Flow of knowledge in information networks

Jean-François BAFFIER[1]

[1] Japan Society for the Promotion of Science
French National Center for Scientific Research
Hosted at the Tokyo Institute of Technology

Monday, December 18, 2018

Network Optimization course

Table of contents

1 Introduction

2 Information Networks

3 Flow of knowledge

4 Multiplex Networks

5 Cycles, connected component and ethics

6 Conclusion

Contextualization

A not very funny joke

Three scientits, one in computer science, one in humanities, and one in biology enter a café. . .

Contextualization

A not very funny joke

Three scientits, one in computer science, one in humanities, and one in biology enter a café...

$\Rightarrow$ They have the same number of citations !

J.-F. Baffier (2018)

Contextualization

A not very funny joke

Three scientits, one in computer science, one in humanities, and one in biology enter a café...

$\Rightarrow$ They have the same number of citations !

J.-F. Baffier (2018)

The analyzis of academic citation networks depends strongly on

- Research field, era, host institution, fundings, ...
- Network size, its shape, ... so its structure
- Analytical tools used !

Comment en suis-je arrivé lá ?

Table of contents

1 Introduction

2 Information Networks

3 Flow of knowledge

4 Multiplex Networks

5 Cycles, connected component and ethics

6 Conclusion

The value of the transmission of knowledge

Education is the transmission of civilization.

(Areil and Will DURANT *— 1968)*

Structure is more important than content in the transmission of information. (The medium is the message)

(Abbie HOFFMAN *— 1968)*

An academic citations network

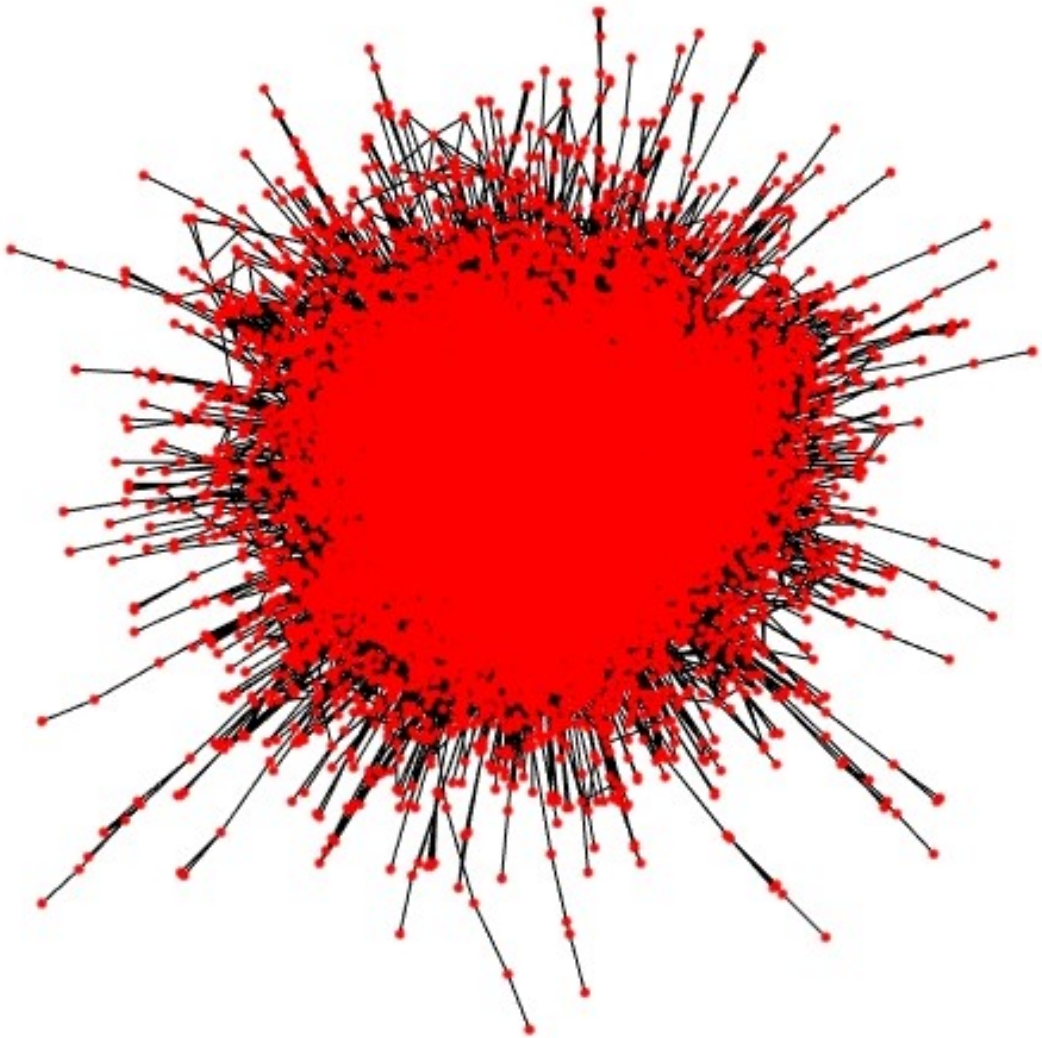


FIGURE – Arxiv HEP-TH (high energy physics theory) : 27770 nodes (articles) — 352807 arcs (citations)

An academic citations network

Expected properties :

- **Directed**
- (Pseudo-)**Acyclic**
- **Dynamic**
- **Stable** nodes (articles) : the papers are not modified after publication
- **Large** et **complex**
  - According to the University of Ottawa, in 2009 the amount of 50 millions of academic publications was reached (starting from 1665). Currently there is about 2.5 millions of new publications per year.
  - With the constant growth in the number of researcher and the global growing population, it is believed that the total number of publications will double every 9 years.
- **Unweighted** (on the links)
- **Information** is generated on each node (or at least on many)

