Scientometrics (2016) 109:533–550
DOI 10.1007/s11192-016-1942-1

Bonded-communities in *HantaVirus* research: a research collaboration network (RCN) analysis

Sameer Kumar[1] · **Bernd Markscheffel**[2]

Received: 26 February 2016 / Published online: 7 April 2016

Abstract Hantavirus, one of the deadliest viruses known to humans, hospitalizes tens of thousands of people each year in Asia, Europe and the Americas. Transmitted by infected rodents and their excreta, Hantavirus are identified as etiologic agents of two main types of diseases—Hemorrhagic fever with renal syndrome and hantavirus pulmonary syndrome, the latter having a fatality rate of above 40 %. Although considerable research for over two decades has been going on in this area, bibliometric studies to gauge the state of research of this field have been rare. An analysis of 2631 articles, extracted from WoS databases on Hantavirus between 1980 and 2014, indicated a progressive increase ($R^2 = 0.93$) in the number of papers over the years, with the majority of papers being published in the USA and Europe. About 95 % papers were co-authored and the most common arrangement was 4–6 authors per paper. Co-authorship has seen a steady increase ($R^2 = 0.57$) over the years. We apply research collaboration network analysis to investigate the best-connected authors in the field. The author-based networks have 49 components (connected clump of nodes) with 7373 vertices (authors) and 49,747 edges (co-author associations) between them. The giant component (the largest component) is healthy, occupying 84.19 % or 6208 vertices with 47,117 edges between them. By using edge-weight threshold, we drill down into the network to reveal bonded communities. We find three communities' hotspots—one, led by researchers at University of Helsinki, Finland; a second, led by the Centers of Disease Control and Prevention, USA; and a third, led by Hokkaido University, Japan. Significant correlation was found between author's structural position in the network and research performance, thus further supporting a well-studied phenomenon that centrality effects research productivity. However, it was the PageRank centrality that out-performed degree and betweenness centrality in its strength of correlation with research performance.

✉ Sameer Kumar
sameer@um.edu.my

[1] Asia-Europe Institute, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Department of Economics, Technische Universität Ilmenau, 98693 Ilmenau, Germany

Keywords Research collaboration networks · Co-authorship networks · HantaVirus · Research communities · Hemorrhagic fever with renal syndrome (HFRS) · Hantavirus pulmonary syndrome (HPS)

Introduction

HantaVirus, transmitted to humans through persistently infected rodents and their excreta, is a global public health threat hospitalizing tens of thousands of people every year throughout the world. Since the isolation of the first hanta virus, HTNV (or Hantaan Virus) in 1976, several other Hantaviruses have been identified, with at least 22 being pathogenic to humans (Bi et al. 2008). One of the first major outbreaks of HantaVirus was reported from 1951 to 1954 when close of 3200 American soldiers serving in Korea became infected with the virus. In recent times, several cases have been reported in Asia (Zhang et al. 2004), US, and Europe. HantaVirus genus belongs to Bunyaviridae family and is identified as an etiologic agent of two different types of diseases—Hemorrhagic fever with renal syndrome or HFRS and hantavirus pulmonary syndrome (HPS). HFRS is also known by earlier names like Korean hemorrhagic fever (KHF), epidemic hemorrhagic fever (EHF), nephropathia epidemica (NE) (Bi et al. 2008). HPRS affects close to 1,50,000–2,00,000 people throughout the world each year while HPS infects just about 200. However, the fatalities caused by the latter are above 40 % when compared to 1–12 % in the case of HPRS depending on the severity of the virus (Lednicky 2003; Schmaljohn and Hjelle 1997). HFRS is more prevalent in the Eurasian region and HPS in the Americas. China remains the most endemic nation accounting for close to 70–90 % HFRS cases in the world (Zhang et al. 2004). NE, the mild form of HFRS, is most dominant in Western and Central Europe.

A quick glance at the Web of Science databases reveals a progressive increase in research papers on HantaVirus. The research in the field is paving the way to finding more pathogens, associated diseases, and vaccines. However, bibliometric studies to gauge HantaVirus research are surprisingly rare. Hence, we set out to mainly identify the prominent researchers in the research collaboration network and the bonded communities they were embedded in.

Research collaboration networks (RCN)

Research collaboration, a key mechanism that brings multiple talents together to accomplish a research task, could be effectively gauged through bibliometric records in research papers (Heinze and Kuhlmann 2008). Co-authorship in research papers has long remained the basis of investigating research collaborations (Beaver and Rosen 1978). The co-authors of a research paper could reveal the exchange of knowledge among researchers in their effort to bring out a published paper. Similarly, the affiliation details in the bibliometric records could be extrapolated to reveal collaboration happening at institutional and international levels.

Whether research collaboration could be gauged by just looking at the bibliometric records is a matter of academic debate (Katz and Martin 1997). For example, a collaboration could take place (i.e. through research advise) even if the two researchers do not finally end up penning the research paper together. Then there are some issues of honorary

and ghost authorships (Wislar et al. Wislar et al. 2011). While these concerns are serious, using bibliometric records is still the most concrete piece of evidence to establish a collaboration. Given the fact that co-authorship associations could also help us in understanding the association at institutional, organizational, and international levels, their significance cannot be overlooked.

The number of co-authored papers across disciplines has been growing over the years (Sonnenwald 2008). Better communications facilities, faster commuting, and industrialization brought in significant changes in the way research was conducted. Now there are researchers in large teams working on research projects and naturally these lead to published papers having significantly larger numbers of co-authors. Price (1963) calls these large lab-based research projects, 'big science'. However, big science research projects aren't generally in the social domain, that is, researchers do not have much choice to decide on their co-authors. Like the sciences, research conducted in the social sciences has also seen significant increase in the number of co-authors (Moody 2004). It remains imperative to note that collaboration is sharing of knowledge and may not always suggest improved quality of work. For example, in the humanities there are still a significant proportion of papers that are solo-written.

