

Social networks of author–coauthor relationships

Yasmin H. Said, Edward J. Wegman*, Walid K. Sharabati¹, John T. Rigsby

Department of Computational and Data Sciences, George Mason University, Fairfax, VA, USA

Received 8 July 2007; accepted 14 July 2007

Available online 9 August 2007

Abstract

Social network analysis has proven to be a useful tool in analysis of many situations. We begin by giving an overview of social network analysis. We then illustrate the concepts by examining the networks of authors of scholarly publications. Scholarly publication is in many ways the lifeblood of academic institutions and there are strong incentives, both in terms of prestige and financial compensation, for faculty members to publish. Different disciplines and individuals have evolved distinguishable mechanisms for coping with the publication pressures. We examine the co-authorship networks of a number of prominent scholars. Based on the clustering within the co-author social network, we distinguish several styles of co-authorship including solo models (no co-authors), mentor models, entrepreneurial models, and team models. We conjecture that certain styles of co-authorship lead to the possibility of group-think, reduced creativity, and the possibility of less rigorous reviewing processes. Published by Elsevier B.V.

Keywords: Clustering; Network science; Allegiance model; Entrepreneurial model; Mentor model; Solo model; Laboratory model; Co-occurrence matrix

1. Introduction

A social network is an emerging tool frequently used on quantitative social science to understand how individuals or organizations are related. The basic mathematical structure for visualizing the social network is a graph. A graph is a pair (V, E) where V is a set of nodes or vertices and E is a set of edges or links. Social network analysis (also called network theory) has emerged as a key technique and a topic of study in modern sociology, anthropology, social psychology and organizational theory. The shape of the social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More “open” networks, with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties. See Granovetter (1973).

Social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages. The social entities are referred to as actors that are represented by the vertices of the graph. Most social network applications consider a collection of actors that are all of the same type. These are known as one-mode networks. Social ties link actors to one another. The range and type of social ties can be quite extensive. A tie establishes a linkage between a pair of actors. Linkages are represented by edges of the graph. Examples of linkages include the

* Corresponding author. Tel.: +1 703 993 1691.

E-mail address: ewegman@gmail.com (E.J. Wegman).

¹ The author is also affiliated with American University.

evaluation of one person by another (such as expressed friendship, liking, respect), transfer of material resources (such as business transactions, lending or borrowing things), association or affiliation (such as jointly attending the same social event or belonging to the same social club), behavioral interaction (talking together, sending messages), movement between places or states (migration, social or physical mobility), physical connection (a road, river, bridge connecting two points), formal relations such as authority and biological relationships such as kinship or descent. A linkage or relationship establishes a tie at the most basic level between a pair of actors. The tie is an inherent property of the pair. Many kinds of network analysis are concerned with understanding ties among pairs and are based on the dyad as the unit of analysis.

A social network consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a significant feature of a social network. A partition of a network is a classification or clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster. Partitions may specify some property that depends on attributes of the vertices. Partitions divide the vertices of a network into a number of mutually exclusive subsets. That is, a partition splits a network into parts. Partitions are also sometimes called blocks or block models. These are essentially a way to cluster actors together in groups that behave in a similar way. Allegiance measures the support that an actor provides for the structure of his block. An actor supports his block by having internal block edges. A measure of this is the total number of edges that an actor has internal to his block. An actor supports his block by not having external edges from the block to other actors or blocks. A measure of this is the total number of possible external edges minus the total number of existing external edges. The allegiance for a block is a weighted sum of a measure of internal allegiance and a measure of external allegiance. The overall allegiance for a social network is the sum of the allegiances for the individual blocks. If the overall allegiance is positive, then a good partition was made. The partitioning continues recursively until a new partition no longer contributes to a positive allegiance.

Centrality is one of the oldest concepts in network analysis. Most social networks contain people or organizations that are central. Because of their position, they have better access to information, and better opportunity to spread information. This is known as the ego-centered approach to centrality. The network is centralized from socio-centered perspective. The notion of centrality refers to the positions of individual vertices within the network, while centralization is used to characterize an entire network. A network is highly centralized if there is a clear boundary between the center and the periphery. In a highly centralized network, information spreads easily, but the center is indispensable for the transmission of information.

