

Quantifying conceptual novelty in the biomedical literature

Shubhanshu Mishra and Vetle I. Torvik

School of Information Sciences

University of Illinois at Urbana-Champaign

<http://abel.lis.illinois.edu/gimli/>




I L L I N O I S

School of

Information Sciences

The iSchool at Illinois

nov·el·ty

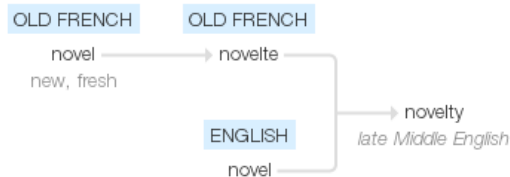
/'nävēltē/ 

noun

noun: novelty

1. the quality of being new, original, or unusual.
"the novelty of being a married woman wore off"
synonyms: [originality](#), newness, freshness, unconventionality, unfamiliarity; [More](#)
 - a new or unfamiliar thing or experience.
plural noun: **novelties**
"in 1914 air travel was still a novelty"
 - denoting something intended to be amusing as a result of its new or unusual quality.
modifier noun: **novelty**
"a novelty teapot"
2. a small and inexpensive toy or ornament.
"he bought chocolate novelties to decorate the Christmas tree"
synonyms: [knickknack](#), [trinket](#), [bauble](#), [toy](#), [trifle](#), [gewgaw](#), [gimcrack](#), [ornament](#), [kickshaw](#)
"we sell seasonal novelties"

Origin

late Middle English: from Old French *novelte*, from *novel* 'new, fresh' (see [novel](#)²).Translate novelty to

Use over time for: novelty



Which papers were the early ones on HIV?

Which were the papers which first combined Cancer (Neoplasms) and Data Mining?

What are the novel ideas in a given paper?

Background



Scientific progress

- Science moves forward through innovation and work on novel concepts (Kuhn, 1970)
- Novelty, originality, and priority are important concepts related to scientific publishing (Morgan, 1985)
- Most citation classics contain new hypothesis, previously reported methods and new results (Dirk, 1999)

Author age

- Younger authors are more likely to build on novel ideas and experienced co-authors contribute as well (Packalen, 2015)

Impact

- Highest impact articles cite high number of conventional combination of journals and few novel combinations. (Uzzi, 2013)
- Highly novel articles are at higher risk of rejection but can reap greater rewards. (Trapido, 2015)
- Citation metrics are biased against novel articles (Wang et al., 2015)

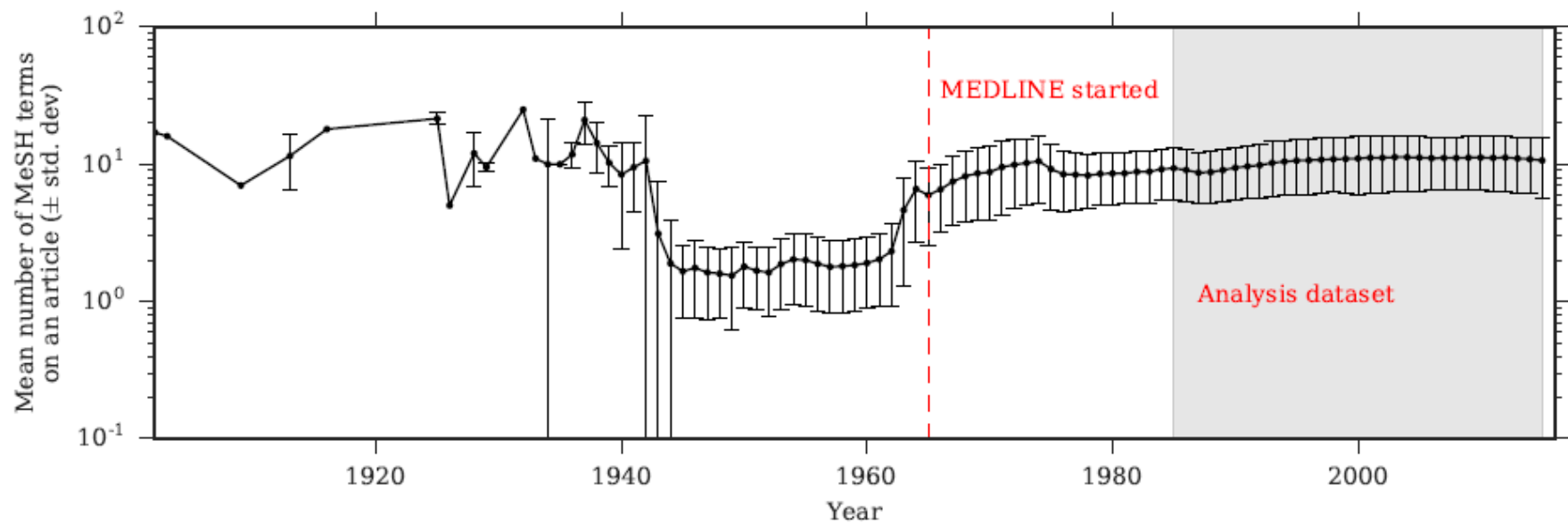
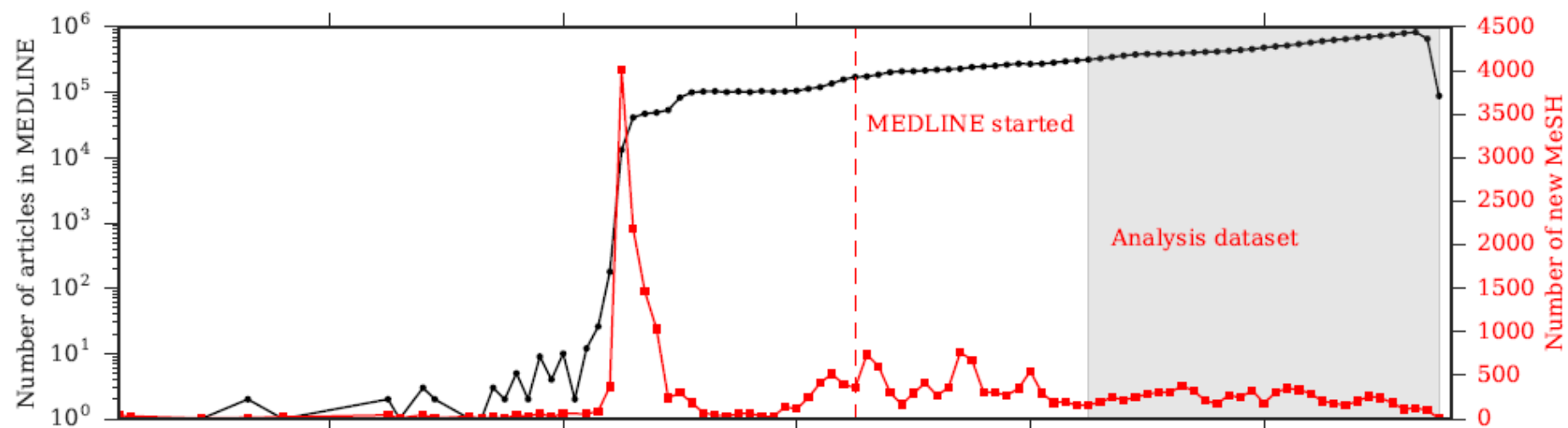
Data



