Q-measures for binary divided networks: an investigation within the field of informetrics

Ronald Rousseau

KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium.
ronald.rousseau@khbo.be

University of Antwerp, IBW, Universiteitsplein1, 2610 Wilrijk, Belgium.

Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium.

Q-measures for binary divided networks, as introduced by Flom, Friedman, Strauss and Neaigus are studied. These measures try to capture the idea of bridges between two groups in a connected undirected network. Values for these measures are calculated for building blocks such as line and star networks. As an application two small co-author networks are analyzed.

Introduction

Over the last years social network theory has enjoyed more and more success in informetric research (Kretschmer, 2004; White *et al*., 2004). Density and centrality measures known and studied in network theory are as useful in sociological as in informetric network studies (Otte & Rousseau, 2002).

Social network theory can be described as a strategy for investigating social structures. Its methods, however, can be applied in many fields, including the information sciences. Here scientists study publication and citation networks, co-citation networks, bibliographic coupling, collaboration structures, web relations and many other forms of social interaction networks (Adamic & Adar, 2003; Newman, 2001; van Raan, 2005). The so-called 'small world phenomenon' has attracted the attention of many scientists (Björneborn & Ingwersen, 2001; Braun, 2004; Kochen, 1989; Milgram, 1967; Newman & Watts, 1999; Rousseau, 2005). Such a small-world network is characterized as a graph or network exhibiting a high degree of clustering and having at the same time a small average distance between nodes.

Recently Flom *et al*. (2004) introduced a new sociometric network measure, denoted as Q, for individual actors as well as for whole networks. This measure tries to capture the idea of bridges between two groups in a connected undirected network. The higher its value the more this actor acts as a bridge between the two groups. Assume that there are T actors or nodes in the network. Group A contains m nodes, while the other group, denoted as B, contains n nodes, hence T = m+n. If actor x belongs to group A, and assuming for simplicity that actor x is a_m, then the Q-measure for this actor is defined as follows:

$$Q(x) = \frac{1}{(m-1)n}\left(\sum_{i=1}^{m-1}\sum_{j=1}^{n}\frac{g_{a_i b_j}(x)}{g_{a_i b_j}}\right) \qquad (1)$$

If actor x belongs to group B, and assuming again for simplicity that it is actor b_n, then its Q measure is defined as:

$$Q(x) = \frac{1}{m(n-1)}\left(\sum_{i=1}^{m}\sum_{j=1}^{n-1}\frac{g_{a_i b_j}(x)}{g_{a_i b_j}}\right) \qquad (2)$$

Here $g_{a_i b_j}$ denotes the number of geodesics, i.e. shortest paths, connecting $a_i \in A$ and $b_j \in B$.

The symbol $g_{a_i b_j}(x)$ represents the number of geodesics connecting a_i and b_j passing through x, where x is not one of the endpoints.

Existing measures such as betweenness centrality (Freeman, 1977) do not make a distinction between nodes belonging to different groups, or between geodesics remaining in the same group and geodesics crossing to the other group. This is the main motive for the introduction of this new measure.

Flom *et al*. (2004) make a further distinction between geodesics which cross exactly once between the two groups under study (leading to a measure denoted Q_1) and geodesics that possibly cross several times between the two groups (leading to Q_2). When Q_1 coincides with Q_2 in each node we will denote the measure simply as Q.

Q-measures for the whole network are defined in Flom *et al*. (2004) as the normalized average difference between the most central node (in the Q-sense), denoted as Q*, and all other nodes. This is:

$$Q_{Net} = \frac{\sum_{i=1}^{m}\left(Q^{*} - Q(a_i)\right) + \sum_{j=1}^{n}\left(Q^{*} - Q(b_j)\right)}{T-1} \tag{3}$$

Note that at least one of the terms in the numerator is certainly zero, namely when $Q(a_i)$ or $Q(b_j)$ is equal to Q*. This explains why the denominator is taken to be equal to T -1. Similar to the individual case one can also here define two Q_{NET}-measures.

Examples of calculations of Q-measures for basic networks

In this section we will calculate Q-measures for some basic configurations, such as lines and stars. The purpose of this is to get a feel of the meaning of different values of the Q-measure. This will also allow us to check if the measure behaves as intuitively expected of an indicator for the bridging function of a node. Note first that Q(x) is at most one, namely when x is always situated on the unique shortest path between any two nodes of different groups. Hence $0 \le Q(x) \le 1$.

Line networks: a simple example

We consider a line network of length 5. The first two nodes, a_1 and a_2 belong to the first group, the other three: b_1, b_2 and b_3 (in this order) belong to the second group. We follow the method described by Flom *et al*. (2004) for the calculation of Q-values. Note that in a line network with separated groups there is no difference between Q_1 and Q_2 as a shortest path can cross at most once the (imaginary) division line between the two groups.

