Journal of Informetrics 12 (2018) 1042–1058

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Topological metrics in academic genealogy graphs

Luciano Rossi[a], Rafael J.P. Damaceno[a], Igor L. Freire[a], Etelvino J.H. Bechara[b], Jesús P. Mena-Chalco[a,*]

[a] Center for Mathematics, Computation and Cognition, Federal University of ABC, Santo André, Brazil
[b] Department of Fundamental Chemistry, Chemistry Institute, University of São Paulo, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 10 March 2018
Received in revised form 4 August 2018
Accepted 4 August 2018
Available online 10 September 2018

Keywords:
Topological metrics
Academic genealogy
Advisor–advisee relationship
Genealogical graph

ABSTRACT

Academic genealogy aims to structure and analyze the mentoring relationships between advisor and advisee. The representation of this structure results in academic genealogy graphs. For the analysis and characterization of these graphs, we present a set of metrics and their corresponding mirror metrics that capture the characteristics of its topological structure and represent them as quantitative attributes. The metrics of fecundity, fertility, descendants, cousins, generations, and relationships consider the descendants of the academics represented in the graph. The mirror metric of these topological metrics considers the ascendancy of academics. Individually, the metrics have strong semantic intuition and define characteristics regarding the performance in the mentoring of an academic. Together, the metrics are useful for the identification, characterization, and classification of communities and their members. The genealogical data available through the platforms of the Mathematics Genealogy Project and the Academic Family Tree were used as case studies. Two hundred thirteen thousand and 675,000 academic records were obtained for each project. We analyze the capacity of characterization of the metrics using the structuring of a similarity graph and through the distribution of the nodes in principal components. We observed that the set of metrics is capable of capturing the configuration pattern existing in genealogy graphs independently of its scale.

1. Introduction

Research that has scientific knowledge and its development as the object of study considers the characterization of academics and researchers through the analysis of their publications. Thus, it is natural to use bibliometric indicators for this purpose (Agarwal et al., 2016; Todeschini & Baccini, 2016). On the other hand, there is a growing interest in the way these actors relate to each other and transmit scientific knowledge through these relationships. Thus, relationships of academic mentoring have been considered as a complement to the analysis of publications to obtain better results in the context of scientific production.

The activity of academic mentoring promotes the evolution of the advisee, of the institution, of the science, and of the society. Currently, we observe different initiatives to documenting, analyzing and classifying structures of academic genealogy (Didegah & Thelwall, 2013; Gargiulo, Caen, Lambiotte, & Carletti, 2016; Malmgren, Ottino, & Amaral, 2010). Analyzing

* Corresponding author.
E-mail addresses: luciano.rossi@ufabc.edu.br (L. Rossi), rafael.damaceno@ufabc.edu.br (R.J.P. Damaceno), igor.freire@ufabc.edu.br (I.L. Freire), ebechara@iq.usp.br (E.J.H. Bechara), jesus.mena@ufabc.edu.br (J.P. Mena-Chalco).

https://doi.org/10.1016/j.joi.2018.08.004
1751-1577/

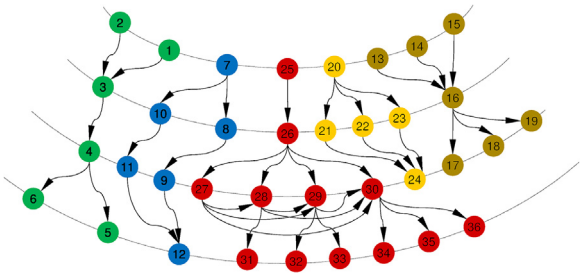


Fig. 1. Example of a general genealogical graph. Each vertex represents an academic and are positioned in hierarchical order (generations) defined by the direction of the edges. The same color indicates a connected component or academic family. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

these relationships in the form of a genealogical structure allows a deeper understanding of the scientific community, and of the individual through their relationships.

Academic genealogy (AG) is defined as a quantitative study of the intellectual heritage perpetuated through academic mentoring relationships between researchers (advisors/supervisors) and their students (advisee) (Sugimoto, 2014). The AG may be a viable way to classify scholars based on their contributions to training human resources and to identify groups with similar performances. The motivation underlying the use of AG as a form of evaluation is the belief that the most efficient way for a scientist to work beyond his or her time is to provide to the next generation of academics mentored by them scientific ideas that may influence youngers to continue their works, contributions and visions.

Analysis considering AG involves the following steps: (i) genealogical data mining (Wang et al., 2010), (ii) genealogical graph structuring (Vicknair et al., 2010), (iii) graph characterization (David & Hayden, 2012), and (iv) the discovery of knowledge in these structures (Sugimoto, Ni, Russell, & Bychowski, 2011). The characterization of AG structures can be performed using metrics that reflect characteristics referring to the topology of these structures and are useful for the classification of academics (Rossi, Freire, & Mena-Chalco, 2017).

The purpose of this paper is to describe a set of metrics that are useful for the characterization of academic genealogy graphs and that have the following characteristics: (i) are simple to obtain and interpret, (ii) individually provide information about the topological structure of the graph and/or that are semantically strong, (iii) globally foster the identification of academic communities with apparent characteristics and (iv) are independent of the scale of the data considered. In the context of bibliometric analysis, it is possible to identify a wide range of publications that describe metrics for their characterization, such as co-authorship networks. However, to the documentation of AG reviews is limited. The graphs of AG have particular characteristics regarding structure and semantics, so it is essential to develop specific tools for the exploration of these structures.

We consider the use of a set of 12 topological metrics. The first six metrics are based on the descent of a vertex of interest. Next, we introduce new other metrics that are mirror symmetries of the first group. They are based on the ascendancy of a vertex of interest. The latter is useful for differentiating vertices that have no descendants. As case studies, we use an example graph (artificial) that contributes to the definition of the metrics and two sets of real genealogical data. The datasets in question are the Mathematics Genealogy Project and the Academic Family Tree project. The characterization of these datasets is not part of the scope of this paper. The analysis presented shows the ability of the metrics to characterize genealogy graphs. Thus, the focus is on the proposed method and not on the characterization of data.

2. Materials and methods

The analyses performed for this paper are based on genealogy graphs (directed graph). A directed graph $\vec{G}$ is a pair of (V, E), where V is a finite set of vertices and E, edges, is a binary relation in V. Vertices (V) represent individuals (academics) and the directed edges (E) represent the mentoring relationships.

The academic lineage is a path in the genealogy graph that connects with the forbears or descendants. Formally, a path with length k from a vertex source to a vertex destination ($u \rightsquigarrow u'$) in a directed graph ($\vec{G}$) is a sequence of vertices ($v_0, v_1, v_2, \ldots, v_k$) such that $u = v_0$, $u' = v_k$ and $(v_{i-1}, v_i) \in E$ for $i = 1, 2, 3, \ldots, k$.

Fig. 1 shows a general genealogy graph where the vertices (academics) are labeled with numbers, and the edges (relationships) connect the vertices. The direction of the edges indicates the origin and the destination of mentoring and represent the advisor and the advisee, respectively. The position of the vertex in the graph indicates the generation to which it belongs (hierarchical order defined by the direction of the edges). The vertices that have the same color form an academic family (connected component).

2.1. Genealogical metrics

A genealogical metric is a quantitative attribute that defines a topological feature of the genealogical graph structure and is semantically stable. A genealogy graph presents particular characteristics, which may be different from those observed in other types of social networks. Thus, developing specific metrics for the characterization of genealogical graphs is useful for the characterization of individuals and groups.

The metrics may consider the descendants and/or the ascendants of an academic. In this paper, we examine six metrics on descendants and their respective symmetric parts, or best, the mirror symmetries of the metrics on ascendants, providing twelve quantitative attributes. The descendant metrics consider a genealogy graph $\bar{G}(V, E)$ and a vertex of interest $v \in V$. The ascendant metrics refer to $\bar{G}(V, E^T)$ where $E^T = \{(u, v) : (v, u) \in E\}$.

- *Fecundity*:

  Fecundity is the number of academic children (direct descendants) that an academic has and refers to the number of advisees of an academic. The metric defines the productivity of an academic and reflects the direct contribution exercised in the formation of the community, by the number of people trained by a given advisor. An important feature of fecundity is that its scope is strictly local, i.e., the metric does not consider the advisee performance that cannot be distinguished. Formally, the direct descendants set of v is

  $$F^+(v) = \{u \in V : (v, u) \in E\}, \tag{1}$$

  and the metric fecundity is the cardinality of this set:

  $$f^+(v) = |F^+(v)|. \tag{2}$$

  In Eq. (2), $|A|$ means the number of elements, or the cardinality, of the set A. Higher fecundity values may indicate maturity in the activity of academic mentoring. Academics who have many students during their career are those who make a significant contribution to the development of the community to which they belong (mentoring). On the other hand, low values for metrics may indicate a young academic or someone that, for some reason, may not have all needed ingredients to develop his/her career regarding mentoring. Analyzing the evolution of the number of advisees over time makes it possible to infer the academic career phase (activity life cycle), considering as a hypothesis, that the number of advisees increases according to the advisor's seniority.