Other information networks

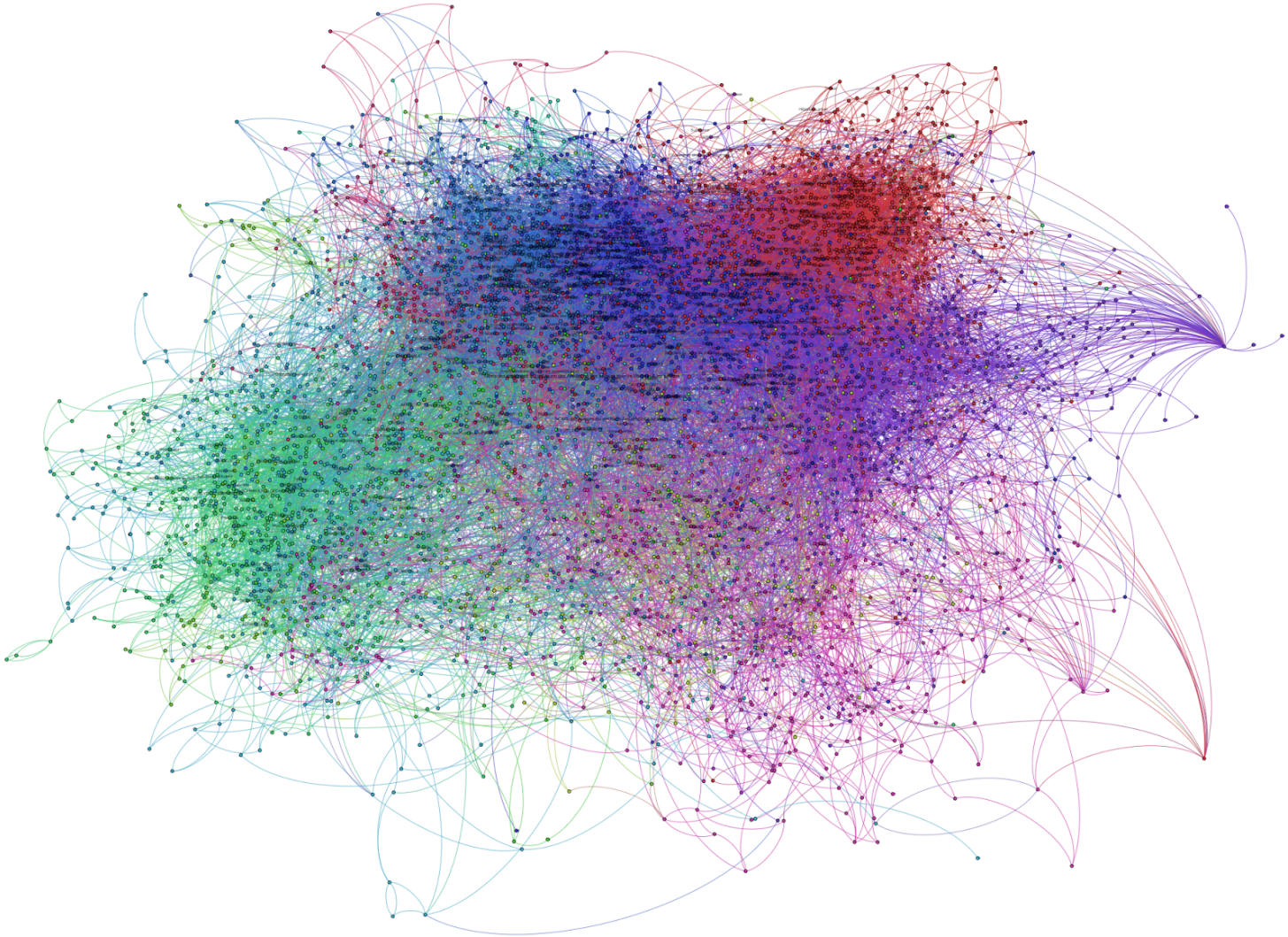
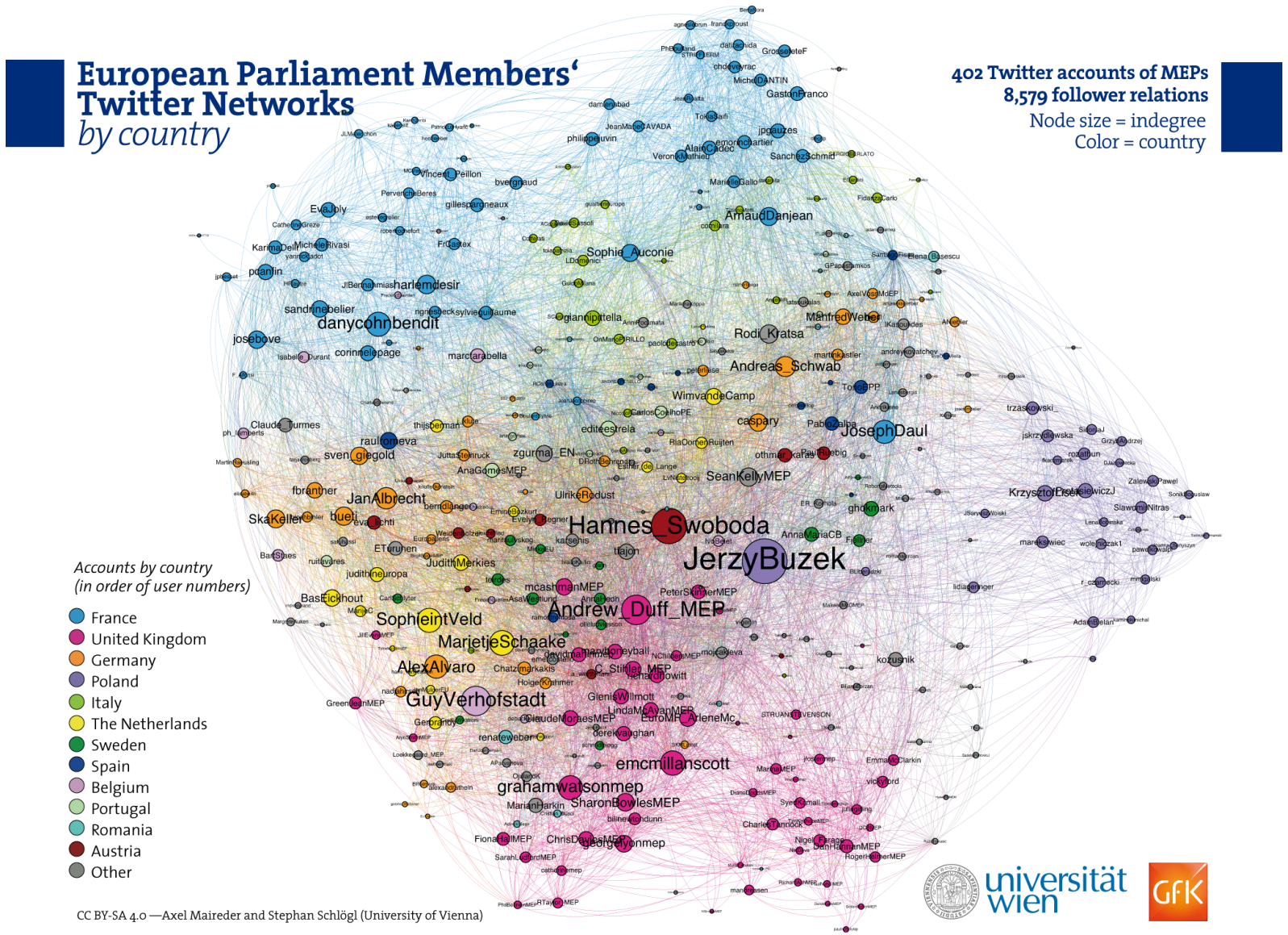