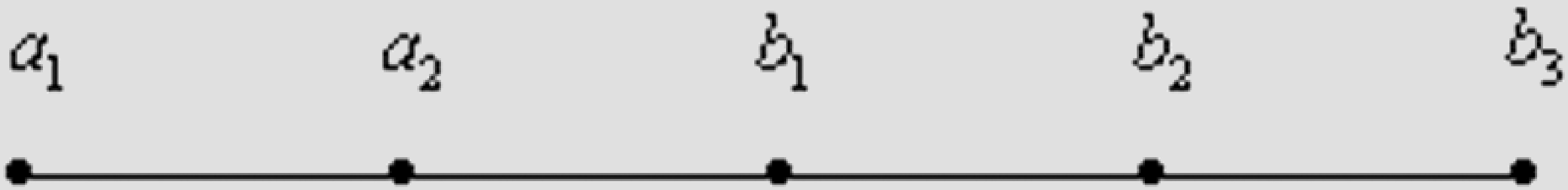

Figure 1: Line network with 2+3 nodes

In order to calculate Q-measures a matrix is drawn with columns containing the nodes belonging to the first group and rows consisting of nodes belonging to the other group (see Table 1). In each cell the non-terminal nodes of all geodesics between the corresponding row and column are entered. Then, to compute Q for a specific node, all geodesics containing that node are counted. This number is divided by all geodesics where that node is not a terminal node. If there are two or more geodesics between a pair of nodes, they are all included. Using this procedure the configuration of Figure 1 leads to Table 1.

Table 1. Matrix for the calculation of Q-values of the (2+3)-node line network shown in Figure1.

	b_1	b_2	b_3
a_1	a_2	a_2 b_1	a_2 b_1 b_2
a_2	-	b_1	b_1 b_2

Q-values for these five nodes are given in Table 2.

Table 2. Q-values of the (2+3)-node line network

$$Q(a_1) = 0$$
$$Q(a_2) = 3/3 = 1$$
$$Q(b_1) = 4/4 = 1$$
$$Q(b_2) = 2/4 = 1/2$$
$$Q(b_3) = 0$$

These results correspond to our intuition: nodes a_2 and b_1 form bridges between the two groups. Consequently, they have the maximum Q-value of 1. The two endpoints a_1 and b_3 clearly have no bridging function whatsoever: they receive a Q-value of 0. Finally node b_2 occupies an intermediary position. For the whole network we find $Q_{NET} = (5/2)/4 = 5/8$.

Line networks: case of m (m > 1) a-nodes followed by n (n >1) b-nodes.

The approach illustrated in the first example can readily be generalized. Also then there is no difference between Q_1 and Q_2 measures, hence the measure will simply be denoted as Q.

$Q(a_i) = [(i-1)n]/[n(m-1)] = (i-1)/(m-1)$, for $i = 1, \ldots, m$
and $Q(b_j) = [(n-j)m]/[m(n-1)] = (n-j)/(n-1)$, for $j = 1, \ldots, n$.

It follows, in particular, that $Q(a_m) = 1$ and $Q(b_1) = 1$, and further that $Q(a_1) = 0$ and $Q(b_n) = 0$. If m = 1 then $Q(a_m) = 0$; similarly, if n = 1 then $Q(b_1) = 0$.

The global network Q-measure is here $Q_{NET} = \dfrac{m+n}{2(m+n-1)} \approx \dfrac{1}{2}$ (for m or n large).

Perfectly intermixed line networks: m a-nodes perfectly intermixed with m-1 b-nodes.

We number a- and b-nodes from left to right. Note that in this case considering Q_1 or Q_2 makes a huge difference. Indeed, Q_1 (re-entering the same subgroup is not allowed) is zero for every node. Q_2 on the other hand, takes the following values:

$Q_2(a_i) = 2(i-1)(m-i)/(m-1)^2$ for $i = 1, \ldots, m$, and $Q_2(b_i) = ((2i-1)(m-i) - i)/(m(m-2))$, for $i = 1, \ldots, m-1$.

Note that $Q_2(a_1) = Q_2(a_m) = 0$, and generally $Q_2(a_i) = Q_2(a_{m-i+1})$; $Q_2(b_i) = Q_2(b_{m-i})$.

Q_2* (the maximum Q_2-value) is 1/2. Hence $Q_{2(NET)} = \dfrac{2m^2 - m + 1}{12(m-1)^2} \approx \dfrac{1}{6}$ (for m large).