Research Collaborations could also be seen from the perspective of networks (Kumar and Jan 2014). In a network, two entities form a connection if there is some kind of association between them (Newman 2001). Using bibliometric data, these associations could be constructed to understand knowledge flows at multiple levels. Co-authorship in published papers is considered a reliable proxy to gauge research collaborations (Sonnenwald 2008; Melin and Persson 1996; Katz and Martin 1997). Social Network Analysis, an established research method to analyse social networks, is a set of mathematical algorithms that quantitatively analyse these relationships between nodes (Wasserman and Faust 1994). In a co-authorship network, for example, it could be applied to identify various patterns—i.e. the best connected nodes or key actors (Taba et al. 2015) or the communities that the researchers form through their associations. Specifically, these analyses reveal the pattern of network at both global and local levels. At the global level, the network pattern is seen from a whole network perspective, revealing, for example, the density, transitivity, scale-free pattern, small-world pattern, or the communities or clusters that the nodes form. At the local level, things are seen from the node perspective. Centrality is an important concept when looking from the perspective of a node and its context in the entire network. Centrality determines the relative importance (through centrality measures such as betweenness, closeness, and PageRank) and connectedness (through 'degree' metric) of nodes. Hence, those with higher centrality scores are those who are the most prominent players in the network. Another interesting aspect of social networks is that of the ties that the node is directly connected to. The strength of connection (depicted by a thicker line on a network graph) demonstrates a more frequent and stronger relationship than those that have an association of just a single or very few times. The idea of strength of relationship (Coleman 1988) is challenged by the notion of structural holes (Burt 1997). Structural holes theory postulates that the absence of ties in an ego network (network of ego—central node- and alters or immediate connections and those immediate connections connecting to one another) brings in more opportunities to the ego (the central node) as the ego then acts a bridge for the flow of resources between the 'alters'. Yet another idea of ties is postulated by the concept of 'weak ties'. The theory argues that in contrast to strong ties, which bring in trust, weak ties bring in new knowledge in the network. Growth and preferential attachment are the prime features of self-organising networks (Barabasi and Bonabeau 2003). Preferential attachment (Kumar and Jan 2015) is

defined by the preference of nodes (due to affinity or similarity) to attach to another node. In the context of co-authorship network, it may be due to the fact that one author connects to another author because he or she is a well-known researcher or has the same nationality as others. Preferential attachment causes some nodes to have much higher number of connections than most other nodes in the network. These hubs are kind of 'power houses' that tie together the network. This is the very reason why a self-organising network are small worlds (has a short path between any two random nodes) (Watts and Strogatz 1998). A targeted attack or absence due to some other reason could break the network down into pieces, which could severely affect the flow of resources in the network. Nonetheless, these self-organising networks are quite tolerant to random attacks (Albert et al. 2000).

Why do researchers collaborate? There are several benefits to collaboration (Beaver 2001). Sharing of expertise and division of work are among the most prominent. Collaboration also allows sharing of resources. For example, it is possible that certain equipment may not be available to certain researcher and collaborating with someone who has access to this equipment enables the conduct of research. Collaboration, due to division of labour, technically reduces the duration for the completion of research project, enabling researchers to publish more papers. Due to requirement for promotion and tenureships, which require papers to be published in high impact journals, collaboration does really help.

The research objectives

Our goals here are two pronged. First, we are interested in knowing the prominent and most connected authors in the field. A number of studies in recent times have found that the relative position significantly correlates with the research performance of researchers (Abbasi et al. 2011; Kumar and Jan 2013a). We want to check if this stands true for our (Hantavirus) dataset. However, another significant goal of this study is to detect the bonded communities of hantavirus research. With bonded communities we simply mean the cluster of researchers who interact more often with each other. A network of thousands of nodes otherwise only results in a hairball-like network that hardly provides much understanding or meaning.

Thus, in addition to common bibliometric analyses (i.e. annual paper production, average citations, top papers, number of papers per country, author research productivity, etc.), the present study has the following main objectives:

a. Investigate the prominent authors and the bonded-research communities clusters in Hantavirus research.
b. Investigate if there is relationship between players or actors structural position in the network and research productivity.

The study has significance as this would be perhaps one of the first studies to investigate research performance and bonded communities in hantavirus research from the perspective of research collaborations and networks. The idea of reaching out to bonded communities may be helpful to scientometricians wanting to get to the core of researchers who thickly interact with one another. They are the 'nucleus' or the real seat of knowledge of the network. Gauging and mapping of research performance of a crucial area such as hantavirus is of immense relevance and importance to health and research policy makers.

In addition, it attempts to understand if indeed the structural position in the network (i.e. the connectedness of actors in the network) has any significant correlation with research productivity. Such results would add to the existing body of knowledge about whether or not structural connectedness in a network does affect academic performance.

The rest of the paper is structured as follows: in the Material and Method section, we next discuss the data harvesting method and the keywords used to select the records. Subsequently, we discuss the findings and finally we draw our conclusions.

Materials and methods

Data harvesting

Records were harvested from the Web of Sciences databases from 1980 to 2014. Important hantavirus related keywords such as "hantavirus", "hantaan virus", "hemorrhagic fever with renal syndrome", "hantavirus pulmonary syndrome", "Korean hemorrhagic fever", "epidemic hemorrhagic fever" and "nephropathia epidemica" were used to refine the records selection.

Following search command was used:

> TOPIC: ("hantavirus" OR "hantaan virus" OR "hemorrhagic fever with renal syndrome" OR "hantavirus pulmonary syndrome" OR "Korean hemorrhagic fever" OR "epidemic hemorrhagic fever" OR "nephropathia epidemica"). Refined by: DOCUMENT TYPES: (ARTICLE). Timespan: 1980–2014. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH.

The above keywords search and data cleaning resulted in the final availability of 2631 records for analysis.

Data cleaning is an arduous task in bibliometric studies. Author name variations are among the most complicated as two or more authors may have same name and some even have the same institutional affiliation and hence their publications could be combined and shown as coming from a single author. On the other hand, an author may have different name variations and his or her publication may get split across these different name variations. At the institution and country levels, there is a need to make the names uniform. For example, in the present set of records, at the institution level, USA actually is an abbreviation of "US Army". In older data some of the country names are not mentioned, hence, by manual checking, they were appended. By manual checking much of these issues were resolved and errors minimised.

Methods

Social network analysis (Wasserman and Faust 1994) is a main research method applied in this study. As mentioned earlier, a network could be constructed when two entities are related in some way. On the graph, nodes are represented by a 'dot' and the connection between nodes, as a line passing between them. Hence, if two or more authors associate to co-write a research paper, the authors would be represented as nodes and the co-authored paper (the basis of relationship) is represented with a line passing between them. Nodes are also referred to as 'vertices' and relationship between them as 'edges'. It is obvious that just one representation of co-authored with dots and lines on a graph does not reveal much but when hundreds and (at times thousands) of papers are represented in a graph, an interplay of association is revealed and how seemingly invisible associations become visible. Data elements from the records are extracted and the co-authorship network constructed (see Fig. 1).