  Obtaining fecundity considering the ascendants of an academic of interest (mirror fecundity) means that we are counting the number of advisors the academic had. Commonly, advisees have a single advisor, but there are many cases where one or more co-advisors are present. In most cases, the co-mentoring is justified by the interdisciplinarity demanded by the subject of research of the advisee. Thus, mirror fecundity may indicate the interdisciplinarity in the formation of the advisee or at least this shows that the advisee needed skills or knowledge that were not available in a single advisor. On the other hand, individuals who have mirror fecundity equal to zero, in a theoretical context, are an academic root, i.e., they are the origins of their scientific community. Considering a real context, it is more common for an academic to have mirror fecundity equals to 0 due to the incompleteness of genealogical data. Formally, the direct ascendants set of v is:

  $$F^-(v) = \{u \in V : (v, u) \in E^T\}, \tag{3}$$

  and the mirror fecundity is the cardinality of this set:

  $$f^-(v) = |F^-(v)|. \tag{4}$$

  For each metric defined from here, an example will follow based on the graph in Fig. 1. Considering vertex 29, we have the following: $F^+(29) = \{30, 32, 33\}$ and $F^-(29) = \{26, 27, 28\}$.

  It is worth noting that the definitions considered for the metric fecundities f^+ and f^- correspond, in Graph Theory, to the denominations of in-degree and out-degree, respectively.

- *Fertility*:

  Fecundity and fertility are semantically synonymous. Considering demographic studies, fertility is related to the number of individuals able to reproduce and fecundity refers to the number of descendants of the individual. In this paper, we consider fertility as the number of academic children who have non-zero fecundity. In other words, the fertility of an academic is the number of advisees that became advisors. Consider only the academics who carried out mentoring aims to demonstrate the propagation of the mentoring activity in the descendants of the academic of interest. The set of fecund advisees of v is:

  $$FT^+(v) = \{u \in V : (v, u) \in E \text{ and } f^+(u) > 0\}, \tag{5}$$

  and the metric fertility is:

  $$ft^+(v) = |FT^+(v)|. \tag{6}$$

  Measuring the fertility of an academic is essential to observe how effective the mentoring was. This effectiveness defines the perpetuation of mentoring activity and is fundamental to the development of the science. This metric reveals the quality

of the mentoring for fostering in advisees the desire to transmit to future generations the characteristics of current scientific thinking.

The mirror fertility, that is, the fertility for ascendants is the number of advisors who had at least one advisor. The set of fecund advisors of v is:

$$FT^{-}(v) = \{u \in V : (v,u) \in E^{T} \text{ and } f^{-}(u) > 0\}, \tag{7}$$

and the metric mirror fertility is:

$$ft^{-}(v) = |FT^{-}(v)|. \tag{8}$$

Individuals v with $ft^{-}(v) = 0$ belong to the group of those who initiated a community or line of academic research. Consider the vertex 29 as an example, we have the following: $FT^{+}(29) = \{30\}$ and $FT^{-}(29) = \{26, 27, 28\}$ (Fig. 1).

- *Descendants*:

Descendants are all advisees who have a direct or indirect mentoring relationship with the academic of interest. By descendants we consider the academic children, academic grandchildren, academics great-grandchildren, and so on, of a given researcher. This metric is important to check the impact of a given academic in the formation of the scientific community. On the other hand, the perceived relevance of an academic is always subject to the performance of their offspring. Thus, the number of descendants can presuppose inaccurate ratings, when considered individually. The set of descendants of v is:

$$D^{+}(v) = \{u \in V : \exists v \rightsquigarrow u\}, \tag{9}$$

and the metric descendants is:

$$d^{+}(v) = |D^{+}(v)|. \tag{10}$$

The set of inverted descendants of v is

$$D^{-}(v) = \{u \in V : \exists u \rightsquigarrow v\}, \tag{11}$$

and its corresponding mirror metric is

$$d^{-}(v) = |D^{-}(v)|. \tag{12}$$

Consider as an example vertex 29, we have the following: $D^{+}(29) = \{30, 32, 33, 34, 35, 36\}$ and $D^{-}(29) = \{25, 26, 27, 28\}$ (Fig. 1).

- *Cousins*:

In the traditional genealogical context, two people are cousins if they share the same grandfather and have different parents. Adapting this kinship to the academic world, academic cousins are those who have different mentors sharing the same advisor. An academic who has a large number of cousins belongs to a large family and has prolific ancestors. The set of cousins of v is

$$C^{+}(v) = \{u \in V, \quad u \neq v : (w,u) \in E, \quad \text{where} \quad w \in F^{+}(x), \quad x \in F^{-}(z) \quad \text{and} \quad z \in F^{-}(v)\}, \tag{13}$$

the cousins' metric is:

$$c^{+}(v) = |C^{+}(v)|. \tag{14}$$

Academic cousins have similar origins. Thus, having many cousins means that there are many academics with similar backgrounds. This similarity considers a common ascendant or a common descendant. For the latter case, similarity among academics is defined by sharing the same academic grandchild, provided that the relationship is made through different academic child (mirror cousins). The set of inverted cousins of v is

$$C^{-}(v) = \{u \in V, \quad u \neq v : (u,w) \in E, \quad \text{where} \quad w \in F^{-}(x), \quad x \in F^{+}(z) \quad \text{and} \quad z \in F^{+}(v)\}, \tag{15}$$

and the mirror cousins' metric is

$$c^{-}(v) = |C^{-}(v)|. \tag{16}$$

It is worth emphasizing that the cousins' metric c^{+} considers the ascendants to infer similarity while the mirror cousins' metric c^{-} considers the descendants for this purpose. We chose not to adjust the metric representation to maintain the semantics related to the concept of the cousin in the context of traditional genealogy, which considers co-blood relationships. Consider an as example the vertex 29, we have the following: $C^{+}(29) = \{27, 28, 30\}$ and $C^{-}(29) = \{26, 27, 28\}$ (Fig. 1).

- *Generations*:

The direct descendants or academic children of an advisor constitute the first generation formed by that person. Consequently, the second generation corresponds to the academic grandchildren of the advisor, and the pattern continues. The number of generations of an academic is an indicative of the impact, perpetuation and evolution of his/her ideas and knowledge in the community in which s/he is inserted. Higher values represent major academic chains interconnected

Table 1
Results of the topological metrics for some vertices selected in the example graph (Fig. 1).

v	Descendancy						Ascendancy					
	$f^+(v)$	$ft^+(v)$	$d^+(v)$	$c^+(v)$	$g^+(v)$	$r^+(v)$	$f^-(v)$	$ft^-(v)$	$d^-(v)$	$c^-(v)$	$g^-(v)$	$r^-(v)$
1	1	1	4	0	3	4	0	0	0	1	0	0
4	2	0	2	0	1	2	1	1	3	0	2	3
8	1	1	2	0	2	2	1	0	1	1	1	1
12	0	0	0	0	0	0	2	2	5	0	3	6
16	3	0	3	0	1	3	3	0	3	0	1	3
20	3	3	4	0	2	6	0	0	0	0	0	0
26	4	4	10	0	2	16	1	0	1	3	1	1
29	3	1	6	3	2	6	3	3	4	3	4	7

by their mentoring relationships. The generations metric represents the size of the largest chain in the offspring of an academic, and we define it as

$$g^+(v) = \max\{k \in \mathbb{N} \cup \{0\} : \exists v \rightsquigarrow u \quad \text{of length} \quad k\}. \tag{17}$$

Generations give an idea of the topological order of academics since we are not considering any external attributes, such as the year in which the academic got his/her degree. This metric is also useful for identifying the perpetuation of academic influence over time, information that cannot be obtained from the other genealogical metrics. The symmetry of this metric in the ascendants of the academic of interest is called mirror generations (g^-), and we define it as

$$g^-(v) = \max\{k \in \mathbb{N} \cup \{0\} : \exists u \rightsquigarrow v \quad \text{of length} \quad k\}. \tag{18}$$

Consider again the vertex 29. We have the following: $g^+(29) = 2$ considering the set of edges $\{(29, 30), (30, 34)\}$ among other sets possible, and $g^-(29) = 4$ considering the set of edges $\{(28, 29), (27, 28), (26, 27), (25, 26)\}$ (Fig. 1).

- *Relationships*:

Academic groups may be more or less cohesive according to the number of relationships that interconnect them. Groups with greater cohesion presuppose academic co-mentoring, and consequently, individual knowledge flows through the links thus increasing the group experience. The metric represents the number of connections among the academics belonging to offspring of an academic of interest. Formally, we define the set of relationships as

$$R^{\pm}(v) = \{(u, w) \in E : u \in D^{\pm}(v) \cup \{v\}\}, \tag{19}$$

and the metric relationship is

$$r^{\pm}(v) = |R^{\pm}(v)|. \tag{20}$$

Academic groups that have many relationships suggest an increased flow of knowledge between them. In this context, a group is formed by the descendants of the academic of interest. For the ascendants, the semantic is the same, and the denomination of the metric is the mirror relationships. Consider as an example the vertex 29, we have the following: $R^+(29) = \{(29, 32), (29, 33), (29, 30), (30, 34), (30, 35), (30, 36)\}$ and $R^-(29) = \{(27, 29), (28, 29), (26, 29), (26, 27), (26, 28), (27, 28), (25, 26)\}$ (Fig. 1).

Individually, the metrics identify a particular characteristic regarding the performance of the academic in the formation of human resources. This feature can also be defined as a topological attribute of the graph. For example, fertility indicates the individual performance of a certain academic, as it takes into account the number of advisees who were instructed by him/her. On the other hand, fertility refers to the descendants who exercised the activity of mentoring, considering only the fecund vertices.

When considered together, the metrics can characterize the vertices globally. In this paper, we will use the metrics together (globally) to identify similar groups and classify academics according to their performance in the formation of academic advisors.

To illustrate each one of the metrics, we consider the graph in Fig. 1 as an example. The results obtained by the metrics for some of the vertices in the graph are represented in Table 1. For each line, we identify the vertex and its metric ascending and descending. Note that, the metrics based on the ascendants of a vertex are necessary to characterize those vertices that do not have descendants. Thus, it is possible to apply the method to academics with more or less experience in mentoring.

