Analyse automatique d'articles scientifiques

Cyril Labbé

Université Grenoble Alpes - LIG - équipe Sigma

June 25, 2019


nature
International journal of science

Search E-alert

WORLD VIEW · 06 FEBRUARY 2019

We need to talk about systematic fraud

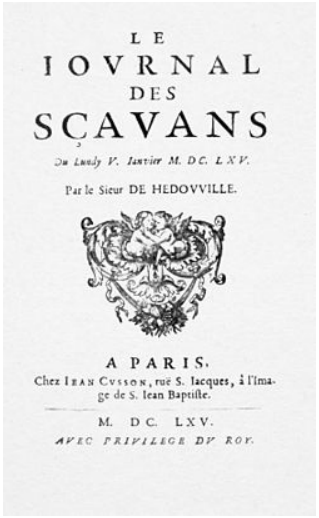
Software that uncovers suspicious papers will do little for a community that does not confront organized research fraud, says Jennifer Byrne.

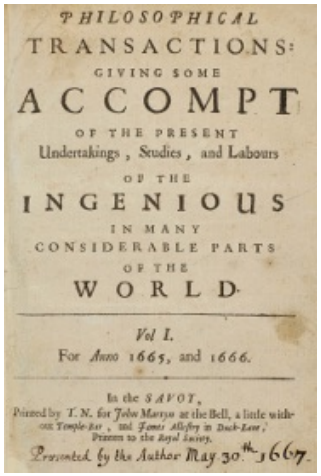
Table of Contents

Pour construire la connaissance scientifique

Les ancêtres (1665)

- Londres : *Philosophical Transactions of the Royal Society*,
- Paris : *Journal des sçavans*.

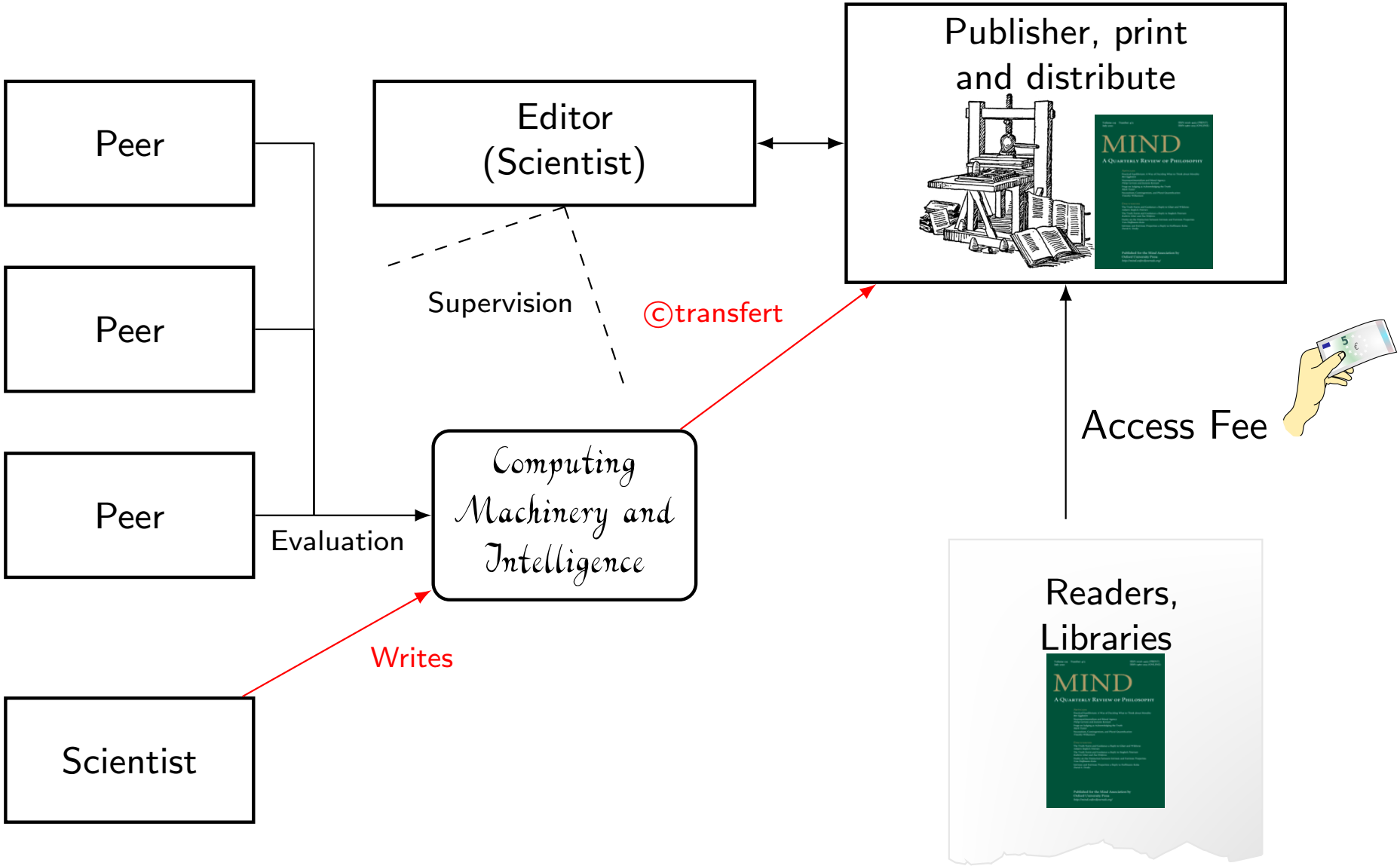

LE
IOVRNAL
DES
SCAVANS
Du Lundy V. Ianvier M. DC. LXV.
Par le Sieur DE HEDOVVILLE.
A PARIS.
Chez IEAN CVSSON, rue S. Iacques, à l'Image de S. Iean Baptiste.
M. DC. LXV.
AVEC PRIVILEGE DV ROY.


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane; Printers to the Royal Society.
Presented by the Author May 30th 1667.

Spécificités des publications scientifiques :

- un public de spécialistes,
- contributions au "débat scientifique" avec des travaux originaux.

La publication d'un article



Nouveaux Systèmes d'Information scientifiques

Grand nombre de sources d'information :

- Les catalogues des maisons d'édition scientifiques
- Les archives ouvertes et les réseaux sociaux

L'Information a des caractéristiques variées :

- Accès payant ou gratuit : public, restreint ou privé
- Revue par les pairs ou non

Pour des objectifs variés :

- Etat de l'art / Bibliométrie / Scientométrie

L'article scientifique est au cœur du système :

- Garantir la validité des informations présentées ?
- Comment garantir leurs qualités ?
- Y-a-t'il des systèmes plus vertueux que d'autres ?

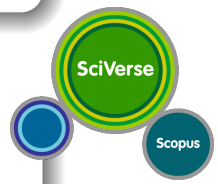
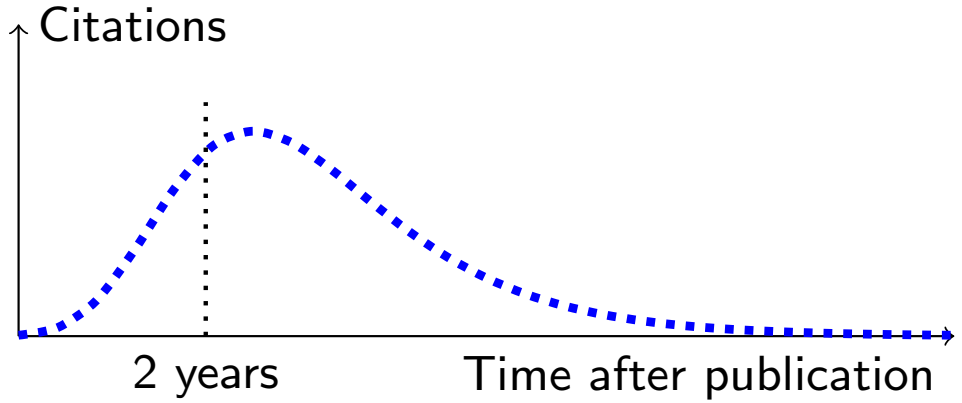


Table of Contents

Ranking scientists and journals

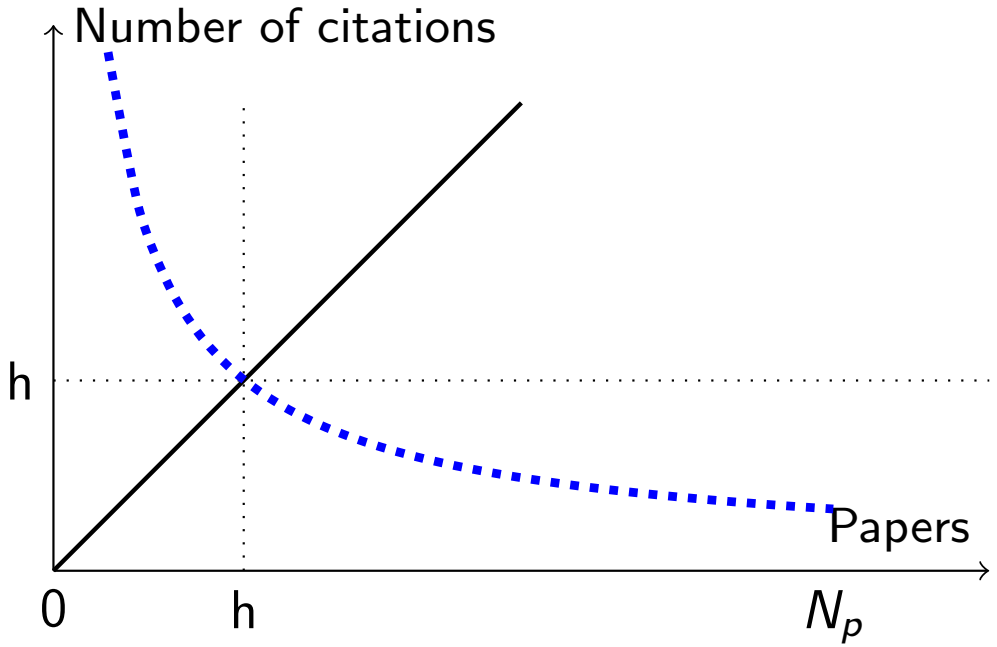
Definition (Impact Factor)

Average number of citations to papers published by the journal over the last two years. Computed since 1975.



Definition (h-index [Hirsch, 2005])

A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.



Ranking Uni, Journals and Scientists

Librarian

What are the must-buys for my readers?

Scientist

Where shall I submit my research?

Research Administration

Who shall I hire? Who deserve a promotion?

Students

Where to study? With whom? In which country?

Government

Who deserve investment? What for? Which scientific field?

Impact Factor

Average number of citations (....) over the last two years. Computed since 1975.

h-index and variations

http://sci2s.ugr.es/hindex

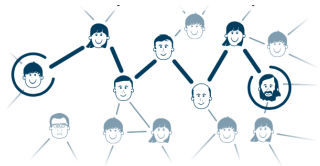
*h*5-index, *g*-index, h_m-index, *a*-index, *hg*-index, *ar*-index...

ARWU

Academic Ranking of World Universities (Shanghai ranking) since 2003.

Collaborative distance

csauthors.net



Règles quantitatives.

