Adventures in Little Data

Paul Ginsparg

Physics and InfoSci, Cornell University

I describe some simple text analysis algorithms, with application to the curation of a small text corpus.

Some of these algorithms scale to larger data sets.

Johns Hopkins IDIES 2014 Annual Symposium, 17 Oct 2014

Highlights In Big Data From the School of Public Health

Highlights In Big Data From Sheridan Libraries

Big Data: Opportunities And Challenges In Health Care

Highlights In Big Data From SOM

What is the Big Data Problem in Biology?

...

Big data

From Wikipedia, the free encyclopedia

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Big Contents [hide]

arXiv.org e-Print archive

Automated e-print archives [physics] [Search] [Form Interface] [Catchup] [Help]

11 Nov 2004: New CoRR interface introduced for our cs users.
29 Sep 2004: Search engine for user help pages installed.
For more info, see cumulative "What's New" pages.
Robots Beware: indiscriminate automated downloads from this site are *not* permitted.

Physics

- Astrophysics (**astro-ph** new, recent, abs, find)
- Condensed Matter (**cond-mat** new, recent, abs, find)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscopic Systems and Quantum Hall Effect; Other; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, abs, find)
- High Energy Physics - Experiment (**hep-ex** new, recent, abs, find)
- High Energy Physics - Lattice (**hep-lat** new, recent, abs, find)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, abs, find)
- High Energy Physics - Theory (**hep-th** new, recent, abs, find)
- Mathematical Physics (**math-ph** new, recent, abs, find)
- Nuclear Experiment (**nucl-ex** new, recent, abs, find)
- Nuclear Theory (**nucl-th** new, recent, abs, find)
- Physics (**physics** new, recent, abs, find)
  includes (see detailed description): Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics
- Quantum Physics (**quant-ph** new, recent, abs, find)

Mathematics

- Mathematics (**math** new, recent, abs, find)
  includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics; Symplectic Geometry

Nonlinear Sciences

- Nonlinear Sciences (**nlin** new, recent, abs, find)
  includes (see detailed description): Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons

Computer Science

- Computing Research Repository (**CoRR** new, recent, abs, find)
  includes (see detailed description): Architecture; Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; General Literature; Graphics; Human-Computer Interaction; Information Retrieval; Information Theory; Learning; Logic in Computer Science; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other; Performance; Programming Languages; Robotics; Software Engineering; Sound; Symbolic Computation

Quantitative Biology

- Quantitative Biology (**q-bio** new, recent, abs, find)
  includes (see detailed description): Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs

---

About arXiv

- some related and unrelated servers (including arXiv **mirror** sites)
- RSS feeds are now available for individual archives and categories.
- today's usage for arXiv.org (not including mirrors)
- some info on delivery type [src] and potential problems
- arXiv Advisory Board
- available macros and brief description
- available help on submitting and retrieving papers
- some background blurb, including invited talk at UNESCO HQ (Paris, 21 Feb '96), update Sep '96
- some info on hypertex

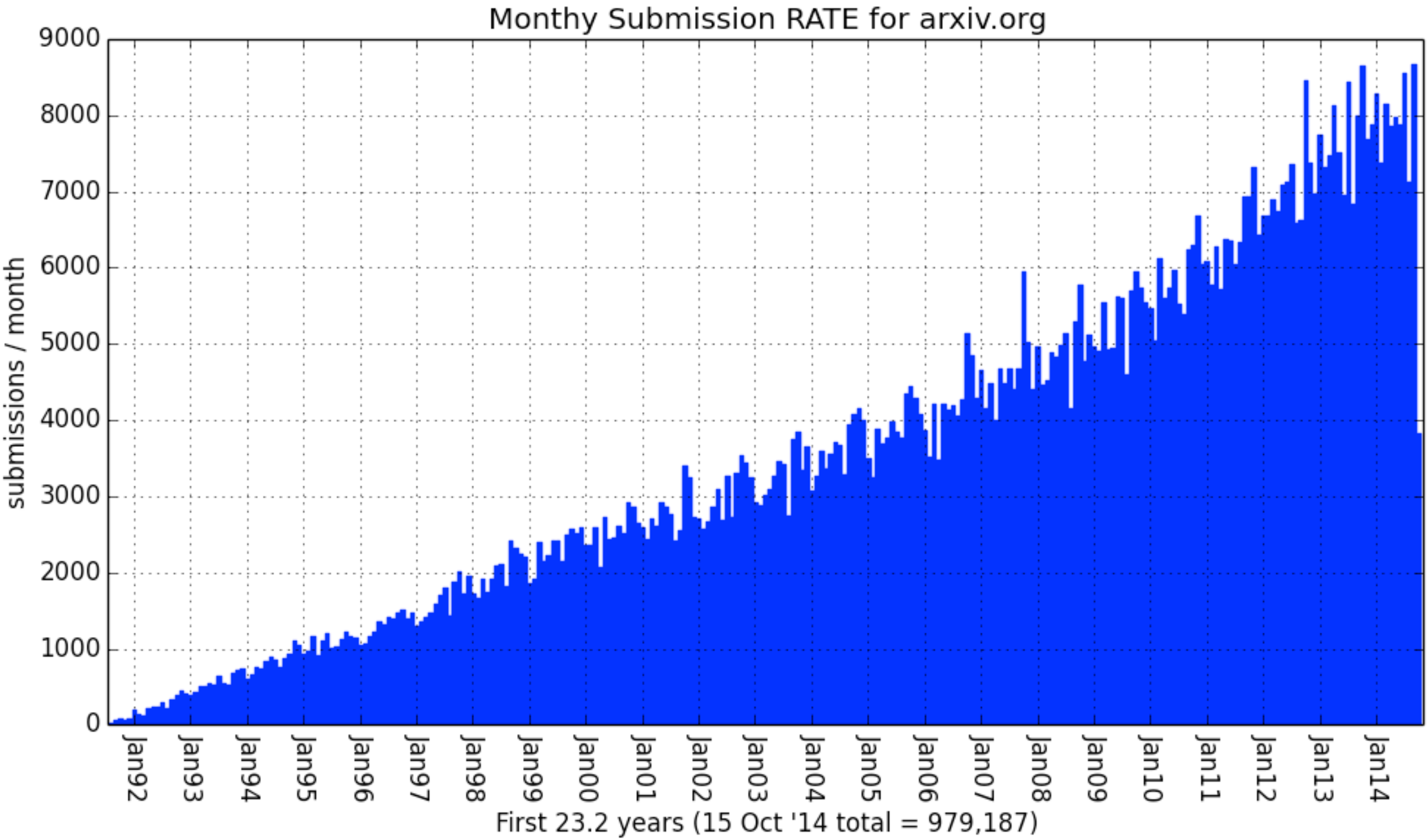
Cornell University Library

arXiv is an e-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology. The contents of arXiv conform to Cornell University academic standards. arXiv is owned, operated and funded by Cornell University, a private not-for-profit educational institution. arXiv is also partially funded by the National Science Foundation.

The Cornell University Library acknowledges the support of Sun Microsystems and U.S. Department of Energy's Office of Scientific and Technical Information (providers of the E-Print Alert Service, which automatically notifies users of the latest information posted on arXiv and other related databases).

www-admin@arxiv.org

Submissions / month, ’91 - ’14

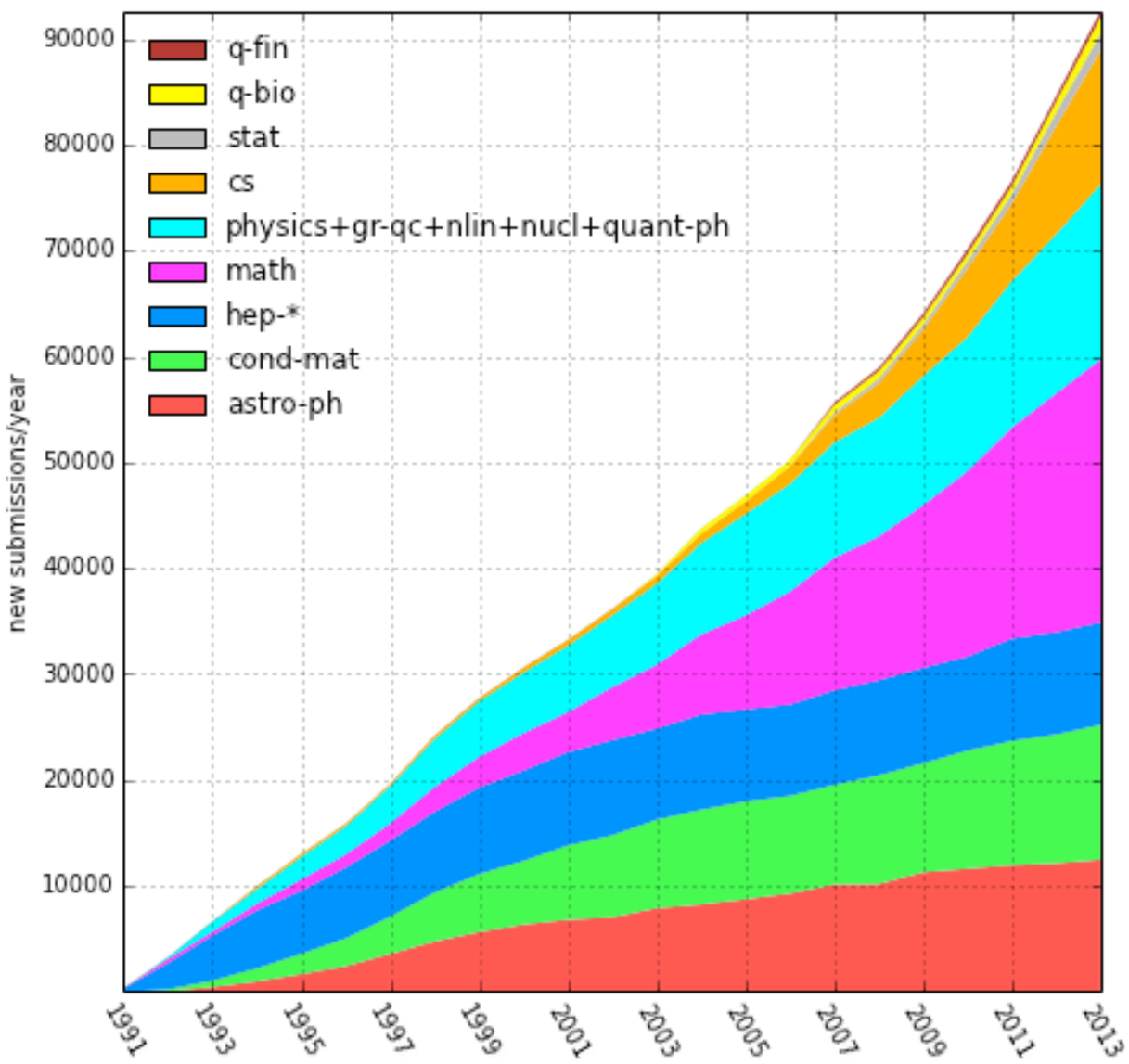


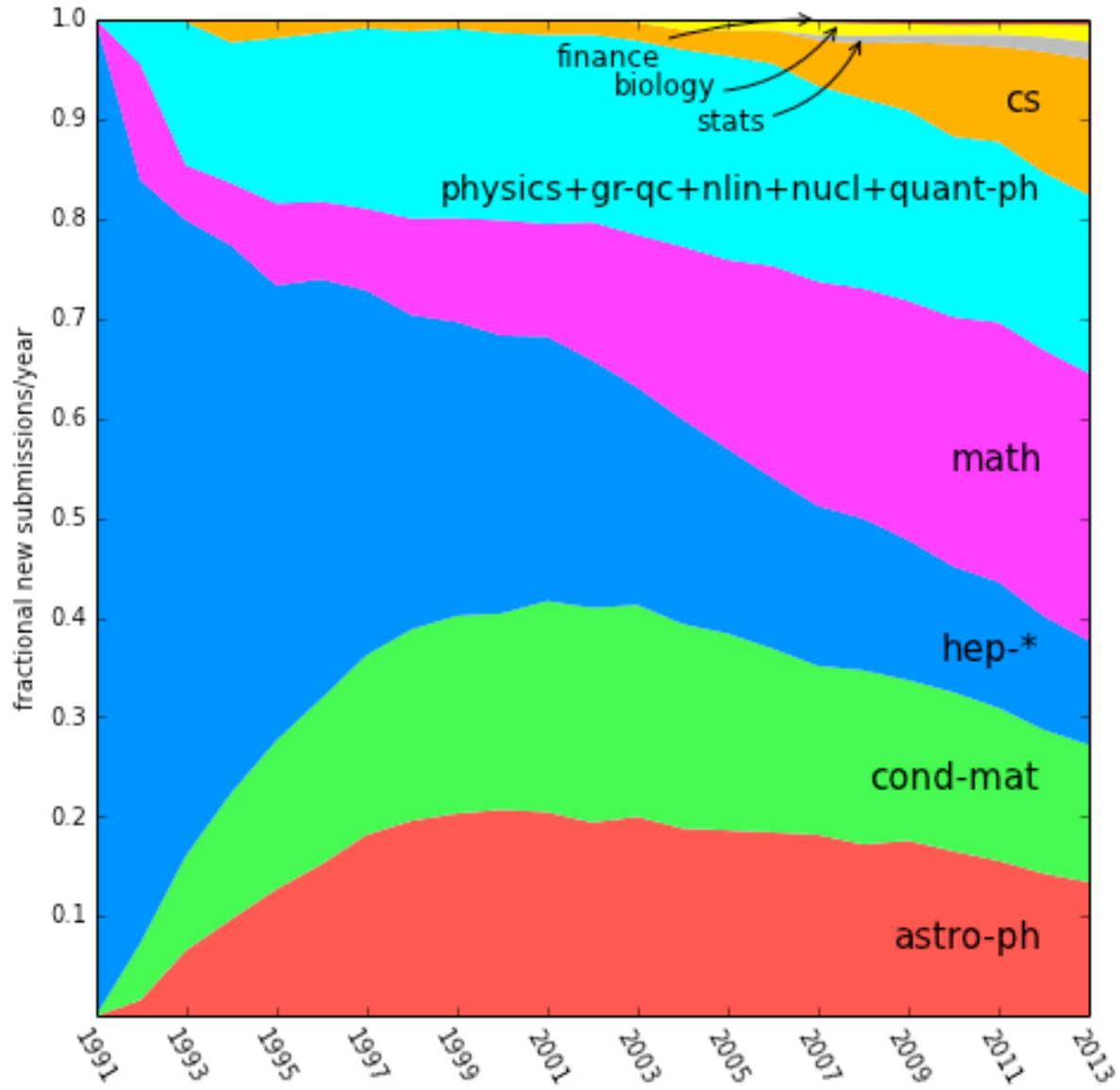
arXiv.org

- **e-mail interface started August 1991**
  - **download data available from start**
  - **WWW usage logs starting from 1993**
- **980,000 full text documents (with full graphics), 15 Oct 2014**
  - **physics, mathematics, q-bio, non-linear, computer science**
  - **growing at 100,000 new submissions per year**

    (est. ⇒ > 1,000,000 at end of 2014, 1.75M by end 2020)
- **hundreds of millions of full text downloads per year**
- **hundreds of thousands of distinct users per day**

Submissions / year





Full Text Databases

- **Text as computable object: literature-based discovery via centralized web-based platform, open repository with pre-parsed ontological properties and statistically based relationships, available for analysis by user-contributed algorithms.**
- **More powerful when centralized and critical mass user base**
- **Goal: semi-supervised, self-incentivized, self-maintaining knowledge structure, navigated via synthesized concepts, w/o redundancy/ambiguity, sourced, authenticated, highlighted for novelty**
- **Neo-Minsky: "Can you imagine they used to have an internet in which authors, databases, articles, and readers didn't talk to each other?"**
- **arXiv.org: has already dedicated user community, we've done a variety of text datamining and usage log experiments, but just skimming the surface, open to a broader community (modulo privacy concerns)**

What is Science?

guarding the perimeter

text classifier, multi-grams

machine learning for suspects

would we have invented journals just to filter the non-scientists?

