Supplementary Material for: “Popularity Spikes Hurt Future Chances for Viral Propagation of Protomemes”

Michele Coscia*

*CID - Harvard University
79 JFK St, Cambridge MA, US 02138*

November 7, 2017

Contents

1	Historic Trends	1
2	Protomeme Examples	3
3	Front Page Uncertainty	5
4	MAX Model	6
5	STD Model	7
6	Canonicity	7
7	ϕ, One and Zero Model Coefficient Interpretation	10
8	One and Zero Models: Negative Binomial	11

1 Historic Trends

In the main body of the article we claimed that the total score awarded is not constant over time, the average score per post is not constant over time, and

*To whom correspondence should be addressed. E-mail: michele.coscia@hks.harvard.edu, Work Telephone number: +1 (857) 241 7563, Work Fax number: +1 (617) 496 8753 (include a header with the name “Michele Coscia”).

Figure 1: Weekly evolution of Reddit activity.

the overall number of daily posts in our data grows over time. We present here supporting evidence for the claim.

We start by reporting Reddit’s statistics. Figure 1 reports the evolution in weekly number of posts (Figure 1(a)) and total score awarded (Figure 1(b)) during the observation period. Both types of activity roughly doubled during the period of observation. In total we analyzed 22,329,506 posts with a total aggregate score of 860,956,257. Figure 1(c) reports the evolution of the average popularity of a Reddit post per week, proving that, over time, Reddit users granted more upvotes on average to their posts.

In our dataset, we found a total of 2,652,609 usernames. However, to preserve anonymity, many users create alternative accounts only to join a particular conversation (these accounts are called “throwaway”). This way, the number of active users is inflated. A more accurate estimate is to just calculate the number of users which submitted at least a post per week, which is depicted in Figure 1(d). The figure shows the user base going from around 170k users to almost 300k during the period of observation. The average number of posts and average score gathered per user per week are depicted in Figures 1(e) and 1(f) respectively, where we show that user activity is roughly constant over time.

Moving to Hacker News, Figure 2 the weekly number of posts (Figure 2(a)) and total score awarded (Figure 2(b)) are significantly lower than Reddit, given the smaller audience of the website. Just like Reddit, also Hacker News roughly doubled during the period of observation (although it has to be noted that the observation period for Hacker News is longer). The Hacker News dataset includes 1,194,436 posts with a total aggregate score of 11,870,900. Figure 2(c) reports similar trends in Hacker News when averaging the score per post, with

Figure 2: Weekly evolution of Hacker News activity.

a 33% increase.

In our dataset, we found a total of 154,082 usernames. Just like in Reddit, there are duplicate and inactive accounts. When calculating the number of users which submitted at least a post per week, we obtain the timeline depicted in Figure 2(d). The user base grows from 1,500 weekly active users to around 3,500. Differently from Reddit, the average number of posts (Figure 2(e)) is constantly declining. The average score gathered per user (Figure 2(f)) is roughly constant.

2 Protomeme Examples

We report in this section a handful of protomeme examples. We focus on popular protomemes, i.e. memes that have been widely used. Among these protomemes, we randomly pick three from Reddit and three from Hacker News. The three Reddit protomemes are:

- “world problems”: this is a template protomeme that has been used especially in its “first world problems” variant¹, but also “second world problems”² and “nerd world problems”³ have had some traction, along many more.
- “bad luck brian”: very popular meme which went viral immediately before our observation period (in January 2012⁴) and stayed popular throughout

¹<http://knowyourmeme.com/memes/first-world-problems>

²<https://www.reddit.com/r/SecondWorldProblems/>

³<http://knowyourmeme.com/memes/nerd-world-problems>

⁴<http://knowyourmeme.com/memes/bad-luck-brian>

Figure 3: Timelines of the protomeme examples.

the year, making it one of the most prominent memes in our dataset.

- “hurricane sandy”⁵: an example of how news items can take a life of their own on social media.

The three Hacker News protomemes are:

- “node js”: a popular runtime environment for developing applications⁶.
- “big data”: a hot topic in scientific and technological environments for its computational challenges and popular results.
- “leasson learned”: a catchphrase that is often used in posts that typically do a postmortem analysis of some product/project, which shows the startup cultural focus of extracting useful knowledge from failures.

We can see that what a protomeme is in different social media platforms can be rather different. While Reddit focuses on pop Internet culture, Hacker News is dominated by technology trends – which can be memes themselves, as the rise and fall of programming languages and environments follows dynamics that have emerging properties not dissimilar from memes[1].

