



ARTICLE



<https://doi.org/10.1057/s41599-023-01947-w>

OPEN

# A mathematical model for the process of accumulation of scientific knowledge in the early modern period

Maryam Zamani <sup>1,2</sup>✉, Hassan El-Hajj <sup>1,2</sup>, Malte Vogl <sup>2,3</sup>, Holger Kantz<sup>4</sup> & Matteo Valleriani <sup>1,2,5,6</sup>✉

In the present work, we model the diffusion of scientific knowledge embodied in the textbooks of the *Sphaera* corpus. This corpus consists of more than 350 different editions of textbooks used for teaching astronomy in European universities during the early modern period. Connections between the editions are based on mutual semantic knowledge and are arranged in a multiplex network of four layers, with each layer representing a different semantic relation. The modeling aims for a better understanding (and possible prediction) of the process of knowledge accumulation in the various editions. We consider semantic text-parts as knowledge units transferred between the editions, and show how these units spread using both an SI model and its modified version, the Bass model. Both models include a parameter representing the rate of transfer, which is interpreted as the mechanism underlying the process of knowledge accumulation; the Bass model has an extra parameter that represents the rate of external influence and stands out as the “resistance to adoption from” and “resistance to be influenced by” other knowledge systems. The modeling has helped us to chart the path and mechanisms of knowledge transformation in the early modern period. Networks are identified by adding further layers whose graphs express socioeconomic relationships and conditional sub-networks. The comparison between the model and these layers enables us to conclude that the accumulation of knowledge was highly dependent on the institutional embedding of scientific production because the diffusion of knowledge was mostly determined by the economic constraints of early modern printers and publishers. It further suggests that geographic proximity played a role—although secondary—in the diffusion of knowledge but only under the condition that the book producers involved were still living. The transformation of early modern scientific knowledge is, therefore, highly dependent on the institutional and economic contexts of the book producers.

<sup>1</sup> BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany. <sup>2</sup> Max Planck Institute for the History of Science, Boltzmannstr. 22, 14195 Berlin, Germany. <sup>3</sup> Max Planck Institute for Geoanthropology, Kahlaische Straße 10, 07745 Jena, Germany. <sup>4</sup> Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, 01187 Dresden, Germany. <sup>5</sup> Institute of History and Philosophy of Science, Technology, and Literature, Faculty I—Humanities and Educational Sciences, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany. <sup>6</sup> The Cohn Institute for the History and Philosophy of Science and Ideas, Faculty of Humanities, Tel Aviv University, P.O.B. 39040, Ramat Aviv, Tel Aviv 6139001, Israel. ✉email: [zamani.marya@gmail.com](mailto:zamani.marya@gmail.com); [valleriani@mpiwg-berlin.mpg.de](mailto:valleriani@mpiwg-berlin.mpg.de)

## Introduction

Research in the field of knowledge economy has convincingly argued that modernity is characterized by a process of knowledge accumulation that has lasted at least three centuries (Mokyr, 2005). Current evolutionary epistemology, as grounded on the work of Karl R. Popper (Popper, 1972), increasingly discloses how the most determinant parameter for the accumulation of knowledge is its potential to circulate (Castells, 2010; Clark, 1998; Marginson, 2016; Renn, 2020). In particular, studies in knowledge economy referring to technology's capacity to innovate are now able to show a direct correlation between the circulation of knowledge within an urban area and potential to innovate of the latter (Consoli and Patrucco, 2003).

The present work contributes to this debate by linking it to the history of science and by presenting, for the first time, a mathematical model of knowledge accumulation based on an extensive and detailed dataset collected through years of analysis of historical sources. The sources are textbooks used at European universities from the 15th to the 17th century for introductory classes on astronomy, thus enabling this discipline to be traced as it developed over time. We model the spread of knowledge and, correspondingly, we interpret it as a factor that determines the rate of knowledge accumulation. The process shows a non-linear dynamic in the time interval under consideration.

In linking knowledge circulation to knowledge accumulation, we refer to the conceptual architecture first developed by Stephen Cole in 1983 (Cole, 1983). In particular, we refer to knowledge economy processes concerning scientific “core knowledge” as distinguished from “research-frontier knowledge.” The concept of scientific core knowledge is specifically designed to denote widely accepted fundamental knowledge; it is characterized by a finite number of transparent assumptions, principles, and theories (always referring to a specific discipline). In this research, as in Cole's, such knowledge is best represented by what is widely taught at educational institutions. Cole's aim was to describe a mechanism of knowledge economy that would allow a transparent definition of how scientific disciplines are organized hierarchically. He described a) the principle defining a higher position in the hierarchy as proportional to the degree of consensus and to the acceptance of the fundamental theories of a given discipline, and b) following Thomas Kuhn (Kuhn, 1970), the accumulation of knowledge in one disciplinary area as being more effective the higher the position of that disciplinary area in the hierarchy. He therefore concludes that accumulation of knowledge takes place especially in reference to core knowledge and it depends on the degree of consensus.

Our goal, unlike Cole's, is not to find a method for compiling hierarchies among scientific disciplines but to quantify the accumulation of knowledge by modeling it. The link between our argument and Cole's conceptual architecture is what we call the homogenization of knowledge (see Section 2 of the Supplement). As has been shown elsewhere, by analyzing a series of multiplex networks (Valleriani et al., 2019, 2022; Zamani et al., 2020), from the end of the 1530s, the core knowledge of astronomy underwent profound homogenization whereby students throughout Europe were increasingly taught the same theories and concepts. We describe this process by showing the increasing homogenization of the content of the textbooks used at the universities and, in line with our results, that phases of a higher rate of homogenization correlate to a higher rate of circulation.

For the purpose of the present study, we consider the degree of homogenization of scientific knowledge as the degree of its acceptance and consensus. If, therefore, the mechanisms of circulation of knowledge determine its homogenization, and if the homogenization of knowledge represents its acceptance, and if a

consensus in turn drives the accumulation of knowledge, then we can state that the accumulation fundamentally depends on the dynamics of knowledge circulation. On the basis of this argument, in the following, we model the circulation of knowledge. Knowing that this process led to a higher consensus, we, therefore, interpret the results in terms of a knowledge accumulation process.

How information, scientific knowledge, ideas, concepts, and even rumors spread across a population has been investigated for many years (Bettencourt et al., 2006; Goffman, 1966; Li et al., 2017; Meel and Vishwakarma, 2020). One approach to this subject is based on the assumption of similarity between the nature of the spread of information and epidemic processes (SI model) (Pastor-Satorras et al., 2015), and (as an extension) the phenomenon of adoption of innovations (Bass model) (Bass, 1969, 2004). Both are used in the present work to explain the spread of scientific knowledge and, therefore, the accumulation of scientific knowledge (Kiss et al., 2010). We conclude that the Bass model, with its extra parameter representing the effect of external factors, works best for our dataset.

The *Sphaera* corpus (see Section 1 of the Supplement) is constituted of 359 different editions of textbooks on astronomy used all over the European continent between 1472 and 1650. For each edition, we analyze one copy and consider it representative of the entire print run. Considering an average print run of about 1000 copies (Gingerich, 1990), our corpus represents approximately 350,000 books circulating in the period under consideration. In total, the 359 copies examined amount to about 76,000 pages of historical material.

