A New Centrality Measure for Social Network Analysis Applicable to Bibliometric and Webometric Data

HILDRUN KRETSCHMER[1, 2, 3],
THEO KRETSCHMER[3]

[1]Department of Library and Information Science, Humboldt-University Berlin,
Dorotheenstr. 26, D-10117 Berlin, Germany

[2]The School of Humanities and Social Sciences, Dalian University of Technology, Dalian, 116023, China

[3]COLLNET Center, Hohen Neuendorf, Germany

Abstract

In the literature there are a large number of publications in sociology, in computer science or in information sciences, as well as in studies of collaboration in science describing the studies of social networks with unweighted ties because measures involving unweighted ties are easier to calculate. It is not surprising that there are few studies on networks with weighted ties since they not only need more complex formulas but need a process of quantification when quantitative empirical data are not directly available.

However quantitative empirical data are directly available under the condition of using bibliometric or webometric data.

In conclusion new complex measures of the degree centrality are introduced including weighted ties possible for use of the analysis of co-authorship or citation networks. Both co-authorship relations and citations are well quantified data (weighted ties).

These new measures are applied to a co-authorship network as an example.

1. Introduction

There is a rapid increase of network analysis in several scientific disciplines beginning some decades ago. The social network analysis (SNA) is developed especially in sociology and in social psychology in collaboration with mathematics, statistics and computer science.

Otte and Rousseau [1] showed that social network analysis (SNA) can also be used successfully in the information sciences, as well as in studies of collaboration in science. The authors showed interesting results by the way of an example of the co-authorship network of those scientists who work in the area of social network analysis.

A social network is a set of nodes (social actors) connected by a set of ties. The ties between the nodes can be directed or undirected and weighted or unweighted.

SNA is used to extract patterns of relationships between social actors in order to discover the underlying social structure. Various measures are offered by network analysis. The most used measures are *density* of the network and the *centrality measures*: *degree centrality, betweenness and closeness.*

This paper is focused on *degree centrality*.

Coulon has pointed out [2] in his literature review about the use of social network analysis there is a large number of publications describing the studies of networks with unweighted ties because measures involving unweighted ties are easier to calculate. According to Coulon's opinion it is not surprising that there are few studies on networks with weighted ties since they not only need more complex formulas but need a process of quantification when quantitative empirical data are not directly available.

However quantitative empirical data are directly available under the condition of using bibliometric or webometric data.

In conclusion a ***new complex measure of degree centrality*** is introduced including weighted

ties suitable for analyzing co-authorship-, citation- or Web networks. Co-authorship relations, citations, Web visibility rates or hyperlinks are well quantified data (weighted ties).

In this paper the new measure is applied to a co-authorship network as an example.

2. Presentation of the Original Measure for Degree Centrality

The nodes of a social network can be individuals, teams, institutions, etc. The relationships (ties) between the nodes can also be of many kinds for example, friendship, business, economic, etc. In this paper we are looking at scientists as nodes and at co-authorship relations, citations or hyperlinks as ties.

The original used measures of social network analysis (SNA) are related to Wassermann & Faust [3]:

- *The Degree Centrality* (DC_A) of a node A is equal to the number of nodes (or ties) to which this node is connected.

For example, in collaboration networks in science the Degree Centrality of a node A is equal to the number of his/her collaborators or co-authors. An actor (node) with a high degree centrality is active in collaboration. He/she has collaborated with many scientists.

- In correspondence with Wassermann and Faust the ***Group Degree Centralization*** quantifies the variability or dispersion of the individual Degree Centralities of the nodes. Centralization describes the extent to which the links (ties) are organized around particular focal nodes, i.e. it provides a measure on the extent to which a whole network has a centralized structure.

There are several degree based measures of graph centralization. One of them is as follows:

$$GDC = \frac{\sum_{i=1}^{v} (DC_L - DC_i)}{(v-1)(v-2)} \qquad (1)$$

The DC_i in the numerator are the v Degree Centralities of the nodes and DC_L is the largest observed value.

This index reaches its maximum value of 1 when one actor (node) has collaborated with all other v-1 actor, and the other actors interact only with this one, central actor. This is exactly the case in a star graph. The index attains its minimum value of 0 when all degrees are equal.