2.2. Datasets

We consider two datasets as case study: the dataset 1 is the Mathematics Genealogy Project[1] (MGP) and dataset 2 is The Academic Family Tree[2] (AFT). Both projects aim to identify and register the genealogical data of academics in web platforms. We performed data mining through recursive queries to the platforms between May and June 2017. The MGP is an initiative of the University of North Dakota and gathers data on PhDs in Mathematics and related fields (e.g., Computer Science). We identified 212,989 academic records in addition to the attributes: (i) academic name, (ii) year of the degree, (iii) the institution and (iv) the country of titling (Chang, 2011; Malmgren et al., 2010).

The AFT project brings together data about PhDs from different fields of knowledge such as Neuroscience, Astronomy, Oceanography, among others, that total more than 50 areas. We identified a total of 674,852 academic records. This platform provides an extension to the attributes described for MGP, additional information about the area of operation, and provides a connection with other platforms (David & Hayden, 2012; Tenn, 2016).

Despite the importance and extensive use, both databases may have biases. Academics fill their records on the platforms, which does not guarantee the quality of the information. Data regarding past academics are difficult to obtain and are recorded based on information that cannot be validated. Thus, the case studies presented in this paper aim to illustrate the potential characterization of the metrics and not the characterization of the datasets specifically.

3. Results

The metrics infer quantitative attributes to the vertices and represent some topological characteristics about the descent and the ascendancy of these vertices. These features reveal how the vertex connects in the graph in which it is inserted. The identification of the connection pattern of vertices allows classification by similarity, that is, to create groups of vertices that have the same topological model. The metrics were calculated for the vertices of the graph in Fig. 1, and the results of some of these vertices are presented in Table 1. These results were used to form a new graph in which two vertices that have similar values are connected by an edge. This graph of similarity is called a Gabriel graph.

Gabriel and Sokal (1969) proposed and defined the Gabriel graph as follows: any two vertices v and u are connected if and only if all other vertices are outside the circle on whose diameter v and u are at opposite points. Formally, v and $u \in V$, and $d_{(v,u)}$ is the Euclidean distance between v and u considering the respective attribute vectors, v and u are connected in the Gabriel graph if and only if:

$$d^2_{(v,u)} < d^2_{(v,w)} + d^2_{(w,u)} : \forall w \in V, \quad w \neq v, \quad w \neq u. \tag{21}$$

The Gabriel graph contains the same vertices of the genealogical graph, but the edges define the proximity between the vertices. Fig. 2 shows the Gabriel graph that represents the similarity of the vertices of the graph shown in Fig. 1. The distribution of its vertices was obtained through the Force Atlas 2 algorithm proposed by Jacomy, Venturini, Heymann, and Bastian (2014). The algorithm simulates a physical system where the vertices repel each other while the edges attract these vertices. These opposing forces converge to a configuration of the graph in which the proximity of the vertices defines communities.

The application of the algorithm of detection of communities introduced by Blondel, Guillaume, Lambiotte, and Lefebvre (2008) allowed the identification of distinct groups, where we considered a standard resolution equal to 1 (Lambiotte, Delvenne, & Barahona, 2008), which resulted in 4 clusters.

Clusters are formed according to the vertices metric values. Vertices with similar results position themselves close to each other. Each group was characterized according to the observed average for each metric and its normalized values were represented in radar charts. The normalization method used was the Min-Max. Consider $C_1, C_2, \ldots, C_q$ the clusters identified and each $C_i = \{\mu_i^1, \mu_i^2, \ldots, \mu_i^p\}$, where μ_i^j is the average of metric j in the cluster i, then the normalized average considered for the radar charts is

$$\|\mu_i^j\| = \frac{\mu_i^j - \min(\mu_q^j)}{\max(\mu_q^j) - \min(\mu_q^j)}. \tag{22}$$

Consider as an illustrative example the Gabriel graph represented in Fig. 2a and Table 2 which presents the averages of the metrics for each cluster and the respective normalized averages with their values between 0 and 1.

In Fig. 2a, the groups A (blue) and B (orange) represented 27.8% and 33.3% of the vertices, respectively and gathered the vertices that did not have detachable values. Both had intermediate results. However, in group A we observed a higher prominence for the ascending metrics, while group B had better performance in the descending metrics. On the other hand, group C (green) gathered the vertices that had the highest average values for the upward composition metrics. This group represented 19.44% of the total vertex in the graph. This was the same representation observed for group D (purple), for

[1] https://genealogy.math.ndsu.nodak.edu/ accessed on September 23, 2017.
[2] https://academictree.org/ accessed on September 23, 2017.

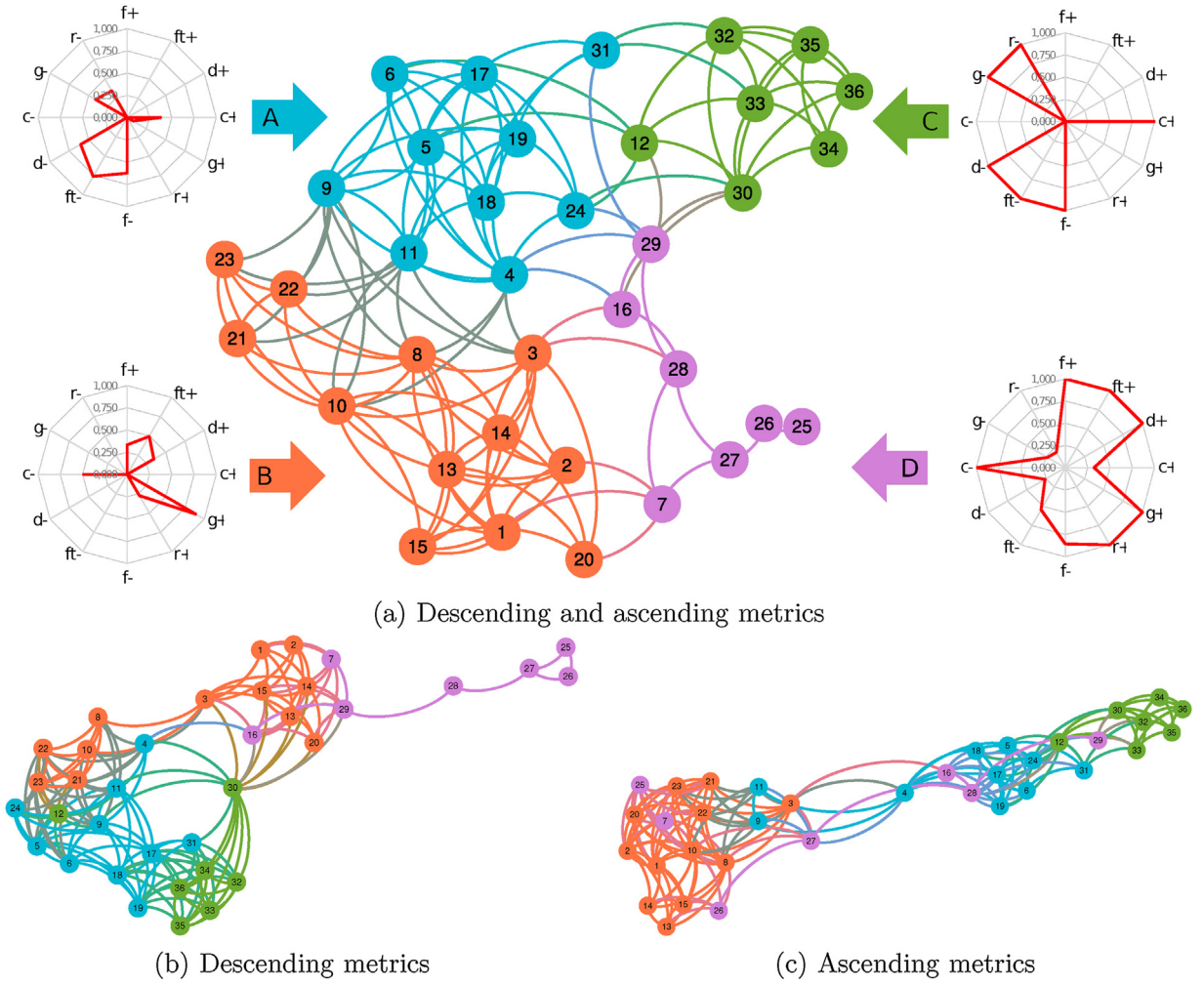


Fig. 2. Gabriel graph obtained from the topological metrics of the vertices belonging to the genealogical graph of Fig. 1. Graph (a) considers all 12 metrics. The other two graphs (b) and (c) identify descending and ascending metrics respectively. Radar charts present the (normalized) metric means for each identified group.

Table 2
Description of the average of metric in each cluster and the respective normalized average in the example graph.