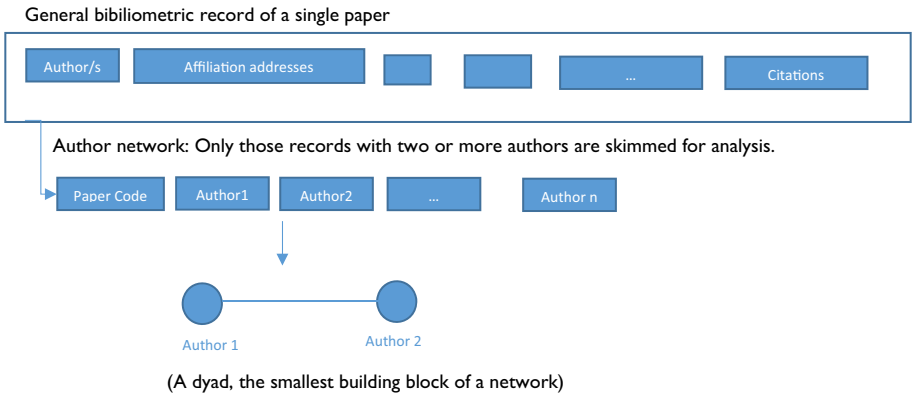

Fig. 1 The extraction of data elements from bibliographic records and construction of co-authorship network

Three centrality measures are calculated—The degree, betweenness centrality, and PageRank centrality. We also calculate the local clustering coefficient and the average geodesic distance of the network. We have not calculated closeness centrality as this centrality gives accurate results for one component (typically a giant component) at a time. It tends to give misleading results if the calculation is made for all the components in the network (for example, those in the dyadic network will have high closeness centrality that those nodes with high degree in the main component). Since we are interested in all the authors in the network (and not just those in the giant component), we have chosen to leave out closeness centrality in our graph metrics calculations.

Degree, a popularity measure, is simply the number of direct connections a node has. Betweenness centrality is path-based and checks how much 'in-between' a node is in the network. Those with high betweenness centrality have positional advantage and work as bridges between communities. Removal of these nodes could severely affect the flow of resources in the network. A PageRank measure is a prestige metric that not only checks the number of connection a node has but also the number of connection of alters.

The mathematical formulae used to calculate are standard and are thus provided in "Appendix".

The centrality values are then correlated using MS-Excel's correlation statistical function, with number of papers produced and citations accumulated, to check if there is any significant association between the two.

NodeXL (Smith et al. 2009) was used to calculate graph metrics and visualize the network diagrams.

Results and analysis

Research productivity

The yearly paper production shows an upward trend. The worldwide alarm raised by the deadly virus has had researchers looking for the pathogens, its geographical reach, and its potential cure. From just four related papers published in 1980, the number grew to 180 in

2014. A linear trendline ($R^2 = 0.93$) shows a good-fit, meaning that the growth in paper production on hantavirus has been steady over the years (Fig. 2).

However, a large proportion of paper production has been concentrated in certain regions of the world. Majority of the research is going on in Europe and the USA (see Fig. 3) When contrasted with the actual occurrence of Hantavirus infection cases (see Fig. 4), we find that China (although a distant second in terms of number of paper produced) is probably doing comparatively much less research when compared to the number of hantavirus cases reported from the region. As mentioned earlier, China accounts for close to 70–90 % of all hanta virus cases in the world. The top ten countries in terms of research productivity are, USA (1003 papers), Peoples Republic of China (271), Finland (255), Sweden (235), Germany (226), Japan (161), South Korea (143), Argentina (116), France (110) and Belgium (105).

Sixty eight percentage of the institutions (or 1566 institutions of the total of 2305 institutions) have just contributed one publication to the dataset. Similarly, 295 institutions have contributed two publications and 145 institutions have contributed three publications. A power law pattern is seen here where a large number of institutions (about 87 %) have contributed just 1–3 publications each and few institutions (299 institutions or about 13 %) have contributed four or more publications. Ten institutions have over 50 papers each. The highest number of institutions doing research on HantaVirus is located in the USA (411 institutions), followed by 246 institutions in China. Although most universities names are distinct, some centres names (i.e. Ctr Dis Control & Prevent) are generic and are also found in other countries around the world. To reduce duplication to the minimum, we thus follow

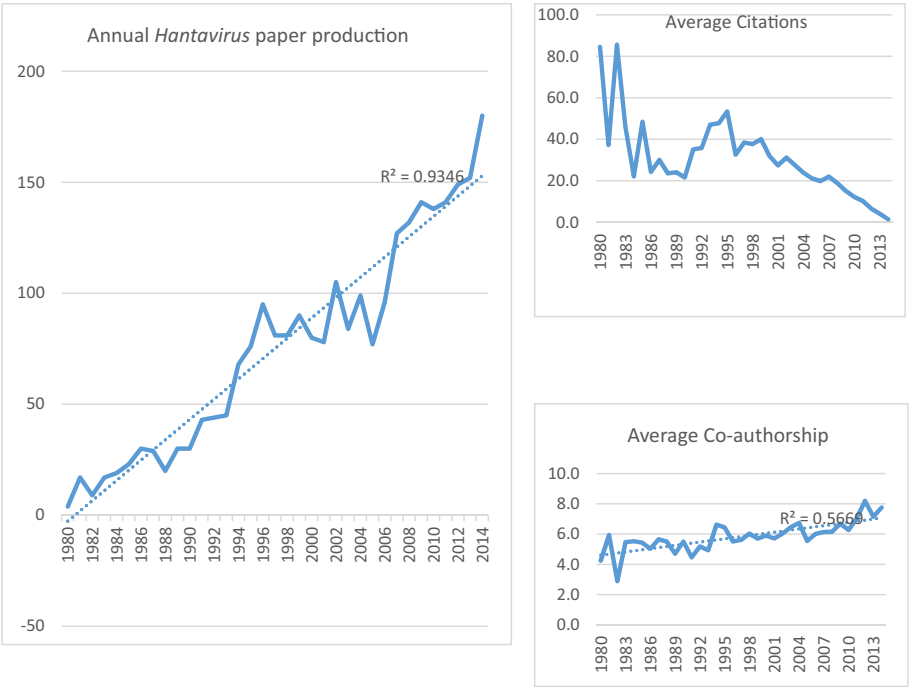

Fig. 2 The annual paper production, average citations received per paper annually and average co-authorship per paper annually of *HantaVirus* research

Fig. 3 Geographical depiction of research productivity (drawn using https://gpsvisualizer.com)