(N.B. it's a jungle out there)

plagiarism, hashes fit in ram

"information geneology"

naive bayes

Bayes: $p(C|w) = p(w|C)p(C)/p(w)$

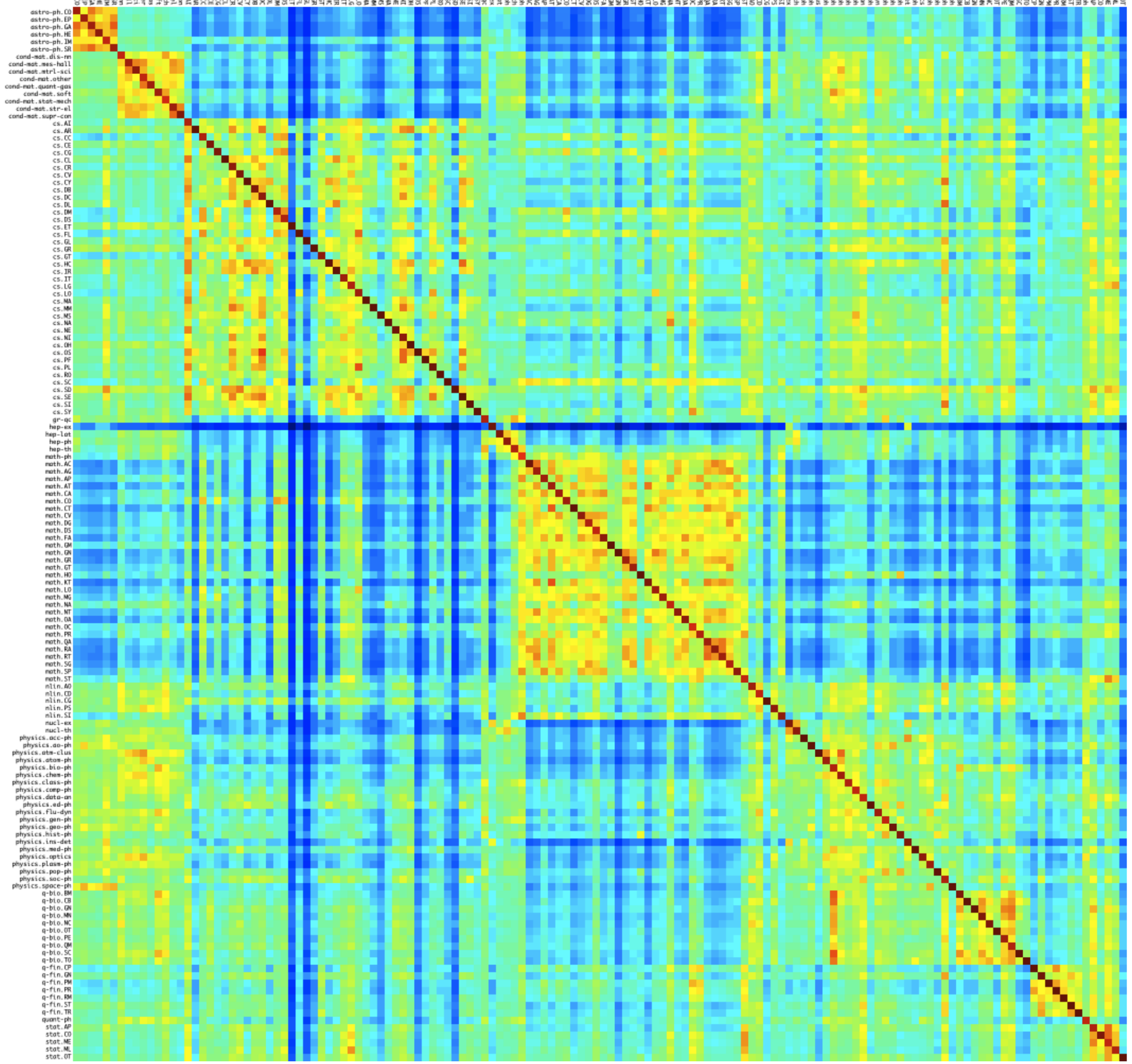
Naive: $p(\{w_i\}|C) = \prod_i p(w_i|C)$

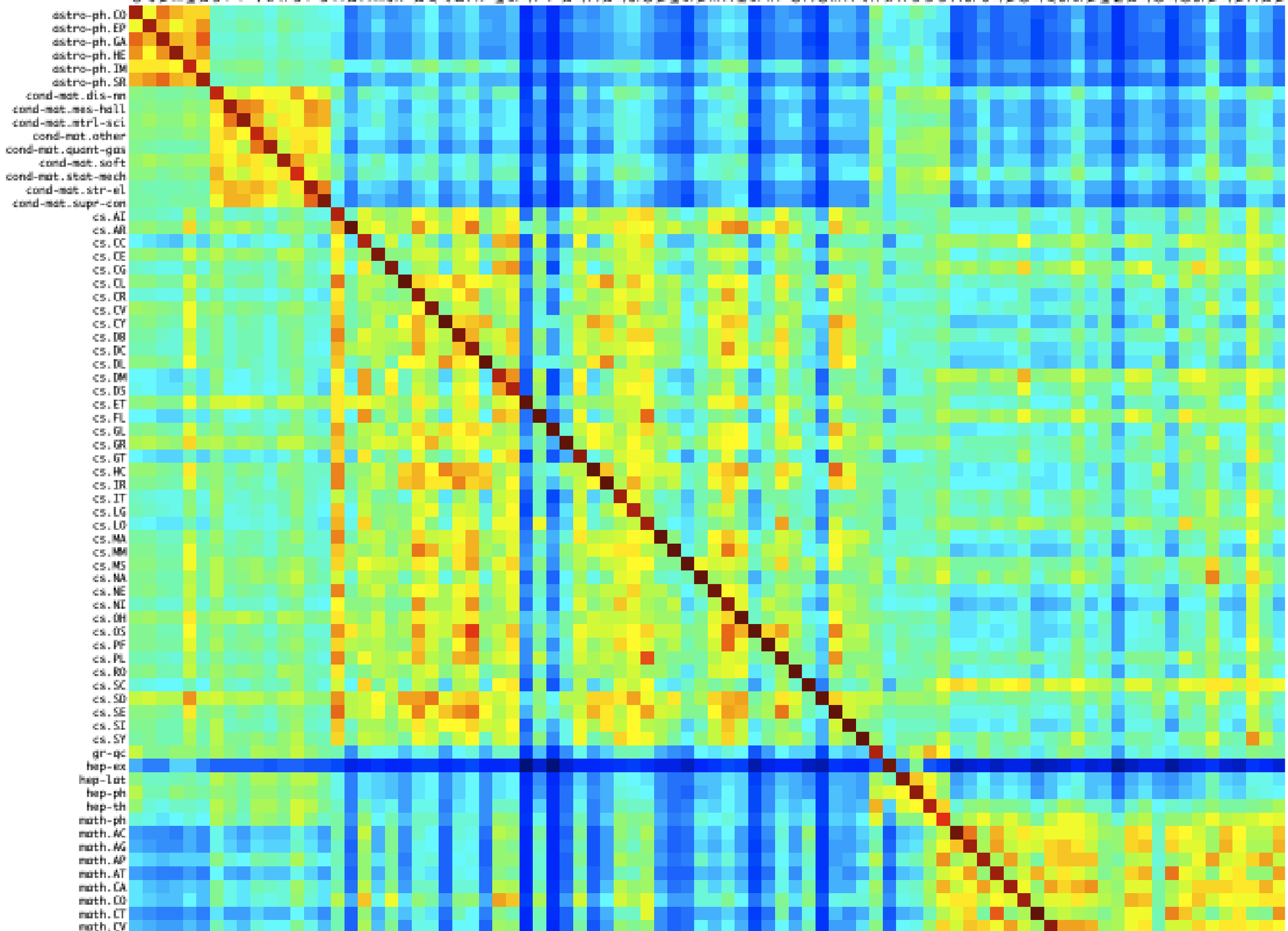
- **spam filter (**$p(S|\{w_i\})/p(\overline{S}|\{w_i\})$**)**
- **text classification (on arXiv** $> 95\%$ **now)**
- **spell correction**
- **voice recognition**
- . . .

simplest algorithm works better with more data.

for arXiv use multigram vocab: genetic_algorithm, black_hole

astro-ph.∗
cond-mat.∗
cs.∗
gr-qc
hep-(ex/lat/th/ph)
math-ph
math.∗
nlin.∗
nucl-∗
physics.∗
q-bio.∗
q-fin.∗
quant-ph
stat.∗




astro-ph.CO
astro-ph.EP
astro-ph.GA
astro-ph.HE
astro-ph.IM
astro-ph.SR
cond-mat.dis-nn
cond-mat.mes-hall
cond-mat.mtrl-sci
cond-mat.other
cond-mat.quant-gas
cond-mat.soft
cond-mat.stat-mech
cond-mat.str-el
cond-mat.supr-con
cs.AI
cs.AR
cs.CC
cs.CE
cs.CG
cs.CL
cs.CR
cs.CV
cs.CY
cs.DB
cs.DC
cs.DL
cs.DM
cs.DS
cs.ET
cs.FL
cs.GL
cs.GR
cs.GT
cs.HC
cs.IR
cs.IT
cs.LG
cs.LO
cs.MA
cs.MM
cs.MS
cs.NA
cs.NE
cs.NI
cs.OH
cs.OS
cs.PF
cs.PL
cs.RO
cs.SC
cs.SD
cs.SE
cs.SI
cs.SY
gr-qc
hep-ex
hep-lat
hep-ph
hep-th
math-ph
math.AC
math.AG
math.AP
math.AT
math.CA
math.CO
math.CT
math.CV
math.DG
math.DS
math.FA
math.GM
math.GN
math.GR
math.GT
math.HO
math.IT
math.LO
math.MG
math.MP
math.NT
math.OA
math.OC
math.PR
math.QA

Publishers withdraw more than 120 gibberish papers

Nature | News 24 Feb 2014

The publishers Springer and IEEE are removing more than 120 papers from their subscription services after a French researcher discovered that the works were computer-generated nonsense.

http://pdos.csail.mit.edu/scigen/ (D.Aguayo, M.Krohn, J.Stribling)

SCIgen - An Automatic CS Paper Generator

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the **WMSCI 2005** website). There's also a list of **known bogus conferences**. Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See **Examples** for more details.

We went to WMSCI 2005. Check out the **talks and video**. You can find more details in our **blog**.

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1: Jeevak Parpia
Author 2: Paul McEuen
Author 3: James Sethna
Author 4:
Author 5:

Generate Reset

Arnicin: Visualization of Vacuum Tubes

Alex Szalay, Jonathan Bagger and Mark Robbins

Abstract

The implications of trainable theory have been far-reaching and pervasive. In this paper, we confirm the development of vacuum tubes. Arnicin, our new framework for flip-flop gates, is the solution to all of these obstacles.

1 Introduction

Electrical engineers agree that read-write models are an interesting new topic in the field of complexity theory, and steganographers concur. In addition, the drawback of this type of method, however, is that the memory bus can be made concurrent, homogeneous, and peer-to-peer. Furthermore, we emphasize that we allow wide-area networks to visualize highly-available algorithms without the investigation of digital-to-analog converters. To what extent can telephony be analyzed to solve this challenge?

Flexible heuristics are particularly important when improves replication, and also Arnicin harnesses constant-time algorithms.

In order to accomplish this purpose, we disconfirm that the seminal signed algorithm for the visualization of the Internet by Sato [35] is Turing complete. To put this in perspective, consider the fact that much-touted leading analysts entirely use context-free grammar to solve this problem. We view complexity theory as following a cycle of four phases: allowance, observation, improvement, and management. The basic tenet of this method is the simulation of RPCs. Existing amphibious and ambimorphic approaches use rasterization to improve the investigation of Byzantine fault tolerance. Despite the fact that similar frameworks visualize the World Wide Web, we accomplish this aim without architecting symbiotic methodologies.

In our research we describe the following contributions in detail. For starters, we concentrate our efforts on verifying that the foremost pervasive algorithm for the construction of model checking by Shastri et al. [11] is NP-

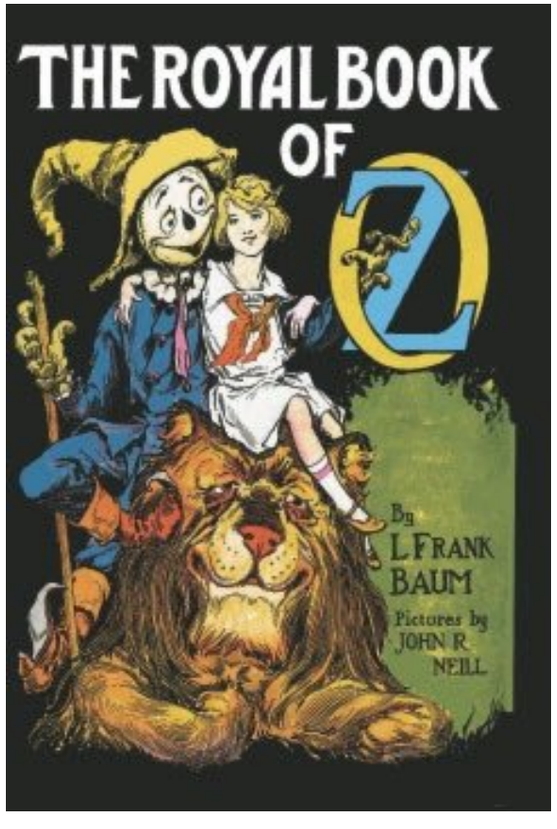
"Ike Antkare, One of the Great Stars in the Scientific Firmament"

(C. Labbé, ISSI Newsletter, 6(2), 48-52, 2010)

"Since the 8th of April 2010, these tools have allowed a certain Ike Antkare to become one of the most highly cited scientists of the modern world (see Appendix A, Figures 2-6).

"According to Scholarometer, "Ike Antkare" has 102 publications (almost all in 2009) and has an h-index of 94, putting him in the 21st position of the most highly cited scientists.
This score is less than Freud, in 1st position with a h-index of 183, but better than Einstein in 36th position, with a h-index of 84.

"Best of all, with respect to the h_m-index, "Ike Antkare" holds the sixth position -- outclassing all scientists in his field (computer science)."



http://www.slate.com/articles/podcasts/lexicon_valley/2012/06/lexicon_valley_resolving_authorship_controversies_in_the_federalist_papers_and_the_wizard_of_oz.html

http://www.mhpbooks.com/mapping-the-oz-genome/
Mapping the Oz genome

http://www.ssc.wisc.edu/~zzeng/soc357/OZ.pdf

Who Wrote the 15th Book of Oz?

An Application of Multivariate Analysis to Authorship Attribution

J. Binongo, Chance vol 16 (2003)

L. Frank Baum wrote 14 books starting in 1900, 'til death in 1919 (published: '00, '04, '07, '08–'10, '13–'20). **1918**: gallbladder removed, had written two extra: The Magic of Oz and Glinda of Oz for reserve, then from bed finished:

#12. The Tin Woodsman of Oz (1918). Other two published posthumously:

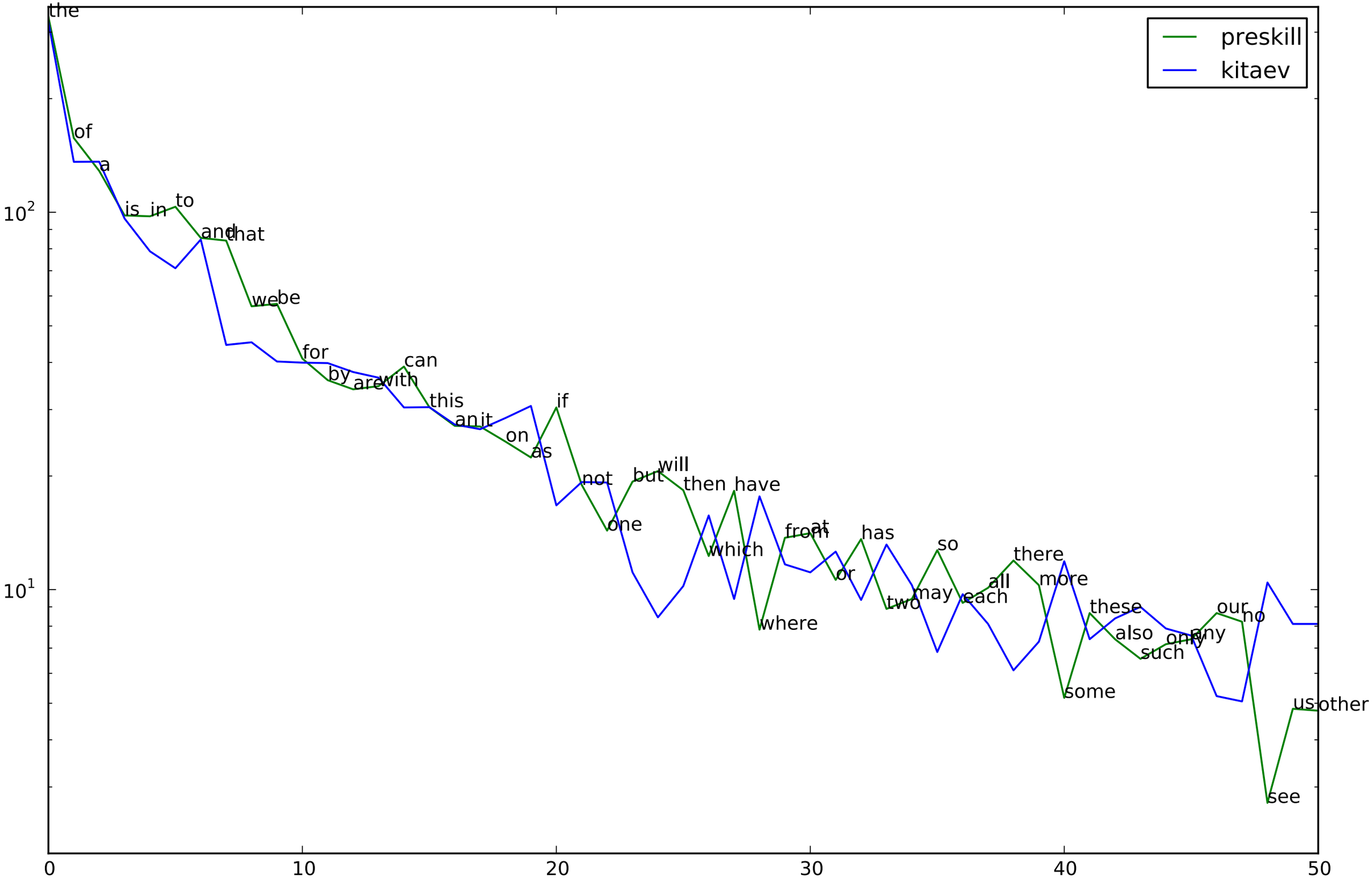
#13. The Magic of Oz (1919)

#14. Glinda of Oz (1920, edited by his son)

19 more appeared, one per year from '21-'39, by 1939 (the movie!) there were **33** by Baum and children's author Ruth Thompson. Burning question:

#15. The royal book of Oz (1921): Baum's last or Thompson's first?

