Trends in Gaming Indicators: On Failed Attempts at Deception and their Computerised Detection

Cyril Labbé

Université Grenoble Alpes - LIG - équipe Sigma

March 26, 2018

BIR-ECIR 2018

1. Of Publications and Gamming
   - Scientometrics: what for?
   - Medley
   - SCIgen a Probabilistic Context Free Grammar
2. Of the use of fake publications
   - h-index hacking
   - Resume Padding
   - Journal Hijacking
3. Detection of SCIgen papers
   - Google Search
   - SciDetect: Automatic detection
4. Automatic detection of questionable research papers
   - Fact checking science
   - Seek & Blastn tool

Table of Contents

1. Of Publications and Gamming
   - Scientometrics: what for?
   - Medley
   - SCIgen a Probabilistic Context Free Grammar
2. Of the use of fake publications
   - h-index hacking
   - Resume Padding
   - Journal Hijacking
3. Detection of SCIgen papers
   - Google Search
   - SciDetect: Automatic detection
4. Automatic detection of questionable research papers
   - Fact checking science
   - Seek & Blastn tool

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

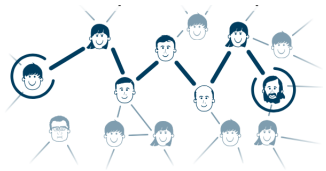
http://sci2s.ugr.es/hindex

*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



Information Systems for science

Scientific publications are at the heart of the system:

- Knowledge diffusion.
- Counting unit.

Increasing number of information sources:

- Publishers repositories
- Open archive and dedicated social networks

Various caracteristics:

- free or toll acces
- Peer review vs non-Peer review

Various goals:

- Spreading knowledge / State of the art / Bibliometry / Scientometrics

Tools that count citations.

Toll based tools.

- Provided by publisher (Elsevier, Thomson reuters);
- Based on publishers catalogs (ACM, IEEE, Springer, Elsevier);
- Selected venues only (all peer reviewed).

Free tools:

- Google Scholar, CiteSeerX,...
- Crawling the web / selected catalogs / added by users;
- Social media (Google+, Scholarometer, Microsoft Academics...).

Free tools that computes indicators

Publish or Perish; Scholarometer; Microsoft Academics; Google+; and many more...

Cases an possible countermeasure

King Abdulaziz University

Recruiting massively highly cited authors in a field.

Hacking peer-review process

Peer-review ring to bypass real peer review and avoid rejection by gaining an easy and quick acceptation.

Academic search engine optimization

Search engine spoofing [Lopez-Cozar et al., 2012, Beel and Gipp, 2010, Beel et al., 2010].

Paper mills

Pay for someone to write and present your paper at a conference/journal.

Citations Analysis

Track down potential manipulations

- h-index and self citations [Bartneck and Servaas, 2011]
- Editors misbehavior [Herteliu et al., 2017]
- Citation cartels [Fister jr et al., 2016]

Content similarity

Track down

- Paper mills (authorship)
- Scientific errors
- Content reuse

The Holy Grail of a lazy scientist

Automatic evaluation (and generation) of (real) scientific papers.

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

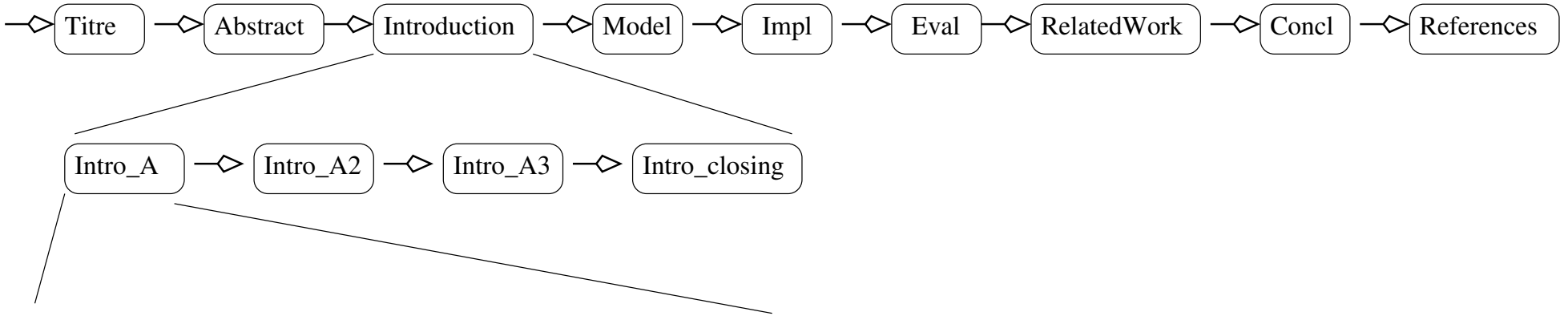
Terminal string example:

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

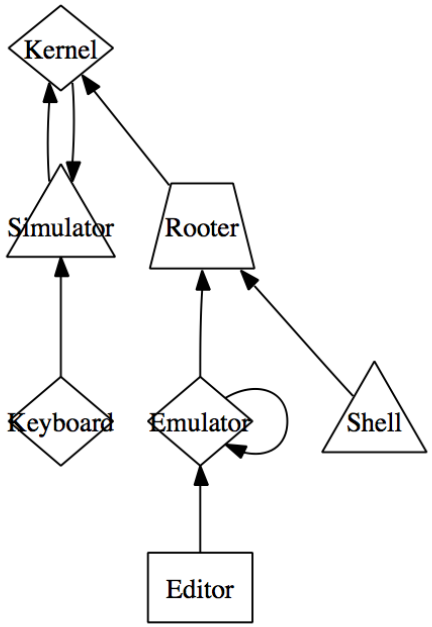


Fig. 2. The schematic used by our methodology.

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.

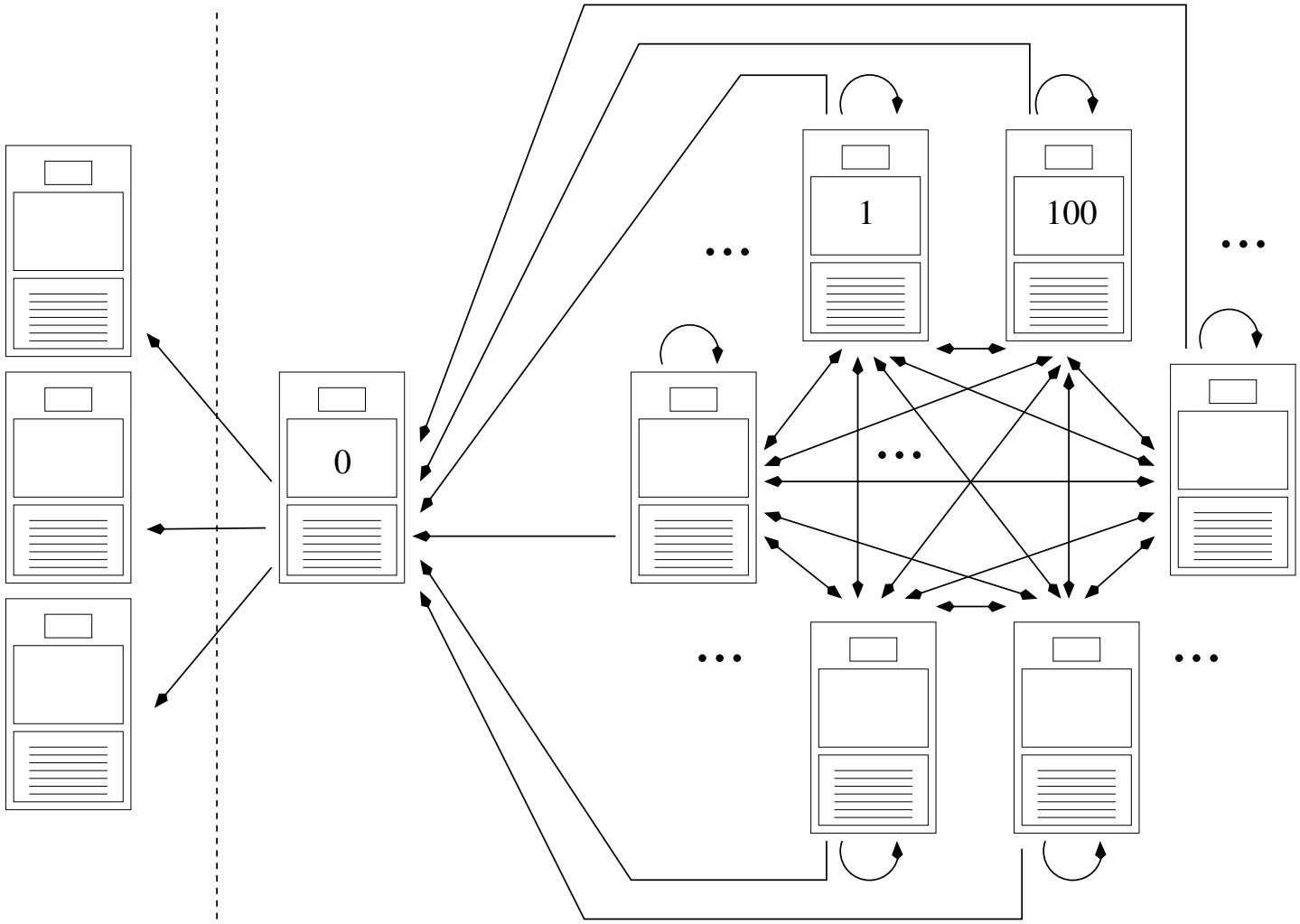
Table of Contents

1. Of Publications and Gamming
   - Scientometrics: what for?
   - Medley
   - SCIgen a Probabilistic Context Free Grammar
2. Of the use of fake publications
   - h-index hacking
   - Resume Padding
   - Journal Hijacking
3. Detection of SCIgen papers
   - Google Search
   - SciDetect: Automatic detection
4. Automatic detection of questionable research papers
   - Fact checking science
   - Seek & Blastn tool

Building a *citation farm*

[Labbé, 2010]

