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/
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

<table>
<thead>
<tr>
<th>Scientific progress</th>
<th>Author age</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science moves forward through innovation and work on novel concepts (Kuhn, 1970)</td>
<td>Younger authors are more likely to build on novel ideas and experienced co-authors contribute as well (Packalen, 2015)</td>
<td>Highest impact articles cite high number of conventional combination of journals and few novel combinations. (Uzzi, 2013)</td>
</tr>
<tr>
<td>Novelty, originality, and priority are important concepts related to scientific publishing (Morgan, 1985)</td>
<td></td>
<td>Highly novel articles are at higher risk of rejection but can reap greater rewards. (Trapido, 2015)</td>
</tr>
<tr>
<td>Most citation classics contain new hypothesis, previously reported methods and new results (Dirk, 1999)</td>
<td></td>
<td>Citation metrics are biased against novel articles (Wang et al., 2015)</td>
</tr>
</tbody>
</table>
Data
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] \( \rightarrow \) Neoplasms by Site [C04.588] \( \rightarrow \) Breast Neoplasms [C04.588.180]
    - Skin and Connective Tissue Diseases [C17] \( \rightarrow \) Skin Diseases [C17.800] \( \rightarrow \) Breast Diseases [C17.800.090] \( \rightarrow \) 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?
• 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**)
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

\[ f(t) = \frac{N_o}{1 + \exp\left(-\frac{(t - t_o)}{s}\right)} \]

- \( N_o \) : 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?
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MicroRNAs</td>
<td>3</td>
<td>331</td>
<td>MicroRNAs (0.778248)</td>
<td>(-0.0186093)</td>
</tr>
<tr>
<td>Genetics, Population</td>
<td>41</td>
<td>12,819</td>
<td>Computational Biology</td>
<td></td>
</tr>
<tr>
<td>Computational Biology</td>
<td>17</td>
<td>15,412</td>
<td>Computational Biology</td>
<td></td>
</tr>
<tr>
<td>RNA, Messenger</td>
<td>43</td>
<td>212,862</td>
<td>Computational Biology</td>
<td>(-0.0202041)</td>
</tr>
<tr>
<td>Humans</td>
<td>60</td>
<td>8,755,350</td>
<td>Humans (0.00367279)</td>
<td>Humans (-0.000106644)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemicals</td>
<td>MicroRNAs (3)</td>
<td>MicroRNAs (331.0)</td>
<td>MicroRNAs (0.778248)</td>
<td>(-0.0186093)</td>
</tr>
<tr>
<td></td>
<td>Computational Biology</td>
<td></td>
<td>Computational Biology</td>
<td></td>
</tr>
<tr>
<td>InfoSci</td>
<td>Computational Biology (17) (15412.0)</td>
<td></td>
<td>(0.180842)</td>
<td>(-0.0202041)</td>
</tr>
<tr>
<td>Organisms</td>
<td>Humans (60)</td>
<td>Humans (8755350.0)</td>
<td>Humans (0.00367279)</td>
<td>Humans (-0.000106644)</td>
</tr>
</tbody>
</table>
A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions.

**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
<table>
<thead>
<tr>
<th>Novelty type</th>
<th>% of novel papers</th>
<th>Growth</th>
<th>% of novel papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>By time</td>
<td>By volume</td>
<td>Accelerating</td>
</tr>
<tr>
<td>Individual concept</td>
<td>2.73% (&lt;3)</td>
<td>1.0% (&lt;20)</td>
<td>61.1%</td>
</tr>
<tr>
<td>Pair of concepts</td>
<td>68.0% (&lt;3)</td>
<td>89.6% (&lt;10)</td>
<td></td>
</tr>
</tbody>
</table>
Novelty and author career

How does the novelty of an author’s papers change over their career?
Erminio Costa (700+ papers in Author-ity 2009 dataset)
http://abel.lis.illinois.edu/gimli/author_profile?au_id=13258841_3
Erminio Costa (700+ papers in Author-ity 2009 dataset)
http://abel.lis.illinois.edu/gimli/author_profile?au_id=13258841_3
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?
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
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
AUlDS: 207360_1 1094466_4

FIND THE TOPICAL EXPERTISE OF AUTHORS ON THIS PAPER

Most Novel Mesh Terms per category

Show 10 entries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemicals</td>
<td>MicroRNAs (341)</td>
<td>MicroRNAs (292.775)</td>
<td>RNA, Messenger (0.0235525)</td>
<td>RNA, Messenger (-0.0015207)</td>
<td>MicroRNAs (4)</td>
<td>MicroRNAs (672.0)</td>
</tr>
<tr>
<td>InfoSci</td>
<td>Computational Biology</td>
<td>Computational Biology</td>
<td>Molecular Sequence Data</td>
<td>Molecular Sequence Data</td>
<td>Computational</td>
<td>Computational Biology</td>
</tr>
</tbody>
</table>
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 10944466 4

FIND THE TOPICAL EXPERTISE OF AUTHORS ON THIS PAPER

Most Novel Mesh Terms per category

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemicals</td>
<td>MicroRNAs (341)</td>
<td>MicroRNAs (292.775)</td>
<td>RNA, Messenger (0.0235526)</td>
<td>RNA, Messenger (-0.0015207)</td>
<td>MicroRNAs (4)</td>
<td>MicroRNAs (672.0)</td>
</tr>
</tbody>
</table>
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 10944436_4

Most Novel Mesh Terms per category

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemicals</td>
<td>MicroRNAs (341)</td>
<td>MicroRNAs (292.775)</td>
<td>RNA, Messenger (0.0236525)</td>
<td>RNA, Messenger (-0.0015207)</td>
<td>MicroRNAs (4)</td>
<td>MicroRNAs (672.0)</td>
</tr>
<tr>
<td>InfoSci</td>
<td>Computational Biology</td>
<td>Computational Biology</td>
<td>Molecular Sequence</td>
<td>Molecular Sequence Data</td>
<td>Computational</td>
<td>Computational Biology</td>
</tr>
</tbody>
</table>
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.