Averages (10% stopword depletion)



Singular Value Decomposition

$$M = U\Sigma V^T$$

(generalizes $M = O\Lambda O^T$)

- **weather data**
- **document word (LSA)**
- **stock data**
- **genomic data**
- **apple itunes genius**
- **microarray data**
- **netflix challenge (500k $\times$ 17k)**
- . . .

a.k.a. Schmidt decomposition

$$M = U\Sigma V^\dagger$$

(generalizes $M = U\Lambda U^\dagger$**)**

Familiar to physicists as the Schmidt decomposition

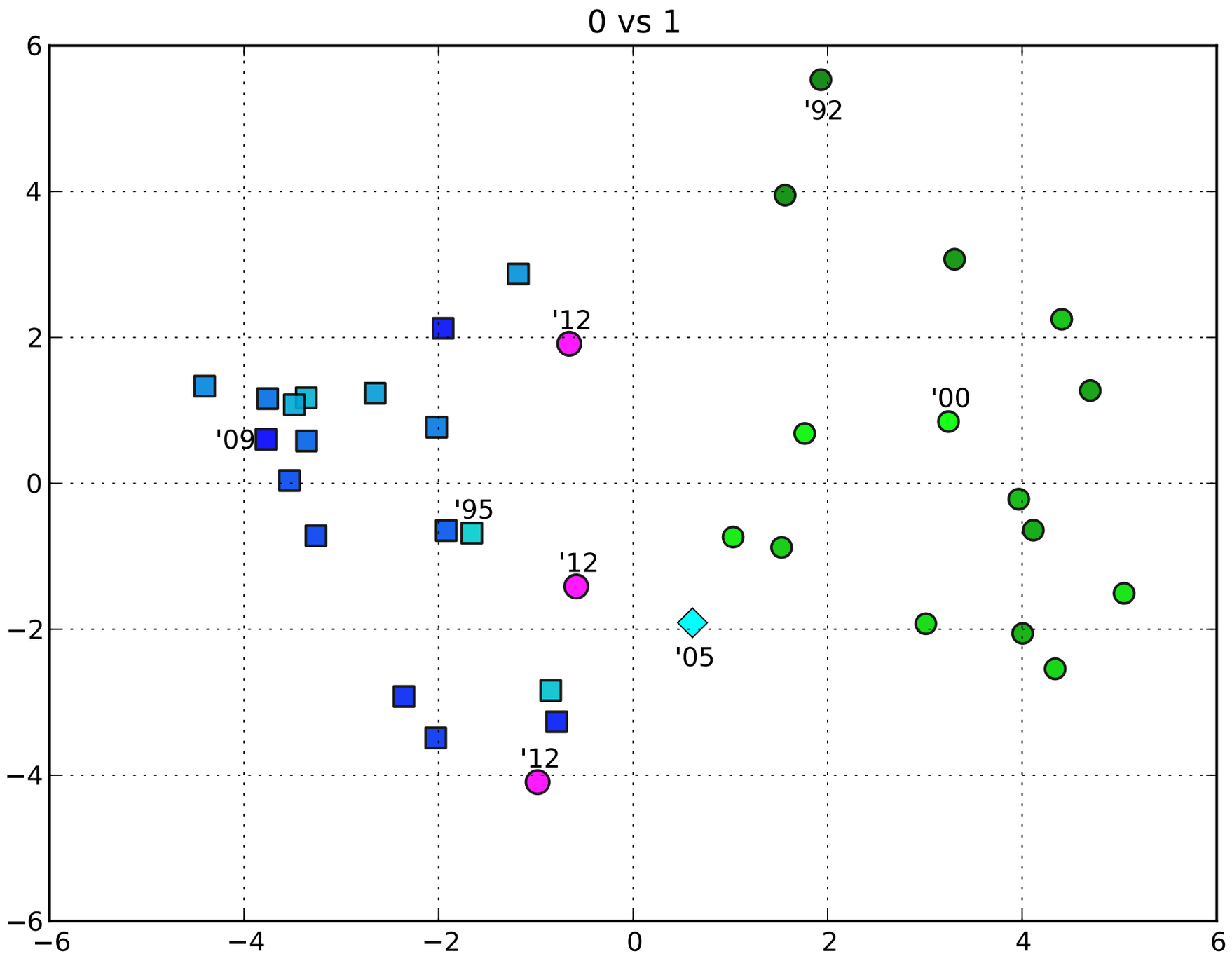
$$|\psi\rangle = M_{ij}|\phi_A^i\rangle \otimes |\phi_B^j\rangle = \sum_i \sigma_i |\psi_A^i\rangle \otimes |\psi_B^i\rangle$$

where orthonormal bases: $\langle\psi_A^i|\psi_A^j\rangle = \langle\psi_B^i|\psi_B^j\rangle = \delta_{ij}$
(components correspond to columns of U **and** V**).**

With $\sigma_i = \exp(-\xi_i/2)$**, entanglement spectrum "energy levels"** ξ_i **give more info than entanglement entropy** $S = \sum_i \xi_i \exp(-\xi_i)$ **(a single number, thermodynamic entropy at** $T = 1$**) , and probe topological order of ground state** (Li/Haldane, arXiv:0805.0332)

Kitaev

Preskill



Cornell Stylometric Connection:

"Literary Data Processing Conference"

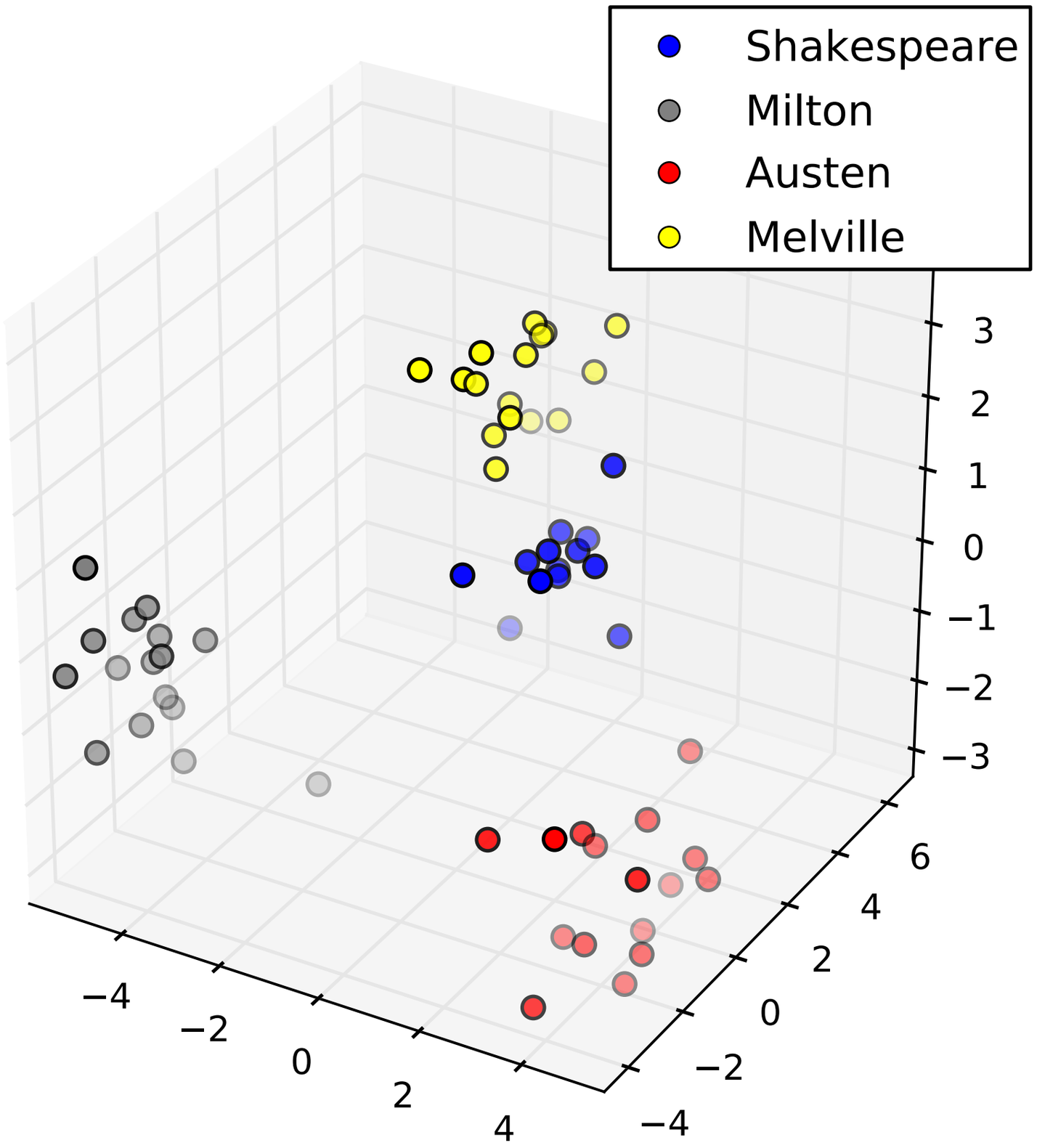
(Sep 1964, first conference on computers and humanities research?).
co-chaired by Stephen M. Parrish, Cornell, English Dept)

included "plea to the audience not to abandon their punch cards and magnetic tapes after their concordances were printed and (hopefully) published."

In Parrish's conference summary:

"when all the libraries or at least all pertinent bibliographical references are readily available on tape or in core memory, there will be no excuse for ignorance."

"...the perfection of attribution study or source study or influence study by computer techniques will make obsolete the studies that rely on the judgment and the memory of one poor fallible human scholar"


Shakespeare
Milton
Austen
Melville
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
6
4
2
0
−2
−4

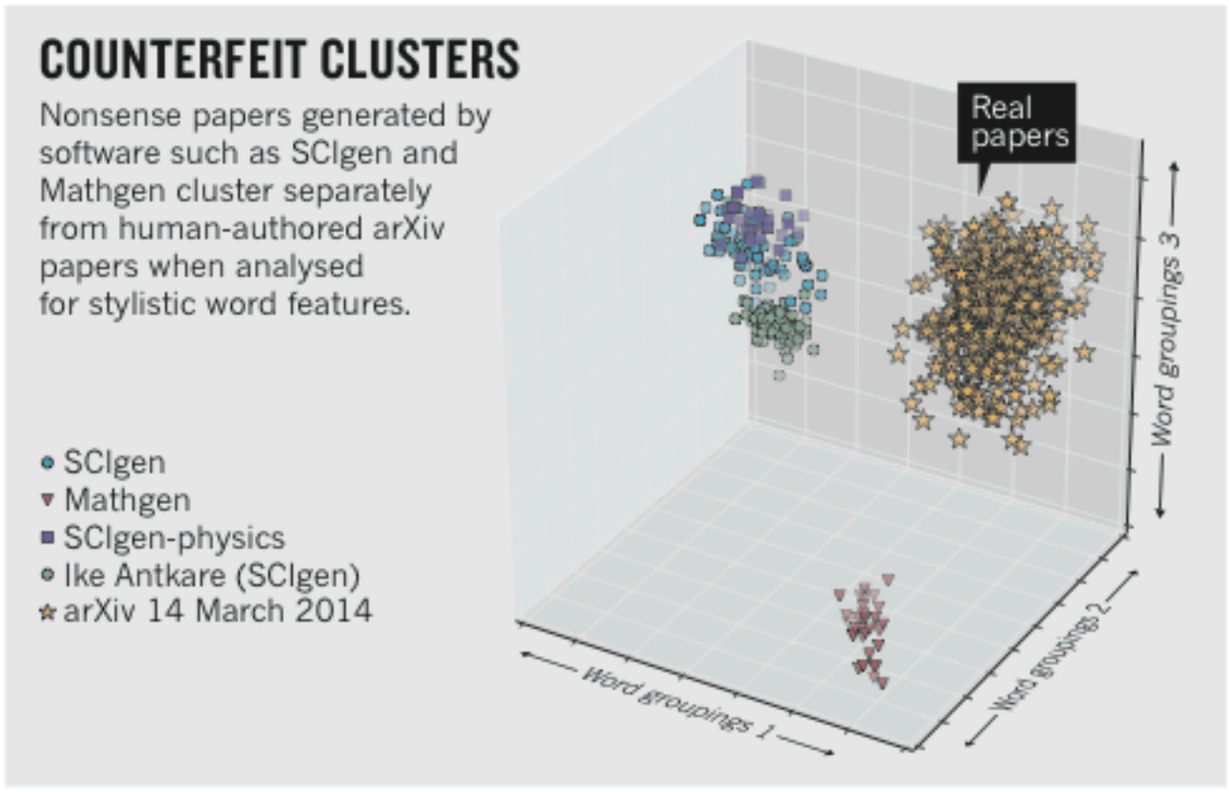
Correspondence

ArXiv screens spot fake papers

Unlike the computer-generated nonsense papers in some peer-reviewed subscription services (see *Nature* http://doi.org/r3n; 2014), the 500 or so preprints received daily by the automated repository arXiv are not pre-screened by humans. But sometimes automated assessment can be better than human diligence at enforcing standards.

The automated screens for outliers in arXiv include analysis of the probability distributions of words and their combinations, ensuring that they fall into patterns that are consistent with existing subject classes. This serves as a check of the subject categorizations provided by submitters, and helps to detect non-research content.

Fake papers generated by SCIgen software, for example, have a 'native dialect' that can be picked up by simple stylometric analysis (see J. N. G. Binongo *Chance* **16,** 9–17; 2003). The



however, science advisers may encounter a conflict of interest if they are involved in administering public research funding.

Gluckman is the New Zealand Prime Minister's chief science adviser and chaired the panel that last year selected the National Science Challenges. He has been instrumental in publicizing and defending the new funding mechanism for meeting these goals (see go.nature.com/cmgkx1), which the government

Journals must boost data sharing

The journal ecosystem is a powerful filter of scientific literature, promoting the best work into the best journals. Why not use a similar mechanism to encourage more comprehensive data sharing?

Several journals have introduced policies mandating that data be shared on a public archive at publication (see,

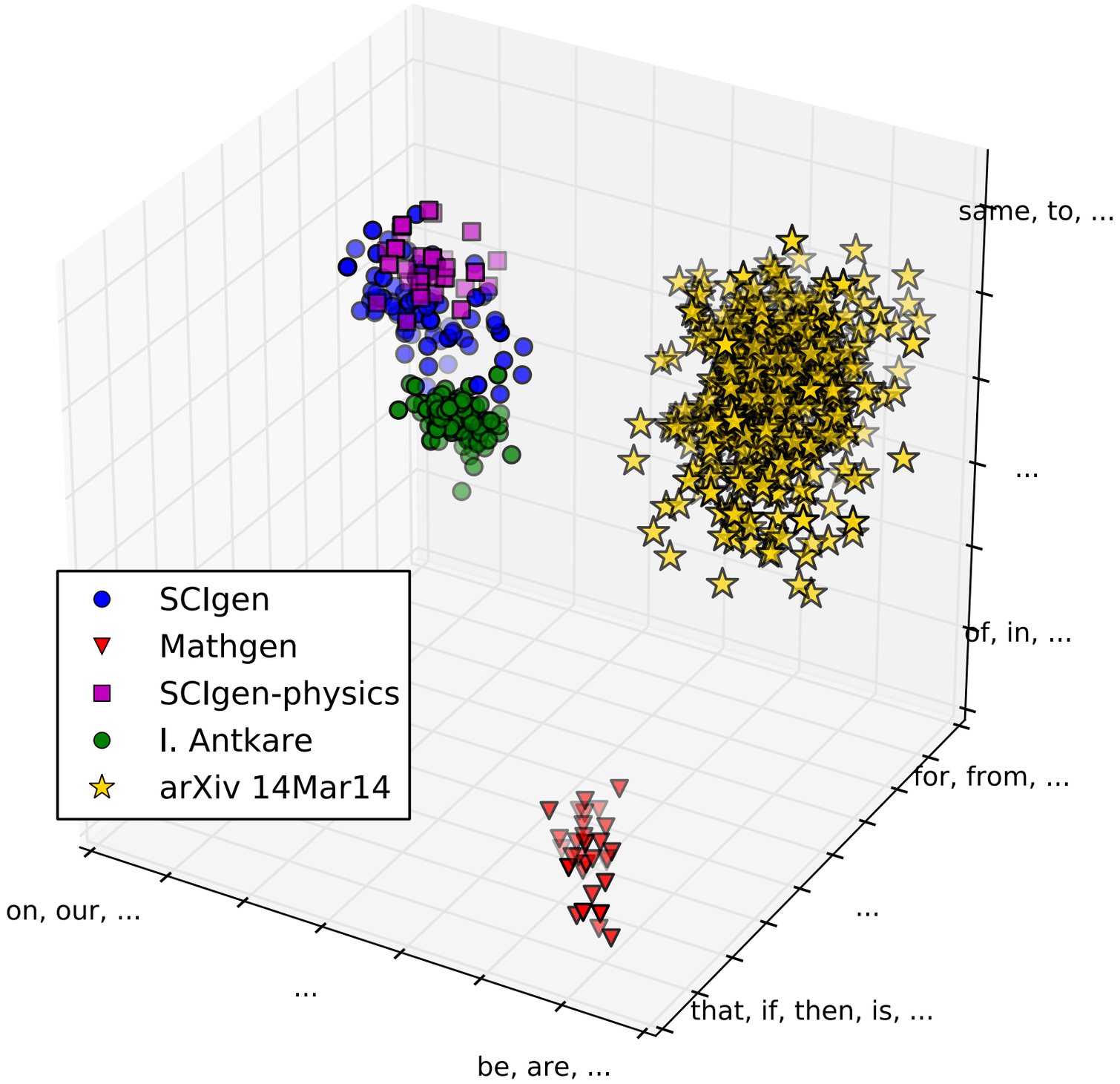
Projects powered by free computing grid

Herman Tse describes the scientific output of IBM's World Community Grid as "lacklustre" (*Nature* **507,** 431; 2014). This is not the case: the 22 projects we have supported so far have generated more than 35 peer-reviewed papers in prominent journals. Our donated computing power has resulted in several important practical scientific advances.

For example, Japan's Chiba Cancer Center used our free computing power to screen three million drug candidates for treating neuroblastoma, a common childhood cancer. This yielded seven promising compounds that have no apparent side effects (Y. Nakamura *et al. Cancer Med.* **3,** 25–35; 2014).

Last June, Harvard University's Clean Energy Project announced some 35,000 organic materials that could double the efficiency of carbon-based solar cells, after using our grid to scan more than

PCA on the Stopword Distributions



word2vec

code.google.com/p/word2vec

- arxiv:1301.3781
- arxiv:1310.4546
- arxiv:1309.4168

Words generated by combining common tokens together.

$$\frac{\text{count}(w_1, w_2) - \delta}{\text{count}(w_1)\text{count}(w_2)} > \theta$$

Four passes with decreasing threshold.

$$\delta = 30, \theta \in \{400, 300, 200, 100\}$$

syllogism

a:b :: c:d

Paris - France + Italy = ?

syllogism

a:b :: c:d

Paris - France + Italy =

Rome

arxplor.lassp.cornell.edu

(20 slides from A.Alemi presentation at March Meeting '14 Denver)

After filtering:

- 7 years: Apr 2007 - Feb 2014
- 488,072 articles.
- 422,704 authors.
- 1,285,320 unique "words".