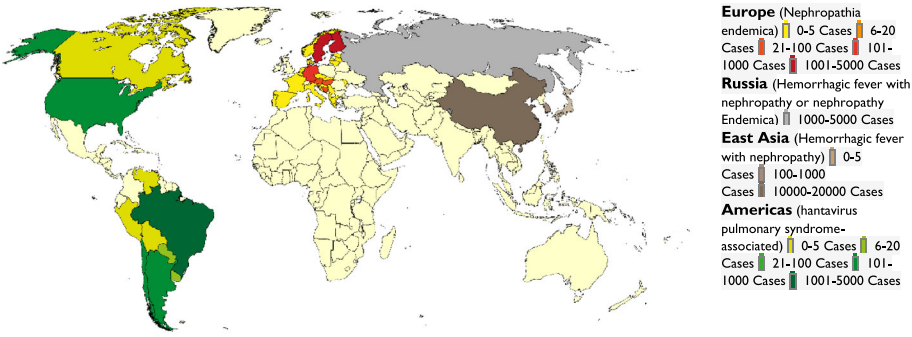

Fig. 4 Worldwide distribution and approximate incidence per country per year (if known) of hantavirus infections (based on data derived from data as in Jonsson et al. 2010). Figure as in https://commons.wikimedia.org/wiki/File:Hantaviren_weltweit.svg available in the public domain)

WoS nomenclature for all institutional names. In terms of the number of papers produced, University of Helsinki ranks the highest with 219 papers, followed by University of New Mexico (194 papers), Center for Disease Control and Prevention (173 papers), US Army (142 papers) and Karolinska Institute (128 papers). However, in terms of citations it is the Center for Disease Control and Prevention (8443 citations) that has the highest citations count, closely followed by US Army (7648 citations) and University of New Mexico (7301 citations).

The entire publication base of hantavirus's 2631 papers received a total of 58,078 citations or an average of 21.09 citations per paper. These are good averages and indicate sound 'health' of research in the field. The papers written have a downward trajectory in terms of average citations received—those papers that have been written earlier are cited significantly more than those that are published in later years (see Fig. 2). This is of course

practical as papers that are written earlier have stayed in the knowledge base for a much longer time than the recent ones and thus have more opportunity to get more cited. Some also get a chance (depending on its influence to the field) to enter the very 'seminal knowledge'. Once these papers are in this select group, they are cited considerably more than the rest of the papers. 1333 papers published during 2005–2014 time frame were cited 11.39 times, compared to 1298 papers published older time period during 1980–2004 that were cited 33.10 times on average. 92 % papers in the older time period (1980–2004) had received at least one citations when compared to 85.67 % in the newer timeframe (2005–2014).

Author productivity

Of the 7426 authors, a large proportion or 5152 authors (69.37 %) have produced just one paper. 1034 authors (13.92 %) have produced two papers each, 1230 authors (16.56 %) have produced 3 papers each, 213 authors (2.86 %) 4 papers each, and 617 authors (8.30 %) 5 papers and above. 19 authors are highly productive and have produced 50 papers and more. Vaheri A (160 papers), Lundkvist A (134), Plyusnin A (105), Arikawa J (103) and Hjelle, B (94) are the most productive authors in the dataset.

In our dataset, 701 authors have received no citations, 2608 authors had between 1 and 10 citations each, 3298 authors had between 11 and 100 citations each and the rest (819 authors) have 100 citations and more. 44 authors had 1000 and more citations each with Peters Cj (6671 citations), Ksiazak, Tg (6048), Vaheri A (5969), Lundkist A (4612), Rollin Pe (4590) garnering the top five slots as the most cited authors.

In both number of papers/author and citations/author, we notice few authors have been significantly more productive than the rest of the block, a common feature of research productivity in most disciplines. Of 428,546 cumulative citations (if a paper has four co-authors and has received ten citations for the paper, cumulative citations for the authors would be ten for each author) by authors, 350,588 citations (or 81.80 %) are garnered by top 20 % of authors, thus, almost fitting 80/20 rule or power law.

Collaboration per paper (number of authors per paper)

As noted earlier, there is a whole host of research that has shown that the co-authorship in paper across disciplines has gone up especially in the last two decades. An analysis of co-authorship (or average number of co-authors on each paper) of our dataset shows that the same is true for publications in the field. However, there hasn't been a striking increase in the number of co-authors in the two time periods—1980–2004 timeframe had an average of 5.29 authors per paper when compared to 6.82 in the time period between 2005 and 2014. About 95 % papers were co-authored (or had at least two authors on a paper). The most common arrangement was 4–6 authors per paper. There were 333 4-author papers, 326 5-author papers and 332 6-author papers. More authors per paper are symbolic of experimental research. Two papers had 67 and 86 co-authors respectively.

Top papers

Table 1 shows the list of top ten most cited papers in hantavirus research. The most cited paper is the year 1993 paper by Nichol et al. (1993) that was published after the outbreak of HantaVirus in the four corners region of the United States. Their study showed that the

Table 1 Top ten highest cited papers on HantaVirus

Title	Authors	Journal	Publication year	Total times cited
1 Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness	Nichol S.T, Spiropoulou C.F, Morzunov S, Rollin P.E, Ksiazek T.G, Feldmann H, et al.	Science	1993	782
2. Hantaviruses: a global disease problem	Schmaljohn C, Hjelle B.	Emerging Infectious Diseases	1997	653
3. Hantavirus pulmonary syndrome—a clinical description of 17 patients with a newly recognized disease	Duchin J.S, Koster F.T, Peters C.J, Simpson G.L, Tempest B, Zaki S.R, et al.	New England Journal of Medicine	1994	450
4. Hantavirus pulmonary syndrome—pathogenesis of an emerging infectious-disease	Zaki S.R, Greer P.W, Coffield L.M, Goldsmith C.S, Nolte K.B, Foucar K, et al.	American Journal of Pathology	1995	422
5. Antigenic and genetic properties of viruses linked to hemorrhagic-fever with renal syndrome	Schmaljohn C.S, Hasty S.E, Dalrymple J.M, Leduc J.W, Lee H.W, Vonbonsdorff C.H, et al.	Science	1985	385
6. Serologic and genetic identification of peromyscus-maniculatus as the primary rodent reservoir for a new hantavirus in the southwestern united-states	Childs J.E, Ksiazek T.G, Spiropoulou C.F, Krebs J.W, Morzunov S, Maupin G.O, et al.	Journal of Infectious Diseases	1994	377
7. Factors in the emergence of infectious-diseases	Morse S.S	Emerging Infectious Diseases	1995	358
8. Nephropathia epidemica—detection of antigen in bank voles and serologic diagnosis of human infection	Brummerkorvenkontio M, Vaheri A, Hovi T, Vonbonsdorff C.H, Vuorimies J, Manni T, et al.	Journal of Infectious Diseases	1980	333
9. Isolation of hantaan virus, the etiologic agent of korean hemorrhagic-fever, from wild urban rats	Lee H.W, Baek L.J, Johnson K.M	Journal of infectious diseases	1982	238
10. Beta(3) integrins mediate the cellular entry of hantaviruses that cause respiratory failure	Gavrilovskaya I.N, Shepley M, Shaw R, Ginsberg M.H, Mackow E.R	Proceedings of the National Academy of Sciences of the United States of America	1998	232