Figure 3 depicts the timelines of number of posts and upvotes gathered by these protomemes in our observation period. There are several interesting observations we can gather from them. The first is a comparison between the “world problems” protomeme and “bad luck brian”. Both protomemes were

⁵<http://knowyourmeme.com/memes/events/2012-hurricane-sandy>

⁶<https://en.wikipedia.org/wiki/Node.js>

Figure 4: Distributions of the scores of top rated posts, averaged per rank.

very popular, but at different times: “world problems” peaked at the end of 2011, “bad luck brian” in middle 2012. So for the former we only observe the tail of the distribution, while for the latter we also see the peak. It is interesting how “bad luck brian” eventually settles on similar levels of posts and votes of “world problems”.

“hurricane sandy” is an interesting case on itself. It is obviously very dependent on the news cycle, peaking when the storm hit the east coast of the US. However, the protomeme shows a longer tail than the simple news cycle: it gathered 5k-6k votes twenty weeks after the event, when the quick lifecycle of the news already moved on through several different topics (Obama’s re-election, pope resignation, Boston marathon bombings). This relationship between memes and the news cycle has been explored in depth [2].

In Hacker News, there seems to be a difference in how protomemes come to prominence. Node JS was initially released in 2009 – before our observation window – and it grew in popularity in 2010 and 2011 (with the involvement of Microsoft). This is reflected by the increased number of posts about it, which levels off around week 2175 (around September 2011). Similar dynamics can be observed for the big data protomeme, which goes from barely any mention to around 25 posts per week, although in this case it is harder to pinpoint this interest to specific events. The “lesson learned” protomeme is likely to be in the tail of its lifecycle, an endemic protomeme which is not used very often, but which is also unlikely to disappear completely.

3 Front Page Uncertainty

In our main regression, we use $FP_{m,i}^l$ as a binary function, telling us if meme m was on the front page in day i . We have to introduce the parameter l , which represents the number of posts hitting front page in a day. As reported in the paper, the default front page in Reddit contains 25 posts, 30 posts for Hacker News. Therefore, we are sure that every day at least 25 (30) posts hit the front page. However, users can increase the length of their front page. Moreover, as

Figure 5: Evolution of β_1 coefficients for increasing l in MAX Model.

time passes, more posts will float on the top in a given day and other posts sink. There is no way to know how many posts hit any particular user’s front page in a given day. To show this, in Figure 4 we report the average score of the top 500 ranked posts for Reddit, and the top 100 ranked post for Hacker News – the difference in range due to the much smaller average number of posts per day submitted to the latter platform. No discontinuities in scores can be found in either case, proving that without actually recording the front page every day there is no *a posteriori* trace of a post being in the front page. For this reason, we run the regression multiple times for different l values. In this way, we verify that our results are robust, regardless that in a day there were 25 or 50 posts in the front page. The interval we set for l is $25 \leq l \leq 100$.

Note that this front page dynamism does not affect significantly our results. We still expect to find stronger results for posts in the top 25 than in the post between 25th and 50th rank. This is because, the higher the post’s rank, the more stable it is. This ranking stability has been studied in several contexts [3, 4], and it ensures the robustness of our results even in presence of front page uncertainty for the medium-high ranked posts.

4 MAX Model

In the main paper, our model (MED Model) uses as target value the median score of posts. Here, we regress on $\max_{w_{m,i}}(p_{m,i})$ instead of $E_{w_{m,i}}(p_{m,i})$. In practice, we investigate the effect of hitting the front page on day $i - 1$ for the best scoring post including the protomeme on day i . For this reason, this model is referred to as MAX Model. The values of β_1 for the MAX Model are depicted in Figures 5 (a) and (b), for Reddit and Hacker News respectively.

In Reddit (Figure 5 (a)) the sign of β_1 is the opposite of the β_1 sign for the MED Model. The interpretation is that for a protomeme m hitting the front page on day $i - 1$ improves the score of the best post containing protomeme m on day i . The best post’s score receives a bonus of +4%. Interestingly, in Hacker News (Figure 5 (b)) this is not true: β_1 still carries a negative sign. This

Figure 6: Distribution of the STD Model’s β for varying l . Thin lines represent the 95% confidence intervals.

can be interpreted as a resilience of Hacker News users to fads: the fact that a protomeme was popular the day before does not mean that best scoring post containing it the day after will gather more votes. It is likely that the explanation lies in Hacker News having a specific purpose for tech-oriented people, which Reddit lacks, being more focused on humor and leisure time activities.

5 STD Model

Given the increase of number of posts containing a protomeme after it hits the front page, we also would expect not only changes in a protomeme’s average and maximum score, but also in the dispersion of the scores. Hitting the frontpage makes a protomeme more popular not only among its consumers, but also among its producers. To test for this hypothesis, we fit a model similar to the MAX model, but using the standard deviation of the scores as the target variable. We also control for the standard deviation of the scores the previous day. For this reason, we call this the STD Model.

