News Selection by News Aggregators and Incentives for Newspapers to Invest in Journalism: The Case of South Korea

Dong Ook Choi (Sangmyung Univ.)

Doh-Shin Jeon (Toulouse School of Economics)

Changmin Oh (Konan Tech.)

Introduction: Access to Online News

- South Korea is an exception in the world:
  - The proportion of people for whom **search and news aggregators** are the main gateway to online news: 66%
  - The proportion of people for whom **direct access to websites and apps** is the main gateway to online news: 6%
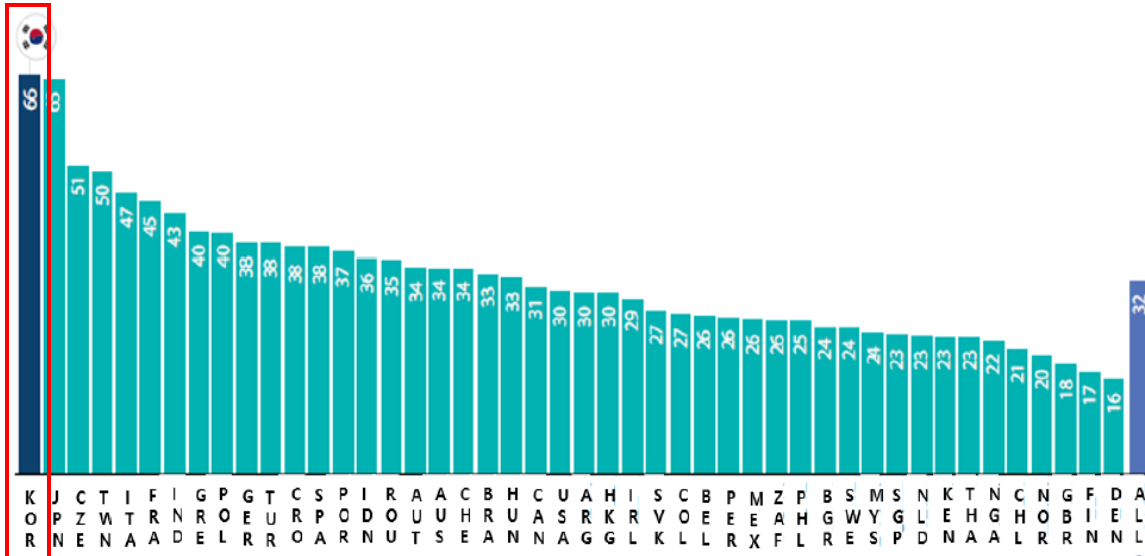
Search engine and news aggregators
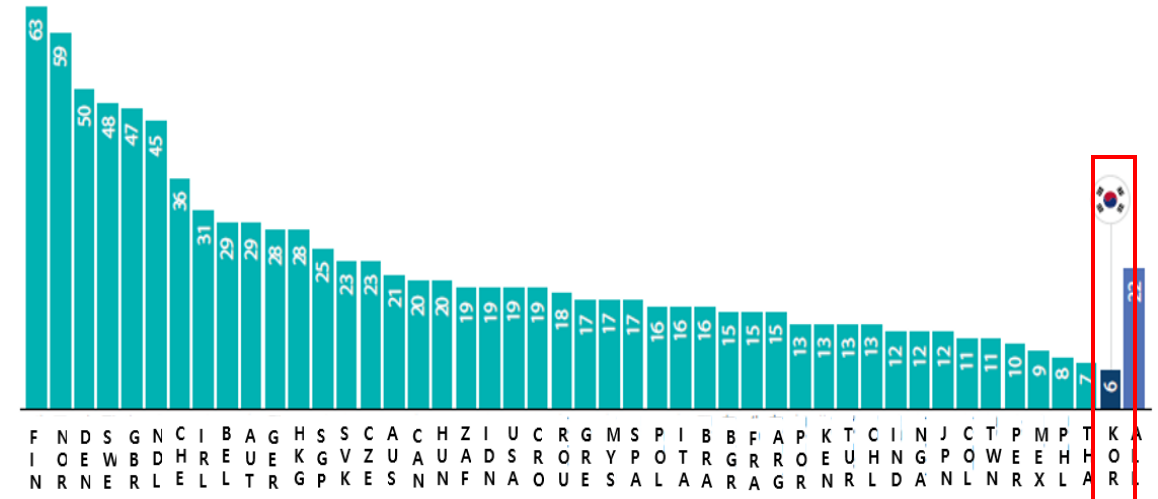
Direct access through websites and apps