Metrics	Clusters							
	A		B		C		D	
	μ	$\|\|\mu\|\|$	μ	$\|\|\mu\|\|$	μ	$\|\|\mu\|\|$	μ	$\|\|\mu\|\|$
f^+	0.40	0.00	1.17	0.33	0.43	0.01	2.71	1.00
ft^+	0.00	0.00	0.92	0.49	0.00	0.00	1.86	1.00
d^+	0.40	0.00	2.83	0.35	0.43	0.00	7.43	1.00
c^+	1.50	0.38	0.00	0.00	4.00	1.00	1.29	0.32
g^+	0.30	0.08	1.92	0.89	0.14	0.00	2.14	1.00
r^+	0.40	0.00	3.00	0.27	0.43	0.00	9.86	1.00
f^-	1.20	0.62	0.58	0.00	1.57	1.00	1.43	0.86
ft^-	1.20	0.76	0.00	0.00	1.57	1.00	0.86	0.55
d^-	3.50	0.60	0.58	0.00	5.43	1.00	1.86	0.26
c^-	0.00	0.00	0.83	0.49	0.00	0.00	1.71	1.00
g^-	2.40	0.41	0.50	0.00	5.14	1.00	1.57	0.23
r^-	3.80	0.35	0.58	0.00	9.86	1.00	2.43	0.20

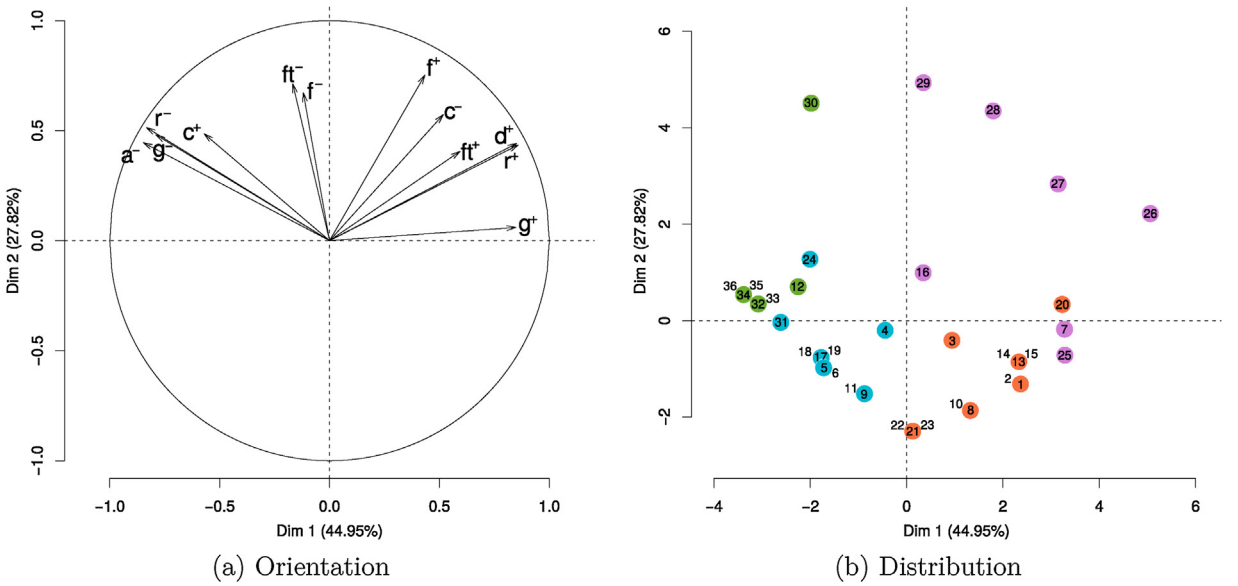


Fig. 3. Principal Component Analysis (PCA) applied to the vector of quantitative attributes (topological metrics) of the vertices in example graph (Fig. 1). Diagram (a) shows the orientation of the axes. Diagram (b) shows the distribution of the vertices as a function of the two principal components.

which the average values of the descending composition metrics were evident. The c^+ considered the ascendancy for the identification of cousins, so it was necessary to rely on this feature for the interpretation of groups.

The grouping provided by the Gabriel graph represents the vertices that were evident in their descending and ascending metrics. That is, the B and D groups were formed by the vertices with the most relevant offspring. This quantitative relevance was the most substantial in Group D, where we identified the vertices that contributed most to the formation of the graph. On the other hand, the vertices that show the ascending metrics located in groups A and C. Group C had the highest results.

The descending composition metrics appeared to contribute more significantly to the group presented. Fig. 2b and c represents the Gabriel graphs by individually considering the descending and ascending metrics, respectively. The configuration remains constant in the graph of the descending metrics compared to the graph of the ascending metrics.

The identification of communities according to the performance of their academics in the training of human resources is essential for the classification of homogeneous groups of advisors according to their quantitative attributes. These attributes can be considered for the individual ranking of these academics. In this paper, we used principal component analysis (PCA) for this purpose.

The PCA is a traditional mathematical procedure that aims to reduce the dimensionality of multidimensional data (Hair, Black, Babin, Anderson, & Tatham, 1998). The dimensions obtained through the application of PCA are called principal components. The first significant component concentrates the most substantial variance of the data. The other components address the greater variation, being orthogonal to the previous ones.

Algebraically the PCA is a linear combination of the original variables or, geometrically, the coordinates of the sample points on new axes that are the result of the rotation of the axes, in the direction of maximum variation. The main components are obtained using a rotation matrix that makes it possible to rotate the original coordinate system.

Fig. 3 shows the result of the application of PCA to the vectors of quantitative attributes,[3] obtained through the topological metrics, of the vertices of Fig. 1. The first component (or dimension) concentrates 44.95% of the data variance, while the second component accounts for 27.82%. The axes are represented in Fig. 3a. The first quadrant (upper right) of the diagram concentrates the descending composition metrics while the ascending composition metrics are in the second quadrant.

The distribution of the vertices in the diagram, shown in Fig. 3b, follows a pattern very similar to that observed in the Gabriel graph. In the first quadrant (upper right) are the vertices identified as group D in the Gabriel graph, see Fig. 2. The other quadrants (counterclockwise) are consistent with groups C, A, and B, respectively. Considering the distribution analyzed, it is possible to obtain a classification of the vertices in one of the principal components, for example, by defining an ordering of the vertices of Fig. 3b. To each vertex we associate its point of location on (x, y)-plane and we use the following order relation between any of the vertices: let u and v two vertices with coordinates (x_1, y_1) and (x_2, y_2), respectively. We say that $u \preceq v$ if $x_1 < x_2$. If $x_1 = x_2$, no matter the values of y_1 and y_2, we say that $u = v$. With this relation we are prioritizing

[3] The vector values were normalized by using the logarithm to improve the distribution of vertices.

Table 3
Clusters identified in similarity graphs with their respective numbers of vertices.

Cluster	Dataset 1		Dataset 2	
	Vertices	Percentage	Vertices	Percentage
A	2114	19.69%	1992	14.95%
B	6652	61.95%	4899	36.77%
C	821	7.65%	502	3.77%
D	1151	10.72%	5931	44.51%

descending composition metrics. Taking this relation into account, we have the following ordering of the vertices of Fig. 3b (see also Fig. 1):

$$34 \preceq 35 \preceq 36 \preceq 32 \preceq 33 \preceq 31 \preceq \ldots \preceq 13 \preceq 27 \preceq 20 \preceq 7 \preceq 25 \preceq 26.$$

Each component can be used as the basis for classification, or we can consider the orientation of an individual metric that, in a specific context, is a priority for the ordering of vertices.

For both the cluster analysis (Gabriel graph) and the two-dimensional distribution (PCA), the proposed metrics were able to differentiate the vertices considering the topology of the graph. The same method was applied to genealogical datasets of greater magnitude. As already mentioned, we used two databases for characterization of topological metrics such as case studies. The MGP (dataset 1) and the AFT (dataset 2) are collections of academic genealogical data widely exploited in many scientific papers. However, few publications consider the development and/or application of topological metrics for the characterization of academic data. These data are considered in contexts where senior academics are honored (honorary genealogy) or when young academics seek to identify their lineages to identify famous people who are their ancestors (egocentric genealogy). We consider the use of this type of data essential for the production of relevant knowledge regarding different academic-scientific communities (analytical genealogy).

The considered datasets are essential to register the continuation of academics in their respective areas. However, the data obtained cannot be consistently validated. These data have some degree of bias. In this context, the evaluation of the datasets is presented globally to verify the ability of the metrics to characterize genealogical data. More accurate data are necessary to describe the academics.

3.1. Case studies according to graphs of similarity

The objective of this work is those academics of greater relevance regarding academic advisor–advisee relationships. Thus, for the structuring of the Gabriel graph, we used a subset of vertices belonging to the databases. The vertices considered were those that had non-zero values for all their metrics. The subset belonging to the dataset 1 had 10,738 vertices (5.04%) with non-zero results, while the subset extracted from dataset 2 considered 13,324 vertices (1.97%). The metrics calculation considered the complete databases. The method described above to obtain the Gabriel graph was applied to datasets 1 and 2. To allow an adequate comparison between the graphs, we considered four clusters in both cases, similar to the previous example. Table 3 describes the number of vertices in each cluster.

Both graphs presented clusters consistent with the one observed in the previous example (Fig. 2). For the graph obtained from dataset 1 (Fig. 4), the groups C and A were distinct from the others regarding the descending composition metrics. The first group includes the vertices that had the highest value for this set of metrics, besides being also prominent in metrics f^- and ft^-. The less representative vertices belonged to the second group. The configuration observed in group C indicated greater differentiation of its vertices, which were organized in a linear pattern. A similar interpretation was made for these groups in the graph of dataset 2 (Fig. 5). The vertices highlighted by their ascendant composition metrics were concentrated in groups B and D. The pattern observed in the previous example repeated for this case. Group B was formed by the vertices with the highest values of metrics and the group D by those with the lowest results. This configuration did not repeat for the graph created from dataset 2. In this case, the characterization of both groups did not distinguish the descending and ascending metrics. This configuration can be due to the bias observed for this dataset.

By comparing the results of the metrics for datasets 1 and 2, as shown by the boxplots in Figs. 4 and 5, we observed that there was a pattern of proportional similarity between the results, considering that the genealogical data commonly have a distribution consistent with the power law. In both cases, there was a higher number of descendants and relationships in the ancestry of vertices (d^- and r^-) than in the offspring (d^+ and r^+). This observation shows that the selected groups belong to the most recent generations since there are more academics and relationships in the ancestry of the graphs.

