



Institut für Forschungsinformation und Qualitätssicherung

On the basis of bibliometric coauthorship analysis

Peder Olesen Larsen and Markus von Ins



Institute for Research Information and Quality Assurance

“... It is also interesting to speculate what the effect would be of extending the concept of fractional citation counting to fractional co-citation counting, that is, converting the present integer co-citation approach into a fractional one. By analogy, fractional co-citation counting would assign a single unit of co-citing strength to each citing paper and apportion that unit equally among all the pairs of references cited by that paper. If, for example, a paper cites “n” highly cited items, each pair of cited items would be assigned a weighted co-citation equal to $1/[1/2n(n-1)]$”

Small and Sweeney 1985

Two different networks (Gauffriau et al. 2007)

the underlying network (UN) of actors

the cumulative turnout network (CTN)









The case of bibliometric co-authorship networks

“ ... The purposes for which people collaborate

1 Access to expertise.

2 Access to equipment, resources, or “stuff” one doesn’t have.

3 Improve access to funds.

4 To obtain prestige or visibility; for professional advancement.

5 Efficiency: multiplies hands and minds; easier to learn the tacit knowledge that goes with a technique.

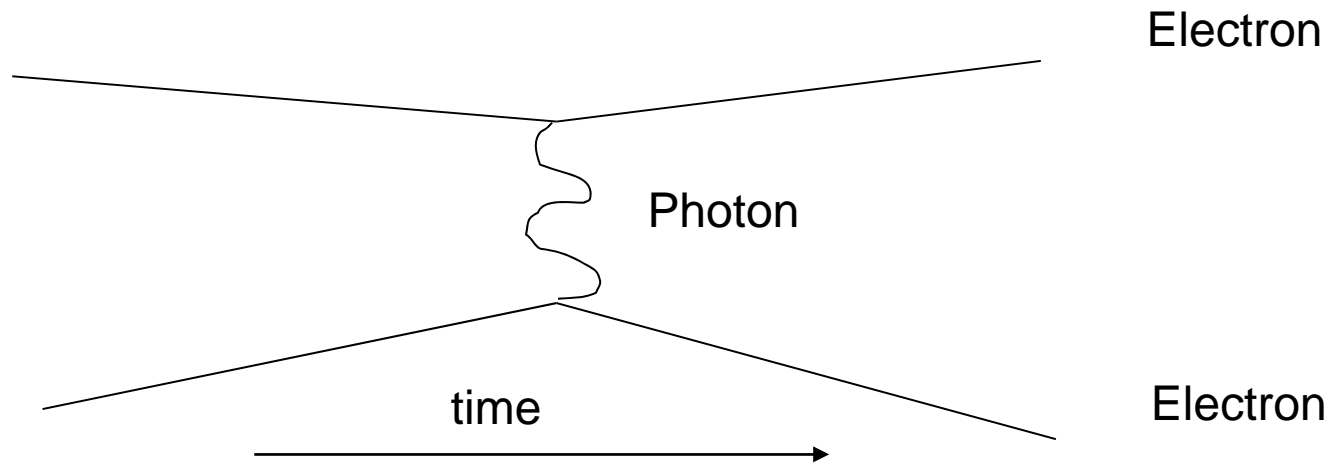
6 To make progress more rapidly.

...”

