Distribution of Co-Author Pairs' Frequencies of the Journal of Information Technology

Hanning Guo[1] Hildrun Kretschmer[2] Zeyuan Liu[1]

14 June 2008

Abstract

Lotka's Law originally calculated only the frequencies of single or first authors. However, given the growth and popularity of scientific collaboration, the distribution of co-author pairs' frequencies should also be considered. Based on the previous researchers' results, this article will demonstrate that the distribution of co-author pairs' frequencies can be reflected by a social gestalt and that there are different shapes of this social gestalt. Results for the distribution of co-author pairs in journals are shown for the co-authorship data of the *Journal of Information Technology* (1994-2007).

1 Introduction

Lotka's law, discovered in 1926, is an important description of the relationship between authors and distribution of their scientific productivity in Bibliometrics, Scientometrics and Informetrics. Calculations treated only single or first authors, and in the latter case other co-authors were ignored. With the development of information technology, collaboration in scientific research has improved. For example, the increase in the rate of collaboration is shown in scientific papers. The phenomenon of scientific collaboration has covered the levels of individuals, institutes and countries. The opportuni ties for collaboration are increasing more and more and the number of authors per paper have grown, from single to double, triple, even tens or more. Given the expansion of scientific research cooperation in the world and the increase of multi-authored scientific papers, using the first-author only principle of Lotka's law no longer sufficiently fit the description of these developments. Thus, to consider the distribution of co-author pairs is urgent and has practical significance.

In recent decades, many researchers have focused on the issue of scientific collaboration (Glänzel, deB. Beaver, Newman, Kretschmer, etc). After studying data from *Science*, *Nature*, *Proc Nat Acad Sci USA* and *Phys Rev B Condensed Matter* from 1980 to 1998, Kret-schmer & Kretschmer discovered there is some regularity or well-ordered structure for the distribution of co-author pairs in journals in comparison with Lotka's law for the distribution of single authors (cf. Fig. 1).

According to Lotka's Law, the frequency of authors A_i with i publications per author is a function of i: $A_i = f(i)$.

However, the distribution of co-author pairs' frequencies (N_{ij}) should be considered. N_{ij} between authors with i publications per author and authors with j publications per author is a function of i and j: $N_{ij} = f(i,j)$, i and j will be counted by using the normal count procedure (Counting how many times an author appears in bibliographies).

[1] WISELAB, Dalian University of Technology, Dalian, 116023, China, blue_hanning at yahoo dot com dot cn
[2] COLLNET Center, Borgsdorfer Str. 5, D-16540 Hohen Neuendorf, Germany, kretschmer dot h at onlinehome dot de

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

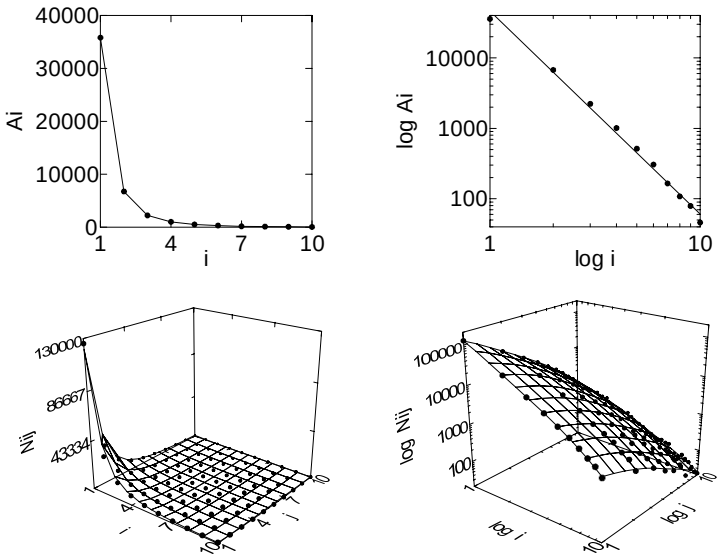

Fig. 1: Lotka's distribution in comparison with the distribution of co-author pairs' frequencies. The data are taken from the journal *Science*)

Among various researches, Kretschmer carried new mathematical approaches into the field of Scientometrics. According to the theory of configuration or Gestalt, she pointed out that the distribution of co-author pairs' frequencies can be reflected by a social gestalt by means of a nonlinear function to describe the three-dimensional pattern of scientists' co-authorship network. This configuration shows general characteristics of the social interaction structure in social networks. That is for example, "Birds of a feather flock together" or "Opposites attract". The distribution of co-author pairs' frequencies based on a social gestalt can better explain the phenomenon of scientific collaboration than Lotka's law.