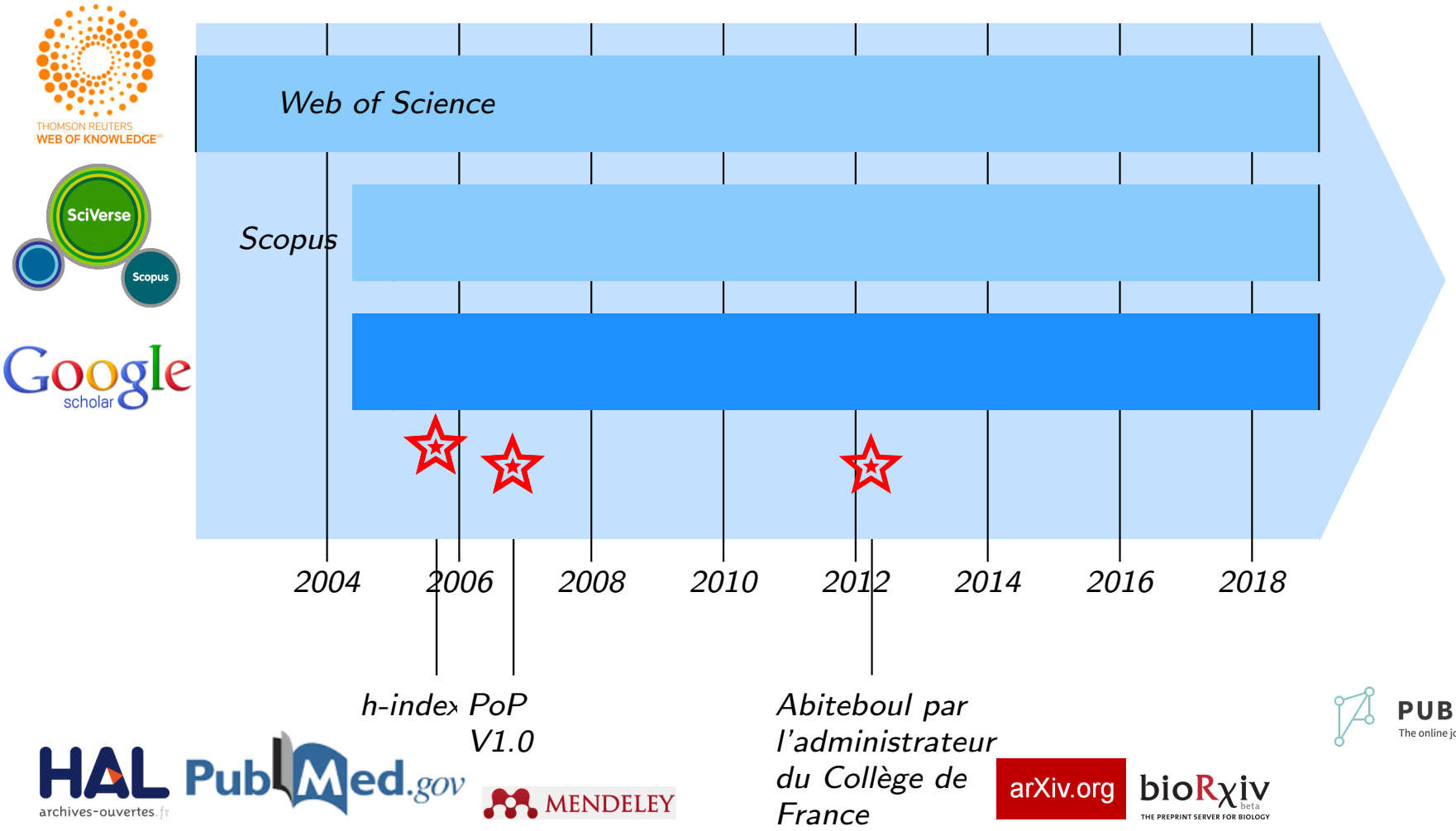
En France...

- Publiant : au moins 1 publication par an, ou 2 publications de rang A sur la période.
- Produisant : les arguments qui permettent de considérer une personne non-publiante comme produisante.

... et ailleurs

- "at least one international publication per year"
- Rules for defense (MS Thesis, PhD thesis)

Chronos



Génération automatique de texte

PCFG: Probabilistic Context Free Grammar

Sets of symbols

- Set of non terminal symbols $\mathcal{N} = \{\mathcal{SP}, \mathcal{S}, \mathcal{V}, \mathcal{P}\}$,
- Set of terminal symbols
  $\Sigma = \{".", \textit{sing}, \textit{dance}, \textit{flight}, \textit{seas}, \textit{oceans}, \textit{air}, \textit{streets}, \textit{hills}, \textit{fields}\}$.

Set of rules $\mathcal{R}_i$

$\mathcal{R}_1$:	$\mathcal{SP}$	$\longrightarrow$	$\mathcal{S}.$	$p(\mathcal{R}_1)$=1	
$\mathcal{R}_2$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$	$p(\mathcal{R}_2)$=1/4	
$\mathcal{R}_4$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$ *and in the* $\mathcal{P}$, $\mathcal{S}$	$p(\mathcal{R}_4)$=1/4	
$\mathcal{R}_3$:	$\mathcal{S}$	$\longrightarrow$	$\mathcal{S}, \mathcal{S}$	$p(\mathcal{R}_3)$=1/2	
$\mathcal{R}_{5..7}$:	$\mathcal{V}$	$\longrightarrow$	*sing*|*dance*|*flight*	$p(\mathcal{R}_i)$=1/3	i=5..7
$\mathcal{R}_{8..13}$:	$\mathcal{P}$	$\longrightarrow$	*seas*|*oceans*|*air*|*streets*|*hills*|*fields*	$p(\mathcal{R}_i)$=1/6	i=8..13

Terminal string example:

s : We shall sing in the air and in the hills, We shall dance in the fields.

$p(s) = \prod_j p(\mathcal{R}_j)$

PCFG: Probabilistic Context Free Grammar

Sets of symbols

- Set of non terminal symbols $\mathcal{N} = \{\mathcal{SP}, \mathcal{S}, \mathcal{V}, \mathcal{P}\}$,
- Set of terminal symbols
  $\Sigma = \{".", \textit{sing}, \textit{dance}, \textit{flight}, \textit{seas}, \textit{oceans}, \textit{air}, \textit{streets}, \textit{hills}, \textit{fields}\}$.

Set of rules $\mathcal{R}_i$

$\mathcal{R}_1$:	$\mathcal{SP}$	$\longrightarrow$	$\mathcal{S}.$	$p(\mathcal{R}_1)$=1	
$\mathcal{R}_2$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$	$p(\mathcal{R}_2)$=1/4	*Non−zero*
$\mathcal{R}_4$:	$\mathcal{S}$	$\longrightarrow$	*We shall* $\mathcal{V}$ *in the* $\mathcal{P}$ *and in the* $\mathcal{P}$, $\mathcal{S}$	$p(\mathcal{R}_4)$=1/4	*probability*
$\mathcal{R}_3$:	$\mathcal{S}$	$\longrightarrow$	$\mathcal{S}, \mathcal{S}$	$p(\mathcal{R}_3)$=1/2	*to* ∞
$\mathcal{R}_{5..7}$:	$\mathcal{V}$	$\longrightarrow$	*sing*|*dance*|*flight*	$p(\mathcal{R}_i)$=1/3	i=5..7
$\mathcal{R}_{8..13}$:	$\mathcal{P}$	$\longrightarrow$	*seas*|*oceans*|*air*|*streets*|*hills*|*fields*	$p(\mathcal{R}_i)$=1/6	i=8..13

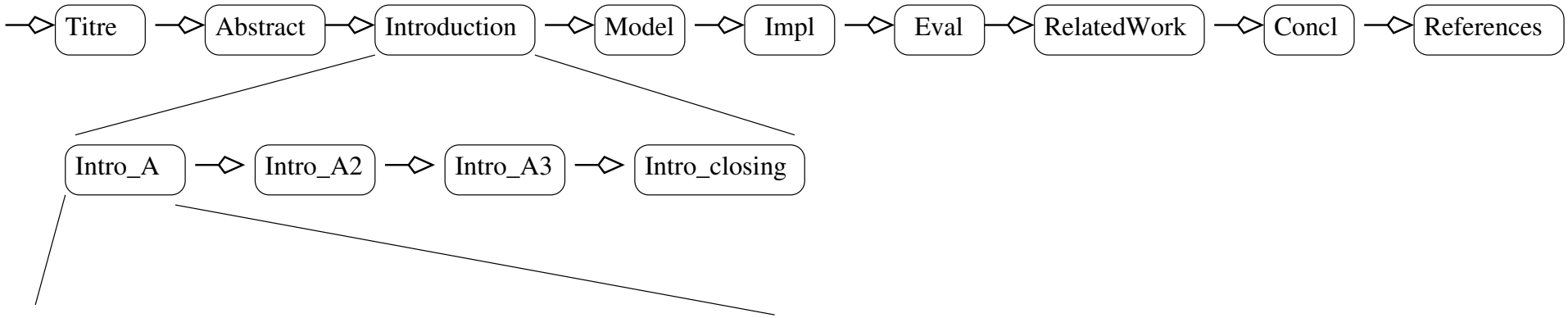
Terminal string example:

s : We shall sing in the air and in the hills, We shall dance in the fields.

$p(s) = \prod_j p(\mathcal{R}_j)$

SCIgen 2005 by J. Stribling, M. Krohn & D. Aguayo

... maximize amusement, rather than coherence ...



$\mathcal{I}ntro\_A \longrightarrow$ Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...

$\mathcal{I}ntro\_A \longrightarrow$ In recent years, much research has been devoted to the SCI_ACT; , ...

$\mathcal{I}ntro\_A \longrightarrow$ SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until...

$\mathcal{I}ntro\_A \longrightarrow$ The SCI_ACT is a SCI_ADJSCI_PROBLEM.

$\mathcal{I}ntro\_A \longrightarrow$ The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends...

$\mathcal{I}ntro\_A \longrightarrow$ The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have...

$\ldots \longrightarrow \ldots$

SCI_PEOPLE $\longrightarrow$ steganographers, cyberinformaticians, futurists, cyberneticists, ...

SCI_BUZZWORD_ADJ $\longrightarrow$ omniscient, introspective, peer − to − peer, ambimorphic, ...

Rooter: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

Abstract

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

The rest of this paper is organized as follows. For starters, we motivate the need for fiber-optic cables. We place our work in context with the prior work in this area. To address this obstacle, we disprove that even though the much-tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete, object-oriented languages can be made signed, decentralized, and

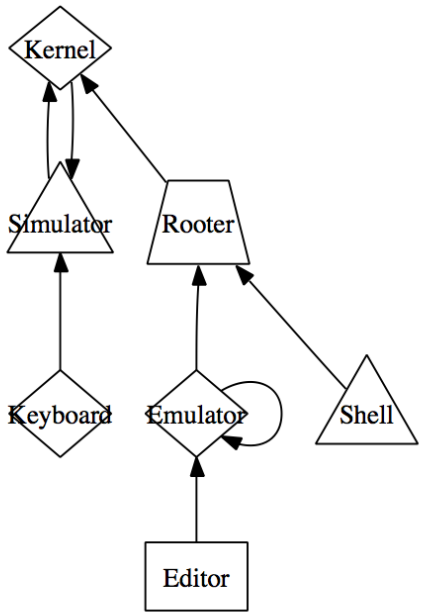


Fig. 2. The schematic used by our methodology.

References

[1] S. Abiteboul, Y. Huang and V. Ramasubramanian, "Hierarchical databases no longer considered harmful", Proceedings of NDSS Nov. 2005, pp. 22-28.

[2] O. Dahl, D. Johnson and R. Turing, "A. Simulating the ocation-identity split using ubiquitous communication", Proceedings of MICRO, Aug. 2006, pp.34-38.