Beaver 2001 Reflections on scientific collaboration

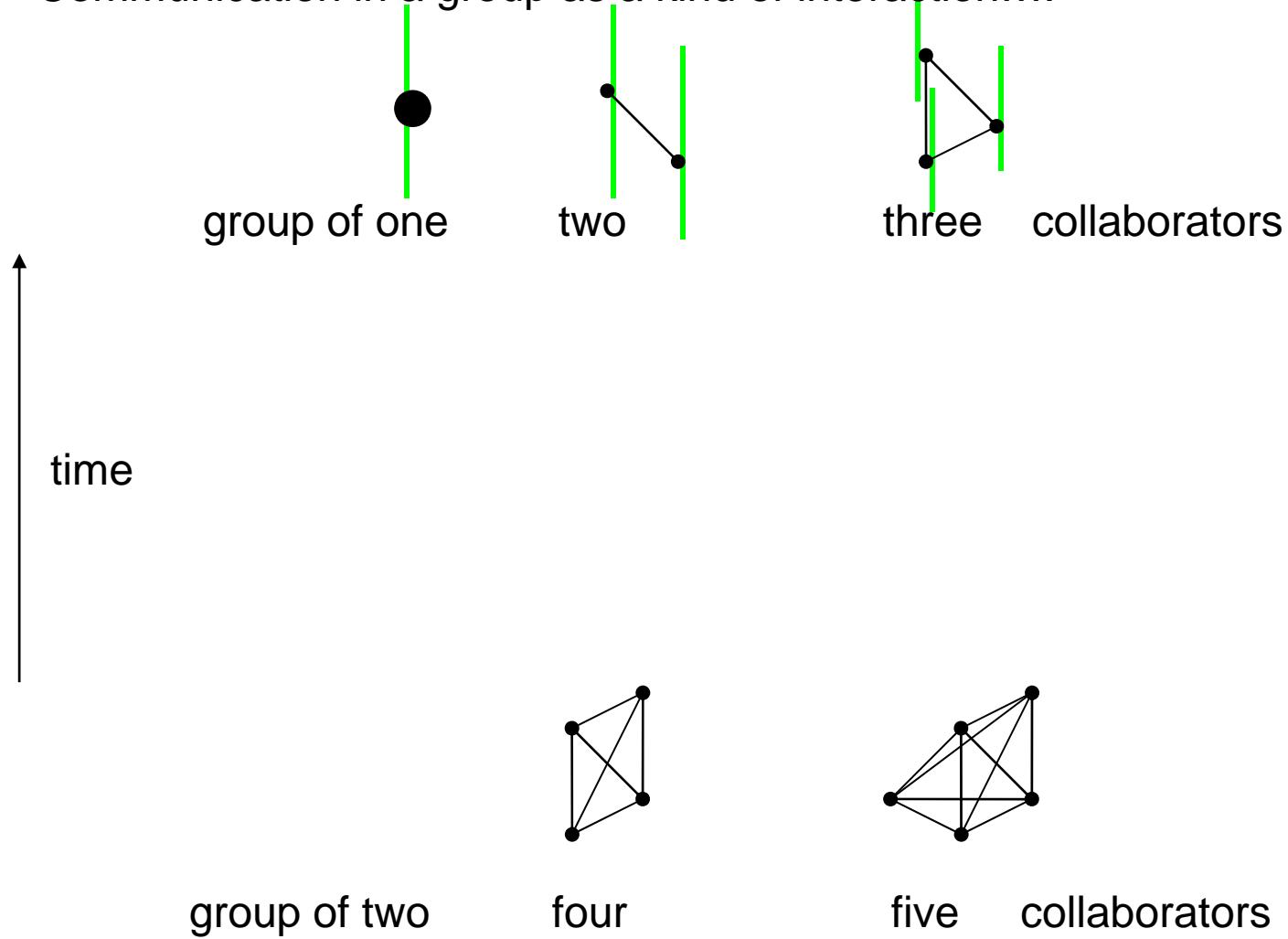


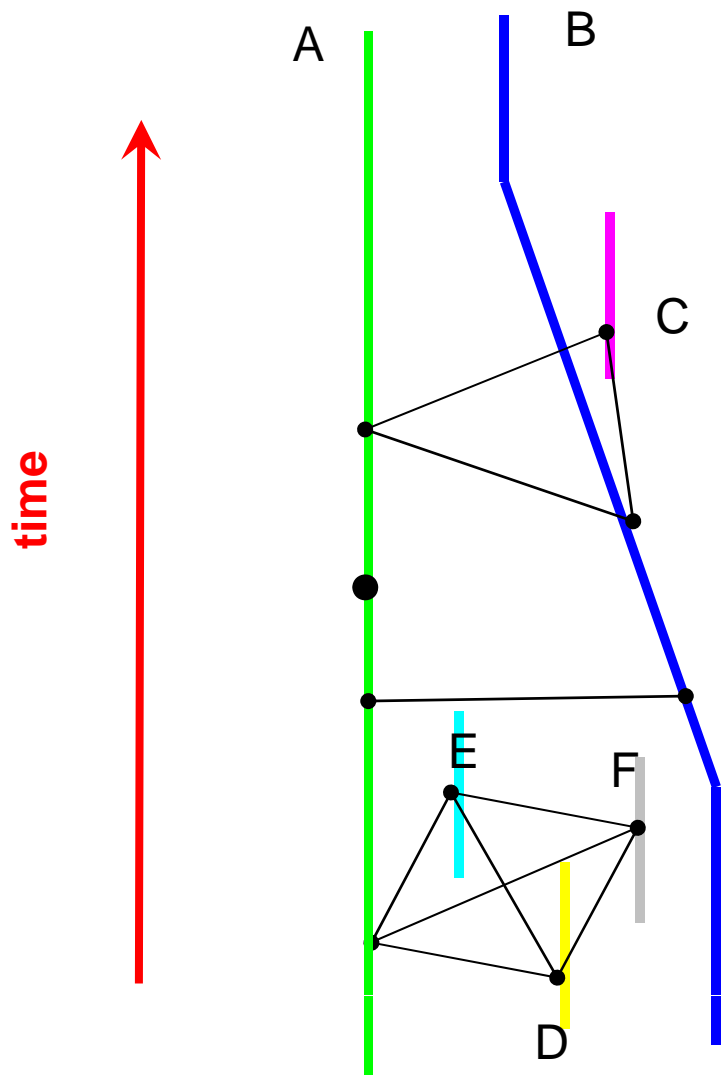