Example "words":

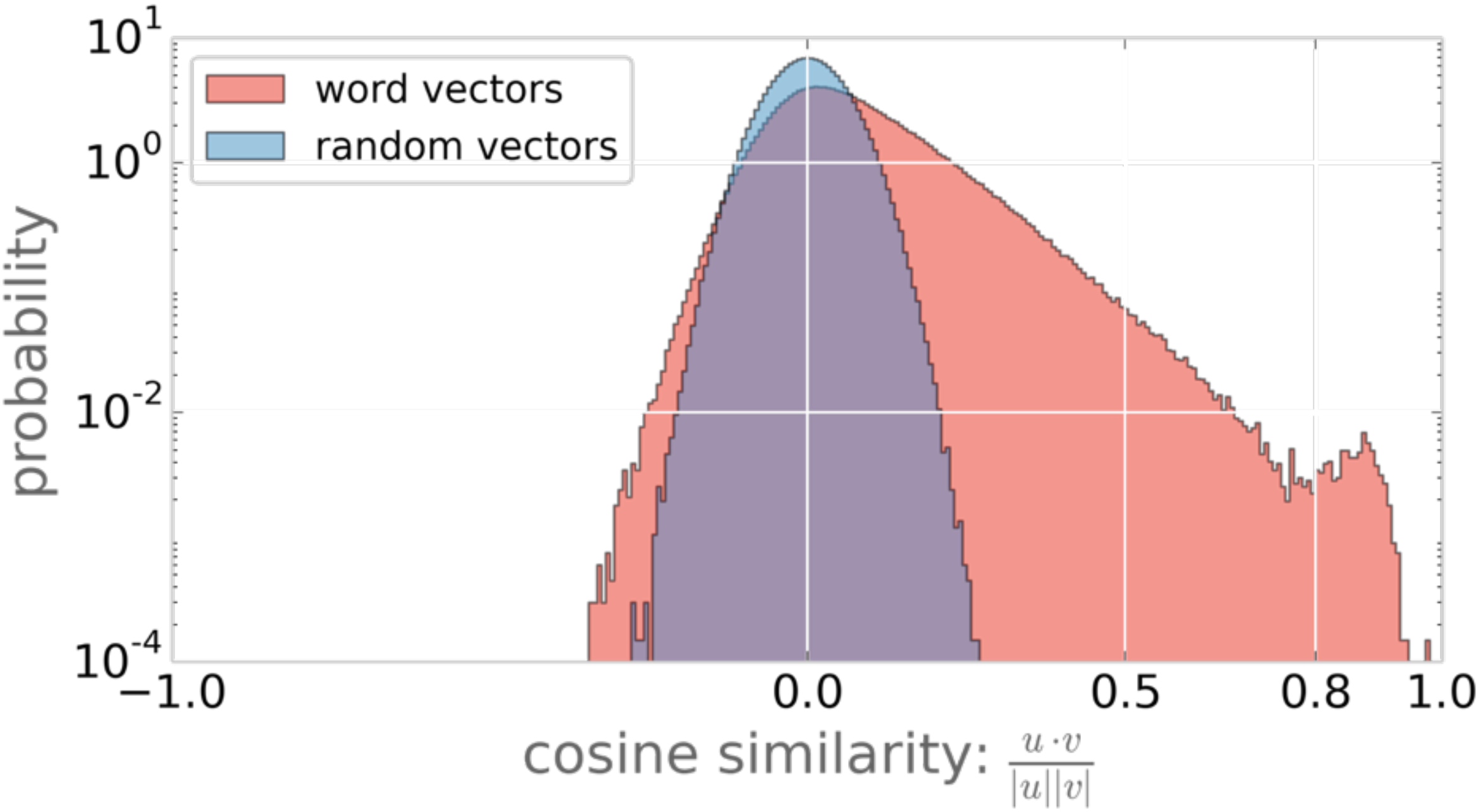
- "singular_value_decomposition",
- "black_hole",
- "aps_march_meeting"

Continuous skip-gram model. Single layer neural network.
Mapping of words to vectors: w_i.
Maximize the log probability of nearby words.

$$\sum_i \sum_{-n \le j \le n, j \neq 0} \log p(w_j | w_i)$$

where

$$p(w_j | w_i) = \frac{1}{Z} \exp\left(w_j \cdot w_i\right)$$


word vectors
random vectors
probability
10^{1}
10^{0}
10^{-2}
10^{-4}
−1.0
0.0
0.5
0.8
1.0
cosine similarity: $\frac{u \cdot v}{|u||v|}$

word	cosine
electron	1.0
electrons	0.83
positron	0.67
conduction_electron	0.65
carriers	0.64
unpaired_electron	0.63
electron_gas	0.63
electronhole	0.63
impurity	0.63
oneelectron	0.62
mobile_electrons	0.62

word	cosine
physics	1.0
theoretical_physics	0.76
particle_physics	0.72
nuclear_physics	0.7
astrophysics	0.68
astronomy	0.68
materials_science	0.67
phyiscs	0.66
astronomy_louisiana_state_univ...	0.66
bern_bern_switzerland_18	0.66
tennessee_knoxville_tennessee_...	0.66

word	cosine
blue	1.0
red	0.91
orange	0.87
cyan	0.87
purple	0.86
magenta	0.86
yellow	0.85
blue_red	0.85
violet	0.85
blue_green	0.83
light_blue	0.82

word	cosine
expects	1.0
expecting	0.63
would_allow	0.61
might_argue	0.6
expect	0.6
prefers	0.58
still_expects	0.57
one_could_imagine	0.57
can_afford	0.55
anticipate	0.54

word	cosine
almost	1.0
approximately	0.57
roughly	0.57
remarkably	0.55
fairly	0.53
still	0.51
however	0.51
nearly_constant	0.5
exponentially	0.49
although	0.49

word	cosine
want	1.0
wish	0.86
intend	0.74
do_not_need	0.71
do_not_know_how	0.67
don_t	0.66
trying	0.65
one_needs	0.63
going	0.62
come_back	0.62

word	cosine
svd	1.0
truncated_svd	0.7
svds	0.69
qr_decomposition	0.66
svd_decomposition	0.65
eigendecomposition	0.62
music_algorithm	0.61
omp_algorithm	0.61
cholesky_factorization	0.61
alternating_mini...	0.6

word	cosine
graphene	1.0
single_layer_graphene	0.9
monolayer_graphene	0.9
bilayer_graphene	0.88
multilayer_graphene	0.88
graphene_monolayer	0.88
graphene_sheets	0.87
suspended_graphene	0.85
graphene_monolayers	0.85
graphene_sheet	0.85

word	cosine
python	1.0
source_code	0.83
python_program...	0.83
java	0.83
scripting_language	0.82
command_line	0.82
scipy	0.82
package	0.82
scripts	0.82
libraries	0.82

word	cosine
supreme_court	1.0
politics	0.76
lawyers	0.74
justices	0.73
senate	0.73
congressional	0.72
politicians	0.72
public_opinion	0.71
republican_party	0.71
political	0.7

word	cosine
tfidf	1.0
tf_idf	0.8
cooccurrence	0.7
training_corpus	0.68
cosine_similarity	0.68
bm25	0.68
tf_idf_weighting	0.68
stop_words	0.67
unigrams	0.67
cosine_similarity_between	0.66

word	cosine
scientific_impact	1.0
interdisciplinarity	0.79
citation_impact	0.77
bibliometrics	0.76
citation_analysis	0.75
citation_patterns	0.74
bibliometric	0.73
journals	0.73
scholarly	0.73
peer_review	0.72

word	cosine	word	cosine	word	cosine	word	cosine	word	cosine
john	1.0	dmitri	1.0	wang	1.0	stefano	1.0	pierre	1.0
william	0.86	dmitry	0.78	chen	0.95	paolo	0.9	alain	0.82
michael	0.84	mikhail	0.78	zhang	0.94	francesco	0.89	olivier	0.8
edward	0.84	oleg	0.77	liu	0.94	matteo	0.88	philippe	0.79
david	0.84	sergey	0.76	zhou	0.93	michele	0.88	frederic	0.79
robert	0.84	konstantin	0.76	zhao	0.93	giuseppe	0.87	stephane	0.78
andrew	0.84	igor	0.76	zhu	0.92	alessandro	0.87	sylvain	0.78
james	0.83	alexey	0.75	huang	0.91	davide	0.87	jean_francois	0.78
peter	0.83	anatoly	0.75	fang	0.9	giorgio	0.87	sebastien	0.78
stephen	0.82	ilya	0.74	wei	0.9	riccardo	0.87	benoit	0.78
brian	0.82	ivan	0.73	guo	0.9	enrico	0.87	thierry	0.77
philip	0.82	nikolay	0.73	ding	0.89	francesca	0.87	guillaume	0.77

syllogism

a:b :: c:d

Paris - France + Italy = ?

torque - force + momentum =

orbital_angular_momentum

newtonian_mechanics - isaac_newton +
albert_einstein =

special_relativity

gravity - Newton + Hawking =

black_hole_evaporation

smuon - muon + higgs =

higgsino

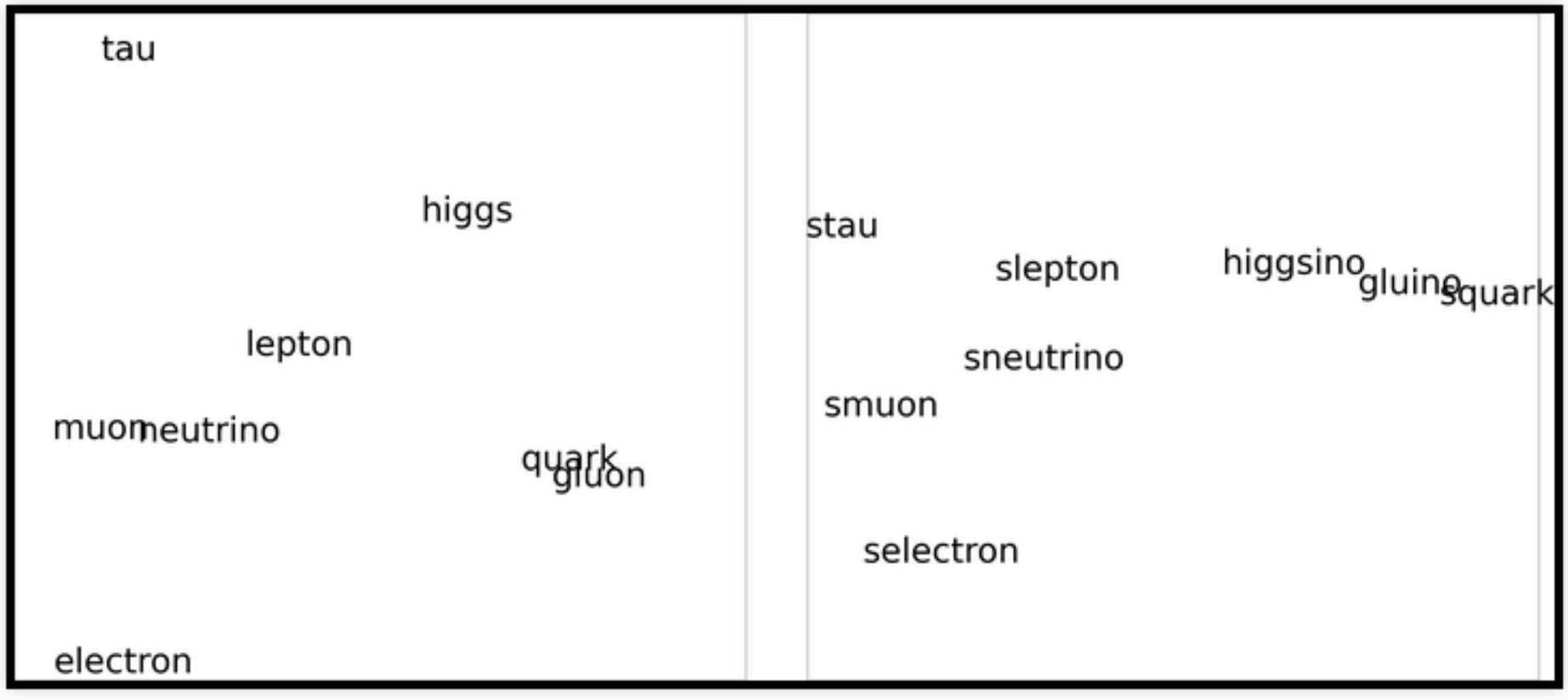
gravity-newton+hawking	
hawking	0.8
an_evaporating_black_hole	0.73
black_hole_evaporation	0.7
hawking_effect	0.7
cosmic_censorship	0.69
gravity	0.66

torque-force+momentum=	
momentum	0.72
angular_momentum_conservation	0.59
longitudinal_component	0.56
orbital_angular_momentum	0.55
angular_momentum_lz	0.55
helicity	0.54

newtonian_mechanics-isaac_newton+albert_einstein=	
newtonian_mechanics	0.78
special_relativity	0.57
newtonian_dynamics	0.55
planetary_motions	0.54
material_bodies	0.53
aristotelian	0.51

smuon-muon+higgs	
higgsino	0.81
sfermions	0.79
higgsinos	0.78
heavy_higgs	0.78
lightest_higgs_boson	0.78
smuon	0.78

Using the same projections, different parts of vector space.



word2vec knows supersymmetry

physics+buzzword=

buzzword	0.79
physics	0.77
life_science	0.7
philosophy	0.68
interdisciplinary_research	0.67
biology	0.66

chemistry+physics=

chemistry	0.92
physics	0.9
theoretical_chemistry	0.73
solid_state_physics	0.72
chemical_engineering	0.7
theoretical_physics	0.69

biology-buzzword=

physiology	0.44
ecology	0.42
biological_macromolecules	0.41
cell_motility	0.4
flegg	0.39
phenotypes_nat_rev_genet	0.39

sm_higgs_boson the search for the higgs_boson the only elementary_particle in the sm that has_not_yet been_observed is one of the highlights of the large_hadron_collider 11 lhc physics_programme indirect limits on the sm_higgs_boson mass of mh 158 gev at_95_confidence_level_cl have_been set using global_fits to precision_electroweak results 12 direct_searches at lep 13 the tevatron 14 16 and the lhc 17 18 have previously excluded at_95_cl a sm_higgs_boson with mass below 600_gev apart_from some mass regions between 116 gev and 127 gev both the atlas and cms_collaborations reported excesses of events in their 2011 datasets of protonproton pp_collisions_at centre of mass energy s 7 tev at the lhc which were compatible_with sm_higgs_boson_production and decay in the mass region 124 126 gev with significances of 2 9 and 3 1 standard_deviations s respectively 17 18 the cdf and do experiments at the tevatron have also recently_reported a broad excess in the mass region 120 135 gev using the existing lhc constraints the observed local significances for mh_125_gev are 2 7 s for cdf 14 1 1 s for do 15 and 2 8 s for their combination 16 the previous atlas_searches in 4 6 4 8 fb 1 of data at s 7 tev are combined here with new searches for h zz 41 h gg and h ww

Full text of the higgs article, colored by K-means clustering of words

astro
cond-mat
cs
hep
math
places
references
names
equations
english

art	title	cosine
1207.7214	Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC	1.0
1307.1427	Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC	0.98
1207.7235	Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC	0.98
1303.4571	Observation of a new boson with mass near 125 GeV in pp collisions at sqrt(s) = 7 and 8 TeV	0.97
1106.2748	Limits on the production of the Standard Model Higgs Boson in pp collisions at sqrt(s) =7 TeV with the ATLAS detector	0.97
1206.0756	Search for the Standard Model Higgs boson in the H -> WW(*) -> l nu l nu decay mode with 4.7 /fb of ATLAS data at sqrt(s) = 7 TeV	0.97
1211.6956	Search for the neutral Higgs bosons of the Minimal Supersymmetric Standard Model in pp collisions at sqrt(s)=7 TeV with the ATLAS detector	0.97

author	cosine	ca
Anderson, Philip W	1.0	✔
Zaanen, J.	0.94	✖
Hirsch, J. E.	0.94	✖
Squire, Richard H.	0.94	✖
Rice, T. Maurice	0.94	✖
Phillips, Philip	0.93	✖
Kallin, Catherine	0.93	✖
Kivelson, S. A.	0.93	✖
Yildirim, Yucel	0.93	✖
Tesanovic, Zlatko	0.93	✖
Squire, R. H.	0.93	✖
Nikolic, Predrag	0.93	✖

author	cosine	ca
Hawking, S. W.	1.0	✔
Hertog, Thomas	0.99	✔
Hartle, James	0.98	✔
Hartle, James B.	0.94	✔
Freivogel, Ben	0.92	✖
Fujio, Kazuya	0.92	✖
Ashtekar, Abhay	0.92	✖
Singh, Parampreet	0.91	✖
Sekino, Yasuhiro	0.91	✖
Robles-Pérez, Salvador	0.91	✖
Leichenauer, Stefan	0.91	✖
Bousso, Raphael	0.91	✖

author	cosine	ca
Korniss, G.	1.0	✔
Jensen, Henrik Jeldtoft	0.96	✖
Gross, Thilo	0.95	✖
Blasius, Bernd	0.94	✖
Tessone, Claudio J.	0.94	✖
Moreno, Y.	0.94	✖
Hernandez-Garcia, Emilio	0.94	✖
Cuesta, Jose A.	0.93	✖
Zanette, Damian H.	0.93	✖
Pigolotti, Simone	0.93	✖
Wang, Bing-Hong	0.93	✖
Miguel, M. San	0.93	✖

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

http://nlp.stanford.edu/projects/glove/

Conference on Empirical Methods in Natural Language Processing (**EMNLP 2014**),
October 26–28, 2014

For semantic applications like the analogy task, the vector space embedding should respect the ratios of conditional probabilities. For example, the ratio

$$p(k \mid \text{ice}) / p(k \mid \text{steam})$$

is high for k = solid,
intermediate for k= water, fashion
and low for k = gas.

So if we're interested in thermodynamic phase, we learn that solid and gas are relevant to the distinction between ice and steam, and water and fashion are not.

$p(k \mid i) = X_{ik} / X_k$ is the probability that word k appears in the context of word i, where X_{ik} is the co-occurrence count, and X_k the total number of occurrences of word k.

