Eigenvector Centrality Based on Shared Research Topics in a Scientific Community

Antonio P. Volpentesta, Alberto M. Felicetti

Eigenvector Centrality Based on Shared Research Topics in a Scientific Community

Antonio P. Volpentesta, and Alberto M. Felicetti,

Department of Electronics, Computer Science and Systems,
University of Calabria
via P. Bucci, 42\C, 87036 Rende (CS), ITALY
{volpentesta, afelicetti}@deis.unical.it

Abstract. In this paper we propose a weighted multi-hypergraph as logical structure to model relationships between researchers and interest groups that join them on the base of shared research topics in a given scientific community. The well known concept of eingenvector centrality for graphs is extended to weighted multi-hypergraphs and we present a model instantiation for centrality analysis in the Pro-VE scientific community.

Keywords: Eigenvector centrality, scientific community, weighted multi-hypergraph.

1. Introduction and Backgrounds

Scientific communities are commonly defined as networks of scientists, researchers and professionals who aim to produce, in a collaborative way, new knowledge within a specific domain or issue-area. However, in many cases, collaboration in scientific environments is restricted, and occurs among a small number of people working in the same group, ignoring in some cases the existence of other researchers who are working on similar projects [1].

Moreover, a scientific community is generally characterized by different research topics and contributions that come from a variety of disciplines and backgrounds. In this context, it might be useful to have an idea of the importance of the different research topics and researchers who work on them within a scientific community.

This has led many scholars to study the concept of centrality in a collaboration network of scientists. As matter of fact, network centrality is a concept widely discussed in literature, especially in social network studies [2], and in general, it refers to the importance of a position within a network.

Several authors have studied the "importance" of a node in a network; according to different approaches, they introduce different measures of centrality, As stated by Freeman [3], "there is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is very little agreement on the proper procedure for its measurement". In literature different centrality measures are presented. Closeness centrality and Graph centrality [4] are based on the distances with the rest of nodes, while Betweenness centrality and Stress centrality [3] emphasize the medium

mediating between a pair of nodes. Another centrality measure that is often used in network analysis is eigenvector centrality [5], called also "rank prestige" [2]. Eigenvector centrality analysis is based on the idea that a node is "more central" if it is in relation with nodes that are themselves central, so the centrality of a node does not only depend on the *number* of its adjacent nodes, but also on their value of centrality.

The usage of centrality measures are particularly interesting in the study of networks formed by researchers belonging to a scientific community [6]. These studies use models based primarily on graph structures, constructed on the basis on the author-topic relationship and, more in general, on the analysis of papers' contents. However, in several cases, models based on graphs do not provide a suitable representation of complex relationships , for instance supra-dyadic relations. The use of more general logical structure as hypergraphs [7] seems to be more appropriate in these situations. Few attempts have been made to utilize hypergraphs in modeling a social network [8], and, more specifically, a scientific community network [9]. However, in our opinion, weighted multi-hypergraphs are the appropriate structures to represent multiple and weighted relationships.

In this paper we propose a model based on a weighted multi-hypergraph to represent relationships between researchers and research interests, grouping researchers with common interests. Moreover, in order to measure the importance of researchers and research topics in a scientific community ,we extend the eigenvector centrality notion to this general logical structure and we present an algorithmic approach. Lastly, we describe a first application of the model to the Pro-Ve community, a scientific community that aims to promote research and production of new knowledge on Collaborative Networks.

2. Eigenvector centrality for weighted multi-hypergraphs

A multi-hypergraph is a generalization of a multi-graph, in which edges, called hyperedges, may connect any positive number of vertices [7]. Formally, a multi-hypergraph $\mathcal{H}$ is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{v_1, \ldots, v_m\}$ is a set of vertices, $\mathcal{E} = \{E_1, \ldots, E_n\}$ is a multi-set of nonempty subsets of $\mathcal{V}$, called hyperedges. Because $\mathcal{E}$ is a multiset, an hyperedge may appear more than once in $\mathcal{E}$. A vertex-hyperedge weighted multi-hypergraph is one in which each couple vertex-hyperedge (v_i, E_j), such that $v_i \in E_j$, is assigned a positive weight.