Using data from the *Journal of Information Technology* (JIT), the intention of this article is to verify further the previous research results of Kretschmer—there are well-ordered structures for the distributions of co-author pairs in journals.

2 Gestalt Theory and 3D Pattern of Co-authorship Network

Gestalt theory which originated in the late 19th century emphasizes the integrity of experience and behavior. That is, the whole does not mean simply a sum of its parts, awareness does not mean merely a collection of sense elements, acts do not mean only the cycle of a reflex arc. People have psychological fields whose can be implemented by two forces, one from the structural function and purpose of the object, the other from the inner driving force of a subject's motives and needs. Gestalt theory considers various psychological processes and it indicates those holistic organizational patterns that play a role in comprising man and the environment. These holistic entities are often designated as psychological fields. Their tendency towards a stable state of order is called conciseness tendency, it is a "tendency towards a good Gestalt". The stable final state is, if possible, built up in a simple, well-ordered, harmonic and uniform manner in accordance with definite rules.

Kretschmer's study makes clear that the balance between field-force and tensile force can explain how a gestalt can exist in social networks according to basic gestalt theories. And we can give it a strict mathematical description on the basis of the general characteristics of social interactions in social networks. Two of the characteristics are well known as

"Birds of a feather flock together" or "Opposites attract". Therefore, Kretschmer chose power functions as the starting point of her research and she founded the nonlinear function model of social network configuration describing scientific collaboration:

Let Dissimilarity = $A=|X-Y|$

$\text{Similarity}=A_{COMPLEMENT}= |X-Y|_{max}+|X-Y|_{min} - |X-Y|$

Analogously to an inverse mathematical operation:

$B= X+Y$

$B_{COMPLEMENT} = |X+Y|_{max}+|X+Y|_{min} - |X+Y|$

and $Z=\text{constant} \cdot (A+1)^{\alpha} \cdot (A_{COMPLEMENT} +1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{COMPLEMENT}+1)^{\delta}$

Z are values of social interactions, X are values of a special personality characteristic of senders, however, Y are values of the same special personality characteristic of receivers and vice versa. And we get a nonlinear function with four corresponding parameters. Various prototypes of social gestalts (cf. Fig.2) can be gotten by changing those four parameters. For example, while in the upper pattern "Birds of the feather flock together" is more likely to be in the foreground, the pattern below reveals that "Opposites attract" is more likely to be accentuated. The patterns in the second row show changes regarding another pair of opposite poles.

This new social gestalt model can reflect interaction among many individuals with particular personal characteristics and we can use three-dimensional configurations to show this interaction vividly.

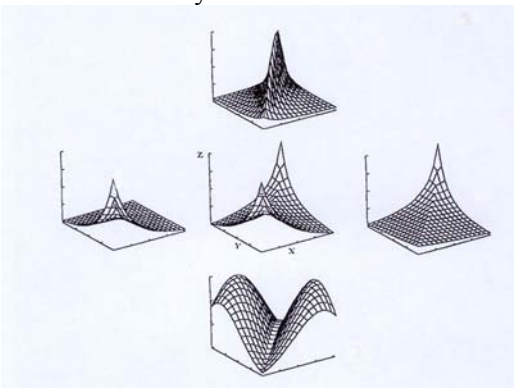

Fig.2: Prototypes of Social Gestalts (Kretschmer & Kretschmer 2007)

From Fig.2, we can observe that there is not only one shape of a social gestalt. In fact, the shape of a social gestalt will change continuously with changes in persons and their environment. Taking co-authorship networks for instance, configurations of these co-authorship networks can be different because of the research environment, method of measurement, research content, sample size, individual characteristic, collaborators' relationship and so on. In other words, these different research characteristics are the reflection of various social gestalts, just as different kinds of mirrors can reflect different distorted but well-ordered images of the same object.

Based on the description above, this article proposes the following two hypotheses through comparison with the *Journal of Biological Chemistry*:

- The distribution of co-author pairs' frequencies of the *Journal of Information Technology* can be reflected by a social gestalt.
- The shape of the social gestalt of the *Journal of Information Technology* (JIT) is different from that of the *Journal of Biological Chemistry* (JBC).