comparison of the human and rodent sequences had a direct genetic link between the virus in infected rodents and infected human 'hantaviral ARDS' cases. On the heels of this study was another highly cited paper by Duchin et al. (1994) who carried out clinical, lab and other analyses on 17 persons infected by newly recognised strain of hantavirus. Their study concluded that the new strain of virus causes HPS. A high fatality rate (76 %) was also reported. One of the earliest studies on disease caused Hantavirus (Nephropathia Epidemia

or NE) and highly cited paper is that of Brummerkorvenkontio et al. (1980). Their study concluded that the detection of NE antigen in rodents (bank voles) facilitates 'specific serologic diagnosis of NE'.

The heat map drawn using Vosviewer (Van Eck and Waltman 2010) in Fig. 5 shows the scientific landscape based on co-citations. Co-citation analysis looks at the relatedness of items based on the number of times they are cited together. We use author (first author only) co-citations for the analysis. Visualization is automatically done by the software after a threshold is provided by the user. Papers Nichol st, 1993 (Science) and Schmalijohn C, 1997 (Emerging infectious diseases) and Lee hw, 1978 (Journal of infectious diseases) are among the most influential papers.

The correlation between structural position of authors in co-authorship networks and research performance

In this section, we investigate if the connectedness and relative position of authors have effect on the research performance and then analyze bonded communities embedded in co-authorship networks. The process of the construction of network is explained in the Materials and Methods section.

The author-based networks have 49 components (connected clump of nodes) with 7373 vertices and 49,747 edges between them (see Fig. 6). The giant component (the largest component) is healthy occupying 84.19 % or 6208 vertices and 47,117 edges between them. A healthy giant component may be an indication of frequent collaborative activity. The giant component is considered the seat of main activity in the research community (Fatt et al. 2010). Knowledge flows in such networks are faster as they are not subject to disruptions which otherwise would have been the case had the giant component been small and the whole network having several small fragments of components. A recent study by Liu and Xia (2015) found that the development of an inter-disciplinary field is

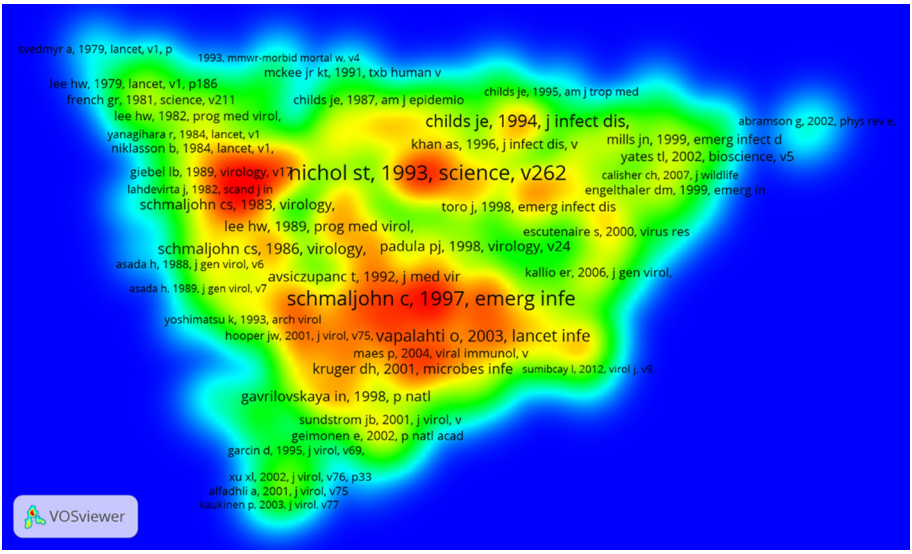

Fig. 5 Heatmap showing co-citation of cited references

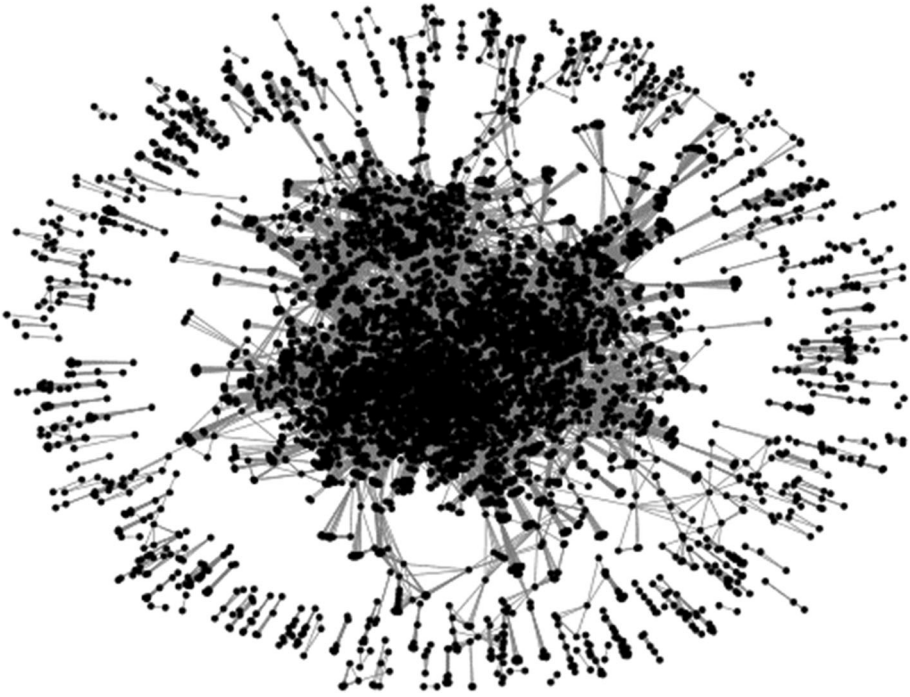

Fig. 6 The overall co-authorship networks of hantavirus dataset (drawn with Fruchterman–Reingold on repulsive force between vertices 4.0 and iterations per layout 50 force directed layout). The darker clump at the center is symbolic of those nodes that are highly connected

characterised by the growth of a giant component of collaboration network, evolving from small clusters to, what they call, 'chained communities', to a more developed giant component that has typical small-world properties.

