# A Simple Model of Knowledge Scaffolding

Franco Bagnoli[1,2(✉)] and Guido de Bonfioli Cavalcabo[1]

[1] Department of Physics and Astronomy and CSDC, University of Florence, via G. Sansone 1, 50019 Sesto Fiorentino, Italy
franco.bagnoli@unifi.it, guido.debonfiolicavalcabo@stud.unifi.it
[2] INFN, Sezione di Firenze, Italy

**Abstract.** We introduce a simple model of knowledge scaffolding, simulating the process of building a corpus of knowledge based on logic derivations starting from a set of "axioms". The starting idea around which we developed the model is that each new contribution, still not present in the corpus of knowledge, can be accepted only if it is based on a given number of items already belonging to the corpus. When a new item is acquired by the corpus we impose a limit to the maximum growth of knowledge for every step that we call the "jump" in knowledge. We analyze the growth with time of the corpus and the maximum knowledge and analyzing the results of our simulations we managed to show that they both follow a power law. Another result is that the number of "holes" in the knowledge corpus always remains limited. Using an approach based on a death-birth Markov process we were able to derive some analytical approximation of it.

**Keywords:** Learner model · Knowledge organization · Scaffolding model · Knowledge modelling · Knowledge visualization

## 1 Introduction

Whether knowledge accumulation is equivalent to scientific progress [3,4] or not [9] or whether this latter concept is broader than either theories suggest [8], we can assume that an increasing amount of knowledge is fundamental for the advancement of science, as actually reported in many fields [5,7,11]. Moreover, accumulation of information and knowledge is as much important to groups as it is for individuals, and, In general, knowledge accumulates over separate yet related learning episodes [2].

In all aspects of scientific disciplines based on deduction, such as most of mathematics and many parts of physics, among others, the derivation of higher-level achievements depends on previous knowledge. In these cases, new pieces of knowledge that are accepted and inserted into the existing corpus are based on previous results. Therefore it is not correct to define this process by the word "accumulation", while "scaffolding" is best suited, in our opinion.


H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 41–49, 2023.
https://doi.org/10.1007/978-3-031-21127-0_4

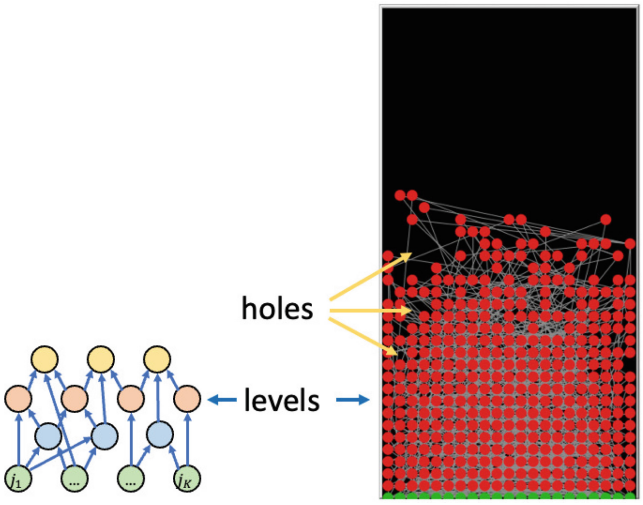



**Fig. 1.** The fundamental knowledge scaffolding model. (left) Knowledge bits are represented as nodes of a network, where different colors represent different levels, and nodes at a certain level only depend on a certain number of nodes at lower levels. Green (basic) nodes represent axioms. (right) Observing the filling of the network (here with fixed width and with fixed number of dependencies), one can detect holes that are filled after the appearance of nodes at higher levels.

A graphical representation of this process is given by an oriented or growing network [1] in which the single bits of knowledge are the nodes, and the links represent connection between the new items and its prerequisites, i.e., the elements of the existing corpus needed to prove the new result.

As shown in Fig. 1, this feed-forward structure can be seen as a layered network in which an item at any level depends on other items at lower levels. In this way one can put into evidence the structures that group together all simultaneous and independent derivations.

