Flows of Knowledge in Citation Networks

Benjamin Renoust, Vivek Claver, Jean-François Baffier

Flows of Knowledge in Citation Networks

Benjamin Renoust [1], Vivek Claver [1,2], and Jean-François Baffier [2,3]

Abstract Knowledge is created and transmitted through generation. Innovation is often seen as a generative process from collective intelligence, but how does innovation emerges from the blending of accumulated knowledge, and from which path an innovation mostly inherit? A citation network can be seen as a perfect example of a generative process leading to innovation. Inspired by the notion of "stream of knowledge", we propose to look at the question of production of knowledge under the lens of DAGs. Although many works look for the evaluation of publications, we propose to look for production of knowledge within a framework for analyzing DAGs. In this framework inspired by the work of Strahler, we can also account for other well known measures of influence such as the *h*-index. We propose then to analyze flows of influence in a citation networks as an ascending flow. We propose an efficient dynamic algorithm for integration with modern graph databases, conducting our experiment with the Arxiv HEP-TH dataset. Our results validate the use of DAG flows for citation flows and show evidence of the relevance of the *h*-index.

1 Introduction

From the ancient times, knowledge passes from individuals to others leading at each step to more discoveries and innovations. In modern times, with the industrialization of research, it has become key to track this production of knowledge [16, 27]. Indeed, it is important for the newly produced innovation to state on which ground it stands, so peers can judge of the quality of the proposed innovation. An innovation must cite its influential sources to give credit to the work it was inspired from and to state its differences with the competing methods. This is one principle at the heart of the peer reviewing system enabling and validating the publication of new knowledge.

This process of citing sources is very important because it makes explicit the transmission of knowledge from prior works to an innovation [5] — and we can

[1]National Institute of Informatics & JFLI CNRS UMI 3527, Tokyo, Japan · [2]University of California Berkeley, Berkeley, USA · [3]JST-ERATO Kawarabayashi Large Graph project, Tokyo, JAPAN, · e-mails: vivek.claver@berkeley.edu, jf_baffier@nii.ac.jp, renoust@nii.ac.jp

consider each new scientific publication as a container of an innovation. Thankfully, this production of scientific knowledge can be easily captured in a citation graph. In this graph, nodes are publications citing other publications. This citation relationship is oriented and corresponds to a borrowing or derivation of knowledge, and we suspect that the impact of a publication can be captured in this graph. The production of knowledge would then be represented as a growing process in a dynamic network.

Key for countries and organizations in modern science, the study of the production of knowledge is mostly considered from partial indicators to establish rankings and compare scientists. This gave rise to the development of many measures deriving from sociometrics [28] including age, field, and other cues. Three major indicators are often used: the number of citations, the impact factor [25] (which is a time-related average number of citations of a collection) and the *h*-index [19]. These are popular indicators used for the evaluation of scientists, however they can be subject to controversy [24] and are designed to reflect only the productivity of a scientist rather than measuring the production of knowledge.

One reason these indicators' popularity is their simplicity in terms of computing. However, when previous network analysis was seen as too complex to deploy, modern graph databases have now grown to ease the analysis of dynamic networks [7]. Inspired by the seminal work from Strahler [26] and from Hirsh [19] we propose to bring a fresh look at the production of knowledge based on the analysis of flows in Directed Acyclic Graphs (DAGs). This view is not limited to the production of indicators but allows a more in-depth analysis of the process and diffusion of knowledge. The traditional indicators are very effective and it is important that our framework allows to establish them, while being easily extended.

We first introduce the Strahler numbers [26] and the *h*-index [19] in a generalized flow framework, and how those two notions belong to one greater notion of flow, and introduce our ascending flow – modeled on the notion of flow of knowledge. We will then discuss parameters of this ascending flow to put it in relation with classical measures. We propose a dynamic algorithm that allows for quick update. We finally run experiments on a publicly available dataset, the ArXiv HEP-TH [15].

2 Related works

The study of the production and transmission of knowledge has attracted quite many scholars in the domains of social and economical science [17], with for example a focus on the population at the origin of production [29], and of transmission to business [14]. These studies come *a posteriori* when observing controlled domains, with well known sociometric indicators. We are instead interested in the modeling of the production and diffusion of knowledge.