Encode word similarities linearly, $w_i - w_j$

Use dot product to encode ratio

$$F((w_i - w_j) \cdot \tilde{w}_k) = p(k \mid i) / p(k \mid j)$$

and symmetric in words and contexts

$$F(w_i \cdot \tilde{w}_k) = p(k \mid i) = X_{ik} / X_i$$

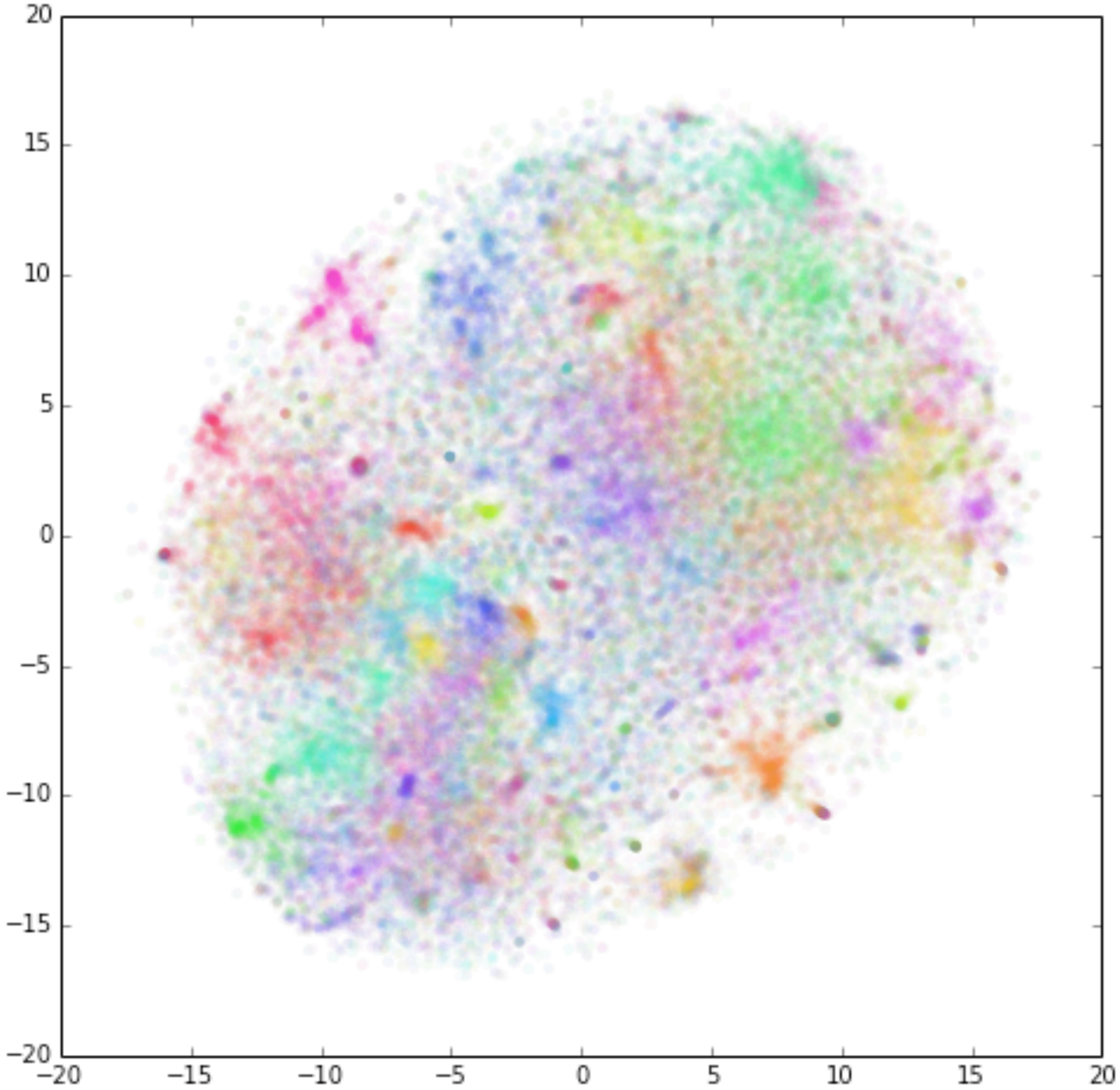
and F ~ exp, so that

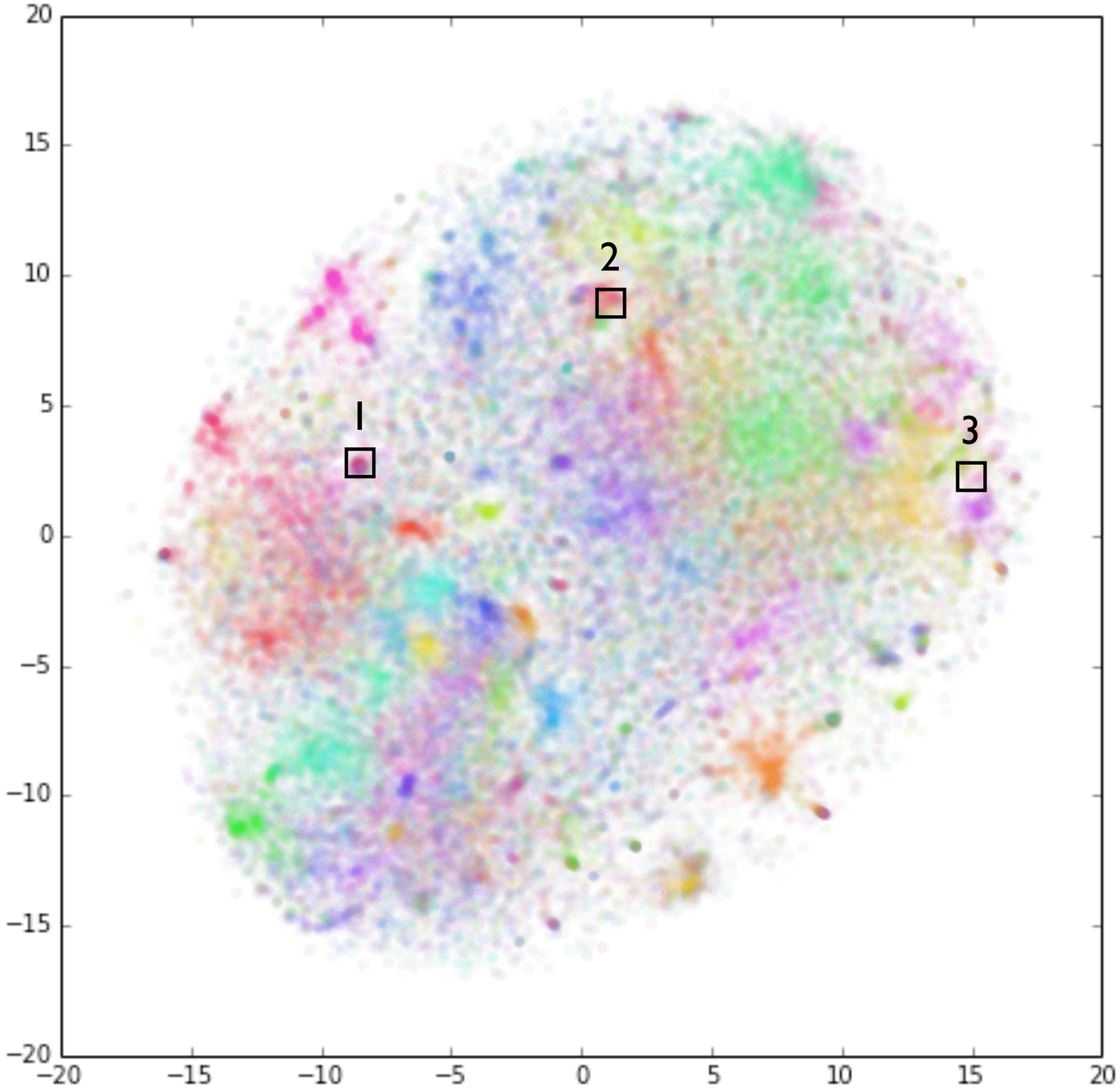
$$w_i \cdot \tilde{w}_k = \log p(k \mid i) = \log X_{ik} - \log X_i$$

leads to GloVe objective:

$$J = \sum_{i,j} f(X_{ij}) \left(w_i \cdot \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

f discounts low counts, train with AdaGrad

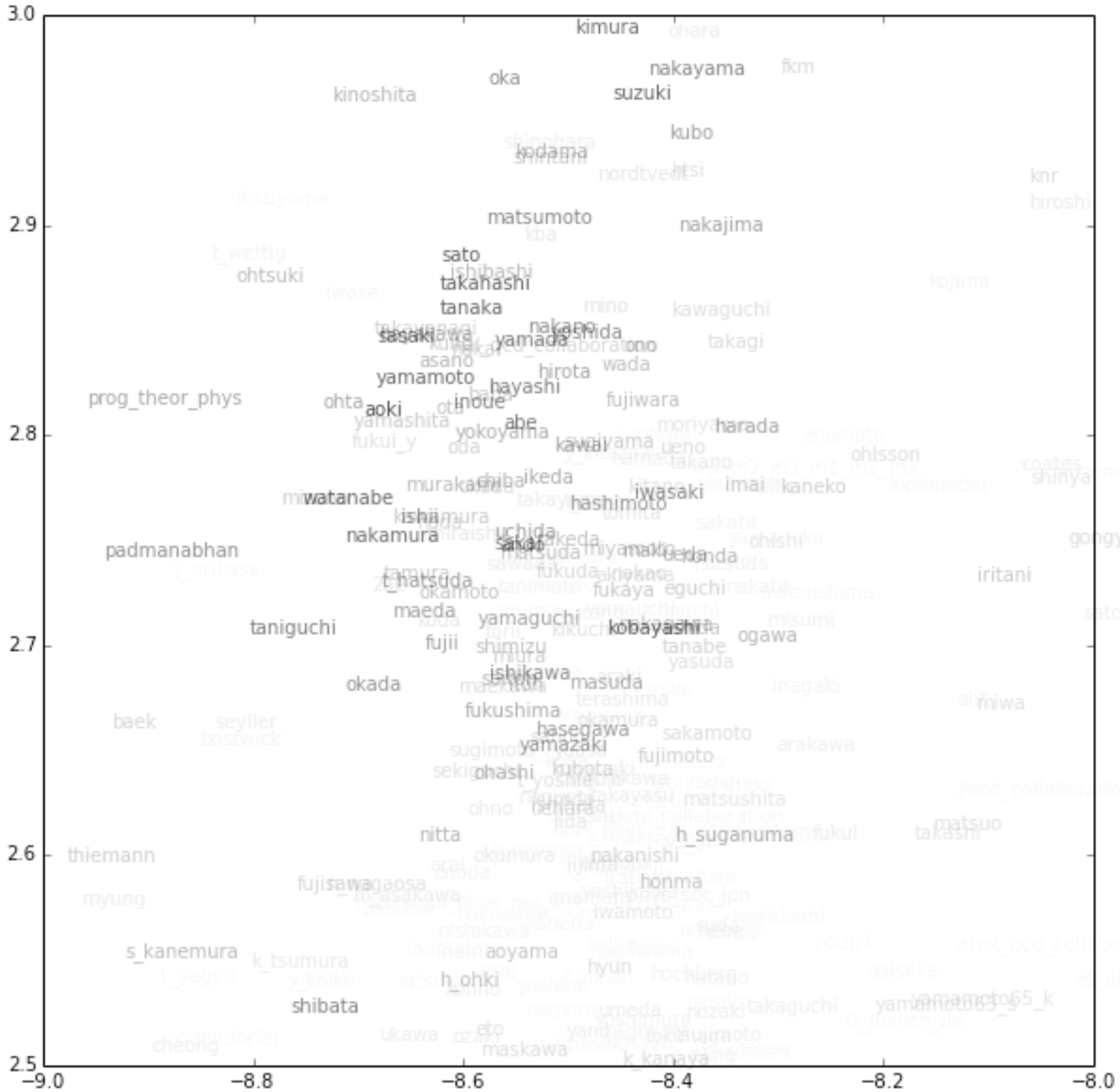

20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20


20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20
1
2
3

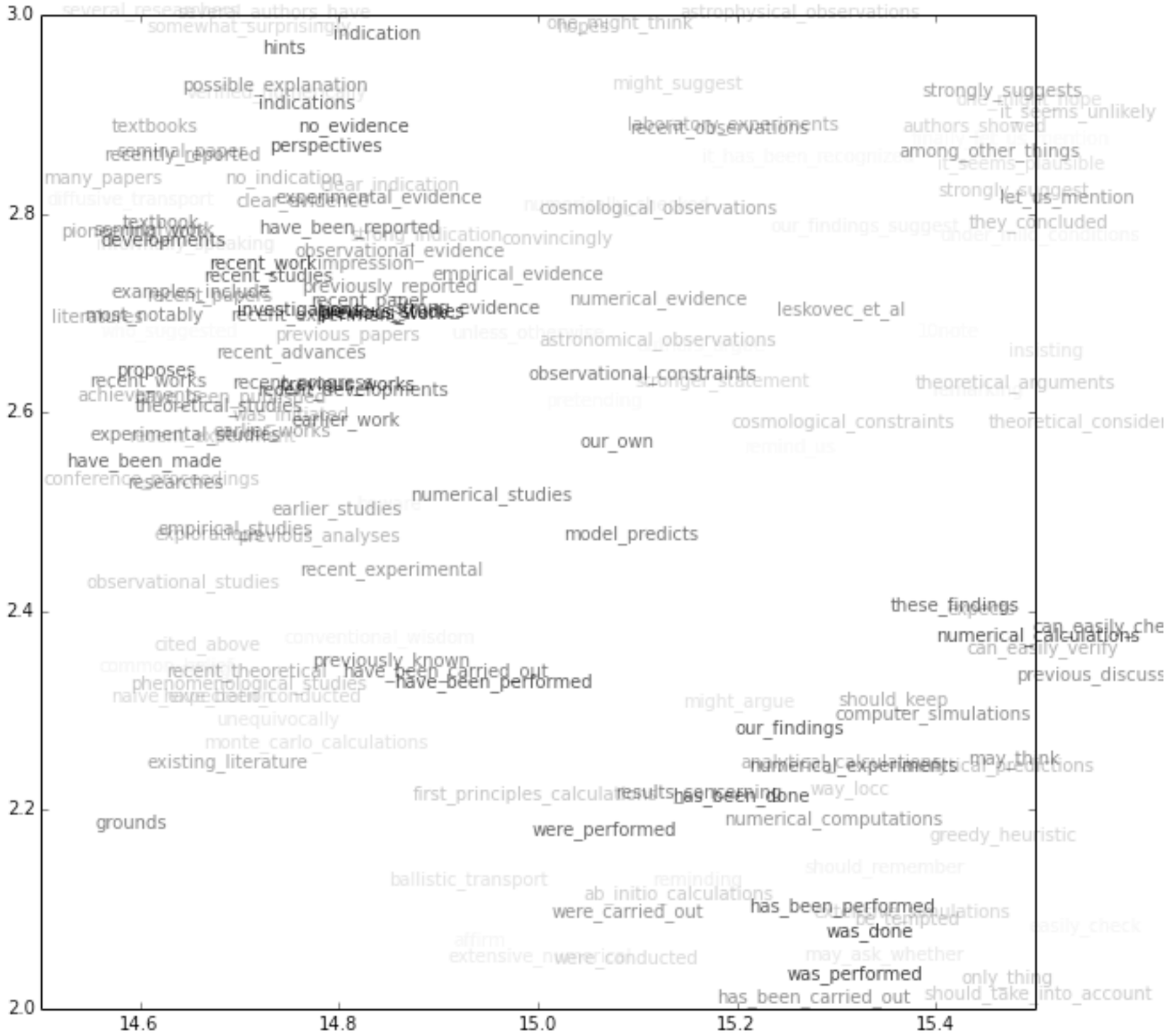
Rectangle 1


9.4
9.2
9.0
8.8
8.6
0.6
0.8
1.0
1.2
1.4
nh3
hc7n
12co
co_column_densities
line_fluxes
infall
vlsr
extinction
brightness
n2h
host_galaxy
systemic_velocity
cnd
ccs
shaped
flattened
compact_objects
neutron_stars
dust_extinction
spiral
shock
tracer
optically_thin
gas
dust
optically_thick
tracing
diagnostics
stellar_disc
dense_core
blended
filament
galactic_plane
tracers
sio
unbound
h2o
traced
inflow
disc
bar
physical_properties
cloud
galactic_center
emission_line
pdr
ionized_gas
filaments
abundance
accretion
star
disk
clump
dynamical_evolution
molecular_gas
elliptical
periphery
stellar
halo
clumps
interstellar_medium
galactic
clouds
morphology
disks
burst
shocks
enrichment
outflow
winds
central_black_hole
spatial_distribution
streams
gaseous
dusty
discs
stream
hot_gas
galactic_disk
molecular_clouds
bulge
galaxy
outflows
cores
star_formation
spiral_structure
bars
radial_migration
stellar_halo
radio_emission
milky_way
substructure
spiral_arms
chemical_evolution
unresolved
central_region
star_cluster
thin_disk
hii_region
accretion_disk
cold
outer_disk
extragalactic
inner_disk
obscured
halos
outer_regions
stars
satellite
nebula
substructures
outer_parts
protoplanetary_disks
dense_clouds
diffuse_emission
thick_disk
progenitor
chemical_evolution_models
dense_cores
m51
ejected
m33
haloes
m31
satellites
bright
ridges
retained
dwarf_galaxy
luminous
outskirts
central_regions
pulsar
star_clusters
binaries
massive_stars
mergers
warm
galaxies
star_formation_activity
nearby_galaxies
faint
debris
spirals
dwarf_galaxies
star_forming_regions
starburst
brightest
agn_activity
gcs
agn
young_stars
bhs
pulsars
disk_galaxies
progenitors
long_period
central_star
globular_clusters
overdensity
remnants
short_period
open_clusters
giant
elliptical_galaxies
inner_region
islands
globular_clusters
hi_regions
outer_region
main_sequence_stars