The textbooks have been “dissected” into text-parts: semantically completed texts taxonomically organized and mutually related by means of the relations “commentary of,” “translation of,” “adaptation of,” and “fragment of.” As these categories indicate, our focus is on the modes of production of scientific content. As text-parts recur (a highly variable range of times) or disappear, the resulting multiplex semantic network shows a series of mutations over time that have been empirically analyzed elsewhere (Valleriani et al., 2019; Zamani et al., 2020).<sup>1</sup>

In the following, we show that in Europe, the early modern period is characterized by two very distinct phases of circulation and accumulation of scientific knowledge, the first ending and the second beginning during the 1530s. We then model these processes and finally also show how the model may be applied in order to reduce the range of historical hypotheses expressed by the number of edges in the graphs, a hitherto insurmountable problem of computational history, when close readings exceed human capacities: for instance, if a series of three or more books contain the same text-part and it is not possible to establish whether the third has been influenced by the first, the second, or both, the usual work-around is to connect each following edition with all previous ones and therefore by potentially creating redundant connections. The application of the results of our model to related economic, social, and institutional data shows that content-related relationships among textbooks were more probable when these overlapped with profitable economic relationships among book producers. Following the latter, it is finally possible to reduce the density of the graphs that express semantic connections.

## Method

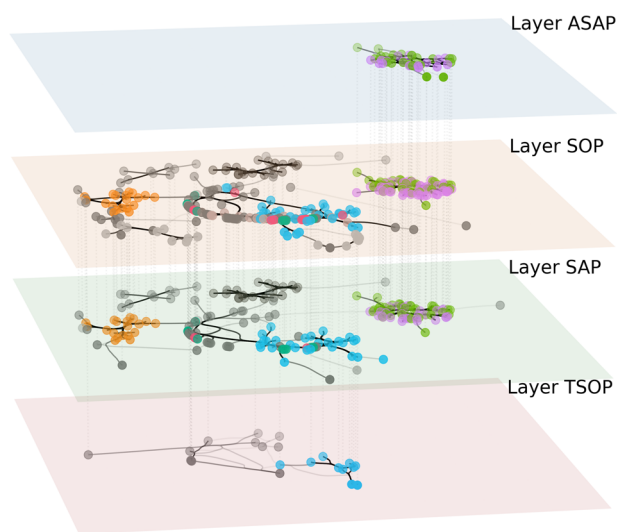
**Data and network structure.** The nodes of the network layers are the historical sources, namely the editions of textbooks connected to one another by semantic relations based on the text-parts they contain and their recurrences over time. The *Sphaera* corpus,

which is considered to be representative of all textbooks of early modern astronomy (Valleriani, 2017), is constituted of editions that contain or are directly related to one specific treatise (one text-part): Johannes de Sacrobosco’s *Tractatus de sphaera* (Thorndike, 1949), compiled during the 13th century. This might artificially enhance the relevance of this treatise in the structure of the network; therefore, after the network was created, this text-part and all connections among editions based thereon were removed (see Section 2 of the Supplement).

In this work, the multiplex network contains four layers determined by four different kinds of semantic interactions between editions. Text-parts are classified thus: original text, commentary, and translation. The connections in each layer of the resulting network express the following semantic relations and historical interpretations:

- SOP (Same Original Part) Two editions are connected if they contain the same original text-part (scientific innovations, not necessarily successful).
- SAP (Same Adaption Part) Two editions are connected if they contain exactly the same adaption part. For instance, the same commentary on the Same Original Part. An Adaption can also be a translation (commentaries and translations shape the *debate* generated by innovations).
- TSOP (Translated Same Original Part) Two editions are related to each other if they both contain a translation, but not the same translation, of the Same Original Part (reach outside the institutional barriers of the universities).
- ASAP (Annotated Same Adaption Part) Two editions have a connection if they both contain commentaries that are not the same but are on the Same Adaption (meta-reflections comparable to current large review articles).

Each kind of connection is expressed by a directed link between editions. The directionality is based on the chronological order of the editions defined by the year of publication. Figure 1 represents the multiplex network of the four layers of the *Sphaera* corpus; only nodes with connections are presented in each layer. The number of nodes (editions) and links (semantic interactions) in each layer is given in Table 1.



**Fig. 1 Visualization of the four layers of the *Sphaera* Corpus.** Components are colored referring to their sizes consistently with Fig. 4: pink (largest), green (second largest), and blue (third largest). Only connected nodes are shown in each layer.

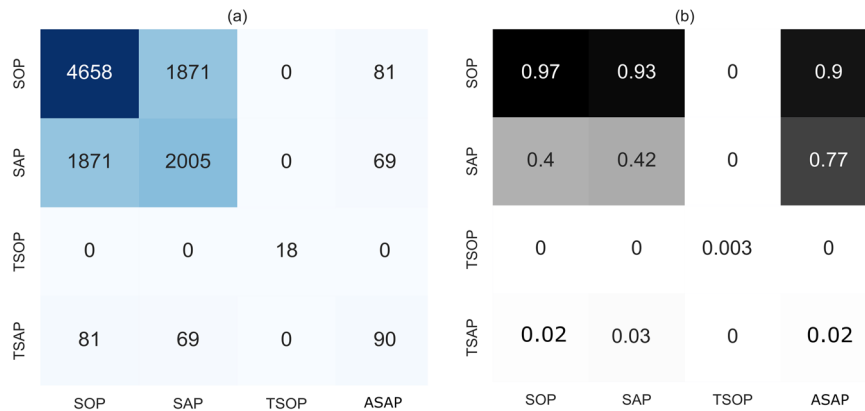
**Relations between the layers.** The multiplex network of the *Sphaera* Corpus with its four layers indicates that their corresponding entities (editions) could be relevant to each other for different aspects. To understand the relation between the layers and their level of connectivity, we investigate the “Layer Interaction Network (LIN).” Each layer is simplified into a node and a new weighted network is built where the nodes are layers and the links between them represent the overlapping edges in the layers; the weight of the links in LIN represents the number of overlapping links between the layers (Renoust et al., 2014). The purpose of the study is to better understand the coupling between the layers themselves and to represent such coupling by a well-defined quantity. This, in turn, results in a value of influence of each layer on the others. If a multiplex network is considered with  $V_M$  number of nodes and  $E_M$  number of links in each layer  $M$ , where  $M = 1, 2, \dots, L$  and  $L$  is the total number of layers, then the LIN network has  $L$  number of nodes and there is a connection between two nodes of LIN, if there is at least a common link in the corresponding layers (Škrlić et al., 2020).

Figure 2a shows the LIN matrix, where the diagonal values are the number of links in each layer and the off-diagonal values show the overlapping links between the layers. Based on the LIN matrix, we can find the “overlapping frequency matrix,” where the off-diagonal values correspond to  $\frac{n_{l,l'}}{n_{l,l}}$  and the diagonal values refer to  $\frac{n_{l,l}}{E}$ .  $n_{l,l'}$  is the number of overlapping links between layers  $l$  and  $l'$ ;  $E$  is the total number of links in the aggregated network of all layers without considering the weight of the links (repeated connections) (Škrlić et al., 2020). In this case, for four layers of the *Sphaera* Corpus, the total number of links in the aggregated graph equals  $E = 4819$ .