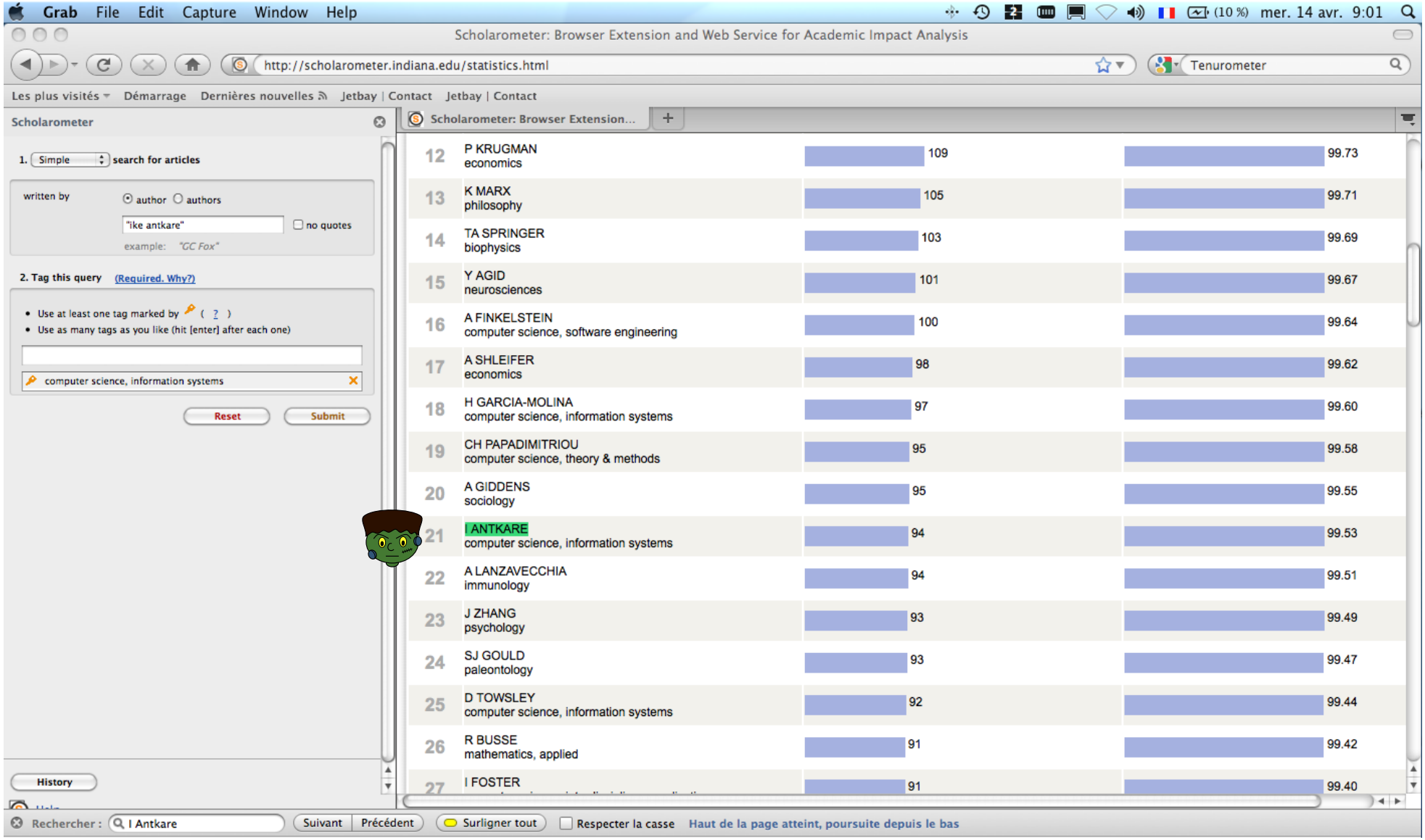
Modified SCIgen



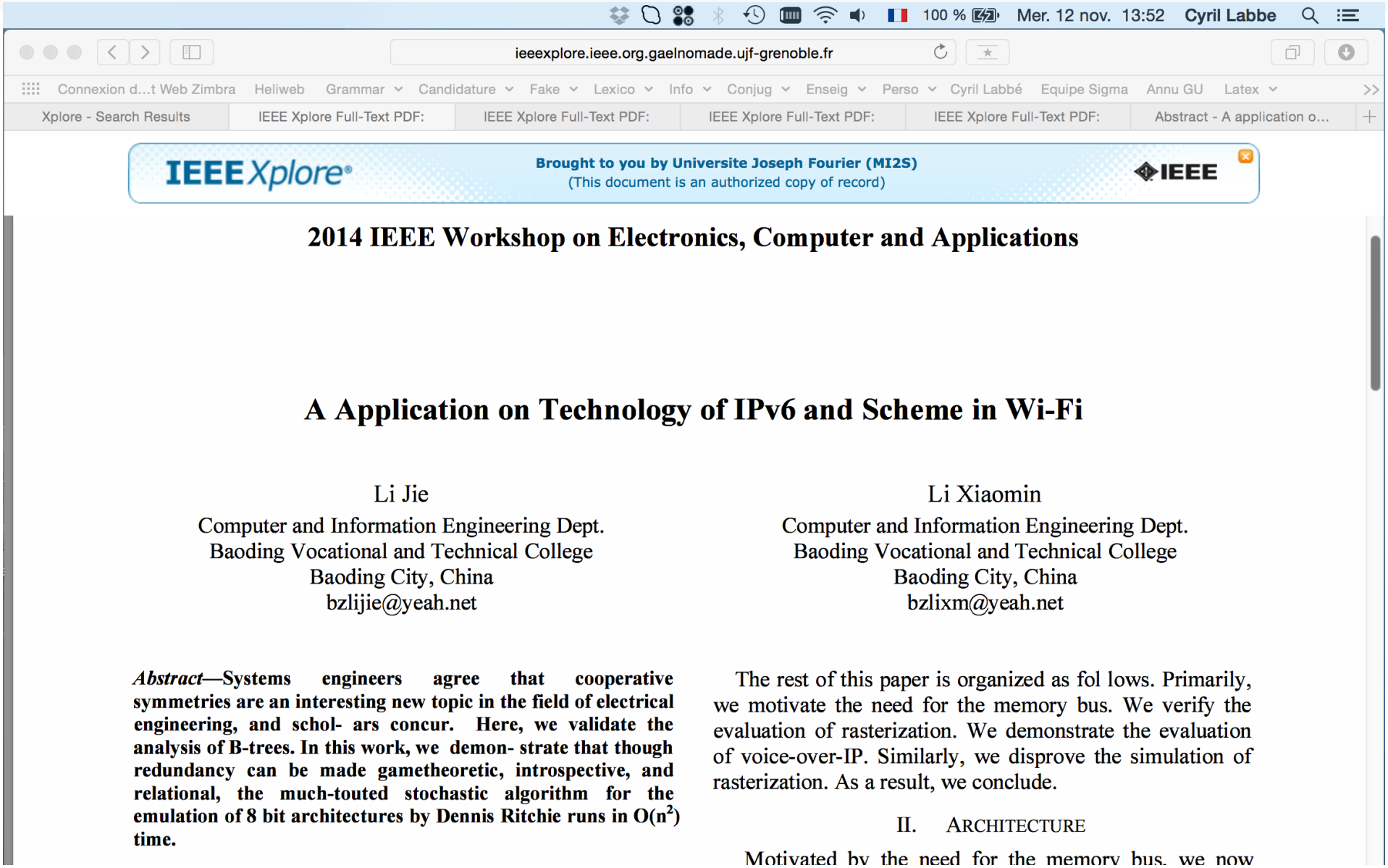
Ike Antkare h-index

Google scholar | Scholarometer

[Labbé, 2010]



IEEEXplore: 12 nov. 2014



IEEEXplore: 2 feb. 2016

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 (Lorem Ipsum)

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

Table of Contents