The giant components continue to grow as more components connect to it. After all, it takes just an edge from a disconnected component to connect to the giant component, thus, making the latter bigger in size. The average geodesic distance (shortest distance between any two random nodes in the network) between the nodes is just 4.15 meaning that, on average, two random authors in the hantavirus dataset are just about four hops away from one another. This is another indication that the authors are closely knit and resource flow and delivery would be faster in this network when compared to networks that are sparse and fragmented. This also confirms the *small-world* nature of this network (Newman 2001). Small world networks typically have shorter geodesic distances.

The centrality values (Degree, Betweenness and PageRank) of authors makes Hjelle B. (Brian Hjelle) the most connected author in the Hantavirus research community (see Table 2). Dr. Brian, a pathologist, is currently the MD/Ph.D program director at the University of New Mexico (USA) and has several awards and recognitions to his credit. He has been conducting research on hantavirus since the 1990s and was also a member of the Hantavirus Pulmonary Syndrome Clinical Trial Committee for the National Institute of Allergy and Infectious Diseases, National Institutes of Health (Collaborative Antiviral Study Group) from 1993–1996 (http://pathology.unm.edu/faculty/faculty/CVs/brian-hjelle.pdf).

Table 2 Centrality measures of authors in the co-authorship network

Authors	Degree	Betweenness centrality	PageRank	Clustering coefficient	No. of co-authored works	Times_cited (on co-authored papers)
Hjelle B	458	3,149,982.76	26.063,417	0.05502948	94	3827
Ksiazek T.G	454	1,485,899.336	21.0358	0.06754772	81	6048
Peters C.J	395	1,596,207.965	17.407885	0.08403264	80	6671
Vaheri A	381	1,456,253.834	23.147631	0.0389004	160	5969
Lundkvist A	373	2,361,250.556	22.232698	0.04102165	134	4612
Rollin P.E	349	509,527.8914	15.174872	0.09994072	59	4590
Arikawa J	289	856,254.7233	14.737556	0.06247597	103	2176
Nichol S.T	273	609,138.5441	14.07499	0.06097824	54	4456
Yoshimatsu K	262	633,352.1306	12.852233	0.07288468	89	1596
Kruger D.H	261	836,905.9949	14.737569	0.05770704	76	1958
Plyusnin A	260	1,142,236.121	15.110315	0.05402435	105	3088
Mills J.N	227	451,290.9653	12.515917	0.07258976	60	2610
Vapalahti O	221	385,782.2516	13.556483	0.06322501	85	2988
Ulrich R.G	207	551,752.1708	10.14514	0.10552976	32	305
Zaki S.R	189	184,382.5531	7.146795	0.25143533	24	1991

The local clustering coefficient provides an interesting picture—those with high degree have low clustering coefficient (correlation −0.404). Why is this the case? Clustering coefficient or transitivity is a measure of prediction that if nodes B and C have common partner A, it is a likelihood that B would eventually connect with C. We surmise that this is due to the fact that a node or ego with many alters, will likely have alters that have less connections among them. This is true in many occasions as the ego with large connections would have these connections from several diverse set of nodes.

Several studies in the recent years have found that centrality measures indeed have significant effect on research performance (Abbasi et al. 2011; Uddin et al. 2012). Hence we set out to investigate if centrality measures have effects on research performance in the dataset of hantavirus research, too. Our correlation test (see Table 3) confirms that indeed in hantavirus datasets there is a significant correlation ($p < 0.01$) between centrality

Table 3 Correlation test between centrality measures and academic performance

	Degree	Betweenness centrality	PageRank	Number of papers	Times cited
Degree	1				
Betweenness	0.719213*	1			
PageRank	0.899933*	0.82809*	1		
Number of papers	0.791788*	0.790944*	0.914009*	1	
Times_cited	0.793115*	0.738517*	0.821499*	0.825798*	1

* Significant at $p < 0.01$

measures (or how well the author is connected with others in the community) and research performance.

However, what stands out is the correlation of Pagerank with research performance. Its correlation coefficient strength with research performance demonstrates its efficacy that is even higher than the well-known measures such as degree and betweenness centrality. The very fact the PageRank is based not just on the connections an author has but the quality of these connections, provides it with a better predictability for research performance.

Detecting bonded-communities

Here we also introduce an idea to detect 'bonded communities'. By increasing the threshold of edge-weight between nodes, a research community could be drilled down to a level where those nodes that frequently interact with one another are revealed. The importance of strength or 'bondedness' needs attention as this may provide new insights into the communities lying within. Drilling down to the desired core (we call it as 'edge-core') is done by progressively increasing edge-weight, till the most bonded communities become visible—it could happen with just three or four in sparse communities and could be ten or more in dense communities.

When our network is reduced and visualized with edge weight ten (edge-core-ten) network (the network only visualizes nodes that have an edge weight of ten and more between them), three distinct 'bonded' communities emerge. Authors (or nodes) in these