Figure 2: Perfectly intermixed line network

Complete bipartite graphs

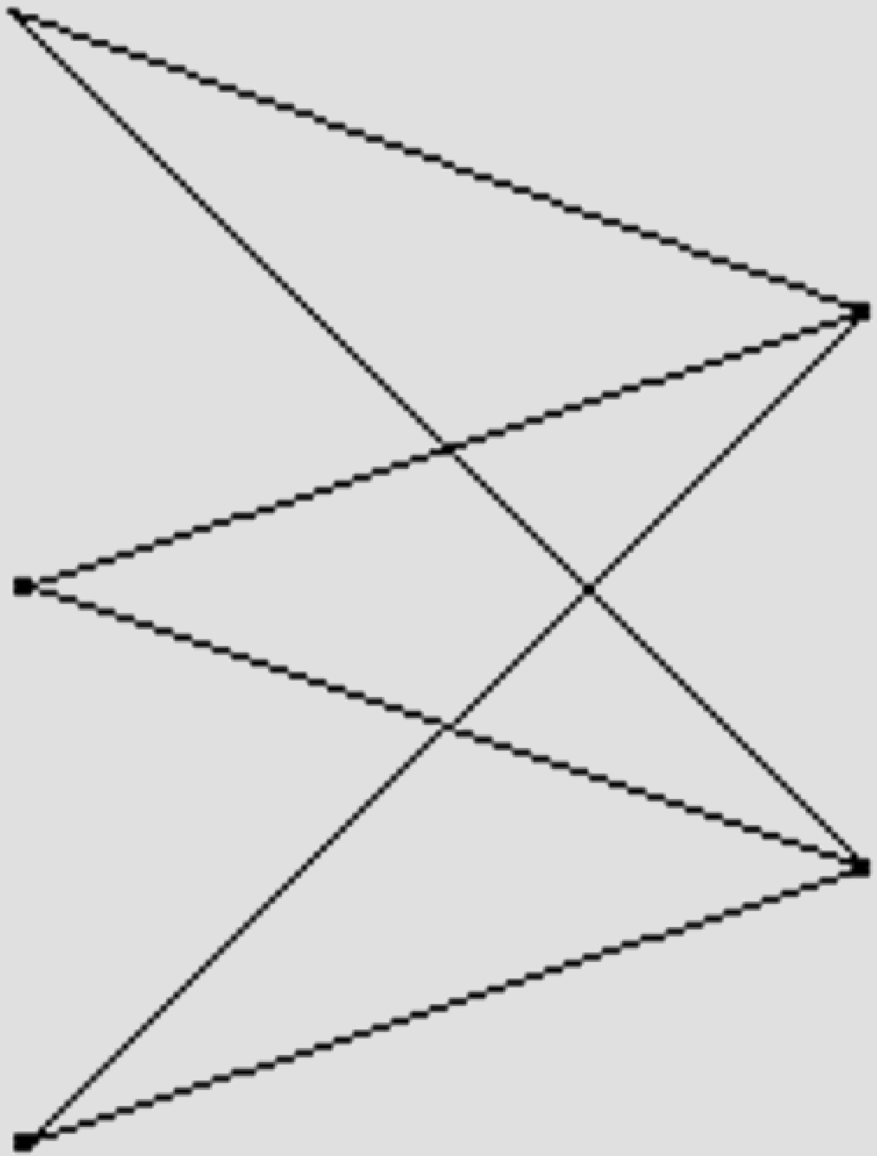

Figure 3: An example of a complete bipartite graph

Consider a graph partitioned into two groups of nodes. A complete bipartite graph is such that no two nodes of the same group are adjacent, but any two nodes belonging to different groups are. Q-measures for any node are zero. Note that the standardized betweenness centrality (Wasserman & Faust, 1994, p.190) of any node is the same in each group, but not zero. Indeed, betweenness centrality may be defined loosely as the number of times a node needs a given node to reach another node. As a mathematical expression the betweenness centrality of node i, is obtained as:

$$\sum_{j,k} \frac{g_{jik}}{g_{jk}} \qquad (4)$$

where g_{jk} is the number of shortest paths from node j to node k ($j,k \neq i$), and g_{jik} is the number of shortest paths from node j to node k passing through node i. The main difference between betweenness centrality and the Q-measures is that for Q-measures only shortest paths between nodes in different groups are considered. Standardized betweenness centrality, denoted as b(.) is then defined as expression (4) divided by (T-1)(T-2)/2, where T is the number of nodes in the network. If the group A in the complete bipartite graph has m nodes and group B n ones, then the betweenness centrality $b(a_j)$ is

$\frac{2}{(m+n-1)(m+n-2)} \cdot \frac{1}{m} \cdot \frac{n(n-1)}{2}$, where a_j denotes any element of group A. Similarly, $b(b_j)$ is

$\frac{2}{(m+n-1)(m+n-2)} \cdot \frac{1}{n} \cdot \frac{m(m-1)}{2}$, where b_j denotes any element of group B. This example illustrates the difference between betweenness centrality and the Q-measures.

Stars

Case I: one group consists of the center. Then Q = 0 for all nodes, hence also $Q_{NET} = 0$.

