CEB Working Paper

A Vietnamese co-authorship network in social sciences: some graphical presentation of 2008-2017 Scopus data

Tung Manh Ho, Ha Viet Nguyen, Thu-Trang Vuong, Quang-Minh Dam, Hiep-Hung Pham and Quan-Hoang Vuong

The application of social network analysis in studying science collaboration has not been used for studying science activities in Vietnam although collaboration is popular among Vietnamese scientists. This paper employs network visualization together with basic network measures to explore characteristics of the network of 412 Vietnamese social scientists whose papers can be found indexed in the Scopus database. Early results show that the network's connections are very sparse, with only 0.52% density, but the clustering coefficient is very high, 58.64%, suggesting that the dissemination of scientific knowledge and expertise in the network is not very efficient. Secondly, the disparity in the levels of connection among individual researchers in the network indicates that it would easily fall apart if a few highly-connected nodes are removed. Finally, upon zooming in on the two largest components of the network, the study found that their characteristics differ from the whole and both of them are led by the most productive researchers who also have the most connections.

Keywords: Social network analysis, science collaboration, network characteristics, research output

JEL Classifications: I23, O32

CEB Working Paper N° 17/020
August 2017

ULB

A Vietnamese co-authorship network in social sciences:

some graphical presentation of 2008-2017 Scopus data

Tung Manh Ho[1,2]; tung.ho@wu.edu.vn

Ha Viet Nguyen[2]; ha.nguyen2@wu.edu.vn

Thu-Trang Vuong[2,3]; Email: thutrang.vuong@sciencespo.fr

Quang-Minh Dam[2]; minh@wu.edu.vn

Hiep-Hung Pham[2]; hiep.pham@wu.edu.vn

Quan-Hoang Vuong[4,2,*]; qvuong@ulb.ac.be

[1] Institute of Philosophy, Vietnam Academy of Social Sciences; No. 59 Lang Ha, Dong Da District, Hanoi, Vietnam.

[2] Centre for Interdisciplinary Social Research, Western University Hanoi (ĐH Thành Tây); Yên Nghĩa Ward, Ha Dong District, Hanoi, Vietnam.

[3] Campus Européen de Dijon, Sciences Po Paris; 14 Avenue Victor Hugo, 21000 Dijon, France.

[4] Centre Emile Bernheim, Université Libre de Bruxelles; 50 F.D. Roosevelt, Brussels 1050, Belgium.

[*] Corresponding author.

Abstracts

The application of social network analysis in studying science collaboration has not been used for studying science activities in Vietnam although collaboration is popular among Vietnamese scientists. This paper employs network visualization together with basic network measures to explore characteristics of the network of 412 Vietnamese social scientists whose papers can be found indexed in the Scopus database. Early results show that the network's connections are very sparse, with only 0.52% density, but the clustering coefficient is very high, 58.64%, suggesting that the dissemination of scientific knowledge and expertise in the network is not very efficient. Secondly, the disparity in the levels of connection among individual researchers in the network indicates that it would easily fall apart if a few highly-connected nodes are removed. Finally, upon zooming in on the two largest components of the network, the study found that their characteristics differ from the whole and both of them are led by the most productive researchers who also have the most connections.

Keywords: Social network analysis, science collaboration, network characteristics, research output

J.E.L. Codes: I23, O32

#

A Vietnamese co-authorship network in social sciences:

some graphical presentation of 2008-2017 Scopus data

I.   Introduction

In the beginning of 2017, the Vietnamese public was once again disappointed to find out there was no Vietnamese universities in the Times Higher Education's ranking of the top 300 universities in Asia. There was no shortage of experts' attempts to explain this disappointing situation; many pointed to the fact that Vietnamese universities have not put enough focus on research (Hong & Le, 2017). Being aware of the demand for improving research capacity, the Ministry of Education and Training has recently issued many policies and proposals addressing the issue head-on. Figuring among the many efforts is the issuance of circular No. 08/2017/TT-BGDĐT on 14 April, 2017 mandating doctoral students must have papers published in Scopus and Web of Science-indexed journals, the doctoral dissertation instructors must also have international publications. There has also been a proposal to mandate that candidates for the titles of Professor and Associate Professor must have international publications. Although these changes and proposals were met with both excitement and dread by the public, it is noteworthy that those who criticize the new regulations do not argue against the changes. Rather, their main concern is the timeline of adopting these policies: whether these changes are too abrupt. In other words, people on both sides of the arguments express their desire to improve research capacity in Vietnam, and only disagree on how.

Thus arose a burning question: How to increase the quantity and quality of scientific publications in Vietnamese social sciences? This scientific question can be answered with quantitative methods. However, the field of quantitative research on scientific activities and research policy in Vietnam is still nascent. Even though there have been several studies on the status of scientific publications in Vietnam as a whole, none has been carried out with a sole focus on social sciences – a field often criticized of low productivity (Pham & Vuong 2017; Vuong et al. 2013). In addition, the technique of social network analysis is yet to be applied in the case of Vietnam, despite its success in explaining and predict scientific performance. Thus, we believe that a study on the nature of scientific co-authorship among Vietnamese social scientists using network statistical analysis will yield valuable insights for policy-makers and educators in Vietnam.

