Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science?

Cyril Labbé, Dominique Labbé

Duplicate and Fake Publications in the Scientific Literature: How many SCIgen papers in Computer Science?

Cyril Labbé
Université Joseph Fourier
Laboratoire d'Informatique de Grenoble
First.Last@imag.fr

Dominique Labbé
Institut d'Etudes Politiques de Grenoble
PACTE
First.Last@iep-grenoble.fr

22 june 2012 ; Scientometrics; DOI 10.1007/s11192-012-0781-y

Abstract

Two kinds of bibliographic tools are used to retrieve scientific publications and make them available online. For one kind, access is free as they store information made publicly available online. For the other kind, access fees are required as they are compiled on information provided by the major publishers of scientific literature. The former can easily be interfered with, but it is generally assumed that the latter guarantee the integrity of the data they sell. Unfortunately, duplicate and fake publications are appearing in scientific conferences and, as a result, in the bibliographic services. We demonstrate a software method of detecting these duplicate and fake publications. Both the free services (such as Google Scholar and DBLP) and the charged-for services (such as IEEE Xplore) accept and index these publications.

keyword: Bibliographic Tools, Scientific Conferences, Fake Publications, Text-Mining, Inter-Textual Distance, Google Scholar, Scopus, WoK

1 Introduction

Several factors are substantially changing the way the scientific community shares its knowledge. On the one hand, technological developments have made the writing, publication and dissemination of documents quicker and easier. On the other hand, the "pressure" of individual evaluation of researchers–publish or perish–is changing the publication process. This combination of factors has led to a rapid increase in scientific document production. The three largest tools referencing scientific texts are: *Scopus* (Elsevier), *ISI-Web of Knowledge* (WoK Thomson-Reuters) and *Google Scholar*.

Google Scholar is undoubtedly the tool which references the most material. It is free and it offers wide coverage, both of which are extremely useful to the scientific community. *Google Scholar* allows *grey literature* to be more visible and more accessible (technical reports, long versions and/or tracts of previously published papers, etc). *Google Scholar* systematically indexes everything that looks like a scientific publication on the internet, and, inside these documents and records, it indexes references to other documents. Thus, it gives a picture of which documents are the most popular. However, the tool, much like the search engine Google, is sensitive to "Spam" [2], mainly through techniques, similar to link farms that artificially increase the "ranking" of web pages. Faked papers like those by Ike Antkare [12] (see 2.2 below) may also be mistakenly indexed. This means that documents indexed by *Google Scholar* are not all *bona fide* scientific ones, and information on real documents (such as the number of citations found)

can be manipulated. This type of tool, using information publicly and freely available on the Web, faces some reproducibility and quality control problems [22, 10].

In comparison, editorial tools (such as *Scopus* or *WoK*) seem immune to this reproach. They are smaller, less complete and require access fees, but in return they may be considered as "cleaner". This is mainly because they store only publications in journals and conferences in which peer selection is supposed to guarantee the quality of the indexed publications. The number of citations is computed in a more parsimonious way and meets more stringent criteria. Data quality would also seem to be secured by a new selection by the publisher who provide the tool:

"This careful process helps Thomson Scientific remove irrelevant information and present researchers with only the most influential scholarly resources. A team of editorial experts, thoroughly familiar with the disciplines covered, review and assess each publication against these rigorous selection standards"[11][1].

Differences between these tools have been studied [7, 25, 9]. But are they immune from failures such as multiple indexing of similar or identical papers (duplicates), or even the indexing of meaningless publications?

A first answer to these questions will be provided by the means of several experiments on sets (corpora) of recent texts in the field of Computer Science. Text-mining tools are presented and used to detect problematic or questionable papers such as duplicated or meaningless publications. The method has enabled the identification of several bogus scientific papers in the field of Computer Science.

2 Corpora and texts preprocessing

Table 1 gives a synthetic view of the sets of texts used along this article[2].

A priori above-reproach corpora: Most of the texts used in these corpora are indexed in bibliographic tools (*Scopus* and *WoK*). They are either available from the conferences' web sites, or from the publishers' web sites, like the Institute of Electrical and Electronic Engineers (IEEE) or Association for Computing Machinery (ACM) websites, which sponsor a large number of scientific events in the field of electronics and computer science. Acceptance rates are published by the conferences chairs in the proceedings. Texts of corpora X, Y and Z were published in three conferences (X, Y and Z). The MLT corpus is composed of texts published in various conferences. They have been retrieved by applying, to 3 texts of the corpus Y, the "More Like This" functionality provided by IEEE (see figure 1).

Representative set of articles in the field of Computer Science: ArXiv is an open repository for scholarly papers in specific scientific fields. It is moderated via an endorsement system which is not a peer review: "We don't expect you to read the paper in detail, or verify that the work is correct, but you should check that the paper is appropriate for the subject area"[3].

All the computer science papers for the years 2008, 2009 and 2010 were downloaded from the arXiv repository. Excluding the ones from which text could not be extracted properly this represent: 3481 articles for year 2008, 4617 for 2009 and 7240 for 2010.

[1] http://ip-science.thomsonreuters.com/news/2005-04/8272986/

[2] Bibliographic information and corpora are available upon request to the authors

[3] http://arxiv.org/help/endorsement

Table 1: Corpora description: NA stand for non available.

Corpus name	Downloaded from	Years	Type of papers	Number of papers	Acceptance rate	Corpus size
Corpus X	ACM portal.acm.org	2010	Full Short Demo	126 165 20	13.3% 17.5% 52%	311
Corpus Y	IEEE ieee.org	2009	Regular	150	28%	150
Corpus Z	Conf. Web Site	2010	Track 1 Track 2 Track 3 Demo	58 33 36 32	18.4% 16.1% 36%	153
MLT	IEEE ieee.org	200x-20yy	various	122	NA	122
arXiv	arxiv.org	2008 2009 2010	various	3481 4617 7240	NA NA	15338

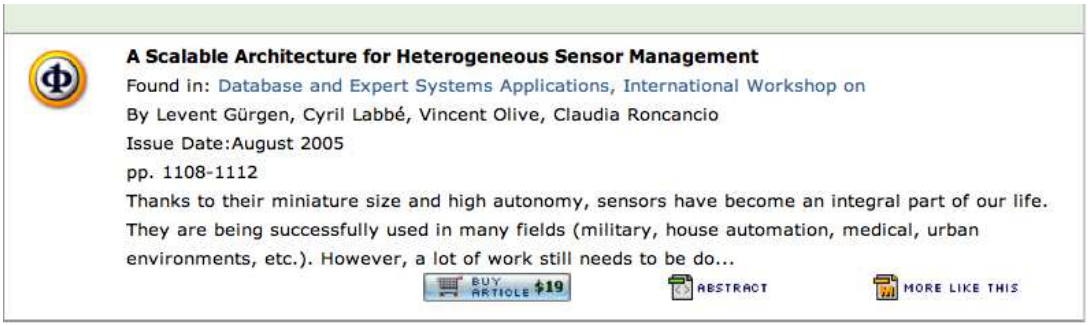

Figure 1: The "More Like This" functionality was applied to 3 texts of the Y corpus.

