




# A Simple Model of Knowledge Percolation

Franco Bagnoli<sup>1,2</sup>  and Guido de Bonfioli Cavalcabo<sup>1</sup>

<sup>1</sup> 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

<sup>2</sup> INFN, sez. Firenze, Firenze, Italy

**Abstract.** We investigate how knowledge percolates and clusters in a given knowledge space. We introduce a simple model of knowledge organization in which each contribution spans a certain number of items. If this contribution overlaps with others above a certain threshold, they form a cluster. A contribution can also merge clusters together. We study the growth of global knowledge and the cluster dynamics, both showing a nontrivial behavior.

**Keywords:** Knowledge modelling · Knowledge visualization · Percolation model · Cluster dynamics · Agent based-model

## 1 Introduction

Knowledge is the set of ideas, emotions, beliefs and experiences, such as facts (descriptive knowledge), skills (procedural knowledge), or objects (acquaintance knowledge) owned by an individual or shared across collaborating individuals [7, 11]. It can be roughly seen as a set of concepts linked by some relationship (e.g. derivations linked to prerequisites or axioms to form theorems). The set of knowledge items that are connected by a path of links can be denoted as a cluster of knowledge.

A good representation of this description is a network [2] where the single knowledge items are the nodes and the links represent connections among items. A connected cluster is a corpus of knowledge. By adding knowledge three things can happen: the new knowledge item is isolated and forms an isolated cluster, it might join an existing cluster, or it may act as a connection between two clusters, fusing them together.

Since percolation describes the patterns of linked elements under a stochastic or semi-stochastic connection mechanisms [8, 10], the process of filling the vector is analogous to a percolation process, and we can refer to it as the knowledge percolation problem [4, 14].

The reference scenario is that of reconstructing the process that has led to the accumulation of a given corpus of knowledge, and, more important, the underlying cluster dynamics. There are many models that interpret the formation of a

collaboration network by the random joining of individuals or contributions, i.e., the formation of a giant component by the establishment of random links. Our model is first of all bipartite, contributions contribute to the knowledge corpus, and the contribution overlaps gives the link among them. Moreover, we require a minimum overlap for establishing the link.

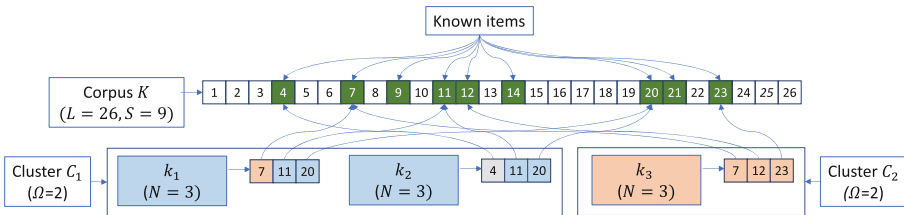
The whole corpus of knowledge can be spanned by several clusters separated by unknown elements of the corpus, or organized in a single cluster where all pieces of knowledge are connected by established relations, the process of acquiring knowledge has many similarity with the formation of a giant components in a random graph [3].

However, in a real case, redundant links are needed for considering concepts as belonging to the same cluster or discipline. So, we assume that a new knowledge item has to have minimum overlap with at least one of the members already belonging to the cluster to be inserted.

Alternatively, this model can be seen also as a collaboration model, in which every agent knows a certain number of concepts, but is able to collaborate with others (i.e. belong to the same group), only if they have a minimum overlap (like speaking the same language and having a shared background [5]), evaluating therefore the possibility of agents to collaborate or to communicate with others, that could be seen as the cooperation of individuals, research groups or societies, to solve a given problem represented by the knowledge vector.

Our model can serve as an interpretation tool for examining, ex-post, how a given corpus assembled and the relative cluster dynamics. For instance, one could study how authors cluster by measuring the overlaps between citations of their papers [9], or compare the evolution of customers of a supermarket by studying the overlap among their buying habits [13]. Our model is however still too rough to be compared with real data.

We study how the number and size of knowledge clusters evolve when adding new items. This can be considered as a k-core growth percolation problem, although normally k-core models are studied by pruning an existing network [6].