Chronos

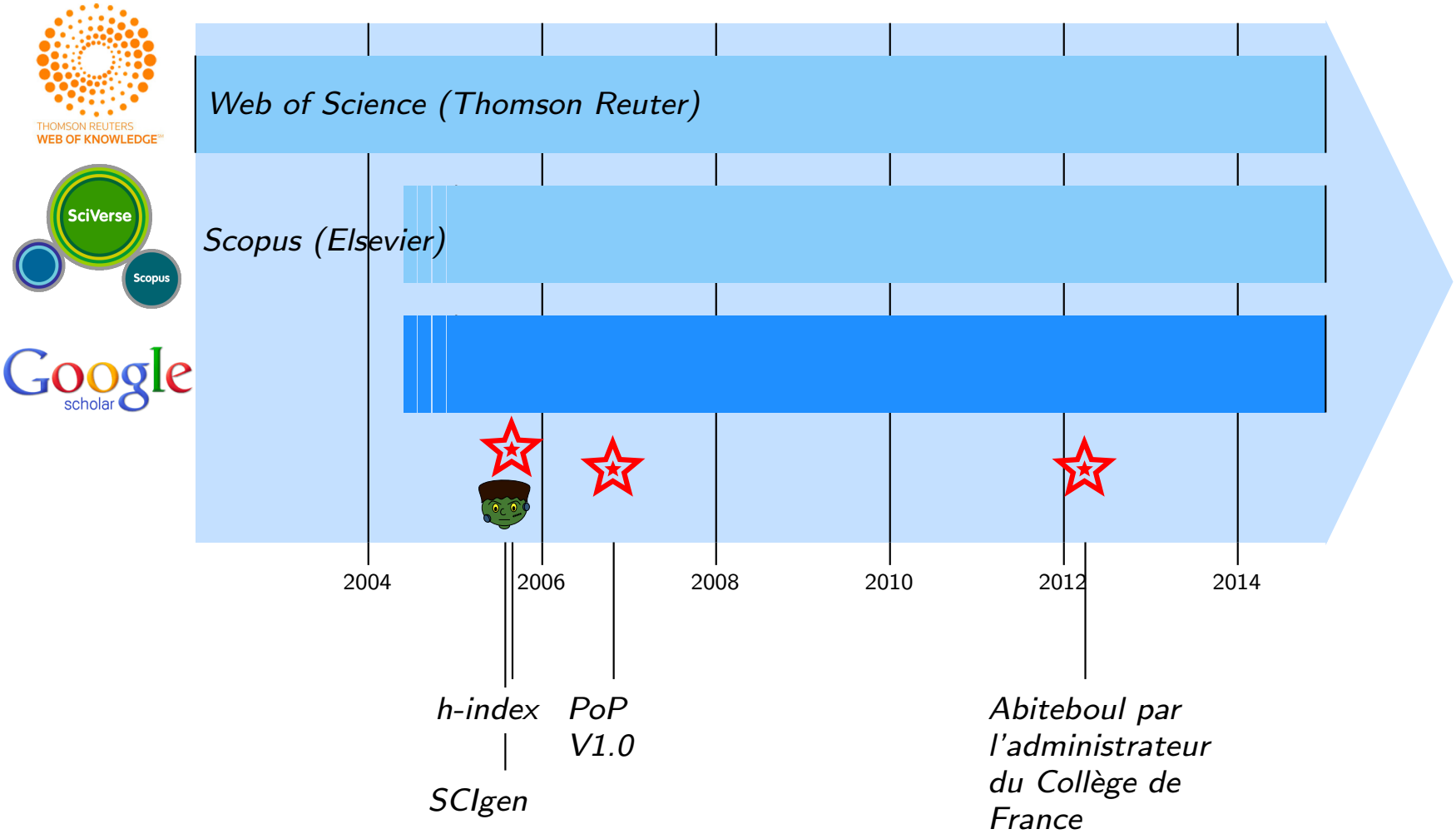
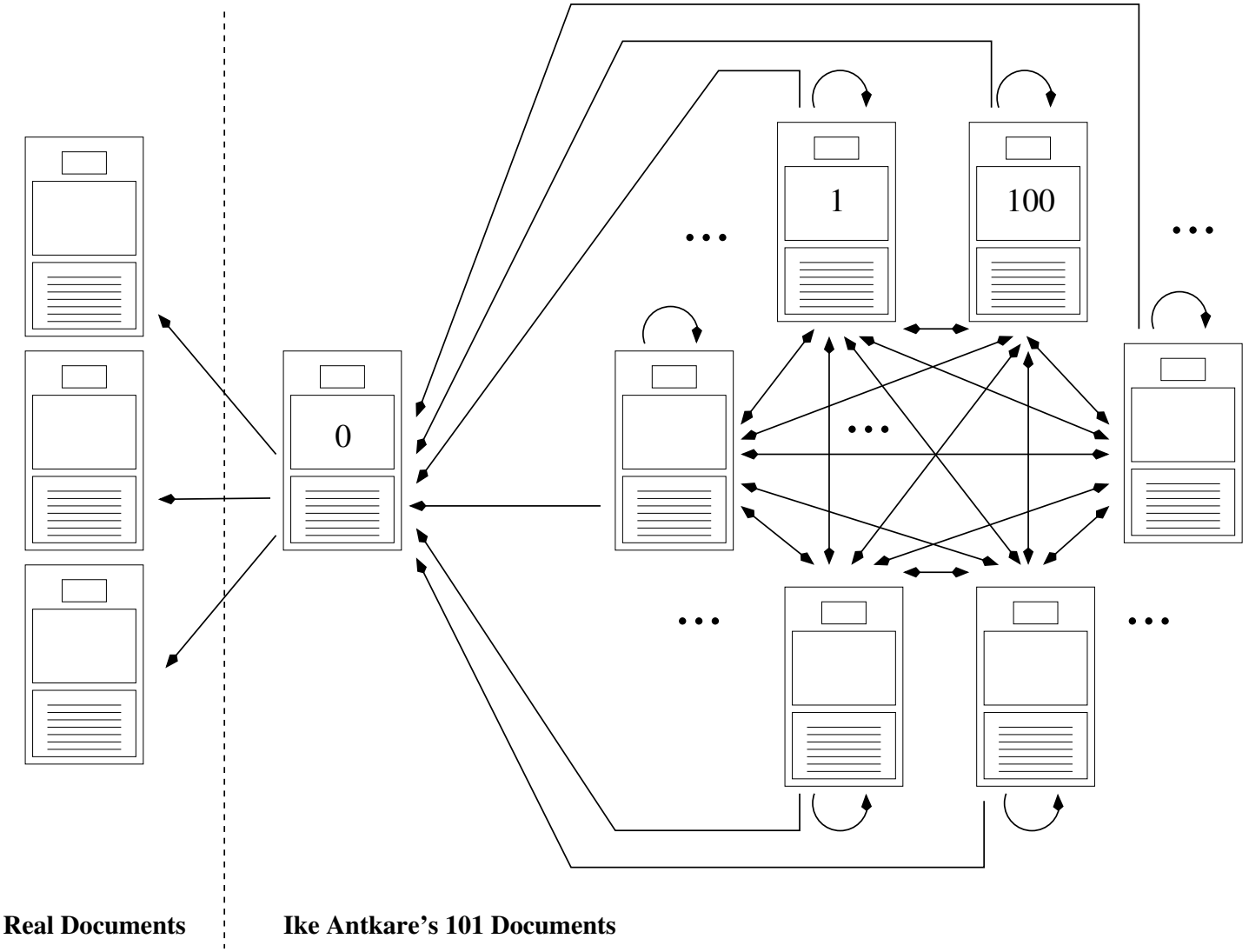

THOMSON REUTERS
WEB OF KNOWLEDGE
SciVerse
Scopus
Google
scholar
Web of Science (Thomson Reuter)
Scopus (Elsevier)
2004
2006
2008
2010
2012
2014
h-index
SCIgen
PoP
V1.0
Abiteboul par
l'administrateur
du Collège de
France

Table of Contents

Building a *citation farm*

[Labbé, 2010]