3. Comparison of Weighted and Unweighted Degree Centrality Measures Explained on the Basis of Co-authorship Networks

3.1. General Remarks

Using the ***unweighted measure*** means the ties (or nodes) are counted independently from the strength of the ties.

However in analyzing bibliometric or webometric networks several sorts of methods are developed ***to measure the strength of a tie*** between a pair of nodes A and B.

Application of similarity coefficients is one of them. While many similarity coefficients were proposed in various research fields, Salton's measure or Jaccard index were most used in scientometric studies (Glänzel [4] , Miquel and Y. Okubo [5] , Katz [6]). Zitt et al. [7] introduced a probabilistic indicator to measure strength of scientific linkages between partners. Yamashita and Okubo [8] have presented a new probabilistic partnership index (PPI).

However in this paper we don't yet consider the above mentioned kinds of measuring the strength of ties. Compared with this we are looking for a co-authorship relation or for a citation, etc., as the ***basic unit of links***:

- The ***strength of a tie between a pair of nodes A and B_i*** can be measured by the number of basic units which exists between A and B_i:

  U_{ABi}

- The ***total strength of all of the ties between a node A and all of the nodes B_i(i=1,2...z) to which this node A is connected*** is equal to the sum of the strengths of these ties:

  $$TR_A=\Sigma_i U_{ABi} \quad (2)$$

- The ***total strength T of all of the ties in a network with v nodes X_j*** is equal to the total sum of TR_{Xj} divided by 2:

  $$T= (\Sigma^{v}_{j=1} TR_{Xj})/2 \quad (3)$$

Let us compare weighted and unweighted degree centrality measures under the following conditions:

- First condition: DC_A=const, TR_A is changing, $U_{AB1} = U_{AB2}, =...=U_{ABz}$
- Second condition: DC_A is changing, TR_A =constant, $U_{AB1} = U_{AB2}, =...=U_{ABz}$
- Third condition: DC_A =constant, TR_A =constant, $U_{ABi} \leq U_{ABj}$ or $U_{ABi} \geq U_{ABj}$

3.2. First Condition

Regarding the variation of the strengths of the ties let us have a view at the following example:

Using the unweighted measure of the degree centrality says the degree centralities of the scientists E and F are equal in both networks (Fig. 1, left side and right side) although on the right side the strengths of the scientist E's ties are several times higher than on the left.

Fig. 1: Comparison between two networks
First condition: DC_E=const, TR_E is changing, $U_{EA}=U_{EB}=U_{EC}=U_{ED}$

What does it mean in co-authorship networks? At first glance scientist E is more centralized in the right side network than in the other network. Additionally, let us take into consideration the theoretical background.

Co-authored research papers are assumed to signal research cooperation and associated knowledge flows and exchanges Calero, van Leeuven & Tijssen [9] . In continuation we assume the knowledge flow between a pair of collaborators A and B is increasing with increasing number of co-authorship relations (strength of the tie). The number of co-authorship relations between a pair of nodes A and B is equal to the number of their joint multi-authored papers.

Analogous considerations can be made in citation or Web networks.

Because of these considerations the centrality of a scientist A is increasing with both increasing number of collaborators (degree centrality) and increasing total number of co-authorship relations with these collaborators. This condition is not fulfilled using the original degree centrality.

Whereas citations or hyperlinks are well defined in our field there maybe differences regarding the term "co-authorship relation". Thus, explanation is necessary as follows:

Counting the total number of co-authorship relations (TR_A) of an author A:

Given one multi-authored paper p_{Ai} of the scientist A with m_{Ai} co-authors then the number of co-authorship relations of A is equal to m_{Ai}.

Let us assume the number of multi-authored papers of the scientist A is equal to z. Following the total number of co-authorship relations of A is equal to the sum of the co-authorship relations of the z multi-authored papers:

$$TR_A=\Sigma^{z}_{i=1} m_{Ai} \quad (4)$$

Whereas the number of co-authorship relations between a pair A and B_i (U_{ABi})

is equal to the number of joint multi-authored papers the total number of co-authorship relations of the author A in collaboration with more than one co-author (TR_A) can be either equal or higher than the total number of multi-authored papers.