These two hypotheses are tested based on specification of X and Y (cf. Appendix I).

3 Data Source, Methods and Results

The methods can be found in Appendix II.

All the data we use in this article are from Science Citation Index (SCI) or from Web of Science. We use the *Journal of Information Technology* as journal source to retrieve all the bibliographic records of this journal from 1994 to 2007. 577 authors and the total sum of N_{ij}=962 co-author pairs will be studied in this article.

Verification of the first Hypothesis

After regression analysis the correlation between the theoretical gestalt and n=27 empirical

values of log N_{ij} is equal to R=0.97, with F-ratio=88.97. The error probability is equal to P=0.000000000 (very high statistical significance). From the regression analysis the four resulting parameters and constant, are used in the model of social gestalts (mathematical function 20, cf. Appendix I). Fig. 3 shows the triple logarithmic presentation of the social gestalt in form of a lattice (Z=log N_{ij}, X=log i, Y=log j). The corresponding empirical data are entered in red colored dots. The first picture of the gestalt in Fig. 3 is rotated twice (3 pictures of the same gestalt in Fig. 3 in total).

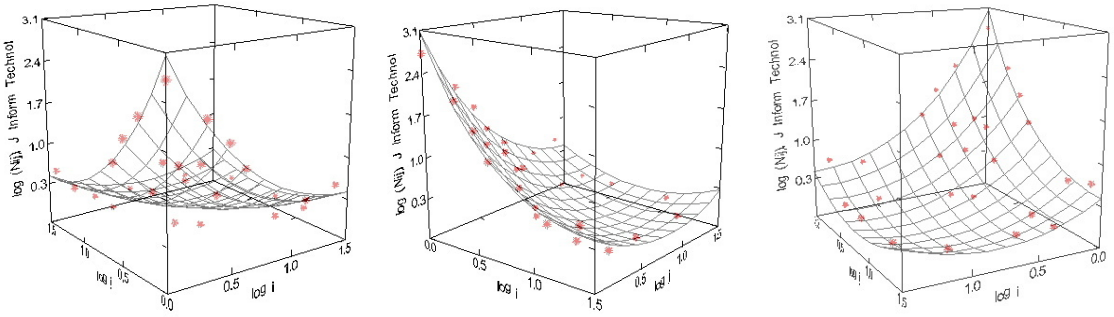

Fig. 3: Gestalt of the *Journal of Information Technology* – Gestalt of Co-author Pairs' Frequencies N_{ij} in Comparison with Distribution of Empirical Data (Triple-logarithmic presentation). The Gestalt is presented by a lattice. The empirical values are red.

Verification of the second Hypothesis

The four pictures in Fig. 4 show the distribution of the empirical data (red colored dots) from the *Journal of Information Technology* (JIT) in comparison with the distribution of the empirical data (blue color) from the *Journal of Biological Chemistry* (JBC). In each of the four pictures both distributions lie on top of each other. The left picture in the first row is rotated three times resulting in the other three pictures.

The widths of the two distributions of the two journals are equal. However, the maximum height of the JBC distribution is higher than the other one, i.e. $Z_{max}(JBC)>Z_{max}(JIT)$.

Further, the proportion of height and width of the distribution of the *Journal of In-formation Technology* has changed in Fig. 4 in comparison with Fig. 3. The same phenomenon is valid regarding the shapes in Fig. 5.

But independently of the different heights, in general the two distributions or the shapes of the social gestalt look different from each other.

The JBC data are taken from Kretschmer, H. & T. Kretschmer 2007: The articles from 1980-1998 were studied with 96,136 authors and the total sum of N_{ij}=779,236 co-author pairs. After regression analysis the correlation between the theoretical gestalt and 469 empirical values of log N_{ij} is equal to R=0.99 with F-ratio=5945.6. The error probability is equal to P=0.000000000 (very high statistical significance).

There are important differences between Fig. 4 and Fig. 5. After regression analyses the lattices in Fig. 5 are produced by the theoretical model of social gestalt. However, the lattices in Fig. 4 emerge independently from this model by a general statistical program used by three-dimensional presentations of empirical data (DWLS: distance-weighted least-squares smoothing). In principle this program can produce a huge number of three-dimensional patterns depending on empirical data. The several shapes of social gestalt are only a tiny part of it. Thus in general, the three-dimensional patterns produced by this program nearly always look very different from a social gestalt.