II.   Literature review

Over the years, the application of network statistical analysis on science collaboration has become pervasive; it has gleamed many insights into the dynamics of scientific activities as well as the properties of scholars' networks. By exploring a number of databases from different fields such as biomedical research, physics and computer science, Newman showed that scientific collaboration networks seem to form "small worlds", in which any randomly chosen pair of scientists would be separated only through a few intermediate collaborators. Another interesting aspect is that there are different degrees of clustering of scientists in different fields, suggesting the differences in social organizations (Newman 2000). In a 2004 study of sociology collaboration networks by exploring of 30 years' worth of data in the field, from 1963 to 1999, Moody discovered that participation in the network depends on the research major, and scholars more inclined to quantitative work are more likely to collaborate than those in non-quantitative work (Moody 2004). In 2008, on the relationship between structural and socioacademic communities of co-authorship networks, Rodriguez and Pepe applied different community detection algorithms into the network of scholars in the field of wireless communication and sensors networks. They found out that even in interdisciplinary fields and multi-institutional research groups, co-authorship is heavily influenced by departments and institutional affiliations. In 2010, a study of network analysis on co-authorship and citation networks using topic-modelling path-finding algorithms showed that productive authors tend to cite and directly collaborate with colleagues sharing the same research interests (Ding 2010).

Not only the application of network statistics is useful in characterizing the nature of scientist networks, it also provides a powerful tool to study and predict scientific performance such as productivity or research impact. A study on the effects of co-authorship on the performance of scholars using regression model and social network analysis shown that researchers who have a strong connection to only one co-author among a group of connected co-authors perform better than those who have many connections to the same group. The study also suggests it is possible to use professional social network of researchers to predict future performance (Abbasi et al. 2011). In 2013, a group of Taiwanese researchers examine co-authorship networks and research impact through social capital perspective. There are six indicators of social capital in the study: degree centrality, closeness centrality, betweenness centrality, prolific co-author count, team exploration, and publishing tenure. The team found that betweenness centrality is the most influential factor affecting citations of publications (Yi et al. 2013). Using data from library and information science in China, a Chinese research team constructed a network of co-authors, then compared an author's centrality values with his/her citations. They found a high correlation between these two elements (Yan et al. 2009).

Meanwhile, in Vietnam, the application of network statistics to study scientific activities appears to have not been done. However, there have been a few attempts to study quantifiable aspects of scientific

activities among Vietnamese scholars. It is shown that Vietnam has a low scientific production rate in South East Asia, only equivalent to 13.33% of Singapore and 29% of Thailand in the period of 1991-2010 (Nguyen et al. 2011; Yi & Pham 2011; Yi at el. 2013). The total scientific output in Vietnam increased about 16 papers per year during the 1996-2001 period and increased by 20% from 2002 to 2010. It is worth noticing that the share of international collaboration was about 77% of the total publications, of which Japan was the largest collaborating country, followed by United States, France, South Korea, and United Kingdom (Manh 2015; Nguyen et al. 2017). Furthermore, most of the key authors of these international projects did not come from Vietnam but from other countries (Manh 2015).#Mathematics was the only field where domestic output proportion was larger than the international. The largest segment was of biology and agriculture, in which 80%-90% published works involved inter-country collaborations. As for social science in Vietnam, a study on a sample of 410 Vietnamese scholars who have international publications in Scopus during the period of 2008-2017 has revealed that more than 90% of social scientists have published at least one co-written article (indexed in Scopus), and they worked in collaborations 13 times on average (Ho at el. 2017).

In short, faced with the current public desire to improve scientific output in social sciences in Vietnam, there is a shortage of in-depth quantitative analysis on the situation of scientific activities and research policy in Vietnam. Given the high frequency of co-authoring among Vietnamese social scientists, a network statistical analysis on collaboration among Vietnamese scholars in the domain of social sciences would prove to be valuable. It would be interesting to see how network analysis, as a technique first developed for studying network existing in natural world, yield valuable insights into the nature of scientific collaborations in Vietnam, despite the limited data.

III.   Objectives

This study aims to describe the basic properties of a co-authorship network in a sample of 410 Vietnamese social scientists who have published in Scopus-indexed journals and have online profiles, in the period of 2008-2017.

First, through analyzing the vertex degree distribution in the network, the study will discuss the concept of robustness of the network, which means how well-connected the network could remain if certain nodes and edges are removed. Then through the number of cliques and components, the study will describe the basic structure of the network. Furthermore, using metrics such as density and clustering coefficient, the status of the communication and exchange of scientific knowledge and expertise in the network will be analyzed.

#

Second, the study doesn't only providing numerical understanding of the network but also shows the various ways in which it can be graphically represented. In doing so, the study will discuss the usefulness of several techniques of network graphical representation that can be applied to facilitate one's understanding of the network.

Finally, the study will extract two of the largest components - one of the largest groups of connected scientists, then explore its characteristics. By comparing this component with the network of 410 Vietnamese social scientists, the study will provide deeper analysis on the concepts visited above.

IV. Materials and Methods

4.1. *Materials: Original data and the network data set*

The data for this study was derived from a dataset on the productivity of Vietnamese scientists in the field of social sciences collected by Vuong & Associates. The survey, which took place within two months from March to April 2017, was conducted under the license V&A/03/2017, issued on 15 March, 2017.

First, we construct a file that contains data on all the attributes of each author, called a *"Nodes list"*. The data collection process was monitored regularly to ensure its reliability, including the following steps: First, the research team used sources such as personal and institutional websites of authors, websites of journals where their works were published, Google Scholar, and Scopus database to collect data. Then, to check the accuracy of the information, we compare various online sources where each author's information can be found; for example, Google scholar versus Scopus, personal websites versus institutional websites. After this process, the research team obtained a complete dataset of 412 scholars' information, consisting of: (i) age, sex, region; (ii) affiliations; (iii) fields of study; (iv) the number of publications in Scopus, (v) the number of research years since the Master graduation; (vi) the number of researchers they collaborated with; (vii) whether or not they have the title of Prof./Assoc. Prof., etc. All of this essentially constitutes the node.

Based on this information, we then construct our relational data, which is called an *"Edges list"*. We consider two authors as exhibiting a co-authorship tie when they appear together in a scientific publication. Each time the same two authors appear together in a paper, it is counted toward the *"weight"* of the tie. The example of an edges list can be seen in the following figure. The data was then processed and analyzed using statistical software R (v3.3.1). Figure 1 shows an example of how relational data is handled in the study. To illustrate, in the first row of the table on the left side, a published paper being co-authored by scientists ID s004, s076 and s079 is recorded into the database first. Then on the right side, co-authorship relations among these three scholars are recorded; and the weight is the count of how many times each pair co-authors.