Many interesting attempts for modeling the production and diffusion of knowledge are actually focused on the producer of knowledge themselves, such as in multi-agent simulation [9, 10]. In these models, the agents are actually interacting to produce knowledge, and the properties of the resulting interaction network of agents are the focus of analysis. The agents can actually be tuned to produce different resulting networks, simulating real world policies [23]. Even on real social networks, the

topology of the networks of the people producing knowledge is the main focus of complex network research [11], because the focus is often to maximize diffusion in such network [1]. In contrast, our focus is on the information produced itself and how it relates to previous works.

A good model for this is the citation graph. It mostly apply to academic research, but have found its way in complex network research. Numerous works actually focus on communities [8], and the characterization of the dynamics of the citation graphs [15]. The closest to the spirit of our research would be the work by Hummon and Dereian [21] who studied the main paths in the citation network in order to extract backbones and areas of interest. The question of the efficient implementation of these cues has been the focus of a previous contribution [4]. An extension of Hummon and Dereian's original work has actually been applied to the study of the development of the h-index [22]. These methods are focused on the path produced by citations and use them as a base for bibliometrics, without capturing the global flow of information. We propose in contrast a natural interpretation of flows in DAGs that can easily capture the same measures used for main path analysis.

One of the most cited work in scientometrics is the *Hirsch index* [19], globally known as the h-index. It originally applies to the authors, and is designed to measures both the quantity and the quality of the authors' production. It was rapidly followed by numerous variants and extensions [28]. The most famous possibly is the g-index of Egghe [13] that is the largest number such that the g articles with the most citations receive at least a total of g^2, averaging the importance of each article. Hirsch [20] proposes a more restrictive version called $\bar{h}$-index, normalized to domain or age. Other variants could be mentioned (such as Bucur *et al.* [6]), but each is designed with specific goals. All-in-all, h-index based measures are measures to analyze the productivity of researchers, but do not allow for the in-depth analysis of production, in contrary to main path analysis approaches.

Our work roots its contribution in the analysis of flows in DAGs. Traditional max-flow approaches are quite far from what we define here, because nodes are always sources of information and edges have infinite capacities — we may be closer to multicommodity flows [2]. Instead, we mostly take our inspiration from a different notion of flows, in river streams, as defined by Strahler [26]. Limited to binary trees, this notion has seen a few extensions [3, 12, 18] with applications to graph visualization. These versions use flows to highlight and extract most relevant paths in DAGs and trees and relatively place elements one to another. We will use this approach and adapt it to the production of knowledge.

In this work we propose to join the different views on knowledge production in a recursive framework. In section 3, we place in this framework different measures such as the h-index and Strahler number. Section 4 introduces our proposition of a flow that captures the production of knowledge: the ascending flow. Finally, we provide experimental comparisons on the ArXiv HEP-TH dataset in section 5.

3 Preliminaries

We consider in our setting a citation graph $G = (V,E)$ in which a node $v \in V$ represents a publication, and a directed edge, hereafter an *arc*, $e(a,b) \in E$ is created when the article a cites an article b. We consider the graph as being directed acyclic (or DAG), although real-world data may introduce cycles, this is a marginal case that we will discard in our study.

In this setting, an author, a journal, proceedings or books can be modeled as collections of publications. Hence, by observing the collective impact of the collection we can characterize the influence this set of publications. In other words, in our citation graph formalism, collections are only sink nodes that can be sourced from the publications themselves. In this work, we will focus on measuring the impact of individual publications only, that can be trivially reported to authors and collections.

Definition 1. For a publication c, its neighborhood $\mathcal{N}(c)$ is the set of all the publications referring to c. The size of $\mathcal{N}(c)$ is simply its in-degree $d^-(c)$.

From its definition, the h-index applies in general trees of depth 3 and can actually be seen as a modified version of the Extended Strahler numbers [3], which generalize Strahler numbers [26] — limited to binary trees — to general trees. In this modification, a root node (*e.g.* an author) does not increase from his maximum valuated nodes, but instead gets weighted by the maximum Extended Strahler number of his direct descendants (*i.e.* the publications).

Strahler numbers have been designed to define the size of river streams based on a hierarchy of dependent streams. Transmission of knowledge is very similar in that sense with publications being tributary to prior works they inherit from, and becoming in turn sources for later works — the h-index then captures the latter quantity. However, we want a finer measure which could capture the impact of a publication across all citations it generated.

