLLNET Journal of Scientometrics and Information Management*, *2*(1), 73–81.

Kretschmer, H. (1999). A new model of scientific collaboration. Part I: Types of two-dimensional and three-dimensional collaboration patterns. *Scientometrics*, *46*(3), 501–518.

Kretschmer, H. (2002). Similarities and dissimilarities in co-authorship networks: Gestalt theory as explanation for well-ordered collaboration structures and production of scientific literature. *Library Trends*, *50*(3), 474–497.

Kretschmer, H., & Kretschmer, T. (2007). Lotka's distribution and distribution of co-author pairs' frequencies. *Journal of Informetrics*, *1*, 308–337.

Kretschmer, H., & Kretschmer, T. (2009). (2009, Invited Keynote Speech): Who is collaborating with whom? Explanation of a fundamental principle. In H. Y. Hou, B. Wang, S. B. Liu, Z. G. Hu, X. Zhang, & M. Z. Li (Eds.), *Proceedings of the 5th international conference on webometrics, informetrics and scientometrics and 10th COLLNET meeting*. China: 13–16 September, Dalian (CD-ROM for all participants and for libraries).

Kretschmer, H., Kundra, R., Beaver D.deB., & Kretschmer, T. (2012). Gender bias in journals of gender studies. *Scientometrics*, *93*, 135–150. http://dx.doi.org/10.1007/s11192-012-0661-5

Kretschmer, H., & Kretschmer, T. (2013). (2013, Invited Paper, open access): Who is collaborating with whom in science? *Explanation of a Fundamental Principle Social Networking*, *2*, 99–137. http://dx.doi.org/10.4236/sn.2013.23011. http://dx.doi.org/10.4236/sn.2013.23011. Published online July 2013

Kretschmer, H., Kretschmer, T., & Stegmann, J. (2014). Growth and structure formation of collaboration patterns obtained from the journals PNAS, SCIENCE and the journal of experimental medicine. *COLLNET Journal of Scientometrics and Information Management*, http://dx.doi.org/10.1080/09737766.2014.916864. Published online: 29 July 2014

Kyvik, S., & Teigen, M. (1996). Child care, research collaboration, and gender differences in scientific productivity. *Science Technology and Human Values*, *21*(6), 54–71.

Kundra, R., Beaver, D. deB., Kretschmer, H., & Kretschmer, T. (2008). Co-author pairs' frequencies distribution in journals of gender studies. *COLLNET Journal of Scientometrics and Information Management*, *2*(1), 63–71.

Lotka, A. J. (1926). The frequency distribution of scientific production. *Journal of the Washington Academy of Science*, *16*, 317–323.

McGrath, W. E. (2002). Introduction. *Library Trends*, *50*(3), 309–316.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, *29*(1), 31–40.

Metzger, W. (1986). *Gestalt-Psychologie. Ausgewählte Werke aus den Jahren 1950 bis 1982 herausgegeben und eingeleitet von Michael Stadler und Heinrich Crabus*. Frankfurt am Main: Verlag Waldemar Kramer.

Morris, S. A., & Goldstein, M. L. (2007). Manifestation of research teams in journal literature: A growth model of papers, authors, collaboration, co-authorship, weak ties, and Lotka's law. *Journal of the American Society for Information Science and Technology*, *58*(12), 1764–1782.

Naldi, F., Luzi, D., Valente, A., & Parenti, I. V. (2004). Scientific and technological performance by gender. In H. F. Moed, H. F. Moed, et al. (Eds.), *Handbook of quantitative science and technology research* (pp. 299–314). Dordrecht: Kluwer Academic.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(September–October (5)), 323–351.

Ozel, B., Kretschmer, H., & Kretschmer, T. (2014). Co-authorship pair distribution patterns by gender. *Scientometrics*, *98*, 703–723. http://dx.doi.org/10.1007/s11192-013-1145-y

Price, D. J. (1963). *Little science, big science*. New York: Columbia University Press.

Pepe, A., & Rodriguez, M. A. (2009). Collaboration in sensor network research: An in-depth longitudinal analysis of assortative mixing patterns. *Scientometrics*, http://dx.doi.org/10.1007/s11192-009-0147-2. Received: 16 September 2009 The Author(s) 2009. This article is published with open access at Springerlink.com

Wolf, Ch. (1996). *Gleich und gleich gesellt sich. Individuelle und strukturelle Einflüsse auf die Entstehung von Freundschaften*. Hamburg: Verlag Dr. Kovac.

Yin, L.-c., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing and Management*, *42*(6), 1599–1613.