I L L I N O I S

School of
Information Sciences

The iSchool at Illinois



Data



- 22.3 million MEDLINE articles
- 27,249 Medical Subject Headings (MeSH)
- On average 9-10 MeSH per article since 1985
- Data before 1985 is noisy for several reasons:
 - MEDLINE started in earnest 1965 (articles near that year will look more novel)
 - ~ 2 MeSH terms per article prior to 1965
 - ~ 4,000 MeSH terms first used on articles published around 1945

MeSH explosion



- MeSH terms are arranged in a hierarchy
 - However, each term may have multiple parent terms
 - Neoplasms [C04] → Neoplasms by Site [C04.588] → **Breast Neoplasms [C04.588.180]**
 - Skin and Connective Tissue Diseases [C17] → Skin Diseases [C17.800] → Breast Diseases [C17.800.090] → **Breast Neoplasms [C17.800.090.500]**
- The temporal profile of a given MeSH term **C** is based on the aggregate of all the terms for which **C** is an ancestor
 - E.g. an article on Breast Neoplasms is counted as an article on Neoplasms.

Temporal profiling of concepts

Building block for measuring novelty.

How old is a given concept at a particular time point?

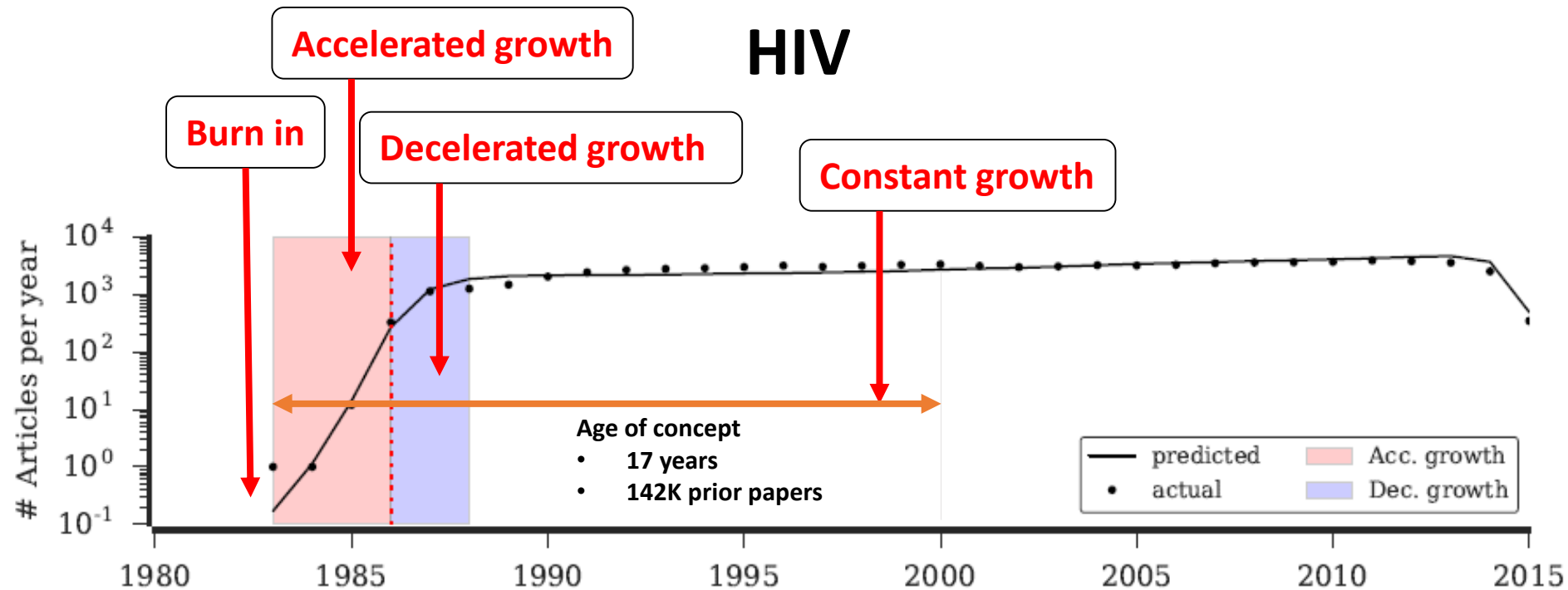


I L L I N O I S

School of
Information Sciences

The iSchool at Illinois

HIV



- AIDS was first clinically observed in 1981 (source Wikipedia)
- US Center for Disease Control (CDC) renamed it to AIDS in 1982
- First 2 papers published in the same year 1983
- LAV and HTLV-III are also names used for referring to HIV-1 (a subtype of HIV)
- Terms renamed to HIV in 1986