1. Of Publications and Gamming
   - Scientometrics: what for?
   - Medley
   - SCIgen a Probabilistic Context Free Grammar
2. Of the use of fake publications
   - h-index hacking
   - Resume Padding
   - Journal Hijacking
3. Detection of SCIgen papers
   - Google Search
   - SciDetect: Automatic detection
4. Automatic detection of questionable research papers
   - Fact checking science
   - Seek & Blastn tool

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...

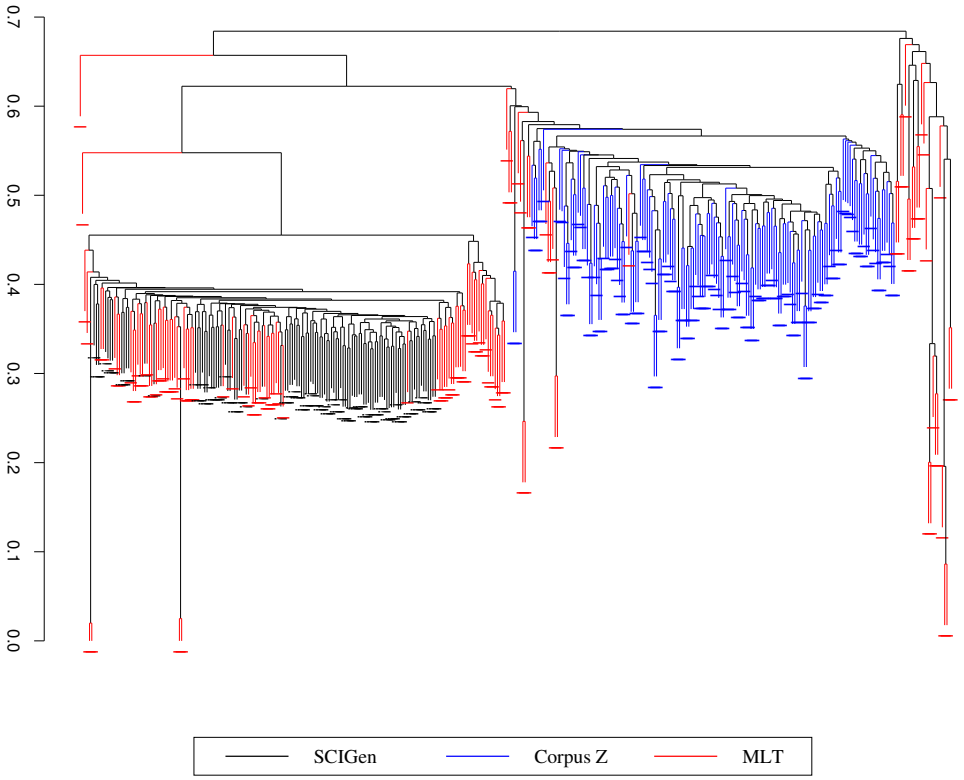
BUY ARTICLE $19 ABSTRACT MORE LIKE THIS

Automatic detection (classification) [Labbé and Labbé, 2013]

Inter-textual Distance:

$\Delta_{(a,b)} = \delta$ proportion of different works (tokens) in the two texts.