We defined above our citations graphs to be DAGs, and fortunately, Strahler numbers have also been extended to DAGs [12, 18]. Herman *et al.* [18] proposes a generic framework to compute the importance K of nodes in DAGs — including Strahler numbers — such as:

$$K(v) = K(\mathcal{N}(v)) = \begin{cases} c, \ \ if \mathcal{N}(v) = \emptyset \\ F(K(s_1), \ldots, K(s_p)) \ s_i \in \mathcal{N}(v) \ o.w. \end{cases} \tag{1}$$

c designates a constant for terminal cases (leafs, often $c = 1$), F is an application of the neighborhood of v. s_i represents the successors (or a_i ancestors) of node v. This framework is nothing but a generic recursive framework, but it allows us to redefine in it other measures. In this context, counting the number of citations would only require to modify the application $F(\mathcal{N}(v))$, such as $F(\mathcal{N}(v)) = |\mathcal{N}(v)| = d^-(v)$. Similarly, the Strahler number of a node v is then defined as:

$$F(\mathcal{N}(v)) = \begin{cases} 1, if \ d^-(v) = 0 \\ \max(K(s_1), \ldots, K(s_p)) + \begin{cases} p-1 \ if \ all \ values \ K(s_i) \ are \ equal \\ p-2 \ otherwise \end{cases} \end{cases} \tag{2}$$

The application for the h-index then becomes:

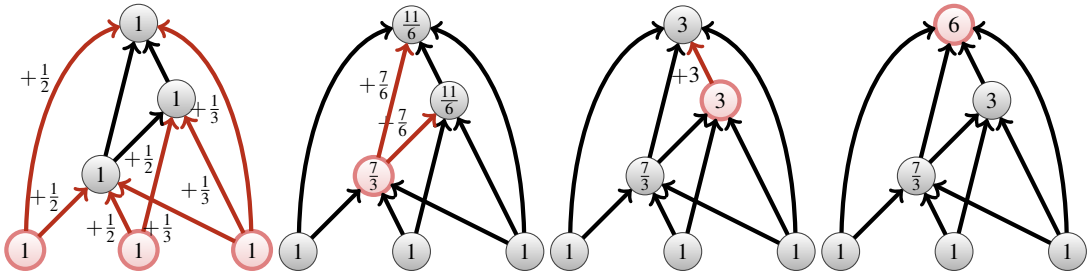


Fig. 1 Ascending flow algorithm: step by step

$$F(\mathscr{N}(v)) = \begin{cases} 0, if\ d^+(v) = 0 \\ 1, if\ d^+(v) = 1 \\ \max(K(k_1), \ldots, K(s_p)) \,|\, |\{K(s_j)\}| = n, with\ K(s_j) = n \end{cases} \tag{3}$$

Strahler numbers, number of citations, and h-index impose a discrete limit in depth which is conceptually an issue — there is no reason not to look for all the extended consequences of a publication. Instead, Herman *et al.* [18] propose in their framework a *Flow metric* for DAGs to emphasize the distribution of information to their successor such as:

$$F(\mathscr{N}(v)) = \begin{cases} 1, if\ d^-(v) = 0 \\ \sum_i K(a_i)/d^-(a_i)\ o.w. \end{cases} \tag{4}$$

In which a_i represents the ancestors of v (instead of the successors k_i). Note that this defines a *descending* flow measure which captures how much information all nodes in the network receive from a root node v, but does not give credit to v for its production of information. In addition, weights are only initialized by the source nodes, so no other node can bring to the flow.

4 Ascending flow in citation networks

We provide now a base measure called *ascending flow* and discuss its complexity. We then extend it to several variants, such as one that is restricted in depth, hence that fits better a dynamic context. Two natural definitions help defining our framework and its integration with existing metrics.

Definition 2 (Related). Two articles a and b are said to be related if and only if there exist a path from a to b or from b to a. They are k-related if they are related and if the shortest path between them is at most of length k.

Definition 3 (k-diffuse). A measure of a node v is k-diffuse when it limits its computation to a subgraph composed of the k-related nodes of v

4.1 Ascending flow

We can now model the stream of knowledge as a flow in our citation network. Indeed, each node — being a publication — produces some information and this production

of information gives credit to their ancestors (in history, or successors in the DAG) as they refer to them. This translates into the framework as:

$$F(\mathcal{N}(v)) = \sum_i K(k_i)/d^+(k_i) + \alpha_v \quad (5)$$

Where α_v represents the information created by the publication v — in practice we set $\alpha_v = 1$. Hence, the more a publication is influential the more credit it will propagate to its ancestors. In contrast to the previous *Flow metric*, our ascendant flow is not only applied to the reversed DAG, but is also equivalent to the sum of the flows computed for each sub-DAG induced by each node.