**Fig. 1.** Schematic representation of the system in the case of a knowledge space with  $L = 26$  items,  $N = 3$  and  $\Omega = 2$  divided into two clusters  $C_1 = \{k_1, k_2\}$  of size  $c_1 = 2$  and  $C_2 = \{k_3\}$  of size  $c_2 = 1$ .

## 2 The Model

We represented the corpus of knowledge  $K$  as a numbered set of  $L$  items, where  $K(n) = 1$  if the knowledge item  $n$  is present in the corpus and  $K(n) = 0$  if it is absent. Each contribution  $k_i$  is given by a set of  $N$  random items  $k_i^{(n)}$ ,  $n = 1, \dots, N$  among the available ones ( $k_i^{(n)} \in \{1, \dots, L\}$ ). When  $k_i$  is added to the corpus, we set  $K(k_i^{(n)}) = 1$  for  $n = 1, \dots, N$ .

The new contribution is added to a group if it has at least an overlap of  $\Omega$  to one of the elements of the group. A new contribution can also cause the fusion of two separated groups. This process is illustrated in Fig. 1.

Once fixed the values of  $L$ ,  $N$  and  $\Omega$ , the algorithm proceeds as follows:

- Randomly generate a contribution  $k_i$  with  $N$  random items  $k_i^{(n)}$  among the  $L$  available ( $1 \leq k_i^{(n)} \leq L$ ) without repetitions;
- Add this contribution to the knowledge corpus  $K(k_i^{(n)}) = 1$  for  $n = 1, \dots, N$ ;
- Check if there is any overlap with all previous contributions  $k_j$  ( $\forall j < i$ ).

By denoting this overlap  $\omega_{ij} = \sum_{nl} \delta_{k_i^{(n)}, k_j^{(l)}}$  (where  $\delta$  is the Kronecker delta), we can have three cases:

1.  $\omega_{ij} \geq \Omega$  and  $k_i$  not belonging to any group:  $k_i$  is added to the cluster  $C_m$  of  $k_j$ ;
2.  $\omega_{ij} \geq \Omega$  and  $k_i$  already belonging to a group: merge the cluster  $C_m$  of  $k_i$  and  $C_q$  of  $k_j$ ;
3.  $\omega_{ij} < \Omega$ : create a new group  $C_q$  and assign  $k_i$  to it.

There are two different dynamics occurring in our model: how the knowledge corpus size grows, and how clusters are formed or merged together. The first one is a representation of how knowledge grows in a single individual or in a group/society, while the second one could be seen as the representation of how different branches of knowledge are intertwined [12].

### 2.1 Knowledge Corpus Dynamics

Let us denote the corpus size by  $S = \sum_n K(n)$ . In the case  $L \gg N$ , at the beginning contributions do not overlap because the probability to have overlapping items is too low. Therefore the corpus size  $S$  grows linearly with time.

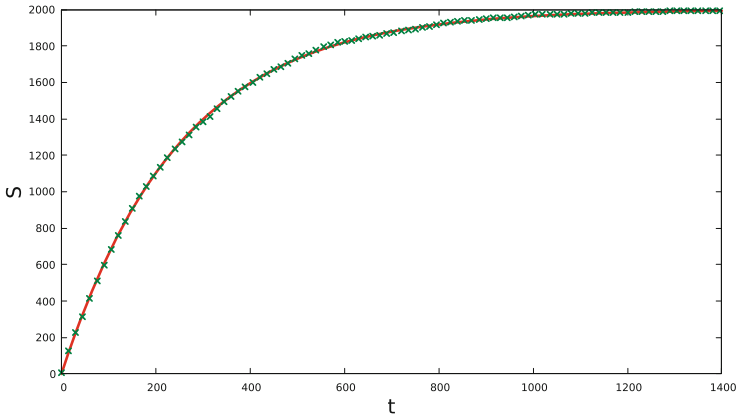
The population dynamics of the corpus is an independent stochastic process, with the probability of adding an original contribution decreasing in time ( $t$ ), while the number of those already present in the corpus ( $x$ ) increases.