Therefore, in this connection we can assume the strong similarities between the lattices in Fig. 4 and Fig. 5 say the theoretical model of a social gestalt is an optimum approximation to the empirical data of social networks.

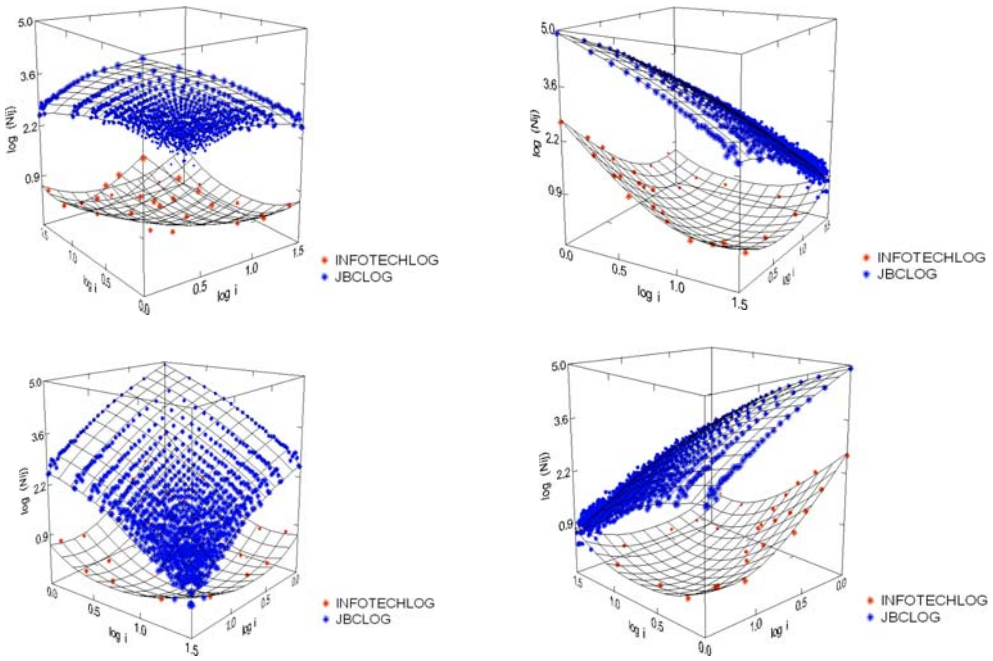

Fig.4: Empirical Patterns' Comparison of the *Journal of Information Technology* with the *Journal of Biological Chemistry*. (log N_{ij} on the Z-axis; log i on the X-Axis and log j on the Y-axis)

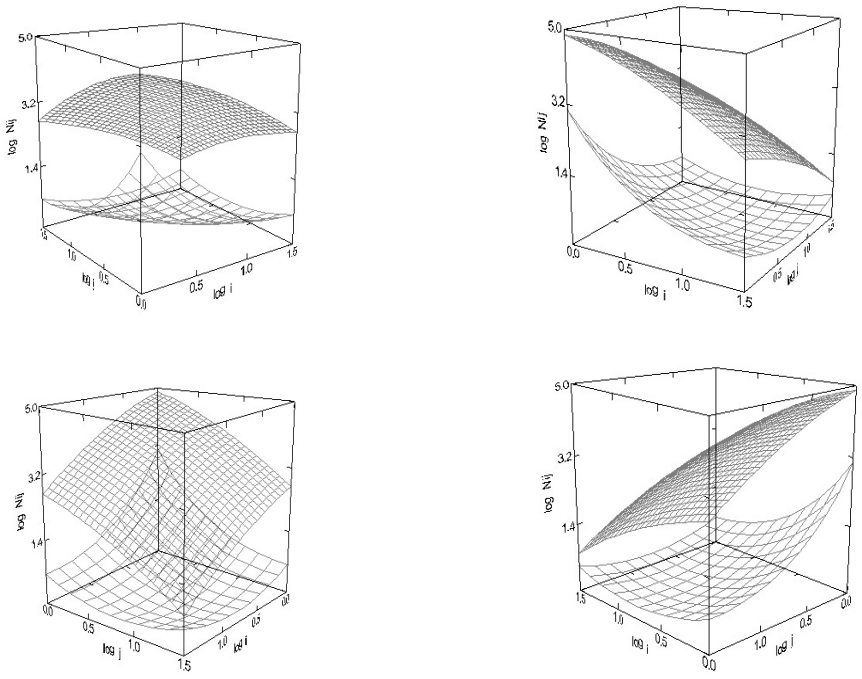

Fig.5: Theoretical Models of Social Gestalts of *Journal of Information Technology* and of *Journal of Biological Chemistry*