Communication in a group as a kind of interaction....



... the classical exchange interactions are **two**-particle interactions

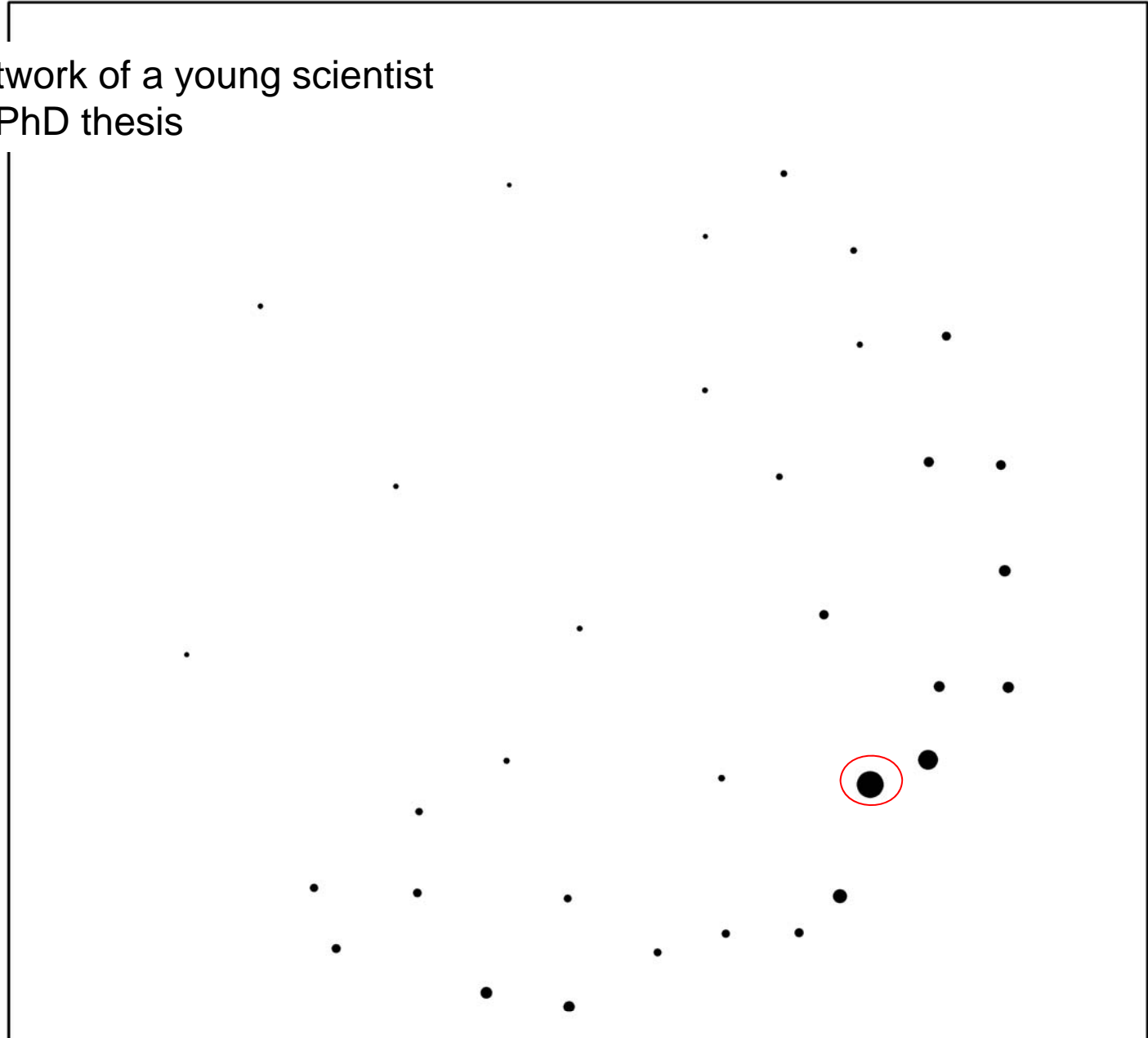
Communication in a group as a kind of interaction....