We use $w_{ij} \in \mathbb{R}_+$ to denote the weight given to (v_i, E_j) and refer to $W=(w_{ij})$, where $w_{ij} = 0$ if $v_i \notin E_j$, as a vertex-hyperedge weighted incidence matrix for $\mathcal{H}$. Notice that W is the classical vertex-hyperedge incidence matrix for $\mathcal{H}$ when $w_{ij} = 0$ or 1.

In order to study the centrality of vertices and hyperedges in $(\mathcal{H}, W)$ we make the well known *mutually reinforcing relationship* assumption [10]: *an important hyperedge is a hyperedge whose elements are important vertices; an important vertex is a vertex that belongs to many important hyperedges.*

Numerically, it is natural to express the mutually reinforcing relationship between hyperedges and vertices as follows:

Let x_i be the 'importance' of vertex v_i and let y_j be the 'importance' of hyperedge E_j. The simplest formulation of the mutually reinforcing relationship assumption is given by these equations:

$$x_i = c_1 \sum_{j=1}^{n} w_{ij}\, y_j \ , \qquad \text{for i} = 1,\ldots,\text{m.} \tag{1}$$

where the constant of proportionality, $c_1>0$, is independent of i.

$$y_j = c_2 \sum_{i=1}^{m} w_{ij}\, x_i \ , \qquad \text{for j} = 1,\ldots,\text{n.} \tag{2}$$

where the constant of proportionality, $c_2>0$, is independent of j.

In matrix notation with $x= (x_1,x_2,\ldots,x_m)$ and $y= (y_1,y_2,\ldots,y_n)$ this yields

$$WW^t x = \lambda x, \qquad W^t W y = \lambda y, \qquad \text{where } \lambda=c_1c_2. \tag{3}$$

Standards results of linear algebra[1] lead to state that (3) is a solvable system of equations. More precisely, a solution is given by setting $\lambda=\lambda^*$, the dominant W^tW's eigenvalue (that is equal to the dominant eigenvalue of of WW^t), $x=x^*$, a nonnegative eigenvector of WW^t in the eigenspace associated with λ^*, $y=y^*$, a nonnegative eigenvector of W^tW in the eigenspace associated with λ^*. We call a normalization of x^* an *eigenvector-centrality measure* of the vertices in ($\mathcal{H}$, W) and a normalization of y^* an *eigenvector-centrality measure* of the hyperedges in ($\mathcal{H}$, W).

If W is the vertex-hyperedge incidence matrix of an hypergraph, the equations (3) are the well known ones that arise when studying eigenvector centrality in hypergraphs, [8]. This means that the notion of eigenvector-centrality we introduced for a weighted multi-hypergraph is the natural extension of the well known one for an hypergraph. Moreover, we can use an adaptation of the Hits (Hyperlink-Induced Topics Search) algorithm, proposed by Kleinberg, [10], in order to calculate eigenvector-centrality of vertices and hyperedges in ($\mathcal{H}$, W).

In the algorithm we have used the sum-norm to range nodes and hyperedge according to their proportion of the centrality within a vertex-hyperedge weighted multi-hypergraph. The effect that different normalization have on the interpretation of eigenvector-centrality within a graph is investigated in [11].

3. The weighted multi-hypergraph model

In order to study the centrality of researchers and research topics in a scientific community, we propose a model whose underlying logical structure is a vertex-hyperedge weighted multi-hypergraph. The components of the model are:

- **D** = {d_1,…, d_p} an ordered set of documents (scientific papers);
- **T** = {t_1,…, t_m}an ordered set of research interests (research topics);
- **R** ={r_1,…, r_n} an ordered set of researchers (authors), members of a scientific community;

[1] WW^t and W^tW share minimum(m,n) eigenvalues; these eigenvalues are all ≥0; due to the theorem of Perron–Frobenius, there exists an eigenvector of the maximal eigenvalue with only nonnegative entries, [15].