Let us examine the case with  $N = 1$  for simplicity. The probability ( $P(S, t)$ ) of having  $S$  items already present at time  $t$  is

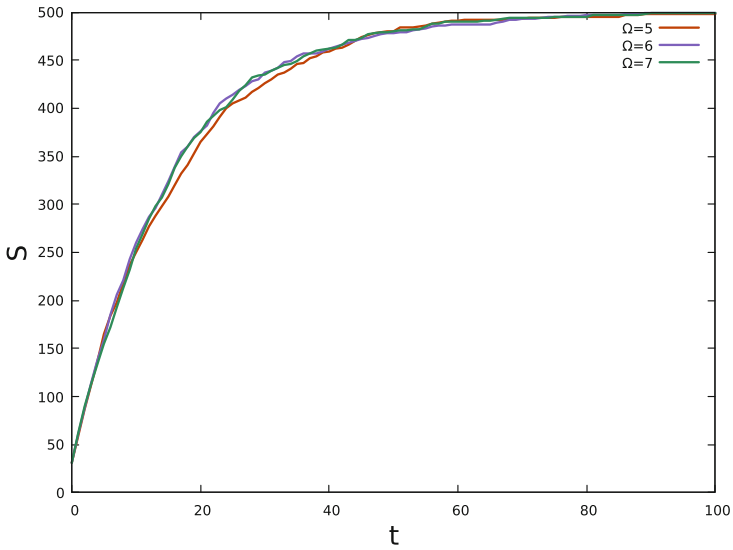
$$P(S, t + 1) = \frac{L - (S - 1)}{L} P(S - 1, t) + \frac{S}{L} P(S, t). \tag{1}$$

In the limit of continuous time and space, we have

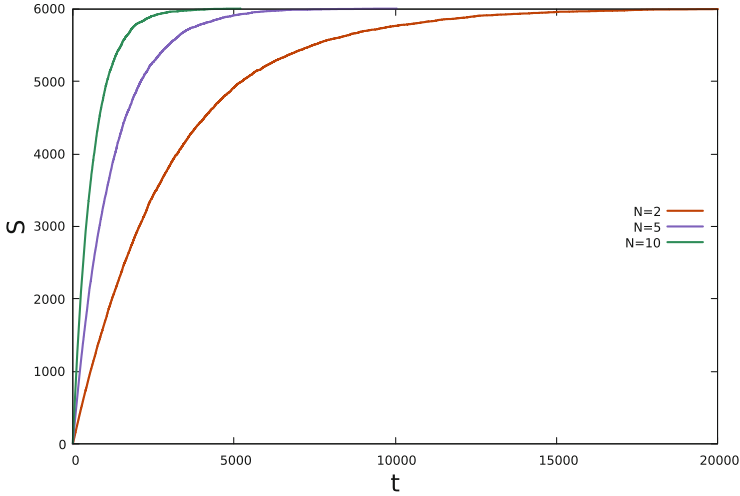
$$\frac{\partial P}{\partial t} = -\frac{L - S}{L} \frac{\partial P}{\partial S} \tag{2}$$



**Fig. 2.** Knowledge size  $S$  vs. time  $t$  for  $L = 2000$ ,  $N = 8$ ,  $\Omega = 3$  (crosses) confronted with Eq. 4 (continuous line) with the same values of  $N$ .



**Fig. 3.** Knowledge  $S$  vs. time  $t$  with  $L = 500$  and  $N = 31$  for several values of  $\Omega$ .



**Fig. 4.** Knowledge  $S$  vs. time  $t$  with  $L = 6000$  and  $\Omega = 1$  for several values of  $N$ .

And using the method of characteristics [1] we obtain that:  $P = f((L - S) \exp(\frac{t}{L}))$  and therefore the average knowledge ( $\bar{S}$ ) grows as  $\bar{S} = L(1 - C \exp(-\frac{t}{L}))$ , where  $C$  is an integration constant, fixed by the initial condition  $\bar{S}(0) = 0$ , so that  $C = 1$ ,

$$\bar{S} = L(1 - e^{-\frac{t}{L}}) \tag{3}$$

Since the addition of knowledge is an independent process we can write Eq. (3) with arbitrary  $N$

$$\bar{S} = L(1 - e^{-\frac{Nt}{L}}). \tag{4}$$

Confronting Eq. (4) with the results of simulations we get an almost complete overlap as shown in Fig. 2, even though in simulations  $\Omega > 1$ .

Indeed, it seems that the knowledge size  $S$  does not depend much on the value of  $\Omega$ , as shown in Fig. 3.