3.2. Case studies according to PCA

Topological metrics are useful for identifying and grouping similar academics, as seen through the structuring of Gabriel's graph and the identification of communities (Fig. 2). Another possibility of analysis is to consider the distribution of the vertices in the multi-dimensional space and, using a rotation of the axes, to obtain a bi-dimensional representation which allows the higher variance. This last form of analysis is performed through PCA. The distribution obtained using PCA was

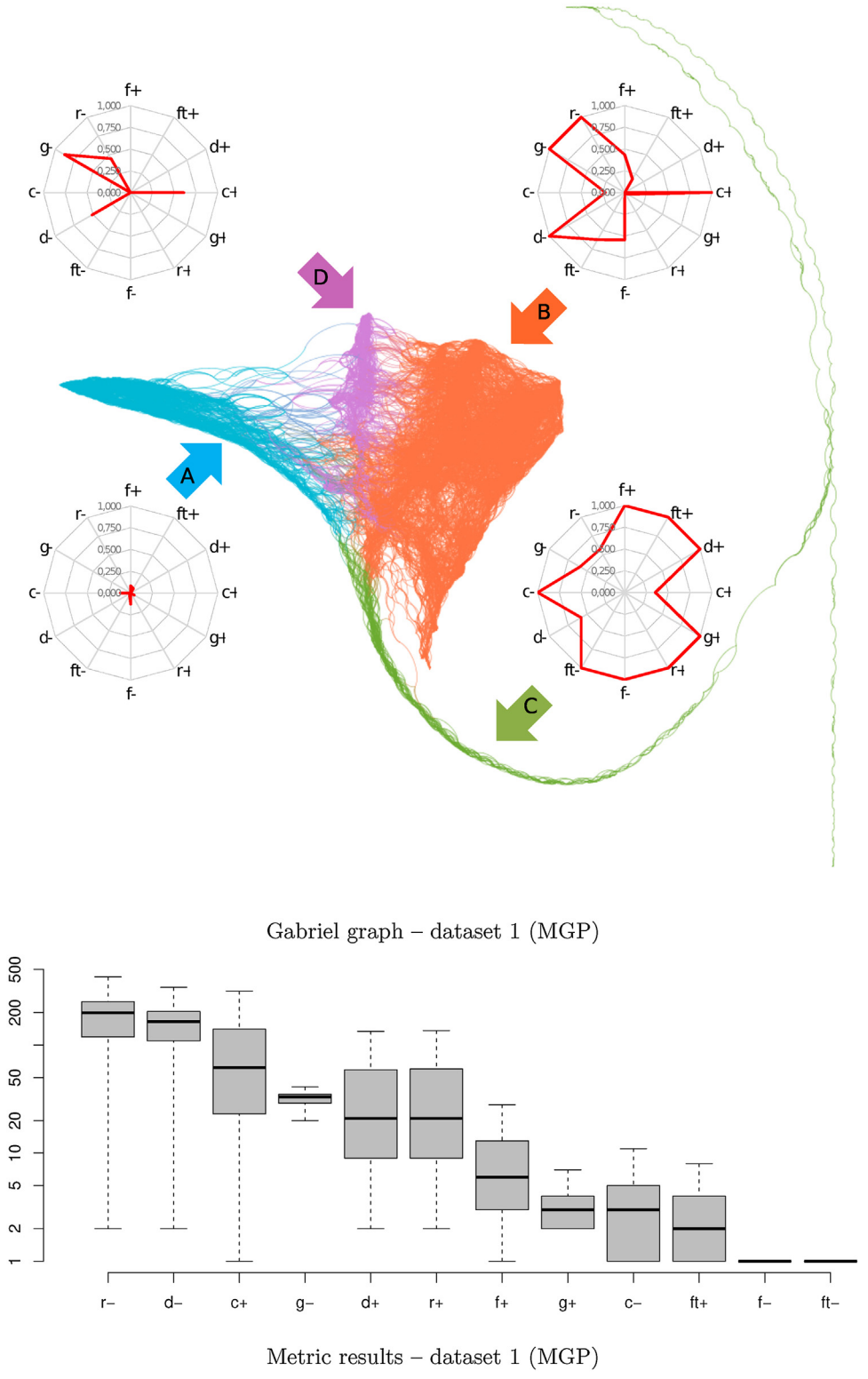
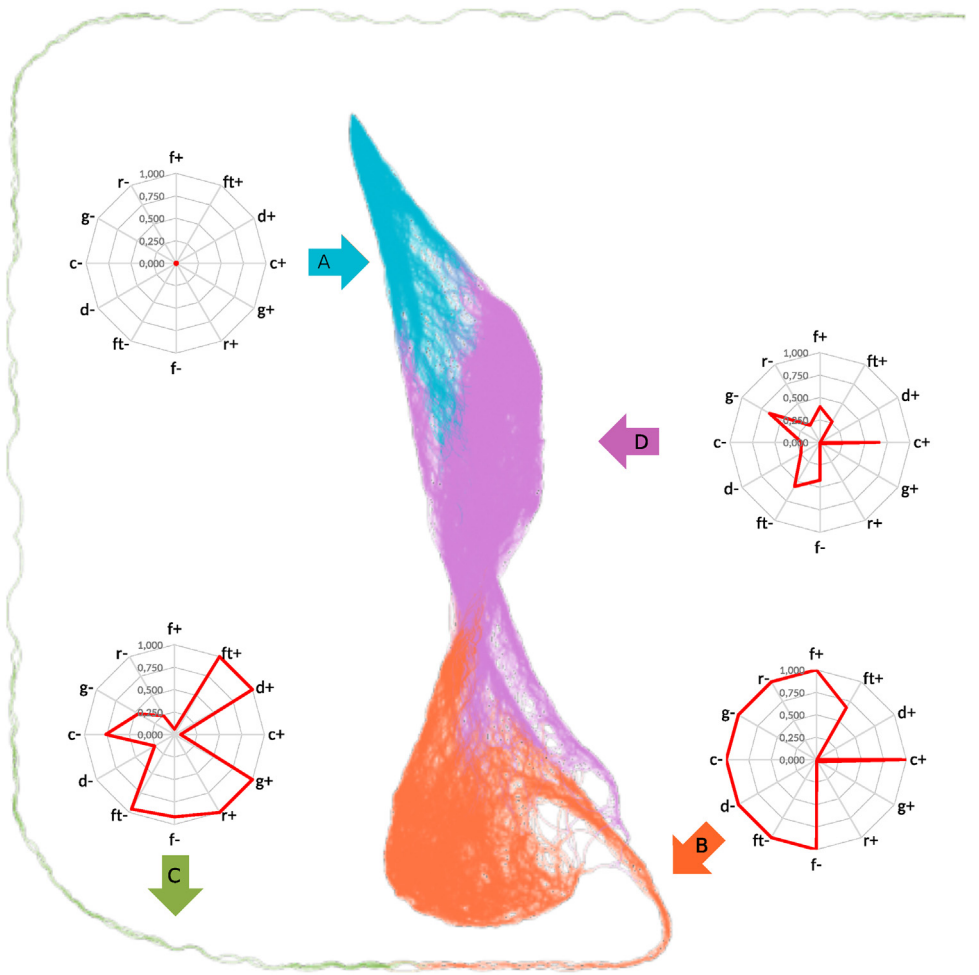
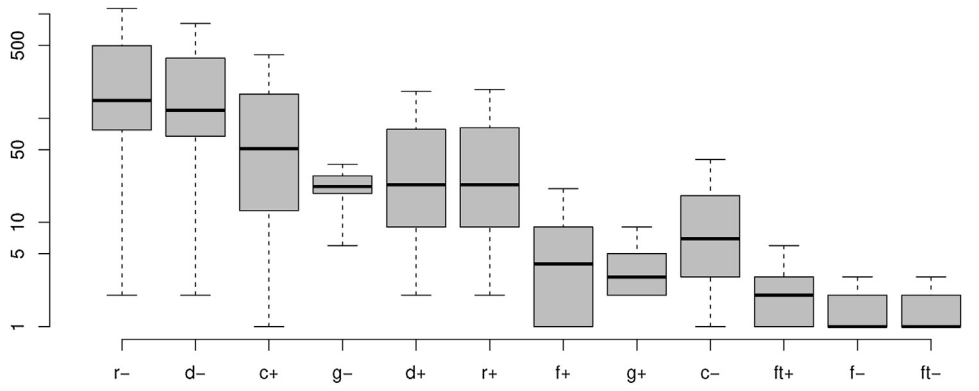


Fig. 4. Similarity graphs (Gabriel's graphs) for dataset 1. The groupings are represented by the different colors. The radar charts refer to the mean observed for the topological metrics in each cluster, representing the lowest (0) observation and the highest (1) observation. Boxplots describe the distribution of results for each metric. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)



Gabriel graph – dataset 2 (AFT)



Metric results – dataset 2 (AFT)

Fig. 5. Similarity graphs (Gabriel's graphs) for dataset 2. The groupings are represented by the different colors. The radar charts refer to the mean observed for the topological metrics in each cluster, representing the lowest (0) observation and the highest (1) observation. Boxplots describe the distribution of results for each metric. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