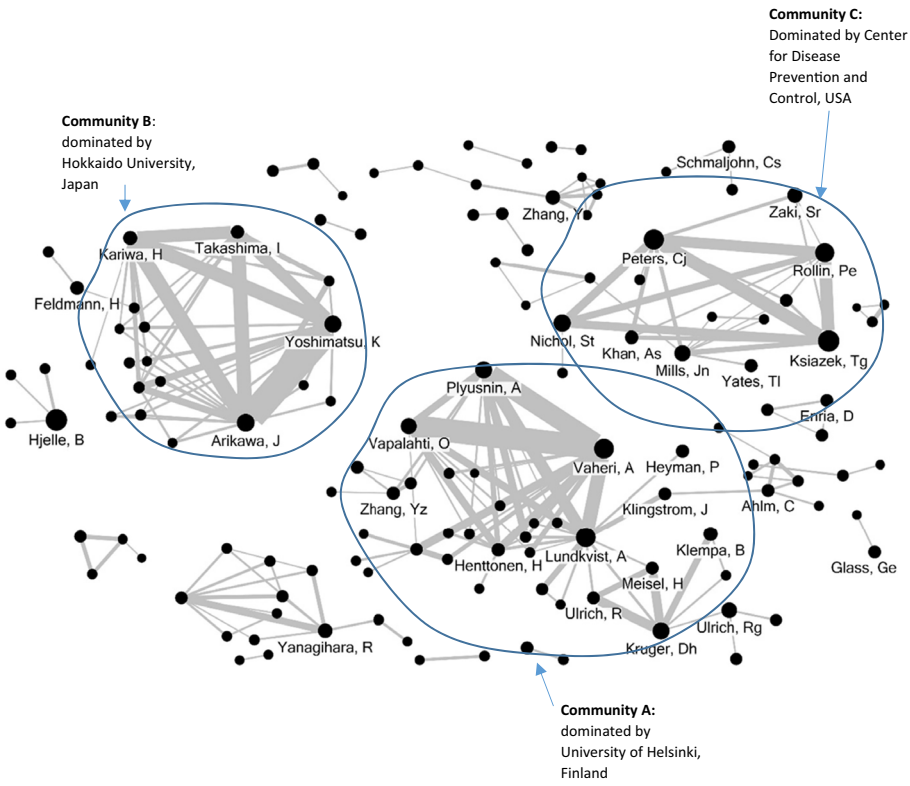

Fig. 7 The bonded communities in co-authorship networks

communities are involved in repeat associations with one another (see Fig. 7). Quite interestingly, at a threshold of edge-weight ten and above, the community of Hjelle B, the most connected author, becomes isolated. This probably goes to show that even best connected author/s may not be embedded in bonded-communities.

Community A, is led by Vaheri A (Vaheri Antti) who has co-authored with Vapalahti O (Vapalahti Olli) (76 times), Plyusnin A (Plyusnin, Alexander) (65 times) and Lundkvist A (Lundkvist Ake) (48 times). Antti Vaheri works at Dept of Virology at University of Helsinki, Finland and has been active since the 1980s. As a matter of fact, his papers, (Brummerkorvenkontio et al. 1980; Schmaljohn et al. 1985) are among the most cited HantaVirus related papers. Olli Vapalahti and Alexander Plyusnin, too, are associated with University of Helsinki while Ake Lundkvist is associated with Karolinska Institute, Sweden. While Community A has Japanese authors, this community has European authors and is dominated by scholars from University of Helsinki. Within the edge-core-ten community, Ake Lundkvist is the author with the highest betweenness. He is a bridge node connecting to the sub-community of Germany-based authors—Ulrich R, Meisel H, Kruger Dh and Klempa B.

Community B that has prominent authors Arikawa J (Arikawa Jiro), Yoshimatsu K (Yoshimatsu Kumiko), Takashima I (Takashima Ikuo) and Kariwa H (Kariwa Hiroaki) are all from Japan's Hokkaido University. Being from the same institution also provides the necessary geographical proximity to carry out joint research.

Community C has prominent authors Ksiazek TG (Ksiazek, Thomas G); Rollin PE (Rollin, Pierre E), Nichol ST (Nichol, Stuart T), Peters CJ (Peters, Clarance James), Zaki Sr and Khan As, all associated with Center for Disease Control & Prevention, Atlanta, USA. Another prominent author Mills, JN (Mills, James N) is associated with Emory University, Atlanta, USA. Clarence James Peters is an accomplished physician who has a well-cited book (Peters and Olshaker 1997), while Zaki Sr has, to his credit, papers (Duchin et al. 1994; Zaki et al. 1995) that are among the top ten most cited papers on hantavirus. However, Peters CJ, Zaki Sr and Khan As have not published (in the dataset) after 2007, 2002 and 2004, respectively. As we see, the community is dominated by authors from the Centers for Disease Control and Prevention and all the prominent authors are stationed in Atlanta, which again shows that geographical proximity is an important factor for deep-bonded association.

A note on institutional and international association

University of Helsinki, Karolinska Institute, and Swedish Inst of Infectious Disease Control dominate the institutional collaborations in Europe. At the same time, the Centers for Disease Control and Prevention and University of New Mexico have a sustained and bonded relationship within USA. Based on collaboration among institutions contributing at least ten or more research papers, University of New Mexico has the maximum degree (collaborating with 51 institutions), followed by University of Helsinki (41), Centers for Disease control and Prevention (36), and Karolinska Institute (32). All the prominent authors as discussed also belong to these Institutions. In the same stride, we thus see Sweden and Finland involved with extensive collaboration (51 repeat associations) in Europe while USA almost controls international collaboration with majority of countries including Argentina (40 repeat associations), South Korea (40), and Peoples Republic of China (32). Germany has a fair share of collaboration with Sweden and Slovakia.

Concluding remarks

Here we scientometrically analysed the research landscape of HantaVirus research. By network reduction or by drilling down into the network based on the strength of ties (or edge weights), we revealed the communities that thickly interact with one other. We demonstrate that these bonded communities actually capture the most prominent authors, too. In our opinion, these bonded communities are the core or "central brain" of the network where central activity takes place. We also theorize that strength of relationship is an equally important criterion (apart from centrality measures) for sustainable research performance. PageRank stands out in its correlation with the research performance which further substantiate the idea that it is not only the number of other authors an author is connected to but the quality of these authors (how well those co-authors are connected) that ensures research visibility.

Acknowledgments Part of the analysis of this study was completed during S.K's research visit to TU-Ilmenau, Germany. The study is supported by High Impact Research, University of Malaya, Grant number UM.C/625/1/HIR/MOHE/SC/13/3.

Appendix

SNA measures (Kumar and Jan 2013a, b).

A *component* is a set of nodes joined in such a way that any single random node in the network could reach out to any other random node by "…traversing a suitable path of intermediate collaborators"(Newman 2004).

Clustering coefficient, C, is also known as 'transitivity' and more accurately as the 'fraction of transitive triples'(Wasserman and Faust 1994). Mathematically, clustering coefficient is calculated as:

$$C = \frac{3 \times \text{no. of triangles}}{\text{no. of connected triples}} \tag{1}$$

where the number of triangles represents trios of nodes in which each node is connected to both others, and connected triples represent trios of nodes in which at least one node is connected to both others (Barabasi et al. 2002; Newman 2004).

Degree is the most common and probably the most effective centrality measure to determine both the influence and importance of a node. A degree is simply the number of edges incident on the vertex. Mathematically, degree k_i of a vertex is

$$k_i = \sum_{j=1}^{n} g_{ij} \tag{2}$$

where $g_{ij} = 1$ if there is a connection between vertices i and j and $g_{ij} = 0$ if there is no such connection.

(Otte and Rousseau 2002).