The ascending flow, formalized above, can be implemented as algorithm 1. It is important to notice that each arc is visited only once and that the total number of visits of all nodes is also equal to the number of arcs. The time complexity of our algorithm is then $\Theta(m)$ where m is the number of arcs. This key property is inherent to the pseudo-DAG nature of our citation network. As described in section 3, citation networks can be converted to DAG with minimum loss of information. However, even a linear time complexity is often too costly for large dynamic network.

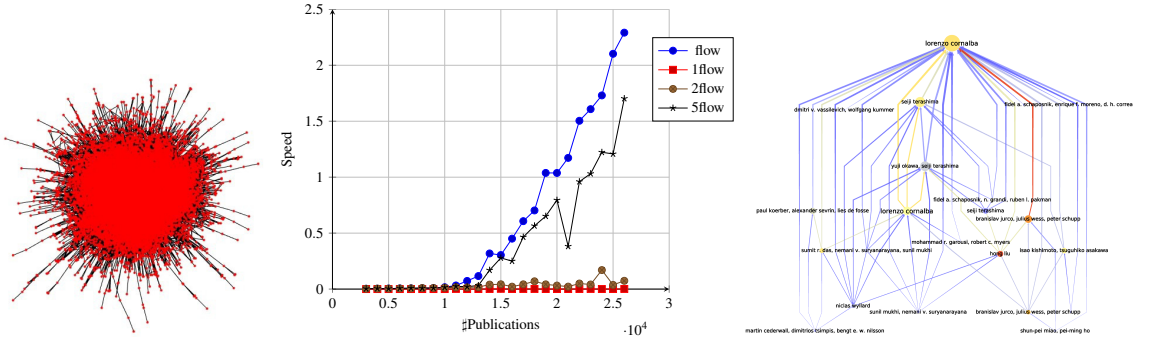
4.2 Depth restriction and dynamic graph

As discussed above, one issue of computing the ascending flow of a node v from our definition is that it needs the computation of all successors own influence. Such a constraint is expansive in the context of a dynamic network, for instance citation networks — in the case of citation network, publication are usually added, not removed. To adapt our previous algorithm, we first need to introduce an update function starting from a single leaf (a new publication). We consider the network initializes as in algorithm 1 but for the flow value on the nodes — that is kept between the updates. We then propagate upwards the flow value in all the subgraphs defined by the ancestors of this publication (Figure 1).

input : A citation network with nodes (articles) and arcs (citations)
An empty dequeue Q (FIFO)
output : The ascending flow on each node (article) and each arc (citation)

```
1  Initialize each article v a with flow value α_v = 1
2  Color each arc in white
3  Add all leaves in Q
4  while Q is not empty do
5  |   v ← pop_first(Q)
6  |   for each w son of v do
7  |   |   Color each (v, w) in blue
8  |   |   α_w ← α_w + α_v / d⁻(v)
9  |   |   if all incoming arcs of w are blue then
10 |   |   |   Q ← push_last(w)
11 |   |   end
12 |   end
13 end
```

Algorithm 1: ascending flow



(a) Arxiv HEP-TH (b) Speed comparisons (c) An example of the ascending flow metric

Fig. 2 (a) The main connected component of the ArXiv HEP-TH (high energy physics theory) citation network with 27770 nodes (articles) and 352807 arcs (citations). (b) Speed comparisons of our algorithm in case of k-diffuse limitations. (c) An example of the ascending flow metric in an excerpt of 22 nodes (60 edges) of our dataset, rooted by a publication by Lorenzo Cornalba. The size of nodes corresponds to their ascending flow in this subgraph. The color of nodes and edges (from blue to red) is actually their ascending flow in the real global dataset — we can see that Hong Liu's publication has probably been a seed for more knowledge than of its ancestor Lorenzo Cornalba. flows

Recall the diffuse property in definition 3. Our base measures, the h-index and the number of citations, are respectively 2- and 1-diffuse by definition, whereas the ascending flow is ∞-diffuse. In the real-world, we can consider that a publication that came a few generations after an original will relatively diverge from the original one, and would marginally contribute to the influence of the previous publication. The k-diffusion property can then take two forms: either we choose a generational limit k that cuts the added influence of nodes generated *after* k generations, or we can set an *evanescence* coefficient that progressively attenuates the contribution of a publication over its ancestors. In the case of a dynamic citation network, a k-diffuse measure is very quick to compute when k is a small constant as in Figure 2b.