There are several ways to measure the centrality of vertices and the centralization of networks. The concepts of vertex centrality and network centralization are best understood by considering undirected communication networks. If social relations are channels that transmit information between people, central people are those people who have access to information circulating in the network or who may control the circulation of information, i.e., they play a brokerage role.

The accessibility of information is linked to the concept of distance. If you are closer to the other people in the network, the paths that information has to follow to reach you are shorter, so it is easier for you to acquire information. If we take into account direct neighbors only, the number of neighbors (the degree of a vertex in a simple undirected network) is a simple measure of centrality. If we also want to consider other indirect contacts, we use closeness centrality, which measures our distance to all other vertices in the network. The closeness centrality of a vertex is higher if the total distance to all other vertices is shorter. The importance of a vertex to the circulation of information is captured by the concept of betweenness centrality. From this perspective, a person is central if he or she is a link in more information chains between other people in the network. High betweenness centrality indicates that a person is an important intermediary in the communication network. Information chains are represented by geodesics and the betweenness centrality of a vertex is simply the proportion of geodesics between other pairs of vertices that include the vertex. The centralization of a network is higher if it contains very central vertices as well as very peripheral vertices.

2. Clustering and allegiance

Clustering in social networks begins with dividing the set of actors into discrete, non-overlapping subsets called partitions. The set of partitions is $P = \{P_1, \dots, P_k\}$ where k is the total number of partitions. We let $P(i, k)$ represent the partition to which actor i belongs when there are k partitions. The partition determines the block model, $B = \{B_{1,1}, B_{1,2}, \dots, B_{k,k-1}, B_{k,k}\}$ (Wasserman and Faust, 1994). The block model is the device that clusters or groups the network data. The block, $B_{i,j}$, is formed from the ties of actors in partition i , P_i , to the actors in partition j , P_j . If $i \neq j$,

the block $B_{i,j}$ represents ties between partitions P_i and P_j . If $i = j$, $B_{i,j}$ represents internal ties of actors within the block. These latter diagonal blocks represent a clustering of the actors. Generally speaking we like to see blocks that are cliques or nearly cliques in the usual graph-theoretic sense of a clique.

Rigsby (2005) developed the concept of allegiance in order to have a systematic way to form the partition and the blocks. As with usual clustering methods the appropriate number of clusters is not usually known. A quantitative measure of block model strength allows us to estimate the true number of partitions. Allegiance measures the support an actor provides for the structure of his block. An actor supports his block by having internal block edges. A measure of this is the total number of internal block edges that an actor has. We denote this by H_{int} where $G(i, k)$ is the group of all actors belonging to the same partition $P(i, k)$.

$$H_{\text{int}}(i, k) = \sum_{j \in G(i, k)} E(i, j),$$

where $E(i, j)$ is the edge weight from actor i to actor j . An actor supports his block by not having external edges from the block. A measure of this is the total number of possible external edges minus the total number of existing external edges. We denote this by H_{ext} . If the total number of actors is N , let $N_{P(i, k)}$ be the number of actors in partition $P(i, k)$. Then

$$H_{\text{ext}}(i, k) = N - N_{P(i, k)} - \sum_{j \notin G(i, k)} E(i, j).$$

We define allegiance $A(i, k)$ as the measure of how much an actor supports his block at a partition size k .

$$A(i, k) = \frac{1}{2} H_{\text{int}}(i, k) + \frac{1}{2} H_{\text{ext}}(i, k).$$

Initially, the data are all in one partition and $A(i, 1)$ is simply half of the out degree of actor i . The allegiance of actors changes as the number of partitions change. We let

$$\Sigma_{A(k)} = \sum_{i=1}^N A(i, k)$$

and

$$O_{A(k)} = \Sigma_{A(k)} - \Sigma_{A(k-1)}.$$

$\Sigma_{A(k)}$ is the summation of allegiance for all actors at k partitions. The first cut divides the data into two partitions; the second cut divides the data into three partitions, and so on. If the overall allegiance $O_{A(k)}$ is positive, then a good partitioning was made. This process is iterated until additional partitioning makes $O_{A(k)}$ negative or zero. Let