Betweenness centrality of a vertex i is the fraction of geodesic paths that pass through i, which could be mathematically represented as

$$b(i) = \sum_{j,k} \frac{m_{jik}}{m_{jk}} \quad (3)$$

where m_{jk} is the number of geodesic paths from vertex j to vertex $k(j, k \neq i)$ and m_{jik} is the number of geodesic paths from vertex j to vertex k, passing through vertex I (Otte and Rousseau 2002; Linton 1977).

PageRank is an importance measure that is calculated based on the premise that 'having links to page p from prominent pages, is a good indication that page p is important one too' (Page et al. 1999).

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics, 5*(4), 594–607.

Albert, R., Jeong, H., & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature, 406*(6794), 378–382.

Barabasi, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American, 288*(5), 60–69.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica a-Statistical Mechanics and Its Applications, 311*(3–4), 590–614.

Beaver, D. B. (2001). Reflections on scientific collaboration, (and its study): Past, present, and future. *Scientometrics, 52*(3), 365–377.

Beaver, D. B., & Rosen, R. (1978). Studies in scientific collaboration. *Scientometrics, 1*(1), 65–84.

Bi, Z. Q., Formenty, P. B. H., & Roth, C. E. (2008). Hantavirus infection: A review and global update. *Journal of Infection in Developing Countries, 2*(1), 3–23.

Brummerkorvenkontio, M., Vaheri, A., Hovi, T., Vonbonsdorff, C. H., Vuorimies, J., Manni, T., et al. (1980). Nephropathia epidemica—Detection of antigen in bank voles and serologic diagnosis of human infection. *Journal of Infectious Diseases, 141*(2), 131–134.

Burt, R. S. (1997). The contingent value of social capital. *Administrative Science Quarterly, 42*(2), 339–365.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*, 95–120. doi:10.1086/228943.

Duchin, J. S., Koster, F. T., Peters, C. J., Simpson, G. L., Tempest, B., Zaki, S. R., et al. (1994). Hantavirus pulmonary syndrome—A clinical description of 17 patients with a newly recognized disease. *New England Journal of Medicine, 330*(14), 949–955. doi:10.1056/nejm199404073301401.

Fatt, C. K., Abu Ujum, E., & Ratnavelu, K. (2010). The structure of collaboration in the Journal of Finance. *Scientometrics, 85*(3), 849–860. doi:10.1007/s11192-010-0254-0.

Heinze, T., & Kuhlmann, S. (2008). Across institutional boundaries? Research collaboration in German public sector nanoscience. *Research Policy, 37*(5), 888–899. doi:10.1016/j.respol.2008.01.009.

Jonsson, C. B., Figueiredo, L. T. M., & Vapalahti, O. (2010). A global perspective on hantavirus ecology, epidemiology, and disease. *Clinical Microbiology Reviews, 23*(2), 412–441. doi:10.1128/cmr.00062-09.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1–18.

Kumar, S., & Jan, J. M. (2013a). Mapping research collaborations in the business and management field in Malaysia, 1980–2010. *Scientometrics, 97*(3), 491–517. doi:10.1007/s11192-013-0994-8.

Kumar, S., & Jan, J. M. (2013b). On giant components in research collaboration networks: Case of engineering disciplines in Malaysia. *Malaysian Journal of Library and Information Science, 18*(2), 65–78.

Kumar, S., & Jan, J. M. (2014). Research collaboration networks of two OIC nations: Comparative study between Turkey and Malaysia in the field of 'Energy Fuels', 2009–2011. *Scientometrics, 98*(1), 387–414. doi:10.1007/s11192-013-1059-8.

Kumar, S., & Jan, J. M. (2015). The assortativity of scholars at a research-intensive university in Malaysia. *The Electronic Library, 33*(2), 162–180.

Lednicky, J. A. (2003). Hantaviruses—A short review. *Archives of Pathology and Laboratory Medicine, 127*(1), 30–35.

Linton, C. F. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41.

Liu, P., & Xia, H. X. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics, 103*(1), 101–134. doi:10.1007/s11192-014-1525-y.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics, 36*(3), 363–377.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213–238.

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

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5200–5205. doi:10.1073/pnas.0307545100.

Nichol, S. T., Spiropoulou, C. F., Morzunov, S., Rollin, P. E., Ksiazek, T. G., Feldmann, H., et al. (1993). Genetic identification of a hantavirus associated with an outbreak of acute respiratory Illness. *Science, 262*(5135), 914–917. doi:10.1126/science.8235615.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441–453.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120.

Peters, C. J., & Olshaker, M. (1997). Virus hunter: thirty years of battling hot viruses around the world. New York: Anchor Books.

Price, D. S. (1963). *Big science, little science*. New York: Columbia University.

Schmaljohn, C. S., Hasty, S. E., Dalrymple, J. M., Leduc, J. W., Lee, H. W., Vonbonsdorff, C. H., et al. (1985). Antigenic and genetic properties of viruses linked to hemorrhagic-fever with renal syndrome. Science, 227(4690), 1041-1044. doi:10.1126/science.2858126.

Schmaljohn, C., & Hjelle, B. (1997). Hantaviruses: A global disease problem. *Emerging Infectious Diseases, 3*(2), 95–104.

Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., et al. (2009). Analyzing (social media) networks with NodeXL. In *Proceedings of the fourth international conference on communities and technologies* (pp. 255–264): ACM.

Sonnenwald, D. H. (2008). Scientific collaboration. *Annual review of information science and technology, 41*(1), 643–681.

Taba, S. T., Hossain, L., Atkinson, S. R., & Lewis, S. (2015). Towards understanding longitudinal collaboration networks: A case of mammography performance research. *Scientometrics, 103*(2), 531–544. doi:10.1007/s11192-015-1560-3.

Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. *Scientometrics, 90*(2), 687–699. doi:10.1007/s11192-011-0511-x.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Wasserman, S., & Faust, K. (1994). *Social network analysis, methods and applications* (1st edition, structural analysis in the social sciences). New York: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

Wislar, J. S., Flanagin, A., Fontanarosa, P. B., & DeAngelis, C. D. (2011). Honorary and ghost authorship in high impact biomedical journals: A cross sectional survey. *British Medical Journal,*. doi:10.1136/bmj.d6128.

Zaki, S. R., Greer, P. W., Coffield, L. M., Goldsmith, C. S., Nolte, K. B., Foucar, K., et al. (1995). Hantavirus pulmonary syndrome: Pathogenesis of an emerging infectious-disease. *American Journal of Pathology, 146*(3), 552–579.

Zhang, Y., Xiao, D., Wang, Y., Wang, H., Sun, L., Tao, X., et al. (2004). The epidemic characteristics and preventive measures of hemorrhagic fever with syndromes in China. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi, 25*(6), 466–469.