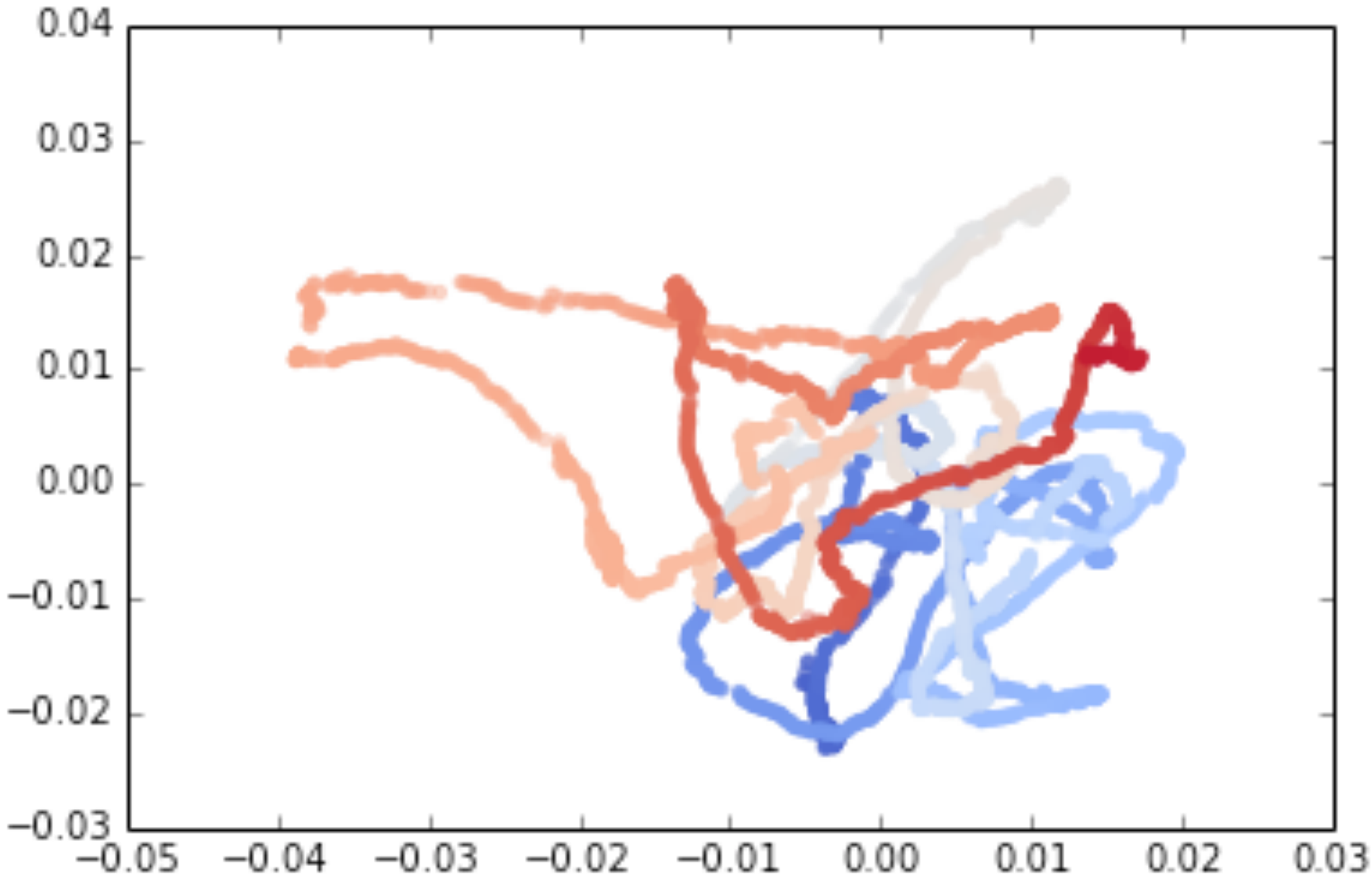
Rectangle 2


3.0
2.9
2.8
2.7
2.6
2.5
−9.0
−8.8
−8.6
−8.4
−8.2
−8.0
kimura
nakayama
oka
suzuki
kinoshita
kubo
kodama
matsumoto
nakajima
sato
ishibashi
takahashi
ohtsuki
tanaka
kawaguchi
nakano
sasaki
yamada
ono
takagi
asano
hirota
wada
yamamoto
hayashi
prog_theor_phys
ohta
inoue
fujiwara
aoki
abe
harada
yokoyama
kawai
ohlsson
ueno
takano
ikeda
chiba
imai
kaneko
watanabe
iwasaki
hashimoto
ishii
nakamura
uchida
sakai
padmanabhan
matsuda
maeda
iritani
t_hatsuda
okamoto
fukaya
eguchi
maeda
yamaguchi
kobayashi
taniguchi
ogawa
fujii
shimizu
tanabe
ishikawa
okada
masuda
fukushima
hasegawa
yamazaki
fujimoto
ohashi
kubota
nitta
h_suganuma
thiemann
nakanishi
honma
s_kanemura
aoyama
h_ohki
shibata

Rectangle 3


3.0
2.8
2.6
2.4
2.2
2.0
14.6
14.8
15.0
15.2
15.4
indication
hints
possible_explanation
indications
no_evidence
perspectives
textbooks
recently_reported
many_papers
no_indication
experimental_evidence
textbook
have_been_reported
developments
observational_evidence
recent_work
recent_studies
empirical_evidence
previously_reported
examples_include
recent_paper
most_notably
recent_advances
proposes
recent_works
theoretical_studies
earlier_work
experimental_studies
have_been_made
researches
numerical_studies
earlier_studies
empirical_studies
previous_analyses
model_predicts
recent_experimental
observational_studies
previously_known
recent_theoretical
have_been_carried_out
have_been_performed
phenomenological_studies
existing_literature
grounds
cosmological_observations
convincingly
numerical_evidence
leskovec_et_al
astronomical_observations
observational_constraints
recent_developments
our_own
cosmological_constraints
strongly_suggests
recent_observations
among_other_things
let_us_mention
they_concluded
these_findings
numerical_calculations
our_findings
numerical_experiments
has_been_done
numerical_computations
were_performed
has_been_performed
were_carried_out
was_done
was_performed
has_been_carried_out

“Paragraphs” live in the same vector space



Introduce article context a

$$J = \sum_i f(X_j)\,(v \cdot w_j + a + b_j - \log X_j)^2$$

linear equation
can be minimized exactly in fixed “background” w

100 dim representation does well on classification task

Readers and Authors live in the same vector space

Extend article context to readers:

Reader vectors => Recommender System

Extend article context to authors:

Author vectors => Referee Selection

Recommendation redux

Complaints about information overload date back 2 millennia

coreadership (proxy for svd or more sophisticated)

- **netflix prize**
- **itunes genius**

Evaluation metric? (free bagels and cream cheese for duration)

The information layer vs. the social layer (Google vs. Facebook): optimal referral mechanism?

Dangers of recommendation systems: local vs. global diversity

- **imperfect filter worse than none?**

Personalization: readers inherit topics from articles

Recommender Systems

Example: NASA ADS (Astrophysical Data System) uses (anonymized) arXiv usage data

(a) infer topics from readership data and keyword assignment

- **classify articles and users (based on past activity) according to topics**
- **measures of proximity of articles to people, and articles to themselves.**
- **reader can be presented with a menu of recent papers on subjects of interest (ordered according to closeness of match, or by importance as measured by readership or citation, ...)**

(b) find the 40 most similar articles (augments data for sparse readership) to make article-based recommendations, via a few algorithms

Text Overlap

Text “reuse” by global researchers in a scholarly corpus

Simple n-gram analysis of the texts in arXiv covering over 20 years

Everything from

- **dozens of pages verbatim from 3rd party lecture notes for PhD theses**
- **large sections of Wikipedia entries for introductory material in articles**
- **series of articles by overlapping authors each greater than 50% overlap with preceding**
- **articles assembled in whole or part from one or more other articles by different authors, with or without attribution**

Majority have found way undetected into conventional publication venues

Shed light on sociology, mentality, methodology, and demography of perps?

Full text analysis

winnowed 7-grams (w/ J. Gehrke, D.Sorokina , S. Warner 2007), after Schleimer et al. (2003)

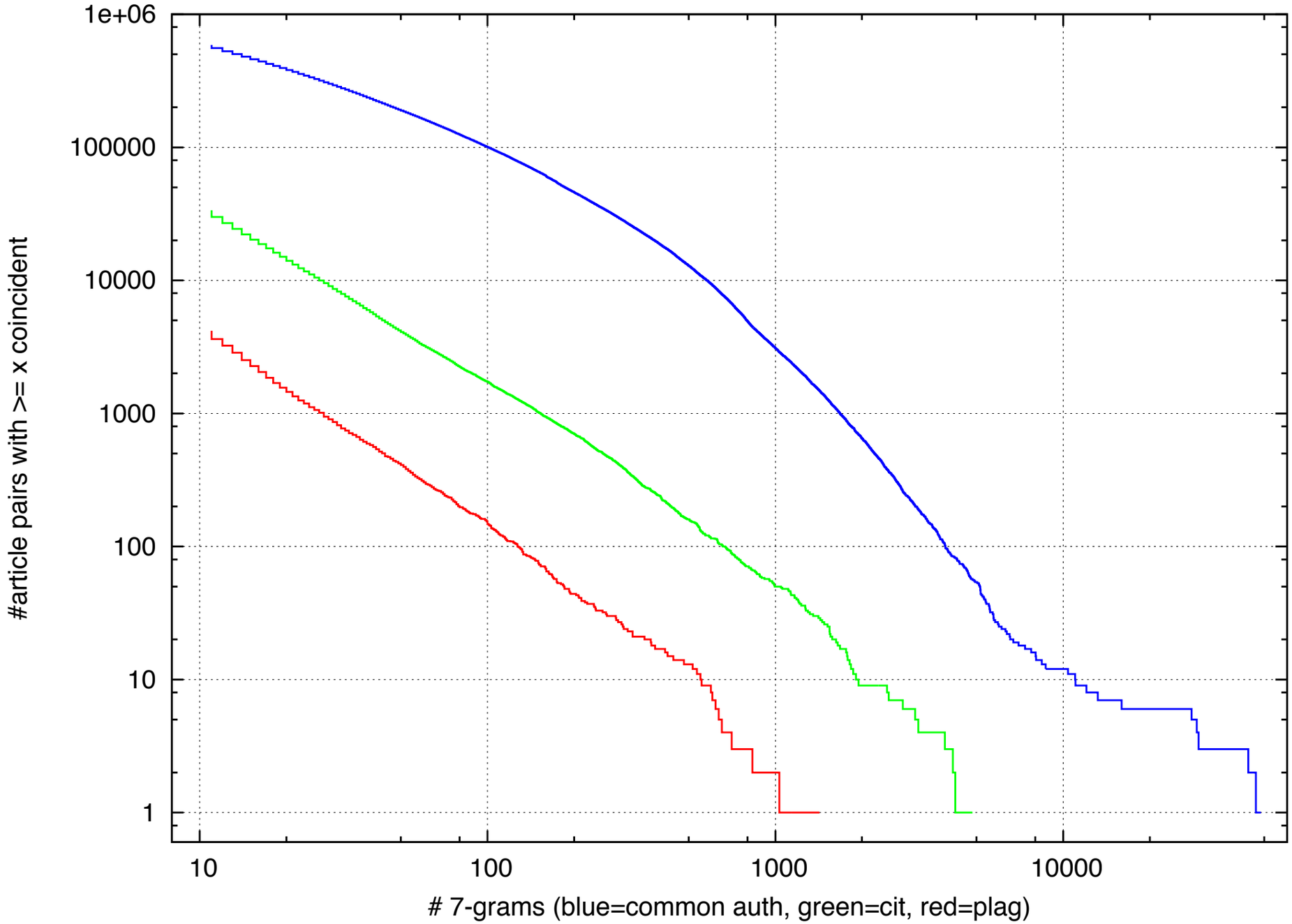
Detection software now more feasible than ever, computation of fingerprint in memory with 96Gb machine, hundreds of lookups per second

practical implications for running arXiv site: problem authors are inconvenience to readers, but screening was haphazard, no systematic baseline to identify outliers

cs Meng project Scott Rogoff (in conjunction with S. Warner), spring 2011

more systematically summer 2012 w/ Daniel Citron

(PNAS, to appear)



Number of article pairs with at least the number of overlapping 7 grams given on horizontal, log–log scale, red signifies without attribution, green with attribution, and blue with at least one common author.

http://arxiv.org/help/overlap

Starting in Jun 2011, some submissions to arXiv marked with an "admin note", indicating text overlap with other arXiv submissions (200-250 / month currently flagged).

"Such notes are intended as informational to readers, and as well to authors from different educational backgrounds. Readers frequently find it useful to know when an article draws heavily from another, or supersedes an earlier article. Some authors, by contrast, are not aware that importing large sections of text either from their earlier articles, or from articles by others, is not common practice."

Caveats

- **Not “plagiarism” in its most general form — i.e., unattributed use of ideas (whether or not text is copied).**
- **no attempt to detect text copied from sources outside of arXiv**
- **simple factual statement regarding textual overlap of materials only within arXiv (not Wikipedia, print literature, web search etc.)**
- **watch out for: famous quotes, experimental articles (author lists), review articles, conf proceedings [but note cs/info ?], other benign (refs not stripped), math (?), explicit quotes, hidden pdf text, ...**

high threshold

Threshold for appearance of the admin note is set quite high – many articles with smaller amounts of detected overlap are not noted.

“The appearance of an arXiv admin note does not suggest misconduct on the part of the author, or that an article does not contain original work. Sometimes it simply serves to suggest a related article, or can serve as a quality flag. (There is a statistically significant correlation between the amount of reused content in an article and a smaller number of citations received years later.)”

high threshold, cont'd

Articles flagged as having text overlap with articles "by other authors" must have at least multiple consecutive sentences in common. Overlap between articles having at least one coauthor in common is permitted an even higher threshold: typically at least roughly 1/3 of the content of the newer text must be taken verbatim from the earlier article in order to be noted.

(in practice also use size of contiguous blocks)

Additional exceptions for articles having a coauthor in common: articles marked by authors in the "Comments:" as review articles, or theses, conference proceedings, book contributions, and so on, are not noted, because such overlaps, whether or not desirable, appear to be common practice.

author reactions

a) none (replace w/o changing, do they even notice?)

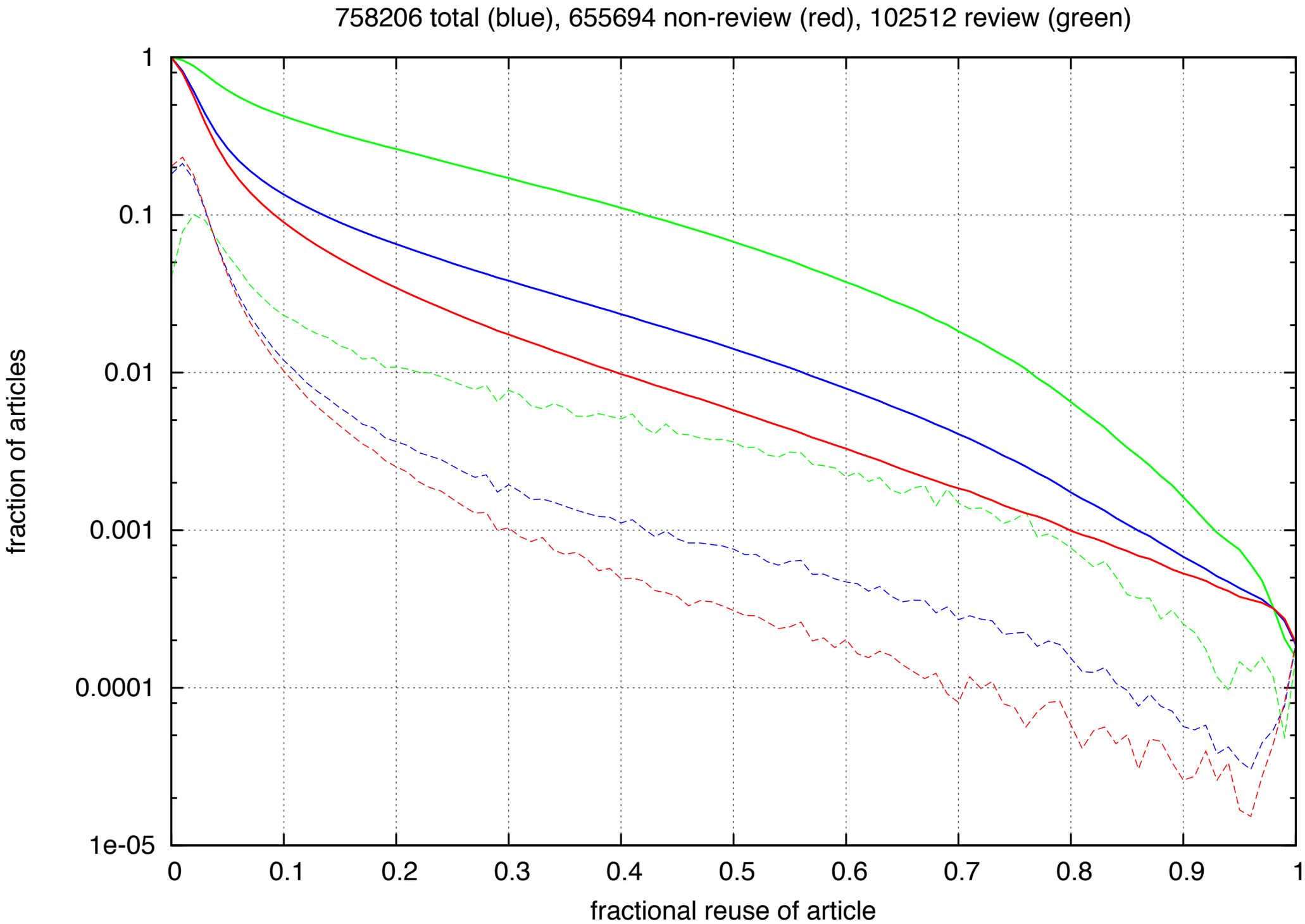
b) try to remove overlaps, not always successful (can't even find ?!?)

c) virulently object (crowdsourced quality control of methodology, though usual complaints misguided, exposing confusion about what is standard practice, as statistically confirmed by arXiv corpus)

[but note Kiesler et al, 2010, "Regulating Behavior in on-line communities":
Design Claim 15: Publicly displaying many examples of inappropriate behavior on the site will lead members to believe this is common and expected.]

start to distinguish

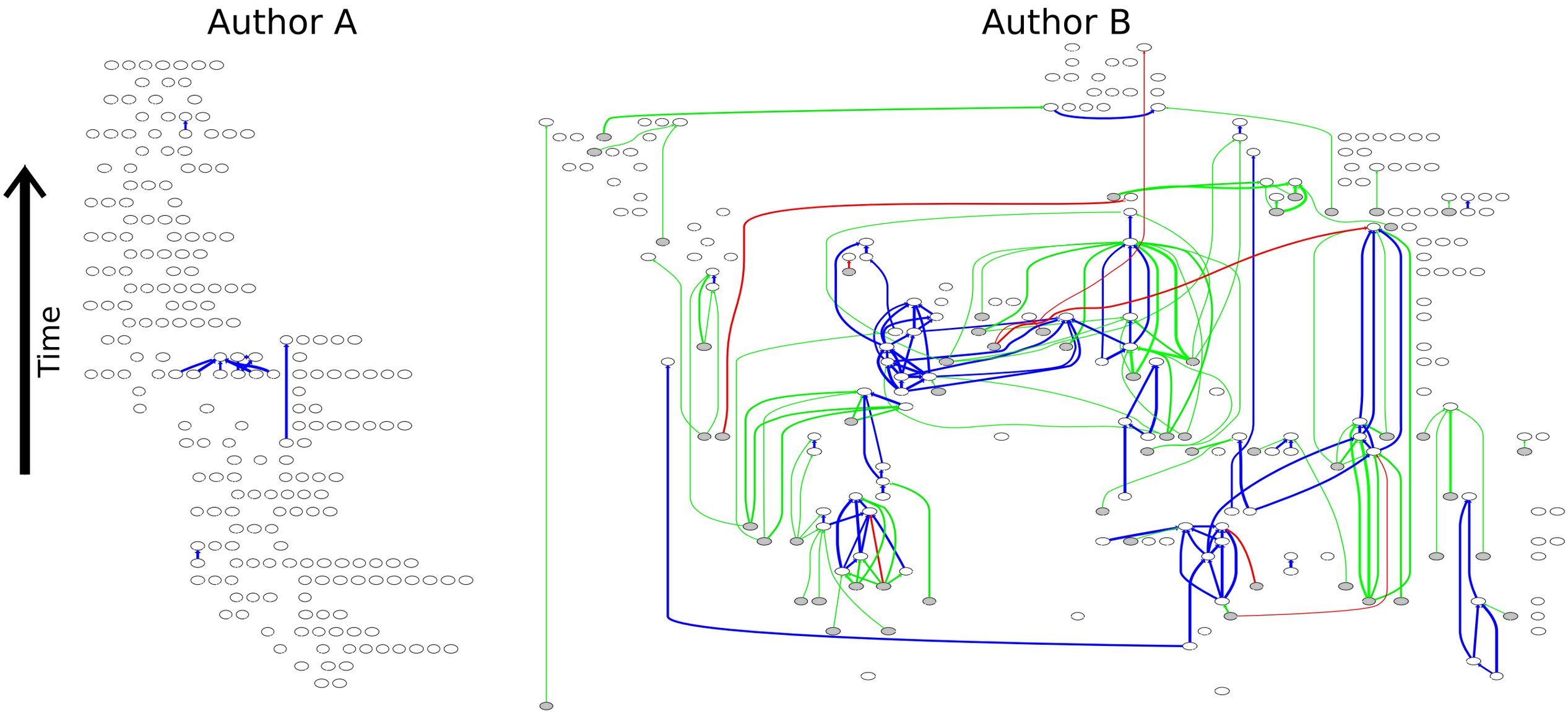
previous graph looked bad, but is some of it "acceptable" recycling?