Figure 2b shows the overlapping frequency matrix. This matrix represents the percentage of the number of links in a layer that overlap with those of the other layers. For example, the first row of Fig. 2b shows that 93% of the links in SAP are shared with SOP, and 90% with TSAP. Layer TSOP does not share any link with any other layer and the value of 0.0037 in the diagonal of this matrix means that 0.37% of the total links of the aggregated graph refer to this layer whereas 97% of the links of the aggregated network are present in layer SOP.

The entanglement of each layer with other layers can be found from the right eigenvector corresponding with the maximum eigenvalue of the “overlapping frequency matrix.” The entanglement value ( $\gamma$ ) for each layer is presented in Table 2. Layer SOP has the maximum entanglement (coupling) with the other layers and layer TSOP has no coupling with the others. These relations of influence were easily captured by looking at the LIN matrix in Fig. 2b as it shows that, among the layers, SOP shares the greatest number of links with the other layers and TSOP does not share any links. However, having a well-defined quantity such as  $\gamma$  gives us a general quantitative understanding of the level of connectivity between the layers and shows the degree to which one layer influences the others.

Layer	Nodes	Links
SOP	300	4658
SAP	202	2005
TSOP	22	18
ASAP	40	90



**Fig. 2 Relations between the layers.** **a** Layer Interaction Network (LIN): diagonal values represent the number of links in each layer and the off-diagonal elements refer to the overlapping links between the layers. **b** Overlapping frequency matrix: It shows the percentage of links in a layer that overlap with those of other layers.

Layer	Entanglement ( $\gamma$ )
SOP	0.91
SAP	0.39
TSOP	0
ASAP	0.02

**SI and bass models.** We use existing epidemic models to model the spread of knowledge among the editions of the *Sphaera* corpus. We treat the text-parts as pathogens transferred from one edition to another. In epidemic modeling susceptible and infectious individuals are defined as follows (Barabasi, 2013; Roberts and Heesterbeek, 2003):

- Susceptible ( $S(t)$ ), number of healthy individuals at time  $t$  who are susceptible to infection.
- Infectious ( $I(t)$ ), number of individuals who have already been infected.

Considering  $N$  as the total number of individuals in the population (in a homogeneous distribution),  $\frac{S(t)}{N}$  is the probability that an infected individual comes into contact with a susceptible one.

Using such a model for the *Sphaera* corpus, we assume all editions published at time  $t$  and contain the text-parts to be individuals that are infected at time  $t$  ( $T$  being the total period of our study); all editions published after time  $t$  (during  $T - t$ ) are considered susceptible at time  $t$ . Since all editions at the end time  $T$  contain text-parts without losing them later on, we do not have any recovered editions in our model. By considering  $\beta$  as the *rate of transmission*, the change in the average number of infectious individuals during the time frame of  $dt$  will be

$$\frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} \tag{1}$$

By substituting the fraction of susceptible and infectious individuals,  $s(t) = \frac{S(t)}{N}$  and  $i(t) = \frac{I(t)}{N}$  in the equation above and using the fact that  $s(t) + i(t) = 1$ , we obtain

$$\frac{di(t)}{dt} = \beta s(t)i(t) = \beta i(t)(1 - i(t)) \tag{2}$$

The solution of the above equation is an s-curve Sigmoid function with a slow start, speeding up in the middle period, and

a slow end, in the following form:

$$i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}} \tag{3}$$

where  $i(t = 0) = i_0$  is the initial condition. In modeling the adoption of innovation, Bass introduced a model similar to SI by adding an extra term referring to the external influence or advertising effect (with rate  $\alpha$ ), whereas the coefficient  $\beta$  is the representative of internal influence (Bass, 1969, 2004):

$$\frac{di(t)}{dt} = (\alpha + \beta i(t))(1 - i(t)) \tag{4}$$

Solving the equation above leads to an s-curve function in the following form:

$$i(t) = \frac{(\beta i_0 + \alpha)e^{(\alpha+\beta)t} + \alpha(i_0 - 1)}{(\beta i_0 + \alpha)e^{(\alpha+\beta)t} - \beta(i_0 - 1)} \tag{5}$$

There are two special cases of the Bass model. The first is the case when  $\alpha = 0$ . In this case, the solution leads to a symmetric logistic equation (as discussed in the SI model). The second case is when  $\beta = 0$ , which renders the solution  $i(t)$  into an exponential saturation. A high value of  $\beta$  indicates a slow start of diffusion, which begins to accelerate after a while. A higher value of  $\alpha$  means a quick start in diffusion but a quick leveling off into saturation at the end (Van den Bulte, 2002).

The case of  $\alpha < 0$  represents a barrier to initial adoption (resistance to external influence). The case  $\beta < 0$  is mathematically well defined but does not make sense because it implies that infected change their status by contact with susceptibles (Ghanbarnejad et al., 2014; Orbach, 2016).

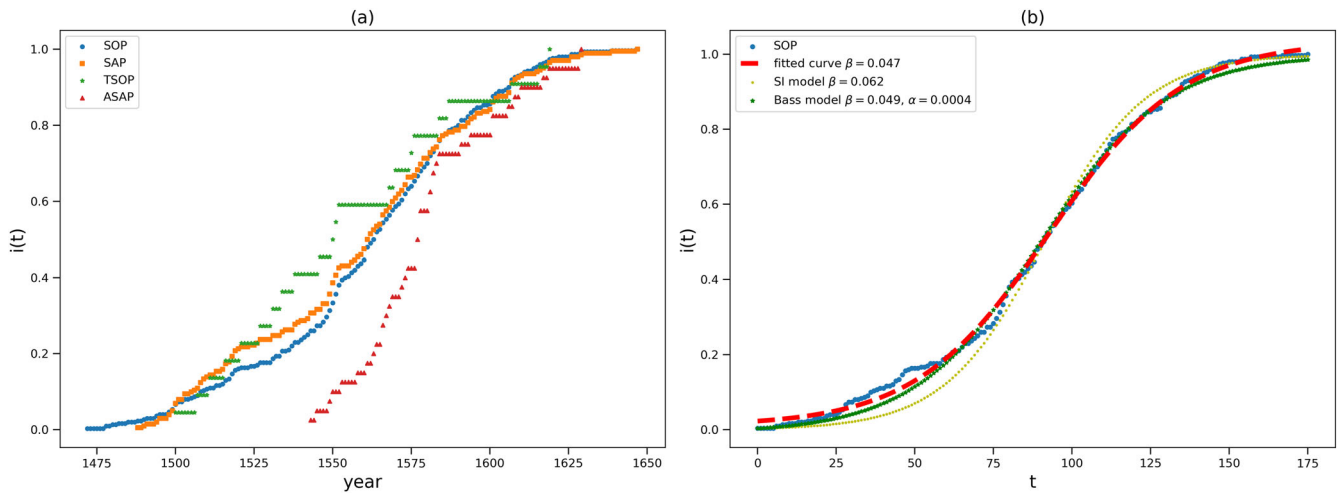
**Results**

In this section, we first present the results of the modeling for the four layers of the *Sphaera* corpus and their comparison. We then consider more closely the components of the aggregated graph of layers SOP and SAP.