Hierarchical Clustering



Let

- t a text under test.
- $\delta_t^{Fake} = \min_{f \in SCIgen} \Delta_{(t,f)}$

Si $(\delta_t^{Fake} < \delta_{th})$ Then

The text is almost surely SCIgen generated.

Else

non-SCIgen.

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

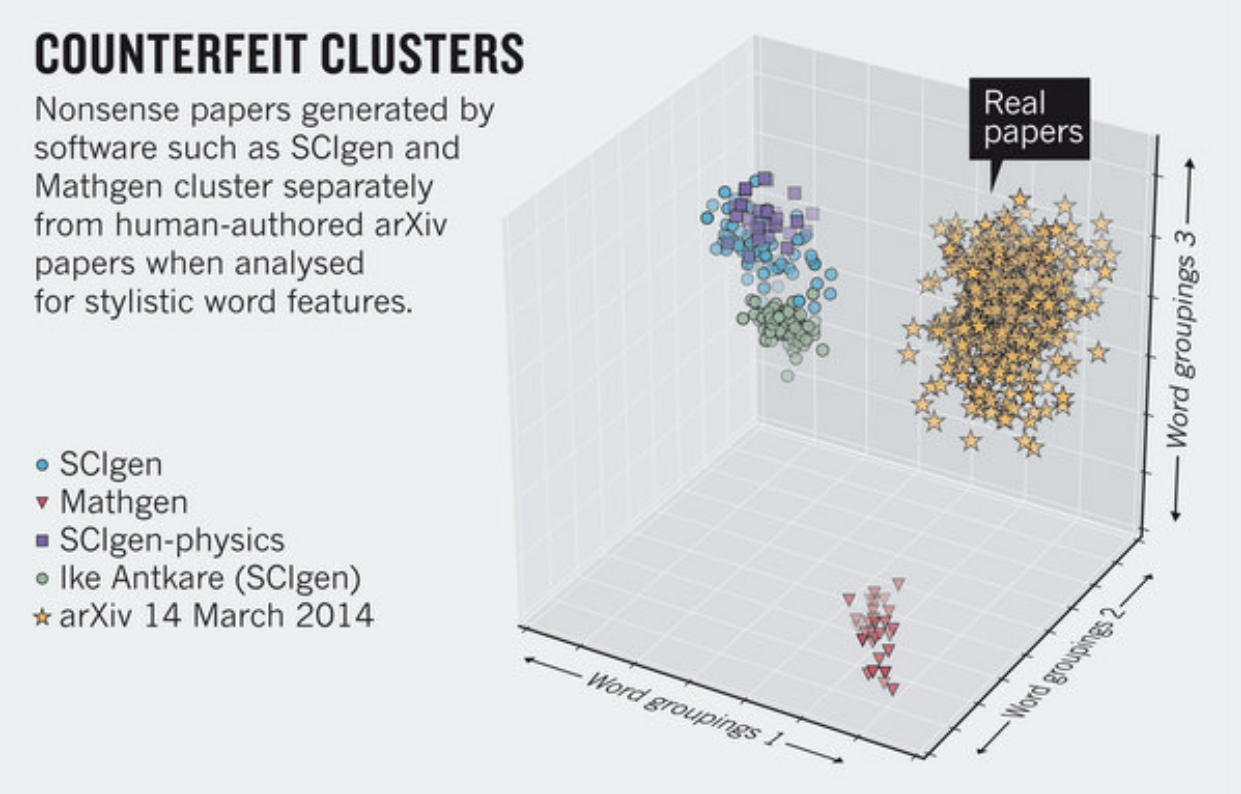
Other generators

- 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/)

Mainstream press (2014), is it arming science?