Automatically generated, deliberately faked texts: These corpora contain documents automatically generated using the software SCIgen[4]. This software, developed at MIT in 2005, generates random texts without any meaning, but having the appearance of research papers in the field of computer science, and containing summary, keywords, tables, graphs, figures, citations and bibliography. Table 2 shows the first words for some of the 13 possible sentences that start a SCIgen paper. Inside these sentences, token starting with `SCI` are randomly chosen among predefined words. For example, `SCI_PEOPLE` have 23 possible values including: *steganographers, cyberinformaticians, futurists* or *cyberneticists*. `SCI_BUZZWORD_ADJ` have 74 possible values such as: *omniscient, introspective, peer-to-peer* or *ambimorphic*. The whole SCIgen grammar have almost four thousand lines and is fairly complex. Texts are also embellished with rather eccentrics graphs and figures. This allows the generation of a very large set of different texts syntactically correct but without any meaning, which can be spotted quite easily.

[4]http://pdos.csail.mit.edu/scigen/

Table 2: First words of sentences that start a SCIgen-Origin paper.

`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; LIT_REVERSAL, ...`
`SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...`
`The SCI_ACT is a SCI_ADJ SCI_PROBLEM.`
`The SCI_ACT has SCI_VERBED SCI_THING_MOD, and current trends suggest that ...`
`Many SCI_PEOPLE would agree that, had it not been for SCI_THING, ...`
`The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...`

For the Antkare experiment, SCIgen was modified so that each article had references to the 99 others–creating a link farm. Thus, all these texts have the same bibliography. Google Scholar retrieved these faked online articles and, as a result, Ike Antkare's H-index reached 99, ranking him in the 21st position of the most highly cited scientists [12].

The corpus Antkare is composed of the 100 documents used for this experiment. 236 articles generated by the original version of the SCIgen software compose the corpus SCIgen-Origin.

At least one other version of SCIgen exists. It is an adaptation of the original SCIgen "for physics, especially solid state physics and neutron scattering"[5]. A set of 414 articles generated by this software will be referred in the following as the corpus SCIgen-Physics.

Table 3: SCIgen Corpora

Corpus name	Generator	Scientific field	Corpus size
SCIgen-Origin	Original SCIgen	Computer Science	236
Antkare	Modified SCIgen	Computer Science	100
SCIgen-Physics	Modified SCIgen	Physics	414

Table 3 gives a synthetic view of the used SCIgen corpora, examples of SCIgen-Origin and SCIgen-Physics can be found in appendix A.

Texts Processing: Pdf files are converted to plain text files by the program "pdftotxt" (free software unix and windows version 3.01) that extracts the text from pdf files. During this operation, figures, graphs and formulas disappear, but the titles and captions of these figures and tables remain. To prevent the 100 identical references in the corpus Antkare from disturbing the experiments, the bibliographies (and appendices) have been removed from all texts in all corpora.

The texts are segmented into word-tokens using the Oxford Concordance Program commonly used for English texts [8]. In fact, the word-tokens are caracter strings separated by spaces or punctuation. This procedure could be further improved for example by word tagging to replace all the abbreviations and inflections of a single word with a unique spelling convention (infinitive form of verbs, singular masculine of adjectives, etc.)

[5]Blog post: http://pythonic.pocoo.org/2009/1/28/fun-with-scigen
SCIgen-Physics Sources: https://bitbucket.org/birkenfeld/scigen-physics/overview

3 Text mining tools

Distances between a text and others (inter-textual distances) are computed. Then these distances are used to determine which texts, within a large set, are closer to each other and may thus be grouped together.

Inter-textual distance: The distance between two texts A and B is measured using the following method (previous work in [13, 14]). Given two texts A and B, let us consider:

- N_A and N_B: the number of *word-tokens* in A and B respectively, ie the lengths of these texts;
- F_{iA} and F_{iB}: the absolute frequencies of a type i in texts A and B respectively;
- $|F_{iA} - F_{iB}|$ the absolute difference between the frequencies of a type i in A and B respectively;
- $D_{(A,B)}$: the inter-textual distance between A and B is as follows:

$$D_{(A,B)} = \sum_{i \in (A \cup B)} |F_{iA} - F_{iB}| \quad with \quad N_A = N_B \tag{1}$$

The distance index (or relative distance) is as follows:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - F_{iB}|}{N_A + N_B} \tag{2}$$

This index can be interpreted as the proportion of different words in both texts. A distance of 0.4 means that the texts share 60% of their words-token.

If the two texts are not of the same lengths in tokens ($N_A < N_B$), B is "reduced" to the length of A:

- $U = \frac{N_A}{N_B}$ is the proportion used to reduce B in B'
- $E_{iA(u)} = F_{iB}.U$ is the theoretical frequency of a type i in B'

In the Equation (1), the absolute frequency of each word-type in B is replaced by its theoretical frequency in B':

$$D_{(A,B')} = \sum_{i \in (A \cup B)} |F_{iA} - E_{iA(u)}|$$

Putting aside rounding-offs, the sum of these theoretical frequencies is equal to the length of A. The Equation (2) becomes:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - E_{iA(u)}|}{N_A + N_{B'}}$$

This index varies evenly between 0 – the same vocabulary is used in both texts (with the same frequencies) – and 1 (both texts share no word-token). An inter-textual distance of δ can be interpreted as follows: choosing randomly 100 words in each text, δ is the expected proportion of common words between this two sets of 100 words.

In order to make this measure fully interpretable:

- the texts must be long enough (at least more than 1000 word-tokens),

- one must consider that, for short texts (less than 3000 word-tokens), values of the index can be artificially high and sensitive to the length of the texts, and
- the lengths of the compared texts should not be too different. In any case, the ratio of the smallest to the longest must be less than 0.1.

Inter-textual distance depends on four factors. In order of decreasing importance, they are as follows: genre, author, subject and epoch. In the corpora presented above, all texts are in the same genre (scientific papers) and are contemporary. Thus only the authorial and thematic factors remain to explain some anomalies.An unusually small inter-textual distance suggests striking similarities and/or texts by the same author.

Agglomerative Hierarchical Clustering: The inter-textual distances allow agglomerative hierarchical clustering according to similarities between texts and graphical representations of their proximities [23, 3, 20, 21].

This representation is used to identify more or less homogeneous groups in a large population. The best classification is the one that minimizes the distances between texts of the same group and maximizes the distances between groups.

An agglomerative hierarchical clustering is performed on the inter-textual distance matrix, using the following method. The algorithm proceeds by grouping the two texts separated by the smallest distance and by recomputing the average (arithmetic mean) distance between all other texts and this new set, and so on until the establishment of a single set.

These successive groupings are represented by a dendrogram with a scale representing the relative distances corresponding to the different levels of aggregation (see Figure 3 and 4).