In this model it is easy to identify the "knowledge holes", which are missing items at levels which are lower than the highest one. Notice that holes are evident in a static snapshot of the fixed-width version (Fig. 1-right), but can be identified also in the variable-width model by observing the temporal filling of the structure.

The "filling" of knowledge holes have been studied for instance in world learning [10]. In this case, however, authors only had at their dispositions a static snapshot (of an arbitrary semantic network) so they had to resort to the analysis of the connectivity (robustness) of the network to identify the location of possible missing items. However, it is not granted that the possible missing items will appear, for instance, in the example provided in the cited article [10], the missing item, identified by the words "large" and "yellow", was a school-bus, but school-buses are not always yellow in every country.

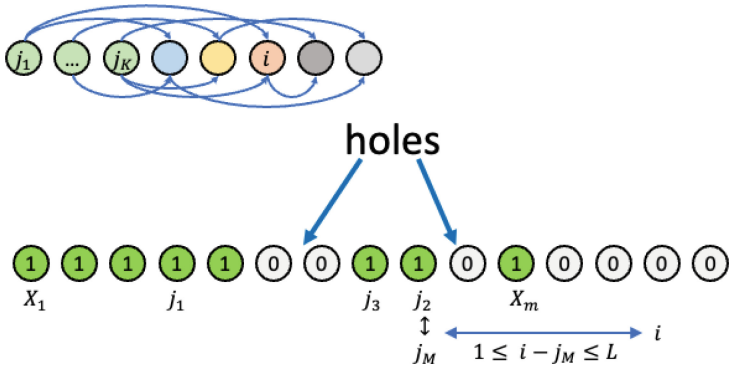



**Fig. 2.** The linear scaffolding model. In this version there are no well-defined levels. Known items are denoted by ones in a linear array of elements ($X$), and unknown ones by zeros. The "highest" known item is at position $m$. New items ($i$) become known ($X_i = 1$) if all random prerequisites $j_k$ are known. The maximum knowledge jump $L$ limits the choice of $i$ so that $1 \leq i - j_M \leq L$, where $j_M$ is the highest of prerequisites $j_k$.

In the following we shall use a simpler, unstructured layout representing a linear progress, illustrated in Fig. 2. This is of course a simplification of the previous process, since even the existence of yet unknown pieces of knowledge is in many cases not knowable: while in the layered structure one can add a variable number of intermediate nodes, in the linear version this is not possible. However, the analysis in this case is easier, and in a practical case it may correspond to an "ex-post" analysis of the temporal development of a knowledge space, for which the final number of items, their dependencies and the temporal sequence of the filling are known.

In summary, in this paper we want to investigate such a model of *linear* knowledge scaffolding, that despite its simplicity shows interesting behavior and can be approximated in an analytic way.

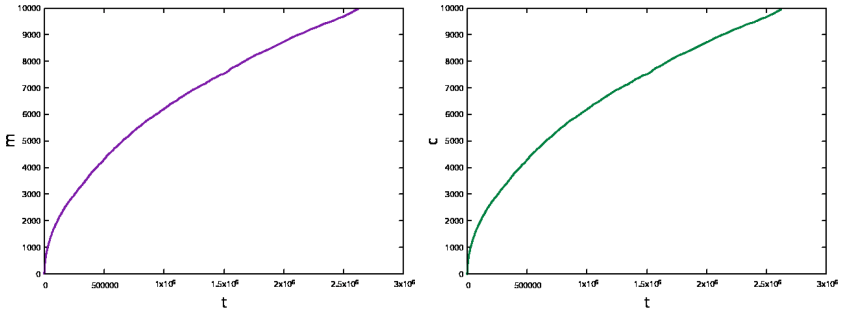
The model is presented in a detailed way in the following section, numerical and analytical results are reported in Sects. 2 and 3, respectively. Conclusions are drawn in the last section.