**Spread of knowledge in each layer of the multiplex network.**

The spread of knowledge in the four layers (SOP, SAP, TSOP, and ASAP) of the *Sphaera* corpus are compared in Fig. 3a. In the spirit of infectious disease modeling, the number  $i(t)$  is simply the fraction of all the editions in a given layer that has been published by time  $t$ . Layers SOP and SAP show an s-curve-like behavior with a slow start, faster growth in the middle period, and a slow saturating end. However, layers TSOP and ASAP behave differently. TSOP grows slowly with zero editions for many years and a





**Fig. 3 Circulation of knowledge and models.** **a** Diffusion of information (text-parts) over time for four different layers of the multiplex network of the *Sphaera* corpus, showing an s-curve in the SOP and SAP layers, with the highest rate of transmission in the middle period of time; **b** Finding the best fit to the s-curve of the SOP layer in order to capture the parameters of the SI and Bass models. R-square value for the red fitted curve is 0.99 and for SI and Bass models equals 0.98 and 0.99, respectively.

sudden cutoff at the end. ASAP starts much later in time than the other layers. The process of knowledge circulation in ASAP has a fast rise that continues for almost 40 years then starts to slow down abruptly.

To fit a curve to each of these plots, we used two different approaches: First, by using an arbitrary Sigmoid function that fits the curve perfectly. Second, by finding the model-based approach that fits the analytical solutions of both SI and Bass models to the s-curve and calculating the corresponding existing parameters in these models. Figure 3b shows each of these fittings to the s-curve of the SOP layer. The red dashed line is the best-fitted curve in the form of a Sigmoid function with the following equation:

$$i(t) = \frac{1.02}{1.01 + 78.05e^{-0.047t}} \quad (6)$$

It follows the original curve perfectly with R-square equal to 0.99 and suggests the value of  $\beta = 0.047$  for the rate of spread. However, to achieve a better understanding of the existing parameters in the process, we need to use the analytical solutions of the SI model (Equation (3)) and the Bass model (Equation (5)).

The yellow dotted curve in Fig. 3b refers to the SI model (Equation (3)) with the set of parameters,  $i_0 = 0.0033$  and  $\beta = 0.062$ . As shown by the figure, apart from the middle and terminal saturating periods, the SI curve cannot explain the spreading process during the first 75 years, which is faster than in the SI curve and it is slower after 100 years, when it reaches the saturation point. On the other hand, the analytical solution of the Bass model (Equation (5)) shows a better coherence to the s-curve (green stars), with parameter values of  $i_0 = 0.0033$ ,  $\beta = 0.049$  and  $\alpha = 0.0004$ . The R-square values for both SI and Bass model equal 0.98 and 0.99, respectively.

Table 3 shows the parameter values of the Bass model related to each of four s-curves of SOP, SAP, TSOP, and ASAP layers.

**Spread of knowledge in the components.** As discussed in section “Relation between the layers”, according to the overlapping frequency matrix, the two layers TSOP and TSAP play a marginal role in the interaction between layers. We, therefore, first create an aggregated graph of only the two main semantic layers SOP and SAP.

Figure 4 depicts the structure of the aggregated graph. The resulting aggregated network is composed of 35 highly intraconnected components. The advantage of looking at the structure of

**Table 3  $\beta$  and  $\alpha$  measured by fitting the analytical equation of the Bass model to the curves of knowledge spread in the four different layers of the *Sphaera* corpus and the first three biggest components of Fig. 4.**

	$\beta$	$\alpha$	$i_0$
SOP	$0.049 \pm 0.0007$	$0.0004 \pm 0.00002$	0.0033
SAP	$0.037 \pm 0.0005$	$0.0023 \pm 0.0001$	0.0049
TSOP	$0.031 \pm 0.0008$	$0.0051 \pm 0.0002$	0.045
ASAP	$0.069 \pm 0.004$	$0.0042 \pm 0.0007$	0.025
Pink-comp.	$0.044 \pm 0.0019$	$0.0122 \pm 0.0005$	0.010
Green-comp.	$0.071 \pm 0.0084$	$0.0072 \pm 0.0014$	0.030
Blue-comp.	$0.194 \pm 0.014$	$-0.0091 \pm 0.0011$	0.05

The results are reported in 95% confidence level.

the aggregated graph with only layers SOP and SAP is that it facilitates a more accurate historical interpretation of the results by pointing exclusively to original texts and the works that elaborate on them (commentaries), these being the two main *modi* of scientific knowledge production during the early modern period. We study the circulation of knowledge in each of the three major components of this network by establishing the rates of diffusion (Fig. 5).

Ordered chronologically and as shown in Fig. 5a, the first component in time (green) is the second in size as its editions are published between 1478 and 1538. The second component in time, the giant component (pink), begins in 1531 and ends in 1629. The third component in time (blue) is also the third in size and temporally partially overlaps with the giant component as it begins in 1570, when the leveling off of the giant component begins, and ends in 1618 (see Section 4 of the Supplement). Note that there is no connection between the components; they, therefore, indicate distinct phases of the process of circulation and accumulation of knowledge.

Figure 5b shows the same plot as Fig. 5a, where the x-axis is the time-step instead of the year, with fitting curves and the parameter values of the Bass model displayed in Table 3 as well. The spread of knowledge in the green component (first in time and second in size), shows a more linear behavior than an s-curve. The speed of spread in the pink component (second in

time and first in size) in the first 38 years is the same as in the green component, but the latter continues to expand with the same speed until its end while the pink component gradually starts to slow down. The rate of transmission  $\beta$  of the green component is higher with the value of 0.071.

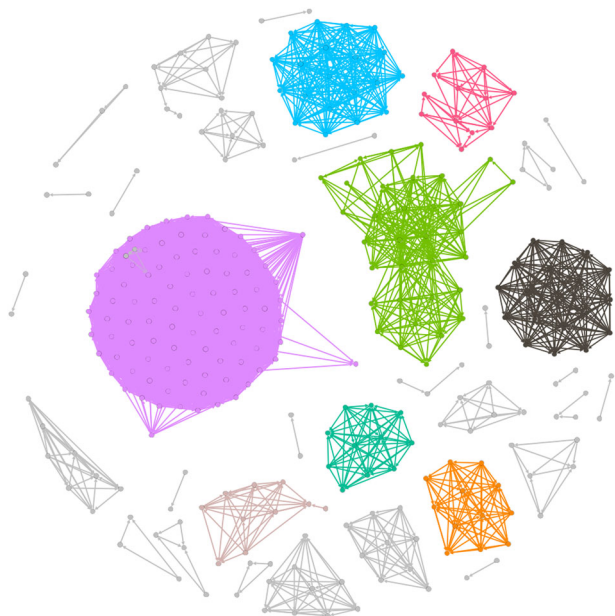
The third component in time and size (blue) behaves differently: a very slow start, rising up fast in the middle period, and slowing down again at the end. The rate of transmission in the third component is the highest among the components with the value of  $\beta = 0.194$ . The negative value for the rate of external influence  $\alpha = -0.009$  suggests the existence of a resistance to external influence in this component. The high value of  $\beta = 0.194$  and the negative value of  $\alpha$  indicate that these editions strongly

learn from each other and even resist adopting knowledge from other external communities.