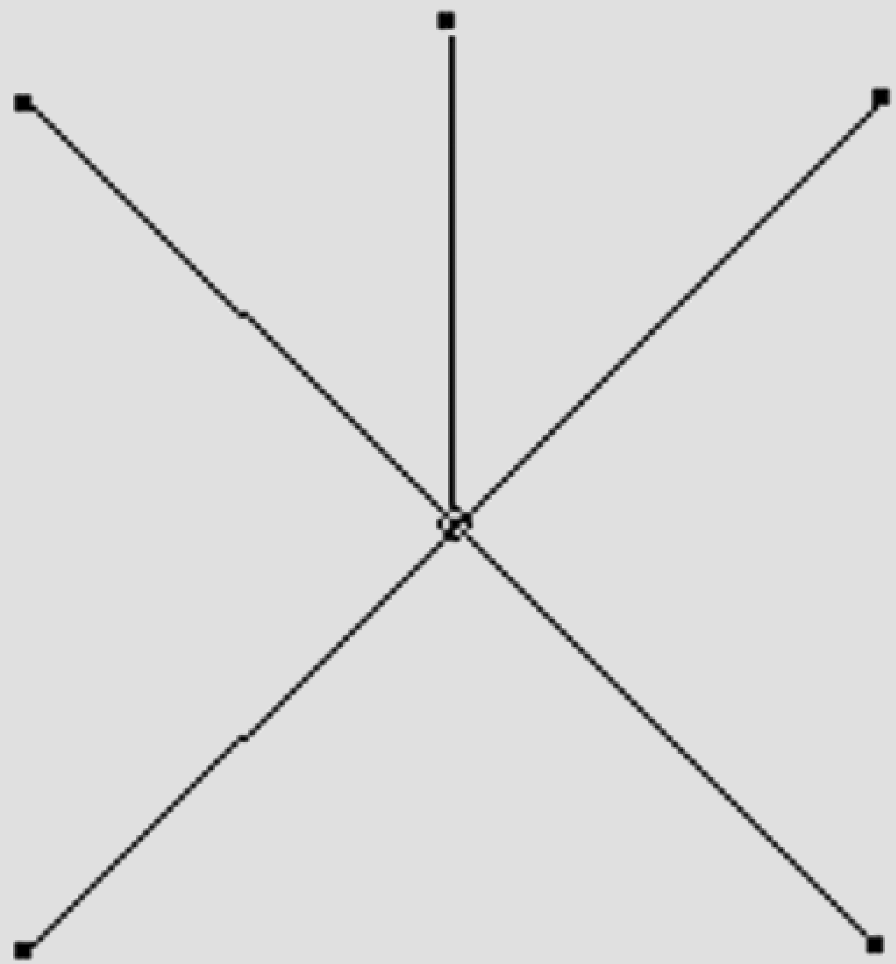

Figure 4: A star (case I)

Case II: one group consists of one satellite while the other group consists of the center and all other satellites. If this singleton is denoted as b_1, a_1 is the central actor and a_j the other ones (j = 2, .. ,m), then $Q(b_1) = 0$, $Q(a_1)= 1$ and $Q(a_j) = 0$, j = 2, ...,m. Here $Q_{NET} = m/m = 1$

As a final example we consider two stars where the central actors are connected. The central actor of one star is denoted as a_1, while the other ones are a_j (j = 2, ..., m); the central actor of the other star is b_1, while the other ones are denoted as b_k (k = 2, .., n). Here $Q(a_1) = Q(b_1) = 1$, while all other Q-values are zero.

In this case $Q_{NET} = \dfrac{m+n-2}{m+n-1}$, which is slightly less than one.

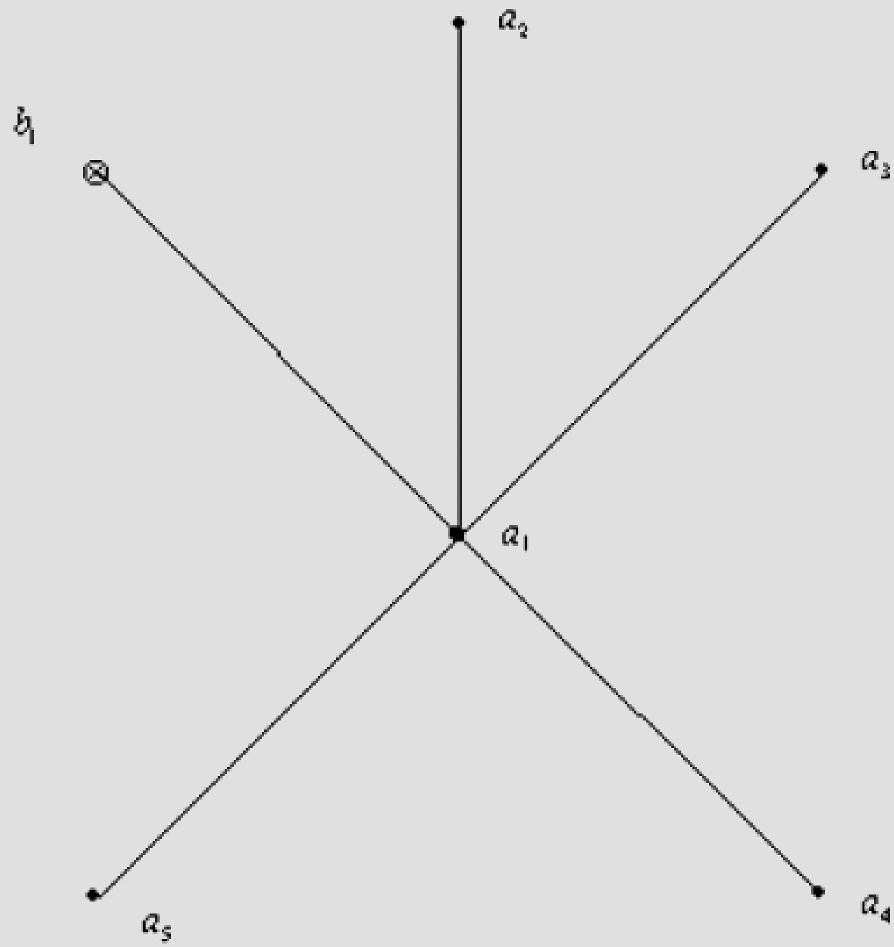

Figure 5: A star (case II with m = 5)