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]

Related/Ongoing Work

Detecting

- Based on references [Xiong and Huang, 2009],
- Compression based and ad-hoc classifier [Dalkilic et al., 2006],
- Ad-hoc similarity and classifier [Lavoie and Krishnamoorthy, 2010],
- Structural distances between texts [Fahrenberg et al., 2014].
- Phrases search [Springer, 2014].
- Topological properties [Amancio, 2015]

Spoofing

- [Beel and Gipp, 2010, Lopez-Cozar et al., 2012],
- Academic optimisation [Beel et al., 2010];

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

1. Of Publications and Gamming
   - Scientometrics: what for?
   - Medley
   - SCIgen a Probabilistic Context Free Grammar
2. Of the use of fake publications
   - h-index hacking
   - Resume Padding
   - Journal Hijacking
3. Detection of SCIgen papers
   - Google Search
   - SciDetect: Automatic detection
4. Automatic detection of questionable research papers
   - Fact checking science
   - Seek & Blastn tool

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

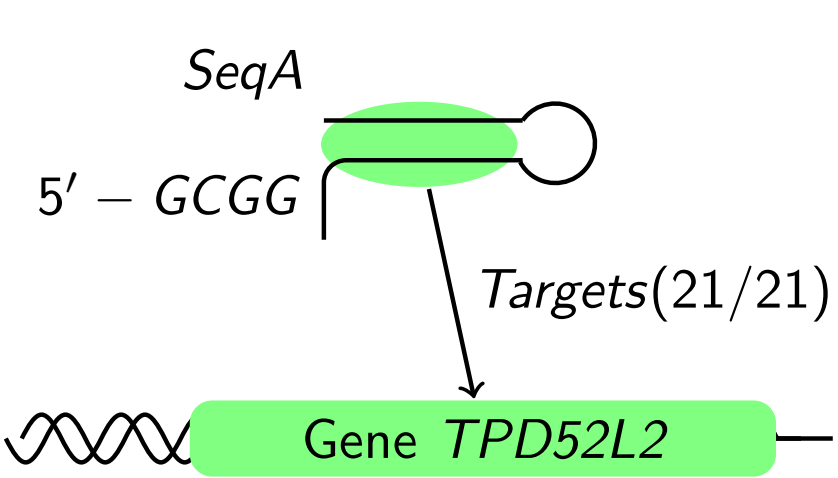
PMID : 25262828

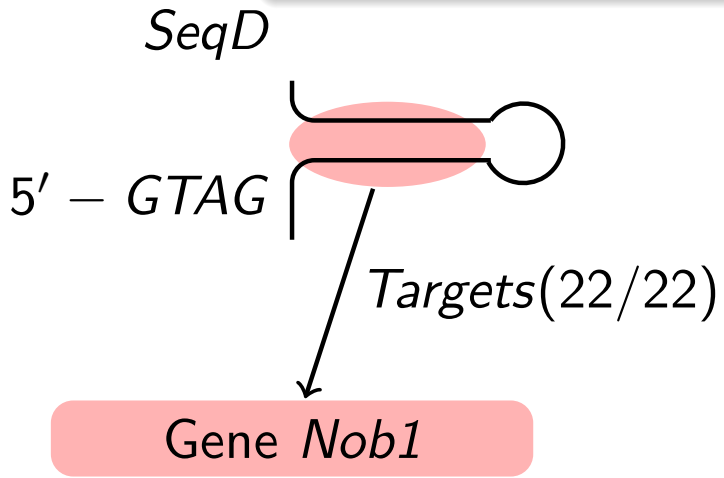
Materials and methods

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

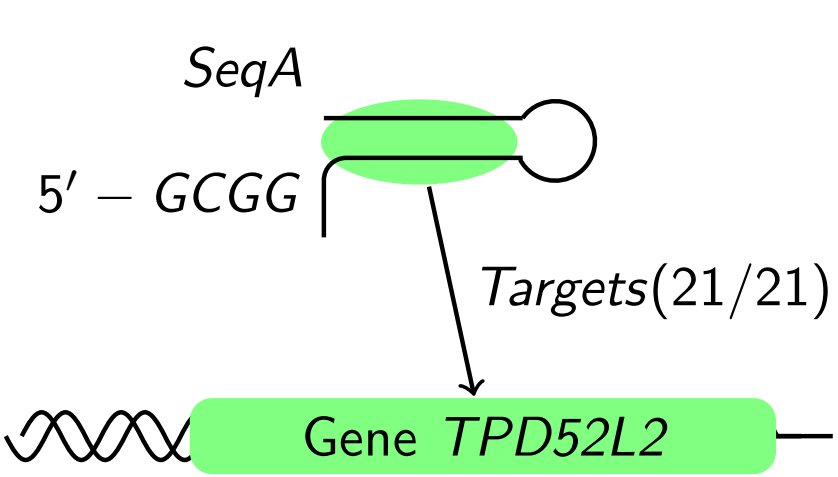
PMID : 25262828

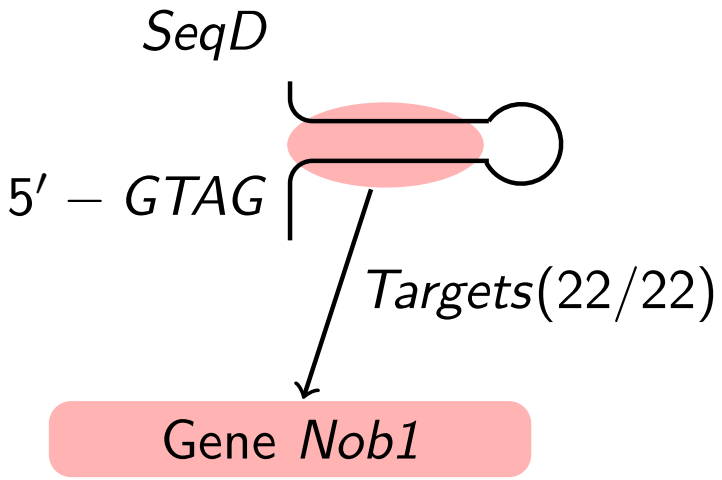
Materials and methods

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
```

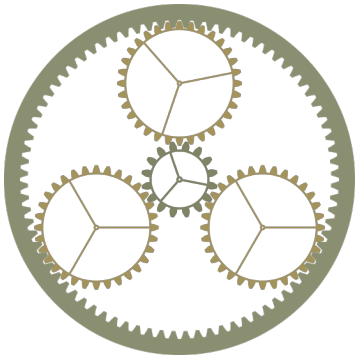




Seek & Blastn at a glance

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
...	...

Tests and results

Used Corpora.

Problematic Paper (CorpusP):

- 38/48 (79%) highly similar publications with nucleotide sequence(s) did not match their experimental use (*blastn*).

Unknown papers (CorpusU):

- 154 papers, retrieved using CorpusP papers and the "PubMed similar" function.

Seek & Blastn performances

- In CorpusU nucleotide sequences were extracted from 111/154 (73%) papers.