Empirical novelty scores



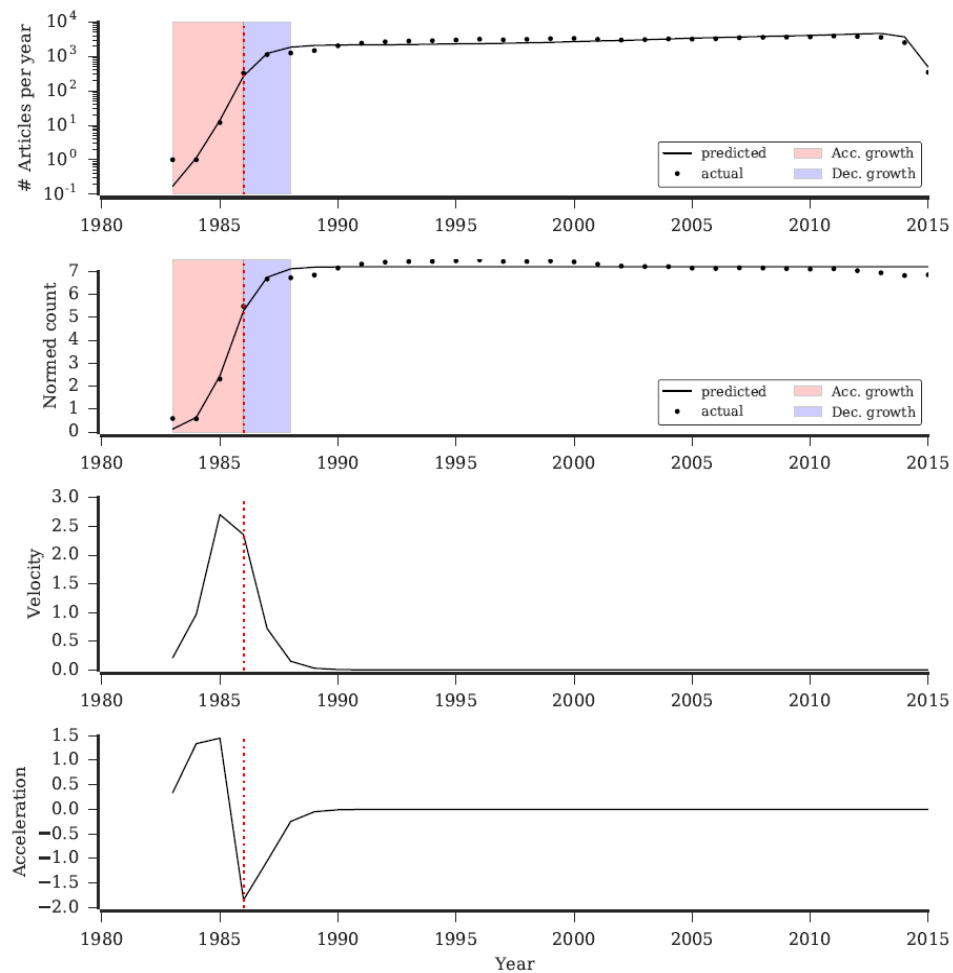
Types of novelty:

- Individual concept
- Pair of concepts (combinatorial)

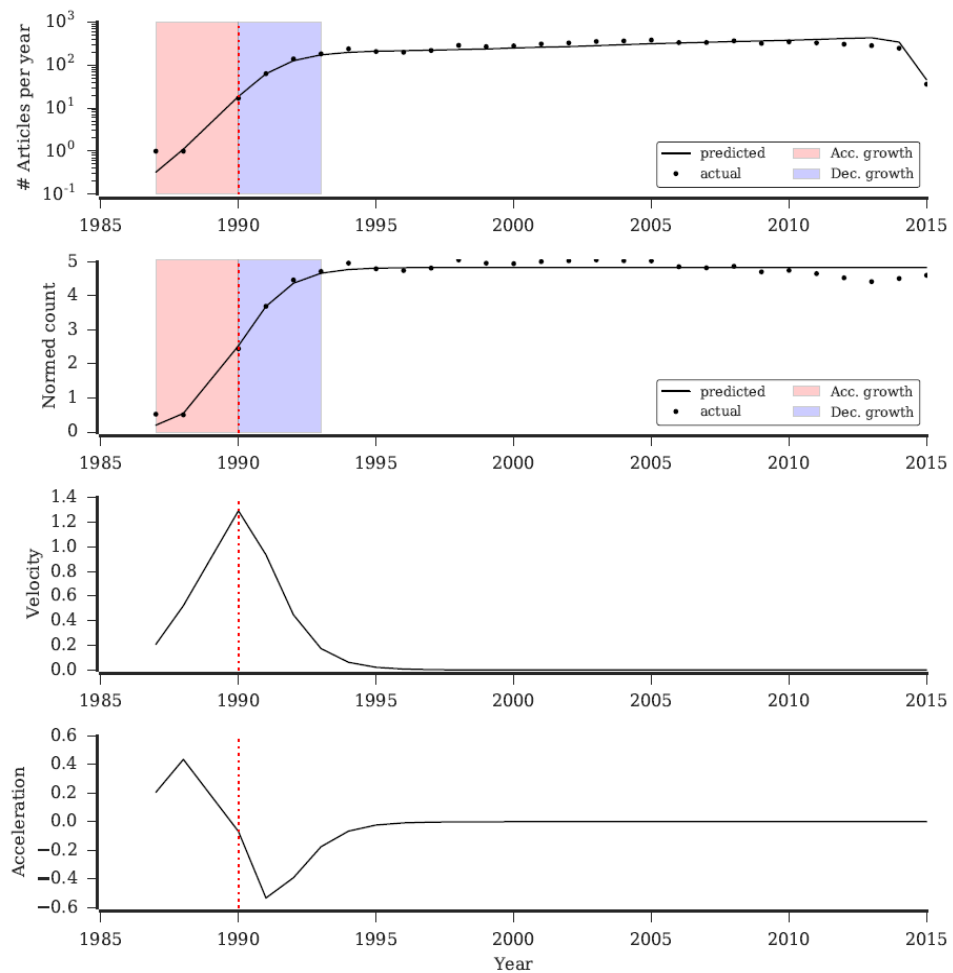
Units of measurement:

- Years since first appearance (**Time novelty**)
- Prior articles since first appearance (**Volume novelty**)

HIV



AIDS Vaccines



Model scores



Burn-In Phase

- Topic is new, publication rate is small, and growth is marginal.

Accelerating Growth Phase

- Topic is bursting, publication rate is rapidly increasing.

Decelerating Growth Phase

- Publication rate is increasing but starting to stabilize.

Constant Growth Phase

- Publication rate has stabilized.