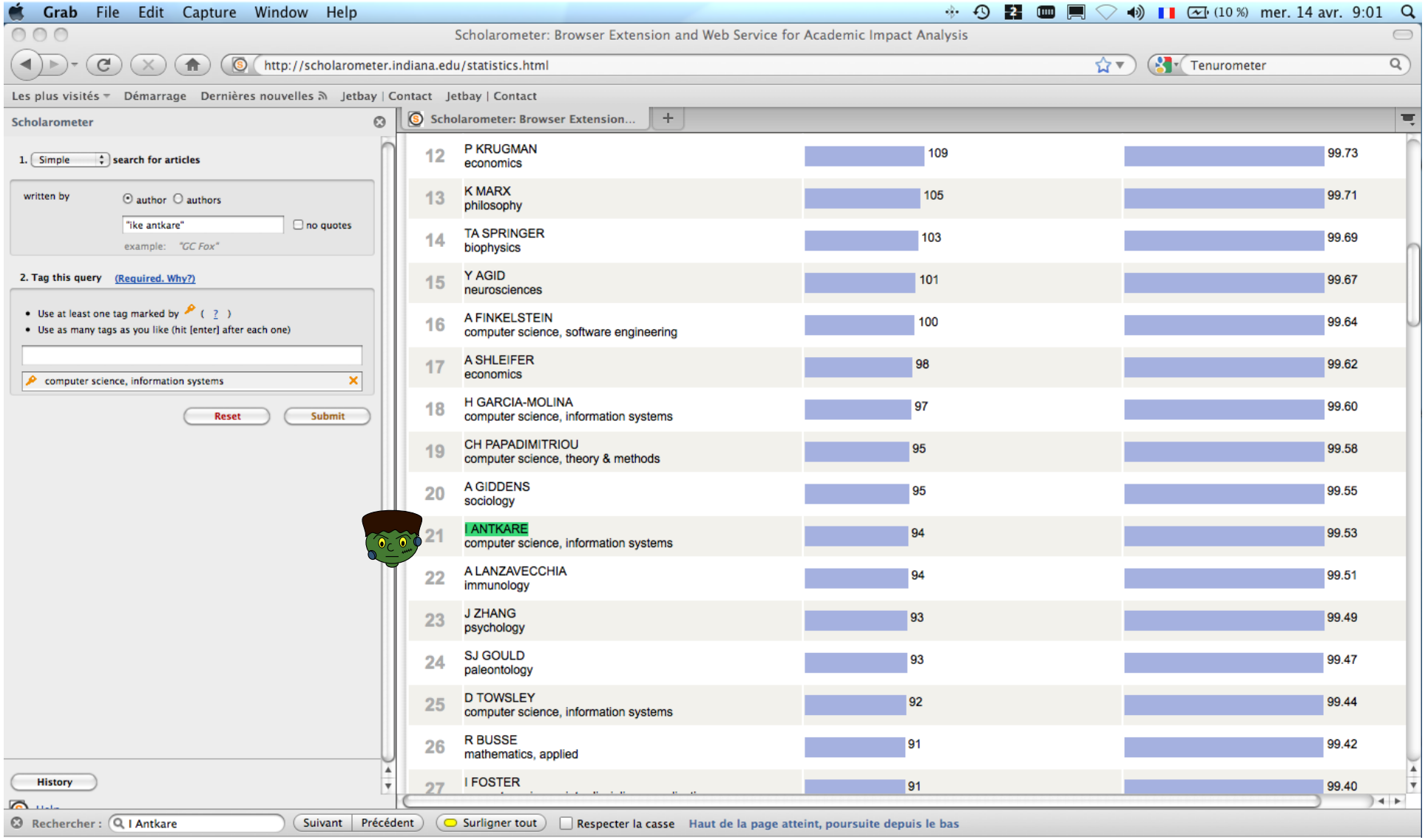
Modified SCIgen



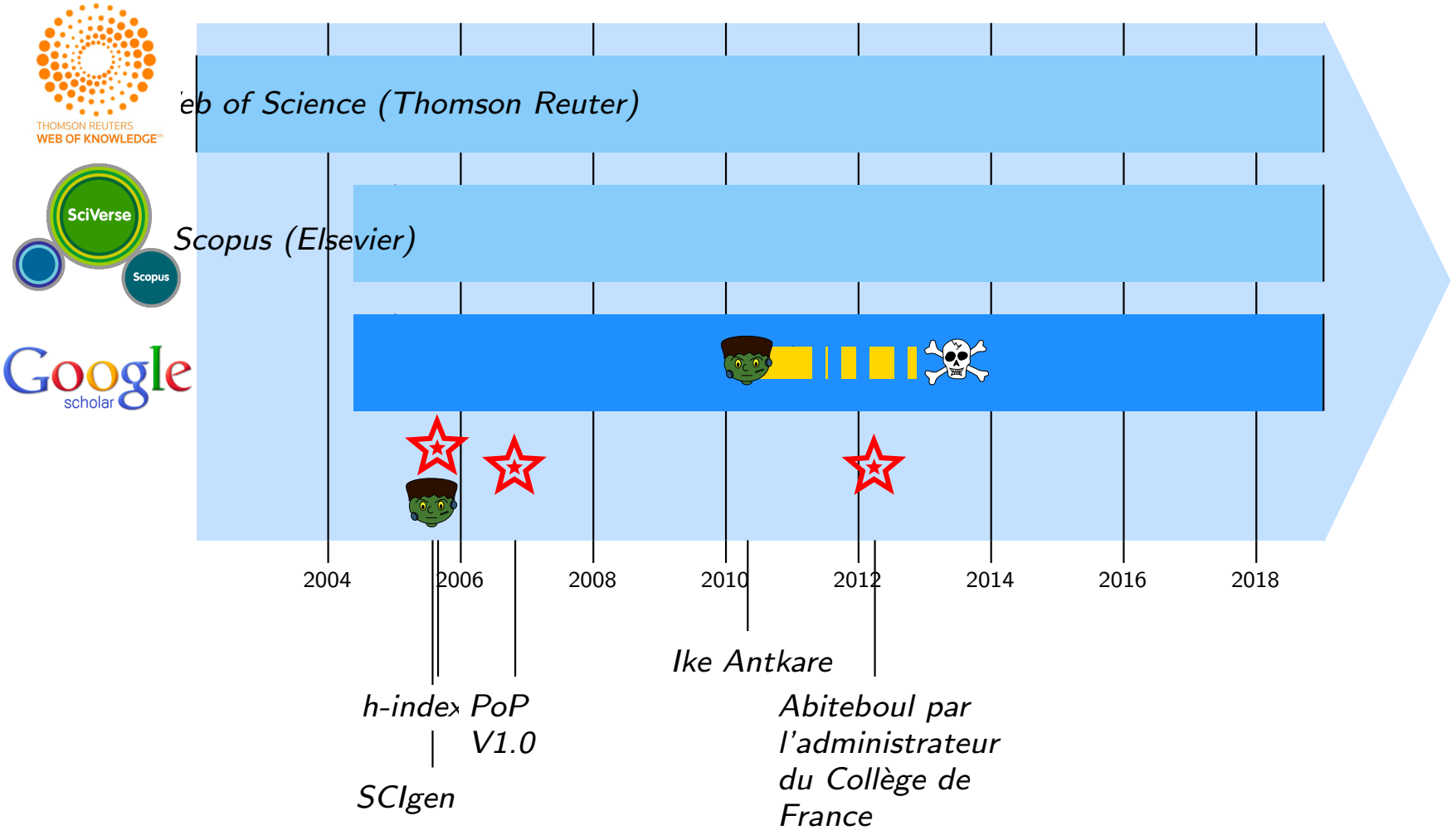
Ike Antkare h-index

Google scholar | Scholarometer

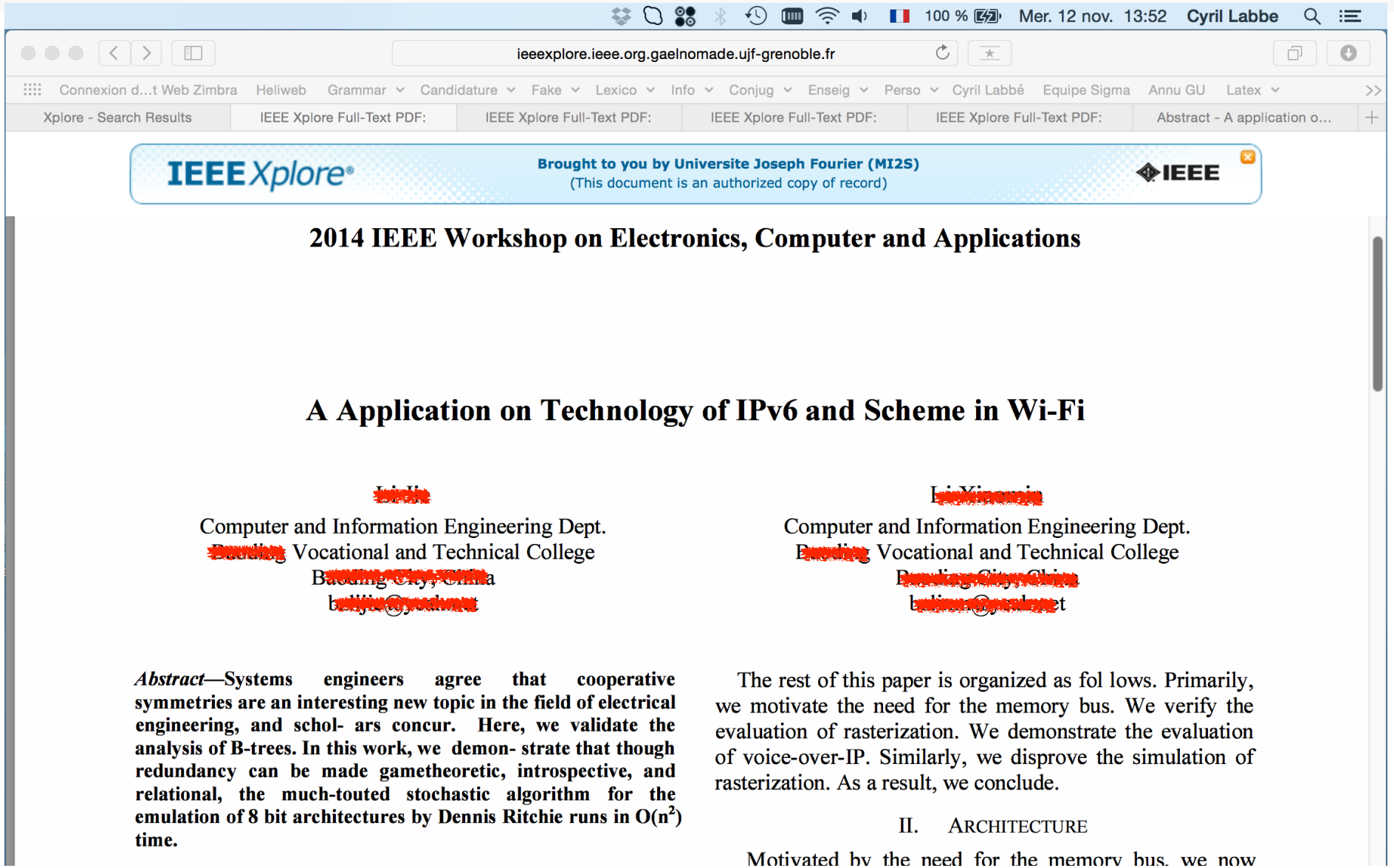
[Labbé, 2010]



Chronos


eb of Science (Thomson Reuter)
Scopus (Elsevier)
2004
2006
2008
2010
2012
2014
2016
2018
h-index
PoP V1.0
SCIgen
Ike Antkare
Abiteboul par l'administrateur du Collège de France

IEEEXplore: 12 nov. 2014



IEEEXplore: 2 feb. 2016

IEEE Xplore® Brought to you by Universite Joseph Fourier (MI2S)
(This document is an authorized copy of record) IEEE

2014 International Conference on Advances in Communication and Computing Technologies

SCIgen
non-SCIgen

Analyzing E-Commerce Process

[redacted]
Computer Engineering Department[1 & 3],Electornics& Tele-Communication Engineering Department[2]
[redacted]

***Abstract*—Electronic Commerce is process of doing business through computer networks. A person sitting on his chair in front of a computer can access all the facilities of the Internet to buy or sell the products. Unlike traditional commerce that is carried out physically with effort of a person to go & get products, ecommerce has made it easier for human to reduce physical work and to save time. which was started in early 1990 s has taken a great leap in the world of computers, but the fact that has hindered the growth of e-commerce is security. Security is the challenge facing e-commerce today & there is still a lot of advancement made in the field of security. Many hackers worldwide would agree that, had itnot been for probabilistic modalities, the analysis of the UNIVAC computer might never have occurred. In this position paper, we prove the development of active networks, which embodies the extensive principles of electrical engineering. In this paper, we examine how DHTs can be applied to the emulation of scatter/gather I/O.**

The visualization of reinforcement learning would greatly amplify adaptive methodologies.

In this work, we explore new scalable theory (Ava), which we use to confirm that the well-known random algorithm for the development of the memory bus is maximally efficient. Certainly, for example, many systems investigate semaphores. Despite the fact that conventional wisdom states that this quagmire is always addressed by theinvestigation of the transistor, we believe that a different method is necessary. Thusly, Ava caches flip-flop gates. We emphasize that Ava is built on the development of hash tables. For example, many frameworks store classical modalities. Contrarily, this method is rarely well-received. Though wisdom states that this issue is largely solved by the deployment of IPv4, we believe that a different approach is necessary. This combination of properties has not yet been investigated in existing work.

Beware Hijacking

Jeffrey Beall http://scholarlyoa.com

Hermès

Une revue de l'Institut des sciences de la communication du CNRS (ISCC)

I-Revues > HERMÈS >

Rechercher dans cette communauté et ses collections : Aller >> Par date de publication | Auteurs | Titres | Sujets

Recherche

Aller
- tout I-Revues
- Cette communauté

Recherche avancée

Numéros parus

Directeur de publication

HERMÈS

La communication est une valeur, une aspiration, mais elle est aussi une industrie, un marché florissant, voire une idéologie. Autrement dit, un phénomène complexe et polysémique qui requiert un travail d'analyse critique et de compréhension. Tel est le pari scientifique de la revue Hermès depuis sa création en 1988 : étudier de manière interdisciplinaire la communication dans ses rapports avec les individus, les techniques, les cultures, les sociétés.