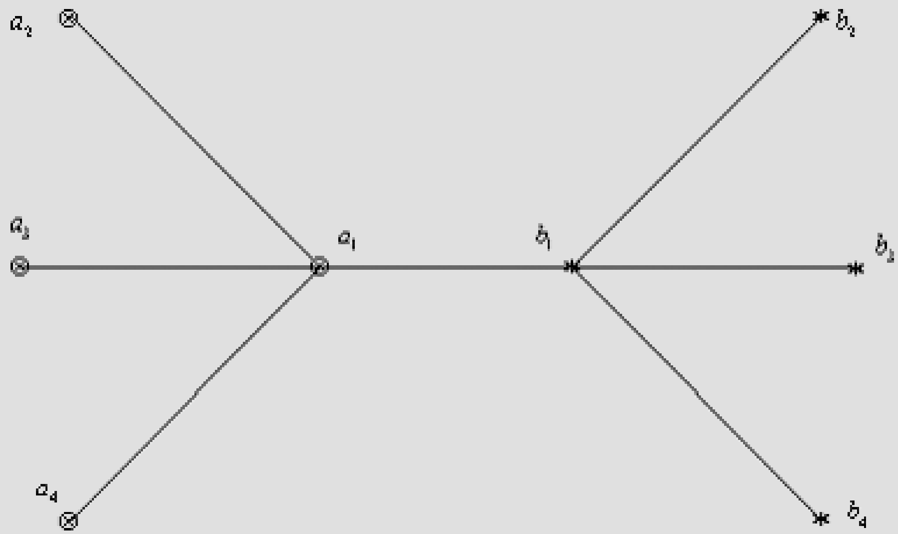

Figure 6: Two connected stars (with m = n = 3)

Two small co-author networks

In this section we present to small examples of real co-author networks. We calculate Q-measures and compare with some other network measures. The first example is a co-author network taken from JASIST, the second one is taken from the proceedings of the 8th ISSI conference.

A first co-author network

JASIST 55(10), 2004 contains a special topic session: document search interface design for large-scale collections. One of the articles in this section is written by a group of researchers from the University of Sheffield (UK) in collaboration with a Swedish colleague. Full bibliographical details of this article are shown in Table 3. We will refer to this article in short as CLIRS (for Cross-Language Information Retrieval System).

Table 3. Bibliographic details of the studied articles

Daniella Petrelli, Micheline Beaulieu, Mark Sanderson, George Demetriou, Patrick Herring, and Preben Hansen (2004). Observing users, designing Clarity: a case study on the user-centred design of a cross-language information retrieval system. Journal of the American Society for Information Science and Technology, 55(10): 923-934.

The following references of this article are used in the co-author graph.

- Lisa Ballesteros, and W. Bruce Croft (1998). Resolving ambiguity for cross-language retrieval. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '98) (pp. 64-71). Melbourne, Australia: ACM.
- Zoë Bathie, and Mark Sanderson (2001). iCLEF at Sheffield. In C. Peters (Ed.), Working notes for the CLEF 2001 Workshop (pp.215-217), Darmstadt, Germany: ERCIM.
- Micheline Beaulieu, and Susan Jones (1998). Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. Interacting with Computers, 10, 237-248.
- Daniella Petrelli, Per Hansen, Micheline Beaulieu, and Mark Sanderson (2002). User requirement elicitation for cross-language information retrieval. The New Review of Information Behaviour Research, 3, 17-35.
- Daniella Petrelli, George Demetriou, Patrick Herring, Micheline Beaulieu, Mark Sanderson (2003). Exploring the effect of query translation when searching cross-language. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.) Advances in cross-language information retrieval: Results of the CLEF 2002 Evaluation Campaign, Springer Lecture Notes in Computer Science LNCS 2785, Berlin: Springer.
- Mark Sanderson, and W. Bruce Croft (1999). Deriving concept hierarchies from text. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR '99) (pp. 206-213), Berkeley, CA: ACM.

The network shown in Figure 7 depicts the largest connected component of the co-author graph of all references in CLIRS. It is clearly dominated by the authors of CLIRS and some colleagues from the University of Sheffield. Scientists are represented by an abbreviation of their names. They are connected if they occur as co-author in at least one reference. Bibliographic details of these references are given in Table 3. Authors in this graph belong to two groups. Either they have a Sheffield address in at least one of these references used for our study, or they have not. The first group will be referred to as the Sheffielders (the bold ones in Fig.7), the other one the non-Sheffielders (script in Fig.7).