Figure 1: An example of the process of handling relational data

Title	year	Journal	ID			Weight
Does Economic Inequality Affect the Quality of Life of Older People in Rural Vietnam?	2017	Journal of Happiness Studies	s076 ;s004; s079	s004	s076	2
The Effect of Having Children on Women's Marital Status: Evidence From Vietnam	2017	Journal of Development Studies	s004; s046	s046	s004	1
Does firm privatisation benefit local households? The case of Vietnam	2015	Post-Communist Economies	s005; s004; s079; s080	s079	s004	2
				s004	s080	1
Firm agglomeration and local poverty reduction: evidence from an economy in transition	2016	Asian-Pacific Economic Literature	s005; s004; s076	s005	s004	2

4.2. Methods of Analysis

The method employed in this study is statistical analysis of network data. There are several reasons why we choose this method. First, the prevalence of co-authorship in research efforts among Vietnamese scientists as shown in the literature review naturally prompts us to ponder on how the co-authors cooperate and the kinds of interactions that exist among them. Second, as we find out that social network analysis has been applied widely all over the world in the study of scientific collaborations, we expect a match between our interest in characterizing collaboration among Vietnamese social scientists and the technical tools this approach provide. Finally, the help of statistical software allows us to create graphic representation of the network, which supplements all the rigorous numerical analysis with a more intuitive way of understanding interactions among actors in the network.

In this study, we will only focus on a descriptive analysis of our network data. The study is strictly limited to the interactions among Vietnamese scholars only. There are two caveats with regards to the method and the scope of the analysis. First, as collaborations with foreign scholars are not accounted for in this study, certain interesting features of the networks can be lost. For example, a foreign scholar could cooperate with two Vietnamese scholars, but these Vietnamese scholars might not publish together. Thus, a link is missing. The cumulative effects of this kind of missing links can make the network appear much less connected than it actually is. Second, network analysis is first developed to solve problems in areas such as mathematics, chemistry, electrical circuits, operational research, and computer science before being applied by sociologists in mid-20th Century to study social network, hence, we can expect there are inherent limits to the explanatory power of the technique.

4.3. Network characterizations

In order to understand the visualization of a network, it is important to familiarize oneself with the terminologies of statistical network analysis. Here, we provide an explanation of terms that are relevant for the scope and purpose of this paper. #More technical explanations of the terms in this paper can be found in Statistical analysis of network data with R by Kolaczyk and Csárdi published in 2014, and Social Network Analysis: A Handbook, Second edition by John Scott published in 2000.

A graph $G= (V, E)$ is a mathematical structure consisting of a set V of *Vertices* (or *nodes*) and a set E of *Edges* (or *links*); elements of E are links between a pair of distinct vertices belongs set V. When two nodes are connected to each other by an edge, they are said to be *adjacent*. In this study, a vertex represents a Vietnamese social scientist, which means the total number of vertices is 410. An edge represents a relationship between two distinct Vietnamese social scientists. A concept that connects edge and vertex is *degree*; a degree of a vertex is the counts of the number of edges incident upon that vertex. For instance, if there are three edges incident upon a vertex, the degree of that vertex is three.

Notice that depending on the attribute of the relationships between two vertices, an edge might or might not have a direction, thus there might be a need to specify the ordering of the pair of vertices in each edge in set E. *A directed graph* is a graph where each edge in E has an ordering to its vertices; *an undirected graph* is a graph where an edge needs not to be defined by the ordering in the vertices. In this study, since the relationship among co-authors is considered to be neutral, the graph that shows their relational ties will be undirected.

To understand the structure of a network, two fundamental concepts are *clique* and *component*. A *clique* is a subset of vertices that are fully cohesive, in that, all vertices within this subset are connected by edges. For example, a node is a clique of size one, an edge is a clique of size two, a triangle is a clique of size three, and so on. A *component* is a subgraph, in which, every vertex can be reached from every other. It is easy to see the different between a clique and a component. In a clique, every two nodes must be connected by an edge or in other words, they must be *adjacent*; while in a component, every two nodes might or might not be connected by an edge, but they must be somehow connected through a path consisting of a number of other edges and nodes.

Regarding the structure of a network, it is natural to wonder about the level of cohesion of the network: How frequent do the edges appear? How likely do three connected nodes close into a clique size 3? These questions can be answered using the concept of *density* and *global clustering coefficient*, also known as transitivity. The *density* of a graph is the frequency of realized edges relative to potential edges. It can be calculated using the formula: density=$2l/[n(n-1)]$, in which l is the numbers of links (or edges), and n is the number of nodes (or vertices). The *clustering coefficient* (or transitivity) measures the relative

frequency with which connected triples of vertices form triangles: $cl_T(G)=3\tau\Delta(G)/\tau_3(G)$, in which $\tau\Delta(G)$ is the number of triangles in the graph G; and $\tau_3(G)$ the number of subgraphs consist of three vertices connected by two edges, i.e. connected triples.

Armed with understanding of relevant technical concepts, next, we will explore the characteristics of the network of 412 Vietnamese social scientists.

V.   Analysis

5.1. Characterizing the network of Vietnamese social scientists

Using R, the dataset employed in this paper counts 412 vertices in the Nodes list and edges in the Edges list. Each vertex or node can be different in terms of degree. The average vertex degree is 1.95 with standard deviation 2.26. This means on average, one Vietnamese social scientist co-authors with about 2 other Vietnamese authors. In figure 2, the degree distribution is plotted on a graph.