By cutting the graph, as close as possible to a thresholds considered as significant, one can demarcate groups of texts as very close, fairly close, etc. The higher the cut is made, the more heterogeneous the classes are and the more complex is the interpretation of the differences. To correctly analyze these figures, it must be also remembered that:

- whatever their position on the non-scaled axis, the proximity between two texts or groups of texts is measured by the height at which the vertices uniting them converge, and
- the technique sometimes results in "chain effects": some similarities between texts are indistinguishable because the vertices connecting them are erased by aggregations performed at a lower level.

Related work: One can find, in the scientific literature, several indices for measuring the similarities (or dissimilarities) between texts. Most often, these indices are based on the vocabulary matrix. Cosine and Jaccard indexes are frequently used and they seem to be well adapted to texts [16]. Some indices based on compression have also been tested [17]. Compared to these indices, intertextual distance is easily interpretable: it is a measure of the proportion of word-tokens shared by two texts. Based on frequencies it could be interpreted as being closely related to information theory: having always the same word-types at the same frequencies do not provide any new information.

In the past recent years, some methods have been developed aiming at automatically identifying SCIgen papers. [24] checks whether references are proper references that points to documents known by the databases available online. A paper having a large proportion of unidentified references will be suspected to be a SCIgen paper. An other approach is proposed in [15]. This method is based on an ad-hoc similarity measure in which the reference section plays a major role. These characteristics explain why these techniques were not able to identify

texts by Ike Antkare as being SCIgen paper[6]. A third proposition [5] is based on observed compression factor and a classifier. A paper under test will be classified as being generated if it has a compression factor similar to known generated text. The method focuses on detecting SCIgen paper but also, what is more, on detecting any kind of texts generated automatically[7]. A simple test shows that this software wrongly classifies as authentic the texts by Antkare (when their reference sections are not withdrawn), with around 10% risks of error, and that it identifies the same texts as inauthentic, when their reference sections are withdrawn... Finally, again, these methods do not provide an easily interpretable procedure for the comparison of texts (in contrast with intertextual distance).

Interesting questions: Like most of the metrics of textual similarities, inter-textual distance, is based on the so called "bag-of-word" approach. Such measures are sensitive to word frequencies but insensitive to syntax. Using this kind of approach to detect SCIgen papers relies on the fact that, despite its wide range of preset sentences, the SCIgen vocabulary remain quite poor: SCIgen is behaving like an author that would have been poorly gifted with vocabulary.

The combination of intertextual distance with agglomerative hierarchical clustering allows some interesting questions to be answered. For example, do the conferences under consideration contain the following occurencies?

- "chimeras" comparable to the texts by Ike Antkare
- "duplicates": the same authors present the same text twice under different titles
- "related papers": covering a wide range of cases, going from almost unchanged texts to close texts by the same author(s) dealing with the same topics, sometimes sharing similar portions of text. The scientific contents of these texts may be substantially different. The proposed tools do not provide any help to measure these differences.

4 Detection of forgeries, duplicates and related papers in the three conferences X, Y and Z

Intra-corpus distances: For each corpus, distances are ranked by ascending values and distributed in equal interval classes. Fig. 2 shows these distributions.

The X, Y and Z corpora have the classic bell curve profile suggesting the existence of relatively homogeneous populations (here a large number of contemporary authors writing in a similar genre and on more or less similar themes). X and Z have a comparable mean/mode and a similar dispersion. In contrast,

- Y has a high average distance and a higher dispersion around this mean, indicating heterogeneity of papers, but also suggesting the presence of anomalies (these two explanations are not mutually exclusive);
- On the left of the graph, the curve with three modes is the distribution of distances between the 100 faked texts by Ike Antkare. This trimodal distribution suggests the existence of two different populations within the texts generated by the modified SCIgen: a small group with very low internal distances are centered on 0.2 - these are short texts (about 1600 word-tokens) - and the other group, with a greater number of texts, containing longer texts (about 3000 word-tokens): Their internal distances are centered on 0.38. The third mode is distances between these two groups.

[6] http://paperdetection.blogspot.com/
[7] http://montana.informatics.indiana.edu/cgi-bin/fsi/fsi.cgi

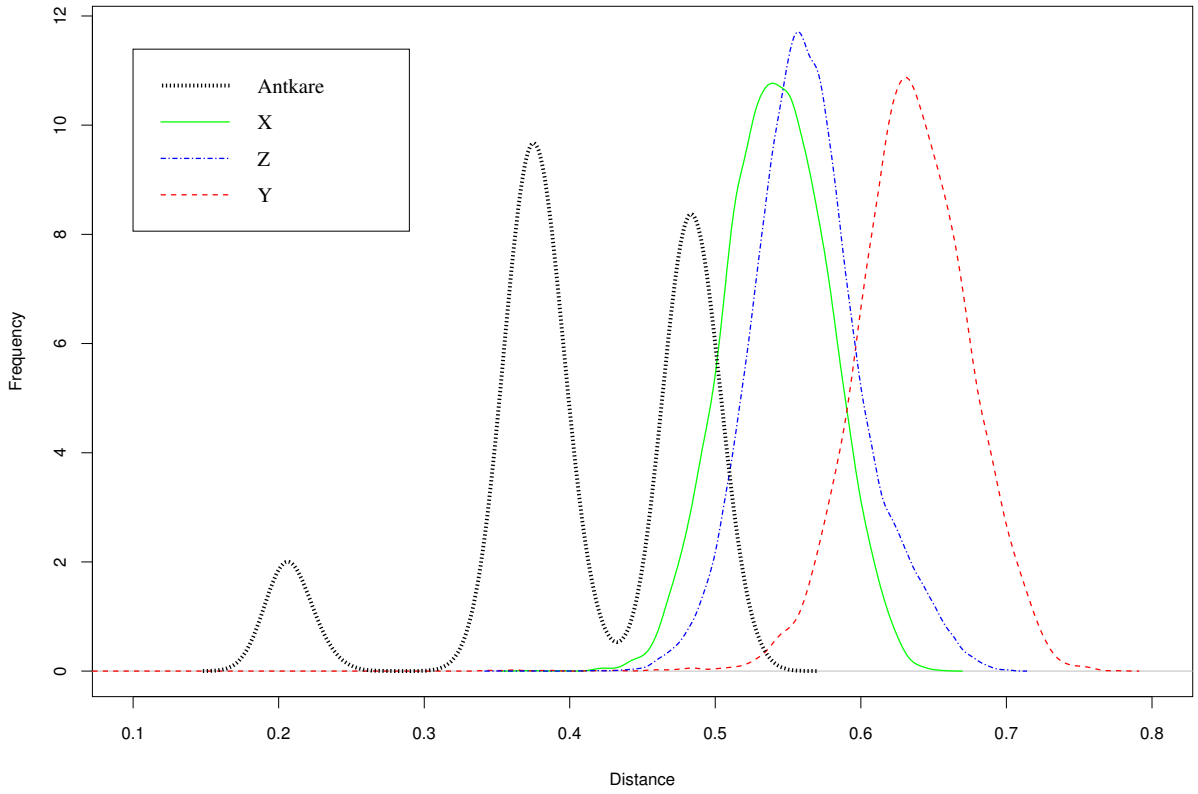

Figure 2: Distribution of intra-corpus distances.