FIGURE – Réseaux d'interconnections des domaines et disciplines dans Wikipédia

Other information networks



Other information networks

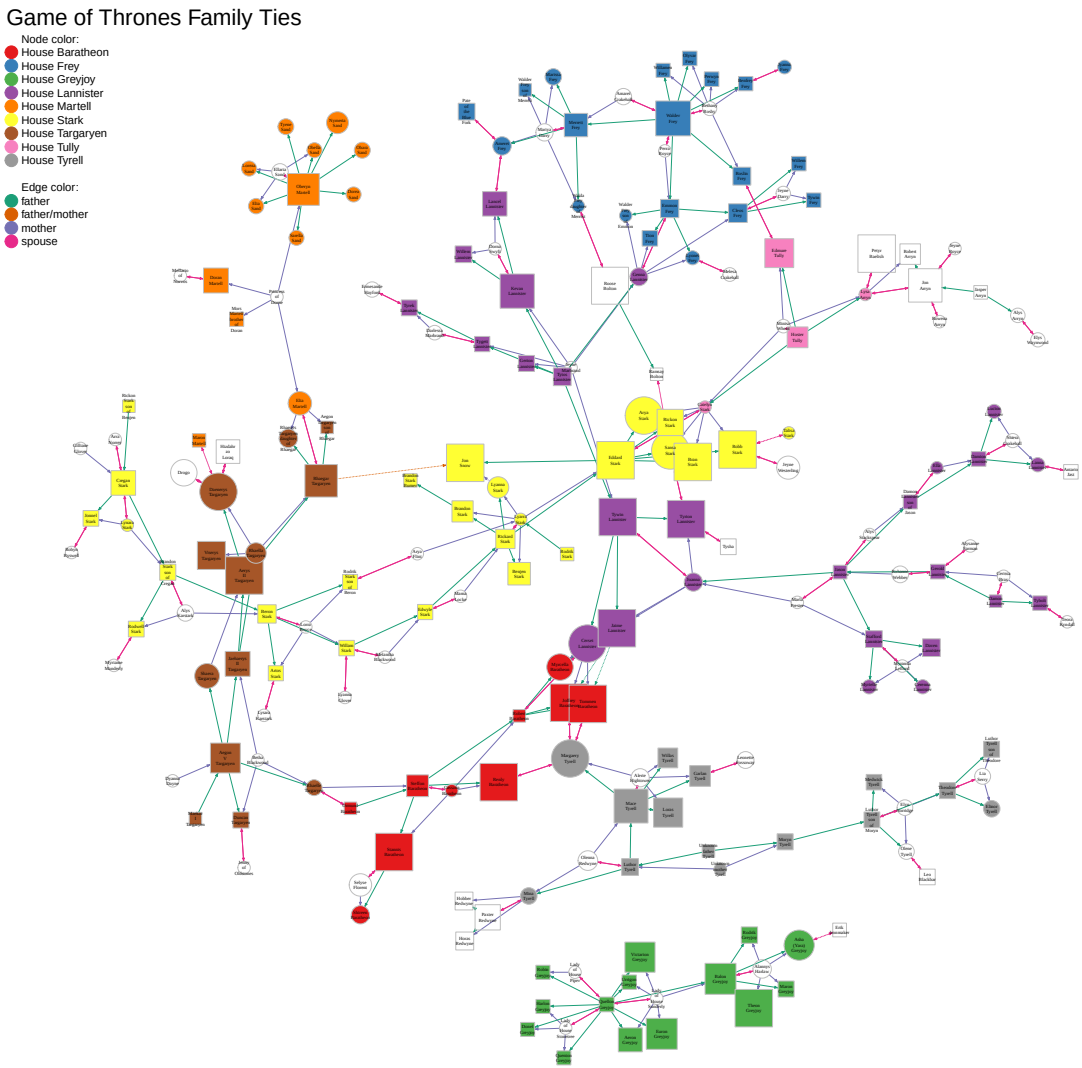


Table of contents

1 Introduction

2 Information Networks

3 Flow of knowledge

4 Multiplex Networks

5 Cycles, connected component and ethics

6 Conclusion

The *h*-index

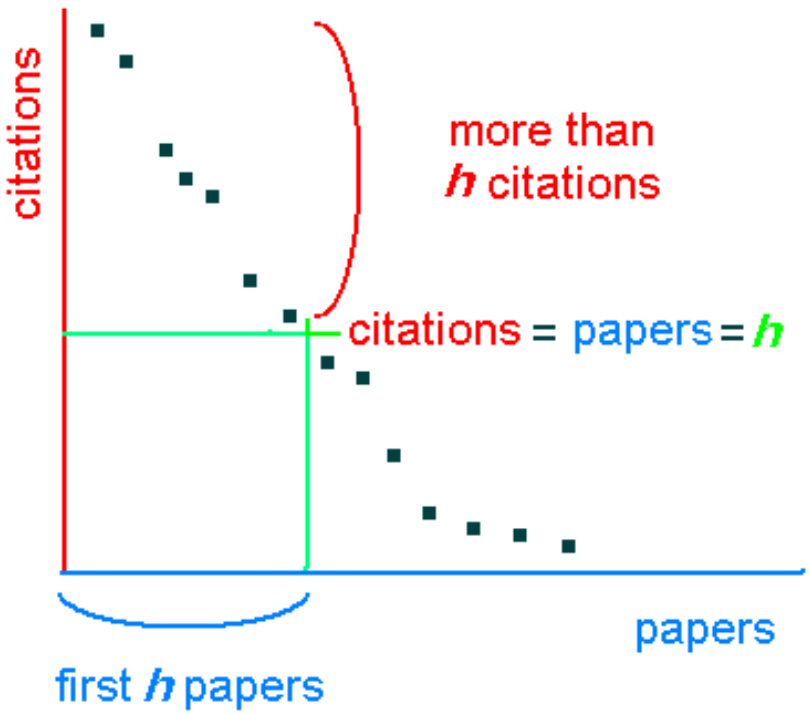


FIGURE – Short exercise : what is the value of *h* here ?

The Hirsch index aims to quantify the impact of a publication (or author).

- Quick computation
- Improvement over the simple metric of the number of direct citations
- Reference metric (good metric. . . on average)

The *h*-index

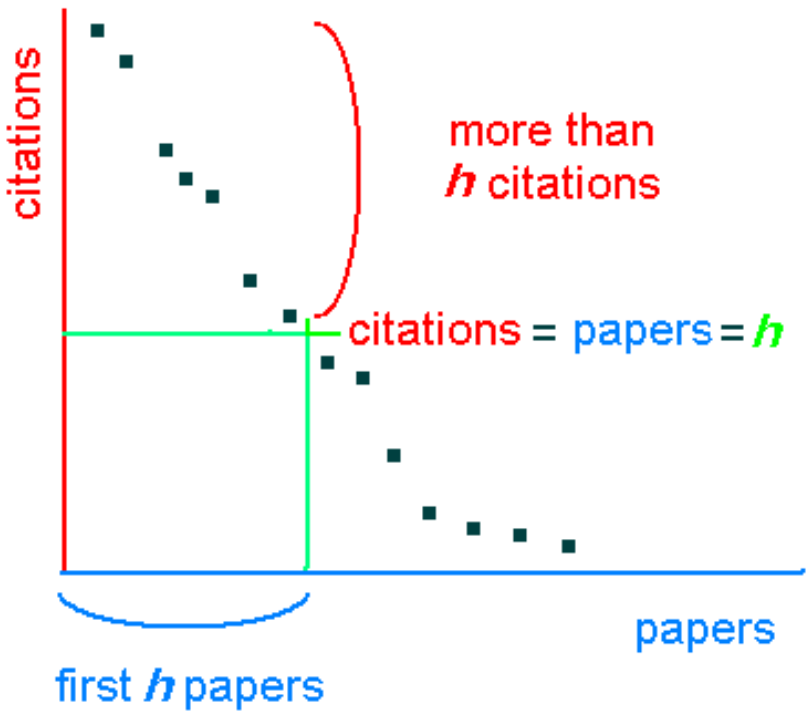


FIGURE – Short exercise : what is the value of *h* here ?

The Hirsch index aims to quantify the impact of a publication (or author).

- Quick computation
- Improvement over the simple metric of the number of direct citations
- Reference metric (good metric. . . on average)
- Only consider a small part of the citation network
- Is it actually useful to evaluate (to grade) the agents creating knowledge ?

Strahler's flow

The flow of Strahler is a first simple way to consider the influence (here the river flow) of an agent on the others.

How to extend it to the context of the transmission of knowledge ?

- Information should possibly be evaluated on an ascendant way
- Classic metrics (i.e. direct citation number, *h*-index, etc.) should be compatible
- Algorithmic cost should be reasonable

Size/Speed ratio			
n	n^2	n^3	2^n
10	100	1000	≈ 1000
100	10^4	10^6	$\approx 10^{30}$
1000	10^6	10^9	$\approx 10^{300}$

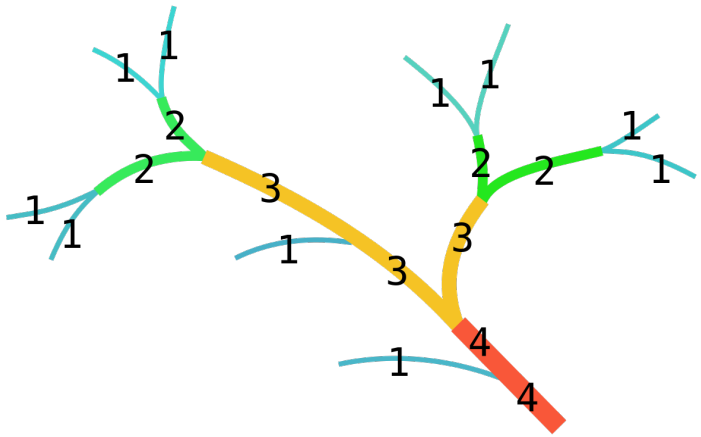


FIGURE – A Strahler flow example to evaluate the influence of each tributary river in the formation of the main river.
Strahler number (4 here) measure the complexity of the branching factor of the river.