$$D_{A(i, k)} = A(i, k) - A(i, k-1)$$

and

$$\Sigma_{D_{A(k)}} = \sum_{i=1}^N D_{A(i, k)}.$$

As noted earlier, $\Sigma_{A(1)}$ is half of the sum of the out degree for all of the actors. The individual actor differences, $D_{A(i, k)}$, in allegiance at each partitioning indicates how each actor is affected by the partitioning. The summation of the actor allegiance differences, $\Sigma_{D_{A(k)}}$, shows the strength change in the block model at partitioning k . When $\Sigma_{D_{A(k)}}$ is negative, the block structure strength is decreased by this partitioning. The first negative value of $\Sigma_{D_{A(k)}}$ yields the maximum number of partitions. If this first negative value of $\Sigma_{D_{A(k)}}$ is significantly negative, then the maximum partition size is $k - 1$. Note that overall allegiance $O_{A(k)} = \Sigma_{D_{A(k)}}$.

3. Co-authorship networks

Co-authorship establishes a linkage or tie between two individuals. These linkages can be examined as a social network and patterns exhibited in the social network of an individual and his co-authors can shed considerable light on how an author works and deals with his colleagues. Using the block model analysis outlined in the previous section, we can cluster the set of co-authors. It should be noted that while a social network is often laid out as a graph for purposes of visualization, it is also possible to simply look at a co-occurrence matrix. The latter view is useful for purposes of understanding the block model structure.

Wegman et al. (2006) undertook a social network analysis of a segment of the paleoclimate research community. This analysis met with considerable criticism in some circles, but it did clearly point out a style of co-authorship that led to intriguing speculation about implications of peer review. Based on this analysis and the concomitant criticism, we undertook to examine a number of author-coauthor networks in order to see if there are other styles of authorship. Based on our analysis we identify four basic styles of co-authorship, which we label, respectively, solo, entrepreneurial, mentor, and laboratory. The individuals we have chosen to represent the styles of co-authorship all have outstanding reputations as publishing scholars. Because of potential for awkwardness in social relationships, we do not identify any of the individuals or their co-authors.

The solo style needs little explanation. Basically solo style characterizes authors who do not have any coauthors. The solo style is very rare and usually is only found among isolated scientists. The almost universal availability of Internet connectivity means that there are few truly isolated scientists. One observation worth mentioning is that papers that are heavily mathematical are often the work of a single individual and therefore often follow the solo style.

Fig. 1 is the example of a block model type we call entrepreneurial style. We have removed co-author names from this network. Notice that in what follows, the matrix representations are symmetric. We identify the blocks by numbering the partitions from upper left to lower right. So for example Block 1 contains the principal author alone. This graphic represents a matrix of a network of the 47 co-authors of a principal author who is represented in the upper left-hand corner. The rows and columns represent the set of actors (authors) and wherever there is a black square in the grid means that there is a co-author relationship. The block diagonal structure indicates that there are strong clusters.

The black borders occur because the principal author, Block 1, is a co-author with every one of actors in this network. In this co-author social network, the block model suggests that the actors are partitioned into nine groups. The principal

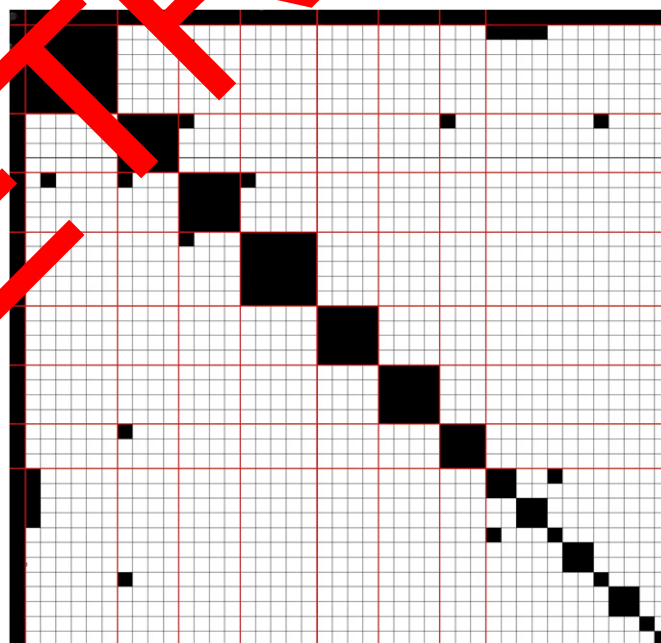


Fig. 1. Block model (matrix) representation of an entrepreneurial co-author social network.