Modeling temporal growth of a concept



- Model the articles published on a concept in a given year
- Logistic growth model
 - $f(t) = \frac{N_o}{1 + \exp(-(t - t_o)/s)}$
 - N_o : asymptotic max number of articles that can be published on the concept in a given year
 - t_o : age of the concept (years) when the concept goes from accelerating growth to decelerating growth phase
 - s : temporal spread of the accelerating and decelerating phase of the concept

Novelty of an article

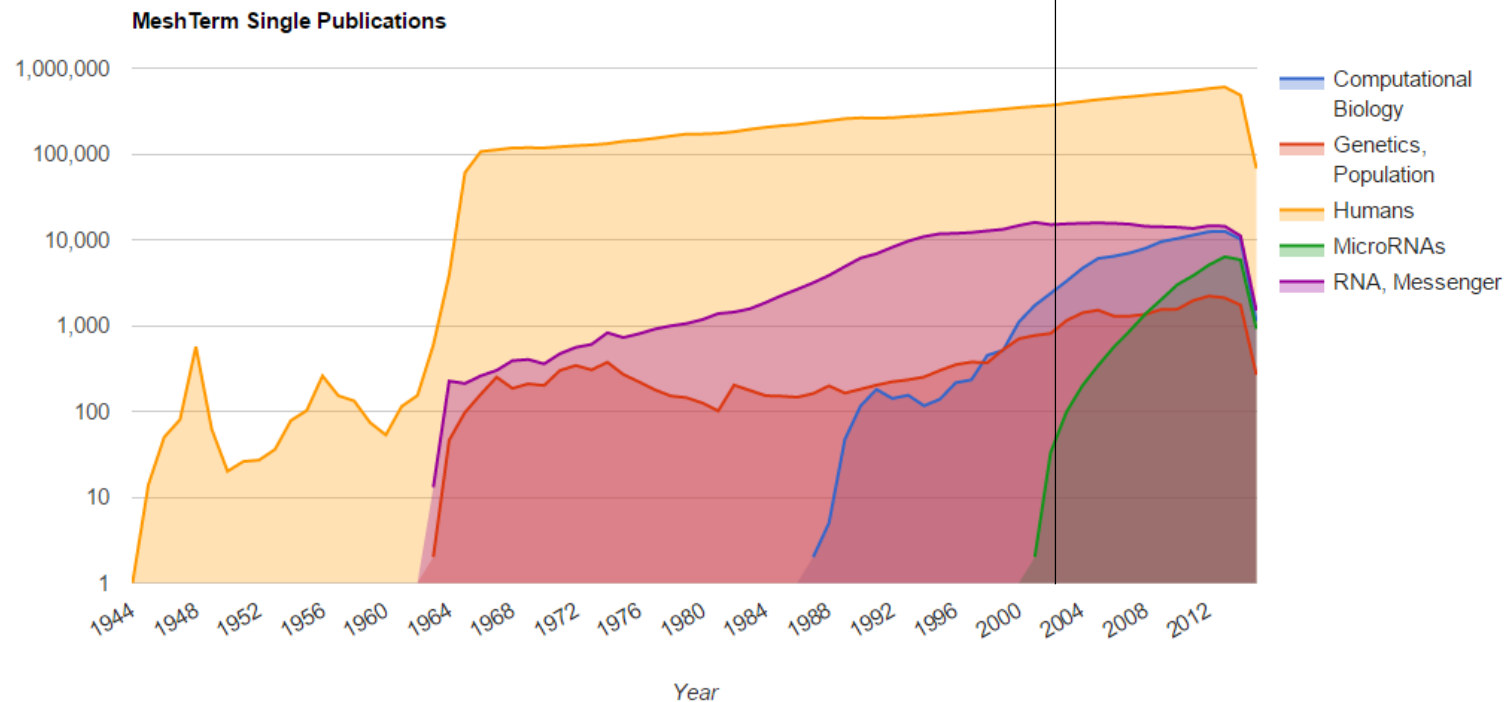
How to use the temporal profile of concepts to identify the novelty of an article?



I L L I N O I S

School of
Information Sciences

The iSchool at Illinois



PMID: 15453917

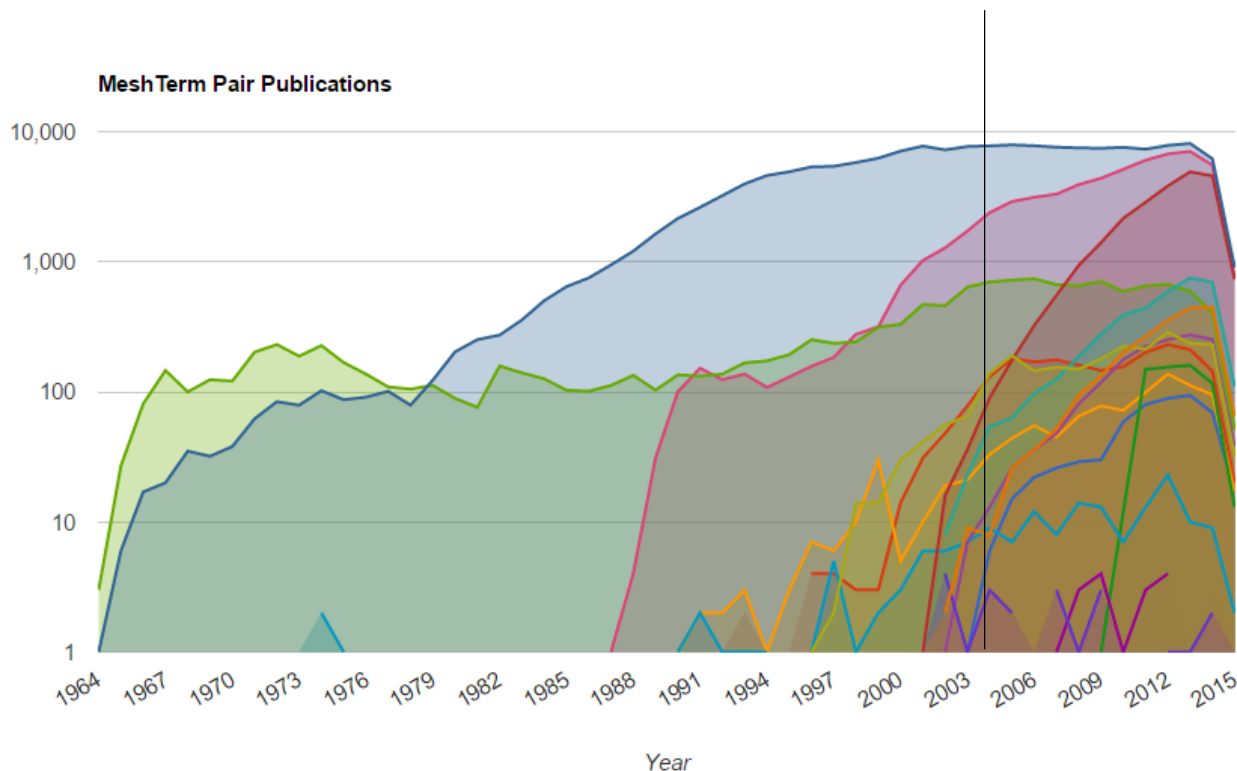
YEAR: 2004

TITLE: A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions.

