What Makes Online Content Viral?

JONAH BERGER and KATHERINE L. MILKMAN

WEB APPENDIX

Data Used

The *New York Times* does not store the content of Associated Press, Reuters, and Bloomberg articles, as well as blogs, and so it was not available for our analyses. We also did not include videos and images with no text.

Modeling Approach

We used a logistic regression model because of the nature of our question and the available data. While more complex panel-type models are appropriate when there is time variation in at least one independent variable and the outcome, we do not have period-by-period variation in the dependent variable. Rather than having the number of e-mails sent in each period, we only have a dummy variable that switches from 0 (not on the most e-mailed list) to 1 (on the most e-mailed list) at some point due to events that happened not primarily in the same period but several periods earlier (e.g., advertising in previous periods). Furthermore, our interest is not in when an article makes the list but whether it ever does so. Finally, although it could be imagined that when an article is featured might affect when it makes the list, such an analysis is far from straightforward. The effects are likely to be delayed (where an article is displayed in a given time period is extremely unlikely to have any effect on whether the article makes the most e-mailed list during that period), but it is difficult to predict a priori what the lag between being featured prominently and making the list would be. Thus, the only way to run an appropriate panel model would be to include the full lag structure on all our time-varying variables (times spent in various positions on the home page). Because we have no priors on the appropriate lag structure, the full lag structure would be the only appropriate solution. So imagine, for example, that there are two slots on the home page (we actually have seven): Position A and Position B. Our model would then need to be something like the following:

Being on the list in period t =

$$\beta_1 \times (\text{being in Position A in period t})$$
$$+ \beta_2 \times (\text{being in Position A in period t} - 1)$$
$$+ \beta_3 \times (\text{being in Position A in period t} - 2) + \ldots$$
$$+ \beta_N \times (\text{being in Position A in period t} - N)$$
$$+ \beta_{N+1} \times (\text{being in Position B in period t})$$
$$+ \beta_{N+2} \times (\text{being in Position B in period t} - 1)$$
$$+ \beta_{N+3} \times (\text{being in Position B in period t} - 2) + \ldots$$
$$+ \beta_{2N} \times (\text{being in Position B in period t} - N)$$
$$+ \beta(\text{a vector of our other time-invariant predictors}).$$

If we estimated this model, we would end up with an equivalent model to our current logistic regression specification in which we have summed all of the different periods for each position. The two are equivalent models unless we include interactions on the lag terms, and it is unclear what interactions it would make sense to include. In addition, there are considerable losses in efficiency from this panel specification compared with our current model. Thus, we rely on a simple logistic regression model to analyze our data set.

Coding Instructions

Anger. Articles vary in how angry they make most readers feel. Certain articles might make people really angry, while others do not make them angry at all. Here is a definition of anger: http://en.wikipedia.org/wiki/Anger. Please code the articles based on how much anger they evoke.

Anxiety. Articles vary in how much anxiety they would evoke in most readers. Certain articles might make people really anxious while others do not make them anxious at all. Here is a definition of anxiety: http://en.wikipedia.org/wiki/Anxiety. Please code the articles based on how much anxiety they evoke.

Awe. Articles vary in how much they inspire awe. Awe is the emotion of self-transcendence, a feeling of admiration and elevation in the face of something greater than the self. It involves the opening or broadening of the mind and an experience of wow that makes you stop and think. Seeing the Grand Canyon, standing in front of a beautiful piece of art, hearing a grand theory, or listening to a beautiful symphony may all inspire awe. So may the revelation of something profound and important in something you may have once seen as ordinary or routine or seeing a causal connection between important things and seemingly remote causes.

Sadness. Articles vary in how much sadness they evoke. Certain articles might make people really sad while others do not make them sad at all. Here is a definition of sadness: http://en.wikipedia.org/wiki/Sadness. Please code the articles based on how much sadness they evoke.

Surprise. Articles vary in how much surprise they evoke. Certain articles might make people really surprised while others do not make them surprised at all. Here is a definition of surprise: http://en.wikipedia.org/wiki/Surprise_(emotion). Please code the articles based on how much surprise they evoke.

Practical utility. Articles vary in how much practical utility they have. Some contain useful information that leads the reader to modify their behavior. For example, reading an article suggesting certain vegetables are good for you might cause a reader to eat more of those vegetables. Similarly, an article talking about a new Personal Digital Assistant may influence what the reader buys. Please code the articles based on how much practical utility they provide.

Interest. Articles vary in how much interest they evoke. Certain articles are really interesting while others are not interesting at all. Please code the articles based on how much interest they evoke.

Additional Robustness Checks

The results are robust to (1) adding squared and/or cubed terms quantifying how long an article spent in each of seven home page regions; (2) adding dummies indicating whether an article ever appeared in a given home page region; (3) splitting the home page region control variables into time spent in each region during the day (6 A.M.–6 P.M. eastern standard time) and night (6 P.M.–6 A.M. eastern standard time); (4) controlling for the day of the week when an article was published in the physical newspaper (instead of online); (5) Winsorizing the top and bottom 1% of outliers for each control variable in our regression; (6) controlling for the first home page region in which an article was featured on the *New York Times*' site; (7) replacing day fixed effects with controls for the average rating of practical utility, awe, anger, anxiety, sadness, surprise, positivity and emotionality in the day's published news stories; and (8) including interaction terms for each our primary predictor variables with dummies for each of the 20 topic areas classified by the *New York Times*.