## 2 The Model

The knowledge space is represented by an array $X_i$ of $N$ items, which can be known ($X_i = 1$) or unknown ($x_i = 0$). We denote by

$$c(t) = \sum_i X_i$$

the size of the knowledge at a certain time.




(a) (b)

**Fig. 3.** Evolution of knowledge corpus and maximum knowledge with $N = 10000$, $K = 4$ and $L = 5$. **a** Maximum knowledge ($m$) versus time. **b** Knowledge corpus ($c$) versus time.

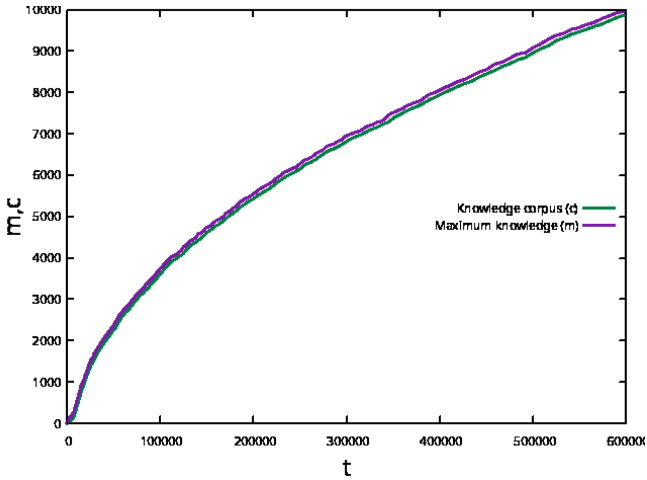



**Fig. 4.** Comparison between the evolution of knowledge corpus ($c$) and maximum knowledge ($m$) versus time for $N = 10000$, $K = 4$ and $L = 20$.

The linear scaffolding process can be modeled by a random proposal of *theorems*, each of which depends on a certain number, say $K$, of prerequisites, and furnishes a *higher* contribution. We assume that there is a limit to *jumps* in knowledge, and therefore the newly proposed item cannot be at a distance greater than $L$ from the highest prerequisite (Fig. 2).

Let us denote by $m$ the index of the highest known item in the corpus, i.e., $X_m = 1$ and $X_j = 0 \forall j > m$. The corpus can contain holes, i.e., $X_j = 0$ with $j < m$.

A theorem cannot be based on prerequisites higher than $m$, but could be rejected because it is based on holes in the present corpus. Redundant theorems (i.e., theorems providing known items) are not considered even if they are valid.

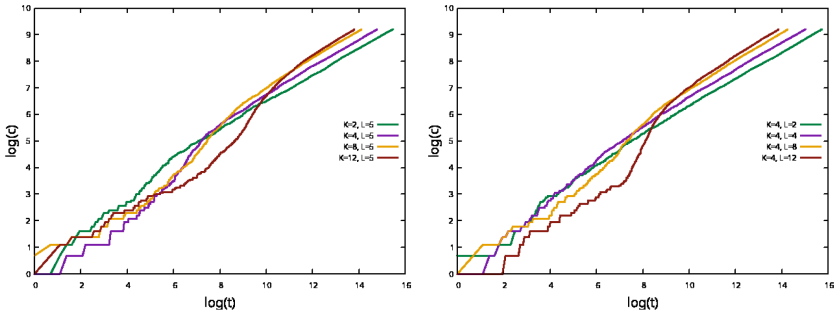



**Fig. 5.** Logarithmic plot of the knowledge corpus $c$ for $N = 10{,}000$. **a** Fixed $L = 5$ and five values of $K$. **b** Fixed $K = 4$ and five values of $L$.