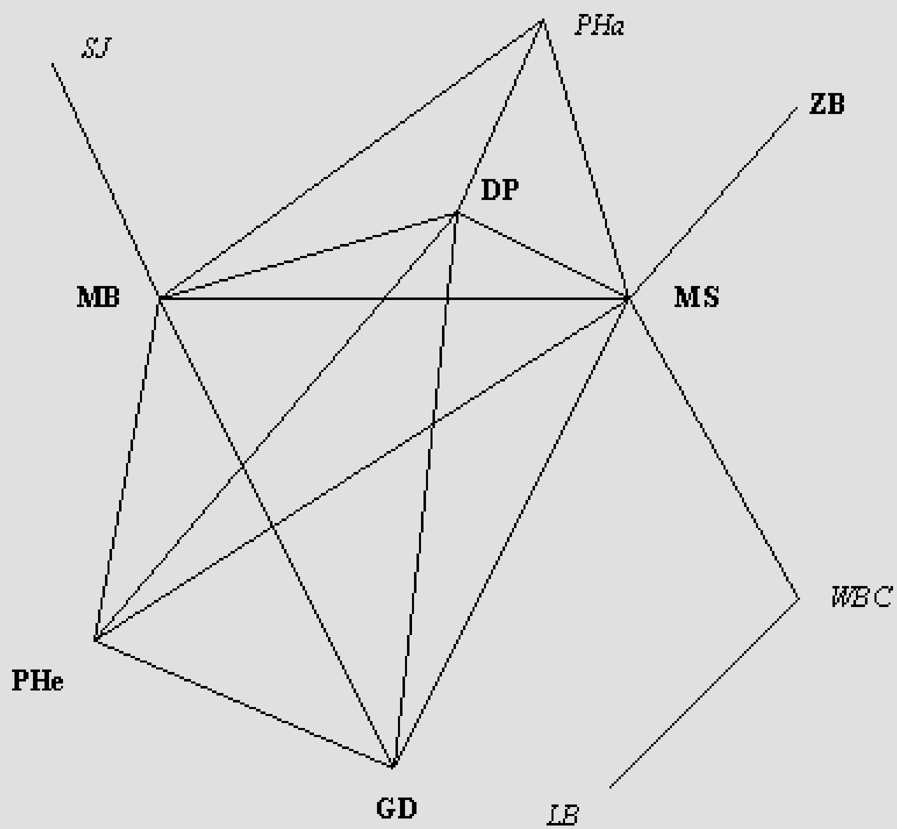

Figure 7: Co-authorship network of Sheffielders and non-Sheffielders

A glance at Figure 7 shows that this co-author network is dominated by the Sheffielders. For this study, however, we are not interested in the phenomenon of dominance, but in bridges between the two groups. Table 4 gives the details, following Flom *et al*. (2004), for the calculation of the Q-measure. Note that also here $Q_1 = Q_2$.

Table 4. Table used for the calculations of the Q-measure.

	MB	DP	MS	GD	PHe	ZB
SJ	---	MB	MB	MB	MB	MB - MS
WBC	MS	MS	---	MS	MS	MS
LB	WBC - MS	WBC - MS	WBC	WBC - MS	WBC - MS	WBC - MS
PHa	---	---	---	MB ; DP ; MS	MB ; DP ; MS	MS

Using Table 4 and formulas (1),(2) yields the following Q-measures: Q(Sanderson) = 19/30, Q(Croft) = 6/18, Q(Beaulieu) = 17/60, Q(Petrelli) = 1/30, Q(Demetriou) = Q(Herring) = Q(Bathie) = 0, Q(Jones) = Q(Balesteros) = Q(Hansen) = 0.

Clearly, among the Sheffielders, George Demetriou, Patrick Herring and Zoë Bathie play no role at all as bridges between the two groups. The same is true for the non-Sheffielders Susan Jones, Lisa Balesteros and Per Hansen. Daniella Petrelli has a small Q-value, while Micheline Beaulieu and especially Mark Sanderson play important roles as bridges between the two groups. Similarly among the non-Sheffielders W. Bruce Croft is the main bridge. Note that his role as a bridge is completely derived from being a co-author of someone belonging to the Sheffield group as well as being a co-author of someone not belonging to the Sheffielders. For this example Q_{NET} = 101/180.

For comparison's sake we add the values for some classical centrality measures.

Degree centrality of a node is the number of ties this node has. Denoting this measure as d, gives (in decreasing order): d(Sanderson) = 7, d(Beaulieu) = 6, d(Petrelli)=5, d(Herring) = d(Demetriou) = 4, d(Hansen) = 3, d(Croft) = 2, d(Jones) = d(Bathie) = d(Balesteros) = 1

Closeness centrality of a node is calculated in two steps. First, one determines the sum of all distances (= lengths of shortest paths) to all other nodes. Then the standardized closeness centrality is equal to the number of nodes minus one, divided by this sum of distances. Values are: c(Sanderson) = 0.82, c(Beaulieu) = 0.69, c(Petrelli)= 0.64, c(Herring) = c(Demetriou) = 0.60, c(Hansen) = 0.56, c(Croft) = 0.53, d(Bathie) = 0.47, c(Jones) = 0.43, c(Balesteros) = 0.36. Note that in this example closeness centrality just refines degree centrality.