Source: Korea Press Foundation (in collaboration with Reuters Institute) 2023

South Korea is an exception

- A large majority of South Koreans consume news through two news aggregators:
  - Naver.com (called Naver, hereafter)
  - Daum.net (called Daum hereafter)
- In 2015 (the year our data is collected), the share of the two aggregators, Naver and Daum, was 77.8%.
- The way the two aggregators select news has a significant impact on the production of news in South Korea.

News Aggregators in South Korea

News in Aggregators

- In-link service:
  - Selected news are shown in the aggregators' own pages.
- Short length of news text:
  - Usually less than 30 seconds reading.

Changes in Korean News Market

- Transition to online news from early in 2000s.
- Competition has shifted towards producing a larger volume of articles at a faster pace.
  - Quality issues: short length of articles and copying behaviors
- A shortage of journalists compared to the number of articles published.
  - The number of media professionals increased by 11% from 2000 to 2012, but the number of published articles surged by 498.6% (Song, 2020).
- In South Korea, despite the dominant role of the aggregators, these problems do not appear to be effectively addressed.

Objectives

- This paper aims to highlight two factors in the news selection process of news aggregators:
  - The presence of incomplete information in selecting issues
  - Being myopic in selecting news articles within an issue
- We also point out the problem of the aggregator's objective function being focused on maximizing clicks.
- Empirical questions:
  - Which issues are selected by the aggregators?
  - Which news articles are selected by the aggregators?
  - Do consumers prefer original news?

Online News Market (Cage et al., 2019)

- Studies online news production in France:
  - Collects all online news content from France in 2013 to measure the extent of copying.
  - Cluster news articles and measure originality of each article.
- However, the authors **do not consider the behavior of news aggregators** in the market.

- Our study employs similar methodology but considers the effect of news aggregators in the market.

Effect of News Aggregators

- Google news increases traffic to newspapers (Athey et al., 2017; Calzada and Gil, 2020; Chiou and Tucker, 2017)
  - Google News is a complement to overall news reading.
  - Newspapers have an incentive to opt-in.

- Theoretical Study: Jeon and Nasr (2016)
  - If readership expansion effect is positive, the aggregator increases quality chosen by each newspaper.

- No studies address **the news selection of aggregators** yet.

Contribution

- The first study on news selection by aggregators
- Novel dataset
  - All internet news content and selection info by aggregators.
  - Sequential order of articles, and their depreciation over time.
- Platform's behavior in online news market
  - Incomplete information about importance of issues, and selection of issues.
  - Ad-financed news aggregators that maximize their profits (clicks)
- Unique case from South Korea
  - Dominance of news aggregators in online news market

Data Source

- Study period: 2015
  - No algorithm for personalizing news provision
- Internet news articles from **BigKinds:**
  - The largest news article database in South Korea: all news articles from 52 newspapers for the year 2015, leading to a total of 4,495,009 articles
- News text from the **two aggregators**:
  - All articles posted on the headline pages of the two aggregators during 2015, leading to 147,000 articles
- Consumer data from **Neilsen:**
  - individual/day level data with approx. 1.5 mil. observations per month

Dispatching Articles

- All newspapers in the dataset have content partnership agreements with both aggregators.
- They are **required to share all their articles with the aggregators.**
- **In-link service**: the aggregators show selected news articles on their own pages.
- Newspapers dispatch their articles <u>at the same time to BigKinds and to the aggregators.</u>
  - We observe **the entry order** of articles in BigKinds.