This depth parameter additionally allows us to reconnect with known measures. For example, the h-index is 2-diffuse and it would not make sense to extend its definition. In turn, the number of citations — which is also the in-degree ($d^-(v)$) — is $1-$diffuse. This can then be easily translated in a k-diffuse measure, the k-degree, which would be the number of publications created until generation k. Then, an ∞-degree would be the number of all publications seeded by v even indirectly.

5 Experimental results

We now study our framework on a real-world setting. We used an available citation graph from 2003 KDD Cup: Arxiv HEP-TH[15][1]. It consists in an archive of 27,770 publications with 352,807 (internal) citations from the well-known ArXiv website of pre-prints in the domain of high energy physics theory, archived between January

[1] avaiable at: http://snap.stanford.edu/data/cit-HepTh.html

Pearson \ *Spearman*	*h*-index	ascending flow	∞-degree	1-degree	2-degree	5-degree	10-degree	20-degree	1-flow	2-flow	5-flow	10-flow	20-flow
h-index	-	0.821	0.765	0.958	0.954	0.849	0.770	0.765	0.776	0.809	0.807	0.807	0.807
ascending flow	0.546	-	0.758	0.858	0.807	0.764	0.759	0.758	0.961	0.990	0.991	0.991	0.991
∞-degree	0.476	0.267	-	0.715	0.809	0.947	1.000	1.000	0.654	0.710	0.714	0.714	0.714
1-degree	0.768	0.648	0.265	-	0.920	0.794	0.719	0.715	0.856	0.863	0.860	0.860	0.860
2-degree	0.850	0.670	0.375	0.766	-	0.908	0.815	0.809	0.725	0.776	0.775	0.775	0.775
5-degree	0.626	0.347	0.856	0.367	0.546	-	0.952	0.947	0.657	0.714	0.716	0.716	0.716
10-degree	0.483	0.270	0.999	0.268	0.381	0.865	-	1.000	0.654	0.710	0.714	0.714	0.714
20-degree	0.476	0.267	1.000	0.265	0.375	0.856	0.999	-	0.654	0.710	0.714	0.714	0.714
1-flow	0.637	0.694	0.330	0.904	0.638	0.367	0.332	0.330	-	0.987	0.985	0.985	0.985
2-flow	0.664	0.814	0.337	0.892	0.712	0.390	0.339	0.337	0.969	-	1.000	1.000	1.000
5-flow	0.656	0.823	0.341	0.879	0.704	0.392	0.344	0.341	0.964	0.999	-	1.000	1.000
10-flow	0.656	0.823	0.341	0.879	0.704	0.392	0.344	0.341	0.964	0.999	1.000	-	1.000
20-flow	0.656	0.823	0.341	0.879	0.704	0.392	0.344	0.341	0.964	0.999	1.000	1.000	-

Table 1 Comparison of Pearson coefficients (bottom left, correlation of values) and Spearman coefficient (top right, correlation of ranks) between all measures.