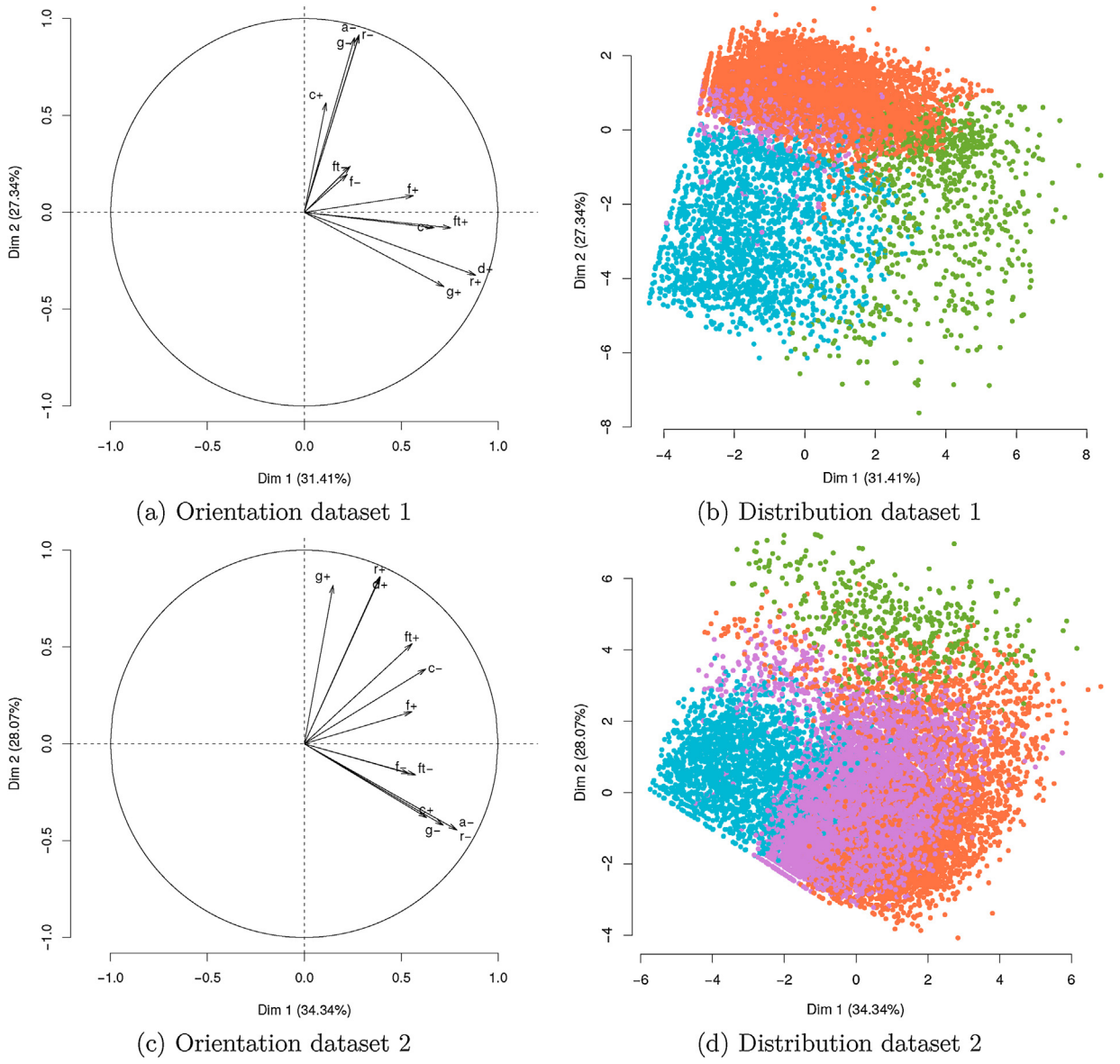


Fig. 6. Principal Component Analysis (PCA) for the case studies considered. Diagrams (a) and (b) show the orientation of the metrics and the distribution of the vertices for dataset 1, respectively. The results for dataset 2 are plotted in diagrams (c) and (d). The vertices in both representations are labeled with the colors that agree with the identified clusters (Figs. 4 and 5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

consistent with the configuration observed in the clusters. This enables us to define a classification for each vertex. For dataset 1, the representation concentrated 48.75% of the total variance of the data, and the orientation of the vertices follows according to the composition of the metrics. The descending metrics scale around the first component and the ascending metrics around of the second component (Fig. 6a).

The vertices represented in Fig. 6b were labeled with the same colors used in the Gabriel graph groups. Thus, it was possible to observe that the distribution obtained was naturally coherent with the clustering. Group C that was described as having the highest means for the descending metrics (green vertices) had its vertices distributed consistently with the previous characterization. These vertices distributed themselves more sparsely than the others, evidencing the singularity that the metrics have given them.

Considering dataset 2 the results were similar. In Fig. 6c the axes defining the orientation of the metrics were configured correlative to that observed for dataset 1. The main difference between the two configurations was the rotation used for the axes. Since the PCA seeks a representation in which there is a prioritization of the components that concentrate the most

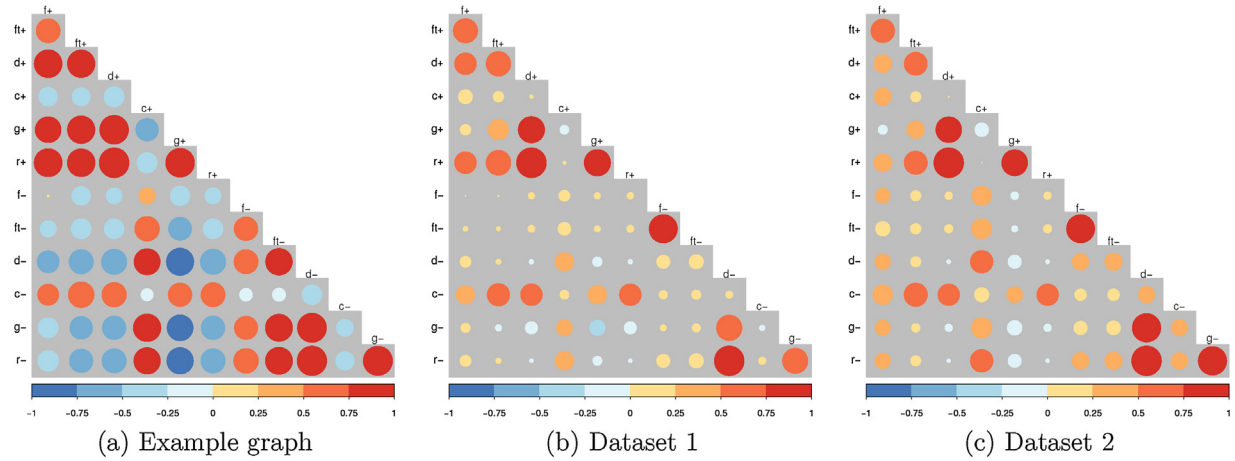


Fig. 7. Correlation coefficients between the results of the topological metrics, for the three cases considered in this paper. The rows and columns represent the metrics, the intersections between them contain a circumference that has a size proportional to the correlation coefficient of that pair and is identified by the color scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

significant variance, which for this dataset was 67.29%, the configurations differ only at this point. However, the orientation pattern of the axes was very close between the two settings.

The distribution of the vertices in the two main components also presented a result consistent with dataset 1 (Fig. 6d). However, the groups previously identified and labeled with their respective colors were less defined in this distribution when compared to the previous one. This divergent point, as for the pattern, can also be attributed to the bias embedded in this dataset, similarly observed in the Gabriel graph.

The axes configuration observed in Figs. 3a, 6a and c followed a pattern where those that defined the orientation of the descending composition metrics were grouped separately from those that guided the ascending composition metrics. Except for c^+ and c^- that present distinct semantics from the others, as discussed above. This configuration of the axes was invariant about the size of the datasets considered.

3.3. Correlation among metrics

A metric is efficient to characterize academic genealogy graphs if the correlation between it and the others is low. A pair of metrics that has a high correlation coefficient indicates that they are not able to differentiate the vertices of the graph. Thus, we look for metrics that do not correlate results, which increasing the capacity of characterization of the metrics.

For the vertices in the example graph (Fig. 2a), the results of the metrics presented high correlation coefficients in most cases (Fig. 7a), in 66.67% of the cases the correlation coefficient was less than −0.5 or greater than 0.5 (high correlation). For the vertices in dataset 1 (Fig. 4), 22.73% of the correlation coefficients (Fig. 7b) presented high correlation. Finally, for the vertices in dataset 2 (Fig. 5), 25.76% of the correlation coefficients (Fig. 7c) showed a high correlation.

Some metrics are correlated naturally with each other. For example, fecundity and fertility have high correlation coefficients for both offspring and ancestry. These cases occur because these variables are dependent on each other. On the other hand, the option to maintain them is justified by the particular semantics that these metrics provide. Thus, even if they are redundant, these metrics provide precise information that is useful for the analysis of genealogy graphs. We considered the Spearman method for the calculation of correlation coefficients.

The high correlation coefficients observed for the example graph were justified due to the limited number of vertices which did not cover the various configurations possible in a real genealogy graph. Note that, for graph based on examples from real data, there was a stable pattern for the observed correlation coefficients. Therefore, the proposed topological metrics were useful for the characterization of the data considered in this paper.

4. Discussion

The clusters identified in the similarity graphs (Gabriel's graphs) are the result of an unsupervised learning process, where the objective is to determine (learn) the classes of objects. The characterization (labeling) of the classes is a later and independent process and, in the context of this work, is done only through the proposed metrics without considering any external attribute to the topology of AG graphs. An important point, as we are presenting a way to obtain quantitative characteristics (metrics) that are based on the graph structure. The analyses carried out are intended to describe the effectiveness of these attributes and not to characterize the databases used.