A general framework

Definition

For a publication p, its **in-neighborhood** $\mathcal{N}^-(p)$ is the set of all the publications referring to p. The size of $\mathcal{N}^-(p)$ is simply its in-degree $d^-(p)$. The corollary implies that $\mathcal{N}^+(p)$, the **out-neighboorhood** of p, corresponds to all publications to which p is referring to (with a size $d^+(p)$, its out-degree).

$$K(p) = \begin{cases} \lambda, \text{ if } \mathcal{N}^+(p) = \emptyset \\ F(K(s_1), \ldots, K(s_{d^+(p)})), \text{ otherwise,} \end{cases} \quad (1)$$

where λ designates a constant for terminal cases (leafs, often $\lambda = 1$), $s_i \in \mathcal{N}^+(p)$ represents the successors of node p, and F is an application depending on the values $K(s_1), \ldots, K(s_{d^+(p)})$. To simplify the notations, we denote $F(\mathcal{N}^+(p)) = F(K(s_1), \ldots, K(s_{d^+(p)}))$.

A general framework

Strahler Flow

$$F(\mathcal{N}^-(p)) = \begin{cases} 1, \text{ if } d^-(p) = 0 \\ \max\limits_{q \in \mathcal{N}^-(p)} (K(q)) + \begin{cases} d^-(p) - 1 \text{ if all values } K(q) \text{ are equal} \\ d^-(p) - 2 \text{ otherwise} \end{cases} \end{cases} \quad (1)$$

h-index

$$F(\mathcal{N}^+(p)) = \begin{cases} 0, \text{ if } d^+(p) = 0 \\ \max\limits_{X \subset \mathcal{N}^+(p)} \min\limits_{q \in X} (d^+(q), |X|) \end{cases} \quad (2)$$

Definitions

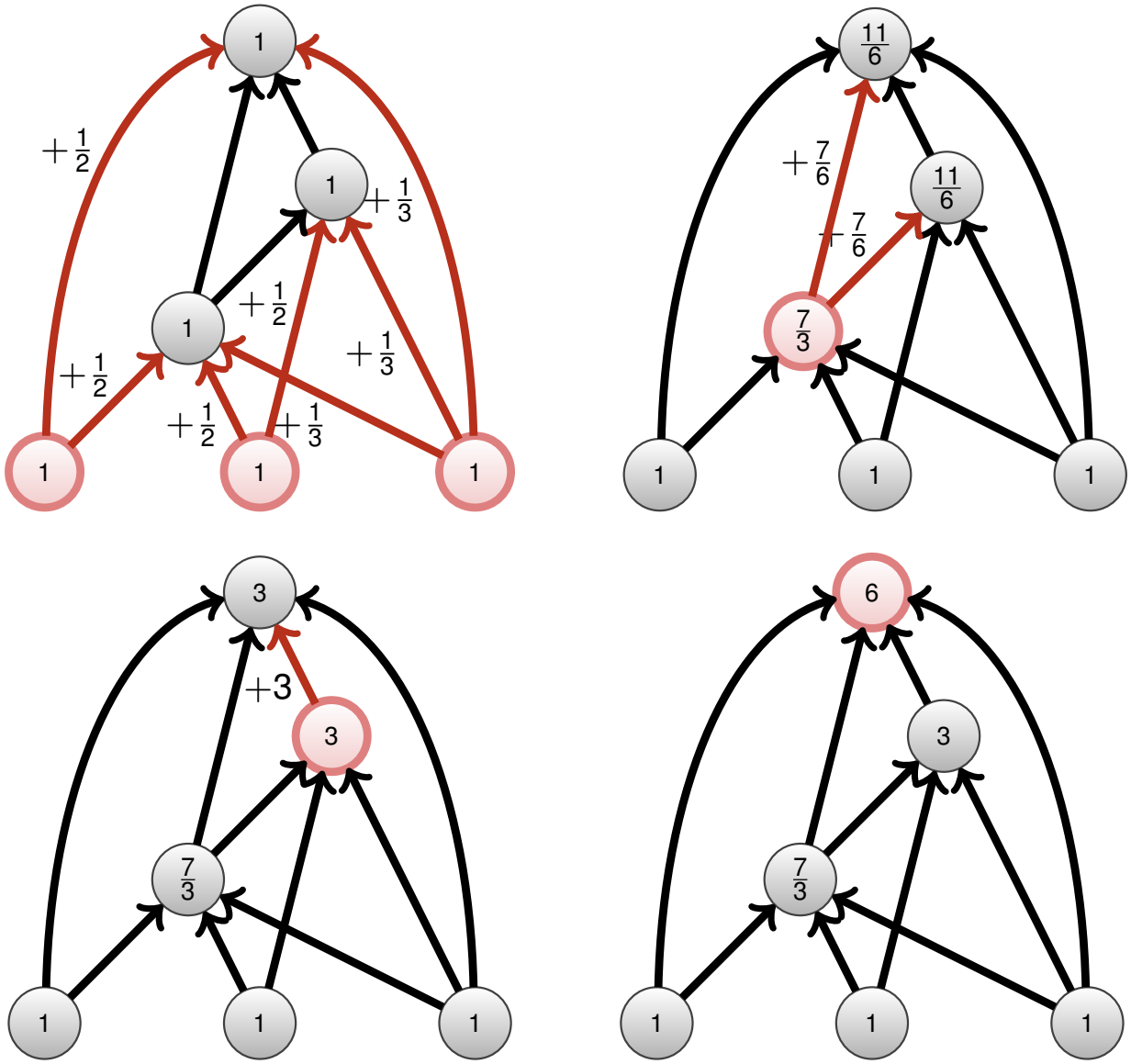
Definition (Related)

Two articles p and q are said to be related if and only if there exist a path from p to q or from q to p. They are k-related if they are related and if the shortest path between them is at most of length k.

Definition (k-diffuse)

A measure of a node p is k-diffuse when it limits its computation to a subgraph composed of the k-related nodes of p