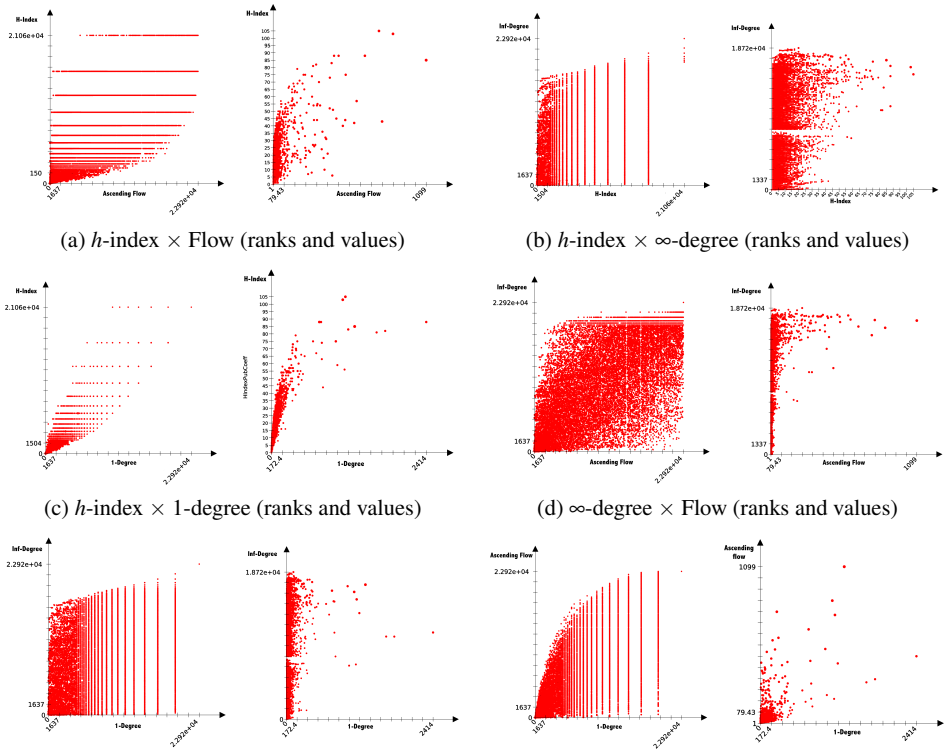
1993 to April 2003. The resulting graph (Figure 2a) is not acyclic due to the nature of publications in ArXiv — some publications have been updated with cross-references to others. We can however consider this graph as pseudo-acyclic because number and size of the cycles are limited (a few cycles of size 2 and 1 cycle of size 3). In our setting we simply remove those edges to keep the properties of a DAG. A resulting excerpt of the graph is shown in Figure 2c.

As we have defined the generalized version of the number of citations in our framework and the *h*-index, we compare these measures altogether. We compare the Pearson and Spearman correlation coefficients of these measures together with the following assumption: if the ascendant flow can reconnect at least partially to the notion of degree and *h*-index, we can then validate the relevance of our framework. Results of the analysis are presented in Table 1 and Figure 3.

First, when comparing the *h*-index, the number of citations, and the total number of publications produced by a work, we can notice a clear difference on our four basic metrics: the number of citations (=1-degree), the number of publications generated ($=\infty$-degree), the *h*-index and the ascendant flow. We additionally varied the depth of degree and flow in $\{1,2,5,10,20,\infty\}$. A second observation is that the limitation in depth of our measure is consistent with what we observe when limiting the depth of the *k*-degree (the most correlated *i*-flow for a *j*-degree is when $i=j$), and the higher *k* for the *k* degree, the more it diverges from the *k*-flow.

Our main observation, is, by value, the *h*-index is most correlated to the 2-degree. This makes complete sense, since the *h*-index is also limited in depth at 2 for which it considers a subset of publications. In contrast, when it comes to rankings, the *h*-index is most correlated to the 1-degree which is equivalent to the number of citations. Interestingly, our ascending flow also shares most correlations with the 2-degree as well and ranks with the 1-degree. This interesting effect may also be observed in Figure 2c showing that most publications bringing influence to the source publication has done it already in depth two. The link between the *h*-index and the degree is further observable in Figure 3.

In terms of computation, from $k=2$, the ranks obtained by the *k*-flow are .99 similar of those of the regular flow so when a gain of computation is needed, one can use *k*-diffuse version of the algorithm (Figure 2b).



(a) h-index $\times$ Flow (ranks and values)

(b) h-index $\times$ ∞-degree (ranks and values)

(c) h-index $\times$ 1-degree (ranks and values)

(d) ∞-degree $\times$ Flow (ranks and values)

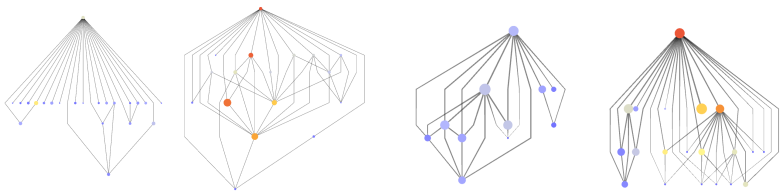
(e) ∞-degree $\times$ 1-degree (ranks and values)

(f) Flow $\times$ 1-degree (ranks and values)

Fig. 3 Comparative distribution of ranks and values among 1-degree (*i.e.* number of citations of a publications, ∞-degree (*i.e.* number total of generated publications), h-index, and ascendant flow. The plots well illustrate the difference between what those statistics are measuring.