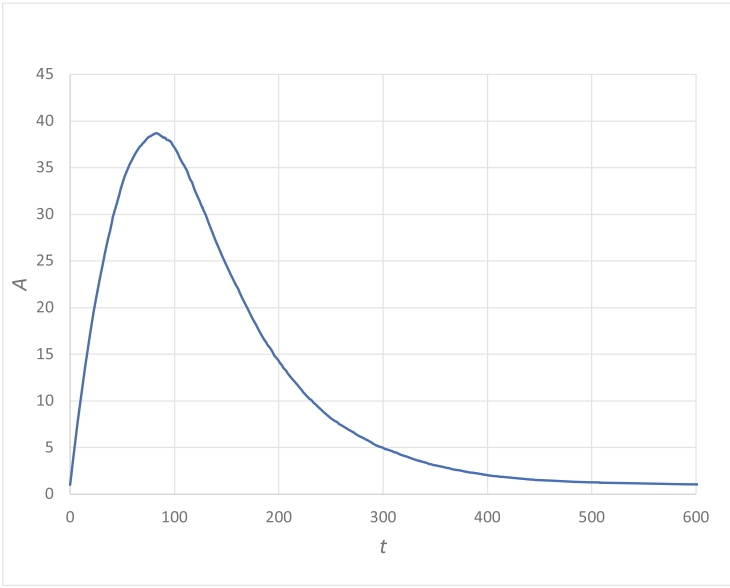
The knowledge size  $S$  grows faster for larger values of  $N$ , as shown in Fig. 4 and consistently with Eq. (4).

### 2.2 Cluster Dynamics

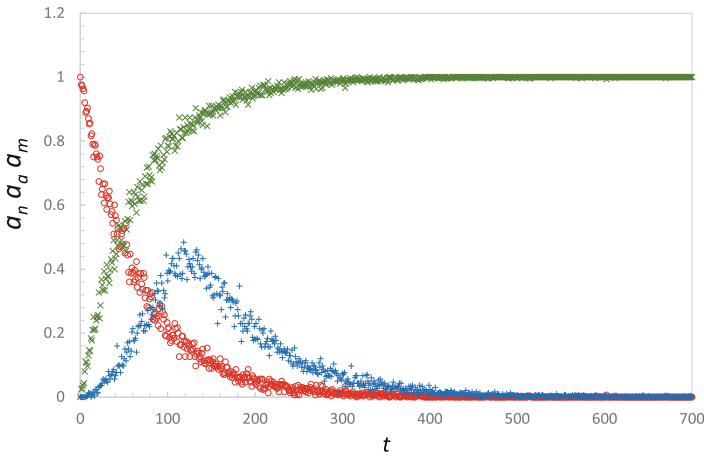
The number of clusters  $A$  at first grows with time, reaches a maximum and then decreases, ending with a single cluster, as reported in Fig. 5.

We can distinguish three phases:

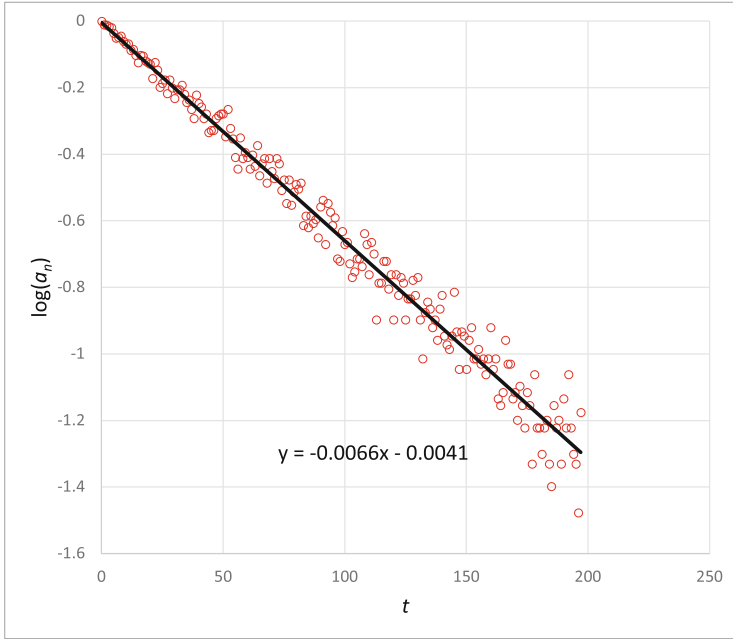
- An initial almost linear growth of  $A$ , where new contributions mostly form a new cluster;
- An intermediate phase with, in which new contributions are mostly added to an existing cluster;