Figure 2: A histogram of vertex degree distribution
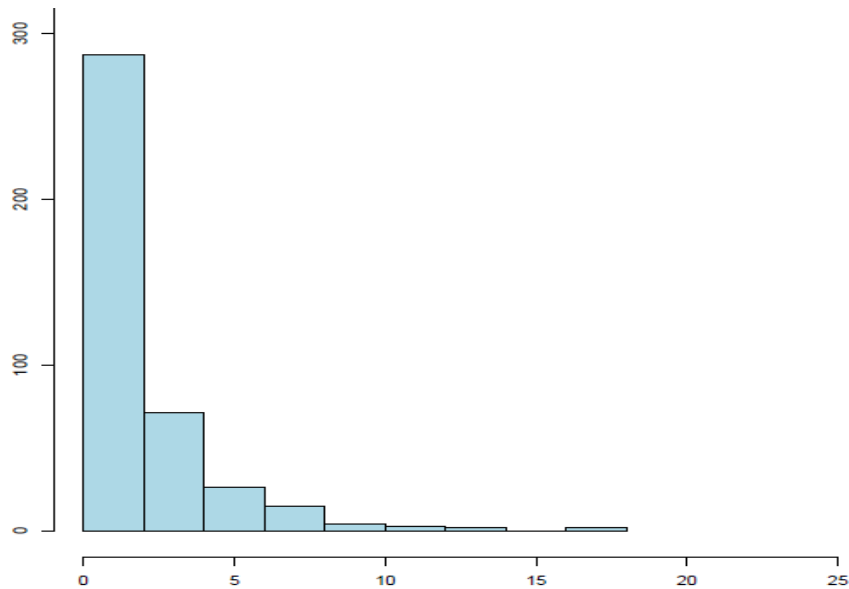
An overwhelming majority of researchers - about 280 out of 412 - possesses degree from 0 to 2. Only about 50 researchers have a vertex degree of 3-4. From degree 4 and up, the number of authors with higher degree decreases dramatically. Clearly, rather than being composed of mostly people with the same level of connections, the network consists of a few very well-connected people, while the rest does not have many connections at all. It can be inferred that it would be possible to break the network into

#

multiples components if we just removed those few well-connected nodes (people of degree higher than 5) or their links. In network analysis literature, how well-connected a network remains when some vertices and edges are taken out is referred to as *robustness* (van Steen 2010). Thus, in this study, the degree distribution reveals that the network of Vietnamese social scientists is not robust. This effect can be seen more visibly when we explore the characteristics of one of the biggest components of this network.

To explore the structure and cohesiveness of the network, it is useful to look at censuses of cliques, components, graph density, and transitivity.

By generating a census of cliques of all sizes, we can get a general sense of the structure of the network:

Table 1: A census of cliques of all sizes for the network of Vietnamese social scientists

Clique size	1	2	3	4	5	6	7	8	9
Numbers	412	401	281	201	144	86	36	9	1

In this network, there are 412 nodes (clique of size 1), 401 edges (clique of size 2), 281 triangles (cliques of size 3), 201 cliques of size 4, and so on. The largest clique is size 9, of which there is only one.

Table 2: A census of components of all size for the network of Vietnamese social scientists

| Component size	1	2	3	4	5	6	7	9	10	11	15	16	27	43
Numbers	125	24	9	3	4	4	2	1	1	2	1	1	1	1

A graph is considered to be connected if every node could be reached by any other node (i.e. if for any two nodes, there is a walk between the two). Looking at table 2, we can see that the network of Vietnamese social scientists is not connected; there are 125 components of size 1. About 30% of the scientists in this study are isolated nodes in the network, possibly because they either work alone or work exclusively with foreigners. On the other hand, the five biggest components (size 11, 15, 16, 27 and 43) together takes up another 30%, while the rest consists of all middle-sized components (size 2-9).

By calculating the density and transitivity of the graph, it can be seen that the network is very sparse. The density of the graph is 0.0052, indicating only about 0.52% of potential edges are realized in this network. On the other hand, when three vertices are connected at all, there is a better than a 50/50 chance they will form a triangle (clique of size 3): The global clustering coefficient of the collaboration graph is 0.5862, indicating that nearly 59% of connected triples have formed triangles. In 2008, a research by Xu and Liu

shows that there is a clear relationship between the speed of the spread of information and clustering coefficient; the higher the clustering coefficient, the slower the information spread. It is reasonable to assume when two scientists co-author in a scientific paper, there is a great deal of knowledge and expertise to be communicated and exchanged. Hence, the low density and high clustering coefficient of the network suggests that the dissemination of knowledge and expertise among 412 Vietnamese social scientists in this study is not happening as smoothly as possible.