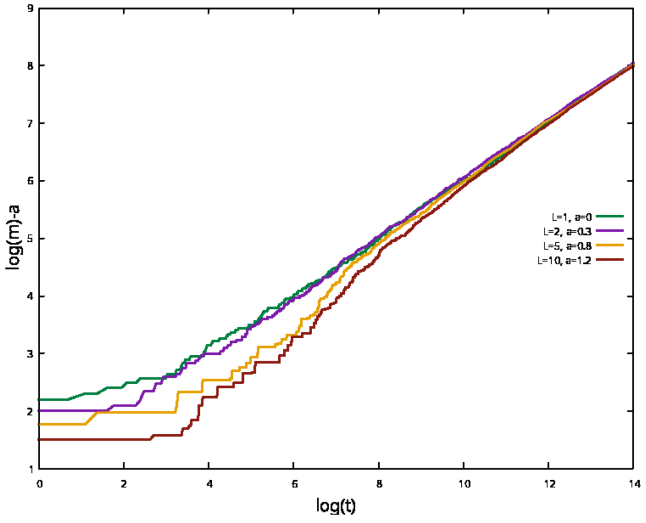



**Fig. 6.** Logarithmic plot of knowledge corpus $c$ for $N = 10000$ and $K = 4$. It is possible to numerically rescale the function with different $L$ to obtain an asymptotic convergence. The same results could be obtained by varying $K$.

The corpus filling proceeds by randomly choosing $K$ items, $j_1, \ldots, j_k, \ldots, j_K$, with $j_k \leq m$, as the prerequisite for the new theorem. It may happen that two or more of the $j_k$ correspond to the same item, since they are chosen at random in the interval $0, \ldots, m$.

We denote by $j_M$ the largest values of the selected $j_k$, and we extract a random integer $i$ such that

$$i \in (j_M + 1, \ldots, j_M + 1 + L)$$

as a candidate for the new piece of knowledge.

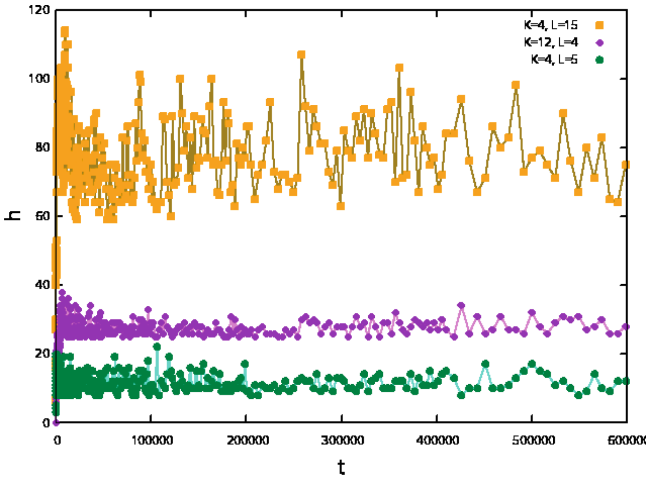



**Fig. 7.** Time evolution of the number of holes ($h$) for $N = 10000$ and different values of $K$ and $L$.

If all the prerequisites are known (i.e., $X_{j_k} = 1 \quad \forall k$) and the piece of knowledge $X_i$ is not already known (i.e., $X_i = 0$), then this derivation is added to the corpus ($X_i = 1$ and $c$ is incremented by one), and if $i > m$ then $m = i$. In any case, the time $t$ is incremented by one.

At the beginning we start with a knowledge vector of zeros, except the $2K$ smallest locations, that represent the axioms from which the knowledge structure is built.

The choice of the number of axioms is arbitrary, and in our case $2K$ was chosen so that the first theorem is not forced to use all present axioms (since each theorem is based on $K$ prerequisites), but the number of starting axioms does not influence the evolution of the corpus.

For $L > 1$ the knowledge corpus may contain holes, i.e., locations $\ell$ with $\ell < m$ and $X_\ell = 0$. We denote by $h(t)$ the number of holes in the corpus at time $t$.

We repeat the previous steps until the maximum knowledge $m$ is equal to $N$.

## 3 Numerical Results

A typical evolution for the maximum knowledge $m(t)$ and corpus size $c(t)$ are reported in Fig. 3. As one can see in Fig. 4, both values grows with the same trend regardless of the value of $K$ and $L$, separated by a gap that remains finite after an initial growth. This gap is related to the number of holes in the knowledge corpus.