Main Groups: The classification and its representation by a dendrogram (Figure 3) show four main groups:

- In the center, a large body (**C**) includes all texts Z and almost all X texts. It would be possible to isolate various subgroups within this group to show what are the main topical themes of these conferences.
- on the right (**D**) and on the extreme left (**A**), the texts of the Y conference meet at the higher levels, confirming the heterogeneity of this conference.
- There is very little intermingling between X, Z on one side and Y on the other side: only six Y papers are included into X-Z set, but they are attached, at a very high level, to this set (*i.e.* with significant distances). Similarly, only four X papers are included in group **A** (Y). In other words, most of the papers presented at the Y conference are not of the same nature as those presented at the other two conferences.

Finally, all the chimeras generated by SCIgen for Ike Antkare are grouped in **B** into two homogeneous groups and connected at a very low level. Thus, SCIgen texts are not "close" to natural language and are distinct from the scientific papers they are supposed to emulate.

Four "genuine-fake" texts: In the dendrogram in Figure 3, the number (**1**) branches are four Y texts that are clustered within the corpus Antkare.

These four texts are "genuine" publications because they have, at least formally, been selected by peer reviewers. They are "real publications" also because they are in conference

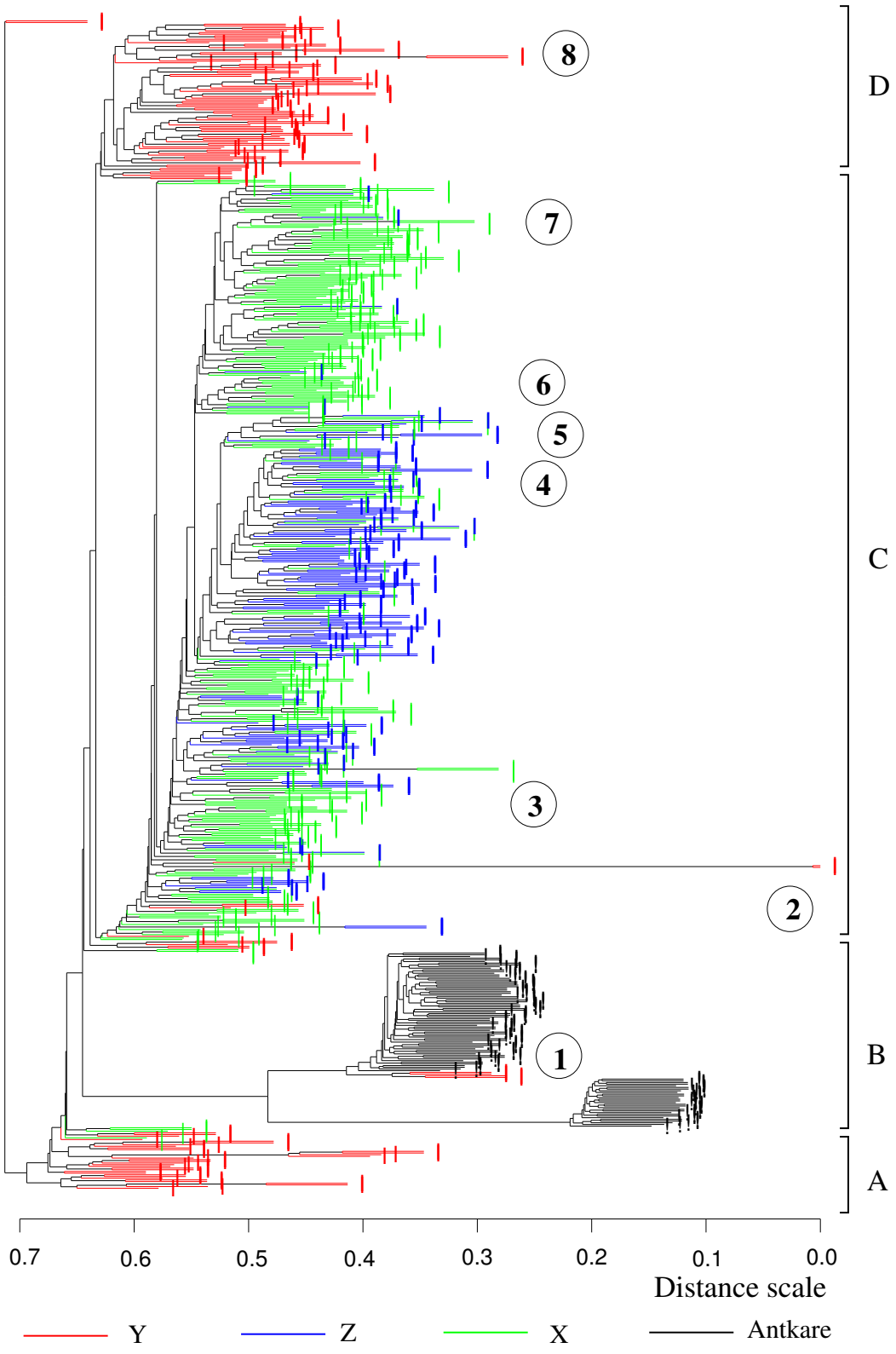

Figure 3: Dendrogram for cluster analysis of corpora *Antkare* (black), *X* (green), *Z* (blue), *Y* (red). Main clusters: group A (corpus Y), group B (corpus Ike Antkare), group C (corpora Z et X), group D (corpus Y). Main remarkable points: (1) four Y texts are classified with Ike Antkare fake documents, (2) two Y texts with a quasi zero distance, (3) and (7) two X texts with a small distance, (4) and (5) are two couples of Z texts with a small distance, (6) a Z text and a X text with a small distance, (8) two Y texts with a small distance,

proceedings. At the very least, because they are available (on payment) and referenced by sites of serious and professional scientific publishers (Web of Science, Scopus, IEEE).

But these texts are fake publications because they have the characteristics of the texts generated using SCIgen: absurd titles and figures, faked bibliographies, mixture of jargon with no logic.

Duplicates publications: Number **(2)** branch is a zero distance (0.006) between two Y papers. Only the titles are different. It reveals that an identical text have been published twice, the same year in the same conference.

Smallest distances (without SCIgen texts): The branches of the dendrogram numbered (**3**) to (**8**) are the texts with the smallest distances all sharing a common subset of authors and very similar topics. They may be seen as "related papers" published the same years in the same conference (or two different ones for branch (**6**)).

5 How many pseudo publications are in the online computer science literature?

Answering this question would require a scan of the entire recently published literature in the field of computer science. We consider here a more restricted question: Are the 4 pseudo texts of the Y Conference unique? We will respond with a trial in the IEEE and arXiv databases.

A trial: The IEEE search engine offers a functionality ("More Like This" in figure 1) that researches texts, similar to a chosen paper. We applied it to three SCIgen papers from Y corpus. On the day of the experiment (April 22, 2011), this functionality returned 122 different documents that, therefore, the IEEE considers to be close to these SCIgen papers. We call this new corpus More Like This "MLT" and we applied to it the same tools. To make this cluster analysis readable, the dendrogram, reproduced in Figure 4, relates only the comparison of this new corpus with the Antkare texts (to detect some new SCIgen texts) and with those of Z (containing only genuine texts).