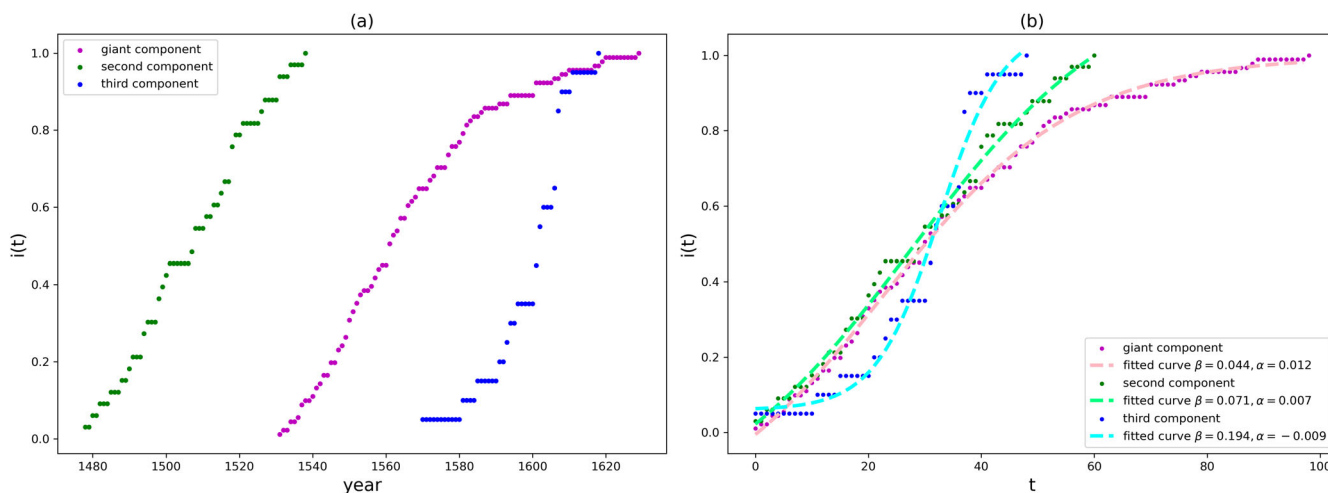
**Historical interpretation and validation: layers and components.** The values of  $\beta$  and  $\alpha$ , measured by fitting the analytical equation of the Bass model to the s-curves in the four different layers, clearly signal a higher external influence in regards to the TSOP layer. This result was expected due to the layer's focus on vernacular translations, as these were by their nature means to spread scientific knowledge outside the institutional barriers of the universities and thus more keenly affected by exogenous factors.  $\alpha$ , representative of the rate of influence or exogenous factors, has the smallest value (0.0004) in reference to the SOP layer. A relatively high value of  $\beta$  and low value of  $\alpha$  in this layer show that the process of knowledge circulation in the SOP layer is mostly determined by  $\beta$  and therefore happens as a result of endogenous factors and is less influenced by outside factors than the other three layers. The highest speed of spread for the ASAP layer, even higher than the SOP layer, represents an interesting historical finding. Editions of this layer are large anthological and integrative works comparable to current large review papers, and the values show that these were mostly contributing to the process of knowledge circulation (see Table 3, as well as Sections 3.2 and 3.3 of the Supplement). This result is moreover confirmed by analyzing the structural parameters of the network (see Section 6 of the Supplement).

Considering now the three major components of the network shown in Fig. 4 and their most frequent edge parameters (text-parts), the chronologically ordered historical analysis (green: first in time and second in size; pink: second in time and first in size; blue: third in time and size) shows that they represent different phases and aspects of one and the same process, namely the one that transformed astronomy, as studied at the universities in Europe, from a natural-philosophical discipline into a proper mathematical one. As the pink (giant) and the blue (third in size) components chronologically overlap, the three components taken together represent two historical and scientific phases of a single historical process (see Section 4.2 of the Supplement). The first phase starts and the second begins during the 1530s.

A closer look at the text-parts reveals the historical meaning of the components (see Section 4.3 of the Supplement). The most



**Fig. 4 Network of the aggregated graph of two layers, SOP and SAP.** It includes 35 separated components with a giant component (pink) of 91 nodes (editions) and 3831 links. The second and third largest components are colored green and blue, respectively.



**Fig. 5 Knowledge circulation process and models for the three major components.** **a** Diffusion process in the three biggest components of the graph shown in Fig. 4. Each component starts at a different time. The beginning of the pink (giant) component roughly corresponds to the end of the green (second in size) component and the blue (third in size) component begins when the giant component starts slowing down. **b** Fitting to each curve of knowledge spread and corresponding parameters of the Bass model for each curve. R-square for both pink and green component equals to 0.99 and for the blue component is 0.98. The high values of R-square indicate that the models explain perfectly the original data sets.

relevant text-parts, around which the pink (giant) component is constituted, show moreover that the mathematization of astronomy was going hand in hand with its integration with sub-disciplines of practical utility (such as nautical astronomy, booming at the time because of the journeys of exploration). This explains the relatively high value of  $\alpha$  for this component. The differentiation between the pink (giant) component and the blue (third in size) component might be due to the specific institutional embedding in which the text-parts of the blue component were produced and used, as they were produced by scholars who belonged to the confessional order of the Jesuits. The Jesuit order was rising in influence along the routes of exploration. Their scientific works were used primarily in their own educational institutions, which were spreading all over the planet in connection with their missionary activity. The highest speed of spread of the third component testifies to the close link between the knowledge production of the Jesuit order and its institutional expansion. The negative value of  $\alpha$  instead indicates the high competition the Jesuits experienced with other educational institutions, including with those of other Catholic orders (Grendler, 2022).

### Finding historically plausible interactions among the editions of the *Sphaera* corpus

As discussed in section “Data and Network Structure”, the connections (links) between the editions (nodes) are based on their shared knowledge (text-parts). Figure 4 illustrates that each component is extremely dense and that almost every edition inside each of them is connected to every other. Such high density is the consequence of the assumption used to build the network, namely an assumption of maximal connection: on the basis of the semantic relationships defined by the layers, every edition is linked to all previously published editions. Because of this assumption, we are certain to include all historically confirmable connections, but might also introduce connections that are artificial. It is indeed very unlikely that each edition acquired its knowledge from all previous editions that contain the same text-part, or that the authors, printers, or publishers of each edition were really informed of all extant editions. Until now, the only way to avoid such assumptions was the close reading of all historical sources and the application of philological analysis (see Section 5 of the Supplement). For the first time, however, we have a method to indicate which connections are likely to be the real historical ones as a sub-group of the connections displayed in Fig. 4 and Table 1. The method is an application of the model described above.

We know how the diffusion of knowledge among the editions of the *Sphaera* corpus proceeds, as we counted the accumulation of “infected” editions (infected through the acquisition of text-parts) over time. We now begin by applying the same method but considering infected editions in the whole aggregated network of four layers. We find infected editions in the structure of the network instead of adding them up by year. For example, imagine an edition “A” as a node in the aggregated network, that it has been published at time  $t$ , and that it has connections to a number of other editions (nodes) in the network. We consider edition “A” and all its target editions (nodes) in the network as the infected editions at time  $t$ . Now imagine edition “B” published at time  $t' > t$ . We check if “B” has been the target of “A” and has therefore been counted among infected editions or not. If not, we add “B” and its targets (if they were not the targets of “A”) as infected editions at time  $t'$ . If “B” was a target of “A” then we just look for the targets of “B”; if they were not already the targets of “A” we add them to the infected editions in time  $t'$ .

In other words, when the new text-part (disease) appears at time  $t$ , all the editions that contain that text-part are counted as infected editions at that time. So, instead of counting the yearly published editions added to the *Sphaera* corpus, we count how many (not yet counted) later editions are linked to the published edition at the corresponding year.