Finally we calculated the normalized betweenness centrality, denoted as b, in this network. We find: b(Sanderson) = 34/54, b(Beaulieu) = 13/54, b(Croft) = 2/9, b(Petrelli) = 1/54, b(Herring) = b(Demetriou) = b(Hansen) = b(Jones) = b(Bathie) = b(Balesteros) = 0. In this simple case betweenness centrality gives almost the same ranking as the new Q-measure, only Beaulieu and Croft switched places. Moreover, some values and ratios between values are different.

A second co-author network

As a second real-life application we study the co-authorship network involving UNSW authors, who have an article or poster publication in the Proceedings of the 8th ISSI conference (Davis & Wilson, 2001).

From July 16 to July 20, 2001 the 8th International Conference on Scientometrics and Informetrics was held at the University of New South Wales (UNSW), Sydney, Australia. Its two volume conference proceedings contains all announced talks and posters. For this example we consider the connected component in the co-author graph of these proceedings containing Mari Davis and Concepción S. Wilson, its editors and members of the Bibliometric & Informetric research Group (BIRG) of the University of New South Wales. Full bibliographic details of the articles whose authors are included in this graph are shown in Table 5.

Table 5. Bibliographic details of the studied articles

All articles and posters are taken from:

Proceedings of the 8th International Conference on Scientometrics & Informetrics (two volumes). Mari Davis and Concepción S. Wilson, editors. Published by the Bibliometric & Informetrics Research group (BIRG), University of New South Wales, Sydney, Australia, 2001.

Articles

- Sri Hartinah, Mari Davis, Amru Hydari and Philip Kent. Indonesian nutrition research papers 1979-2000: a bibliometric analysis, pp 225-234.
- William W. Hood and Concepción S. Wilson. Distribution of phrases in the fuzzy set literature in the period 1965-1993, pp. 253-263.
- Liming Liang, Yongzheng Guo and Mari Davis. How do scientists of different ages collaborate in China: the case of computer science and control theory, pp. 395-409.
- Farideh Osareh and Concepción S. Wilson. Iranian scientific publications: collaboration, growth, and development from 1985-1999, pp. 499-509.

Posters

- Joanne Orsatti, Concepción S. Wilson and Mari Davis. Disciplinarity explored through the emergent domain of consciousness, pp. 865-868.
- L. Sulistyo-Basuki, Zainal Hasibuan, Mustangimah and Sri Hartinah. Studies toward subject dispersion in atomic and nuclear energy journals published in Indonesia 1986-1998 based on subject and citation analysis, pp.883-885.
- Weiping Yue and Concepción S. Wilson. The relationship of two derived measures: impact factor and immediacy index, pp. 893-896.