A first straightforward ascending flow



A first straightforward ascending flow

ALGORITHM 1: ascending flow

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
```

A first straightforward ascending flow

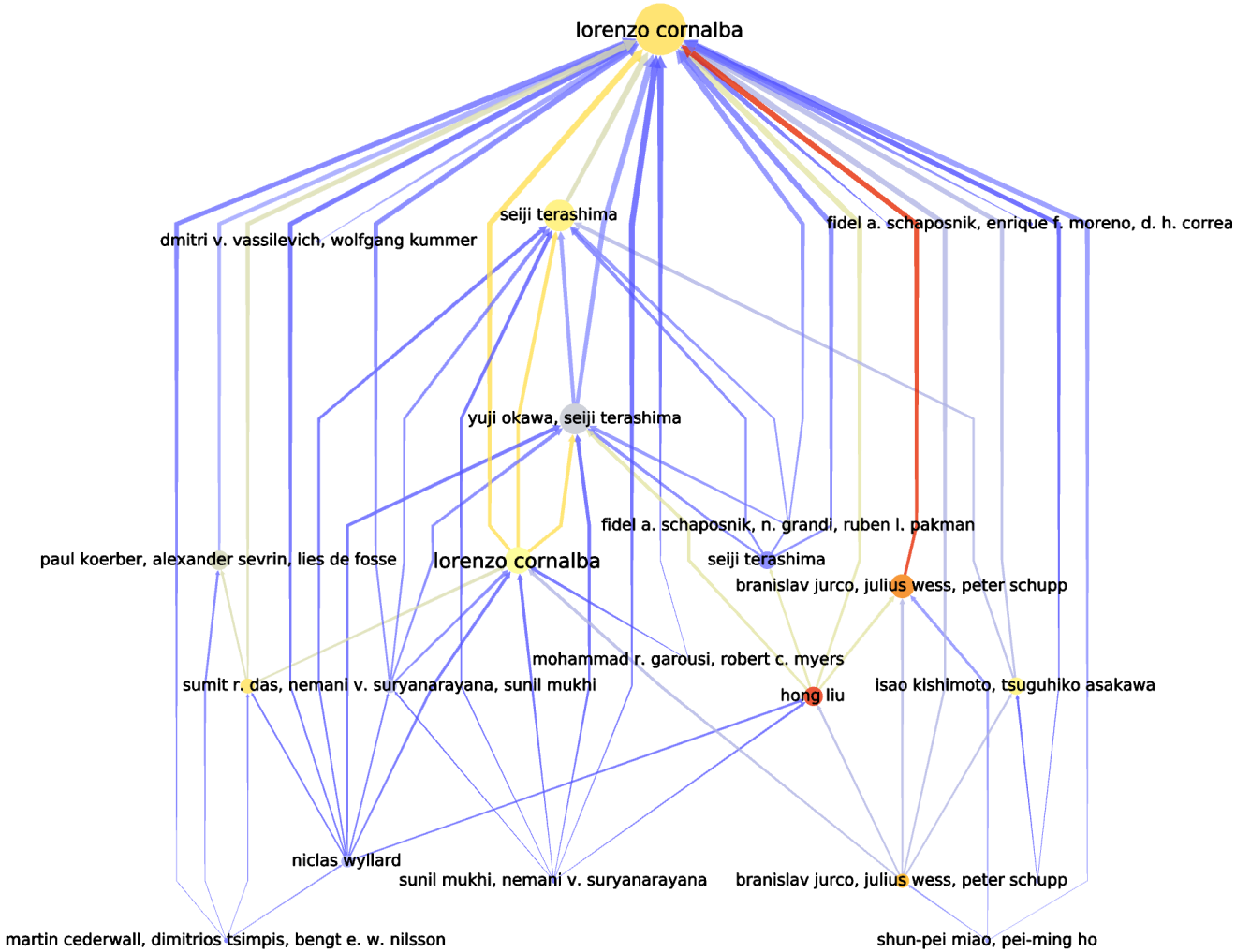
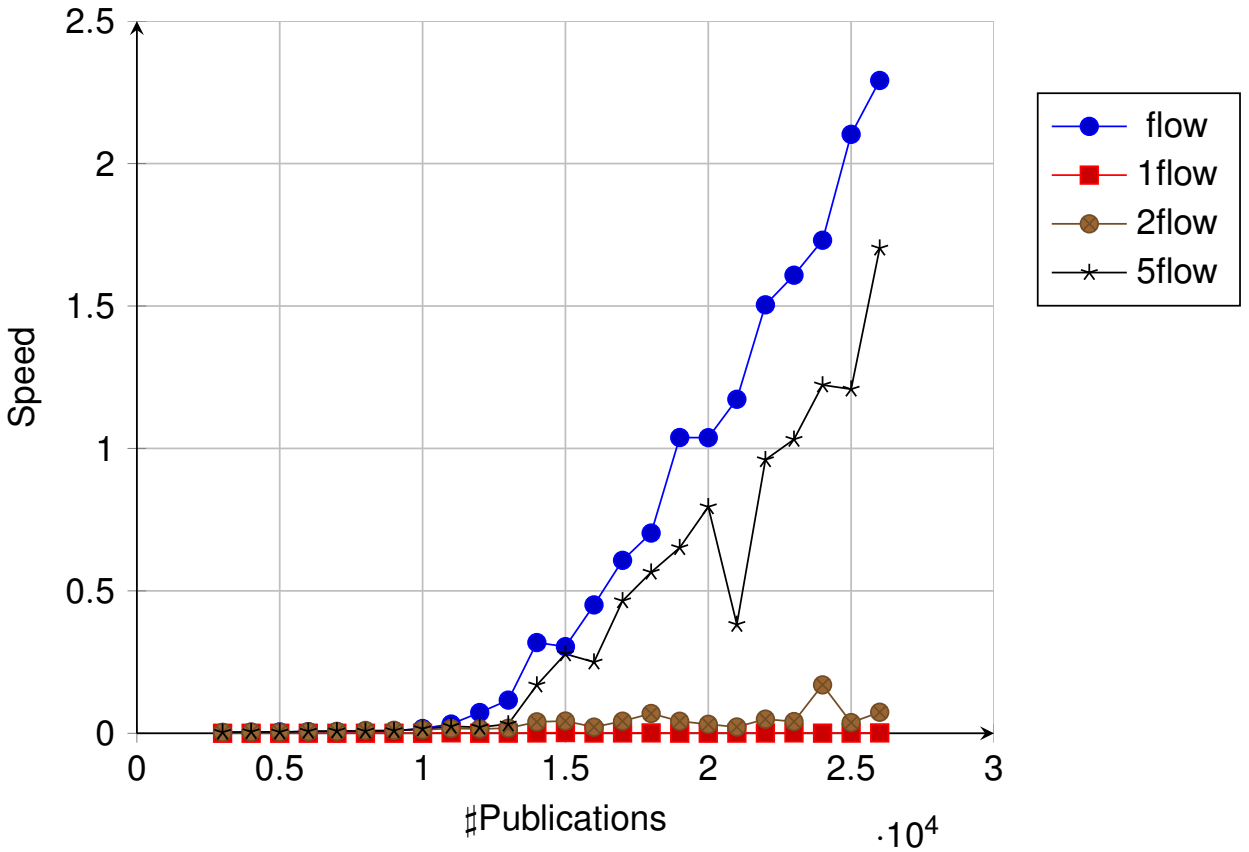


FIGURE – An ascending flow example (publications with a high *h*-index)

A first straightforward ascending flow



In terms of computation, from $k = 2$, the ranks obtained by the k-flow are $r_s = 0.99$ similar of those of the regular flow so when a gain of computation is needed, one can use k-diffuse version of the algorithm .

Table of contents

1 Introduction

2 Information Networks

3 Flow of knowledge

4 Multiplex Networks

5 Cycles, connected component and ethics

6 Conclusion

The intrication of data

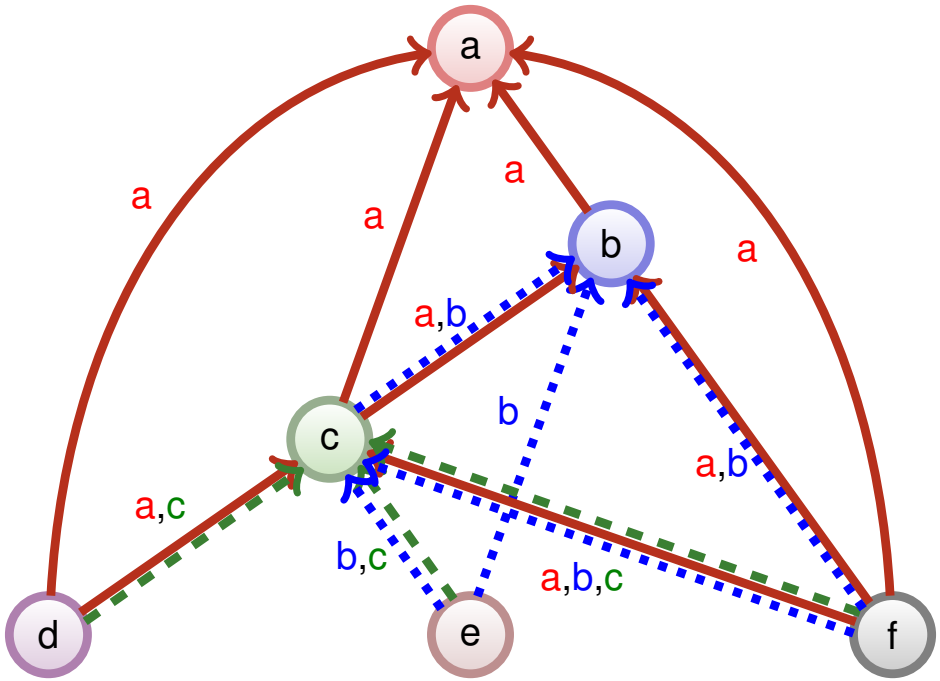


FIGURE – Illustration of the previous citation network transformed to a multiplex network, each citation creates its own layer of interaction.

The intrication of data

Definition

A citation subgraph $G_p = (V_p, E_p)$ is induced by a node p such as $V_p = p \cup \mathcal{N}^-(p) \subset V$, and $E_p \subset E$ is such that, $\forall (q, r) \in V_p^2, \ \exists a(q, r) \in E$.

Let consider, for each publication p, its induced citation subgraph $G_p = (V_p, E_p)$ of G, the multiplex citation network $\mathcal{G}$ results in combining all individual subgraphs G_p together.

Definition

A multiplex network $\mathcal{G} = (V, \mathcal{E}, \mathcal{L})$ connects nodes (p, q) on different layers l such as arcs $a(p, q, l) \in \bigcup_{l \in \mathcal{L}} \mathcal{E}_l$. A multiplex citation network $\mathcal{G} = (V, \mathcal{E}, \mathcal{L})$ is defined such as $\mathcal{G} = \bigcup_p^{p \in V} G_p$, hence $\mathcal{E} = \bigcup_p^{p \in V} E_p$ and $\mathcal{L} = V$.