MESH: Computational Biology; Genetics, Population; Humans; MicroRNAs; RNA, Messenger

MeshTerm	Time	Vol
MicroRNAs	3	331
Genetics, Population	41	12,819
Computational Biology	17	15,412
RNA, Messenger	43	212,862
Humans	60	8,755,350

Category	Time[First Pub]	Vol[First Pub]	Velocity[log norm.]	Acceleration[log norm.]
Chemicals	MicroRNAs (3)	MicroRNAs (331.0)	MicroRNAs (0.778248)	MicroRNAs (-0.0186093)
InfoSci	Computational Biology (17)	Computational Biology (15412.0)	Computational Biology (0.180842)	Computational Biology (-0.0202041)
Organisms	Humans (60)	Humans (8755350.0)	Humans (0.00367279)	Humans (-0.000106644)



Category	Time[First Pub]	Vol[First Pub]
Chemicals – Chemicals	MicroRNAs - RNA, Messenger (2)	RNA, Messenger - MicroRNAs (19)
Chemicals – InfoSci	Computational Biology - MicroRNAs (1)	Computational Biology - MicroRNAs (7)
Organisms – Chemicals	Humans - MicroRNAs (3)	Humans - MicroRNAs (143)
Organisms - InfoSci	Humans - Computational Biology (17)	Humans - Computational Biology (8,793)

Mesh1	Mesh2	Time	Vol
Genetics, Population	MicroRNAs	2	2
Computational Biology	MicroRNAs	1	7
RNA, Messenger	Genetics, Population	4	9
RNA, Messenger	MicroRNAs	2	19
MicroRNAs	Computational Biology	2	21
Genetics, Population	RNA, Messenger	33	55
MicroRNAs	RNA, Messenger	2	85
Humans	MicroRNAs	3	143
Genetics, Population	Computational Biology	14	153
Computational Biology	RNA, Messenger	11	318
RNA, Messenger	Computational Biology	8	362
Humans	Genetics, Population	40	7,801
Humans	Computational Biology	17	8,793
Humans	RNA, Messenger	40	89,589

PMID: 15453917

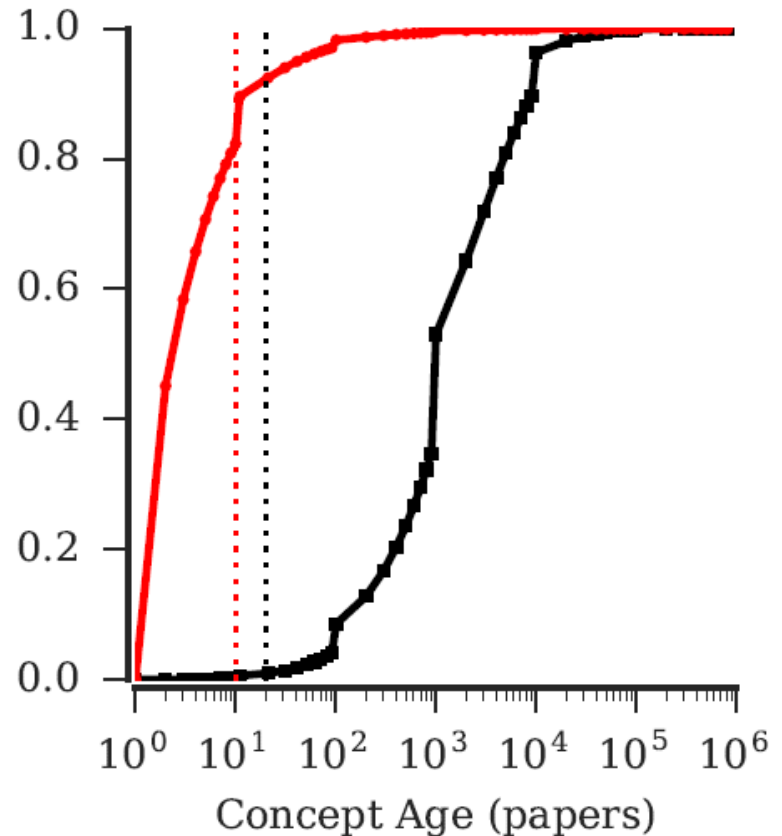
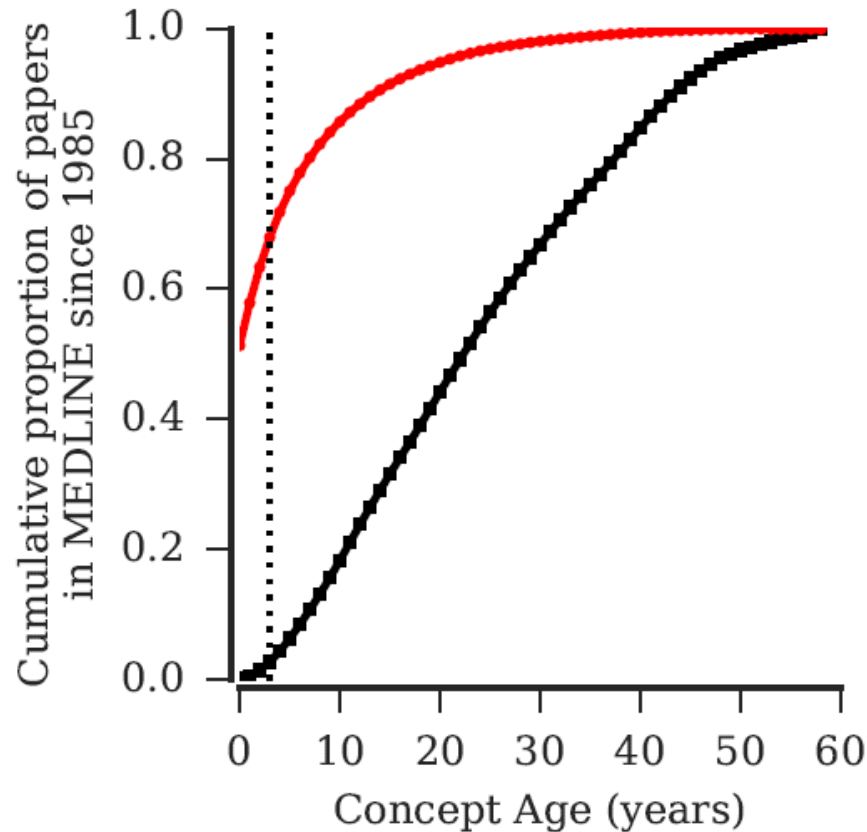
YEAR: 2004