The characterization of clusters uses the metrics results themselves so that we can maintain the independence of external attributes. From the observation of the three graphs of similarity is possible to note that the descendants and the ascendants

Table 4
Topological metrics of typical academics in each dataset (MGP and AFT). The academics (real examples) were selected from datasets considering the lowest Euclidean distance to each cluster.

	Real example	Topological metrics											
		f^+	$ft+$	d^+	c^+	g^+	r^+	f^-	ft^-	d^-	c^-	g^-	r^-
MGP	Cluster A	8	2	35	36	3	36	1	1	13	4	9	14
	Michel A. Melkanoff	9	3	34	35	3	34	2	2	11	10	9	12
	Cluster B	10	3	44	132	3	46	1	1	190	5	34	233
	Harold I. Levine	8	1	43	123	4	44	1	1	188	2	35	233
	Cluster C	14	7	17,712	70	12	19,460	1	1	115	11	24	139
	Johannes Heurnius	2	2	18,240	10	27	19,286	2	2	59	3	15	68
	Cluster D	7	2	20	94	3	21	1	1	104	3	31	113
	Joanne Elliott	4	2	22	93	2	22	1	1	111	2	32	120
AFT	Cluster A	5	2	20	23	3	21	1	1	11	6	7	11
	Paula Fass	8	2	22	26	2	22	1	1	14	3	6	14
	Cluster B	9	4	297	197	4	313	2	2	408	34	28	546
	Leo Postman	8	7	287	205	6	297	3	3	431	41	27	563
	Cluster C	5	4	27,367	35	17	34,655	2	2	112	27	17	140
	Johannes van Horne	3	2	27,057	21	21	33,794	2	2	45	3	16	50
	Cluster D	6	2	70	137	4	72	2	1	108	12	21	130
	James B. Hendrickson	6	4	67	154	4	68	1	1	111	16	21	138

of the vertices seem to govern the structuring of the clusters. An example of this influence is the clusters C and D in Fig. 2 (example graph). This occurs because the correlation coefficients between these two subsets of metrics are low, as shown in Fig. 7. However, if we extend our observation to the other clusters, is possible to note that there are other intermediate clusters in both descendants and ascendants. Clusters B and D (Fig. 2) are governed by descendants, and the scale of this subset of metrics defines these partitions.

The interpretation of each identified cluster depends on the particular semantics of the metrics (as detailed in Section 2.1) and their respective values. Considering cluster A in the similarity graph referring to dataset 1 (MGP – Fig. 4 and Table 4), a typical academic of this group has on average eight advisees (f^+), of which only two carried out academic mentoring (ft^+). The influence of this academic propagates for three generations (g^+), which bring together 35 descendants (d^+) that are interconnected by 36 mentoring relationships (r^+). Four other academics share the same "*grandchildren*" as the typical academic (c^-).

On the other hand, considering the ascendancy, the typical academic of the cluster A has only one advisor (f^- and ft^-). Their ancestors add up to 13 academics (d^-) distributed in nine generations (g^-) that are interconnected by 14 mentoring relationships (r^-). Finally, 36 other academics share the same "*grandfathers*" as the typical academic (c^+). Table 4 presents the typical academics for all clusters identified in the two datasets. Additionally, we show an example of a real academic whose metrics are more close to those of the regular academic, considering Euclidean distance.

An interesting feature to observe is the elongated tails present in both similarity graphs (Figs. 4 and 5). The academics who make up these tails are not similar to other academics. The structuring method of the similarity graph predicts that each vertex will be connected with at least one other vertex, to result in a single connected component. Thus, academics who are dissimilar are connected to the closest one, which sometimes means a single vertex. The identification of which metrics contributed to making these vertices different from the others can be made using the respective distributions obtained by the PCA (Fig. 6). Note that the green vertices that form the tails are represented in the distributions (Fig. 6b and d) with this same color. Although there is a difference in orientation of the axes (Fig. 6a and c), it is possible to observe that the green vertices are differentiated mainly by the descending metrics. The difference of the orientation of the axes is due to the PCA that looks for the two main components that represent the higher variance of the data.

The metrics c^+ and c^- are particularly less intuitive to be analyzed than the other ones. Considering the same example, the typical academic in cluster A has 36 cousins. This result indicates the magnitude of the family in which the typical academic is inserted. This information opens the possibility of comparing performances among academics who have similar ancestry or with their ancestor, whose performance is evidenced by the metrics. Another important feature of these metrics is that they consider the peers of the analyzed academics who are their contemporaries in the same temporal line, while the other measures consider the past or the future.

From the comparisons of the clusters in datasets 1 and 2 is possible to note that there are some divergences in their compositions, which can be attributed to the bias that exists in these data types. We can highlight the incompleteness of the

data as the main bias, and it is observed in both databases. Notably, the MGP is a homogeneous database that considers only relationships at the doctoral level. The historical bias, in this case, is the major problem, since there are records from a time that there was no formalization of scientific fields as is currently known. On the other hand, AFT constitutes a heterogeneous database because it considers different types of relationships, such as postdoctoral studies. This type of bias can cause a density of AG graph, which directly impacts the results of the metrics. In spite of the biases that the data content, the metrics demonstrate that they can capture the topological pattern that these types of structures present, such as the distribution of their values presented in Figs. 4 (Metric results – dataset 1, MGP) and 5 (Metric results – dataset 2, AFT).

Structuring similarity graphs and identifying clusters allows different vertex classes to be created, which can, for example, be used to label a new vertex through the result of its metrics. On the other hand, this method does not allow the classification of the complete set using a topological ordering. Thus, the purpose of the PCA application is to reduce the dimensionality of the data and to provide the desired classification using one of the main components. Considering Fig. 3b, a possible ordering would be the projection of the vertices in the first principal component. In this case, analyzing the vertices from left to right in the first significant component define prominence in ascending and descending metrics, respectively. As in similarity graphs, the interpretation of topological ordering is challenging to perform and depends on the observation of the axis that guides the distributions (Fig. 3a).

5. Conclusions

In this paper, we presented a study on metrics that represented relevant characteristics regarding the topology of genealogy graphs. In this context, AG is a fertile field for obtaining information on the configuration of scientific knowledge and the identification of evolutionary patterns of academic communities from advisor–advisee relationships. These studies can be considered to complement the analyses where the scholarly production of the researchers is found, which provides a new way to evaluate the academic communities and their members.

The analyses showed that the metrics are useful to differentiate academics according to their genealogical characteristics. These metrics can be referred to as primitive because they represent simple concepts about genealogy graphs but are useful in the characterization of graphs and present a robust semantic intuition, precisely because of their simplicity. As future work, we intend to combine different primitive metrics to compose derived metrics that consolidate diverse topological features in a single measure.

For the datasets considered there was an observable pattern regarding the proportionality of the results of the topological metrics. This model was reflected in the formation of academic communities, the distribution of vertices and the results of the metrics. There was a correlation between the metrics that had a similar configuration in both datasets, despite of bias. In the field of research that has AG as its goal, many challenges are still faced. The development of methods that allow genealogical data mining and its validation considering different databases can provide the structuring of more assertive genealogical graphs. Reducing the bias of the data found in the composition of these structures is essential for the quality of the expected results (Damaceno, Rossi, & Mena-Chalco, 2017).

Finally, topological metrics can infer a genealogical graph growth analyses, identifying relevant paths to the flow of scientific knowledge and to the analysis of the similarity between different genealogical structures. The possibility of aggregating external attributes to AG graphs, such as specific areas of expertise of academics, is especially crucial for obtaining graphs of knowledge topics. These last ones reflect the way in which the topics relate, considering the hierarchy provided by AG. The relationships between knowledge topics can highlight: (i) the life cycle of the scientific topics and disciplines, (ii) the emergence or disappearance of specific areas, and (iii) the pattern of unfolding a topic in others with more celebrated specialty.

Author contributions

Conceived and designed the analysis: Luciano Rossi; Jesús P. Mena-Chalco
Collected the data: Luciano Rossi; Rafael J.P. Damaceno
Contributed data or analysis tools: Luciano Rossi; Rafael J.P. Damaceno; Igor L. Freire; Etelvino J.H. Bechara; Jesús P. Mena-Chalco
Performed the analysis: Luciano Rossi
Wrote the paper: Luciano Rossi; Rafael J.P. Damaceno; Igor L. Freire; Etelvino J.H. Bechara; Jesús P. Mena-Chalco
Mathematical review: Igor L. Freire
English review: Etelvino J.H. Bechara
Content review: Jesús P. Mena-Chalco

Acknowledgments

The authors would like to express their thanks to CNPq and CAPES for financial support for this work. We are thankful to the anonymous reviewers for their valuable comments and suggestions which improved considerably our paper. We are also indebted with Professor Ronaldo Cristiano Prati for his useful suggestions.

Appendix

To illustrate the distribution of academics according to PCA, Fig. 8a and b shows the identification of these individuals considering the first component. For dataset 1, this component represents the prominence in the descending composition topological metrics (mostly). On the other hand, for dataset 2, the highlight shows the ascending composition metrics. For each scholar we show his name and the country of origin (flag) (Fig. 9).