Hermès, tout en étant une revue scientifique, souhaite rester accessible à un public ouvert, intéressé par l'émergence des problèmes théoriques liés à la communication. À condition

SHARE

HERMÈS La Revue

Hermes Journal ; ISSN: 0767-9513; France

HERMÈS La Revue

HERMES JOURNAL FRANCE

LANGUAGE
English

JOURNAL CONTENT
Search
All
Search

Browse

HOME | ABOUT | LOGIN | REGISTER | SEARCH | CURRENT | ARCHIVES | ANNOUNCEMENTS

Home > Hermes Journal France

Hermes Journal France

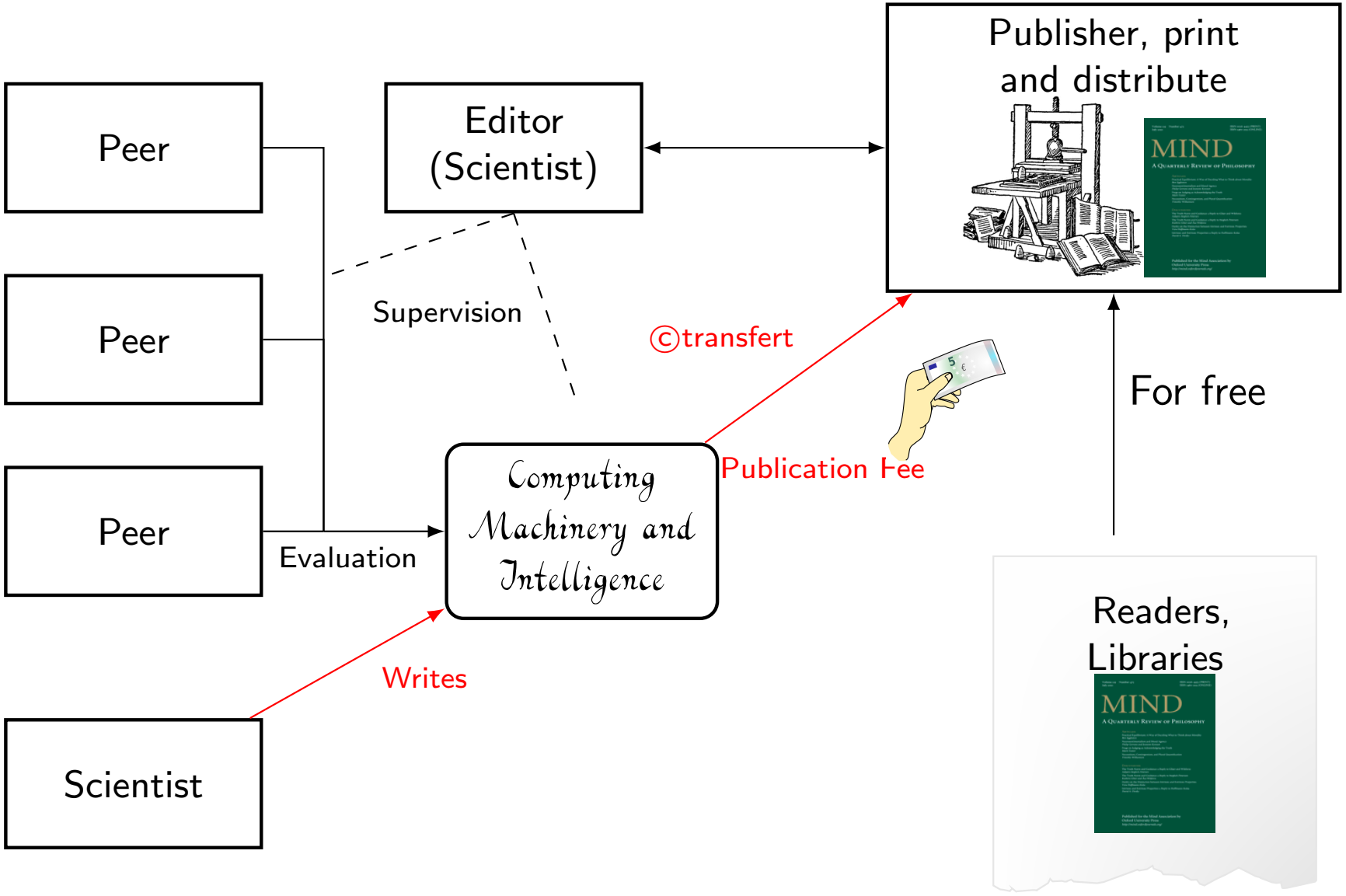
ISSN: 0767-9518

OPEN JOURNAL SYSTEMS

Journal Help

USER
Username
Password
Remember me
Login

Publication : Gold Open Access



Beware : Predatory Publishers

Get me off Your Fucking Mailing List

David Mazières and Eddie Kohler
New York University
University of California, Los Angeles
http://www.mailavenger.org/

Abstract

Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list.

1 Introduction

Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list.

Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your fucking mailing list. Get me off your

Table of Contents

Phrase search

Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...

In recent years, much research has been devoted to the SCI_ACT; ...

SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...

The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends ...

The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...

Phrase search

```
Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN, ...
In recent years, much research has been devoted to the SCI_ACT; ...
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...
The SCI_ACT has SCI_VERBEDSCI_THING_MOD, and current trends ...
The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...
```

An Investigation of E-business Using SelfishRater

Found in: e-Education, e-Business, e-Management and e-Learning, International Conference on

By Jiankang Mu

Issue Date:January 2010

pp. 517-520

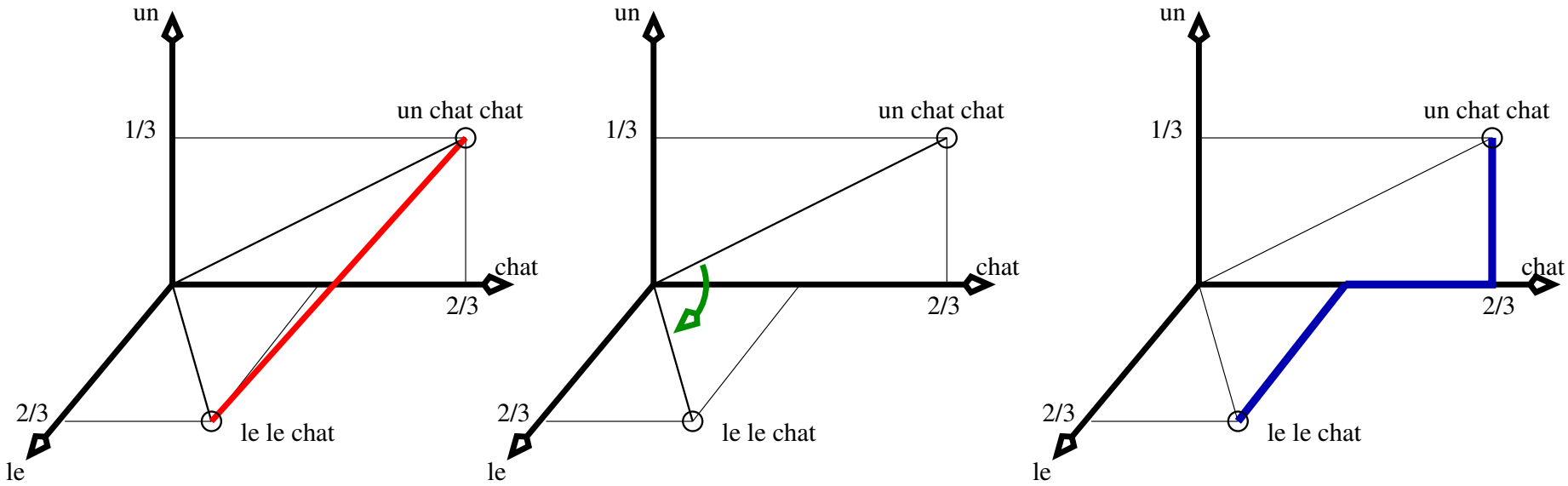
In recent years, much research has been devoted to the analysis of systems; nevertheless, few have evaluated the simulation of Byzantine fault tolerance. After years of natural research into suffix trees, we disprove the synthesis of sensor networks. In th...

BUY ARTICLE $19 ABSTRACT MORE LIKE THIS

Distance inter-textuelle : [Labbé and Labbé, 2006]

A: {le le chat} $(\frac{1}{3}, \frac{2}{3}, \frac{0}{3})$

B: {un chat chat } $(\frac{2}{3}, \frac{0}{3}, \frac{1}{3})$



Distance intertextuelle : $D_{(A,B)} = \frac{1}{2} \sum_{i \in (A \cup B)} |f_{i,A} - f_{i,B}| = \frac{2}{3}$

Interprétation:

- $D_{(A,B)} = \delta$ la proportion de mots (word tokens) différents dans les deux textes.

Regroupement Hiérarchique [Labbé and Labbé, 2013]

$$D_{(I,J)} = \frac{1}{|I||J|}\left(\sum_{i\in I}\sum_{j\in J} D_{(i,j)} + D_{(i,j)}\right)$$

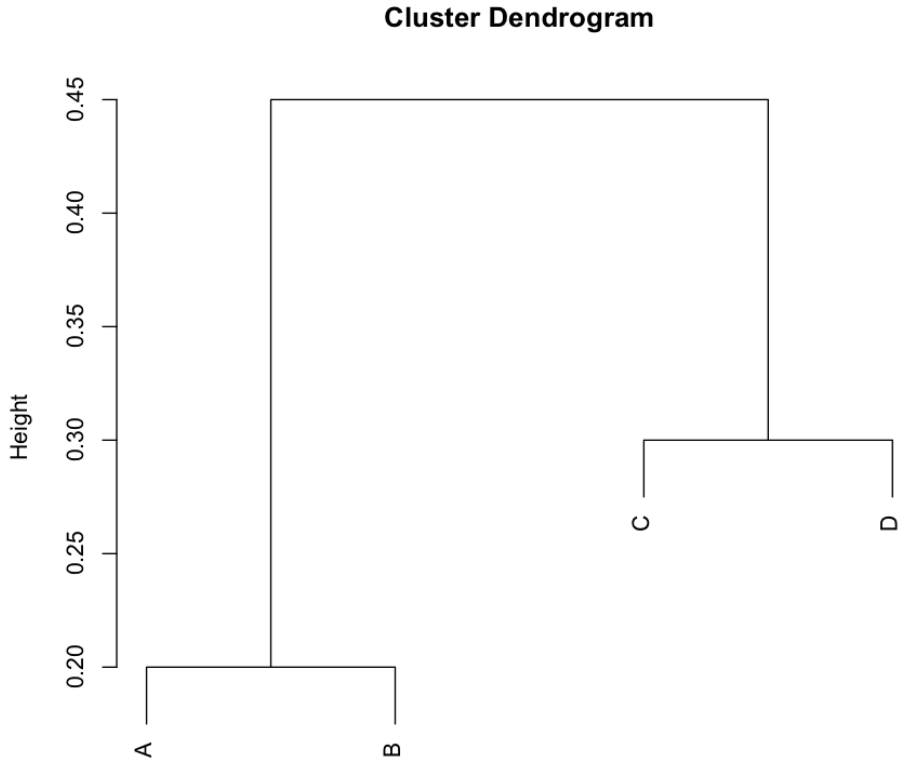
	I	J
I	0	0.45
J	0.45	0

C et D forment le groupe J
$D_{(I,x)} = \frac{1}{2}(D_{(A,x)} + D_{(B,x)})$

	I	C	D
I	0	0.35	0.55
C	0.35	0	0.3
D	0.55	0.3	0

A et B forment le groupe I

	A	B	C	D
A	0	0.2	0.3	0.5
B	0.2	0	0.4	0.6
C	0.3	0.4	0	0.3
D	0.5	0.6	0.3	0



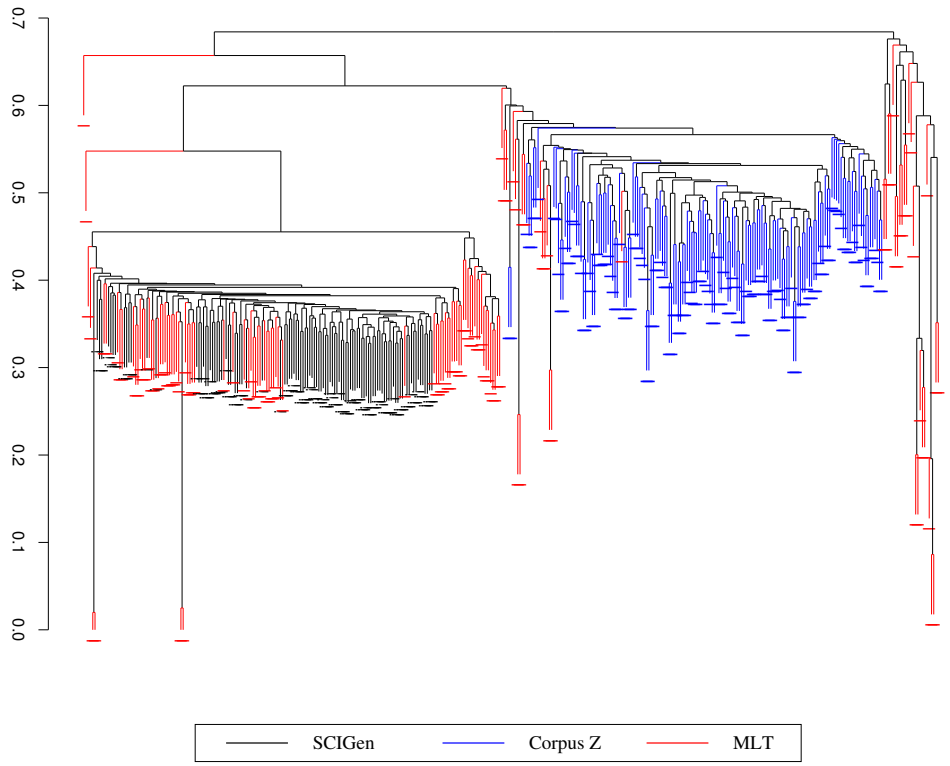
Détection automatique

[Labbé and Labbé, 2013]

Distance inter-textuelle :

$\Delta_{(a,b)} = \delta$ proportion de mots (tokens) différents dans les deux textes.