Figure 6 reports the results, both for Reddit and Hacker News. The confidence intervals of our estimates are too wide for us to make any sort of definitive claim. It seems that the hypothesis proposed could be true, as the coefficients are positive for many l values, but the p-values are too low for any confident conclusion. We leave this particular question open to future work.

6 Canonicity

In the paper we introduce the Canonicity measure to estimate post-post similarity, as the meme-meme similarity measure defined in [5] is not applicable here. In Figure 7 we provide an illustrated example describing the behavior of the Γ function that calculates the canonicity of a post containing a protomeme. We consider a protomeme composed by words μ_1 and μ_2 , which co-appear with

	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	$\Gamma(\mathcal{M}, m)$
$\pi_{m,\mu}$	1	1	0.75	0.66	0.5	0.25	0.1	0.05	0.02	0.01	
$\mathcal{M}_1$											1
$\mathcal{M}_2$											0.782
$\mathcal{M}_3$											0.592
$\mathcal{M}_4$											0.416
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	

Figure 7: Visual illustration of the computation of $\Gamma(\mathcal{M}, m)$.

other 8 words $\mu_3 \dots \mu_{10}$. The corresponding $\pi_{m,\mu}$ values are reported in the second row of the figure. For instance μ_4 appears in 66% of the posts containing the protomeme. The colored boxes represents four possible posts ($\mathcal{M}_1 \dots \mathcal{M}_4$) at decreasing levels of canonicity: maximum ($\mathcal{M}_1$), high ($\mathcal{M}_2$), medium ($\mathcal{M}_3$) and low ($\mathcal{M}_4$). The last column reports the $\Gamma(\mathcal{M}, m)$ scores.

From Figure 7 we can see that, to score a low $\Gamma(\mathcal{M}, m)$, it is necessary to satisfy two conditions: the words included in the post have to have low $\pi_{m,\mu}$ and there have to be many of them. If a post includes only an unusual term, its canonicity is still high, because it is still mostly composed by the protomeme itself.

We integrate this abstract example with a concrete one. We choose “long lost” as a protomeme. This is a sentence usually used in humorous context beyond its literal meaning, to indicate serendipitous discoveries rather than actual serious reunions. We focus on three posts using this catchphrase with high, medium and low canonicity.

The highest canonicity for this protomeme is usually associated with the additional word “friend”⁷. This word has been used in 9.28% of all posts containing the protomeme. As a consequence, the canonicity value is $(2 + 0.0928) / 3 = 0.6976$. A post with medium canonicity⁸ goes with the very common associations with “found” and “brother” (used in 12.22% and 9.15% of its posts, respectively), plus the rather uncommon word “ent” (used only .42% of the times with “long lost”). “Ent” is a particular slang popular in a subreddit, making this usage only mediumly uncommon. The canonicity is then $(2 + 0.1222 + 0.0042 + 0.0915) / 5 = 0.4436$. Finally, low canonicity can be achieved when associating this catchphrase with words not usually used in its context⁹. Here the canonicity drops to about .3.

⁷E.g. <https://www.reddit.com/14ce5r>

⁸<https://www.reddit.com/14vpda>

⁹As in <https://www.reddit.com/1222nt>

Figure 8: Post title length distribution.

Figure 9: Distribution of meanings and canonicities per post.

We now describe how canonicity is distributed in our dataset. First, we start by showing the distribution of post lengths, since the post length is in the denominator of the canonicity formula. Figure 8 depicts the distributions for Reddit and Hacker News. We can see that most posts in both social media have a short title. In both cases, 90% of post titles contain ten words or less. Since protomemes usually contain two or three words in themselves, this means that the rest of the title usually contributes up to seven-eight additional words to the denominator, typically – in 30% of the cases – just one. Note that here we count only the stopworded and stemmed terms: posts might be much longer than this by having a lot of stopworded elements like “a”, “the”, “is” and so on.

We now scan Reddit’s timeline. For each post, sorted by timestamp, we first calculate the post’s $\Gamma(\mathcal{M}, m)$ value. We then update $\pi_{m, \mu}$, for each meaning $\mu \in \mathcal{M}$. After that, we move to the next post. In practice, for each post we first calculate its canonicity and then we update the canon of the protomeme m contained in $\mathcal{M}$.