5.2. Network visualization

*Figure 3: A visual representation of the network of 412 Vietnamese social scientist*
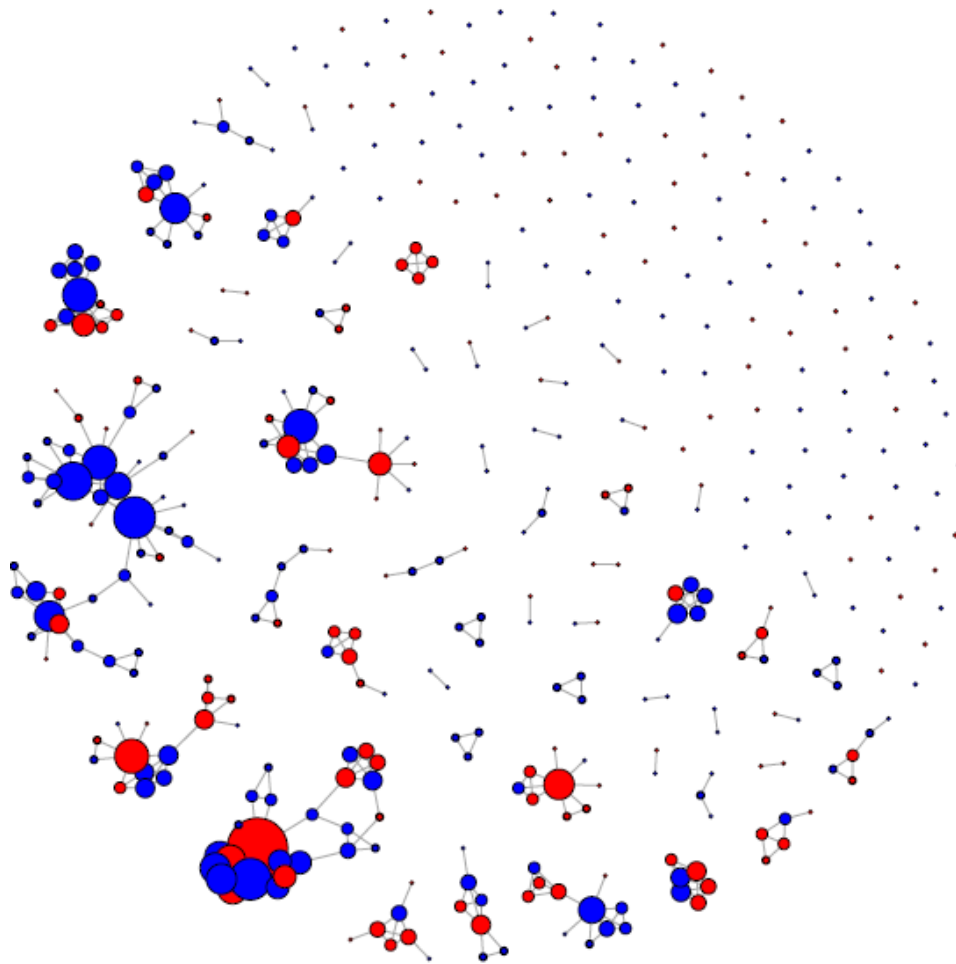
There are several ways to create visual representation of the network. Here, the study aims to strike a balance between creating a graph both visually attractive and useful in facilitating the statistical understanding of the previous histogram and analysis.

For coloring, among many attributes of the nodes that have been collected (region, age, title, etc.), biological gender has been chosen as the basis because of its relatively simple binary nature. In this study,

#

blue color represents male and red represents female. It is hoped that such simplicity will make the graph more aesthetically appealing. Regarding size, the size of each vertex is determined by the number of edges incident on each node – in other words, by the vertex degree. Hence, the higher the number of edges incident upon a vertex, the bigger the vertex is. This is to make visible the gap between the well-connected scientists and the more isolated ones, one of the most striking features of the network as shown in section 5.1. For layouts, among all those available in R(v3.1.1)'s igraph packages, layout Fruchterman-Reingold is chosen because it makes the structure of the network nicely perceptible: 30% of nodes fall into three largest components, 40% are middle-size components, and the 125 left are isolated nodes (recall the statistics on components in section 5.1).

Figure 4: Network visualization with community detection algorithm
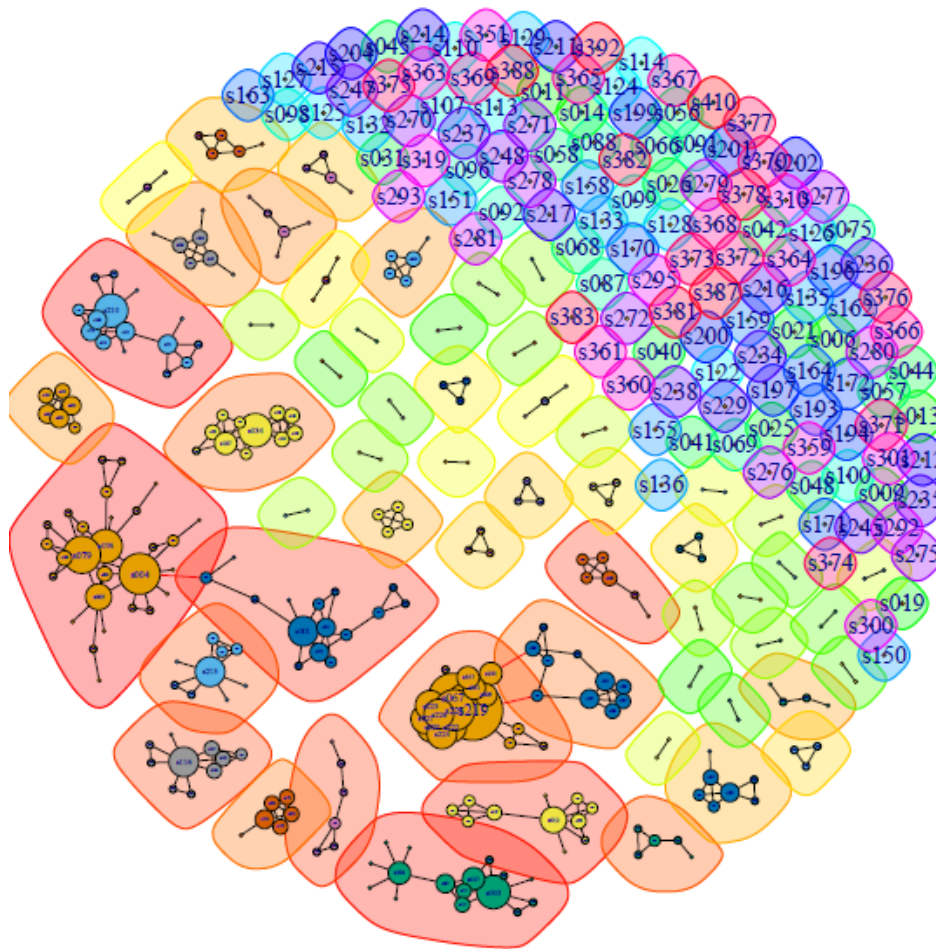
Looking at the biggest components in figure 3, a new pattern emerges: though one can realize the big components are fully connected, they do not seem to be one big close group; rather, they seem to consist of a few smaller communities of very closely connected scientists, and these communities are linked together by one or two vertices acting as weak links. This observation leads us to run the community

#

detection algorithm, which generates another visualization as presented in figure 4. The algorithm does indeed break the two big groups into smaller communities with one or two vertices that connect these communities. It also shows that there are 181 communities.