Now we can compare publications of a same h-index and published around the same date which have very different flow measures. We took 2 publications with very different ascending flows: the first one shows a flow at 11.23 (Figure 4a, left), while the second one displays a flow measure at 425.44 (Figure 4a, right). Their in-degree does not vary that much (21 *vs.* 16 for the most influential), however, the 2-degree makes the difference (151, vs. 707). That means in average, the publications citing the most influential work produce more than four times more citations in turn – average h-index is 3.2 vs. 10.6. Note also that our measure takes into account how the information is spread out. In the first case, we have 390 citing edges out, while we have 171 in the other case.

We repeated the same experiment with two varying 2-flow measures (h-index =6 and similar date of publication): the first one is 2.25 with 10 citations (Figure 4b, left), and the second one is 21.59 with 20 citation (Figure 4b, right). The average h-index in the least influential one is actually higher (3.45) than of the most influential (1.80). However, the most influential has seeded 102 citations (2-degree) vs. 17 edges outs,



(a) ascending flows (b) 2-flows

Fig. 4 Comparison of direct citations of four publications with h-index =6. The top node is one original publication, and all other nodes its citing nodes (a) Comparison of the general ascending flows with two extreme values: left ID920426 (flow=425.4), right ID9201019 (flow=11.2) (b) Comparison of 2−flows with two extreme values: left ID9201079 (2−flow=2.3), right ID9201058 (2−flow=21.6). Relative node size (between couples of pictures) correspond to h-index values for each node. Node color correspond to, (a) ascending flow, (b) 2-flow.

when the first one 68 citations for 182 citing out. The flow measures then capture much more details of the graph of produced by citations than the h-index allows.

6 Discussion and conclusion

We have shown that the production and diffusion of knowledge can be modeled in a recursive framework that studies flows in DAGs, with a natural interpretation of the notion stream of knowledge. The framework allows for other known metrics to be embedded, and for efficient computation on large dynamic graphs. We applied our different flows and compared them with other known measures. By comparing the ascendant flow with the h-index we clearly see a correlation. The h-index has been a very popular indicator and useful for predictions and scientometrics. Our measure's interpretation is straightforward, and this correlation goes in favor of the relevance of the h-index. But we do not fully correlate with the h-index, and many cases that are oversimplified by the h-index can be finer described by the ascending flow.

We looked for differences in flow when the h-index gives a same value. We found cases with large differences, and explain the differences as follows: the h-index gives a rough estimation of a publication's production of knowledge, but it does not take into account how each citation refer to the original work. The flow measure, even 2−diffuse, is reinforced by two factors. A first one is something similar to a "community" effect in citations, *i.e.* when the citations produced also cite each other in relative proportion, in comparison to citations "outside" that "community" of citations. For example, this happens when a paper has an influence in developing a community of research, the large the community, the greater the flow. The second effect gets more relevant as the depth of diffusion is greater. It is somewhat close to the hubs and authorities effect: the more citations a paper gets from influential papers the more influential it will get.

The interpretation of flow we propose is much more flexible than the h-index, and can fairly support a wide range of parameters for scientists to conduct further experiments (such as additional weights, edge filtering, depth of influence, *etc.*). More

than a metric, when studying the influence of a work (or a collection of works), we argue that the structure of the flow of knowledge it produces, *i.e.* the DAG generated by a publication and its citations should be taken into account.

Although our study does not hold for an evaluation for which a comparison with many other metrics and regression would have been necessary, we still have set and validated the basis of our framework – in that it comprises well other known measures. Now, this will allow to take our graphs to another level of complexity – namely multiplex DAGs. H-index would apply with difficulty in a multiplex network, but we are currently focusing our effort in studying the ascendant flow in a version of our citation graphs where different routes could be considered in parallel (because knowledge does not flow equally in all citation sources). Among our future work is also the application to the analysis of news documents. Indeed, DAGs also apply to the study of closely related documents – even if there is no citation relationship, the time dependency between closely related documents can maintain the DAG assumption. Extending our study to other databases, such as DBLP, we would like to conduct case studies on authors and journals this time, to observe the influence of Nobel prizes or high standard journals.