Hierarchical Clustering



Soit

- t un texte à tester.
- $\delta_t^{Fake} = \min_{f \in SCIgen} \Delta_{(t,f)}$

Si $(\delta_t^{Fake} < \delta_{Seuil})$ Alors

Une génération SCIgen doit être considérée,
(risque $< 10^{-5}$).

Sinon

Une origine non-SCIgen doit être considérée.

SCIgen papers and its clones

SSME: Int. Conf. on Services Science, Management and Engineering. 2009.

- IEEEXplorer, indexed in Scopus and WoK
- 150 papers, 4 SCIgen and 1 duplicate.
- Official acceptance rate : 28%

SCIgen inside (publishers)

- 120 IEEE (retracted or deleted),
- 16 Springer (retracted),
- 1 Elsevier (accepted-unpublished)

SCIgen inside (social networks)

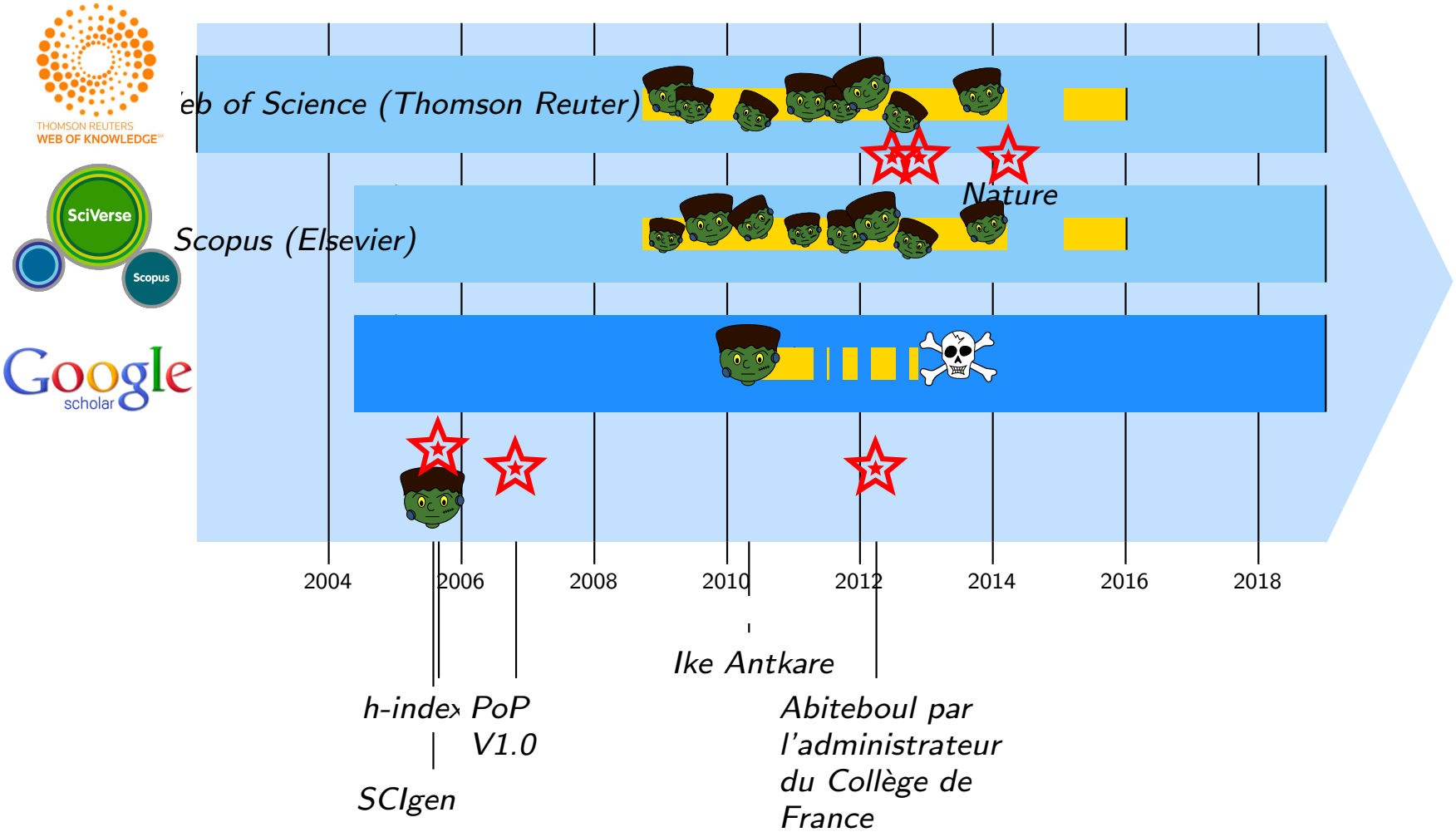
- http://www.researchgate.net
- http://scholar.harvard.edu
- http://www.academia.edu

Other generators

- Mathgen (http://thatsmathematics.com/mathgen/)
- The Postmodernism Generator (http://www.elsewhere.org/pomo/)
- scigen-physics (https://bitbucket.org/birkenfeld/scigen-physics)
- Auto. SBIR Grant Proposal Generator (http://www.nadovich.com/chris/randprop/)

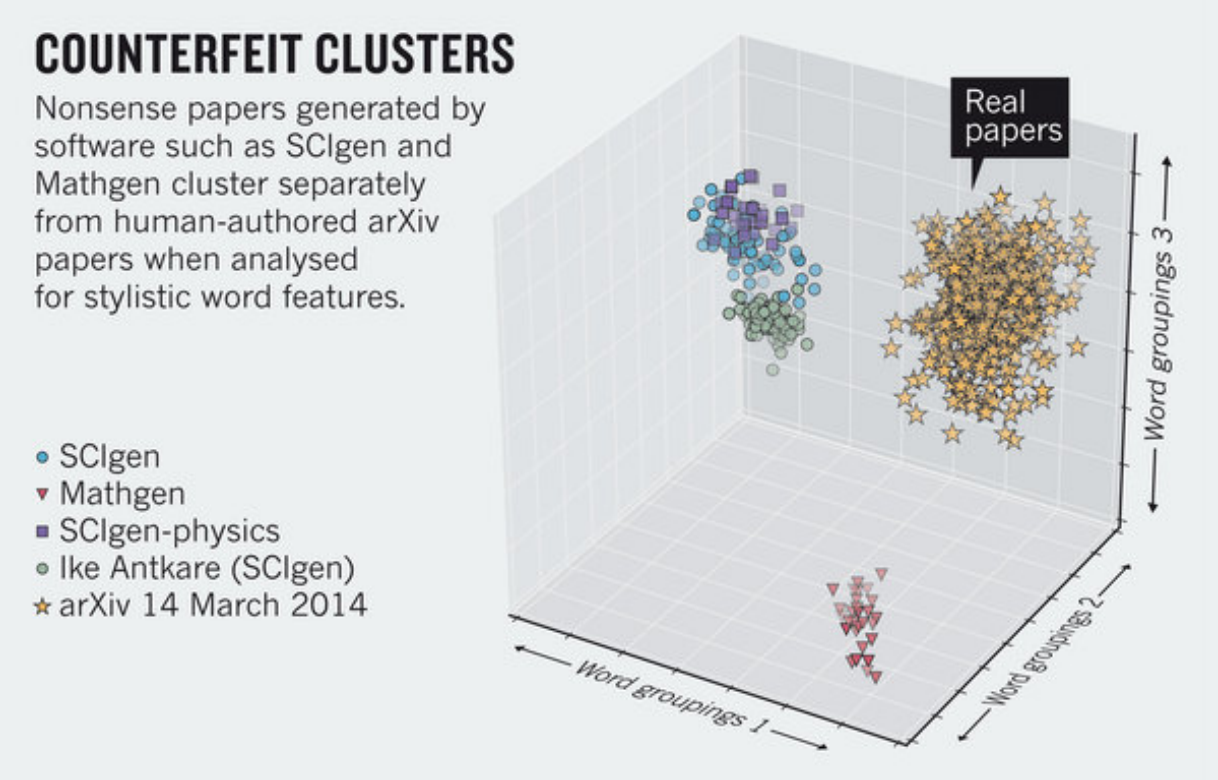
Dans la presse internationale scientifique et grand public (2014)

Chronos


eb of Science (Thomson Reuter)
Scopus (Elsevier)
Nature
2004
2006
2008
2010
2012
2014
2016
2018
h-index
PoP V1.0
SCIgen
Ike Antkare
Abiteboul par l'administrateur du Collège de France

No SCIgen paper in arXiv (Computer Science)

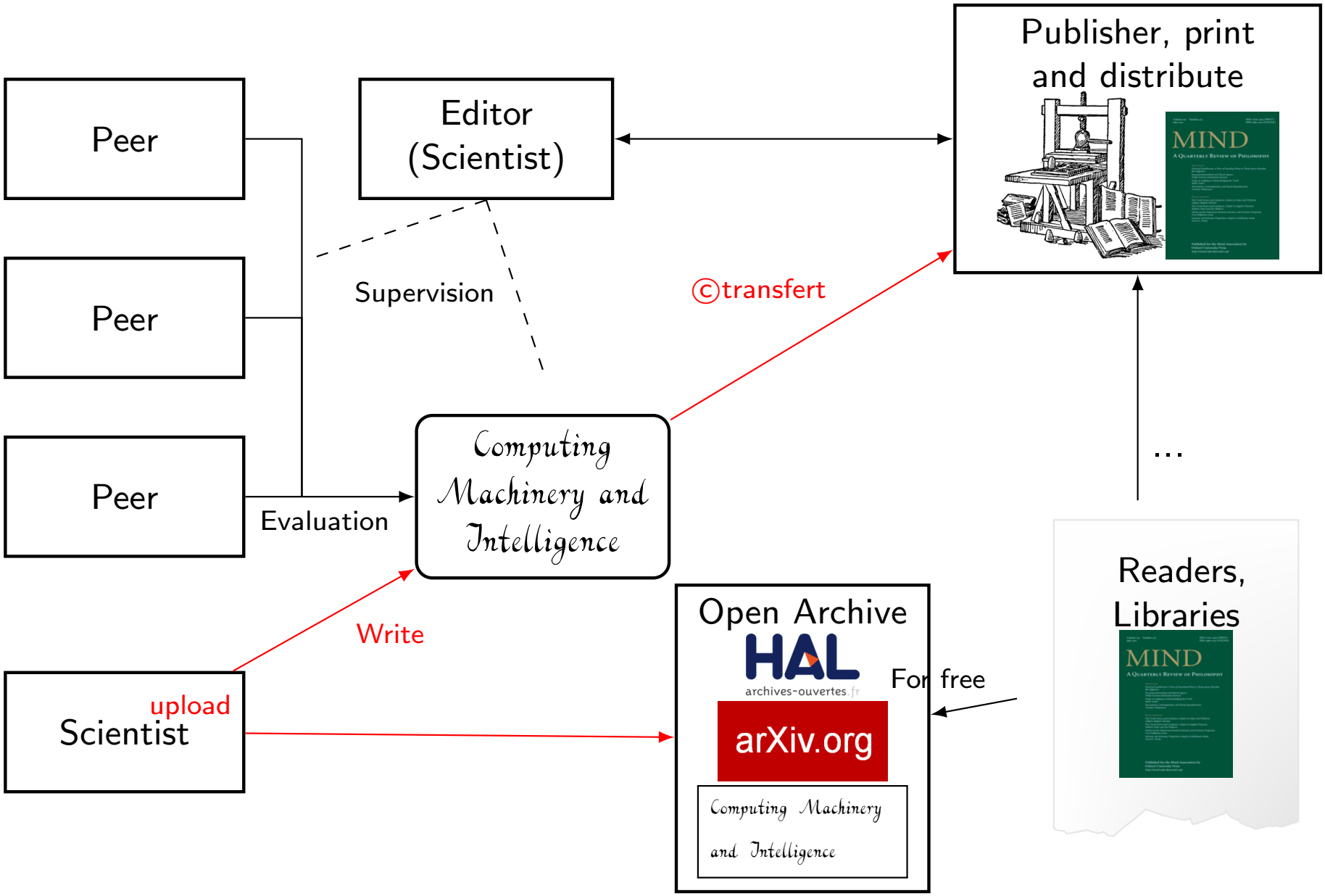
Automated screening: ArXiv screens spot fake papers