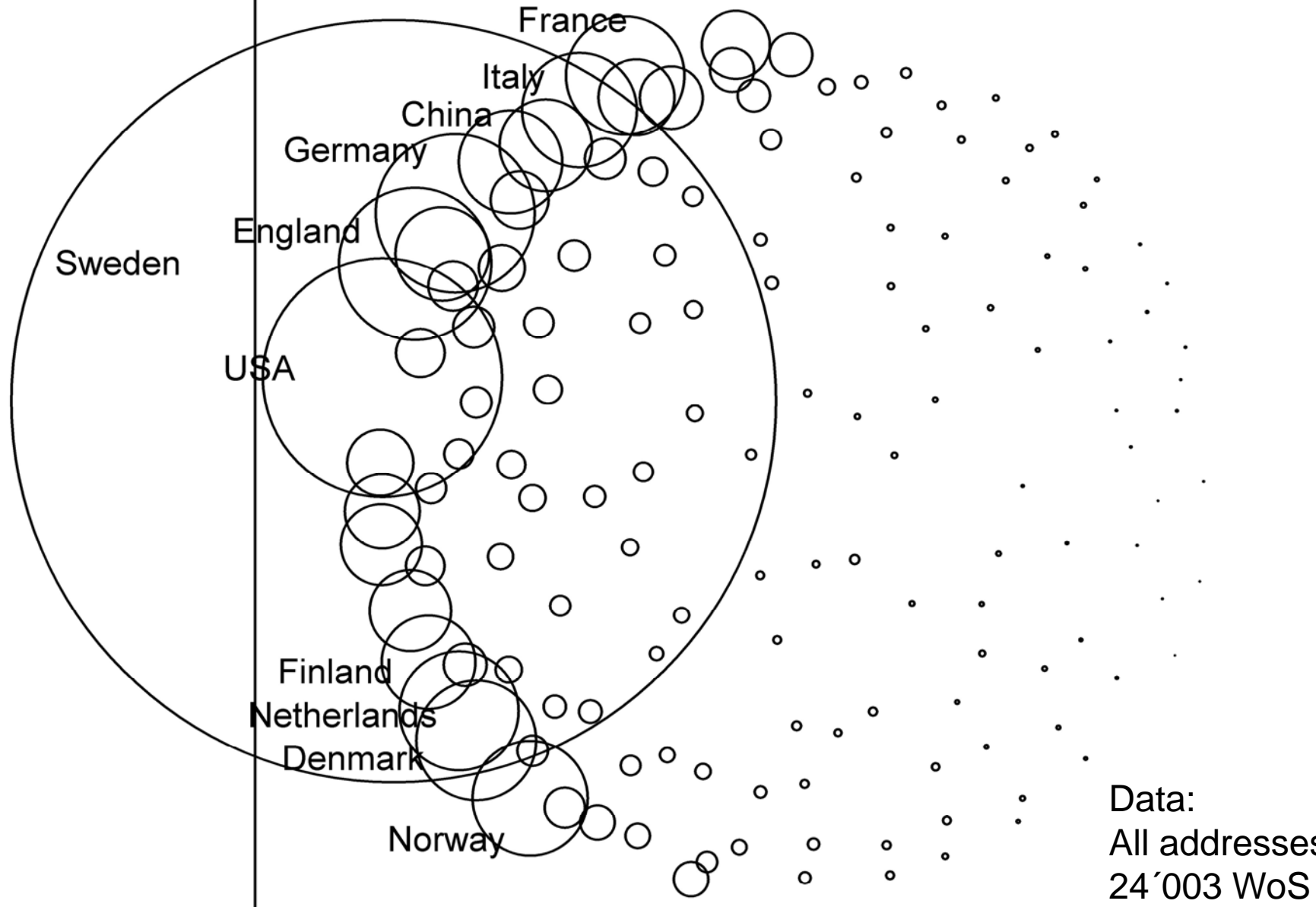


	A	B	C	D	E	F	degree centrality
A	1	0.67	0.17	0.08	0.08	0.08	2.08
B	0.67		0.17				0.83
C	0.17	0.17					0.33
D	0.08				0.08	0.08	0.25
E	0.08			0.08		0.08	0.25
F	0.08			0.08	0.08		0.25
							4

Co-author network of a young scientist 4 years after PhD thesis



The research network of Sweden 2008



Data:
All addresses in
24'003 WoS papers
with at least one
Swedish address

Thank you for your attention.

On the basis of bibliometric coauthorship analysis

Peder Olesen Larsen and Markus von Ins

Title-Slide 1

In this presentation we would like to present you some recent work on the description of the cooperative structure of science and research.

Slide 2

This research began with a problem of normalization in co-citation clustering and bibliographic coupling processes pointed out by first Small and Sweeney in 1985. They ask, what would happen, if one would change the counting method in co-citation clustering.

Later in 2007 we showed in Gauffriau et al. that counting methods cannot be arbitrarily changed, but must be chosen according to the problem addressed and that the results of counting methods depend strongly on the cooperation or co-occurrence of the basic units addresses and author names. Therefore, it was near to ask, in what way the concepts of counting methods could be extended to cooperation and other co-occurrence structures.

Slide 3

In Gauffriau et al. 2007 we distinguished two networks arising from cooperative research.

Commonly known from social network analysis is the underlying network UN of actors as for examples authors, institution or countries .

But in bibliometric studies this underlying network is a priori not known nor visible – rather what we find in scholarly publications are lists of authors and their institutional addresses producing results or turnouts. Therefore, what we recognize in bibliometric studies is a kind of resulting network or cumulative turnout network CTN.

This distinction leads us to reconstruct the underlying network out of its turnouts in the publications of the network.

In order to find, what we shall reconstruct and what the problems are, we first consider an example of a traffic network with a known underlying network.

Slide 4

In this example we find a schematic map of the underlying railway network of stations and lines or ties in Stockholm. This underlying railway network has the following properties:

a) it does virtually not change in time and therefore, we will call this type of network a static network

b) the corresponding cumulative turnout network is given in a database called the time-table

c) In the schematic map we cannot recognize the true distances , rather we have to reconstruct these distances from for example the time-table

d) As indicated by its name, the central station is more central than the other ones; Therefore our task is to reconstruct the notion of centrality and its interpretation.