- $\mathbf{A} \in \mathbb{R}^{m \times p}$ a binary matrix that represents the relationships between authors and documents produced by them, i.e.: $a_{ik} = 1$, if researcher r_i is one of the authors of document d_k, otherwise $a_{ik} = 0$.
- $\mathbf{B} \in \mathbb{R}^{p \times n}$ a nonnegative matrix that gives a measure of how much documents are devoted to research topics. More precisely, the generic entry b_{kj}, measures the portion of the document d_k that deals with research topic t_j and it's required that $0 \leq b_{kj} \leq 1$, for any j, k, and $\sum_j b_{kj} = 1$, for any k.
- $\mathbf{C} = (c_1, c_2, \ldots, c_p)$ a positive vector, where the generic entry c_k represents a measure of the popularity [2] of d_k in the scientific community.

We introduce the multi-hypergraph $\mathcal{H}(\mathcal{R}, \mathcal{E})$, where:

$\mathcal{R} = \mathbf{R} = \{r_1, \ldots, r_n\}$;

$\mathcal{E} = \{E_1, \ldots, E_n\}$, with $E_j = E(t_j) = E(tj) = \{r_i \in \mathbf{R}: \exists\ k$ such that $a_{ik} = 1$ and $b_{kj} > 0\}$.

By assuming that research interests of any researcher r_i are manifested on documents whose r_i is an author, E_j represents an interest group on a research topic t_j; in other words E_j is the subset of $\mathbf{R}$ consisting of all researchers that share the research topic t_j. Of course, a researcher may belong to many interest groups and many interest groups may be constituted by the same subset of researchers (this is the reason why $\mathcal{H}$ is a multi-hypergraph). The relationship between researchers and interest groups may be derived through a semantic analysis of the documents' content.

In order to assign a weight[3] to any couple researcher-interest group (r_i, E_j), we make the following assumptions and settings:

- The content of a document is due in equal measure to all its authors. More precisely, the fraction a_{ik}/h_k, where h_k is the number of authors of d_k, measures the document portion that is attributed to r_i and the research topics of d_k are also research interests of its authors.
- The number $b_{kj} \cdot c_k$ measures the contribution given by the research topic t_j to the popularity of the document d_k.
- The number $(a_{ik}/h_k) \cdot (b_{kj} \cdot c_k)$ measures the contribution given by the portion of d_k, dealing with t_j and attributed to r_i, to the popularity of d_k.

According to these assumptions and settings, we propose to estimate the weight associated to the couple (r_i, t_j), as follows:

$$w_{ij} = \sum_{k=1}^{p} (a_{ik}/h_k) \cdot (b_{kj} \cdot c_k)$$

In order to calculate eigenvector-centrality of researchers and research topics in a scientific community, we may consider the weighted multi-hypergraph $(\mathcal{H}, W)$, where $W=(w_{ij})$, i=1,…m and j=1,…n, is the matrix that represents the weighted relationships between researchers and research topics. We observe that the characteristic matrix associated with W is the incidence matrix of $\mathcal{H}$.

---

[2] Researches in bibliometrics have long been concerned with the concept of popularity (or importance or impact) of individual scientific papers and journals and they have provided quantitative estimates based on the use of citations. The most well-known measure in this field is Garfield's impact factor, [13].

[3] Some authors explain the meaning of a weight in terms of strength of endorsement within a community, [10].

Example.
Let us consider the following instantiations of the model components:
D = $\{d_1,d_2,d_3\}$; T = $\{t_1, t_2, t_3, t_4\}$; R = $\{r_1, r_2, r_3, r_4, r_5\}$; C =(1,2,3);

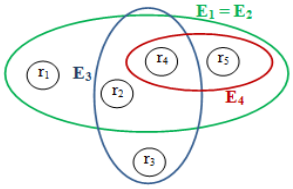
A	d_1	d_2	d_3
r_1	1	0	0
r_2	1	1	0
r_3	0	1	0
r_4	0	1	1
r_5	0	0	1

B	t_1	t_2	t_3	t_4
d_1	3/4	1/4	0	0
d_2	0	0	1	0
d_3	1/4	1/4	0	1/1

We have: $h_k = \frac{1}{\sum_i a_{ik}}$ k=1,2,3, i.e. $(h_1, h_2, h_3) = (2, 3, 2)$;