3.3. Second Condition

In social networks usually the number of actors to which an actor A is connected can vary ***independently*** from the total strength of the ties.

For example in Fig. 1, the number of collaborators of E or F is constant but the total strength of the ties (total number of co-authorship relations) is different.

Vice versa, in another network the authors G and J can have the same total number of co-authorship relations but the number of collaborators is different. In Fig. 2 the total number of co-authorship relations of G (equal to 8) is spread out over 2 collaborators but the same total number of co-authorship relations of J (equal to 8) is spread out over 4 collaborators.

Fig. 2: Comparison between two networks
Second condition: DC_X is changing, TR_X =constant, U_{XY1} =U_{XY2}, =...=U_{XYz} with X=G=J, Y_i are all of the nodes Y_i(i=1,2...z) to which the special X is connected

General Stipulation for a weighted degree centrality measure of an actor:

If two variables can vary independently of each other, the following condition has to be fulfilled: If one variable remains constant and the other variable assumes a higher value, then a weighted degree centrality measure must assume a higher value. This requirement will be met by the geometric mean of the magnitudes of the two variables:

A weighted degree centrality measure of a node A is equal to the geometric mean of the number of nodes to which this node is connected and the total strength of the ties.

Whereas the original Degree Centralities of the scientists E and F are equal in both networks in Fig. 1 a weighted degree centrality measure of E is higher in the network on the right side of Fig. 1. However in Fig. 2 both the Degree Centrality and a weighted degree centrality measure of J are higher than both corresponding values of G because the total sum of co-authorship relations remains constant.

3.4. Third Condition

In the examples in Fig. 1 and in Fig. 2 we have counted the number of collaborators of a scientist A on the basis of equal strengths of the ties between the pairs of collaborators. However how to measure the number of collaborators on the basis of unequally weighted ties?

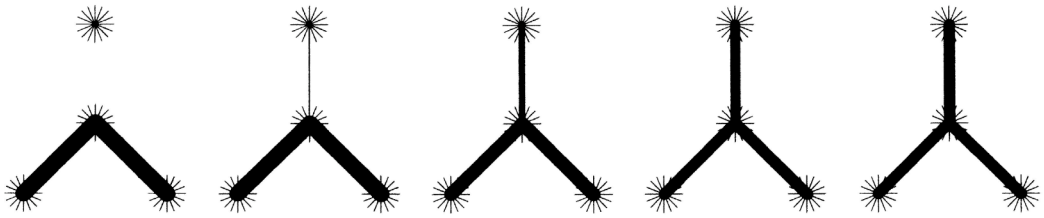

Fig. 3: Comparison between five networks with four nodes per network and node A in the middle of each of the networks

Third condition: DC_A=constant, TR_A=constant, $U_{ABi} \leq U_{ABj}$ or $U_{ABi} \geq U_{ABj}$

There are five networks in Fig. 3 with four nodes per network. The first network on the left side shows the number of collaborators of the scientist A (node in the middle of the four nodes) is equal to 2 but the last network on the right side says it is equal to 3. Moving our eyes from left to right at first glance we have the impression there is a continuous change of the "number" of collaborators from 2 to 3, i.e. in the middle networks the "number" of collaborators of A is between 2 and 3.

In Fig.3 is valid: The total sum of co-authorship relations remains constant however the spread over the (possible) collaborators is changing.

How to measure the "number" of collaborators under these kinds of conditions?

In Table 1 the total number of co-authorship relations (TR_X) is equal for all of the scientists in the column (A, B, C or D) in co-authorship with the scientists in the row (E or F). However the strengths of the ties between the pairs of collaborators (values in the cells of the matrix) are different.

The number of collaborators of A: (DC_A) is clearly 1 and the number of collaborators of D: (DC_D) is also clearly 2. However, if you ask B for the number of his collaborators, he will possibly give the answer „one" since the number of co-authorship relations between him and the collaborator E (equal to1) in relation to the number of co-authorship relations between him and collaborator F (equal to 99), is so small that this fact could be neglected. By contrast, C might say that the number of his collaborators is 2 because the co-authorship relations are almost equally distributed.