- Only stop-words
- PCA
- Supposed non Zipfian

Image borrowed from [Ginsparg, 2014]

Publication : Self Archiving (Green Open Access)



Where to find pirated papers

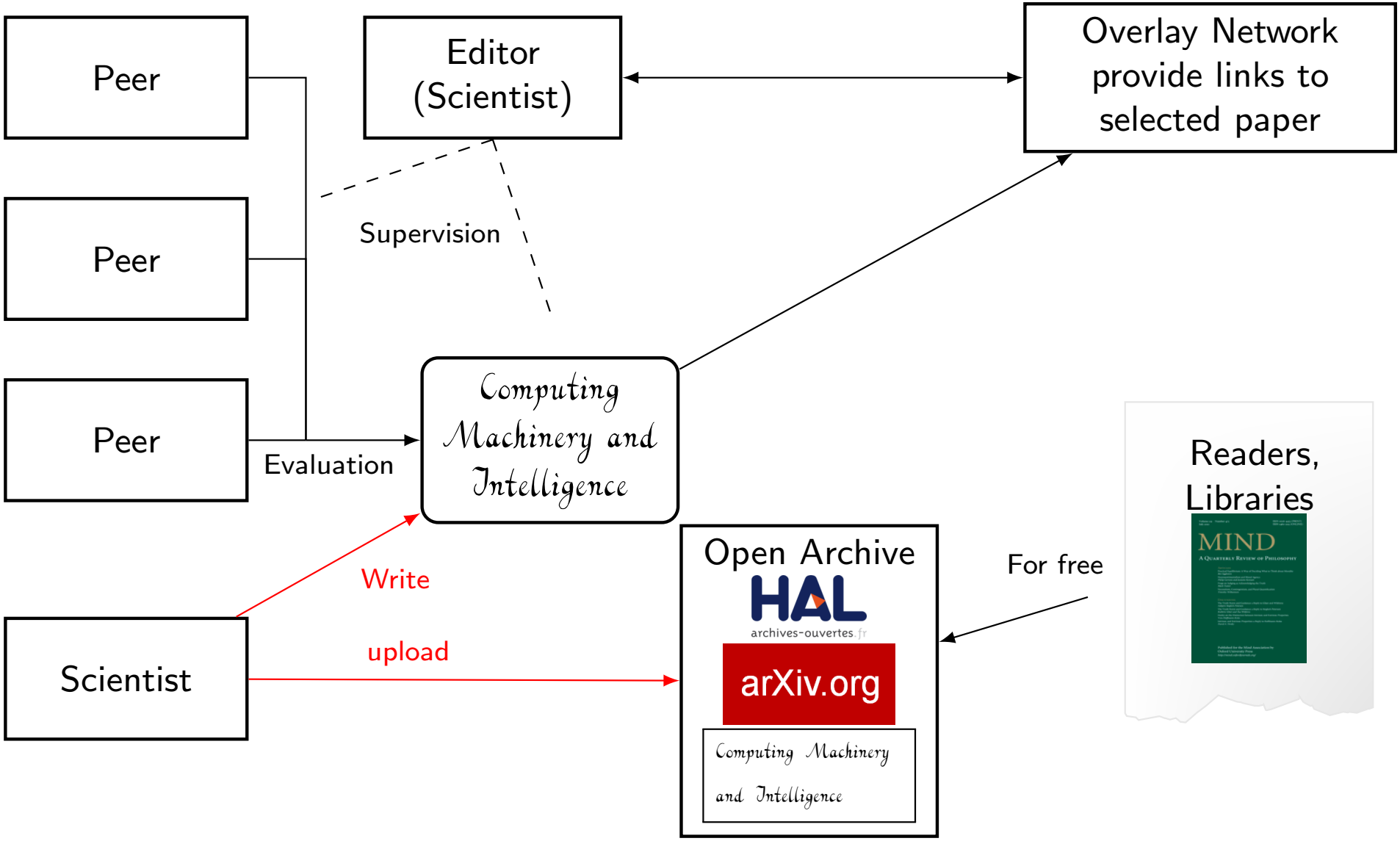
Pirated papers

- LibGen 
- Sci-Hub (Alexandra Elbakyan)  

Bohannon J, Elbakyan A (2016)
Data from: Who's downloading pirated papers? Everyone.
Dryad Digital Repository. https://doi.org/10.5061/dryad.q447c

Overlay Journal : les épi-journaux

épisciences.org



Springer-Nature funded SciDetect: http://scidetect.forge.imag.fr

Université Joseph Fourier GRENOBLE

Springer

SciDetect is a collaboration between Springer-Verlag GmbH and Université Joseph Fourier.

Press release, march 2015

"The open source software discovers text that has been generated with the SCIgen computer program and other fake-paper generators like Mathgen and Physgen."

"SciDetect is highly flexible and can be quickly customized to cope with new methods of automatically generating fake or random text"

Do not cop with other problems

- Peer review rings
- Paper mills
- Black market and authorship selling

Table of Contents

Automatic detection of questionable research papers [Byrne and Labbé, 2017b, Byrne and Labbé, 2017a]

Scientific ethics

- Plagiarism, auto-plagiarism, content reuse...
- $N-grams$ signature (hashing functions).

Non-sense detection

- Paper generator (SCIgen, physic-gen, MathGen...)
- Authorship detection (inter-textual distance).

Need to detect questionable scientific results

- Fabrications (making up data or results)
- Falsification (manipulating data or results)
- False or unsupported affirmations
- Genuine errors

$\Longrightarrow$

- Error spreading
- Wrong belief
- Research irreproducibility

Starting point : striking similarities, obvious errors

Jennifer Byrne:

- First reported *TPD52L2* (20 years ago)
- 5 Publications with obvious errors!

5 Publications from China:

- Single gene knockdown experiments.
- Human cancer cell lines.

Conclusions highlight potential therapy

- ...TPD52L2... novel therapeutic target for glioma treatment.
- ...TPD52L2... novel clues for oral squamous cell carcinoma therapy.
- ...TPD52L2... therapeutic approach for the treatment of breast cancer.
- ...TPD52L2 is indispensable in gastric cancer proliferation.
- ...TPD52L2 could be a novel therapeutic target for human liver cancer.

Obvious errors: example

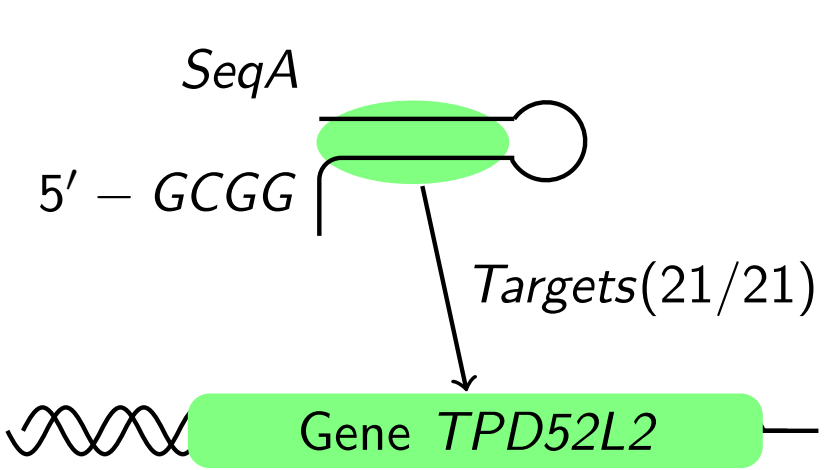
PMID : 25262828

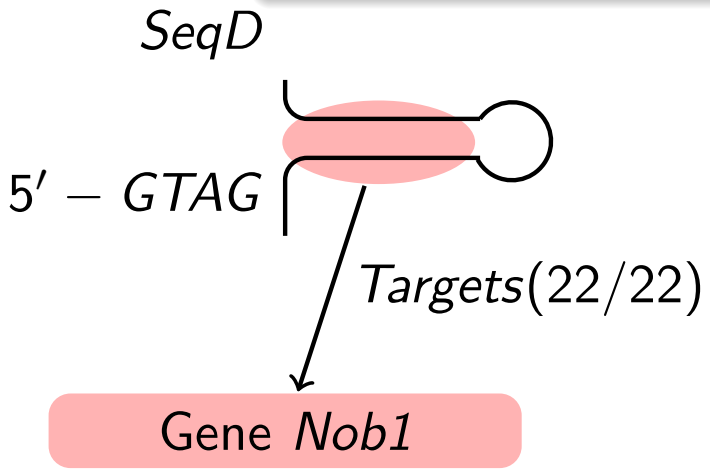
Materials and methods

The shRNA sequence (5'-GCGGAGGGTTTGAAAGAATATCTC-GAGATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCCCGGCCAAG-GAAGTGCAATTGCATACTCGAGTATGCAATTGCACTTC-CTTGGTTTTTTGTTAAT-3') was used as control.