Slide 5

The notion of centrality and the different degrees of centrality or “betweenness” were studied by Wassermann and Faust in 1994.

First we recognize the red circles for the line-end-stations with lowest centrality and only one tie.

Slide 6

In blue circles a few examples of nodes or stations with two ties are given and in a yellow circle we find the only station with three ties.

More towards the central station the number of ties increases and reaches a maximal centrality of the central station with 14 ties – four blue ties, four red ties and six green ties.

In this example we recognize two further properties of centrality. The more central nodes or stations show of course a higher train frequency and are also more *important*.

Slide 7

The importance of a node can be recognized when it is omitted from the network. If for example the end-station of T11 "Akalla" in the red circle is omitted, a walk to the next station solves the problem – the network stays connected and virtually intact.

But if there is an outage in the central station in the green circle, then there is a major traffic problem, because the network decays into four disconnected pieces. This degree of importance is often measured with the betweenness centrality or degree centrality according to Wassermann and Faust 1994.

In conclusion our task is twofold: We have to reconstruct the distances of the nodes or stations and we have to reconstruct their centrality and importance.

Slide 8

Next we consider the situation in bibliometric co-authorship networks.

As already pointed out by Price in 1961 and 1963 co-authorship is a very recent phenomenon in science and research close to absent before the advent of the twentieth century.

Compared with the history of science of more than 2500 years, the history of cooperative work in research is very short. Therefore, a description by a static network would of course fail to describe the research activities before 1900.

Price 1963 also argues, that publications as well as collaborations in his "invisible colleges" could be due to communication processes. In this rationale the author list represents the "inner" communication networks and the publications form communications to the outside or the scientific community outside the group.

Nonetheless, this rationale cannot explain the reasons for the co-operations or invisible colleges.

Slide 9

These reasons or purposes of collaboration were later studied by Donald de Beaver 2001 in a study resulting in 18 purposes. The first six reasons are copied to the slide and show, that a static description would fail: Tasks and goals as indicated can only be described in dynamic pictures and theories.

Because the tasks are different in different research projects, an adequate network must be chosen according to the projects and ergo will change all the time.

Slide 10

If we once more consider the communication rationale by Price, we have to formulate this as an interactive theory.

In theoretical physics we find of course a lot of such interactive dynamic theories as for example Feynman's famous Quantum Electrodynamics with interactions of charged particles by exchange of photons. One such example is copied to the slide.

The dynamics is indicated by the time-scale.

The figure shows, that such exchange theories can only describe *two*-particle or *two*-body interactions.

But in contrast to the well known interactive theories in physics of particles and fields we find in communication interactions another type of *multi*-body interactions.

Slide 11

Some of these multi-body interactions/communications are displayed.

The communicating or interacting authors are shown as green lines, there communication interactions are indicated as black lines.

In contrast to the example from particle physics we find in research also singly-authored papers or inner reasoning or communication processes of one author.

In the case of two authors the "exchange rationale from particle physics fits good, but in the case of three, four or five authors in the group with find a of multi-body interaction.

Moreover, such interacting groups can be rather large: Groups of several hundred scientists are known in many sciences such as for example clinical medicine, molecular biology and particle physics.

For example, also our workshop forms according to the list of participants a forty-one node communication network and the corresponding interaction picture would form a forty-dimensional simplex or tetrahedron with 820 ties.

Therefore, a new question shows up: All the collaborations have the same distances in this figure. But what is the correct length of the ties or intensity of the communication or collaboration? An answer can be borrowed from educational studies. If we assume, that we cannot have more than one speaker a time, then the communication intensity in our group of forty is much lower than in twenty small groups of two.

Therefore a normalization of the tie-strength or intensity with the inverse number of ties seems adequate in order to measure cooperation strength or intensity.