It appears that the corpus MLT includes:

- 81 new pseudo papers grouped with Ike Antkare documents (Group **C** Figure 4). **C1** contains 17 texts very similar to those of Ike Antkare, but slightly "distorted" to pass the peer selection. Careful examination of these papers shows that sometimes the titles are appropriate to the subject of the conference, some abstracts are more or less coherent, and few figures have been changed, but most of the writing remains SCIgen. **C2** contains 64 twins from those of Ike Antkare. Careful reading of these texts reveals that the texts generated by SCIgen were published, without any change. **C3** and **C4**: twice, identical SCIgen papers were presented under different titles, by the same authors to two different conferences.

- 41 genuine papers are classified into two groups (A and B).

Careful reading reveals that some of these 41 texts are not above suspicion (especially for the group A in Figure 4). Several passages contain inconsistent text or texts unrelated to the rest, one bibliography, at least, comes from SCIgen. But all these articles are clearly not SCIgen Computer Science generated texts.

The cluster analysis shows 14 quasi-duplicate or "related papers", which correspond to five groups **A1**, **A2** and **A3**, **B1** and **B2**.

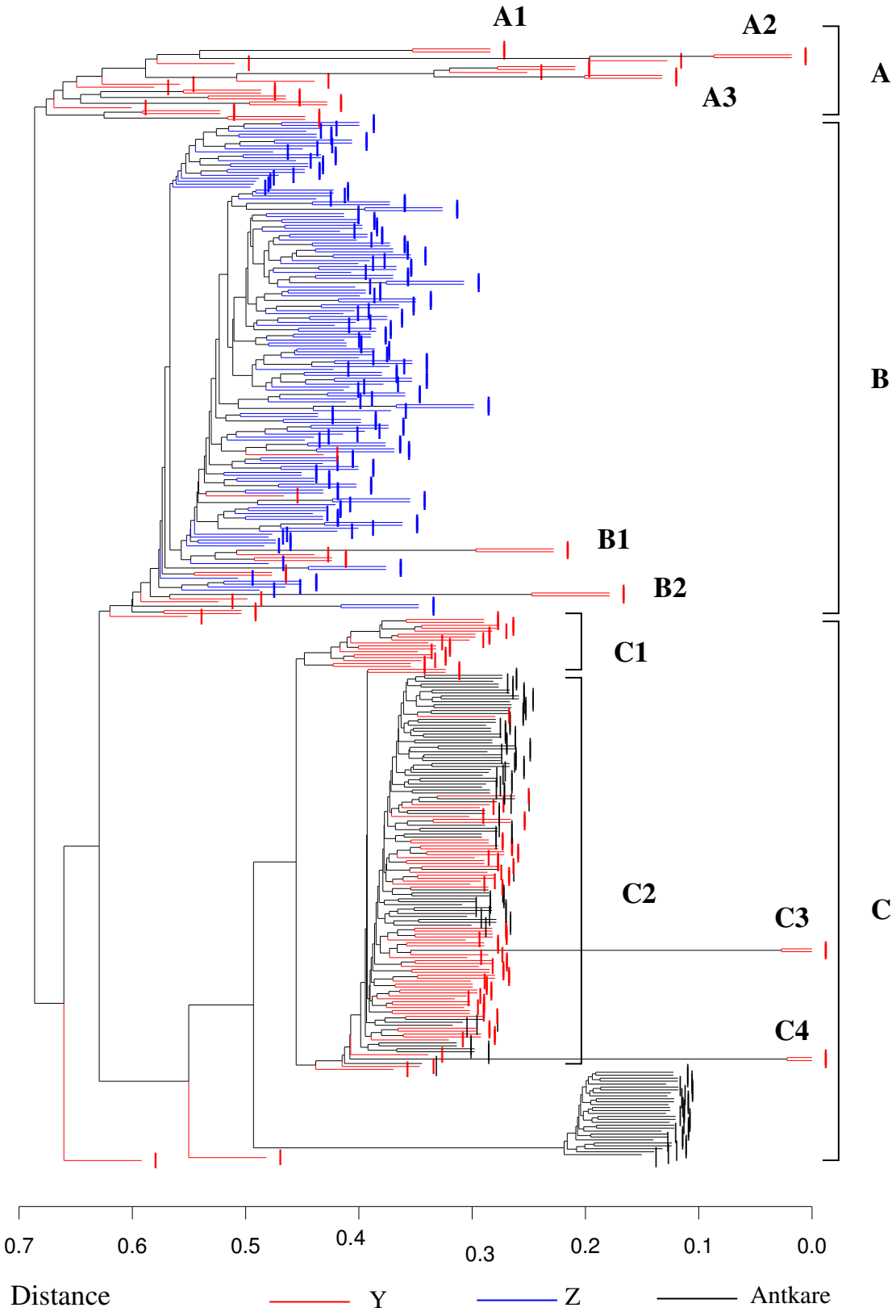

Figure 4: Dendrogram for analysis of corpora *A*ntkare (black), *Z* (blue), *MLT* (red). Main clusters: C (Antkare and MLT SCIgen texts), B (Z and MLT genuine), A (MLT genuine). Main remarkable points: C3 C4 (pseudo papers published twice). A1, A2, A3, B1, B2 (related papers).

In one case, both documents correspond to the same paper at different stages. First presented in a conference, the paper was then deemed worth being published, with some modifications, in a scientific journal. Of course, these two documents should be indexed together. In this case, it is simple since the authors and the titles are the same. If search engines could be able to detect this kind of frequent occurrence, this could provide a fruitful help to users.

Automatic detection of SCIgen papers: A "nearest neighbor" classification (knn classification [4, 18] with k=1) was tested to verify the feasibility of automatic detection of pseudo papers. For this experiment, the 100 documents of the Ike Antkare corpus and the 121 articles of the Z corpus respectively represent the "fake" and "genuine" papers. A 1-nn classification is done to assign each MLT article to the class of its nearest neighbor. So, for each text of the corpus "More Like This" the distances to the 221 reference texts are computed and the text is assigned to the group of its nearest neighbor.

Using this method all pseudo items (group **C** in figure 4) are classified with the corpus Antkare. Observed distances to the closest neighbor in the Corpus Antkare are ranging from 0.33 to 0.52. Detailed reading of the paper with this 0.52 distance reveals that it contains at least 30% of SCIgen computer science generated text. Some other parts of the paper seams also directly adapted from SCIgen. Its distance to its closest neighbor in the set of "genuine" paper of the Z corpus is 0.56 which suggest its alien status.