Table 1: Unequally weighted ties between pairs of collaborators

X/Y	E	F	TR_X	DC_X	$2^{H(Ki)}$
A		100	100	1	1
B	1	99	100	>1≈1	1.06
C	49	51	100	<2≈2	1.99
D	50	50	100	2	2

Note: The calculation of $2^{H(Ki)}$ can be found below.

Conclusions:

The idea is to search for a function that upon an equal distribution of weights on the elements to be counted is equal to the "number" of elements. The greater the deviation is from this equal distribution, the smaller shall be the value of the function.

In the example the elements to be counted were the collaborators E or F and the weights (or strength of ties) were the number of co-authorship relations existing between the corresponding scientists A, B, C, or D and each of these collaborators.

According to the definition for the sought-after function the „number" of the collaborators of C is to be smaller than 2 but almost 2. The „number" of collaborators of B is to be greater than 1, but almost 1. A function that meets those requirements is the transformed entropy 2^H. The results are shown in Table 1. The calculation is shown below.

Calculation of the entropy H(K_i*):*

There is a series of numbers $K_i (i=1,2,\ldots z)$, $K_i \neq 0$

$$h_i = K_i / \Sigma^z_{i=1} K_i \quad (5)$$

$$H(K_i) = - \Sigma^z_{i=1} \; h_i \cdot lg_2 h_i \quad (6)$$

Stipulation: If $\Sigma^z_{i=1} K_i = 0$ then $2^{H(Ki)}=0$ (7)

In calculation of the "number" of collaborators of the scientist A the term K_i is equal to the strength of a tie between a pair of the nodes A and B_i:

$$K_i = U_{ABi} \quad (8)$$

The "number" of collaborators of the scientist A is equal to $2^{H(Ki)}$:

$$\text{"DC"}_A = 2^{H(Ki)} \quad (9)$$

4. *The Complex Measure of Degree Centrality (CDC_A) of a Node A*

The considerations in paragraph 3 are resulting in the following definition for the Complex Degree Centrality CDC_A measure of a node A:

The Complex Measure of Degree Centrality CDC_A of a node A is equal to the geometric mean of the "number" of nodes to which this node is connected and the total strength of the ties.

Therefore, for the analysis of networks based on co-authorship data the Complex Degree Centrality of a scientist A is defined as:

CDC_A= ("Number" of collaborators of scientist A · Total sum of co-authorship relations of the scientist A)$^{1/2}$

$$CDC_A = (\text{"DC"}_A \cdot TR_A)^{1/2} \quad (10)$$

As mentioned above Group Degree Centralization quantifies the variability or dispersion of the individual Degree Centralities of the nodes. One of the several degree based measures of graph centralization was described in formula 1.

The Complex Centralization describes the extent to which the links (ties) are organized around both particular focal nodes and ties, i.e. it provides a measure on the extent to which a whole network has a centralized structure.

Using the Complex Degree Centrality a new ***Complex Group Degree Centralization*** measure is proposed describing the dispersion in a connected graph, i.e., there is a "path" between each pair of nodes X_i and X_j:

$$GCDC= 1 - (H(CDC_{Xi}) – H_{min})/(H_{max}-H_{min}) \quad (11)$$

with $H_{max}=lg_2\ v$

Stipulation for calculation of the minimum Entropy H_{min}:

Given a network with v nodes X_i and with the total strength T of all of the ties in this network:

- The total strength of all of the ties between the node X_1 and all of the other nodes should be equal to T:

  $TR_{X1}=T$

- The strength of the tie between X_1 and X_2 should be equal to T-(v-2):

  $U_{X1X2}=T-(v-2)$

- The strength of a tie between X_1 and X_i with i (i=3,4,..v) is equal to 1:

  $U_{X1,Xi}=1$

GCDC reaches its maximum value of 1 when one actor (node) has collaborated with all other v-1 actors, and the other actors interact only with this one, central actor and when there is a maximum strength of one of the ties. This is exactly the case in a star graph with a maximum strength of the tie between the central node and one of the other nodes. The index attains its minimum value of 0 when all Complex Degree Centrality degrees are equal.

The Complex Measure of Degree Centrality is applied to a co-authorship network as an example. The new measure is compared with the original.