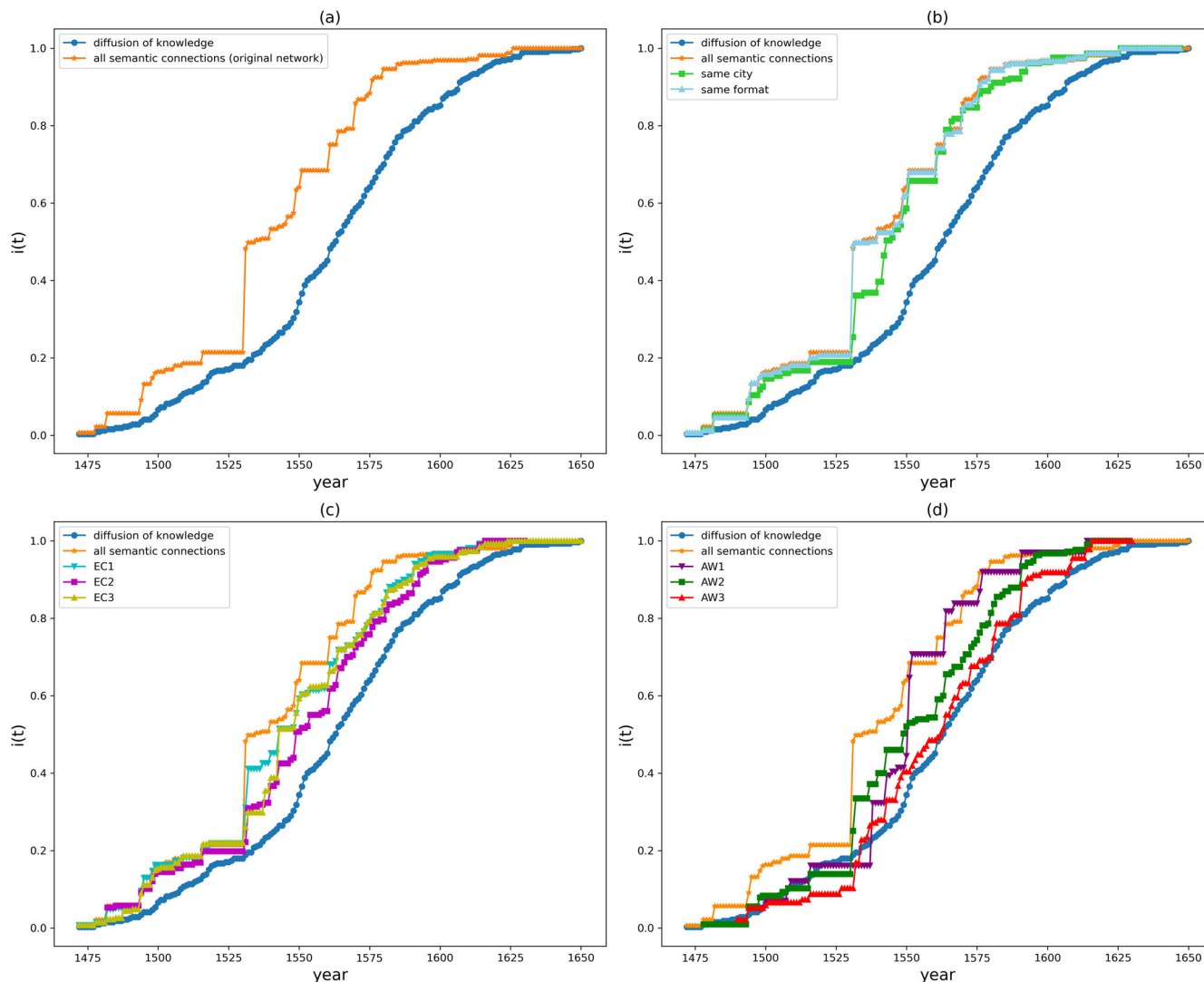
Figure 6a shows the diffusion of knowledge in the network of the aggregated graph (orange), measured by the method explained above, and compares it with the s-curve, which is the result of counting and adding up the editions by year. The infection curve of the aggregated graph does not fit to the s-curve (blue). As noted, we assume the reason to be the fact that the original aggregated graph contains connections that are not real due to the assumption of maximal connection. In order to indicate which of the connections are likely to be the real ones, we would like to search for a network of connections between editions that converges to the blue s-curve. For this purpose, we need to remove some links from the aggregated graph and we let us guide by the domain knowledge, namely by specific historical meaningful hypotheses. We control whether:

- geographical aspects affect the transmission of knowledge, for example, whether authors, printers, or publishers tended to refer to editions that had been published in the same city, region, or country (local market);
- connections between editions are correlated to the book formats sexto-decimo, octavo, quarto, and folio (price and marketing position of the book on the market);
- the fact that the producers of older editions were alive at the time of a new publication could bring the new producer to primarily refer to those older editions (social awareness).

Figure 6b compares the transmission of knowledge in two newly generated networks with the original network and the s-curve. The first network considers only connections between editions published in the same location and the second one includes only the connections between editions with the same book format (Valleriani et al., 2019). These two graphs do not relevantly differ from the behavior of the original aggregated network. The curve for the spread of knowledge in the network of the same city shifts a little toward the blue s-curve but does not converge to it. We decided, therefore, to make use of these and other pieces of information to create new layers, which are, however, less generic. In particular, all the following layers, which will be compared to the s-curve, are created under the additional condition that the editions connected to each other have also to share at least one semantic relation.

In Fig. 6c, the comparison has been executed with the transmission inside three other new networks that express economic aspects related to the production of the editions. These layers are named EC1, EC2, and EC3.

- In EC1, two editions with shared semantic knowledge and the same format are connected if the printers or publishers of these two editions were alive at the time of the publication of both editions. The links express the fact that printers and publishers were producing the same content during more or less the same range of time for the same price segment of the market.
- In EC2, two editions are connected if they share the same semantic knowledge and were published in the same place by different printers and/or publishers who were alive at the times of publication of both editions. The links in EC2 refer to the fact that two printers and/or publishers were producing the same content in almost the same range of time for the same local market.
- In EC3, we first distinguish between five types of editions:



**Fig. 6 Finding true connections between editions of the *Sphaera* corpus by searching for a network that converges to the s-curve.** **a** Comparison of the s-curve of transmission of knowledge among all editions with the transmission inside the original network of the aggregated graph that contains all the semantic connections. R-square of comparison of these two curves shows the value of 0.82. **b**) Knowledge transmission in the networks built on the connections among editions published in the same city and those editions with the same book format and their comparison with the s-curve does not show a good match and not much improvement with respect to figure (a). **c** Diffusion of knowledge in three networks built on the basis of data with economic relevance; it shows a small shift toward the s-curve. R-square values for EC1 curve is 0.87, for EC2 is 0.93, and for EC3 is 0.91, showing that EC2 better explains the variability of the original blue s-curve than EC1 and EC3. **d** Transmission of knowledge in graphs showing mutual social and economic awareness among authors, printers, and publishers, as well as the graphs' convergence to the s-curve for AW1 and AW2 before 1530, and AW3 after 1530. The convergence of the awareness graphs shows the relevance of the meaning of these networks for the mechanisms of knowledge circulation in the early modern period. In particular, the high value of R-square for AW3, 0.98, among the other tested layers, 0.77 and 0.93 for AW1 and AW2, respectively, indicates the high relevance of AW3 for knowledge circulation and accumulation during that time period.