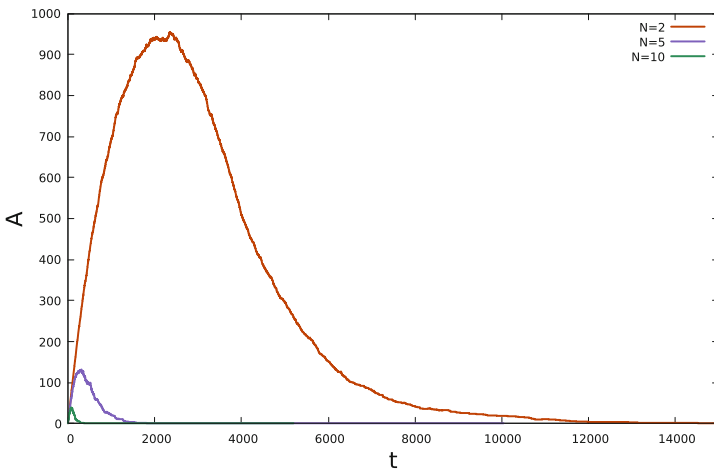
**Fig. 5.** Number of clusters  $A$  vs. time  $t$  for  $L = 600$ ,  $N = 2$  and  $\Omega = 1$ .



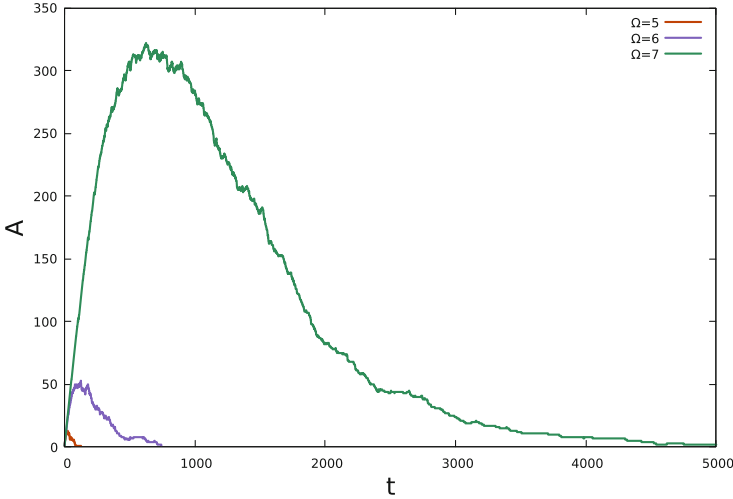
**Fig. 6.** Number of new clusters  $a_n(t)$  (red circles), number of additions  $a_a(t)$  (green times signs) and number of merging  $a_m(t)$  (blue pluses) vs. time for  $L = 600$ ,  $N = 2$  and  $\Omega = 1$ . (Color figure online)



**Fig. 7.** Number of new clusters  $a_n$  vs. time  $t$  in log-log scale for  $L = 600$ ,  $N = 2$  and  $\Omega = 1$ . The exponential decreasing factor is  $(3N - 4)/L$ .



**Fig. 8.** Number of clusters  $A$  vs. time  $t$  for  $L = 6000$ ,  $\Omega = 1$  and varying  $N$ .



**Fig. 9.** Number of cluster  $A$  vs. time  $t$  for  $L = 500$ ,  $N = 31$  and varying  $\Omega$ .

- A decreasing behavior dominated by cluster merging.

We measured the formation of a new cluster (number of new clusters  $a_n(t)$ ), the addition to an existing cluster (number of additions  $a_a(t)$ ) and the merging of two clusters (number of merging  $a_m(t)$ ), as shown in Fig. 6.

The actions of forming new clusters or adding it to an existing one (whether or not it causes a merging) are mutually exclusive, therefore  $a_n + a_a = 1$  and the total number of clusters  $A(t)$  is given by

$$A(t) = \sum_{\tau=1}^t a_n(\tau) - a_m(\tau) \tag{5}$$

We can develop a simple approximation for  $a_n$  in the case  $N = 1$ ,  $\Omega = 1$ , for which we have either the formation of a new cluster (of size one) or the addition of another cluster, and no cluster merging.