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

4 Discussion and Conclusion

We have discovered directly that the *Journal of Information Technology*'s theoretical model of the social gestalt is consistent with its empirical distribution (cf. Fig. 3 and results of the regression analysis). In other words, we have reason to believe that the distribution of co-author pairs' frequencies of the *Journal of Information Technology* can be reflected by a social gestalt, which verifies our first hypothesis above.

And what's more, we also have discovered that the shapes of gestalts as well as the distributions of the empirical data of two journals are different (cf. Figures 4 and 5). The shape of the JBC gestalt is closer to the shape of the gestalt of the journal *Science* (Compare the upper shape in the picture on the right side, first line, of the Figures 4 and 5 with the picture in Fig. 1, second line, right side). There are at least two possible reasons for the explanation of this phenomenon. First, the research fields and contents of the two journals are very different. Second, the sample size of the Journal of *Information Technology* is much smaller than that of *Journal of Biological Chemistry*. As we say in part two, the shape of gestalts of co-authorship networks can be different because of the research environment, method of measurement, research content, sample size, individual characteristics, collaborators' relationships, and so on. And this verifies Kretschmer's opinion that the shapes of gestalts are diverse according to the theoretical model of social gestalts. Simultaneously, it also verifies our second hypothesis above.

This study has shown that Gestalt theory is not only applicable to the study of psychological issues, but it is also applicable to explain the phenomenon of scientific collaboration generated by bibliometric data. In fact, Gestalt theory has been widely applied to a number of disciplines' studies, just as studies in literature, aesthetics, architecture, music and so on.

In any case, we realize that scientific collaboration is a long term issue to be studied with the progress of science and technology, because the development of collaboration depends on scientific fields, age, education, economics, politics, time and space, and so on. And what's more, we selected just one specific journal, to see whether previous results were similarly valid in bibliographies of less scientific journals. In future, investigations of interest will be in the patterns of collaboration visible in books and patents.

Acknowledgement

The authors wish to acknowledge contributions of Prof. Donald deB. Beaver and Dr. Chen Yue, also the other colleagues who gave suggestions to our paper.

This research was supported by the National Natural Science Foundation of China under Grant 70773015, 70431001, 70620140115, National Social Sciences Foundation under Grant 07CTQ008, Project of DUT under Grant DUTHS1002, Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20070141059, projects of League of Social Sciences in Liaoning Province under Grant 2007lslktglx-52.

References

Beaver, D. deB., & Rosen, R. (1978). Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. *Scientometrics*, 1, 65-84.

Beaver, D. deB., & Rosen, R. (1979). Stud ies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite.1799-1830. *Scientometrics*, 1(2), 133-149.

Glänzel,W. (2002). Coauthorship patterns and trends in the sciences(1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, 50, 461-473.

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. & T. Kretschmer. (2006). Well-ordered collaboration structures of co-author pairs in journals. *Proceedings*

of the 2nd National Conference on S&T Policy and Management & International Forum on Science Studies and Scientometrics, September 2006, Dalian University of Technology, Dalian China, 14-30.

Kretschmer, H. & T. (2007). Kretschmer. Lotka's distribution and distribution of co-author pairs' frequencies. *Journal of Informetrics*, 1, 1308-337.

Kretschmer, H. (In Chinese): Configurations in international co-authorship networks[A]. In: Evaluation and its Indicators. Edited by G. Jiang. *Hongqi Publishing House*, Beijing 2000：95-116.

Kretschmer, H. & T. Kretschmer (2007): Lotka's distribution and distribution of co-author pairs. In: *Book of Papers of Third International Conference on Webometrics, Informetrics, Scientometrics and Science and Society & Eighth COLLNET Meeting*, COLLNET 2007, March 6-9, 2007, New Delhi. Divy Srivastava, Ramesh Kundra, Hildrun Kretschmer (Eds.). Sonu Printing Press Pvt. Ltd.: New Delhi, 2007,164-176.