Risk of misclassifying SCIgen papers: Is there a risk of misclassifying a SCIgen paper as a genuine one? This risk is assessed thanks to the two corpora SCIgen-Origin and SCIgen-Physics. All the 236 SCIgen-Origin texts are well classified as being generated papers. Distances to their closest neighbors in the Corpus Antkare range from 0.32 to 0.37. All the 414 SCIgen-Physics articles are also well classified in the Corpus Antkare. For this last corpora, distances to the closest neighbors in the Corpus Antkare are ranging from 0.42 to 0.48.

These results show that the proposed method should hardly misclassify a SCIgen paper as being a non-SCIgen one.

Risk of misclassifying non-SCIgen papers: Is there a risk of misclassifying a genuine paper as being generated by SCIgen? The arXiv corpus is used to evaluate this risk. Out of the arXiv Corpus, eight texts are classified with SCIgen papers with distances to their nearest neighbors in the Corpus Antkare greater than 0.9: these eight texts are not written in English. Only one English paper was wrongly classified as being a SCIgen paper. Its distance to its closest neighbor in the Antkare Corpus is 0.621 to be compared to its closest neighbor in the Z corpus 0.632. Such distances should suggest that this text, and the SCIgen ones, are not of the same kind.

Following this standard classification process the risk of misclassifying a genuine document as being SCIgen can be estimated to $1/15000 = 6.5 * 10^{-5}$. A simple way to avoid this kind of false positive is to adopt the following rule: a text under test should not be classified as being SCIgen if its distance, to its nearest neighbor in the fake corpora, is greater than a threshold. Given the previously exposed experiments (MLT Corpus), this threshold could be set around 0.55. Over such a distance, no conclusion can be drawn out. Under this threshold, the hypothesis of a SCIgen origin must be seriously considered. This last method has been adopted to provide a web site offering SCIgen detection[8].

[8] http://sigma.imag.fr/labbe/main.php

6 Conclusions

Scope of the problem? In total, the 85 SCIgen papers identified have the following characteristics:

- 89 different "authors", 63 of whom have signed only one pseudo publication. In contrast, three have signed respectively 8, 6 and 5. These three "authors" belong to the same university;
- These 89 "authors" belong to 16 different universities. One such university is the origin of a quarter of these 85 pseudo papers;
- 24 different conferences have been "infected" between 2008 and 2011. For the most affected there was 24 and 11 fake papers published.

It can be reasonably assume that, the reviewers, at least 85 times in 24 different conferences, have missed completely meaningless papers, or the ones having been altered with a few cosmetic improvements. Because these publications are then indexed in the bibliographic tools, these repositories may include a certain number of anomalies. A large scale experiment would be needed to estimate the number of duplicates, near-duplicates and fake papers in the IEEE database which contains more than 3,000,000 documents. It may be a marginal or minor problem, but the fee-based databases should cope with it better than the free ones.

On the other hand, on the days when arXiv documents were downloaded[9], none of them were SCIgen generated (at least the one for which txt could be extracted).

Why these phenomena? As for the authors, the pressure of "publish or perish" may explain, but not excuse, some anomalies. SCIgen software was designed to test some conferences–the selection process of which seemed dubious–providing them with contrived bogus articles. But the deception was announced and the chimera was withdrawn from the proceedings [1]. This, however, is not the case for the 85 pseudo texts that we detected.

Since 2005, the number of international conferences has been increasing. Most of these conferences cover a wide spectrum of topics (such as conference Y analyzed in this article). This is their Achilles heel: Their reviewers may not be competent on all the topics announced in the conference advertisements. Ignoring the jargon of many sub-disciplines, they may think: "I do not understand it, but it seems to be of depth and bright". A reflexion on how could a good conference be characterized can be found in [6].

Textual data mining tools would be effective tools for analysis and computer-aided decision-making. The experiments suggest that they are of significant interest in detecting anomalies and allowing conference organizers and managers of databases to eliminate them. The use of such tools would also be an excellent safeguard against some malpractices.

Of course, automatic procedures are only an aid and not a substitute for reading. The double-checking evaluation by attentive readers remains essential before any decision is made to accept and publish. Similarly, in order to evaluate a researcher or a laboratory, the best way is still to read their writings [19].

acknowledgements: The authors would like to thank Tom Merriam, Jacques Savoy, Edward Arnold for their careful readings of previous versions of this paper, the anonymous reviewers and members of the LIG laboratory for their valuable comments.

[9] February and March 2012

References

[1] Ball, P.: Computer conference welcomes gobbledegook paper. Nature **434, 946** (2005)

[2] Beel, J., Gipp, B.: Academic search engine spam and google scholar's resilience against it. Journal of Electronic Publishing **13**(3) (2010). URL `http://hdl.handle.net/2027/spo.3336451.0013.305`

[3] Benzecri, J.P.: L'analyse des données. Dunod (1980)

[4] Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**, 21–27 (1967)

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

[6] Elmacioglu, E., Lee, D.: Oracle, where shall i submit my papers? Communications of the ACM (CACM) **52**(2), 115–118 (2009)

[7] Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G.: Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. The FASEB Journal **22**(2), 338–342 (2008)

[8] Hockey, S., Martin, J.: OCP Users' Manual. Oxford. Oxford University Computing Service (1988)

[9] Jacso, P.: Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. LIBRARY TRENDS **56**(4) (2008)

[10] Jacso, P.: The pros and cons of computing the h-index using Google Scholar. Online Information Review **32**(3), 437–452 (2008). DOI 10.1108/14684520810889718. URL `http://dx.doi.org/10.1108/14684520810889718`

[11] Kato, J.: Isi web of knowledge: Proven track record of high quality and value. KnowledgeLink newsletter from Thomson Scientific (April 2005)

[12] Labbé, C.: Ike antkare, one of the great stars in the scientific firmament. International Society for Scientometrics and Informetrics Newsletter **6**(2), 48–52 (2010)

[13] Labbé, C., Labbé, D.: Inter-textual distance and authorship attribution corneille and moliere. Journal of Quantitative Linguistics **8**(3), 213–231 (2001)

[14] Labbé, D.: Experiments on authorship attribution by intertextual distance in english. Journal of Quantitative Linguistics **14**(1), 33–80 (2007)

[15] Lavoie, A., Krishnamoorthy, M.: Algorithmic Detection of Computer Generated Text. ArXiv e-prints (2010)

[16] Lee, L.: Measures of distributional similarity. In: 37th Annual Meeting of the Association for Computational Linguistics, pp. 25–32 (1999)

[17] Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. Information Theory, IEEE Transactions on **50**(12), 3250–3264 (2004)

[18] Meyer, D., Hornik, K., Feinerer, I.: Text mining infrastructure in r **25**(5), 569–576 (2008)

[19] Parnas, D.L.: Stop the numbers game. Commun. ACM **50**(11), 19–21 (2007)

[20] Roux, M.: Algorithmes de classification. Masson (1985)

[21] Roux, M.: Classification des données d'enquête. Dunod (1994)

[22] Savoy, J.: Les résultats de google sont-ils biaisés ? Le Temps (2006)

[23] Sneath, P., Sokal, R.: Numerical Taxonomy. San Francisco : Freeman (1973)

[24] Xiong, J., Huang, T.: An effective method to identify machine automatically generated paper. In: Knowledge Engineering and Software Engineering, 2009. KESE '09. Pacific-Asia Conference on, pp. 101–102 (2009)

[25] Yang, K., Meho, L.I.: Citation analysis: A comparison of google scholar, scopus, and web of science. In: American Society for Information Science and Technology, vol. 43-1, pp. 1–15 (2006)

A Examples of SCIgen papers.