By denoting with  $P(A, t)$  the probability of having  $A$  clusters at time  $t$ , we have

$$P(A, t + 1) = \frac{L - A + 1}{L} P(A - 1, t) + \frac{A}{L} P(A, t)$$

which can be approximated by

$$\frac{\partial P}{\partial t} = \frac{A - L}{L} \frac{\partial P}{\partial A}$$

and therefore

$$P = f \left( (A - L) \exp \left( \frac{t}{L} \right) \right)$$



so that, for large  $A$  and small  $t$ , we have

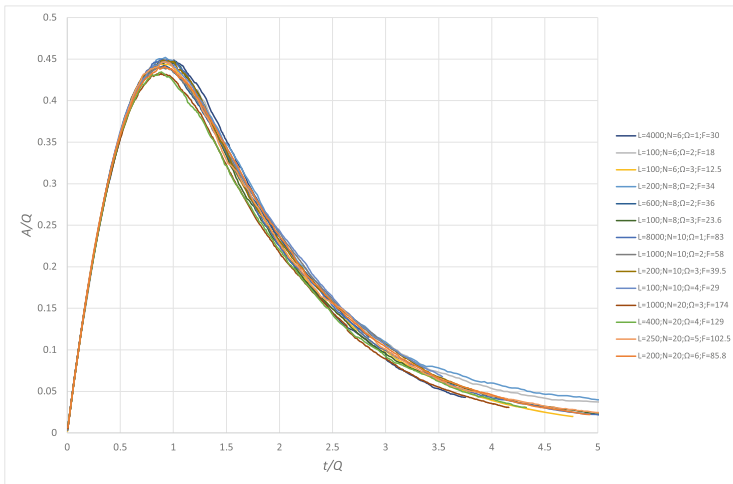
$$a_n(t) = \frac{d\bar{A}}{dt} \propto \exp\left(-\frac{t}{L}\right)$$

which, for  $N > 1$  corresponds to

$$a_n(t) \propto \exp\left(-\frac{g(N)t}{L}\right)$$

and numerically, as shown in Fig. 7, we have roughly

$$g(N) = 3N - 4.$$



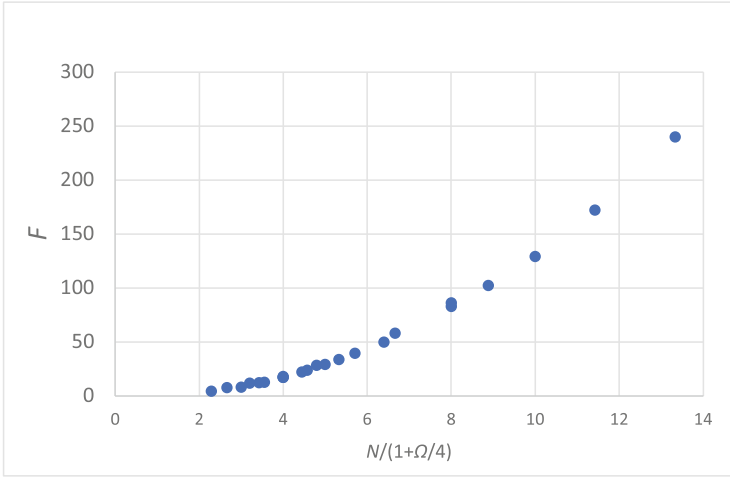
**Fig. 10.** Scaled  $A/Q$  vs  $t/Q$  with  $Q = (L/F)^\Omega$ .

The time distribution of the numbers of clusters  $A(t)$  depends on  $N$  and  $\Omega$ . In particular with a smaller number of items in each contribution (smaller value of  $N$ ) is less probable to get items have an  $\Omega$  overlap and therefore a larger number of clusters  $A$  will form before merging, as shown in Fig. 8.

On the contrary, having a smaller number of elements needed to match for merging (smaller values of  $\Omega$ ), means that the dynamics will reach the final state much faster, and the maximum number of clusters reached will be much smaller, as shown in Fig. 9.