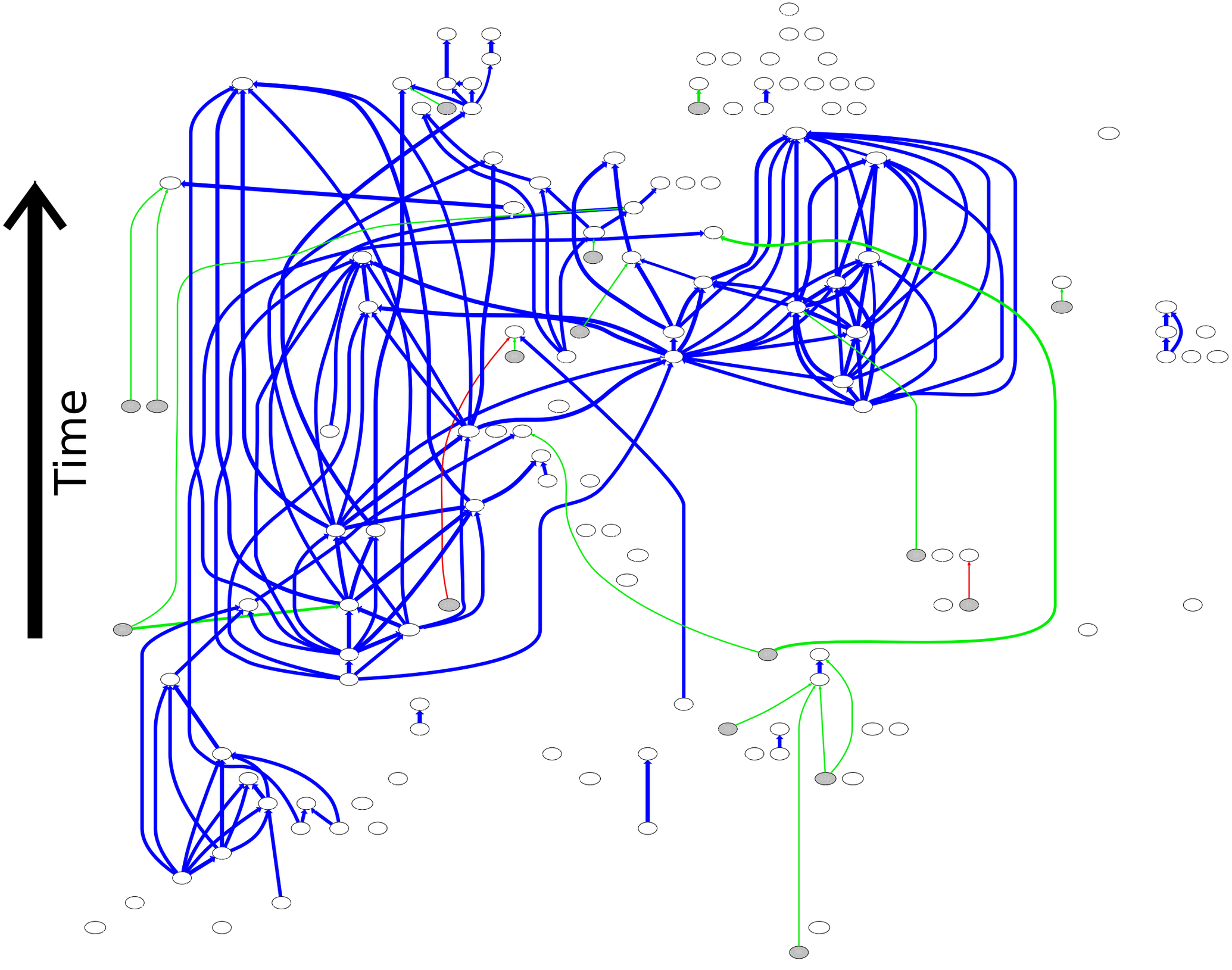
Fraction of articles on vertical with at least the indicated fraction of reused 7-grams on horizontal. green = “review”, red = non-“review”, blue= all

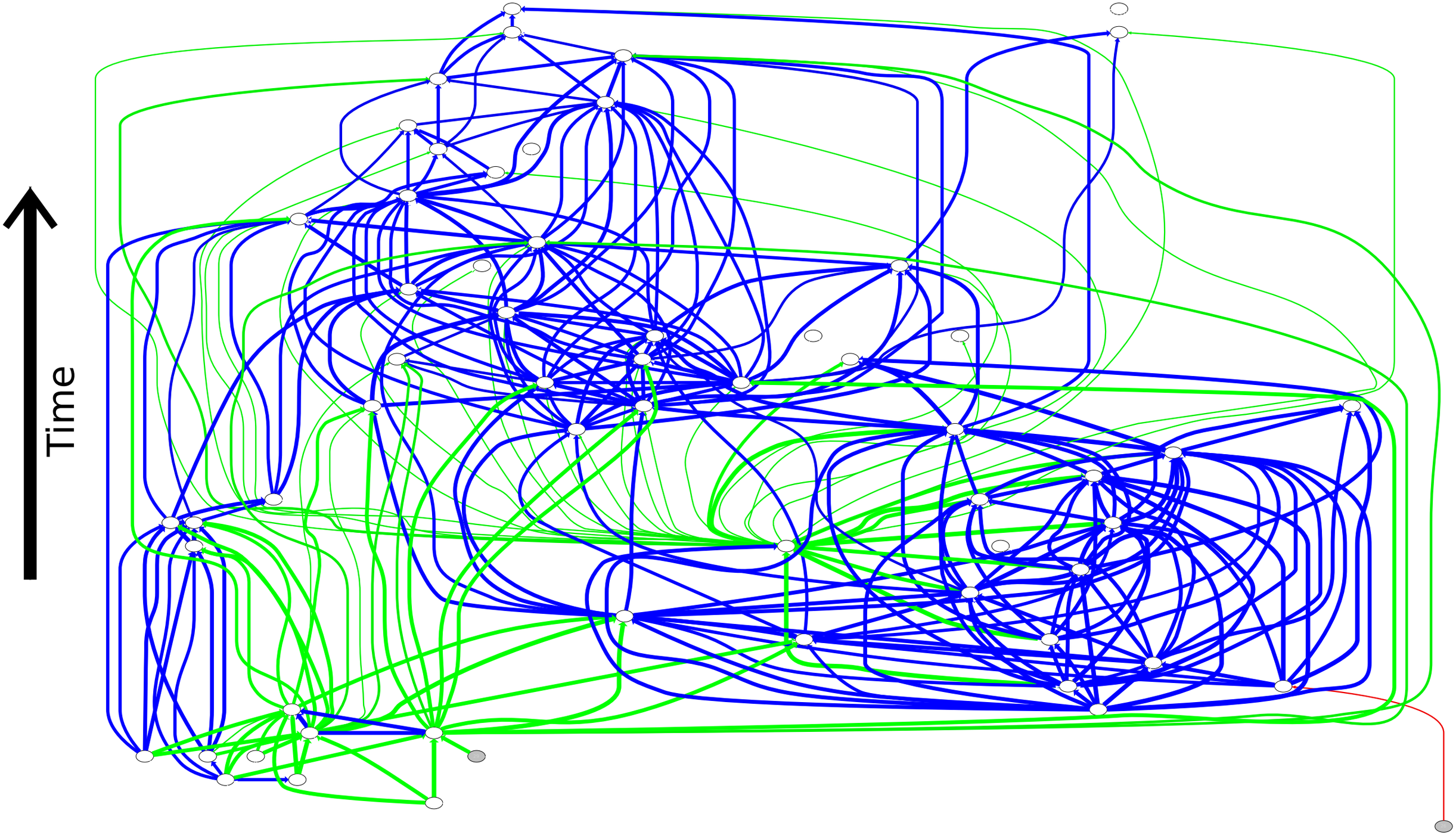
but how distributed?

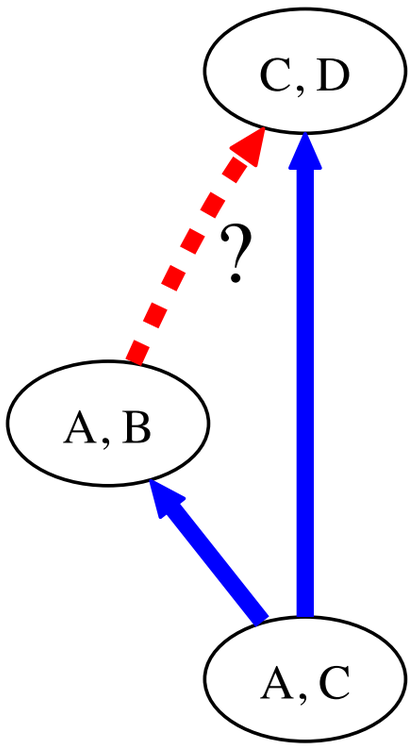
previous graph looked bad, but is it all of the authors some of the time, or some of the authors all of the time?



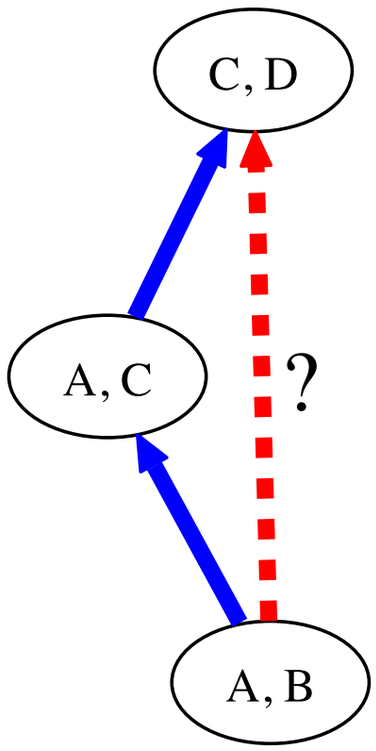
Tale of two authors: edges representing self-copying, cited material, and material recopied without citation are colored **blue**, **green**, and **red**, respectively. The edge thickness increases with the amount of overlap between the two articles. Nodes colored grey are attributed to other authors.


Time

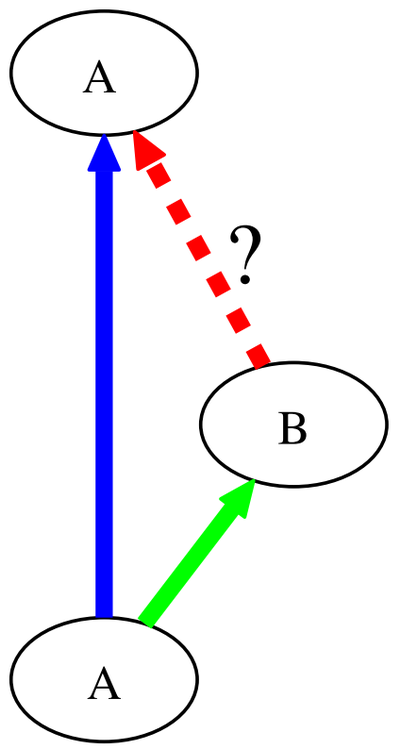

Time



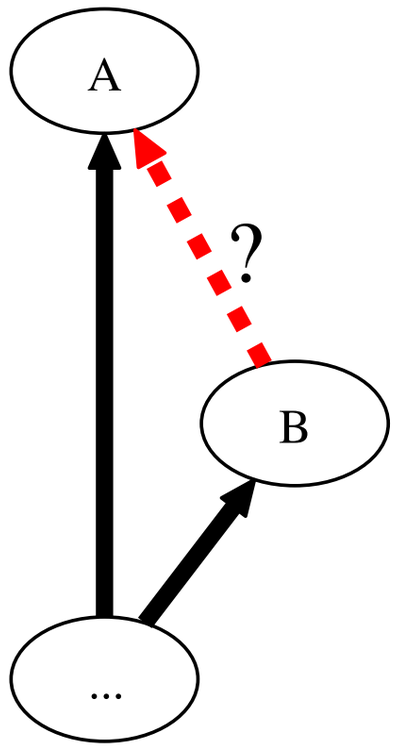
prior co-authored,
but “ok”



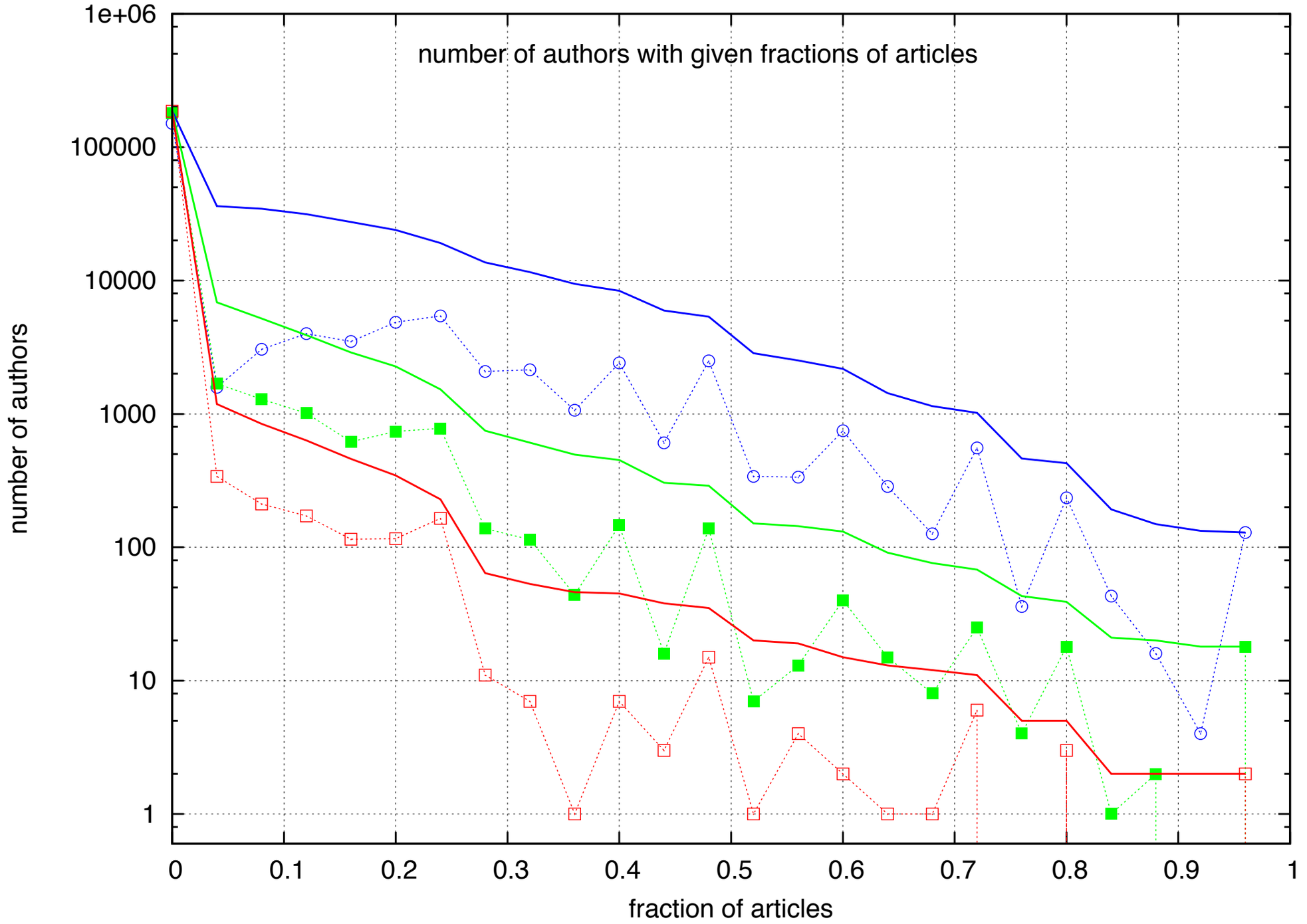
demonstrably not
by C,D but ?



actually self



common source



Authors vs. fraction of articles that include text overlaps. $\sim$ 1720 authors have at least 50% of articles with significant CA text overlap. Of 392,850 authors, only 49,830 have at least 1% CA, only 8990 with CI, and only 1630 contain PI (vast majority OK). Moreover only 10,550, 1130, and 130 authors have at least 25%, resp.

Country correlation?

Depends whether sort by % copied content, links data, % problematic authors, . . .