The multi-hypergraph $\mathcal{H}$ has the following vertex-hyperedge incidence matrix and graphical representation:

E	t_1	t_2	t_3	t_4
r_1	1	1	0	0
r_2	1	1	1	0
r_3	0	0	1	0
r_4	1	1	1	1
r_5	1	1	0	1



and the matrix W is the following:

W	t_1	t_2	t_3	t_4
r_1	3/8	1/8	0	0
r_2	3/8	1/8	2/3	0
r_3	0	0	2/3	0
r_4	3/8	3/8	2/3	3/4
r_5	3/8	3/8	0	3/4

Through the application of an adaptation of HITS Algoritm to W, we obtain the eigenvector-centrality of researchers and research topics:

x =(0,0656; 0,1966; 0,1309; 0,3689; 0,2379)
y =(0,2165; 0,1729; 0,3082; 0,3024)

4. A model instantiation for the Pro-Ve community

The Pro-VE community is a scientific community that aims to promote research and production of new knowledge on Collaborative Networks (shortly, CN). PRO-VE conferences offer researchers and practitioners opportunities to meet together, present and discuss both latest research developments and industrial practice case studies.

In what follows we briefly describe the instantiations of the model components as well the path taken to them. In order to instantiate the sets **D**, **R** and the matrix **A**, we have considered scientific papers presented at Pro-Ve conferences. More specifically, **D** consists of all selected papers that were published in the books of the last five Pro-Ve conferences (2005-2009), [14], **R** is the set of those researchers who appeared as an author of at least one scientific article published in such books and **A** represents their authorship relation to their Pro-Ve papers.

In order to instantiate **T**, we have modeled a research topic t_j in the Pro-Ve community as a triple $t_j = (OF, DA, \mathrm{ES})$, where:

- OF is the set of CN Organizational Forms. OF is a flat set whose elements are substantially derived from the classification provided in [15];
- DA is the set of Dimensional Aspects of a CN. DA is a flat set whose elements are substantially derived from the reference model described in [16], and widely accepted in the Pro-Ve community;
- ES is the set of the economic sectors, each one encompassing real business environments, where CN models, mechanisms, methodologies, principles and supporting tools are instantiated and implemented. According to the well known four-sector hypothesis, ES consists of primary, secondary, tertiary and quaternary sector and a "dummy" element denoting that no real world application is addressed by the research.

In other words a research topic is characterized by a dimensional aspect of a CN organizational form and possibly a case study or an application in primary industry, manufacturing, industrial services or intellectual services, (see the following table):

Table 1. Instances of research topics components

ORGANIZATIONAL FORMS	Collaborative Network (*), Supply chain, Virtual Government, Virtual Enterprise, Virtual Organization, Extended Enterprise, Virtual team, Human breeding environments (communities), Organizational Breeding Environments (VBE), Industry Cluster, Industrial District, Business Ecosystem, Collaborative Virtual Lab, Disaster rescue Net, Innovation networks.
DIMENSIONAL ASPECTS	Actors/relationships, Roles, Hardware / software resources, Human resources, Information / knowledge resources, Ontology resources, Processes, Auxiliary processes, Methodologies, Prescriptive behavior, Obligatory behavior, Constraints and conditions, Contracts and cooperation agreements, Meta dimension (**), External view (***)
ECONOMIC SECTORS	No real world application, Primary Economic Sector (****), Secondary Economic Sector (****), Tertiary Economic Sector (****), Quaternary Economic Sector (****).
(*) The focus is on general forms of CN rather than on specific organizational forms. (**) This dimension addresses to the analysis of principles, models and theories applicable and useful for modeling Structural, Componential, Functional and Behavioral dimension of CN. (***) This dimension deals with exogenous interactions with CN surrounding environment, such as Market (customers, competitors, other CNs) and/or Society (third party institutions, Governments, No Profit Organizations). (****)Primary sector (i.e.): Agriculture, Fishing, Forestry, etc… Secondary sector (i.e.): Automotive, Construction, Electronics, Mechanical, Textile, etc… Tertiary sector (i.e.): Industrial Services, Commerce, Transportation, Hospitality, Maintenance, etc.. Quaternary sector (i.e.): Banking, Consulting, Education, Government Services, Healthcare, etc…	