5. Data

COLLNET is a global interdisciplinary research network under the title: “Collaboration in science and in technology” (www.collnet.de). In a former study (Kretschmer & Aguillo [10]) a request was made to all the 64 COLLNET members for their complete bibliographies, independently of the type of the publications and independently from the date of appearance of these publications.

From these bibliographies all publications were selected that appeared in co-authorship between at least two COLLNET members. Thus, it concerns 223 bibliographic multi-authored publications. From this, the respective number of common publications between two members was determined as the basis for the analysis of the co-authorship network.

The co-authorship network developed according to this method covers the entire lifetime collaboration between the COLLNET members.

The last COLLNET data are from June 2003. These data are used in this paper, too.

6. Results and Conclusion

The structure of the co-authorship network obtained from the bibliographies is shown in Fig. 4 with unweighted ties.

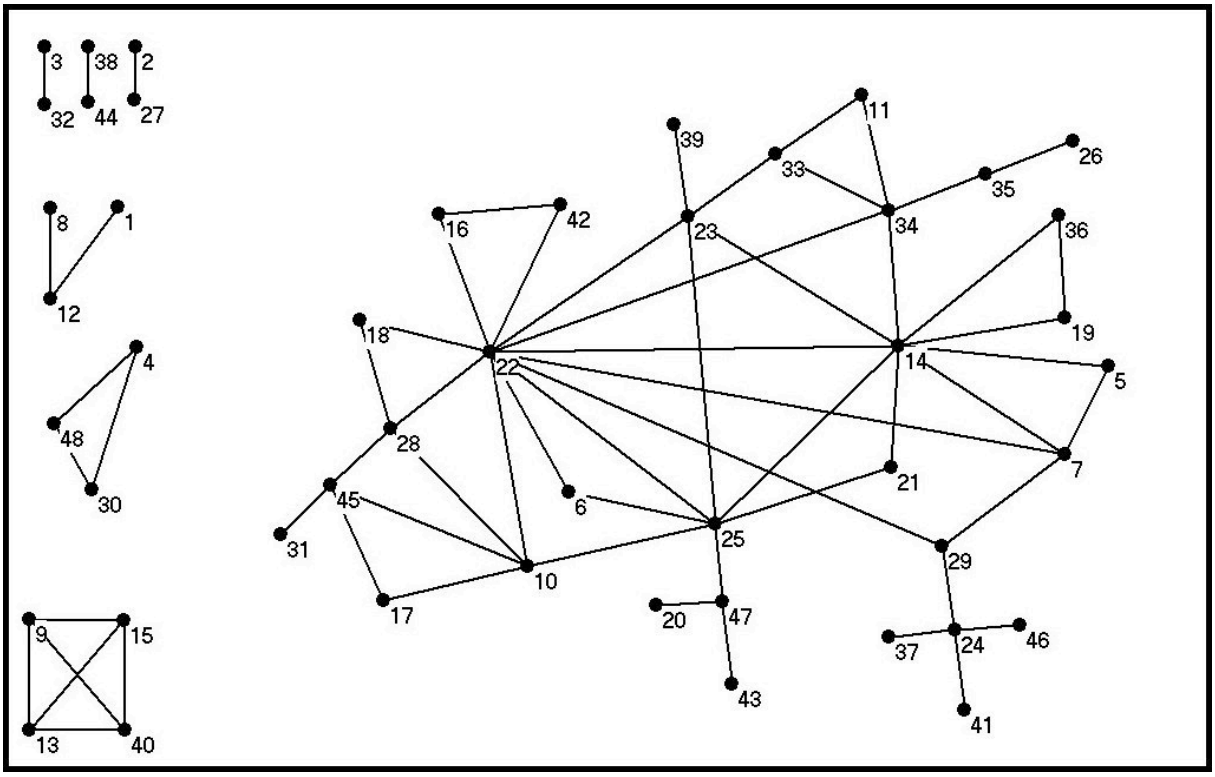

Fig. 4: Co-authorship network of the COLLNET members (Kretschmer & Aguillo, 2004)

Legend (Fig. 4):