By plotting $c(t)$ in a log-log scale, as shown in Fig. 5, we can notice that, regardless of the values chosen for $L$ and $K$, it is asymptotically approximated

by an power law

$$m(t) = \mu(K, L)t^{\alpha}$$

and

$$c(t) = \gamma(K, L)t^{\alpha},$$

with $\alpha = 1/2$.

Looking at Fig. 5, we can notice that the transient depends on $K$ and $L$: increasing these values the curve becomes less linear in the left part of the graph. But the asymptotic trend is independent of $K$ and $L$, as shown in Fig. 6.

By investigating the number of holes $h(t)$, one sees that the value remains quite constant and that this effect is present regardless of the value of $L$ and $K$, as reported in Fig. 7. Indeed, $h$ increases at first, and the value around which the number of holes oscillates increases with $L$ and, in a less evident way, also with of $K$, but eventually, for every $K$ and $L$, the the filling of inner holes takes place, in correspondence with a slowing growth of $c(t)$.

## 4 Markov Birth-Death Approach

Let us consider the simplest case possible: $K = L = 1$, for which there is no hole in the corpus, since every piece of knowledge depends on the immediately previous one. In this case $c = m$.

For each time step, either the knowledge increases by one, or stays constant. Therefore it is an example of a birth-death Markov Chain, with no deaths [6].

Let us denote by $P(y, t)$ the probability that at time $t$ the corpus or maximum knowledge is $y$. The evolution of $P$ is given by

$$P(y, t+1) = P(y, t)\left(\frac{y-1}{y}\right) + P(y-1, t)\left(\frac{1}{y-1}\right), \tag{1}$$

where there are two processes: the knowledge increases by one if the $j_1$ prerequisite is equal to $y - 1$ (with probability $1/(y-1)$, or stays constant and equal to $y$ if $j_1$ is one of the other $y - 1$ possibilities.

The average knowledge $\bar{y}(t)$ at time $t$ is defined as

$$\bar{y}(t) = \sum_{y} yP(y, t). \tag{2}$$

The Markov process starts from the condition $P(y, 0) = \delta_{y,2K}$.

The time and space continuous approximation of the Markov equation (1) (valid far from the initial conditions, for $y \gg 2K$ and $t \gg 0$) is:

$$\frac{\partial P}{\partial t} = -\frac{1}{y}\frac{\partial P}{\partial y}, \tag{3}$$

which implies that $P$ is a function of $y^2 - 2t$.

Therefore, in this approximation

$$\bar{y}(t) \propto \sqrt{2t} \tag{4}$$

for large $t$.

Indeed, this is the case, as shown in Fig. 8.

The approximation Eq (1) has been obtained in the case $L = 1$, $K = 1$, for which there is no hole. It is possible to generalize it to the case $L > 1$, imposing the absence of holes, as

$$P(y, t+1) = \frac{1}{Z(t)} \sum_{k=0}^{L} P(y-k-1, t) \left( \frac{L-k+1}{L(y-k-1)} \right),$$

where $y-k-1 > 0$ and $Z$ is a normalization constant, such that $\sum_y P(y,t) = 1$. This generalization is justified by the fact that even for $L > 1$ the number of holes stay limited. This approximation reproduces the power-law growth of the corpus.