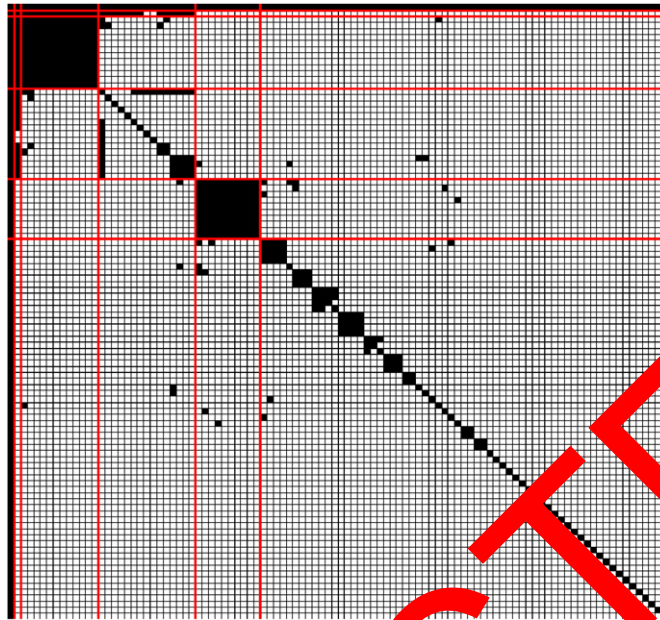


Fig. 2. Block model (matrix) representation of a mentor-style co-author social network.

author is in a group of his own. The larger black squares mean that within the group, every actor is a co-author with every other actor in the same group albeit not necessarily on the same paper. That is, say two or three of nine may be on a given paper, but not all nine are on the same paper. The implication is that the principal author collaborates with closely coupled groups. We have referred to this as an entrepreneurial style because in essence the principal author has sought out groups with sub-specialties needed for a particular paper. In essence these specialty groups represent subdisciplines and are tightly coupled within the subspecialty. These specialty groups tend not to collaborate with each other as indicated by the block diagonal structure in the social network. The last group on the bottom right side in Fig. 1 is not a coherent group, but represents something of a miscellaneous set of co-authors of the principal author.

Fig. 2 is an example of a block model type we call the mentor style. Here the principal author has 101 co-authors. Again the black borders mean that the principal author, who occupies Block 1 alone, has collaborated with all 101 co-authors. Conspicuously absent from this matrix representation is the strong block diagonal structure seen in the entrepreneurial model. Block 2 is another single co-author who is the principal author's most frequent collaborator. There are two large blocks, Block 3 and 5. Both of these represent single papers in which two different teams collaborated with the principal author. Unlike the entrepreneurial example, all of the actors in these blocks were co-authors on the one paper. Other than those two examples, there is essentially no block diagonal structure. We note that using the allegiance methodology, there are only six groups and even within these groups, the clustering structure is not very strong. This principal author tends to co-author papers with younger colleagues who were his students or other young associates. His basic strategy is to work with associates individually getting them started in scholarship. Once they have the experience of writing a few papers, they begin to write on their own or with other colleagues in a new venue. The horizontal and vertical bars in the fourth block indicate that the co-authors in that block have moved on from the mentor and have collaborations with other authors. Thus the reason we call such a block model co-author social network the mentor style is fairly clear.

Fig. 3 represents a block model type we call the laboratory style. The main scholar in this co-author social network is a biostatistician, who is a consultant on a number of National Institutes of Health contracts and thus has a distinctive block diagonal structure representing papers in which most of the members of a particular laboratory project have their names on the paper. In this example, the principal author has 29 collaborators. Unlike the entrepreneurial style where the cliques tend not to have every member of the block on every paper, in the laboratory model, the blocks tend to have most if not all of the members of the block on every paper coming from the laboratory. Notice that the first three blocks in Fig. 3 are individuals. These are the other consultants with the principal author who tend to collaborate with the principal author and with certain blocks. For example the author in Block 2 collaborates principally laboratory component represented