Note that an arc $a(p, q, l)$ exists if and only if both p and q cite l or if $l = q$. As a consequence, the multiplex network once "flattened", has the exact same topology as the original citation network. The difference lies in the multiple edges.

The intrication of data

In our multiplex citation network, the notion of neighborhood remains the same as in the monoplex case. However we can refer to a different notion of multiplex degrees $\delta^+(p)$ and $\delta^-(p)$ that takes into account the number of arcs connecting a node to its neighborhood.

Definition

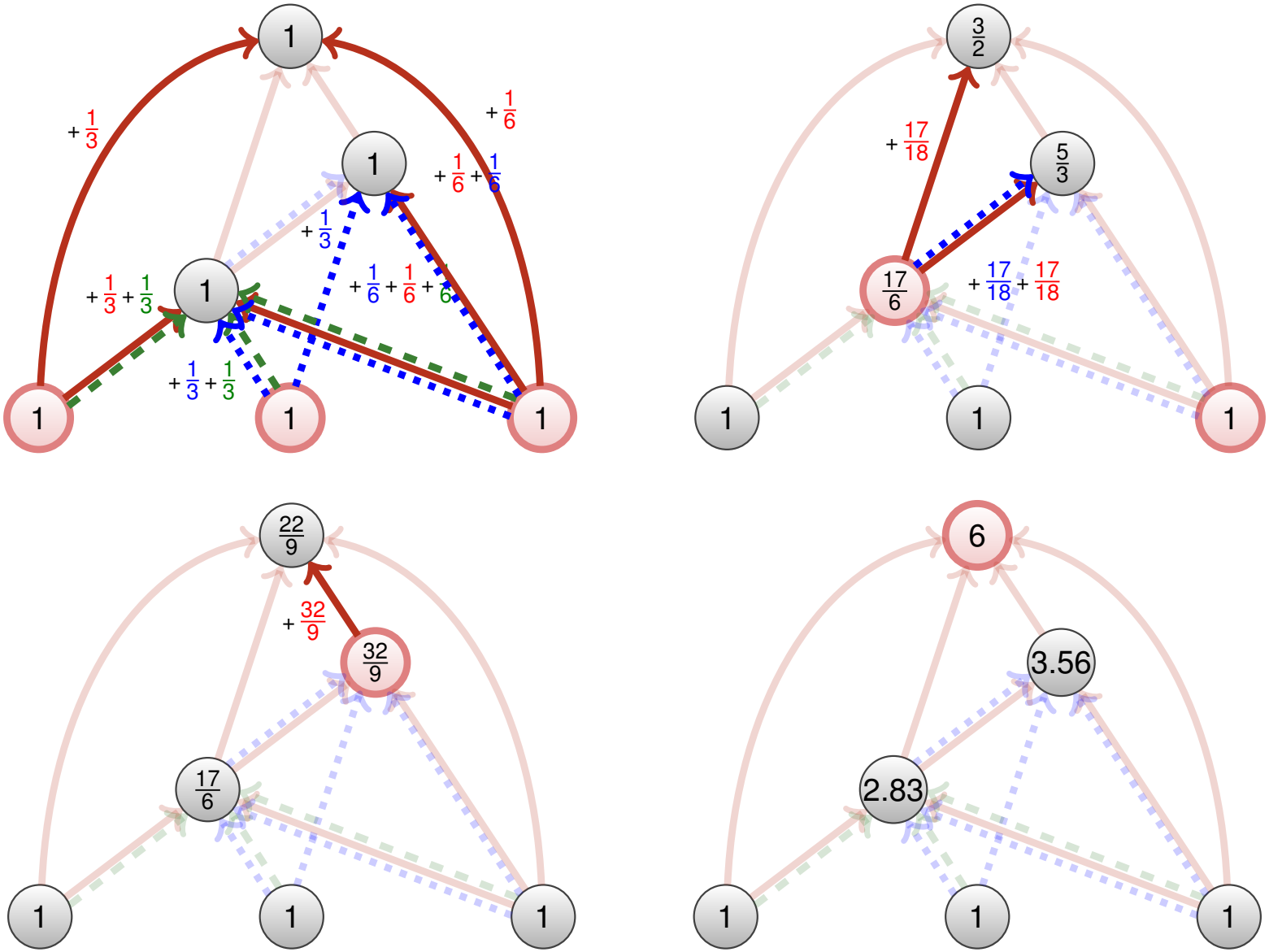
Denote the multiplex out-degree $\delta^+(p)$ (respectively the multiplex in-degree $\delta^-(p)$) a node p in the multiplex network $\mathcal{G}$.
$\delta^+(p) = |a(p,q,l)|, \forall q \in V,\ \forall l \in \mathcal{L},\ s.a.\ \exists a(p,q,l) \in \mathcal{E}$ (respectively $\delta^-(p) = |a(q,p,l)|$)

The degrees $d^+(p)$ and $d^-(p)$ still refer to the degree in the monoplex network G, **i.e.** the size of the neighborhood. We then introduce the degrees $d_l^+(p)$ and $d_l^-(p)$ corresponding to the degree in the subgraph G_l, **i.e.** the number of arc adjacent to p on the layer l.

Definition

Denote the layer out-degree $\delta_l^+(p)$ (respectively the layer in-degree $\delta_l^-(p)$) a node p in the subgraph $\mathcal{G}_{\updownarrow}$. $\delta_l^+(p) = |a(p,q,l)|, \forall q \in V,\ s.a.\ \exists a(p,q,l) \in \mathcal{E}$ (respectively $\delta_l^-(p) = |a(q,p,l)|$)

Aggregated Flow



Aggregated Flow

ALGORITHM 2: Aggregated flow (multiplex)

input : A citation network with nodes (articles) and arcs (citations)
An empty dequeue Q (FIFO)
A function α_{init} over the nodes
output: The ascending flow on each node (article) and each arc (citation)