In the next section, the two largest components, component size 43 and component size 27, will be studied more deeply.

5.3. Exploring the characteristics of the two largest components

Recall that component is a technical term in network theory that refers to a maximally connected subgraph, in which any two vertices can be reached from another via a path consisting of any number of edges and nodes. Thus, any graph can be constituted by many different components. In this study, the network of 412 Vietnamese social scientists is the sum total of 179 components of various size, ranging from 1 to 43; the two largest components have 43 and 27 nodes each. One can treat such components as independent networks in and of themselves. In this section, the characteristics of these two largest components will be explored and compared with the whole network. From this point on, the components will be called Comp43 and Comp27, and the original network will be dubbed Net412.

As one might expect, as we zoom in, there will be differences in the properties of the components in question and that of the network as a whole. Table 3 summarizes and compares the basic metrics of Comp43, Comp27 and Net412.

Table 3: Comparison of basic network metrics of Net412, Comp43 and Comp27

| Metrics	Net412	Comp43	Comp27
Graph density	0.52%	7.20%	22.51%
Mean degree	1.95	3.02	5.58
Transitivity	58.62%	32.43%	70.43%
Mean total publications	3.56	5.53	2.00

Recall in section 4.1, based on density and transitivity, there have been meaningful remarks about the situation of information dissemination in the network. As three networks are compared, interesting patterns start to emerge. In all network metrics, Comp27 scores the highest. Specifically, in terms of density of connections, Net412 is the sparsest, 0.52%; it has about 14 times fewer connections than Comp43 (7.20%) and about 44 times less than Comp27 (22.51%). Regarding average vertex degree, Comp27 is the highest followed by Comp43 then Net412. Concerning global clustering coefficient (or

transitivity), Comp27 towers over Net412 by 11 percentage points (70% versus 50%), while the latter is in turn over 2 times higher than Comp43 (70% versus 32%).

High clustering and low density suggest a certain level of inefficiency in the spread of knowledge and expertise (as explained in section 4.1 on the characteristics of the network of 412 Vietnamese social scientists); either could be the cause of the other. Thus, from the network metrics, one would expect Comp27's dissemination of scientific knowledge and expertise to be less efficient than Comp43. In fact, even though the density of connection in Comp27 is about 3 times that of Comp43, its effects would be limited because of the higher clustering. One can then ask how to verify that high clustering cancels the good effects of even high density. Supposing that better dissemination of scientific knowledge and expertise can be observed in a better scientific output, we could look at the mean value of total publications of scientists in each network for insights on the aforementioned question. Indeed, as table 4 shows, Comp43 performs better than Comp27 in terms of scientific output – almost 3 times higher, 5.53 versus 2.00.

The difference in scientific output between Comp43 and Comp27 can be viewed in figure 5.

Figure 5: Visualization of Comp43 (left) and Comp27 (right) with node size equals scientific output
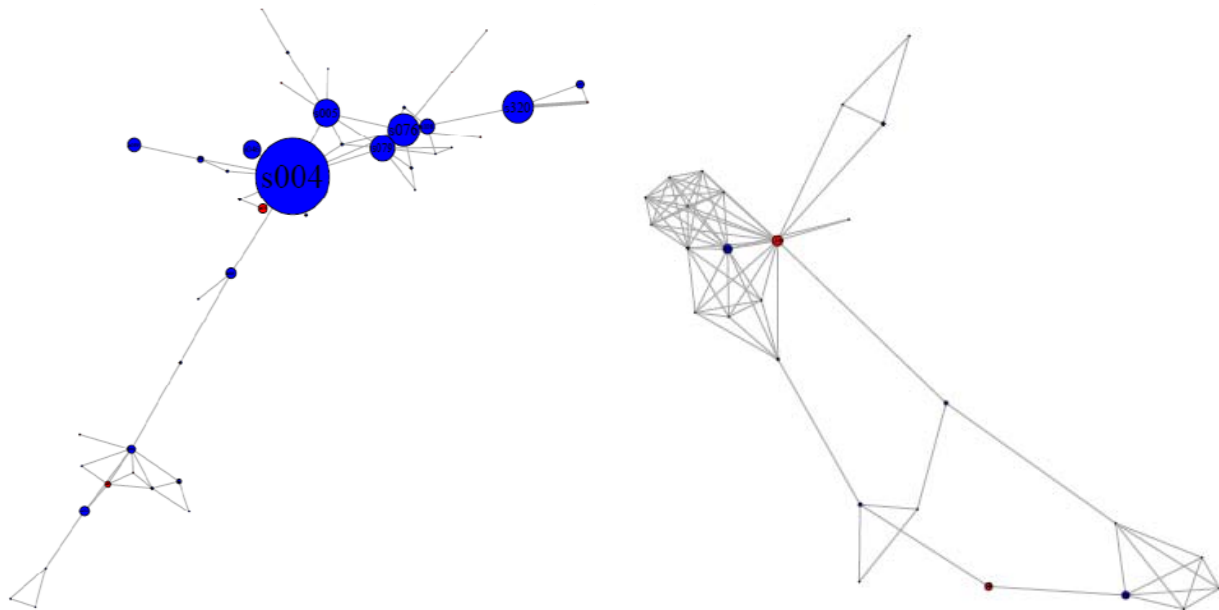
Besides revealing the differences in scientific output of the two networks, figure 4 also reveals that nodes in both networks seems to revolve around one or two important nodes with higher level of scientific output. In Comp43, it is node s004 and in Comp27, it is node s067 and s219 (the visible blue and red dots on the left side of figure 4). It is interesting that these three nodes have highest numbers of edges incidents upon them in their respective networks; s004 has degree 11, highest in Comp43; s067 has degree 13 and

s319 has degree 16, also highest in Comp27. If these important vertices are to be removed, the networks would break apart into several smaller components. This feature was referred to in section 5.1 through the concept of *robustness*, and it should be noted that Net412 is not robust. The situation is the same for Comp43 and Comp27. In figure 5, the histogram distributing the degrees of nodes in these networks shows a clear disparity in vertex degree.