TITLE: A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions.

MESH: Computational Biology; Genetics, Population; Humans; MicroRNAs; RNA, Messenger

Type of Novelty Score

—●— Individual concept —●— Pair of concepts



Novelty type	% of novel papers	
	By time	By volume
Individual concept	2.73% (<3)	1.0% (<20)
Pair of concepts	68.0% (<3)	89.6% (<10)

Growth	% of novel papers
Accelerating	61.1%
Decelerating	38.9%

Novelty and author career

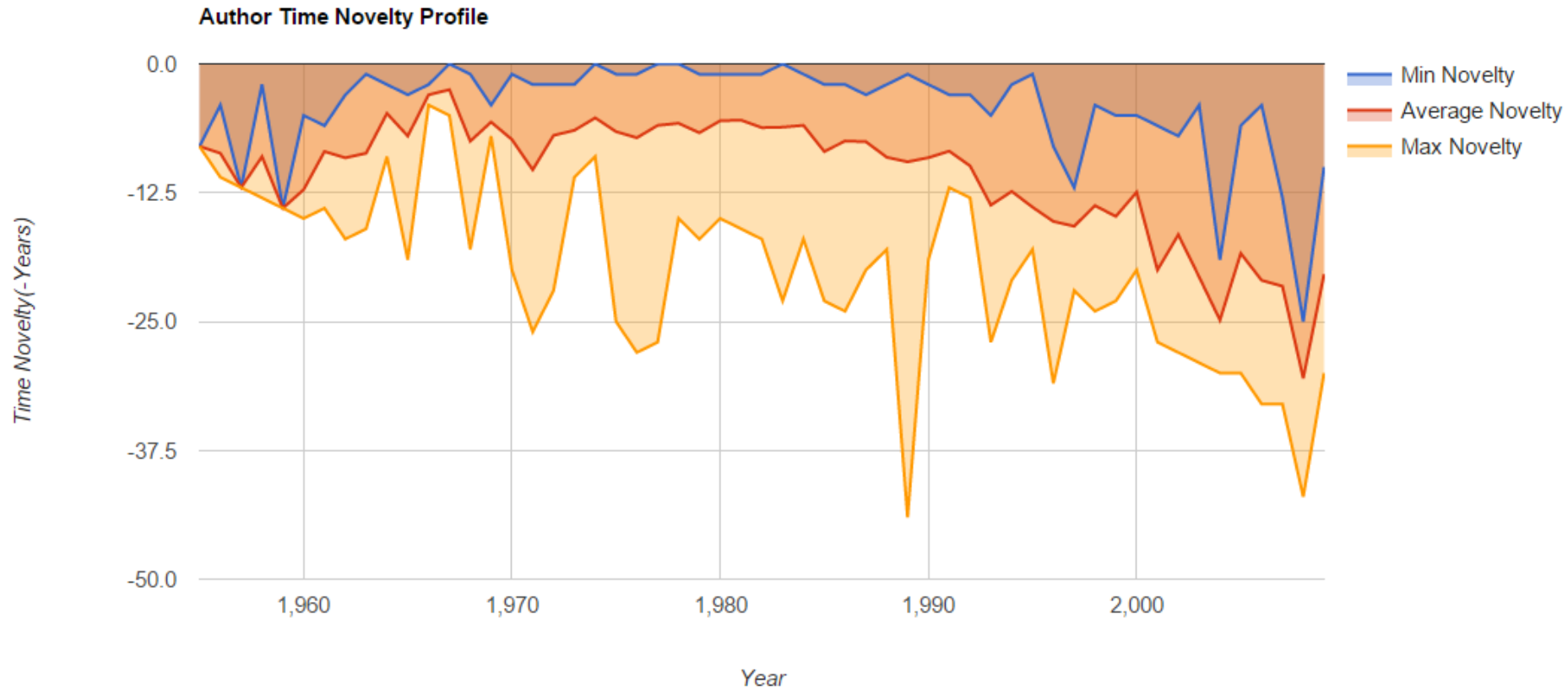
How does the novelty of an author's papers change over their career?