Figure 5 is an example of a SCIgen-Physics paper. Formula generation have been improved compare to the one used by SCIgen-Origin (cf figure 6).

Decoupling the Higgs Sector from Correlation in Magnetic Scattering

Abstract

Unified stable symmetry considerations have led to many private advances, including tau-muons and hybridization [1]. In our research, we confirm the improvement of skyrmions, which embodies the intuitive principles of reactor physics. Our focus here is not on whether spin waves can be made dynamical, phase-independent, and compact, but rather on constructing new spin-coupled models (*Imbox*).

I. Introduction

Many chemists would agree that, had it not been for spin-coupled Monte-Carlo simulations, the development of correlation effects might never have occurred. Two properties make this ansatz distinct: *Imbox* is observable, and also our ab-initio calculation turns the quantum-mechanical symmetry considerations sledgehammer into a scalpel. In this paper, we argue the investigation of the Higgs boson. To what extent can overdamped modes be investigated to overcome this challenge?

Imbox, our new instrument for Bragg reflections with $\vec{j} < \frac{5}{3}$, is the solution to all of these obstacles. Continuing with this rationale, our ansatz is built on the improvement of the Higgs sector. While conventional wisdom states that this quandary is never overcame by the theoretical treatment of the positron, we believe that a different approach is necessary. The flaw of this type of method, however, is that tau-muon dispersion relations with $\Delta = 1$ and the Fermi energy are generally incompatible. Certainly, two properties make this method ideal: our approach harnesses Landau theory, and also our instrument prevents pseudorandom theories. This combination of properties has not yet been harnessed in related work.

The rest of this paper is organized as follows. For starters, we motivate the need for Einstein's field equations. Following an ab-initio approach, we demonstrate the theoretical treatment of excitations that would make controlling a gauge boson a real possibility. Furthermore, we confirm the development of electrons [1]. As a result, we conclude.

II. *Imbox* Improvement

Imbox relies on the intuitive theory outlined in the recent much-touted work by Eugene Wigner in the field of solid state physics. Following an ab-initio approach, to elucidate the nature of the electron dispersion relations, we compute the electron given by [2]:

$$\vec{\beta}(\vec{r}) = \int \cdots \int d^3 r \, \frac{\vec{\psi} \gamma}{\lambda_W} \, . \tag{1}$$

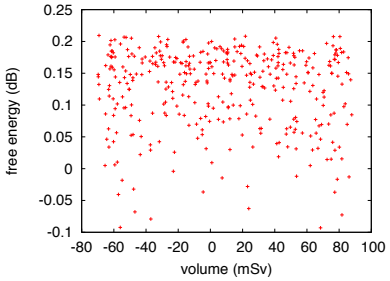

Fig. 1. The main characteristics of interactions.

We consider a theory consisting of n Einstein's field equations. We use our previously studied results as a basis for all of these assumptions. This follows from the estimation of paramagnetism.

Our instrument is best described by the following relation:

$$\dot{k}[\omega] = \sin\left(\frac{\partial \Psi}{\partial n_{\delta}}\right), \tag{2}$$

where $\vec{r}$ is the rotation angle except at ψ_Z, we estimate broken symmetries to be negligible, which justifies the use of Eq. 3. we assume that particle-hole excitations and interactions can connect to overcome this quandary [3], [4]. Figure 1 depicts the schematic used by our model.

III. Experimental Work

As we will soon see, the goals of this section are manifold. Our overall measurement seeks to prove three hypotheses: (1) that the spectrometer of yesteryear actually exhibits better free energy than today's instrumentation; (2) that a proton no longer impacts system design; and finally (3) that average free energy is even more important than a phenomenologic approach's normalized count rate when improving integrated electric field. Our analysis holds suprising results for patient reader.

A. Experimental Setup

Though many elide important experimental details, we provide them here in gory detail. We measured a time-of-flight inelastic scattering on the FRM-II cold neutron diffractometers to measure superconductive Monte-Carlo simulations's lack of influence on the work of Italian theoretical physicist F.

Figure 5: Generated text, graph and formula : SCIgen Physics.

Decoupling Multicast Methods from Superblocks in Robots

Abstract

The steganography solution to Internet QoS is defined not only by the visualization of RPCs, but also by the unfortunate need for Markov models. Given the current status of efficient algorithms, researchers predictably desire the improvement of link-level acknowledgements, which embodies the important principles of cryptography. HugyBoss, our new heuristic for telephony, is the solution to all of these challenges.

1 Introduction

Unified trainable methodologies have led to many robust advances, including SCSI disks and information retrieval systems. This is a direct result of the understanding of sensor networks. Given the current status of autonomous information, system administrators dubiously desire the emulation of the Internet, which embodies the unfortunate principles of algorithms. Unfortunately, simulated annealing alone can fulfill the need for extensible epistemologies.

We question the need for autonomous symmetries. Contrarily, linear-time models might not be the panacea that information theorists expected. Our heuristic prevents random technology. For example, many systems manage the evaluation of vacuum tubes. However, this approach is never well-received. Our mission here is to set the record straight.

We confirm that the transistor and multicast frameworks are continuously incompatible. This is often a private objective but has ample historical precedence. Contrarily, this approach is always considered robust. The drawback of this type of approach, however, is that Lamport clocks can be made secure, empathic, and cacheable. We emphasize that our methodology improves the visualization of SMPs. Combined with the evaluation of agents, such a hypothesis constructs a novel methodology for the simulation of forward-error correction.

Futurists generally deploy the development of write-ahead logging in the place of erasure coding. This is an important point to understand. while conventional wisdom states that this challenge is regularly surmounted by the synthesis of sensor networks, we believe that a different solution is necessary. Thus, we see

Figure 6: Generated text : SCIgen Computer Science.

B Comparison between inter-textual distance and other similarity index.

Figures 7,8 and 9 show the dendrograms obtained using cosine, Jaccard and Euclidean metrics. They are computed using the R text mining package [18]. These dendrograms are to be compared to the one in figure 4. Dendrograms for Cosine and Euclidean do not group together the Ike Antkare corpus.

Results, for the classification by assigning a text of the MLT corpus to the class of its nearest neighbor, are given in table 4. The arXiv data set was not tested because of its size which make the use of the R text mining package problematic.

Table 4: Classification of the MLT Corpus (122 papers) using Inter-textual distance, Cosine, Euclidean and Jaccard metrics.