- OT: editions that contain only one original text (17 editions);
- COT: editions that contain one original text and a commentary on it, usually printed by fragmenting the original text and commenting on each fragment on the same page (48 editions);
- COMP: editions that contain several original texts that do not directly refer to but rather enrich each other (43 editions);
- COT + COMP: editions that combine the characteristics of COT and COMP (124 editions);
- ADAPT: editions that do not contain the reference text around which the corpus was built, but contain at least one text that has been designed by following it as a model

(same subjects and order with at least partially overlapping visual apparatus) (127 editions).

The taxonomy of the five types of editions directly expresses five different modes of knowledge production and, therefore, a semantic value. EC3 is then built by connections realized if two editions have the same semantic knowledge and are of the same type and were published by different printers or publishers who were alive at the times of publication of both editions. The links express that both printers and publishers were offering the same kind of intellectual product to the market, be this local or trans-regional.

Figure 6c shows that the three EC networks are close to the s-curve but do not converge to it. Among these, EC2 is closer to



the s-curve, showing the possible relevance of the local market in the stimulation of academic book production in the early modern period, but with the condition that printers and/or publishers were alive at the times of publication of both editions (Gehl, 2013; Valleriani and Ottone, 2022).

In Fig. 6d, we compare the diffusion process with another set of new graphs, which identify networks of social awareness among book producers and that lead us to more certain conclusions. First, we will introduce each of the “awareness graphs” and their meaning.

- AW1, two editions are connected if they share the same semantic knowledge (at least two text-parts) authored by different scholars who were alive at the times of publication of both editions. The links in AW1 indicate the awareness of the two authors of each other’s works (Valleriani, 2020);
- AW2, two editions are connected if they share at least one paratext and are published by two different printers or publishers who were alive at the times of publication of both editions. Paratexts are texts such as literary compositions or dedication letters to patrons, and which therefore indicate whether the two book producers belonged to the same institutional or political context, such as an early modern court (Valleriani and Sander, 2022);
- AW3, two editions are linked if they were produced by different printers or publishers who were alive at the times of publication of both editions and are identified by a similar fingerprint, the latter being a characteristic that denotes similarity in content and layout between both editions (Valleriani et al., 2022).

Figure 6d shows how AW1 and AW2 converge to the s-curve from 1473, the very beginning of the period of study, until 1530. After 1530, it is AW3 that mostly converges to the s-curve and continues until the end of the period of study. This reveals that, while the role of the network among scholars and the institutional embedding of the book producers remained more or less equally relevant over time, after 1530 it became fundamental to produce editions that were similar to previous ones *both* in content and layout. This peculiar form of imitation was the consequence of very precise economic and material constraints that early modern printers and publishers had to cope with to maintain their business running (Valleriani et al., 2022). This finally implies that in order to shape networks of diffusion of scientific knowledge in the early modern period, historians need to reduce their semantic networks by considering the economic relationships among their producers rather than their content. This research, therefore, shows that the content of the textbooks became increasingly homogeneous on the continent because of such imitation process. Thus, a step was realized toward the formation of a modern scientific identity in Western culture (see Section 5.1 of the Supplement).

## Discussion

We have shown how the application of epidemiological models to a historical dataset could help in identifying different phases and aspects of circulation of knowledge, including the discovery that this process was highly dependent on the institutional embedding of scientific production. By integrating these results with those obtained in (Valleriani et al., 2019, 2022; Zamani et al., 2020), we can now conclude that the higher rates of circulation coincide with phases of more rapid homogenization. This, in turn, means that a higher consensus was built around the fundamental astronomic knowledge (core knowledge) taught at European universities, though this changed along a temporal interval of about 180 years, and that the accumulation of knowledge

accelerated and became fundamental in the temporal frame during which the major component (pink) developed.

Applying the model to the reduced network and pairwise comparison between Fig. 6c and 6d revealed that neither the relations among scholars and their scientific ideas, nor their geographic locations *as a whole* played a significant role in the way scientific knowledge was transformed during the early modern period. The diffusion of knowledge was mostly determined by the economic constraints of early modern printers and publishers that caused them to imitate each other in both content and layout (AW3). It could be added, however, that geographic proximity might have facilitated this as well, but only on condition that both printers and publishers were alive at the time of their respective publications (EC2).

By correlating these findings, it is possible to reach a conclusion concerning the relation between global and local knowledge circulation. First, the circulation of books (a condition required to imitate their content and layout) was rapid and unhindered by geographic obstacles. This is in line with results achieved by book historians who were able to demonstrate how printed books immediately entered the large trans-regional markets of Europe and were transported along with other goods traded in that market. Second, the early establishment of international book fairs, such as those in Frankfurt and Leipzig during the 16th century, certainly contributed to this process, too (Maclean, 2022). Third, the circulation of knowledge was enhanced, though with less impact, by personal contacts - not among scholars but among printers and publishers. These were obviously more frequent on a local basis.

In our approach, we retained the differentiation between the semantic and other layers. An alternative approach is possible and will be executed in the future in order to compare the results with those obtained by (Granell et al., 2013) and (Yang et al., 2022). These works study the interrelation between the spread of an epidemic and the spread of awareness about the same epidemic as a multiplex network of two layers. One layer represents virtual contacts among individuals who mutually inform each other about the epidemic while the other layer represents real contacts that are responsible for the actual spread of the epidemic. In our future work, we will consider all semantic interactions as one layer and the awareness connections between the editions as the second layer.

The ability to model the spread and circulation of scientific knowledge over long periods of history opens the door to studies whose aim is to predict the relation between communication channels and scientific developments. From the perspective of the historical disciplines, moreover, modeling means finding abstract, mathematically formulated relationships that are characteristic of epochs and cultures, potentially redrawing commonly accepted divisions in human epochs. As shown, if further data concerned with the social, institutional, and economic contexts are available, the modeling results can be further applied to better investigate the past and discover relevant causes for the developments under consideration. On the basis of this study, modeling itself can become a new working tool for historians.

## Data availability

The necessary data and code to replicate the results presented in this paper can be found here: <https://gitlab.gwdg.de/MPIWG/Department-I/sphaera/sphaera-SI-model>.

Received: 12 October 2022; Accepted: 20 July 2023;

Published online: 25 August 2023

**Note**

- 1 The database can be reached through the project website *The Sphere. Knowledge System Evolution and the Shared Scientific Identity of Europe* (<https://sphaera.mpiwg-berlin.mpg.de>).