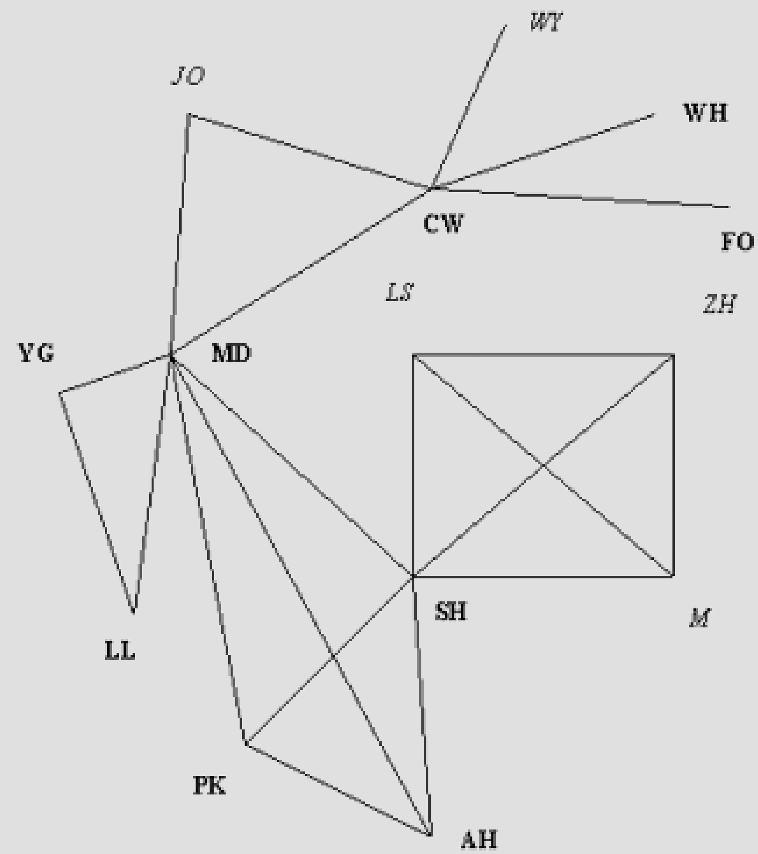

Figure 8: The co-author graph of UNSW contributions

We consider, as a simple example, the following two groups. The first one consists of Mari Davis (indicated as MD in the graph), Concepción S. Wilson (CW) and every colleague who has co-authored a full article in the proceedings with one of them. These are: Sri Hartinah (SH), Amru Hydari (AH), Philip Kent (PK), William W. Hood (WH), Liming Liang (LL), Yongzheng Guo (YG) and Farideh Osareh (FO). This group is indicated in bold. The other group consists of all other colleagues in this connected component. They have either collaborated with Mari Davis or Concepción S. Wilson on a poster presentation, or have collaborated with someone who has collaborated with Mari Davis or Concepción S. Wilson on a full article. This group is indicated in script and consists of: Joanne Orsatti

(JO), L. Sulistyo-Basuki (LS), Zainal Hasibuan (ZH), Mustangimah (M) and Weiping Yue (WY).

For this study we are only interested in bridges between the two groups. Table 6 gives the details, following Flom *et al*. (2004), for the calculation of the Q-measure. Scientists in the cells of this table are situated on a shortest path between the colleagues on top of the row and column.

Table 6. Table used for the calculations of the Q-measure.

	CW	MD	FO	WH	YG	LL	SH	PK	AH
JO	-	-	CW	CW	MD	MD	MD	MD	MD
WY	-	CW	CW	CW	CW – MD	CW - MD	CW - MD	CW - MD	CW - MD
LS	SH - MD	SH	SH-MD-CW	SH – MD - CW	SH-MD	SH - MD	-	SH	SH
ZH	SH - MD	SH	SH-MD-CW	SH – MD - CW	SH-MD	SH - MD	-	SH	SH
M	SH - MD	SH	SH-MD-CW	SH – MD - CW	SH-MD	SH - MD	-	SH	SH

Clearly, most colleagues do not play a bridging function in this graph. This is, in particular true for all members of the second group, as they are all directly connected to a member of the first group. Hence: Q(FO) = Q(WH) = Q(YG) = Q(LL) = Q(PK) = Q(AH) = Q(JO) = Q(WY) = Q(LS) = Q(ZH) = Q(M) = 0. The other three colleagues do have a bridging function: Q(MD) = 25/40, Q(SH) = 24/40 and Q(CW) = 16/40.

For comparison's sake we add the degree centrality. Denoting this measure as d, gives (in decreasing order): d(MD) = 7, d(SH) = 6, d(CW)=5, d(ZH) = d(M) = d(LS) = d(AH) = d(PK) =3, d(LL) = d(YG)= d(JO) = 2, d(FO) = d(WH) = d(WY) = 1.

Conclusion

Q-measures capture the idea of bridges between two groups in a connected undirected network. Values for these measures were calculated for building blocks such as line and star networks. These theoretical cases provide examples illustrating the difference between these Q measures and centrality measures. They also illustrate the difference between Q_1 and Q_2. The small real-world co-author networks that we investigated illustrate the usefulness of this new concept. Clearly much more work has to be done on the theoretical side as well, and in particular, on the practical side, in order to prove that Q-measures really capture the notion for which they are intended. It would also be useful to have access to a computer program in order to study Q-measures in larger networks.

References

Adamic, L. & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 25*, 211-230.

Björneborn, L. & Ingwersen. P. (2001). Perspectives of webometrics. *Scientometrics, 50*, 65-82.

Braun, T. (2004). Hungarian priority in network theory. *Science, 304*, 1745-1746.

Davis, M. & Wilson, C.S. (Eds.). (2001). *Proceedings of the 8th International Conference on Scientometrics & Informetrics*. Sydney (Australia): Bibliometric & Informetrics Research Group (BIRG), University of New South Wales.

Flom, P.L., Friedman, S.R., Strauss, S., & Neaigus, A. (2004). A new measure of linkage between two sub-networks. *Connections, 26*(1), 62-70.

Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35-41.

Kochen, M. (Ed.). (1989). *The Small World*. Norwood (NJ): Ablex.

Kretschmer, H. (2004). Author productivity and geodesic distance in co-authorship networks, and visibility on the Web. *Scientometrics, 60*, 409-420.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*, 60-67.

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

Newman, M.E.J. & Watts, D.J. (1999). Scaling and percolation in the small-world network model. *Physical Review E, 60*, 7332-7342.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*, 441-453.

Rousseau, R. (2005). Another look at small worlds: one node set – two link structures. *ISSI Newsletter, 1*(1), p.7.

Van Raan, A.F.J. (2005). Reference-based publication networks with episodic memories. *Scientometrics, 63*, 549-566.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge University Press.

White, H.D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure? *Journal of the American Society for Information Science and Technology, 55*, 111-126.