It is therefore expected that  $A(t)$  scales with  $N$  and  $\Omega$ . Numerically, we found that all numerical curves overlap, as shown in Fig. 10, by rescaling  $A/Q$  and  $t/Q$ , with

$$Q = \left(\frac{L}{F}\right)^\Omega$$



**Fig. 11.** Scaling factor  $F$  vs.  $X = \frac{N}{1+\Omega/4}$ .

and

$$F = aX^2 + bX + c$$

with

$$X = \frac{4K}{4 + \Omega}$$

as shown in Fig. 11. We have not found a valid approximation for this behavior.

### 3 Conclusions

We investigate a problem related to knowledge percolation, clustering and formation of a giant component, somewhat related to k-core percolation problems.

We studied a simple model in which each contribution is constituted by a certain number of items, joining a cluster or even fusing two of them when the overlap exceeds a given threshold. We showed that the growth of global knowledge and the cluster dynamics has a nontrivial time behavior, providing some analytical approximations.

### References

1. Abbott, M.B.: An Introduction to the Method of Characteristics. American Elsevier (1966)
2. Barabási, A.L.: Network science: luck or reason. *Nature* **489**(7417), 507–508 (2012). <https://doi.org/10.1038/nature11486>
3. Ding, J., Kim, J.H., Lubetzky, E., Peres, Y.: Anatomy of a young giant component in the random graph. *Rand. Struct. Algorithm.* **39**(2), 139–178 (2010). <https://doi.org/10.1002/rsa.20342>

4. Essam, J.W.: Percolation theory. *Rep. Progr. Phys.* **43**(7), 833–912 (1980). <https://doi.org/10.1088/0034-4885/43/7/001>
5. Fleming, L., Frenken, K.: The evolution of inventor networks in the silicon valley and boston regions. *Adv. Complex Syst.* **10**(1), 53–71 (2007). <https://doi.org/10.1142/s0219525907000921>
6. Kong, Y.X., Shi, G.Y., Wu, R.J., Zhang, Y.C.: k-core: theories and applications. *Phys. Rep.* **832**, 1–32 (2019). <https://doi.org/10.1016/j.physrep.2019.10.004>
7. Kumbhar, R.: Knowledge organisation and knowledge organisation systems. In: *Library Classification Trends in the 21st Century*, pp. 1–6. Elsevier (2012). <https://doi.org/10.1016/b978-1-84334-660-9.50001-1>
8. Lee, D., Kahng, B., Cho, Y.S., Goh, K.I., Lee, D.S.: Recent advances of percolation theory in complex networks. *J. Korean Phys. Soc.* **73**(2), 152–164 (2018). <https://doi.org/10.3938/jkps.73.152>
9. Leicht, E.A., Clarkson, G., Shedden, K., Newman, M.E.: Large-scale structure of time evolving citation networks. *Eur. Phys. J. B* **59**(1), 75–83 (2007). <https://doi.org/10.1140/epjb/e2007-00271-7>
10. Li, M., Liu, R.R., Lü, L., Hu, M.B., Xu, S., Zhang, Y.C.: Percolation on complex networks: Theory and application. *Phys. Rep.* **907**, 1–68 (2021). <https://doi.org/10.1016/j.physrep.2020.12.003>
11. Renn, J.: *The Evolution of Knowledge: Rethinking Science for the Anthropocene*. Princeton University Press, Princeton (2020). <https://press.princeton.edu/books/hardcover/9780691171982/the-evolution-of-knowledge>
12. Shen, Z., et al.: Interrelations among scientific fields and their relative influences revealed by an input–output analysis. *J. Informet.* **10**(1), 82–97 (2016). <https://doi.org/10.1016/j.joi.2015.11.002>
13. Stassen, R.E., Mittelstaedt, J.D., Mittelstaedt, R.A.: Assortment overlap: its effect on shopping patterns in a retail market when the distributions of prices and goods are known. *J. Retail.* **75**(3), 371–386 (1999). [https://doi.org/10.1016/s0022-4359\(99\)00013-5](https://doi.org/10.1016/s0022-4359(99)00013-5)
14. Stauffer, D., Aharony, A.: *Introduction to Percolation Theory*. Taylor & Francis (1985). <https://doi.org/10.4324/9780203211595>