References

[1] Alkemade, F., Castaldi, C.: Strategies for the diffusion of innovations on social networks. Computational Economics **25**(1-2), 3–23 (2005)

[2] Assad, A.: Multicommodity network flowsa survey. Networks **8**(1), 37–91 (1978)

[3] Auber, D.: Using strahler numbers for real time visual exploration of huge graphs. In: International Conference on Computer Vision and Graphics, vol. 1, p. 3 (2002)

[4] B., V.: Efficient algorithms for citation network analysis. CoRR **cs.DL/0309023** (2003)

[5] Bornmann, L., Daniel, H.: What do citation counts measure? a review of studies on citing behavior. Journal of Documentation **64**(1), 45–80 (2008)

[6] Bucur, O., Almasan, A., Zubarev, R., et al.: An updated h-index measures both the primary and total scientific output of a researcher. Discoveries **3**(3) (2015)

[7] Cattuto, C., Quaggiotto, M., et al.: Time-varying social networks in a graph database: A neo4j use case. In: First Int. Workshop on Graph Data Management Experiences and Systems, GRADES '13, pp. 11:1–11:6. ACM (2013)

[8] Chen, P., Redner, S.: Community structure of the physical review citation network. Journal of Informetrics **4**(3), 278–290 (2010)

[9] Cointet, J., Roth, C.: How realistic should knowledge diffusion models be? Journal of Artificial Societies and Social Simulation **10**(3), 5 (2007)

[10] Cowan, R., Jonard, N.: Knowledge creation, knowledge diffusion and network structure. In: Economics with heterogeneous interacting agents, pp. 327–343. Springer (2001)

[11] Cowan, R., Jonard, N.: Network structure and the diffusion of knowledge. Journal of economic Dynamics and Control **28**(8), 1557–1575 (2004)

[12] Delest, M., Don, A., Benois-Pineau, J.: Dag-based visual interfaces for navigation in indexed video content. Multimedia Tools and Applications **31**(1), 51–72 (2006)
[13] Egghe, L.: Theory and practise of the g-index. Scientometrics **69**(1), 131–152 (2006)
[14] Ernst, D., Kim, L.: Global production networks, knowledge diffusion, and local capability formation. Research policy **31**(8), 1417–1429 (2002)
[15] Gehrke, J., Ginsparg, P., Kleinberg, J.: Overview of the 2003 kdd cup. ACM SIGKDD Explorations Newsletter **5**(2), 149–151 (2003)
[16] Gibbons, M., Johnston, R.: The roles of science in technological innovation. Research Policy **3**(3), 220–242 (1974)
[17] Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M.: The new production of knowledge: The dynamics of science and research in contemporary societies. Sage (1994)
[18] Herman, I., Marshall, M.S., Melançon, G., et al.: Skeletal Images as Visual Cues in Graph Visualization, pp. 13–22. Springer Vienna (1999)
[19] Hirsch, J.E.: An index to quantify an individual's scientific research output **102**(46), 16,569–16,572 (2005)
[20] Hirsch, J.E.: An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics **85**(3), 741–754 (2010)
[21] Hummon, N., Dereian, P.: Connectivity in a citation network: The development of dna theory. Social networks **11**(1), 39–63 (1989)
[22] Liu, J., Lu, L.: An integrated approach for main path analysis: Development of the hirsch index as an example. Journal of the American Society for Information Science and Technology **63**(3), 528–542 (2012)
[23] Mueller, M., Bogner, K., Buchmann, T., et al.: Simulating knowledge diffusion in four structurally distinct networks: An agent-based simulation model (2015)
[24] Pendlebury, D.A.: The use and misuse of journal metrics and other citation indicators. Archivum immunologiae et therapiae experimentalis **57**(1), 1–11 (2009)
[25] Reuters, T.: The thomson reuters impact factor. thomsonreuters.com/products_services/science/free/essays/impact_factor/ (2012)
[26] Strahler, A.N.: Quantitative analysis of watershed geomorphology. Eos, Transactions American Geophysical Union **38**(6), 913–920 (1957)
[27] Van Raan, A.F.: Measuring science. In: Handbook of quantitative science and technology research, pp. 19–50. Springer (2004)
[28] Waltman, L.: A review of the literature on citation impact indicators. Journal of Informetrics **10**(2), 365 – 391 (2016)
[29] Wuchty, S., Jones, B.F., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science **316**(5827), 1036–1039 (2007)