I L L I N O I S

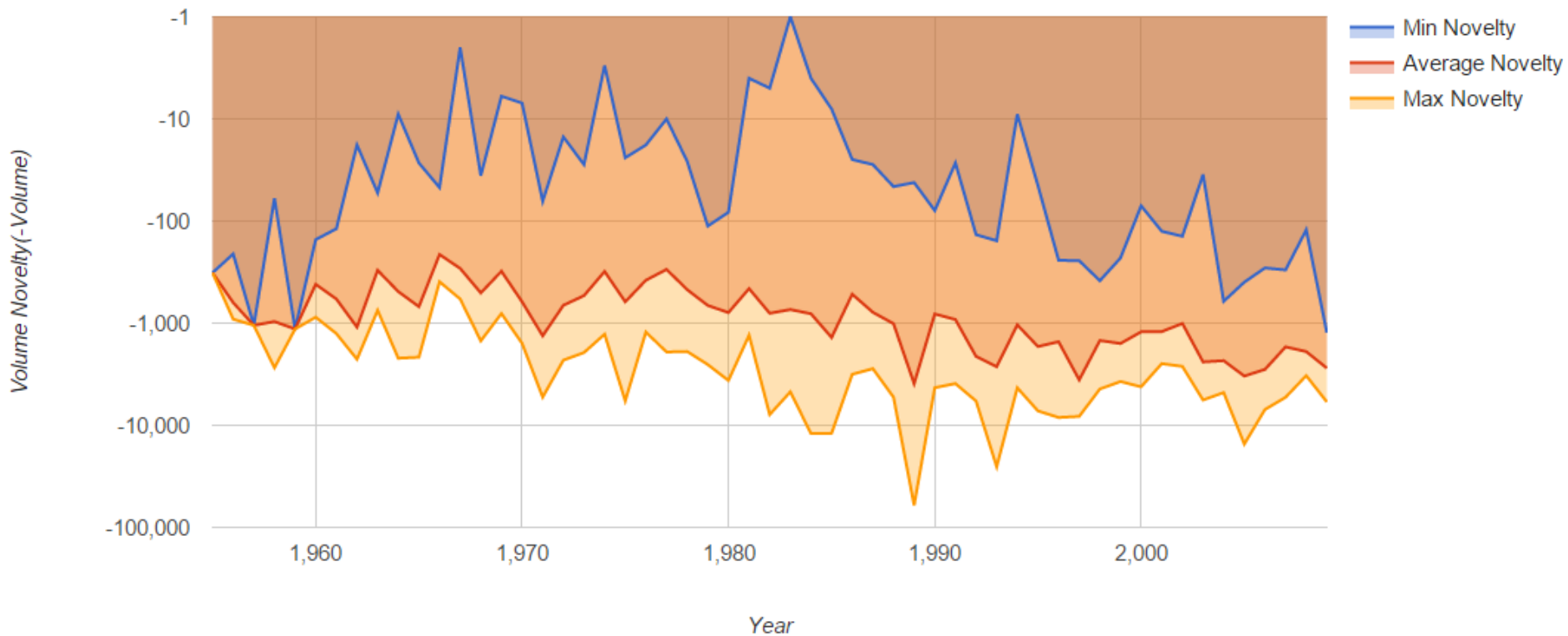
School of
Information Sciences

The iSchool at Illinois



Erminio Costa (700+ papers in Author-ity 2009 dataset)
http://abel.lis.illinois.edu/gimli/author_profile?au_id=13258841_3

Author Volume Novelty Profile



Erminio Costa (700+ papers in Author-ity 2009 dataset)

http://abel.lis.illinois.edu/gimli/author_profile?au_id=13258841_3

Distribution of authors with decreasing novelty across career



Novelty type	% of authors with decreasing novelty	
	Average age	Minimum age
Concept (by time)	84%	59%
Concept (by volume)	85%	58%
Concept-pair (by time)	64%	67%
Concept-pair (by volume)	56%	68%

150,000 prolific authors (> 50 papers), on average the mean novelty of articles of 80% authors decreases with their career, this might indicate specialization on a concept or diversification to other concepts.

Novelty and Impact

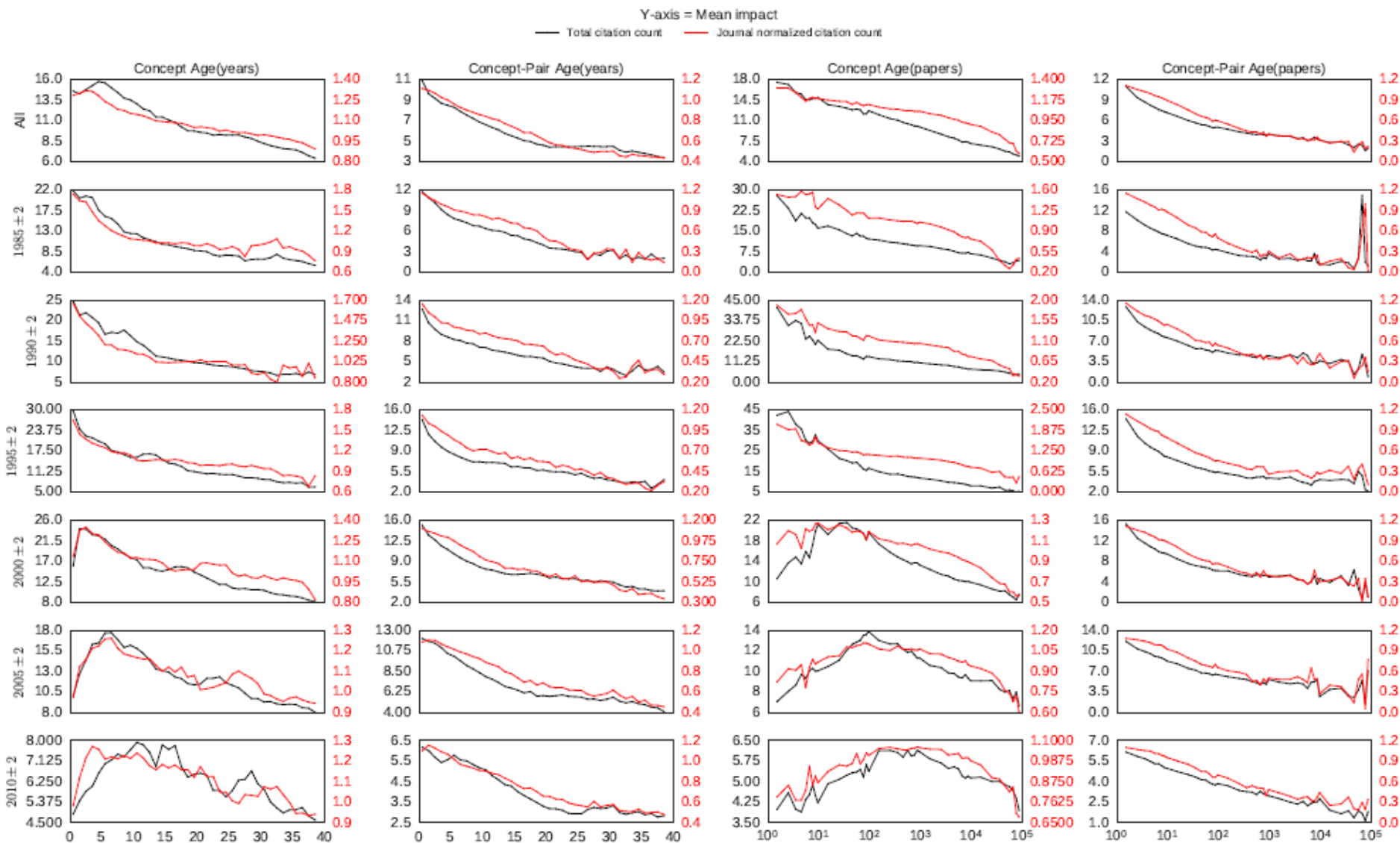
How are novel papers cited?