**References**

- Barabási A-L (2013) Network science. *Philosophical Trans R Soc A Math Phys Eng Sci* 371(1987):20120375
- Bass FM (1969) A new product growth for model consumer durables. *Manag Sci* 15(5):215–227
- Bass FM (2004) Comments on “a new product growth for model consumer durables the Bass model”. *Manag Science* 50(12):1833–1840
- Bettencourt LM, Cintrón-Arias A, Kaiser DI, Castillo-Chávez C (2006) The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Phys A Stat Mech Appl* 364:513–536
- Castells M (2010) *The rise of the network society*. Wiley-Blackwell, Hoboken, NJ
- Clark BR (1998) *Creating entrepreneurial universities: organizational pathways of transformation*. Emerald, Bingley
- Consoli D, Patrucco PP (2003) *The knowledge trade-off: circulation, growth and role of knowledge-intensive business services in urban innovation systems*. Torino: Dipartimento di Economia, Università di Torino
- Gehl PF (2013) Advertising or fama? Local markets for schoolbooks in sixteenth-century Italy. In: Costas BR (ed) *Print culture and peripheries in early modern Europe. A contribution to the history of printing and the book trade in small European and Spanish cities*, Brill, Leiden, pp. 69–100
- Ghanbarnejad F, Gerlach M, Miotto JM, Altmann EG (2014) Extracting information from s-curves of language change. *J R. Soc. Interface* 11(101):20141044
- Gingerich O (1990) Five centuries of astronomical textbooks and their role in teaching. In: Pasachoff JM, Percy JR (eds) *The teaching of astronomy*, Cambridge University Press, Cambridge, pp. 189–211
- Goffman W (1966) Mathematical approach to the spread of scientific ideas—the history of mast cell research. *Nature* 212(5061):449–452
- Granell C, Gómez S, Arenas A (2013) Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Phys Rev Lett* 111(12):128701
- Grendler PF (2022) *The Sphaera in the Jesuit education*. In: Valleriani M, Ottone A (eds) *Publishing Sacrobosco's De Sphaera in early modern Europe. Modes of material and scientific exchange*, Springer, Cham, pp 369–407. [https://doi.org/10.1007/978-3-030-86600-6\\_11](https://doi.org/10.1007/978-3-030-86600-6_11)
- Kiss IZ, Broom M, Craze PG, Rafols I (2010) Can epidemic models describe the diffusion of topics across disciplines? *J Informetr* 4(1):74–82
- Kuhn T (1970) *The structure of scientific revolutions*. University of Chicago Press, Chicago, IL
- Li M, Wang X, Gao K, Zhang S (2017) A survey on information diffusion in online social networks: models and methods. *Information* 8(4):118
- Yang B, Shang K, Small M, Chao N (2022) Information overload: How hot topics distract from news—COVID-19 spread in the USA. *Natl Sci Open*. <https://doi.org/10.1360/nso/20220051>
- Macleán I (2022) Sacrobosco at the book fairs, 1576–1624: the pedagogical marketplace. In: Valleriani M, Ottone A (eds) *Publishing Sacrobosco's De Sphaera in early modern Europe. Modes of material and scientific exchange*, Springer, Cham, pp. 187–224. [https://doi.org/10.1007/978-3-030-86600-6\\_6](https://doi.org/10.1007/978-3-030-86600-6_6)
- Marginson S (2016) *The dream is over*. University of California Press, Oakland, CA
- Meel P, Vishwakarma DK (2020) Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-art, challenges and opportunities. *Expert Syst Appl* 153:112986
- Mokyr J (2005) *The gifts of Athena. Historical origins of the knowledge economy*. Princeton University Press, Princeton, USA
- Orbach Y (2016) Parametric analysis of the Bass model. *Innov Marketing* 12(1):29–40
- Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. *Rev Mod Phys* 87(3):925
- Popper KR (1972) *Objective knowledge: an evolutionary approach*. Oxford University Press, Oxford
- Renn J (2020) *The evolution of knowledge. Rethinking science for the Anthropocene*. Princeton University Press, Princeton, NJ
- Renoust B, Melançon G, Viaud M (2014) Entanglement in multiplex networks: understanding group cohesion in homophily networks. *Social Network Analysis-Community Detection and Evolution*, Springer, pp. 89–117
- Roberts M, Heesterbeek J (2003) *Mathematical models in epidemiology* vol. 215. Paris, EOLSS
- Škrlić B, Renoust B (2020) Layer entanglement in multiplex, temporal multiplex, and coupled multilayer networks. *Appl Netw Sci* 5:1–34
- Cole S (1983) The hierarchy of the sciences? *Am J Sociol* 89(1):111–139
- Thorndike L (1949) *The Sphere of Sacrobosco and its commentators*. University of Chicago Press, Chicago
- Valleriani M (2017) The tracts on *The Sphere*. Knowledge restructured over a network. In: Valleriani M (ed) *The structures of practical knowledge*. Springer, Cham, pp. 421–473
- Valleriani M (2020) Prolegomena to the study of early modern commentators on Johannes de Sacrobosco's *Tractatus de sphaera*. In: Valleriani M (ed) *De Sphaera of Johannes de Sacrobosco in the early modern period: the authors of the commentaries*. Springer, Cham, pp. 1–23. [https://doi.org/10.1007/978-3-030-30833-9\\_1](https://doi.org/10.1007/978-3-030-30833-9_1)
- Valleriani M, Kräutli F, Zamani M, Tejedor A, Sander C, Vogl M, Bertram S, Funke G, Kantz H (2019) The emergence of epistemic communities in the *Sphaera* corpus: mechanisms of knowledge evolution. *J Hist Netw Res* 3:50–91. <https://doi.org/10.25517/jhnr.v3i1.63>
- Valleriani M, Ottone A (2022) Printers, publishers, and sellers: Actors in the process of consolidation of epistemic communities in the early modern academic world. In: Valleriani M, Ottone A (eds) *Publishing Sacrobosco's De Sphaera in early modern Europe. Modes of material and scientific exchange*. Springer, Cham, pp. 1–24. [https://doi.org/10.1007/978-3-030-86600-6\\_1](https://doi.org/10.1007/978-3-030-86600-6_1)
- Valleriani M, Sander C (2022) Paratexts, printers, and publishers: Book production in social context. In: Valleriani M, Ottone A (eds) *Publishing Sacrobosco's De Sphaera in early modern Europe. Modes of material and scientific exchange*. Springer, Cham, pp. 337–367. [https://doi.org/10.1007/978-3-030-86600-6\\_10](https://doi.org/10.1007/978-3-030-86600-6_10)
- Valleriani M, Vogl M, El-Hajj H, Pham K (2022) The network of early modern printers and its impact on the evolution of scientific knowledge: automatic detection of awareness relations. *Histories* 2(4):466–503. <https://doi.org/10.3390/histories2040033>
- Van den Bulte C (2002) Want to know how diffusion speed varies across countries and products? Try using a Bass model. *PDMA Vis* 26(4):12–15
- Zamani M, Tejedor A, Vogl M, Kräutli F, Valleriani M, Kantz H (2020) Evolution and transformation of early modern cosmological knowledge: a network study. *Sci Rep Nat* 10:19822. <https://doi.org/10.1038/s41598-020-76916-3>

**Acknowledgements**

This work was mostly supported by the German Ministry for Education and Research as BIFOLD—Berlin Institute for the Foundations of Learning and Data (ref. 01IS18037A), by the BMBF Grant No. 01 UG2131, Project ModelSEN, by the Max Planck Institute for the History of Science, and by the Max Planck Institute for the Physics of Complex Systems.

**Author contributions**

Conceptualization: MZ and MVa; methodology: MZ, HK, and MVo; validation: MVa and MZ; formal analysis: MZ and HK; data curation: HH; writing—original draft preparation: MZ and MVa; writing—review and editing: MZ, MVa, HK, MVo, and HH. All authors have read and agreed to the published version of the manuscript.

**Funding**

Open Access funding enabled and organized by Projekt DEAL.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-023-01947-w>.

**Correspondence** and requests for materials should be addressed to Maryam Zamani or Matteo Valleriani.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023