Alternate Dependent Measures

Making the 24-hour most e-mailed list is a binary variable (an article either makes it or it does not), and while we do not have access to the actual number of times articles are e-mailed, we know the highest rank an article achieves on the most e-mailed list. Drawing strong conclusions from an analysis of this outcome measure is problematic, however, for several reasons. First, when an article earns a position on the most e-mailed list, it receives considerably more "advertising" than other stories. Some people look to the most e-mailed list every day to determine what articles to read. It is unclear, however, exactly how to properly control for this issue. For example, the top ten most e-mailed stories over 24 hours are featured prominently on the *New York Times*' home page, but readers must then click on a link to see the rest of the most e-mailed list (articles 11–25). This suggests that it may be inappropriate to assume that the same model predicts performance from rank 11–25 as rank 1–10. Second, any model assuming equal spacing between ranked categories is problematic because the difference in virality between stories ranked 22 and 23 may be very small compared with the difference in virality between stories ranked 4 and 5, thus reducing the ease of interpretation of any results involving rank as an outcome variable. That said, using an ordered logit model and coding articles that never make the most e-mailed list as earning a rank of 26 (leaving these articles out of the analysis introduces additional selection problems), we find nearly identical results to our primary analyses presented in Table 5 (Table A3).

Another question is persistence, or how long articles continue to be shared. This is an interesting issue, but unfortunately it cannot be easily addressed with our data. We do not have information about when articles were shared over time, only how long they spent on the most e-mailed list. Analyzing time spent on the most e-mailed list shows that both more affect-laden and more interesting content spends longer on the list (Table A3). However, this alternative outcome variable also has several problems. First, there is a selection problem: Only articles that make the most e-mailed list have an opportunity to spend time on the list. This both restricts the number of articles available for analysis and ensures that all articles studied contain highly viral content. Second, as we discussed previously, articles that make the most e-mailed list receive different amounts of additional "advertising" on the *New York Times* home page, depending on what rank they achieve (top ten articles are displayed prominently). Consequently, although it is difficult to infer too much from these ancillary results, they highlight an opportunity for further research.

Table WA1
HOME PAGE LOCATION ARTICLE SUMMARY STATISTICS

	% of Articles That Ever Occupy This Location	*For Articles That Ever Occupy Location*		
		% That Make List	*Mean Hours*	*Hours Standard Deviation*
Top feature	28%	33%	2.61	2.94
Near top feature	32%	31%	5.05	5.11
Right column	22%	31%	3.85	5.11
Middle feature bar	25%	32%	11.65	11.63
Bulleted subfeature	29%	26%	3.14	3.91
More news	31%	24%	3.69	4.18
Bottom list	88%	20%	23.31	28.40

Notes: The average article in our data set appeared somewhere on the *New York Times*' home page for a total of 29 hours (SD = 30 hours).

Table WA2
PHYSICAL NEWSPAPER ARTICLE LOCATION SUMMARY STATISTICS

	% of Articles That Ever Occupy This Location	*For Articles That Ever Occupy Location*		
		% That Make List	*Mean Page Hours*	*Mean Page Number for Articles That Make List*
Section A	39%	25%	15.84	10.64
Section B	15%	10%	6.59	5.76
Section C	10%	16%	4.12	5.38
Section D	7%	17%	3.05	2.27
Section E	4%	22%	4.78	7.62
Section F	2%	42%	3.28	3.43
Other section	13%	24%	9.59	14.87
Never in paper	10%	11%	—	—

Table WA3

AN ARTICLE'S HIGHEST RANK AND LONGEVITY ON THE *NEW YORK TIMES'* MOST E-MAILED LIST AS A FUNCTION OF ITS CONTENT CHARACTERISTICS

Outcome Variable	*Highest Rank (7)*	*Hours on List (8)*
Emotion Predictors		
Emotionality	.22***	2.25**
	(.04)	(.85)
Positivity	.15***	.72
	(.04)	(.81)
Specific Emotions		
Awe	.25***	–1.47
	(.05)	(1.11)
Anger	.35***	.35
	(.08)	(1.14)
Anxiety	.19**	.36
	(.06)	(.95)
Sadness	–.16**	–.77
	(.06)	(.93)
Content Controls		
Practical utility	.31***	.38
	(.05)	(1.07)
Interest	.27***	1.85†
	(.06)	(1.00)
Surprise	.17***	1.04
	(.05)	(.85)
Homepage Location Control Variables		
Top feature	.11***	–.18
	(.02)	(.18)
Near top feature	.11***	.21†
	(.01)	(.13)
Right column	.15***	.88***
	(.01)	(.17)
Middle feature bar	.05***	–.01
	(.00)	(.06)
Bulleted subfeature	.03*	–.21
	(.01)	(.22)
More news	.01	.32
	(.01)	(.24)
Bottom list × 10	.04*	.07
	(.02)	(.22)
Other Control Variables		
Word count × 10^{-3}	.37***	4.67*
	(.08)	(1.99)
Complexity	.01	–1.10
	(.03)	(.95)
First author fame	.21***	1.89***
	(.02)	(.55)
Female first author	.37***	4.07**
	(.07)	(1.35)
Uncredited	.74***	13.29†
	(.26)	(7.53)
Newspaper location and web timing controls	Yes	Yes
Article section dummies (e.g., arts, books)	No	No
Observations	6956	1391
Regression modeling approach	Ordered Logit	Ordinary Least Squares
Pseudo-R^2/R^2	.13	.23
Log-pseudo-likelihood	–6929.97	N.A.

†Significant at the 10% level.
*Significant at 5% level.
**Significant at 1% level.
***Significant at the .1% level.

Notes: The regressions models examine the content characteristics of an article associated with its highest rank achieved on the *New York Times*' most e-mailed list (reverse-scored such that 25 = the top of the list and 0 = never on the list) and its longevity on the list. Both models rely on our primary specification (see Table 5, Model 4) and include day fixed effects. N.A. = not applicable.