Figure 6: Histogram of degree distribution of Comp43 (left) and Comp27 (right)
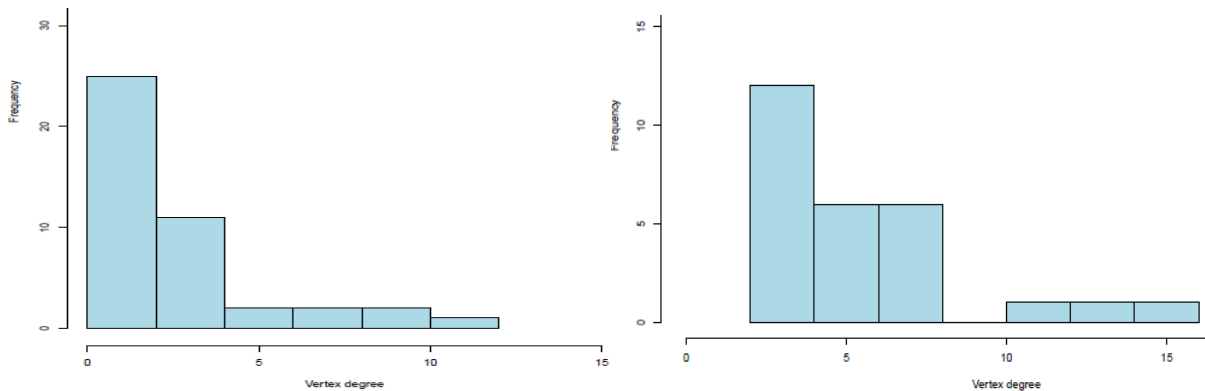
VI.   Discussion

With the purpose of understanding the structure and characteristics of the network of 412 Vietnamese social scientists, the study has applied the technique of social network analysis to give a sense of the structure of the network, the level of connection as well as the level of clustering in the network. Remarks corresponding to each characteristic along with insights into the robustness of the network and the spread of scientific knowledge and expertise in the network have been extracted and discussed. Furthermore, in the last parts of this paper, we zoom into the two largest components of the network and compare their relevant characteristics together with the network of the entire sample.

This paper cannot claim to have exhausted the toolkits that social network analysis could provide. There are still many other aspects of the network worthy of further investigation. For example, though the study has hinted at the difference in scientific output of two networks (comparison of Comp43 and Comp27 in section 5.3), it is worth considering a more systematic examination of the relationship between a network's properties and the scientific output of the vertices it contains. On a more ego-centric and contextual note, there seems to be a relationship between the social status (their position in an institution, for example) of an individual in the network (whether he/she has the most connections or being central to many connections in some ways) and his or her scientific output, as suggested by the examples of node s004, s067 and s219 in the analysis in section 5.3. Thus, finding out whether a correlation among these variables exists does merit further investigation. Another promising area of research is the exploration of

diversity in scientific co-authorship. In the study, node color is coded by gender (section 5.2), but other attributes such as age, region, work, titles, etc. can also be added to the analysis as well.

VII. Conclusions

First, as social network analysis is applied to study the sample of 412 social scientists, the study has shown the network has a low level of connection with only 0.52% of all potential edges realized, and high in clustering with 59% chance a connected triple would close into a triangle. These two statistics together suggest a reality that the communication and exchange of knowledge and expertise among the Vietnamese social scientists are not very efficient. In addition, the degree distribution reveals that it would be difficult for the network to stay well-connected when a few highly-connected nodes and their edges are removed; or, in network theory's terminology, the network is not very robust.

Second, in this study, network visualization is shown to be useful not only in facilitating quantitative understanding but also in discovering new insights into the structures of the network. By applying appropriate techniques of graph plotting, the disparity of the level of connections and the structure of the network can be easily visualized. Using the community detection algorithm, an interesting fact about these biggest groups is unraveled: they mostly comprised of smaller and tightly connected communities with one or two vertices connecting these communities together.

Third, close investigations show that the two largest components in the network have different characteristics from the 412-node-graph. Both smaller networks have more connections than the big one, but in terms of clustering, the 43-node-graph has a much higher level of clustering. Despite these differences, all the three networks resemble in low level of robustness and high disparity in terms of degree distribution, which means when the most connected people are removed from the networks, these latter would immediately be decomposed into several smaller graphs. Most strikingly, the two smaller networks seem to be led by the most productive researchers in them, who also have the most connections.

Finally, there is still much to be learned from both the dataset of 412 social scientists and the network that can be constructed from the raw data. How would the network turn out if other dimensions such as weights or durability of the relational data are added to the analysis? How useful are certain aspects of the network in predicting scientific performance? How would this network evolve over time? Not only intellectually stimulating, these are important questions that are of tremendous practical value for policy-makers, educators and academics, particularly when their decision-making concerns education policies

#

and research organizations. Further investigation in this area of research and on this topic is thus necessary.

Acknowledgement

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.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Journal of Informetrics, 5(1), 187-203.