Newspapers

Own Websites

News aggregators → Select articles* → Selected news by aggregators

Articles are dispatched at the same time

BigKinds: Public Archive* → all internet news articles with entry order
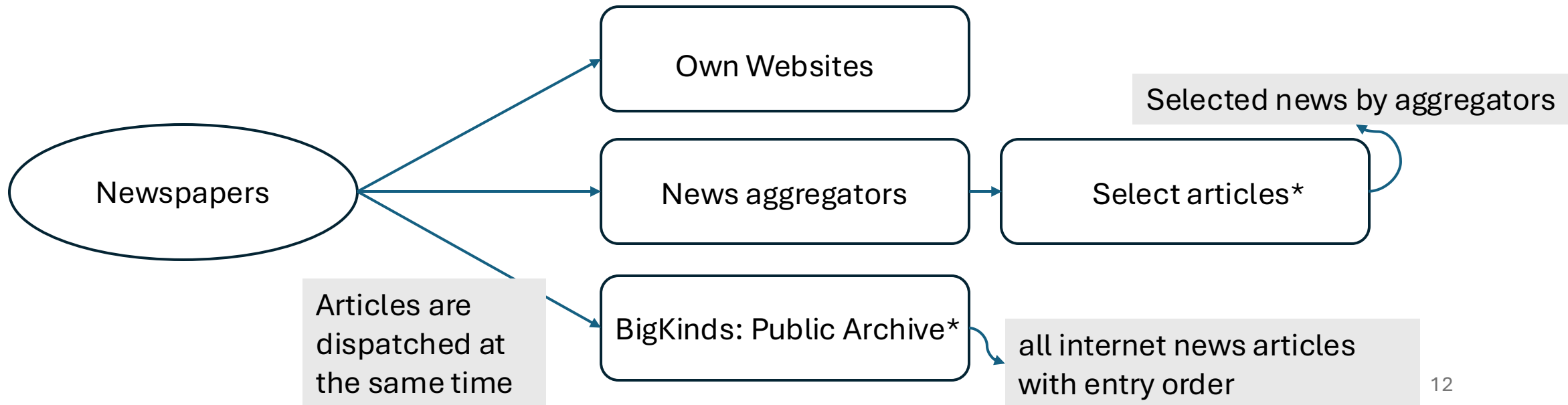
Methods

- **Clustering**:
  - Building on Cagé, Hervé and Viaud (RES, 2019)
  - Group news articles by topic through a clustering process (DBSCAN algorithm)
  - Compute originality rate of each article within each cluster.
- **Identifying news selection**:
  - the similarity between a news article in BigKinds and an article in the aggregators is calculated
  - Due to some modifications in the aggregators' news text (headers, footers, and copyright notes), articles are considered identical if their Jaccard similarity score is above a certain level.

Description: Clusters

- The number of clusters in the analysis is 45,260.

- The average cluster has 20.7 articles, with duration of 2.28 days. It has 0.3 selected articles with 9.43 newspapers.

- Among clusters in our study, 15% had at least one article selected, and only 2% have the first article in the cluster selected.

Variable	N	Mean	Median	Min	Max
Cluster size	45260	20.71	15.00	10.00	2015.00
Duration (days)	45260	2.28	2.00	1.00	15.00
N of sel	45260	0.30	0.00	0.00	64.00
Select Dummy	45260	0.15	0.00	0.00	1.00
1st rank selected	45260	0.02	0.00	0.00	1.00
N of newspapers	45260	9.43	9.00	2.00	50.00

Description: Articles in the Clusters

- The total number of observations is 937,549, of which 13,573 articles were selected by the aggregators.

