



Finding factors behind potential breakthrough papers

Jos Winnink, Robert Tijssen and Ton van Raan



Universiteit
Leiden

STI-2016, September 14-16, 2016

Jos Winnink

Research question

'What are factors behind potential breakthrough research?'

What is needed?

1. A large scale data source that can be harvested by computer algorithms to search for breakthrough-class discoveries.
2. A formalised analytical method to do the harvesting
3. Suitable bibliographic information
 - Web of Science (SCOPUS is also suitable)
 - Focus is on ‘articles’ and ‘letters’; contain original research

Method

We already developed a method for the early-stage identification of potential breakthrough discoveries

This search method focuses on individual publications; the analytical findings relate to aggregate publication sets

We call publications identified by our method **breakout** publications

Early stage = 24 – 36 months after publication

Algorithms

- Researchers inflow impact (RII)
- Application-oriented research impact (ARI)
- Cross-disciplinary impact (CDI)
- Discoverers intra-group impact (DII)
- Research niche impact (RNI)

Some validation results (papers published in 1990 -1994)

- Nature's *The top 100 papers** contains 13 publications published in 1990-1994
 - 2 publications are not included in this study, 'wrong' document type
 - 11 remaining publications were recognised
- For 5 of the 8 Nobel Prizes in Chemistry, Physics, and Physiology or Medicine for which scholarly work from 1990-1994 forms the scientific basis at least one of the seminal publications was identified

* van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers - nature explores the most-cited research of all time. *Nature*, 514(7524):550–553.

A more extensive validation?

In a forthcoming publication we validate the results of our method:

Winnink, J.J., Tijssen, R.J.W. and van Raan, A.F.J. *(to be published in PLoS One)* Can early-detection algorithms of breakout papers uncover scientific breakthroughs?

Our method succeeds to filter out what seem to be less relevant publications.

Distribution of all papers in the WoS 1990-1994 (articles & letters)

		Citations by patents until 3 rd Q 2013			
		< Top 10%	Top 10%	Top 5%	Top 1%
Citations by review papers until June 2016	Top 1%	0.4%	0.1%	0.1%	0.0%
	Top 5%	4.3%	0.2%	0.1%	0.0%
	Top 10%	15.3%	0.3%	0.1%	0.0%
	< Top 10%	78.6%	0.4%	0.1%	0.0%

0.0% means < 0.05%

Distribution of breakout papers

Articles & letters from 1990-1994

		Citations by patents until 3 rd Q 2013			
		< Top 10%	Top 10%	Top 5%	Top 1%
Citations by review papers until June 2016	Top 1%	17.5%	10.5%	8.3%	2.4%
	Top 5%	15.7%	6.7%	3.4%	0.6%
	Top 10%	10.0%	2.7%	1.2%	0.2%
	< Top 10%	16.5%	3.4%	1.0%	0.1%

Breakouts and all publications (articles & letters from 2007-2011)

Organisational category	Number of papers	Breakout papers	share
'All'	2,660,300	114,778	4.2%
University	1,886,048	79,144	4.2%
Research Institute	225,731	10,046	4.5%
Company	129,950	4,557	3.5%
Hospital	99,178	1,936	2.0%

Breakout and no-breakout papers (1990-1994)*

Organisational category**	All publications (letters & articles)	No breakout	Breakout
University	76.8%	76.8%	77.2%
Research Institute	11.3%	11.3%	13.1%
Company	6.4%	6.4%	6.4%
Hospital	5.5%	5.6%	3.4%

* No Arts, Humanities and Social Science papers

** Excludes papers that could not be assigned to organisational subcategories

Number of authors per paper

	<i>Breakout papers</i>	<i>All papers</i>
<i>Organisational category</i>	<i>Average number of authors per paper</i>	
University (U)	6.3	4.3
Research Institute (R)	5.8	4.2
Company (C)	8.6	4.9
Hospital (H)	6.2	4.4

Collaboration (articles & letters, 2007-2011)

Organisational category	No collaboration		Domestic collaboration		International collaboration	
	Breakout	No breakout	Breakout	No breakout	Breakout	No breakout
University (U)	34%	45%	32%	32%	34%	23%
Research institute (R)	23%	36%	32%	35%	45%	29%
Company (C)	15%	25%	37%	49%	49%	26%
Hospital (H)	11%	31%	44%	51%	45%	18%

Observations and conclusions

- Our method selects publications that have an above average citation impact on science & technology
- Breakout papers have more authors than ordinary papers
- 92% of the breakout papers show this characteristic in the first year after publication

We conclude that our method creates document sets that facilitate the study of the factors behind 'breakthroughs'

Thank you for your attention!



Distribution of *breakouts*: over-or underrepresentation relative to mean

	Medical and Life Sciences	Natural Sciences	Engineering Sciences	Social and Behavioural Sciences	Multidisciplin ary Journals
<i>Organisational category</i>					
University (U)	-	-	-	-	-
Research Institute (R)	-	-	+	-	-
Company (C)	-	-	-	-	+
Hospital (H)	-	-	-	-	-
U + R	+	+	+	+	+
U + H	+	+	-	+	+
U + C	+	+	-	+	+
U + H + R	+	+	-	+	+
U + H + C + R	+	+	-	+	+
U + H + C	+	+	-	+	+

Effect of organisational co-operation (articles + letters, 2007-2011)*

	Medical and Life Sciences	Natural Sciences	Engineering Sciences	Social and Behavioural Sciences	Multidisciplinary journals**
Total number of papers	2,362,512	2,066,802	508,093	360,443	75,047
Number of breakout papers detected	24,277	9,020	466	800	5,873
<i>Share of total</i>	<i>1.0%</i>	<i>0.4%</i>	<i>0.1%</i>	<i>0.2%</i>	<i>7.8%</i>
<i>Organisational category</i>					
University (U)	56.5%	69.8%	69.1%	83.6%	56.5%
Research institute (R)	4.4%	6.5%	9.2%	1.4%	3.3%
Company (C)	2.2%	1.6%	4.1%	0.3%	2.0%
Hospital (H)	1.5%	0.0%	0.0%	0.3%	0.4%
U + R	13.8%	15.7%	9.9%	7.9%	21.1%
U + H	8.0%	0.4%	0.2%	3.3%	4.1%
U + C	6.9%	5.7%	7.5%	2.5%	9.2%
U + H + R	3.6%	0.1%	0.0%	0.4%	2.2%
U + H + C + R	1.6%	0.0%	0.0%	0.1%	0.8%
U + H + C	1.5%	0.1%	0.0%	0.4%	0.5%

* Excludes papers that could not be assigned to organisational subcategories

** Journals assigned to this category, by Thomson Reuters, include *Nature*, *Science* and *PNAS*.