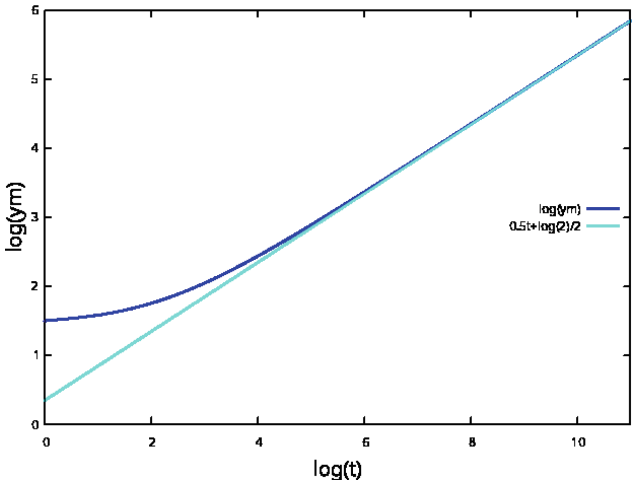



**Fig. 8.** Comparison of the $\bar{y}(t)$ (ym in the graph) with the function $0.5 \cdot t + \frac{\log(2)}{2}$, with $N = 10000$ and the total time of the simulation $T = 400000$. As shown in Eq. (4) there is a linear proportionality between time and $\bar{y}(t)$.

## 5 Conclusions

We have presented and analyzed a simple model of knowledge scaffolding, approximated by the growth and filling of a knowledge vector. The idea is that of providing a model corresponding to an ex-post analysis of the temporal development of a knowledge space.

The elementary step of our model is the proposal of a new theorem depending on at most $K$ prerequisites and providing a knowledge jump of at most $L$ steps.

This move is accepted and inserted in the corpus if the prerequisite does already belong to the known corpus and the result does not.

The size of the corpus, i.e., the number of known items and the maximum knowledge grows in time following a power law with exponent 1/2, regardless of the number of input items $K$ and jump $L$.

An analytical approximation, based on a death-birth Markov process is proposed, reproducing the power law.

We believe that this very simple and approximate model can be used as a basis for a qualitative description of more complex systems.

## References


1. Barabási, A.L.: Network science: Luck or reason. Nature **489**(7417), 507–508 (2012). https://doi.org/10.1038/nature11486
2. Bauer, P.J.: We know more than we ever learned: Processes involved in accumulation of world knowledge. Child Dev. Perspect. **15**(4), 220–227 (2021). https://doi.org/10.1111/cdep.12430
3. Bird, A.: What is scientific progress? Noûs **41**(1), 64–89 (2007). https://doi.org/10.1111/j.1468-0068.2007.00638.x
4. Bird, A.: Scientific progress as accumulation of knowledge: a reply to rowbottom. Stud. History Philos. Sci. Part A **39**(2), 279–281 (2008). https://doi.org/10.1016/j.shpsa.2008.03.019
5. Cong, L.W., Xie, D., Zhang, L.: Knowledge accumulation, privacy, and growth in a data economy. Manage. Sci. **67**(10), 6480–6492 (2021). https://doi.org/10.1287/mnsc.2021.3986
6. Karlin, S.: A First Course in Stochastic Processes. Academic Press (2014)
7. Kuo, C.I., Wu, C.H., Lin, B.W.: Gaining from scientific knowledge: the role of knowledge accumulation and knowledge combination. R&D Manage. **49**(2), 252–263 (2018). https://doi.org/10.1111/radm.12322
8. Mizrahi, M.: What is scientific progress? lessons from scientific practice. J. General Philos. Sci. **44**(2), 375–390 (2013). https://doi.org/10.1007/s10838-013-9229-1
9. Rowbottom, D.P.: N-rays and the semantic view of scientific progress. Stud. History Philos. Sci. Part A **39**(2), 277–278 (2008). https://doi.org/10.1016/j.shpsa.2008.03.010
10. Sizemore, A.E., Karuza, E.A., Giusti, C., Bassett, D.S.: Knowledge gaps in the early growth of semantic feature networks. Nat. Human Behaviour **2**(9), 682–692 (2018). https://doi.org/10.1038/s41562-018-0422-4
11. Wagener, T., Gleeson, T., Coxon, G., Hartmann, A., Howden, N., Pianosi, F., Rahman, M., Rosolem, R., Stein, L., Woods, R.: On doing hydrology with dragons: realizing the value of perceptual models and knowledge accumulation. WIREs Water **8**(6) (2021). https://doi.org/10.1002/wat2.1550