```
1  Initialize each article p with a flow value λ_p = λ_init(p)(= 1 by default)
2  Color each arc in white
3  Add all leaves in Q
4  while Q is not empty do
5  |   p ← pop_first(Q)
6  |   for each q son of p do
7  |   |   for each layer l ∈ L_p do
8  |   |   |   Color a(p, q, l) in blue
9  |   |   |   λ_q ← λ_q + λ_p/δ⁺(p)
10 |   |   end
11 |   |   if all incoming arcs of q are blue then
12 |   |   |   Q ← push_last(q)
13 |   |   end
14 |   end
15 end
```

Aggregated Flow

General framework formulation

$$F(\mathcal{N}^-(p)) = \sum_{q}^{q\in\mathcal{N}^-(p)} K(k_q)/\delta^+(k_q) + \lambda_p \tag{3}$$

The complexity of the monoplex ascending flow is $\Theta(m)$ where m is the number of citations. Since the input is a multiplex network, in the worst case the number of links in the networks is equal to the number of nodes (layers) times the number of citations. Thus, the time-complexity of this aggregated flow is $\Theta(mn)$.

Sum Flow

This second extension of the ascending flow to a multiplex network consists in combining multiple monoplex versions of the ascending flow.
Recall the definition of the ascending flow, adapted to a subgraph G_l :

$$F_{G_l}(\mathcal{N}^-(p)) = \sum_{q}^{q\in\mathcal{N}^-(p)} K_{G_l}(k_q)/d^+_{G_l}(k_q) + \lambda_{(p,l)} \tag{4}$$

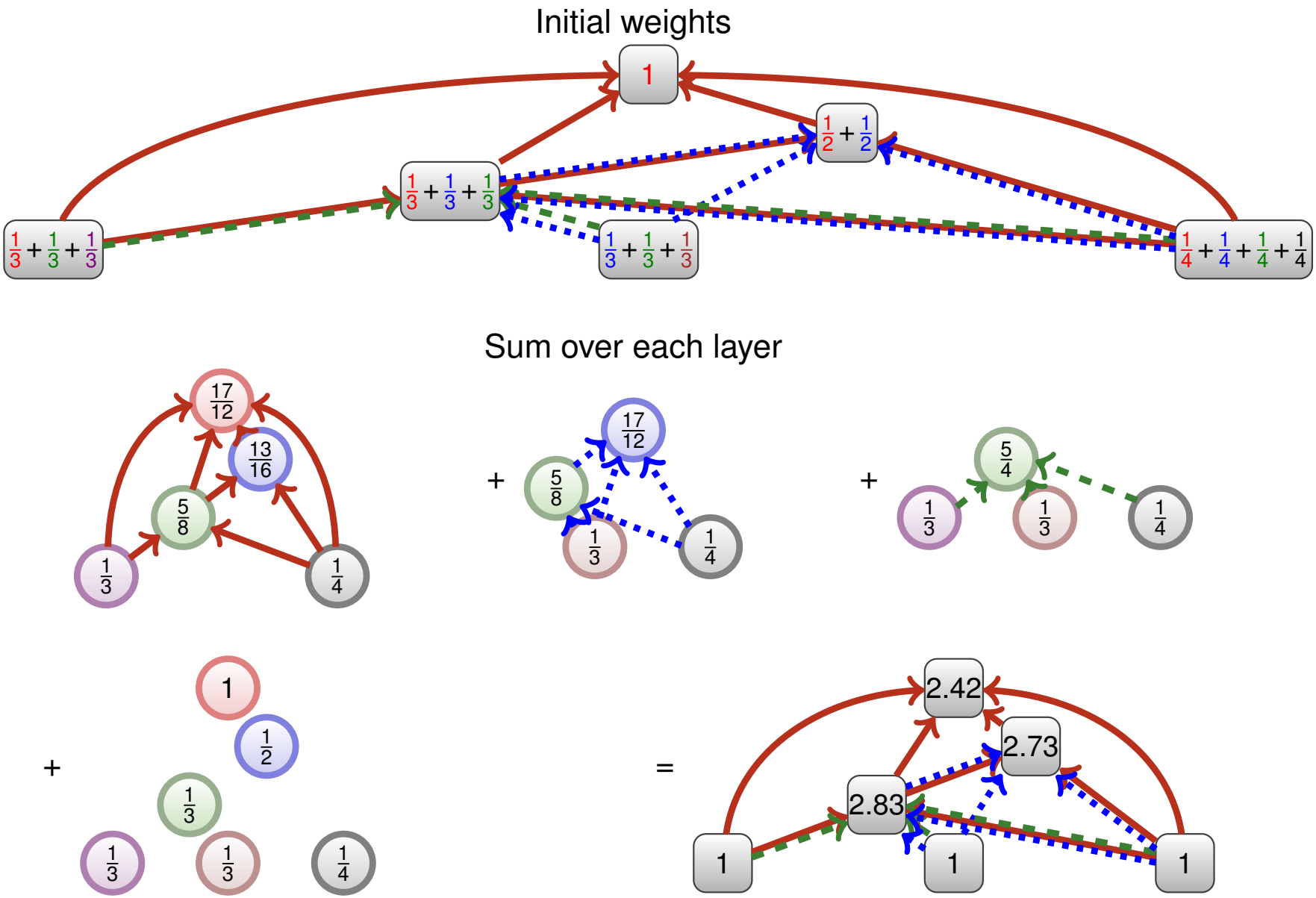
The sum flow will simply sum ascending flows over all subgraphs composed of one layer, determined for a node p by :

$$F_{\mathcal{G}}(p) = \sum_{l}^{l\in\mathcal{L}(p)} F_{G_l}(\mathcal{N}^-(p)) \tag{5}$$

$$\lambda_{(p,l)} = \frac{\lambda_p}{|\mathcal{L}(p)|}$$

If we set the parameter $\lambda_{(p,l)} = 1$, the contribution of one publication to the whole system will be exactly its number of citations, and a publication that cites lots of work will produce a lot of flow. In order to maintain constant the unit of contribution of a publication of a work, we set $\lambda_p = 1$ hence $\sum_l^{l\in\mathcal{L}(p)} \frac{1}{|\mathcal{L}(p)|} = 1$ such as a publication brings exactly 1 unit of contribution to the system.

Sum Flow


Initial weights
1
1/2 + 1/2
1/3 + 1/3 + 1/3
1/3 + 1/3 + 1/3
1/3 + 1/3 + 1/3
1/4 + 1/4 + 1/4 + 1/4
Sum over each layer
17/12
13/16
5/8
1/3
1/4
+
17/12
5/8
1/3
1/4
+
5/4
1/3
1/3
1/4
+
1
1/2
1/3
1/3
1/3
1/4
=
2.42
2.73
2.83
1
1
1

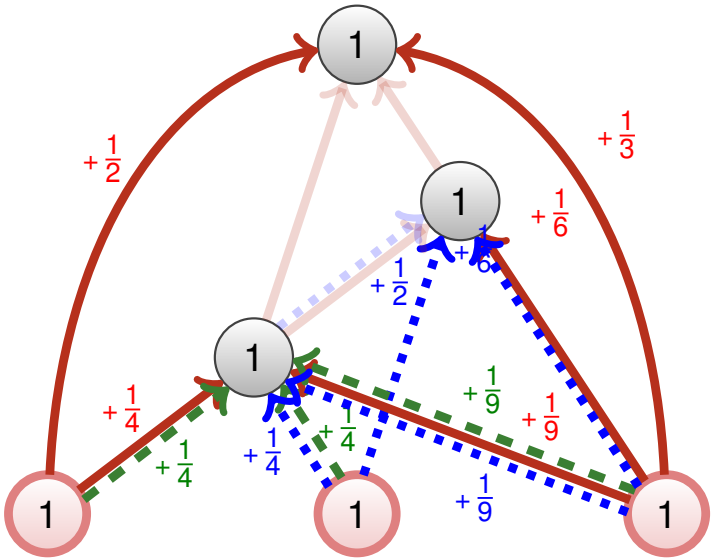
Sum Flow

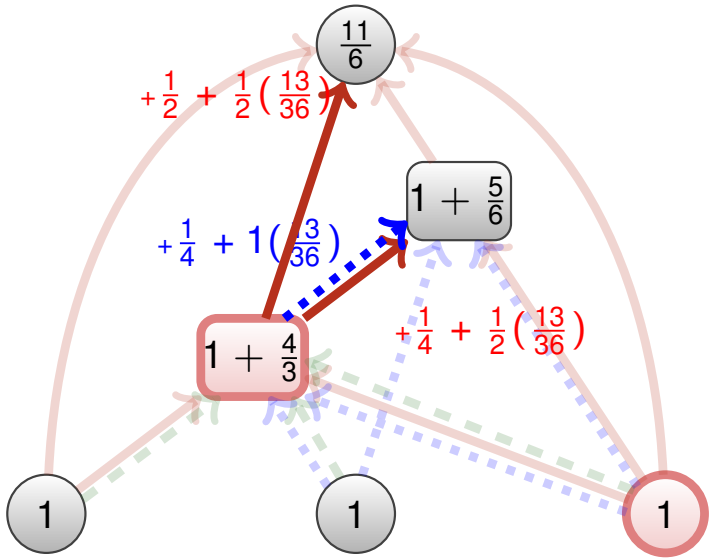
ALGORITHM 3: Sum flow (multiplex)

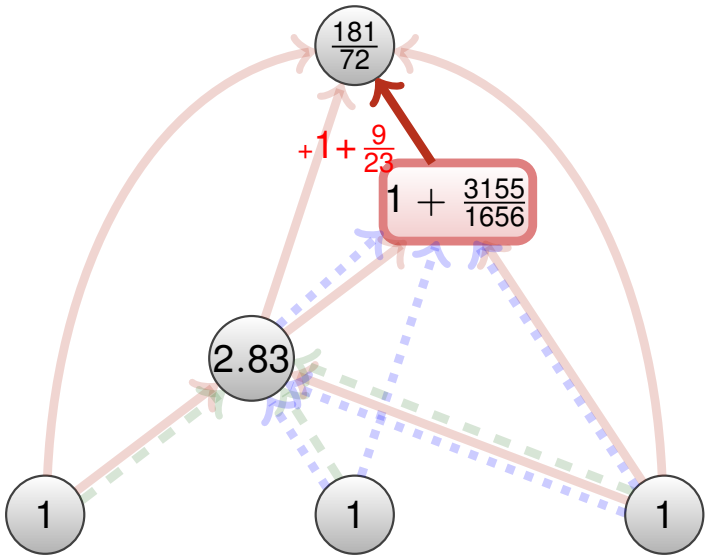
input : A citation network with nodes (articles) and arcs (citations)
output: The sum flow on each node (article) and each arc (citation)

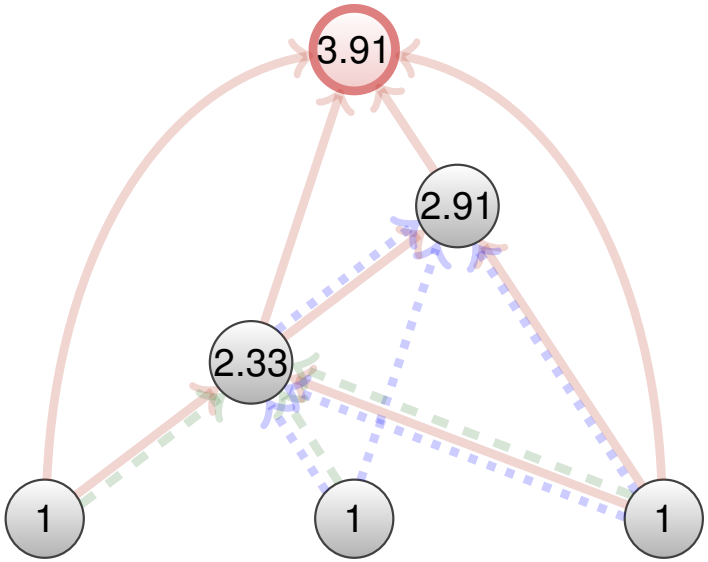
Initialize each article p with a flow value $\lambda_p = 0$