- Articles selected by the aggregators exhibit higher originality and longer text lengths.
  - The average originality of selected articles is 36.6%, while that of non-selected articles is approximately 28%.

	selected by aggregators		
	0	1	Total
N	923,976 (98.6%)	13,573 (1.4%)	937,549 (100.0%)
Originality	27.968 (27.605)	36.600 (29.437)	28.093 (27.652)
length	788.992 (478.252)	1,107.166 (563.088)	793.598 (481.091)
exclusive news dummy	0.001 (0.034)	0.056 (0.231)	0.002 (0.044)
breaking news dummy	0.002 (0.047)	0.005 (0.069)	0.002 (0.048)
byline dummy	0.727 (0.446)	0.908 (0.289)	0.729 (0.444)
soft news dummy	0.225 (0.418)	0.014 (0.117)	0.222 (0.416)
hard news dummy	0.730 (0.444)	0.953 (0.212)	0.733 (0.442)

Description: Selected Articles

- On average, selected articles are the 24.9th article within a cluster and are positioned at the 46.8th percentile.

- The rank of the first selected article is around 10.8, implying that aggregators are not likely to select the first article to break an issue.

Variable	N	Mean	Median	Min	Max
All selected articles					
Rank	13,573	24.88	10	1	1,949
Percentile	13,573	46.84	45.45	0.43	99.7
First selected article dummy	13,573	0.49	0	0	1
Originality	13,573	36.6	35.32	0	100
First-selected articles in clusters					
Rank	6,694	10.82	7	1	338
Originality	6,694	54.21	54.11	0	100

Regression Results: Cluster Level

- Y: selection dummy
- X: cluster characteristics

- The probability for a cluster to be selected increases with
  - The size of the cluster
  - The number of newspapers which contribute articles to the cluster
  - The average originality of articles in the cluster
  - The average length of articles in the cluster

	OLS_1	OLS_2	Probit_1	Probit_2
main				
Cluster size	0.0003**	0.0013**	-0.0011**	0.0125***
	(0.000)	(0.000)	(0.000)	(0.001)
Duration (days)	0.0186***	0.0072**	0.0775***	0.0094
	(0.002)	(0.002)	(0.007)	(0.008)
N of newspapers	0.0179***	0.0099***	0.0757***	0.0331***
	(0.000)	(0.001)	(0.002)	(0.003)
Avg orig		0.0019***		0.0165***
		(0.000)		(0.001)
Avg length		0.0001***		0.0004***
		(0.000)		(0.000)
Byline ratio		0.0022		0.1878***
		(0.006)		(0.045)
Excl. dummy		0.3633***		1.1763***
		(0.017)		(0.050)
Break dummy		0.1359***		0.3611***
		(0.015)		(0.046)
Soft news dummy		-0.1171***		-1.1410***
		(0.004)		(0.035)
Constant	-0.0697***	-0.0989***	-1.9862***	-2.5596***
	(0.004)	(0.009)	(0.020)	(0.053)
R-squared	0.082	0.143		
N	45,260	45,260	45,260	45,260
ll	-15404.584	-13845.369	-17351.218	-15442.620

* p<0.05, ** p<0.01, *** p<0.001

Regression Results: Article Level

- Probit model
  - Y: selection dummy
  - X: article characteristics

- No significant relationship between the selection and rank
  - No premium for being the first to report on an issue

- Once an issue is selected, the aggregator tends to choose higher quality articles.
  - Originality, article length, and byline dummy are positively correlated.

	All articles		Articles in selected clusters	
	Model_1	Model_2	Model_3	Model_4
Selected				
Rank	0.0000	-0.0001	-0.0022**	-0.0020**
	(0.000)	(0.000)	(0.001)	(0.001)
Originality	0.0031***	0.0014***	0.0038***	0.0024***
	(0.000)	(0.000)	(0.000)	(0.000)
length	0.0003***	0.0002***	0.0002***	0.0002***
	(0.000)	(0.000)	(0.000)	(0.000)
byline dummy	0.4466***	0.3457***	0.4280***	0.3428***
	(0.017)	(0.020)	(0.022)	(0.024)
exclusive news dummy		1.6643***		1.7024***
		(0.055)		(0.058)
breaking news dummy		0.1729		-0.0027
		(0.092)		(0.103)
soft news dummy		-0.8809***		0.2518*
		(0.072)		(0.123)
Constant	-2.8881***	-2.7532***	-2.0886***	-2.0961***
	(0.026)	(0.037)	(0.043)	(0.052)
Log-Likelihood	-67642.152	-58038.345	-47704.997	-42346.118
N	937,549	812,280	204,070	175,404