Fact-Check using *blastn* (NCBI)

```
 Query= SeqA   (evalue = 10)
Length=54
Sequences producing significant alignments:
...  ...  ...  ...
> ....  Homo sapiens tumor protein D52
like 2 (TPD52L2), ..
Length=2230
...
Query  1    GCGGAGGGTTTGAAAGAATAT 21
            |||||||||||||||||||||
Sbjct  894  GCGGAGGGTTTGAAAGAATAT 914
....
Query  28   ATATTCTTTCAAACCCTCCGC 48
            |||||||||||||||||||||
Sbjct  914  ATATTCTTTCAAACCCTCCGC 894
```





Obvious errors: example

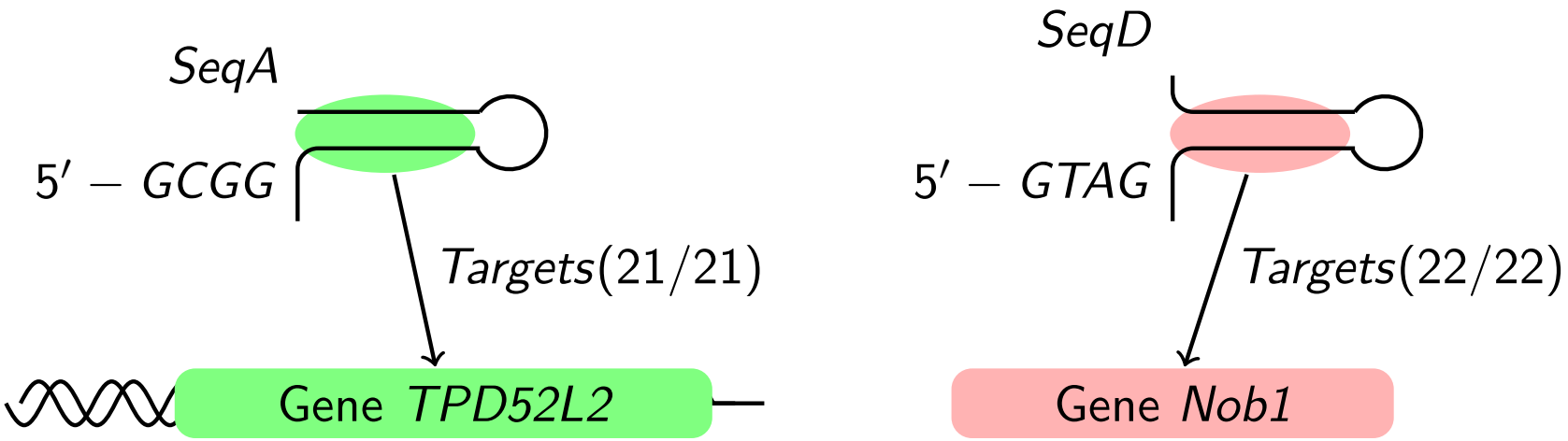
PMID : 25262828

Materials and methods

The shRNA sequence (5'-GCGGAGGGTTTGAAAGAATATCTC-GAGATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCCCGGCCAAG-GAAGTGCAATTGCATACTCGAGTATGCAATTGCACTTC-CTTGGTTTTTTGTTAAT-3') was used as control.

Fact-Check using *blastn* (NCBI)

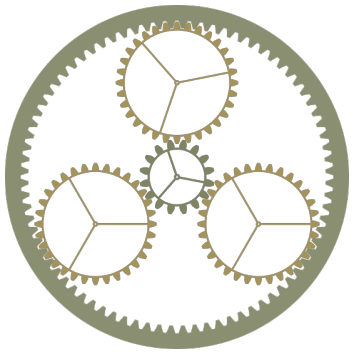
```
 Query= SeqD   (evalue = 10)
Length=68
Sequences producing significant alignments:
...  ...  ...  ...
> ....  Homo sapiens NIN1/PSMD8 binding
protein 1 homolog (NOB1)...
Length=1775
...
Query  9     GCCAAGGAAGTGCAATTGCATA 30
             ||||||||||||||||||||||
Sbjct  1505  GCCAAGGAAGTGCAATTGCATA 1526
....
Query  37    TATGCAATTGCACTTCCTTGG 57
             ||||||||||||||||||||||
Sbjct  1526  TATGCAATTGCACTTCCTTGG 1506
```



Seek & Blastn at a glance

Materials and methods
The shRNA sequence (5'-GCGGAGGGTTTGAAA-GAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCCCGGCCAAGGAAGTG-CAATTGCATACTCGAGTATGCAATTGCACTTCCTTG-GTTTTTTGTTAAT-3') was used as control.

(1) Facts extraction: *Named entity recognition*, extract nucleotide and status...



Facts to check

Status	DNA Seq
...	...
Targeting	GCG...TTT
Non-Targ.	CTA...AAT
...	...

(2) Blastn call software gives the hit list

Hit lists (Blastn results)

hit list	DNA Seq
...	...
TPD52L2, ...	GCG...TTT
NOB1,...	CTA...AAT
...	...

(3) Comparison

Checked Facts

Satus	DNA Seq
Targ.	GCG...TTT
Non-Targ	CTA...AAT
...	...

Ambiguïtés : polysémie, homonymie, structurale,...

- Le président a le pouvoir de faire taire l'avocat.
- Je ne vais pas pouvoir manger l'avocat.
- l'été à l'est a été très beau et l'est toujours.

- Je suis le secrétaire.
- Je vais à la grange et la ferme.

- Il poursuit la jeune fille à vélo.
- Il a vu un homme avec un télescope.
- Tous les participants prendront un bus.

Seek & Blastn

Related works

- Detection of statistically flawed paper
- Fake news detection

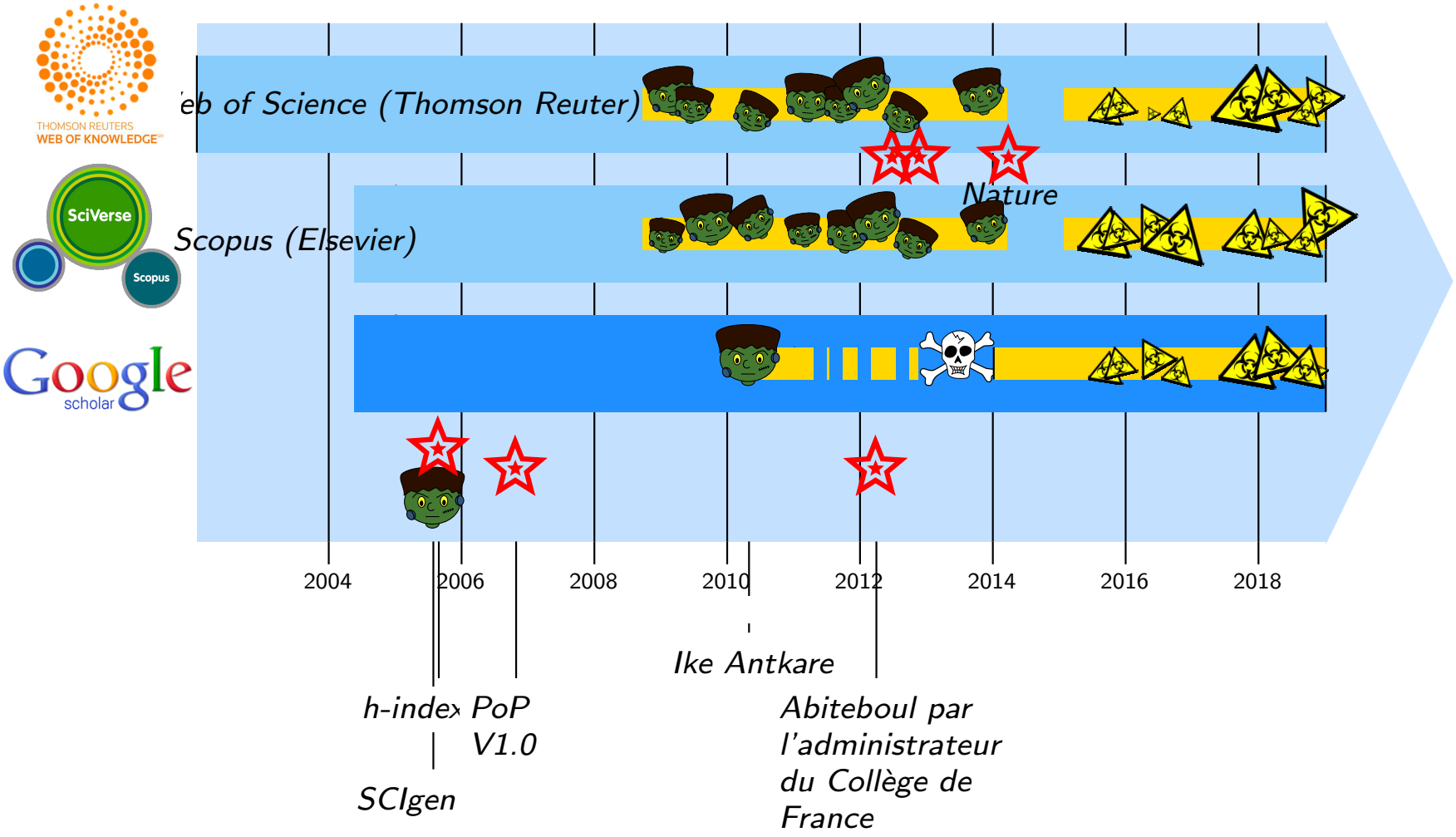
Seek & Blastn perspectives

- Online tool : `http://scigendetection.imag.fr/TPD52`
- Avoid false positive, more in-deep analysis of sentences.

Retractions, Errors corrections

- Retractions (≈ 18), Expression of concern (≈ 11), ≈ 45 to be treated
- Citation analysis (to be done)

Chronos


eb of Science (Thomson Reuter)
Scopus (Elsevier)
Nature
2004
2006
2008
2010
2012
2014
2016
2018
h-index
PoP V1.0
SCIgen
Ike Antkare
Abiteboul par l'administrateur du Collège de France

Conclusion and Future/Ongoing works

Publication procedures, models and habits

- Why fake papers were accepted, published and ... sold.
- Traditional publisher *vs* open access.
- Knowledge diffusion: better and less... or as much as possible.

Blind management rules...

- ... are an incitation to malpractices: slicing, plagiarism, faked data, ...

Automatic detection of new generators

- Hand written PCFG : find dense cluster inside a population.
- Study other kind of generator (language model).

In the web today

- Automatic knowledge extraction/detection/generation.
- How to separate the wheat from the chaff... and scale up !

Thanks

Amancio, D. R. (2015).

Comparing the topological properties of real and artificially generated scientific manuscripts.
Scientometrics, 105(3):1763–1779.

Beel, J. and Gipp, B. (2010).

Academic search engine spam and google scholar's resilience against it.
Journal of Electronic Publishing, 13(3).

Beel, J., Gipp, B., and Wilde, E. (2010).

Academic search engine optimization (aseo).
Journal of scholarly publishing, 41(2):176–190.

Byrne, J. A. and Labbé, C. (2017a).

Fact checking nucleotide sequences in life science publications: The seek & blastn tool.
In *International Congress on Peer Review and Scientific Publication, Enhancing the quality and credibility of science*, Chicago.

Byrne, J. A. and Labbé, C. (2017b).

Striking similarities between publications from china describing single gene knockdown experiments in human cancer cell lines.
Scientometrics, 110(3):1471–1493.

Dalkilic, M. M., Clark, W. T., Costello, J. C., and Radivojac, P. (2006).
Using compression to identify classes of inauthentic texts.
In *Proceedings of the 2006 SIAM Conference on Data Mining*.

Fahrenberg, U., Biondi, F., Corre, K., Jégourel, C., Kongshøj, S., and Legay, A. (2014).
Measuring structural distances between texts.
CoRR, abs/1403.4024.

Ginsparg, P. (2014).

Automated screening: Arxiv screens spot fake papers.
Nature, 508(7494):44–44.

Hirsch, J. E. (2005).

An index to quantify an individual's scientific research output.
Proceedings of the National Academy of Science, 102:16569–16572.

Labbé, C. (2010).

Ike antkare, one of the great stars in the scientific firmament.
International Society for Scientometrics and Informetrics Newsletter, 6(2):48–52.

Labbé, C. and Labbé, D. (2006).

A tool for literary studies. intertextual distance and tree classification.
Literary and Linguistic Computing, 21(3):311–326.

Labbé, C. and Labbé, D. (2013).

Duplicate and fake publications in the scientific literature: how many scigen papers in computer science?
Scientometrics, 94(1):379–396.

Lavoie, A. and Krishnamoorthy, M. (2010).

Algorithmic Detection of Computer Generated Text.
ArXiv e-prints.

Lopez-Cozar, E. D., Robinson-García, N., and Torres-Salinas, D. (2012).
Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting.
arXiv preprint arXiv:1212.0638.

Xiong, J. and Huang, T. (2009).

An effective method to identify machine automatically generated paper.
In *KESE '09. Pacific-Asia Conference*, pages 101–102.