For example in the group of five co-authors in the lower right corner there are ten ties and the normalization factor equals one over ten.

Slide 12

Next we study the dynamics and evolution in time.

In this example we display four publications in the publication list of author A. The author A is displayed as a green line. The first publication had four authors, the second two authors and the third was single-authored and the fourth had three authors. The various co-authors are indicated by lines in different colors.

Co-author B in blue co-authored two of the four publications. All the other co-authors co-authored just one publication. Therefore, it is near to suppose a stronger or more intense cooperation of authors A and B than of other pairs of authors.

We can add the intensities or cooperation strengths in the set of publications obtain and obtain a resulting matrix for normalized cooperation intensities or strengths.

In this example this matrix is a symmetric five times five matrix on the right. The strongest cooperation intensity is found between authors A and B as expected and the sum over all intensities equals four and ergo equals the number of publications observed. The general proof, that the four in the lower right corner of the table *is* the number of publications is straightforward and is omitted.

Similarly we omit the proof, that the row-sums in the column to the right equal the fractional publication numbers of the authors and just check for author C: In this case the fractional publication number is of course one third.

In the table we recognize another important fact about cooperation networks: Most of the cooperation intensities or matrix elements vanish or are very small and therefore, the distances between the corresponding authors are large.

Finally we recall the definition of the degree centrality by Wassermann and Faust: The row-sums in the table are just the degree centralities in the definition of Wassermann and Faust.

In summary we have found a further Corollary: The fractional publication number of each author or object of study equals its degree centrality or importance or visibility in the network.

Slide 13

After this rather theoretical part we turn to a few empirical examples.

In this figure, only the size of the circles and the distances of the circles are meaningful due to the Multi-Dimensional-Scaling (MDS) approach chosen.

In a first example we considered the twelve publications in the publication list of a fellow in the Emmy Noether Programme of the German Research Foundation DFG four years after the PhD thesis.

In these twelve publications we found about 80 co-author-names and calculated the corresponding eighty times eighty cooperation intensity matrix.

This matrix was then displayed with the multi-dimensional scaling routine of SPSS in a two – dimensional plane of projected distances.

The degree centralities or fractional publication numbers are indicated as circle-sizes. The fellow (in a red circle) shows as expected the largest publication number or visibility in his ego-centered network.

Next to the Emmy Noether fellow we find two co-authors with also a large centrality and visibility. These two authors are the heads of the two research groups in which our young post-doc was working.

All the other smaller or in part invisible circles belong to collaborators of the two research groups contributing to the one or other publication.

Finally we observe, that the distances and degree centralities are somewhat correlated – very far away in the upper left corner, there are just small or invisible circles and near the “ego” or Emmy Noether Fellow there are co-authors with large publication number or centrality.

Slide 14

In the second example we considered all the 24000 Publications in Web of Science 2008 with at least one Swedish address. In the publications we found coauthor-addresses in 144 countries. The 24000 publications form the cumulative turnout network of Swedish researchers in 2008.

The degree centralities or fractional publication numbers are indicated as circles and the distances are calculated according to the cooperation intensity.

The largest degree centrality has Sweden with about two third of the total publication number of the network. Most foreign cooperation partners are found in the United States followed by Germany and England.

Amongst the central or large or important collaborators we find as expected the large countries of the G7, but also prominent are the much smaller Nordic Countries Denmark, Finland and Norway. Indicated with name are also the next ranked Netherlands and China. In the figure we find also two disadvantages stemming from an artefact of the calculus for the figure:

The first disadvantage is, that the most central countries are displayed at the boundary of the projection circle, whereas one would expect the central nodes rather in the middle of the figure.

The second disadvantage is, that some distances are falsified by the projection. In that sense the small nodes or circles in the left half are displayed “too near” to the central nodes. In contrast to the first disadvantage, this artefact is common to all projections and cannot be removed.

Despite of these disadvantages the dynamic and interactive interpretation of networks shows a very realistic picture of the underlying networks.

Slide 15

Thank you for your attention.