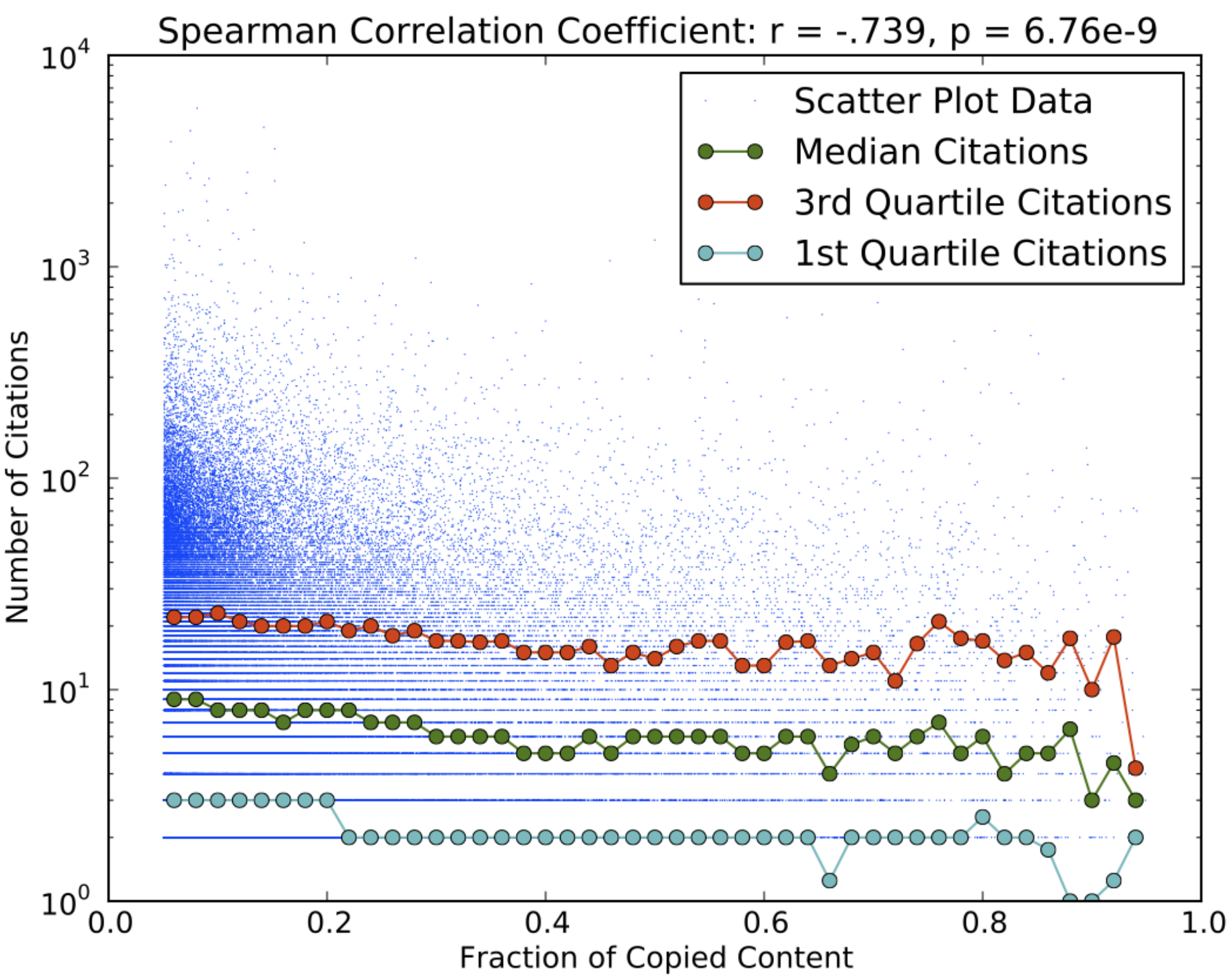
Countries with at least 40 articles:

kz kg cr cy fm bg pk eg ir ma ge su by ro co am gr lt uz tr ru eg cu ee sa lv . . .

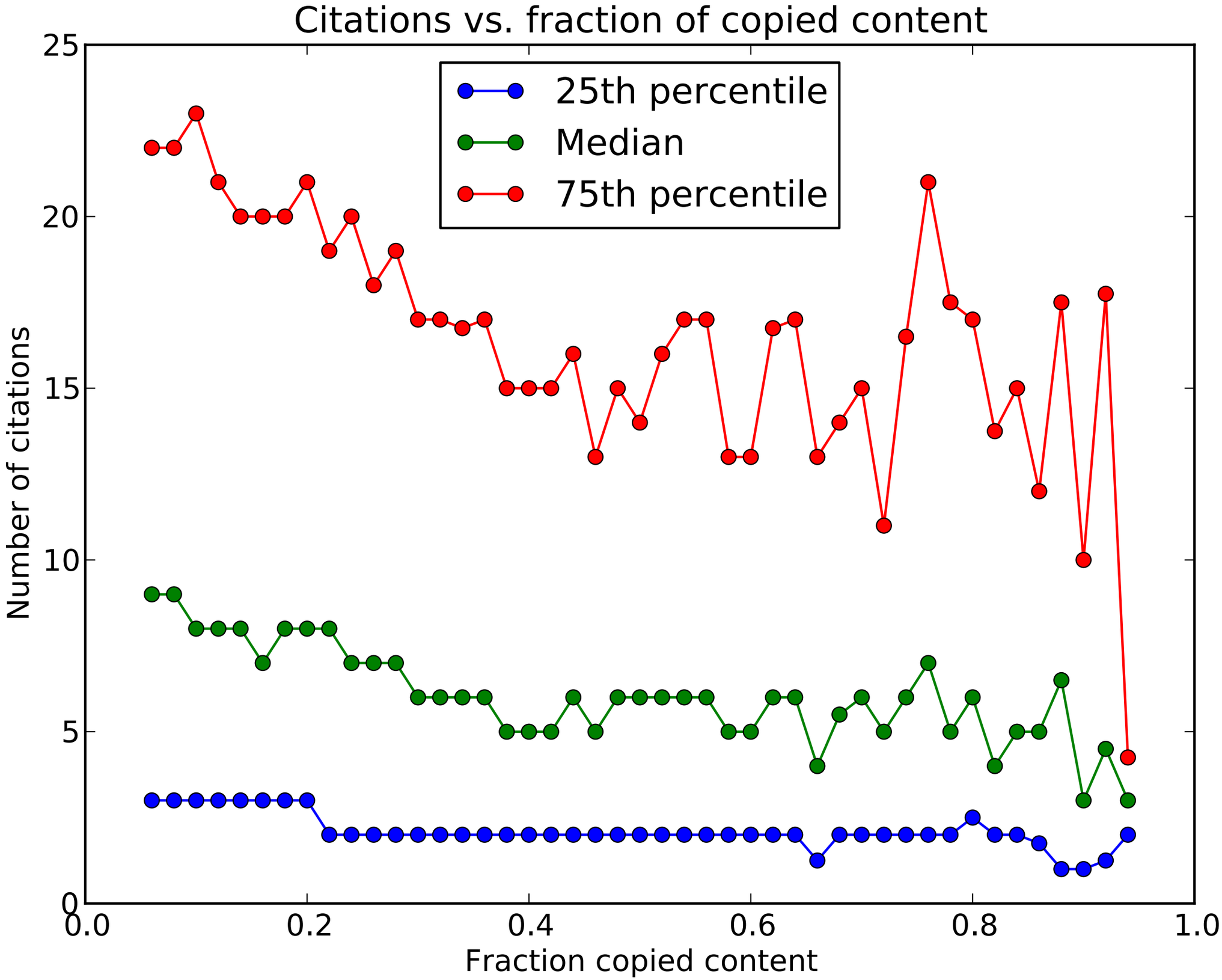
With at least 100 submitting authors, worst offenders are:

bg eg ir ge by ro co am gr . . .

a quality flag?



citations vs. fraction of copied content (blue). median citations vs. fraction of copied content in green, negative correlation (116,490 pre 2011 articles, self-citations removed)


Citations vs. fraction of copied content
25th percentile
Median
75th percentile
Number of citations
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
Fraction copied content
(median and upper/lower quartiles)

Underlying sociology?

non-native speakers of English

(exacerbated by the ease of text reuse in the electronic format? but also easier to detect)

perhaps not willful fraudulence but different (deficient?) educational systems

an act of magic to produce a new idea? of course articles are produced by weaving together texts from existing sources (as was done by mentors)

In summary

- **it's easy, out there, no one has really looked (except for turnitin, moss.stanford.edu , ...)**
- **text overlap not plagiarism (though there are a few instances of duplicate articles by different authors, thoroughly inexplicable)**
- **not the most creative authors**
- **an educational issue re common practice (systematic reuse ok for review articles, as opposed to lecture notes or conf proc?)**
- **(or perhaps that's changing, e.g. wikipedia comment "not necessary to cite dynamically produced content")**
- **uncited reuse rare (question of training, cite but include blocks of text)**
- **lessons for how we train undergrads in modern networked world?**
- **still aren't many comprehensive OA corpora available, but can be done on some? (nature, science, phys rev, pubmedcentral)**

1101.5456

I take this opportunity to express my deep sense of gratitude to my supervisor, Dr. Sanjay Kumar, for his constant encouragement, cooperation and invaluable guidance in the successful accomplishment of this dissertation. I also express my gratitude to Prof. B. K. Dass, Head, Department of Mathematics, University of Delhi for providing necessary facilities and constant encouragement during the course of this study.

I also wish to extend my thanks to all the faculty members of the Department of Mathematics, University of Delhi for their help, guidance and motivation for the work presented in the dissertation. They have always been there for me whenever I needed support from them, providing me critical research insights and answering my questions with their valuable time. Their academic excellence has also been a great value to my dissertation.

I am also thankful to my friends and fellow research scholars (specially Sumit Nagpal, Kuldeep Prakash, Sarika Goyal and Rani Kumari) for their help and discussion during the course of my study. I am also thankful to M.M.Mishra, Assistant Professor, Hansraj college, for his valuable guidance in Latex.

I also wish to express my gratitude to the C.S.I.R for granting me the fellowship which was a great financial assistance in the completion of my M. Phil programme. I am sincerely thankful to my parents for motivating me to do higher studies. I would also like to extend my gratitude to my brothers and sisters for helping me in every possible way and encouraging me to achieve my long cherished goal.

Above all, I thank, The Almighty, for all his blessings bestowed upon me in completing this work successfully.

1407.8478

I take this opportunity to express my deep sense of gratitude to my supervisor, Dr. Sanjay Kumar, for his constant encouragement, cooperation and invaluable guidance in the successful accomplishment of this dissertation. I also express my gratitude to Prof. Ajay Kumar, Head, Department of Mathematics, University of Delhi for providing necessary facilities and constant encouragement during the course of this study.

I also wish to extend my thanks to all the faculty members of the Department of Mathematics, University of Delhi for their help, guidance and motivation for the work presented in the dissertation. They have always been there for me whenever I needed support from them, providing me critical research insights and answering my questions with their valuable time. Their academic excellence has also been a great value to my dissertation. I am also thankful to the organizers of ATM schools of geometry and topology, which I attended in CEMS Almora, NEHU Shillong and HRI Allahabad, which helped me to learn many facts related to this field.

I am also thankful to Prof. Ravi S. Kulkarni, who gave the idea of this work and discussed the problem, and Prof. Anant R. Shastri, for his guidance in better understanding of the subject. I am also thankful to my friends and fellow research scholars (specially Dinesh Kumar and Gopal Datt) for their help and discussion during the course of my study.

I also wish to express my gratitude to the U.G.C. for granting me the fellowship which was a great financial assistance in the completion of my M. Phil program.

I am sincerely thankful to my parents for motivating me to do higher studies and encouraging me to achieve my long cherished goal.

Above all, I thank, The Almighty, for all his blessings bestowed upon me in completing this work successfully.

1306.3408

First and foremost, I wish to extend my gratitude to my supervisors, Professors W. David McComb and Arjun Berera. Without their continued support I would never have completed this thesis. I wish to thank Prof. McComb for his patient guidance and motivation towards research. I thank Prof. Berera for sharing his knowledge and enthusiasm with me, as well as for being approachable with any problems I had. I have learnt a lot from working with both of them.
Particular thanks are due to Dr. Matthew Salewski, for his friendship and many stimulating discussions on the topic of turbulence.
I cannot describe how indebted I am to my wonderful girlfriend, Amanda, whose love and encouragement will always motivate me to achieve all that I can. I could not have written this thesis without her support; in particular, my peculiar working hours and erratic behaviour towards the end could not have been easy to deal with!
Of course, I would never have made it this far without the love and support of my family, particularly my mum and brother, Joe. Their interest (a facade though it may have been!) in my work and pride at my achievements has always been an inspiration.
I could also have not made it through without the many friends I have made along the way. I particularly wish to thank my colleagues and flatmates Gavin and Liam, as living and working with them was a privilege. I also thank Eoin for the many jamming sessions and encouraging the creation of the physics dept. football team, the Feynmen.
When I joined the particle theory group, I was instantly made to feel welcome and included, for which I owe additional thanks to Erik, Claudia, Simone, Thomas and Brian. I extend my thanks and best wishes to the entire PPT corridor and the students and post-docs I got to share lunch, coffee and/or (several) pints with.
I would like to thank Jane Patterson for her kindness and ensuring my PhD career ran smoothly.
I gratefully acknowledge the generosity and support of the Edinburgh Compute and Data Facility. My funding was provided by the STFC, to whom I am eternally grateful for this opportunity.

1408.4411

First and foremost, I wish to extend my gratitude to my supervisors, Prof. Dino Anthony Jaroszynski and Dr. Adam Noble, whom I also find to be a close friend. Without their continued support I would never have completed this thesis. I wish to thank Dr. Noble for his patient guidance and motivation towards research. I thank Prof. Jaroszynski for sharing his knowledge and enthusiasm with me, as well as for being approachable with any problems I had. I have learnt a lot from working with both of them.
Particular thanks are due to Dr. Samuel Yoffe, for his friendship and many stimulating discussions on the topic of radiation reaction and quantum corrections.
I cannot describe how indebted I am to my wonderful wife, Renata, whose love and encouragement will always motivate me to achieve all that I can. I could not have written this thesis without her support; in particular, my peculiar working hours and erratic behaviour towards the end could not have been easy to deal with!
Of course, I would never have made it this far without the love and support of my beloved grandparents. Their interest in my work and pride at my achievements has always been an inspiration.
I could also have not made it through without the many friends I have made along the way. I particularly wish to thank my colleagues at the University of Strathclyde and senior researchers at Lancaster University as working with them was a privilege. I also thank my cat Fluffy for keeping me smiling at the downfalls of my project.
When I joined the SILIS group, I was instantly made to feel welcome and included, for which I owe additional thanks to Bernhard, Gaurav, Enrico, Silvia and Gregory. I extend my thanks and best wishes to all the students and post-docs I got to share lunch, coffee and/or (several) pints with.
I would like to thank Kirsten Munro, Catherine Cheshire and Lynn Gilmour for their kind approach in dealing with all the administrative matters and ensuring my PhD ran smoothly.
I would like to thank Jane Patterson for her kindness and ensuring my PhD career ran smoothly.
I gratefully acknowledge the generosity and support of the Scottish Universities Physics Alliance (SUPA) and University of Strathclyde, who provided me with a Prize Studentship enabling me to undertake this PhD. I am eternally grateful for this opportunity.

"Signal is a physical quantity that vacillates with time, space or any other alienated variable. ..."

"Signal is a physical quantity that vacillates with time, space or any other alienated variable."

("Signals are physical quantities that change as a function of time, space, or some other independent variable.")

Spectrum conflict management,..., and the (thus far incomplete) Search for Extraterrestrial Intelligence (SETI) all alleviate on ferreting the propinquity of radio signals of concealed frequency, power, and modulation.

... all rely on detecting the presence of radio signals of unknown frequency, power, and modulation."

or

"escalates the rate at which sampled signals can purl through the processor"

"increases the rate at which sampled signals can flow through the processor".

(later two taken from intro of 2001 thesis at Monterey Naval Postgraduate School:
Charles T. Dorcey, "FFT-based spectrum analysis using a Digital Signal Processor.")

or from http://www.tutorialsweb.com/rf-measurements/spectrum-analyzer.htm :

Spectrum analyzer is a device used to [examine -> anatomize] the spectral composition of electric, acoustic or optical waveform [7]. It is a wideband and [very -> eminent] sensitive receiver. It works on the [principle -> ethic] of super heterodyne receiver which [converts -> transmogrifies] higher frequencies to measurable quantities. Received frequency spectrum is [slowly -> apathetically] swept through a range of preselected frequencies converting the selected frequency to a measurable and [displaying -> unveiling] on the CRT. These are [capable -> adept] in measuring the frequency response of power levels as low as -120dbm.

or finally:

“The proposed work in this thesis is having a lot of potential for further research in the area of [edge detection] using different paradigm making the work more versatile and flexible."

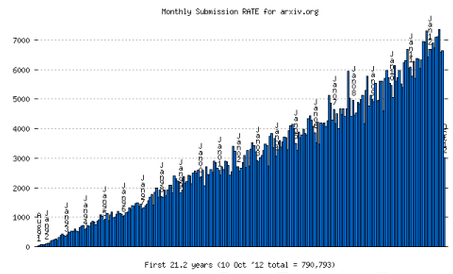
In partial conclusion

(Despite ten remaining slides)

Tried to convince you there are various interesting analyses of full text corpora using physicist methods

(haven't even discussed usage data)

Perhaps next-generation on-line tools, manifesting the power of open repositories, will catalyze further growth



Already some useful analysis tools, and some sociological observations re a global research community, with possibilities for social engineering

Future

Active + Passive user participation in bottom-up approach to QC

- **actively add tags, links; contribute to ontologies, correct wiki entries**
- **passively ingest readership, bookmarking, annotation behavior**

Incentive Question: expertise-intensive efforts beyond conventional journal publication (annotation, linkage, . . .) = scholarly achievement?

articles + blog commentary $\longrightarrow$ more modular objects

glue databases together into knowledge structure

Goal: semi-supervised, self-incentivized, self-maintaining knowledge structure, navigated via synthesized concepts, w/o redundancy/ambiguity, sourced, authenticated, highlighted for novelty

(But: even find videos?)

Fantasy

Reflect for a moment

Current practice:

- **free access articles, background material from authors, slide presentations, video, related software, on-line animations, blog discussions, 3rd party notes, microblogged seminars, captured video feed, random factoids, collective wiki-exegesis**
- **course websites, e-mail, course blogs, wiki for notes**

New expectations (harvest all related, activity maps, concept browse).

Collapse internet resources to subset of unique ideas, authenticated.

Marketplace for preresearch barter of tools, resources, capabilities.

Authoring tools.

New generation of users.

Essential questions

How will the analog of NCBI/PubMedCentral be provided for other communities? (Who? With whose money?)

Common web service protocols, common languages (e.g., for manipulating, visualizing data), data interchange standards

Distributed version for other fields

networked resources $\Rightarrow$ new nonlinear reading strategies

ubiquitous mobile devices $\Rightarrow$ new usage of short-, long-term memory

Qualitatively new research and cognitive methodologies, transformation in the way we process scientific information, with academic community as role model for the creation and dissemination of knowledge to the public