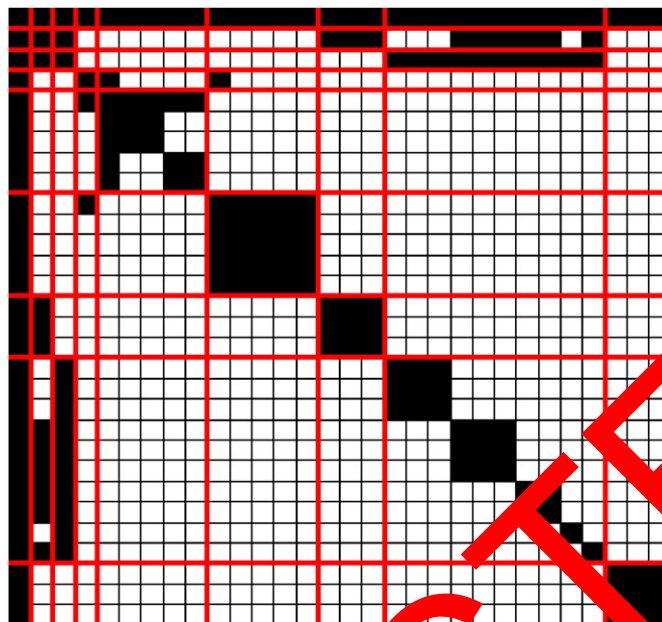


Fig. 3. Block model (matrix) representation of a laboratory-style co-author social network.

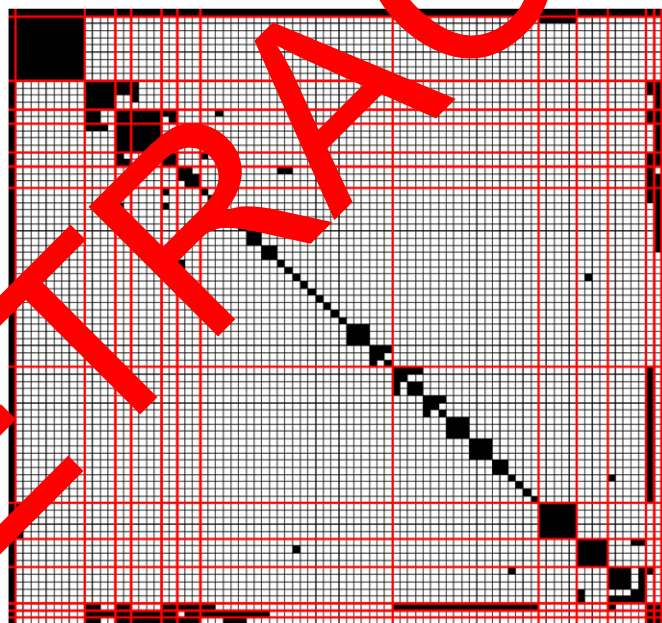


Fig. 4. Block model (matrix) representation of the co-author social network of an individual who changed career status.

by Block 7 and part of Block 8. The author in Block 3 collaborate principally with the laboratories represented by Block 8. Although we have not examined the co-authorship network of a medical/biological laboratory director here, in general, that network would have a very strong block diagonal structure with very few blocks.

Fig. 4 represents another interesting co-author social network. Here the principal author has 88 co-authors in the social network. Again using the allegiance criterion, we obtain a very interesting block model with 15 blocks. There are significant blocks in both the upper left and lower right corners of this matrix. This principal author was for a time embedded in a laboratory setting at one institution, became an academic in a standard department, and later, while

still an academic, became affiliated as a consultant to another laboratory. The purely academic phase is situated in the middle, which bears the hallmarks of the mentor network, while the upper left and lower right resemble the laboratory model. This hybrid model represents something of a combination of the laboratory and mentor networks.