1. Isidro Aguillo	2. Petra Ahrweiler	3. R. Ambuja	4. Bassecoulard
5. Aparna Basu	6. Donald deB. Beaver	7. Sujit Bhattacharya	8. Maria Bordons
9. Martina Brandt	10. Mari Davis	11. Leo C.J. Egghe	12. Isabel Gomez
13. Ulla Grosse	14. Brij Mohan Gupta	15. Frank Hartmann	16. Frank Havemann
17. William W. Hood	18. Margriet Jansz	19. Karisiddappa	20. Sylvan Katz
21. Ved P. Kharbanda	22. Hildrun Kretschmer	23. Ramesh Kundra	24. Loet Leydesdorff
25. Liming Liang	26. Sofía Liberman	27. Zeyuan Liu	28. Valentina Markusova
29. Martin Meyer	30. Yoshiko Okubo	31. Farideh Osareh	32. Koti S. Raghavan
33. Ravichandra Rao	34. Ronald Rousseau	35. Jane Russell	36. Shivappa Sangam
37. Andrea Scharnhorst	38. Annedore Schulze	39. Dimiter Tomov	40. Rainer Voss
41. Caroline Wagner	42. R. Wagner-Döbler	43. Yan Wang	44. Vera Wenzel
45. C. S. Wilson	46. Paul Wouters	47. Yishan Wu	48. Michel Zitt

49.-64. are singletons up to June 2003. The 16 singletons are not included in the figure.

For demonstration of the differences between the original and the complex measures of degree centrality we have extracted two scientists with the same Degree Centrality from Fig. 4: $DC_{24}= DC_{45} = 4$. The node 24 can be found in the middle of the left network of Fig. 5 and the node 45 in the middle of the right network.

The strengths of the ties of the node 24 are as follows:

$U_{24;29}=1$, $U_{24;37}=1$, $U_{24;41}=2$, $U_{24;46}=5$ with the total strength $TR_{24}=9$

The strengths of the ties of the node 45 are:

$U_{45;28}=5$, $U_{45;31}=6$, $U_{45;17}=20$, $U_{45;10}=10$ with the total strength $TR_{45}=41$

The "number" of collaborators "DC"$_{24}$=3.155 is rather similar as the "number" of collaborators "DC"$_{45}$=3.429.

However the Complex Degree Centrality of node 45 is higher than the Complex Degree Centrality of node 24 because of the higher total strength:

$CDC_{24} = (\text{"DC"}_A \cdot 9)^{1/2} = (3.155 \cdot 9)^{1/2} =$ **5.33**

$CDC_{45} = (\text{"DC"}_A \cdot 41)^{1/2} = (3.43 \cdot 41)^{1/2} =$ **11.86**

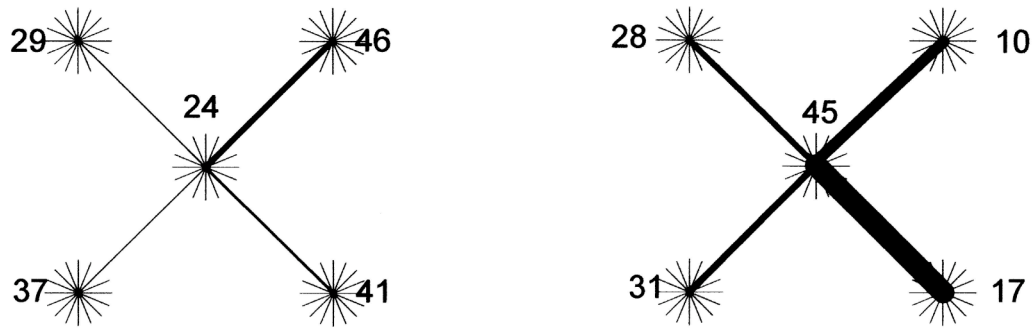

Fig. 5: Comparison between two scientists with equal Degree Centrality

For another comparison we have selected 3 triads for calculation of the degree centralities of the nodes and the group degree centralization (Fig. 6). The original centrality measures say the degree centralities are equal for all of the nodes and the group degree centralization is equal for all of the three triads.