The instantiation of the matrix **B** has been obtained through a collaborative process of semantic analysis of Pro-Ve papers' content. Such a process, widely described in [17], is collaboratively performed by a team of experts that are supported by an automatic paper indexing tool. It is aimed to associate one or more instantiation of the triple (OF, DA, ES) to any Pro-VE paper and it essentially consists of the following interrelated steps:

- Making the list of research topics.
- Developing a structured set of concepts for any research topic.
- Extracting a set of keywords from any paper.
- Associating paper's keywords to concepts.

By assuming an equi-distribution of the content of a paper among its research topics, the matrix **B** has been instantiated as follows: $b_{kj} = 1$, if is a research topic of the k-th paper, otherwise $b_{kj} = 0$. Lastly, any entry c_k of C is instantiated at x_k+1 , where x_k is the number of documents in D that cite d_k.

5. Conclusions and future works

The presented work is a proposal aimed to determine eigenvector centrality in scientific community, starting from research publications. From a theoretical point of view, the model we have introduced can be further exploited by extending other well known concepts of centrality (e.g. closeness or betweenness centrality) to weighted multi-hypergraphs. From a practical point of view, we have presented a model instantiation that allows us to study eingenvector centrality in the Pro-Ve community. This work is still in implementation phase (we are collecting and validating data derived from a semantic analysis of Pro-Ve papers) and one of the future steps is its completion in order to provide measurements and statistucal analysis of the centrality of researcher and research topics within the Pro-Ve community.

References

1. Rodrigues, S., Oliveira, J., Moreira de Souza, J., Competence mining for virtual scientific community creation, Int. J. Web Based Communities, V. 1, N. 1, pp. 90-102, (2004)
2. Wasserman, S., Faust, K., Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK, (1994)
3. Freeman, L.C., A set of measures of centrality based on betweenness. Sociometry, V. 40, pp. 35–41, (1977)
4. Sabidussi, G., The centrality index of a graph. Psychometirka, V. 31, 581–603, (1966).
5. Bonacich, P., Factoring and weighting approaches to clique identification. Journal of Mathematical Sociology, V. 2, pp. 113-120, (1972)
6. Newman, M. E. J., Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical Review E, V. 64, (2001)
7. Berge, C., Graphs and Hypergraphs. Elsevier, (1973)
8. Bonacich, P., Holdren, A. C., Johnston, M., Hyper-edges and multidimensional centrality, Social Networks V. 26, pp. 189–203, (2004)
9. Estrada, E., Rodríguez-Velázquez, J.A., Subgraph centrality and clustering in complex hyper-networks, Physica A, 364, 581–594, (2006)
10. Kleinberg, J., Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5): pp. 604–632, (1999)
11. Ruhnau, B., Eigenvector-centrality: a node-centrality?, Social Networks V. 22, pp. 357–365, (2000)
12. Golub, G., Van Loan, C.F., Matrix Computations, J. Hopkins University Press, (1989)
13. Garfield, E., Citation analysis as a tool in journal evaluation, Science, V. 178, 471-479, (1972)
14. Camarinha-Matos, L.M., et al.(eds): Pro-VE 2005 IFIP Vol. 185, Pro-VE 2006 IFIP Vol. 224, Pro-VE 2007 IFIP Vol. 243, Pro-VE 2008 IFIP Vol. 283, Pro-VE 2009 IFIP Vol. 307, Springer.
15. Camarinha-Matos, L.M. and Afsarmanesh, H., Collaborative Networks: Value Creation in a Knowledge Society, in Knowledge Enterprise, IFIP, V. 207, 26-40, Springer, (2006)
16. Romero, D., Galeano, N., Molina, A., A Virtual Breeding Environment reference model and its instantiation methodology, in Camarinha-Matos, L.M., Picard, W., Pervasive Collaborative Networks, Springer eds., (2008)
17. Volpentesta, A. P., Ammirato, S., Felicetti, A. M., Competence mapping through analysing research papers of a scientific community, Technical Report, DEIS-Unical (2010)