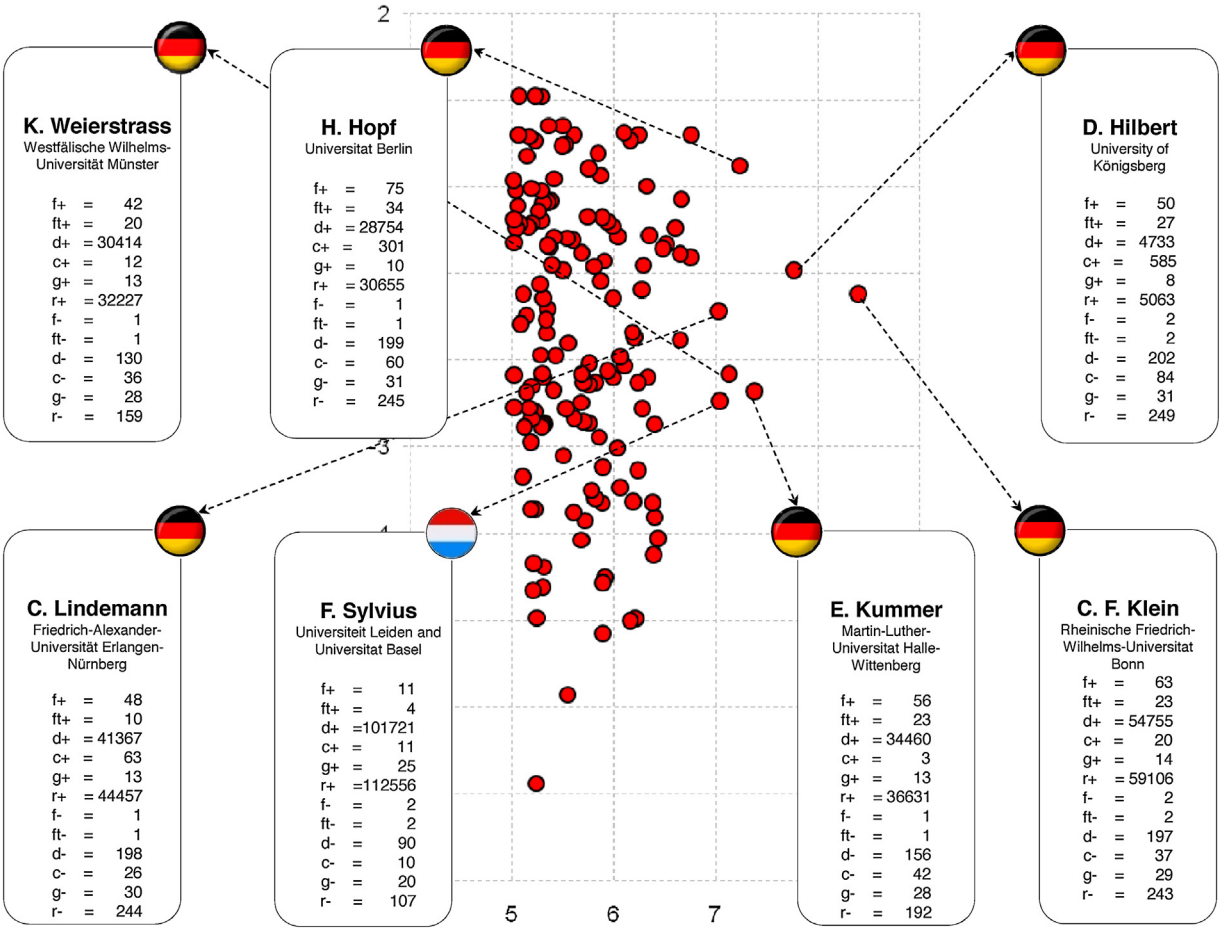


Fig. 8. Academics belonging to dataset 1 (MGP) with prominent offspring, according to the distribution of the first principal component (horizontal axis).

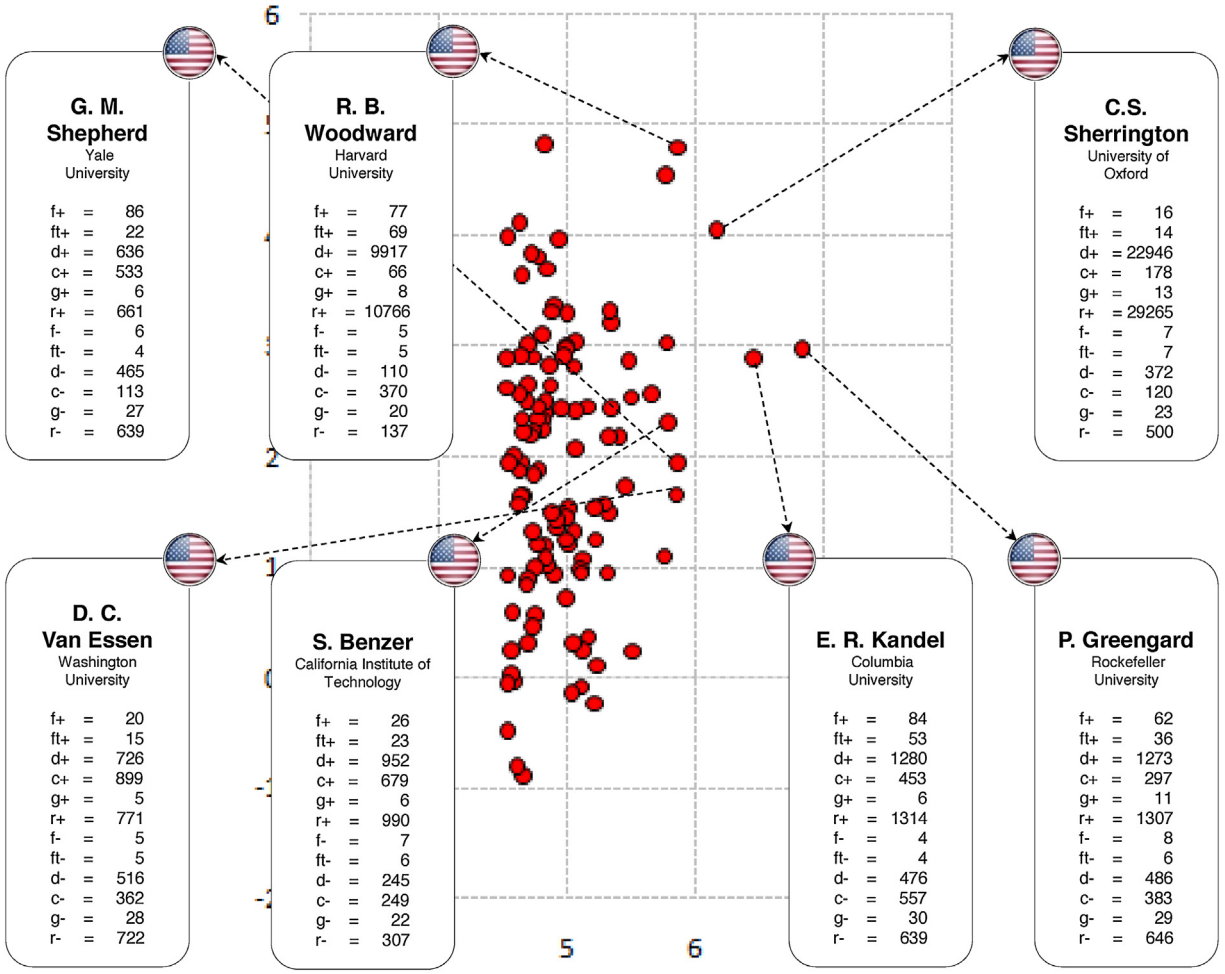


Fig. 9. Academics belonging to dataset 2 (AFT) with prominent offspring, according to the distribution of the first principal component (horizontal axis).

References

Agarwal, A., Durairajanayagam, D., Tatagari, S., Esteves, S. C., Harlev, A., Henkel, R., et al. (2016). Bibliometrics: Tracking research impact by selecting the appropriate metrics. *Asian Journal of Andrology*, *18*, 296.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008.

Chang, S. (2011). *Academic genealogy of mathematicians*. World Scientific.

Damaceno, R. J., Rossi, L., & Mena-Chalco, J. P. (2017). Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes. In *Proceedings of the 32nd Brazilian symposium on databases*

David, S. V., & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE*, *7*, e46608.

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, *7*, 861–873.

Gabriel, K. R., & Sokal, R. R. (1969). A new statistical approach to geographic variation analysis. *Systematic Zoology*, *18*, 259–278.

Gargiulo, F., Caen, A., Lambiotte, R., & Carletti, T. (2016). The classical origin of modern mathematics. *EPJ Data Science*, *5*, 26.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (1998). . *Multivariate data analysis* (Vol. 5) Upper Saddle River, NJ: Prentice Hall.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE*, *9*, e98679.

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2008). *Laplacian dynamics and multiscale modular structure in networks.* , arXiv:0812.1770.

Malmgren, R., Ottino, J., & Amaral, L. (2010). The role of mentorship in protégé performance. *Nature*, *465*, 622–626.

Rossi, L., Freire, I. L., & Mena-Chalco, J. P. (2017). Genealogical index: A metric to analyze advisor–advisee relationships. *Journal of Informetrics*, *11*, 564–582.

Sugimoto, C. R. (2014). Academic genealogy. In B. Cronin, & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (1st ed., Vol. 5, pp. 365–382). MIT Press.

Sugimoto, C. R., Ni, C., Russell, T. G., & Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science. *Journal of the American Society for Information Science and Technology*, *62*, 1808–1828.

Tenn, J. S. (2016). Introducing astrogen: The astronomy genealogy project. *Journal of Astronomical History and Heritage*, *19*, 298–304.

Todeschini, R., & Baccini, A. (2016). . *Handbook of bibliometric indicators* (Vol. 1) Wiley-VCH Verlag GmbH & Co. KGaA.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010). A comparison of a graph database and a relational database: A data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference* (p. 42).

Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., et al. (2010). Mining advisor–advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 203–212).