Kretschmer, H. & T. Kretschmer (2007). Distribution of co-author pairs' frequencies of the journal of Biological Chemistry explained as social Gestalt. *COLLNET Journal of Scientometrics and Information Management*.Vol.1, No.2, 21-25

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

Newman M E.J. (2001). The structure of scientific collaboration networks. *Proceedings of National Academy of Sciences of the United States of America*,98,404-409., 2001, 98(2) :404 -409.

Appendix I

The derivation of the specified function regarding X andY is taken from Kretschmer, H. & T. Kretschmer (2007).

Specified function:

There is a conjecture by de Solla Price (1963), physicist and science historian, that the logarithm of the number of publications is of a higher degree of importance than the number of publications per se.

Thus, using the logarithm of the number of publications (log i or log j respectively) as personal characteristic 'productivity', we define:

$X=\log i$
$Y=\log j$
$A = | \log i - \log j |$
$B = \log i + \log j$

Thus:

$A_{min}=|X-Y|_{min}= | \log 1 - \log 1 |= 0$
$A_{max}=|X-Y|_{max}= | (\log i)_{max} - \log 1 |= | \log 1 - (\log j)_{max} |= (\log i)_{max} =(\log j)_{max}$
$B_{min}=(X+Y)_{min}= \log 1 + \log 1 =0$
$B_{max} =(X+Y)_{max}= (\log i)_{max} +(\log j)_{max}=2(\log i_{)max} =2(\log j)_{max}$

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

A_{max}=log 1000=3
B_{max}=2 A_{max}=6
Following:
$A_{COMPLEMENT}$=3- | log i – log j |
$B_{COMPLEMENT}$=6- (log i + log j)

The theoretical mathematical function for describing the social Gestalts of the distribution of co-author pairs' frequencies is resulting:

$$N_{ij}=\text{constant}\cdot(\ |\log i - \log j| +1)^{\alpha}\cdot(4 - |\log i - \log j|\ )^{\beta}\cdot(\log i + \log j+1)^{\gamma}\cdot(7-\log i - \log j)^{\delta}$$

Appendix II

The methods are taken from Kretschmer, H. & T. Kretschmer (2007).
Method for Counting N_i and N_{ij} :

Given is an artificial bibliography including 8 papers (names of authors: A, B, ...)

1. A
2. B
3. D,E
4. D, A, F
5. C
6. G, H
7. H,G
8. H,G,A

The number of publications i (or j respectively) per author P (or Q respectively) is determined by resorting to the "normal count procedure". Each time the name of an author appears, it is counted (e.g. A three times: once in the first paper, and once each in the 4th and 8th papers).

Pairs P,Q are marked in the cells of the matrix under the condition of both the first authors P count (i) and the second authors Q count (j), i.e. the authors are ordered according to i or j respectively in both the row and the column (cf. Table 1).

Under the condition, the place of the authors in the by-line is not taken into consideration the symmetrical matrix is resulting. For example, the pair G,A is marked two times: once under the condition G count (i) and A count (j) and once under the condition A count (i) and G count (j).

Table 1: Symmetrical matrix of the pairs P,Q

i / j		1				2	3			
	P/Q	B	C	E	F	D	A	G	H	Np
1	B									
	C									
	E					1				1
	F					1	1			2
2	D			1	1		1			3
3	A				1	1		1	1	4
	G						1		1	2
	H						1	1		2
									SUM	14

In the symmetrical matrix, one can determine for each author P the number of his collaborators N_P. N_P is equal to the Degree Centrality in Social Network Analysis (SNA).

The matrix of N_{ij} (Table 2, derived from the symmetrical matrix) is the representation of the number of pairs N_{ij} with authors who have i publications per author, with authors who have j publications per author included in the bibliography.

For example, the pairs E,D and F,D in Table 1 are counted both as $N_{12}=2$ and $N_{21}=2$ in the matrix of N_{ij}

Table 2: Matrix of N_{ij}

i / j	1	2	3	N_i	A_i	SUM
1	0	2	1	3	4	
2	2	0	1	3	1	
3	1	1	6	8	3	
N_j	3	3	8			14
A_j	4	1	3			8
SUM				14	8	

$N_i = \Sigma_j N_{ij}$ is the number of collaborators of all authors with i publications per author.
$N_j = \Sigma_i N_{ij}$ is the number of collaborators of all authors with j publications per author.
A_i is the number of authors with i publications per author.
A_j is the number of authors with j publications per author.