4. Implications for peer review

Wegman et al. (2006) suggested that the entrepreneurial style could potentially lead to peer review abuse. Many took umbrage at this suggestion. Nonetheless, there is some merit to this idea. Peer review is usually regarded as a gold standard for scientific publication. Clearly it is desirable that the peer reviewer have three important traits: independent, unbiased, and knowledgeable in the field. As any hard-working editor or associate editor knows, finding independent, unbiased, and knowledgeable referees for a paper or proposal is a difficult chore. This is especially true in a rather narrow field where there are not many experts so that issues of independence arise quickly. Clearly, as a field becomes increasingly specialized, there are not as many independent experts. The finding someone who is both independent and knowledgeable is difficult. In the past, when many more authors adopted a solo style of authorship, finding someone who was not a co-author was relatively easy. Nonetheless, the issue of unbiasedness still was an issue. The introduction of double-blind refereeing focused on the unbiasedness issue. It is only natural for a referee to act in a favorable way toward someone he or she admires scientifically or with whom he or she may have a friendship. Double-blind refereeing, however imperfect it might be, at least removes the perception of referee bias by removing the name of the authors. In an era of Google Scholar, it is usually not very hard to turn the names of the authors of a paper simply by googling the title of the paper. Even if the authors of a paper are unknown, the topic of a paper and its similarity to the work of the referee can bias the referee to look upon the paper in question favorably. Of course making these observations about potential problems immediately suggests that editors and referees are not acting with integrity. We do not mean to imply that this is the case. Indeed, such attitudes and biases may be entirely subconscious. Interestingly, referees are usually not identified. Presumably this is because of the fear of retaliation against a referee who provides an unfavorable review. Implicit in this procedure is the assumption that the author would not behave with integrity.

Of course because referees are not identified, getting hard evidence of independence, unbiasedness and knowledgeable expertise is not readily available. The social network analysis can therefore only be suggestive. It is our contention, however, that safeguards such as double-blind refereeing and not identifying referees invariably lead to the conclusion that peer review is at best an imperfect system. Any author with a long history of publication in their heart of hearts knows that they have benefited or have been benefited, probably both, by imperfect peer review.

The social network analysis of an entrepreneurial style suggests the following. There are many tightly coupled groups working closely together in a relatively narrow field. It is clear that closely coupled groups have a common perspective. Thus it is very hard to find a referee that is both knowledgeable and independent. Because of the common perspective, in addition it is very hard to find an unbiased referee. Thus this style of co-authorship makes it more likely that peer review will be compromised. One mechanism for selecting referees is to look at papers referenced by the paper in question. This possibility means that a naive associate editor might actually pick someone from the social network of co-authors, who is not obviously a co-author. Indeed, the paleoclimate discussion in Wegman et al. (2006), while showing no hard evidence, does suggest that the papers were refereed with a positive, less-than-critical bias. In contrast, the laboratory style of co-authorship is somewhat less prone to peer-review problems in that the laboratories themselves, rather than individual scientists, become the publishing unit, and as such, tend to be somewhat more competitive with each other. Nonetheless, we note that many discussions of concerns about peer review seem to take place in medical/biological related journals. Finally, the mentor style of co-authorship, while not entirely free of the possibility of bias, does suggest that younger co-authors are generally not editors or associate editors. And often they are not in a position to become referees, so that the possibility of bias is much reduced. Nonetheless, even here, a widely respected principal author has the possibility of smoothing the path for his or her junior collaborators, while the papers of a high reputation principal author may not be as critically reviewed as might be desirable.

5. Conclusions

Social network analysis of author–coauthor networks at the very least gives an interesting insight into the sociology of scientific workers. The fact that there are distinct modes of authorship readily identifiable by the block model, while

interesting in its own right, also provides insight into the why certain fields of study may have migrated into a more politically driven framework.

Acknowledgments

The work of Dr. Yasmin Said was supported in part by the National Institutes on Alcohol Abuse and Alcoholism under grant 1 F32 AA015876-01A1. The work of Dr. Edward Wegman was supported in part by the Army Research Office under contract W911NF-04-1-0447. The work of Dr. Said and Dr. Wegman was also supported in part by the Army Research Laboratory under contract W911NF-07-1-0059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health.

References

- Granovetter, M., 1973. The strength of weak ties. *Amer. J. Sociology* 78, 1360–1380.
- Rigsby, J.T., 2005. Block Models and Allegiance, Thesis submitted to George Mason University in partial fulfillment of the M.S. in Statistical Science.
- Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK.
- Wegman, E.J., Scott, D.W., Said, Y.H., 2006. Ad-hoc Committee Report on the 'Hockey Stick' Global Climate Reconstruction, A Report to Chairman Barton, House Committee on Energy and Commerce and to Chairman Whitfield, House Subcommittee on Oversight and Investigations: Paleoclimate Reconstruction.