To visualize the results, we bin Reddit’s history into weeks, as we did for Figure 1. We also bin the $\Gamma(\mathcal{M}, m)$ values. This results in a count of how many

Figure 10: Evolution of post canonicity during the observation period.

posts were posted in each week and in each $\Gamma(\mathcal{M}, m)$ value interval. To choose the most appropriate bins for canonicity, we recall that $\Gamma(\mathcal{M}, m)$ has the length of the posts in its denominator. Most posts have four or less words as shown in Figure 9(a). As a result, $\Gamma(\mathcal{M}, m)$ distributes with several discontinuity points around simple fractions like $2/3$ or $1/4$, as shown in Figure 9(b). For these reasons, we decided to bin $\Gamma(\mathcal{M}, m)$ using simple fractions. The list of bins is then: $]0, 1/4]$; $]1/4, 1/3]$; $]1/3, 1/2]$; $]1/2, 2/3]$; $]2/3, 3/4]$; $]3/4, 1]$.

The result is depicted in Figure 10, a heatmap where the rows are the different canonicity bins and the columns are the weeks of observation. We removed the first week from the heatmap as it contains mostly noise, because in the first week we had no canon information for the protomeme, thus artificially manipulating canonicity in an unrealistic way. We also exclude the last week as it does not contain complete information. From Figure 10 we see that most of the posts consistently belong to the lowest canonicity bin over time. This means that the protomeme pool is characterized by a very high mutation rate, and this rate appears to be systematically high over time. By looking at Figure 10, we realize one fundamental difference between protomeme dynamics and gene dynamics: protomeme mutations are much faster and more radical, at least in the context of social bookmarking's posts.

7 ϕ , One and Zero Model Coefficient Interpretation

In this section, we briefly discuss how to interpret the coefficient of the ϕ Model presented in the article. We discuss the case of Reddit with $l = 25$. In this case, the β coefficient takes value of -1.38 , with $p < 0.0001$. Since it is a logistic model, the value of the beta represents the change in odds for the post of hitting the front page. In this case, a post $\mathcal{M}_1$ with $\Gamma(\mathcal{M}_1, m) = 1$ has an odd factor of $e^{-1.38} \sim 0.25$ compared to a post $\mathcal{M}_2$ with $\Gamma(\mathcal{M}_2, m) = 0$. From our data, we know that, on average, the odds of a meme experiencing two

Figure 11: Distribution of the One and Zero Models' β for varying l , using a negative binomial model estimation. Thin lines represent the 95% confidence intervals.

popularity spikes in a row are about one against four. If we assume a post $\mathcal{M}_1$ having these odds with $\Gamma(\mathcal{M}_1, m) = .5$, then a post $\mathcal{M}_2$ using the same meme m and $\Gamma(\mathcal{M}_2, m) = 1$ will have odds of $\frac{1}{4} \times e^{-1.38/2} \sim 0.125$, i.e. about one against eight.

As for the One Model, in the main article we saw that the β values carry a negative sign, and they take values in the $] - .6 : -.3[$ interval (all β s are significant with $p < 0.0001$). In this model, an observed β of $-.5$ means that a $\Gamma(\mathcal{M}_1, m) = 1$ post, that was in the top 5% scoring posts, expects to obtain a 50% score penalty compared to a $\Gamma(\mathcal{M}_2, m) = 0$ post.

8 One and Zero Models: Negative Binomial

In the paper, we run the Zero and One models using a Poisson regression. However, since their target variable is a post score, these models are likely facing an overdispersion problem. Even if the model testing for overdispersion did not return a confident answer (as shown in Section 5), we still test for this scenario. In this section, we tackle this issue by fitting the a model that accounts for overdispersion: a negative binomial model. The model specification is the same as the Zero and One models reported in the paper.

Figure 11 reports the results of the negative binomial model, and it is an equivalent of Figure 4 in the paper. The negative binomial model returns approximately the same β coefficients, but with larger confidence intervals. This is expected as we could not specify meme random effects in the model and we had to specify meme fixed effects, which have a more powerful explanatory value. Nevertheless, the One model still confirms a highly significant negative effect, corroborating the main result of the paper. The Zero model is affected by higher uncertainty for small front page sizes, but it is still significant for more realistic estimates of how many posts hit the front page in a given day.

References

- [1] Meyerovich, L. A. & Rabkin, A. S. Empirical analysis of programming language adoption. *ACM SIGPLAN Notices* **48**, 1–18 (2013).
- [2] Leskovec, J., Backstrom, L. & Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 497–506 (ACM, 2009).
- [3] Blumm, N. *et al.* Dynamics of ranking processes in complex systems. *Physical Review Letters* **109**, 128701 (2012).
- [4] Ghoshal, G. & Barabási, A.-L. Ranking stability and super-stable nodes in complex networks. *Nature communications* **2**, 394 (2011).
- [5] Coscia, M. Average is boring: How similarity kills a meme’s success. In *Scientific Reports* (2014).