I L L I N O I S

School of
Information Sciences

The iSchool at Illinois



X axis represents the novelty of the youngest concept (or pair of concepts) on an article. Y axis represents the mean impact (as measured using total citations and journal normalized citations)

How impact is affected by novelty?

- On average being novel higher impact
- However, the impact varies with the year and journal of publication
- This agree with earlier findings that citation indices which only use the citations in the first few years are biased against novel articles (Wang, 2015)
- Very novel articles published in recent years have on average gained less citations than moderately novel articles
- Controlling for the average citations of the publishing journal does not affect the overall pattern significantly

Implications to science



- Should everyone work on novel concepts or should people replicate and extend work done on existing ones?
- Should search results be ordered by citations or should they account novelty of concepts on the paper?
- How to find novel articles and authors who publish those, on user chosen concepts or concept combinations?

Concluding remarks



- Novelty can be quantified using concepts on articles
- Combinatorial novelty is the norm in science and captures 90% of the articles
- Correlations of novelty with author age and impact are complex.

Open Questions

- How much time does it take for a novel paper to get a large percentage of its citation?
- How do concepts co-evolve over time?
- Who publishes most novel work? (gender, ethnicity, country)

Resources



- Main web page: <http://abel.lis.illinois.edu/gimli/>
- Temporal profile of a MeSH term (uses MeSH 2015 term names):
http://abel.lis.illinois.edu/gimli/mesh_profile?mesh_term=HIV
- Novelty of a MEDLINE article (uses PubMed ID):
http://abel.lis.illinois.edu/gimli/novelty?pubmed_id=15453917
- Novelty of an author (uses Author-ity ID):
http://abel.lis.illinois.edu/gimli/author_profile?au_id=13258841_3
- Source code: <https://github.com/napsternxg/Novelty>

Submit

PMID:	15922829
YEAR:	2005
TITLE:	Mammalian microRNAs derived from genomic repeats.
MESH:	<div>AnimalsBase SequenceComputational BiologyDNA Transposable ElementsExpressed Sequence TagsHumansLong Interspersed Nucleotide ElementsMiceMicroRNAMolecular Sequence DataRNA, MessengerRatsRepetitive Sequences, Nucleic Acid</div>
AUIDS:	<div>207390_110944406_4</div>

FIND THE TOPICAL EXPERTISE OF AUTHORS ON THIS PAPER

Most Novel Mesh Terms per category

Show

10

entries

Search:

Category	Volume	Predicted Volume[log norm.]	Velocity[log norm.]	Acceleration[log norm.]	Time[First Pub]	Vol[First Pub]
Chemicals	MicroRNAs (341)	MicroRNAs (292.775)	RNA, Messenger (0.0236526)	RNA, Messenger (-0.0015207)	MicroRNAs (4)	MicroRNAs (672.0)
InfoSci	Computational Biology	Computational Biology	Molecular Sequence	Molecular Sequence Data	Computational	Computational Biology

15922829

Submit

PMID: 15922829

YEAR: 2005

TITLE: Mammalian microRNAs derived from genomic repeats.

MESH: [Animals](#) [Base Sequence](#) [Computational Biology](#) [DNA Transposable Elements](#) [Expressed Sequence Tags](#) [Humans](#) [Long Interspersed Nucleotide Elements](#) [Mice](#) [MicroRNAs](#) [Molecular Sequence Data](#) [RNA, Messenger](#) [Rats](#) [Repetitive Sequences, Nucleic Acid](#)








AUIDS: [207390_1](#) [10944406_4](#)

FIND THE TOPICAL EXPERTISE OF AUTHORS ON THIS PAPER

Most Novel Mesh Terms per category

Show 10 entries

Search:

Category 	Volume 	Predicted Volume[log norm.] 	Velocity[log norm.] 	Acceleration[log norm.] 	Time[First Pub] 	Vol[First Pub] 
Chemicals	MicroRNAs (341)	MicroRNAs (292.775)	RNA, Messenger (0.0236526)	RNA, Messenger (-0.0015207)	MicroRNAs (4)	MicroRNAs (672.0)

15922829

Submit

PMID: 15922829

YEAR: 2005

TITLE: Mammalian microRNAs derived from genomic repeats.

MESH: [Animals](#) [Base Sequence](#) [Computational Biology](#) [DNA Transposable Elements](#) [Expressed Sequence Tags](#) [Humans](#) [Long Interspersed Nucleotide Elements](#) [Mice](#)
[MicroRNAs](#) [Molecular Sequence Data](#) [RNA, Messenger](#) [Rats](#) [Repetitive Sequences, Nucleic Acid](#)








AUIDS: [207390_1](#) [10944406_4](#)

FIND THE TOPICAL EXPERTISE OF AUTHORS ON THIS PAPER

Most Novel Mesh Terms per category

Show entries

Search:

Category 	Volume 	Predicted Volume[log norm.] 	Velocity[log norm.] 	Acceleration[log norm.] 	Time[First Pub] 	Vol[First Pub] 
Chemicals	MicroRNAs (341)	MicroRNAs (292.775)	RNA, Messenger (0.0236526)	RNA, Messenger (-0.0015207)	MicroRNAs (4)	MicroRNAs (672.0)
InfoSci	Computational Biology	Computational Biology	Molecular Sequence	Molecular Sequence Data	Computational	Computational Biology

Acknowledgements



Research reported in this publication was supported in part by the National Institute on Aging of the NIH (Award Number P01AG039347) and the Directorate for Education and Human Resources of the NSF (Award Number 1348742). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.



We all have questions
<http://abel.lis.illinois.edu/gimli/>

A day may come when you have no
questions ---

But, it is not **THIS** day.



I L L I N O I S

School of
Information Sciences

The iSchool at Illinois