- Claims were not (correctly) identified for 19/341 (5.6%) sequences in CorpusP.
- Identification of the 38/48 (79%) CorpusP papers that incorrectly employed sequences.

Error detection in scientific literature

- 38 papers in CorpusP appear to have incorrectly employed nucleotide sequence.
- "seek & blastn" predicted that 30/154 (19%) CorpusU papers may have incorrectly employed nucleotide sequence reagent(s).

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

- A few retractions (≈ 10), ≈ 50 to be treated
- Citation analysis (to be done)

Open Access vs Fee based

- When fee-based, automatically download is not permitted.
- Paywall are hiding good and junk science

Conclusion, Future/Ongoing works

Publication procedures, models and habits

- Why fake papers were accepted, published and ... sold.
- Traditional publisher *vs* open access.
- Blind management rules: incitation to malpractices, slicing, faked data, ...

Automatically Identify and flag scientific errors/breakthrough

- Mutual enrichment of two families of techniques (B+IR).
- Joint analysis of citations and text.

Measurement of *perturbations*...

- ... introduced by measuring science.

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.

Bartneck, C. and Servaas, K. (2011).

Detecting h-index manipulation through self-citation analysis.
Scientometrics, 87(1):85–98.

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.

Fister jr, I., Fister, I., and Perc, M. (2016).
Toward the discovery of citation cartels in citation networks.
4.

Ginsparg, P. (2014).

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

Herteliu, C., Ausloos, M., Ileanu, B. V., Rotundo, G., and Andrei, T. (2017).
Quantitative and qualitative analysis of editor behavior through potentially coercive citations.
Publications, 5(2).

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.