Hong, H., & Le, T. (2017, March 29). Xếp hạng 300 trường ĐH hàng đầu châu Á: Vì sao đại học Việt Nam chưa ghi tên? Retrieved July 03, 2017, from http://dantri.com.vn/giao-duc-khuyen-hoc/xep-hang-300-truong-dh-hang-dau-chau-a-vi-sao-dai-hoc-viet-nam-chua-ghi-ten-20170329162534079.htm

Kolaczyk, E. D., & Csárdi, G. (2014). Statistical analysis of network data with R (Vol. 65). New York: Springer.

Li, E.Y., Liao, C.H., & Yen, H.R. (2013). Co-authorship networks and research impact: A social capital perspective. Research Policy, 42(9), 1515-1530.

Manh, H.D. (2015) Scientific publications in Vietnam as seen from Scopus during 1996–2013. Scientometrics, 105(1), 83-95. doi: 10.1007/s11192-015-1655-x

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. (2001). Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E, 64(1), 016131.

Nguyen, T.V., Ho-Le, T.P., & Le, U.V. (2017) International collaboration in scientific research in Vietnam: an analysis of patterns and impact. Scientometrics, 110(2), 1035–1051. doi: 10.1007/s11192-016-2201-1.

#

Nguyen, T.V., & Pham, L.T. (2011) Scientific output and its relationship to knowledge economy: an analysis of ASEAN countries. Scientometrics, 89(1), 107-117. doi:10.1007/s11192-011-0446-2

Pham, H.H., & Vuong Q.H. (2017, July 16) Cần cơ chế hỗ trợ nghiên cứu sinh. [Support mechanism for doctoral candidates is needed]. Nhân Dân, serial no. 29(1483), 4-5. Available from: <http://nhandan. com.vn/cuoituan/item/33473902-can-co-che-ho-tro-nghien-cuu-sinh.html> (Retrieved Aug 6, 2017)

Phillips, K.W. (2014, October 1). How diversity makes us smarter. Retrieved June 16, 2017, from https://www.scientificamerican.com/article/how-diversity-makes-us-smarter/

Rodriguez, M.A., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. Journal of Informetrics, 2(3), 195-201.

Vuong, Q.H., Ho, M.T., Vuong, T.T., Napier, N.K., Pham, H.H., & Nguyen, V.H. (2017) Gender, age, research experience, leading role, and academic productivity of Vietnamese researchers in the social sciences and humanities: exploring a 2008-2017 Scopus dataset. European Science Editing, 43(3), 51-55. doi: 10.20316/ESE.2017.43.006

Vuong, Q.H., & Napier, N.K. (2017) Academic research: The difficulty of being simple and beautiful. European Science Editing, 43(2), 32-33. doi: 10.20316/ESE.2017.43.002

Vuong, Q.H., Tran, T.D., Napier, N.K., & Dau, T.H. (2013). Business education in the emerging economy of Vietnam: Twenty years of expectations, illusions, and lessons. In Alon I, Jones V, and McIntyre J (Eds.) Innovation in Business Education in Emerging Markets (pp. 96-112). New York, NY: Palgrave Macmillan

Wu, X., & Liu, Z. (2008). How community structure influences epidemic spread in social networks. Physica A: Statistical Mechanics and its Applications, 387(2), 623-630.

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. Scientometrics, 83(1), 115-131.

Yi, Y., Qi, W., & Wu, D. (2013) Are CIVETS the next BRICs? A comparative analysis from scientometrics perspective. Scientometrics 94(2), 615-628. doi:10.1007/s11192-012-0791-9

APPENDIXES

APPENDIX 1: R code for plotting the graphs of leaders and followers among Vietnamese social scientists

#

```
1.  nodes <- read.csv("C:/.../20170615-Socsci-NODES.csv",header=T, as.is=T)

2.  links <- read.csv("C:/.../Network file 20170615_leadvnco_T.csv", header=T, as.is=T)

3.  head(links)

4.  nrow(nodes); length(unique(nodes$id))

5.  nrow(links); nrow(unique(links[,c("from","to")]))

6.  links <- aggregate(links[,3], links[,-3], sum)

7.  library(igraph)

8.  net <- graph_from_data_frame(d=links, vertices=nodes, directed=T)

9.  class(net)

10. E(net)

11. V(net)

12. net <- simplify(net, remove.multiple = T, remove.loops = T)

13. plot(net, layout=layout.kamada.kawai(net),
    edge.arrow.size=.1,vertex.size=3,vertex.label=NA)

14. a=read.csv("C:/.../20170426-Socsci-NODES.csv",header=T, as.is=T)

15. V(net)$Sex=as.character(a$sex[match(V(net)$name,a$id)])

16. V(net)$Sex

17. V(net)$color=V(net)$Sex

18. V(net)$color=gsub("M","blue",V(net)$color)

19. V(net)$color=gsub("F","red",V(net)$color)

20. nodesize=degree(net)

21. plot(net, layout=layout.kamada.kawai(net), edge.arrow.size=.105, vertex.size=nodesize,
    vertex.label=NA)

22. hist(degree(net),col="lightblue", xlim=c(0,25),ylim=c(0,400),xlab="Vertex degree",
    ylab="Frequency", main="")
```
##Figure 3:

```
23. graph.density(net)

24. table(sapply(cliques(net),length))

25. transitivity(net)

26. comps<-decompose.graph(net)

27. table(sapply(comps,vcount))

28. cut.off <- mean(links$weight)

29. net.sp <- delete_edges(net, E(net)[weight<cut.off])

30. plot(net.sp, layout=layout.kamada.kawai(net), edge.arrow.size=.105, vertex.size=nodesize,
    vertex.label=NA)
```
##Figure 4:

```
31. netc<-fastgreedy.community(net)

32. plot(netc, net, edge.arrow.size=.105,vertex.size=nodesize,vertex.label=NA)

33. length(netc)

34. membership(netc)
```

##Figure 5:

#