* p<0.05, ** p<0.01, *** p<0.001

Marginal effects of ranking on selection probability

- Even within the selected clusters, the effect of ranking on selection prob. is very low.

- When the rank changes from 1 to 20, the percentage reduction in selection probability is 4.7.

- This implies that almost **no premium for breaking the first story** and a **strong incentive to copy.**

<Table 7> Marginal effects estimation results

Rank	dy/dx	Standard Error	P>z
1	-0.00029	0.000106	0.007
20	-0.00027	9.66E-05	0.005
Average rank	-0.00026	8.83E-05	0.003
Observations		175,404	

Note: Marginal effects are estimated using the parameter estimate of the rank variable in column (4) in Table 6.

Consumer preferences

	(1) PV	(2) Duration	(3) PV	(4) Duration
Originality	-0.027*	0.207***	-0.016	0.220***
	(0.016)	(0.038)	(0.017)	(0.043)
Length	0.009*	-0.009	0.009*	-0.006
	(0.005)	(0.012)	(0.005)	(0.013)
Length^2	-0.000*	0.000	-0.000*	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Byline	-0.007	0.019	-0.006	0.023
	(0.012)	(0.032)	(0.012)	(0.032)
N. of Excl./Break	0.055**	0.028	0.054**	0.020
	(0.017)	(0.043)	(0.017)	(0.044)
N. of articles	0.003	-0.003	0.002	-0.008
	(0.010)	(0.028)	(0.010)	(0.028)
N. of newspapers	0.004	0.002	0.004	0.001
	(0.016)	(0.041)	(0.016)	(0.041)
N. of news agencies	-0.008*	0.019*	-0.008*	0.019*
	(0.004)	(0.011)	(0.004)	(0.011)
N. of clusters	0.037**	-0.223***	0.041**	-0.217***
	(0.016)	(0.047)	(0.016)	(0.048)
Diff. political stance			-7.145**	-8.375
			(3.524)	(8.635)
Avg. cluster size			0.033	0.074
			(0.025)	(0.069)
Max cluster size			0.022	-0.156
			(0.055)	(0.147)
Observations	380,450	380,450	380,450	380,450
r2	0.003	0.003	0.003	0.003

- Panel FE model results
  - Y: PV or Duration per day of users
  - X: daily average of aggregator's characteristics

- In column (1), a lower originality rate is significantly correlated with the increase in the number of clicks but is not significant in column (3).

- In columns (2) and (4), as the originality increases, duration also increases.

- Duration-based metrics may generate stronger incentives to invest in news quality.

Summary

- On average, the rank of the first selected article within each selected cluster is 10.8th article
- Cluster-level analysis:
  - Higher selection probability for issues with larger sizes (n. of articles) and with higher newspaper participation.
- Article-level analysis:
  - Selection probability has no relationship with article rank, implying that there is no premium for being the first rank in the issue.
  - Low marginal effect of rank implies the strong incentive to copy others' news.
- Consumer preferences:
  - Consumers take time to read and pay more attention to original news articles.

Concluding Remark: Why no premium for breaking a story?

- Incomplete information and belief updating:
  - The aggregators do not know which issues are important.
  - They update their belief about the importance of a issue as additional articles arrive.

- Being myopic:
  - After a good number of articles accumulate on a topic, aggregators recognize its importance but tend to prioritize selecting the most recent articles over the initial ones.

- Furthermore, if the aggregators maximize clicks, they have little incentives to encourage originality.