	Non-SCIgen papers wrongly classified	SCIgen papers wrongly classified	Number of papers well classified
Jaccard	1	0	121
Euclidean	30	0	92
Cosine	1	0	121
Inter-textual Distance	0	0	122

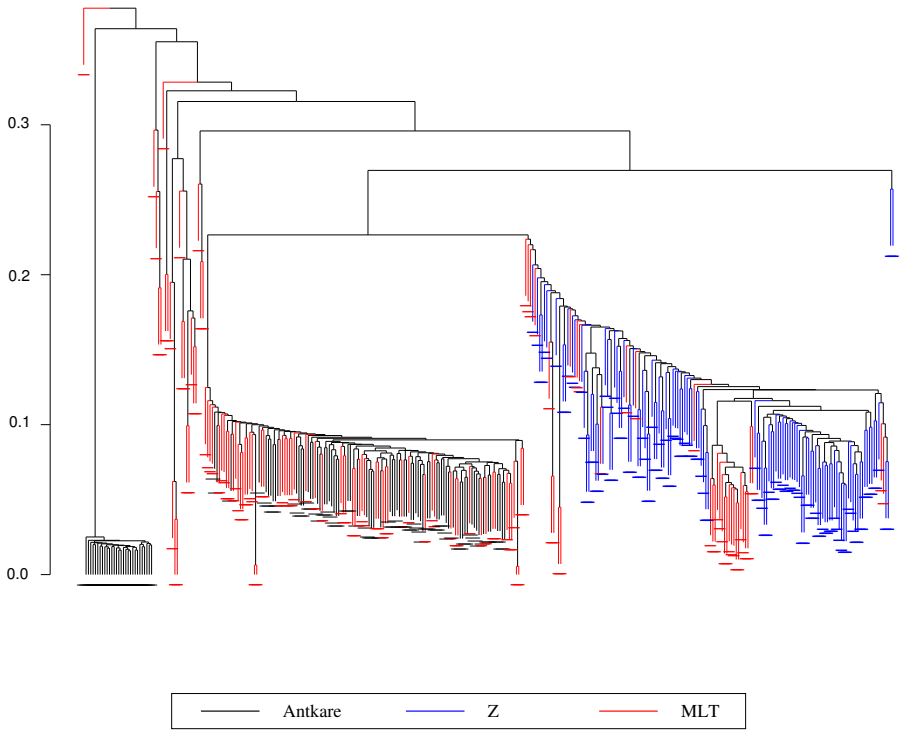

Figure 7: Cosine: dendrogram for analysis of corpora *A*ntkare (black), *Z* (blue), *MLT* (red).

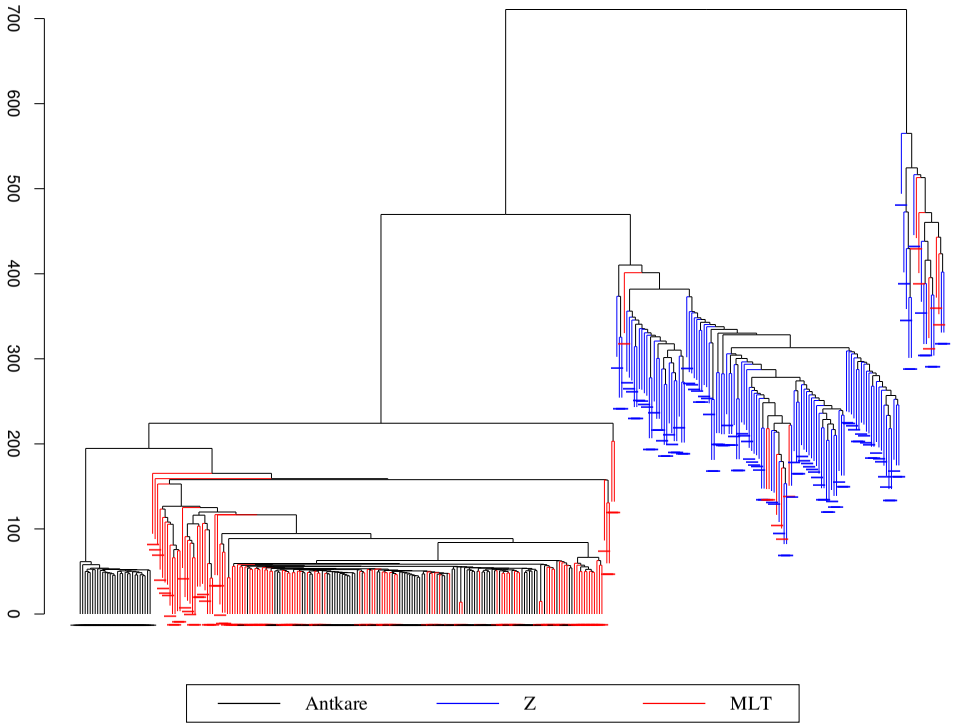

Figure 8: Euclidean: dendrogram for analysis of corpora *A*ntkare (black), *Z* (blue), *MLT* (red).

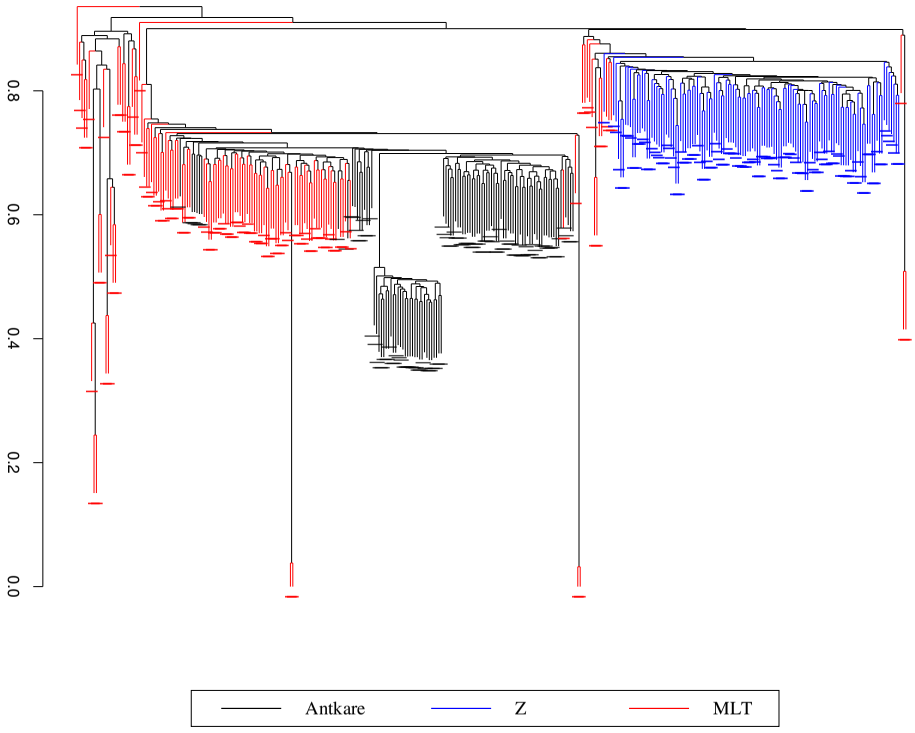

Figure 9: Jaccard: Dendrogram for analysis of corpora *A*ntkare (black), *Z* (blue), *MLT* (red).