for each article p **do**
  Construct the citation subgraph G_p
  $\mathcal{F} \leftarrow$ ascending-flow(G_p) with $\forall q \in V_p, \lambda_{init}(q) = \frac{1}{d^+(q)+1}$
  for each article $q \in V_p$ **do**
    $\lambda_q \leftarrow \lambda_q + \mathcal{F}(q)$
  end
end

Selective Flow









Selective Flow

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

Initialize each article p with the color **white** and the value $\lambda_p = 1$
Color all leaves in **red** and push them into Q
while Q **is not empty do**
    $q \leftarrow pop\_first(Q)$
    for each r **son of** q **do**
        if r **is white then**
            Construct G_r and initialize all $\lambda(a(q, r, l))$ to 0
            $push\_last(Q, r)$
            Color r in **red**
        end
    end
    for each layer $l \in L_q$ **do**
        Initialize $\lambda_{l,q} = 0$
        for each parent p **of** q **do**
            $\lambda_{l,q} \leftarrow \lambda_{l,q} + \lambda(a(p, q, l))$
        end
        for each son r **of** q **do**
            $\lambda(a(p, q, l)) \leftarrow \frac{1}{d^+(q) \times (1+|\mathcal{L}(q,r)|)} + \frac{\lambda_{l,q}}{d_l^+(q)}$
        end
        $\lambda_q \leftarrow \lambda_q + \lambda_{l,q}$
    end
end

Selective Flow

$$F_{\mathcal{G}_l}(\mathcal{N}^-(p)) = \lambda_{(p,l)} + \sum_{q}^{q\in\mathcal{N}^-(p)} \frac{K_{\mathcal{G}_l}(k_q)}{d^+_{G_l}(k_q)}$$

$$F_{\mathcal{G}}(p) = \sum_{l}^{l\in\mathcal{L}(p)} F_{\mathcal{G}_l}(\mathcal{N}^-(p)) \tag{4}$$

$$\alpha_{(p,l)} = \sum_{j}^{j\in\mathcal{N}^+(p)} \frac{\lambda_p}{d^+(p) \times |\mathcal{L}(p,j)|}$$

Table of contents

1 Introduction

2 Information Networks

3 Flow of knowledge

4 Multiplex Networks

5 Cycles, connected component and ethics

6 Conclusion

What ethics in influence measures

We can define some criteria to guarantee an ethic that we hope flows naturally

- **Equity** between articles
- Metrics with bounded values in pratice
- A network (or sub-network) with **consistent** content : for instance articles from a same field or institute.

These criteria are put to the test by the mere presence of **cycles** within the network. That is, the existence of a couple of nodes (items here) that are descended from each other.
We say that a sub-network is a (strongly) connected component if every vertex is reachable from every other vertex.

A costly solution, decycling (with drawings)

A costly solution, decycling (with drawings)

A costly solution, decycling (with drawings)

Solving the time/accuracy ratio in Information Networks

Conclusion

Solutions to the Big Data Science problems are inherently cross-domain.

(Jean-françois BAFFIER *— 2018)*