At first glance the centralization of the right triad is higher (with node 48 in the center) than the centralization of the other triads. Indeed the Complex Degree Centrality of the node 48 is higher than the original measure and the same with the Complex Group Degree Centralization. In detail:

The strengths of the ties are as follows:

- left triad: $U_{5;14}=U_{14;7}=U_{5;7}=1$
- triad in the middle: $U_{15;9}=5$, $U_{9;40}=5$, $U_{40;15}=4$,
- right triad: $U_{48;30}=5$, $U_{30;4}=2$, $U_{4;28}=28$

The Complex Degree Centralities of the nodes are:

- left triad: $CDC_5=2$, $CDC_7=2$, $CDC_{14}=2$
  with the **Complex Group Degree Centralization GCDC=0**
- triad in the middle: $CDC_{15}=4.23$, $CDC_9=4.47$, $CDC_{40}=4.23$
  with the **Complex Group Degree Centralization GCDC=0.00255**
- right triad: $CDC_{48}=7.11$, $CDC_{30}=3.567$, $CDC_4=6.19$
  with the **Complex Group Degree Centralization GCDC=0.20027**

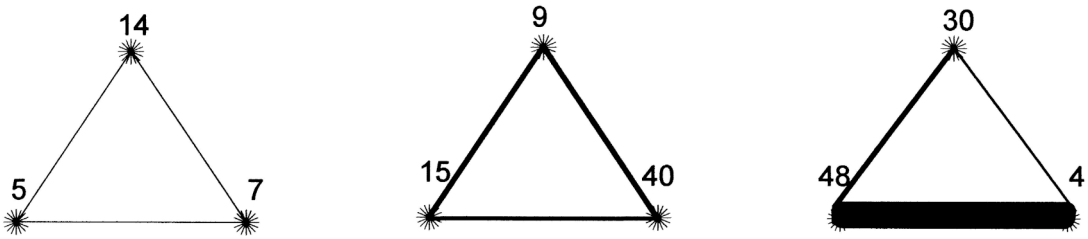

Fig. 6: Selection of three triads

After demonstration of these new introduced Complex Degree Centrality Measures in this pilot study this new method should be empirically tested in future analyses of bibliographic citation or co-authorship networks as well as in analyses of Web networks.

References

1. E. Otte & R. Rousseau, Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28, 443-455, 2002
2. F. Coulon, The use of social network analysis in innovation research: a literature revie, 2005.

   http://www.druid.dk/conferences/winter2005/papers/dw2005-305.pdf (The Web search took place on January 1, 2006)
3. S. Wasserman & K. Faust, Social network analysis. Methods and applications.

   Cambridge: Cambridge University Press, 1994
4. W. Glänzel & A. Schubert, Analyzing scientific networks through co-authorship. In H.F. Moed et. al. (Eds.), *Handbook of QuantitativeScience and Technology Research,* (pp.257-276), The Netherlands: Kluwer Academic Publishers, 2004
5. J.F. Miquel, & Y. Okubo, Structure of International Collaboration in Science-Part II: Comparisons of Profiles in Countries using a Link Indicator, *Scientometrics* 29. No.2, 271-297, 1994
6. J.S. Katz, Geographical proximity and scientific collaboration, *Scientometrics* 31, 31-34, 1994
7. M. Zitt, E. Bassecoulard & Y. Okubo, Shadows of the Past in International Cooperation: Collaboration profiles of the top five producers of science, *Scientometrics*. 47, No. 3, 627-657, 2000
8. Yamashita & Y. Okubo, Patterns of scientific collaboration between Japan and France: Inter-sectoral analysis using probabilistic Partnership Index (PPI). In: Peter Ingwersen & Birger Larsen (Eds.). Proceedings of the 10th ISSI International Conference on Scientometrics and Informetrics, July 24-28, 2005, Stockholm, Sweden, Volume 2. Published by Karolinska University Press: Stockholm, 2005, 517-526, 2005
9. C. Calero, T.N. van Leeuven & J.W. Tijssen (2005). Reseaarch networks of pharmaceutical firms: geographiccal patterns of research collaboration within and between firms. In: Peter Ingwersen & Birger Larsen (Eds.). Proceedings of the 10th ISSI International Conference on Scientometrics and Informetrics, July 24-28, 2005, Stockholm, Sweden, Volume 1. Published by Karolinska University Press: Stockholm, 2005, 310-315, 2005
10. Kretschmer & I. Aguillo, Visibility of collaboration on the Web. *Scientometrics*. Vol. 61, No. 3, 405-426, 2004