Determining the Influence of Reddit Posts on Wikipedia Pageviews

Daniel Moyer
Department of Computer Science
University of Southern California
941 Bloom Walk
Los Angeles, CA 90089-0781

Samuel L. Carson, Thayne Keegan Dye
Salford Systems
9685 Via Excelencia, Suite 208
San Diego, CA 92126

Richard T. Carson
Department of Economics
UC San Diego
9500 Gilman Drive, #0508
La Jolla, CA 92093-0508

David Goldbaum
Economics Discipline Group
University of Technology, Sydney
Broadway, NSW 2001
Australia

Abstract

The activity of passive content consumers on social media sites is typically difficult to measure. This paper explores the activity of a subset of such consumers by looking at the influence on Wikipedia pageviews of one large Reddit community which frequently links to Wikipedia articles. The subreddit used in this analysis, `/r/todayilearned` (TIL), features a large number of posts on an eclectic set of topics, but excludes current events, which helps rule out the primary threat to being able to make causal statements. Wikipedia's public hourly pageview data provides a unique opportunity to study the influence of a Reddit post on a Wikipedia page at different time horizons.

We here present analyses using posts from 2012 in TIL, showing that the week in which a post references a specific Wikipedia article is associated with a substantial increase in pageviews relative to prior and successive weeks. We then perform a higher resolution analysis on the hourly time series, applying functional PCA to characterize pageview dynamics. We also provide a qualitative analysis of the subset of Wikipedia topics posted to Reddit.

Introduction

Wikipedia has a close but complex set of interactions with other social media. Often used as a casual citation or repository of general knowledge, its common use as a reference on forum sites and media aggregators makes its metadata a valuable resource for the analysis of passive user activity on social media websites. On many of those sites, statistics on passive activities such as viewing a page or clicking a link are either not collected or simply unavailable to most researchers. In the case of links to Wikipedia, however, the data is public and thus we are able to measure these activities.

Given data from a social media site, a relationship can be extracted and analyzed between observable attributes of posted links to that site and the number of Wikipedia pageviews. Due to the high temporal resolution of Wikipedia metadata, analyses of activity on the referencing forum, at least with respect to posts using Wikipedia links, may be conducted via analyses of resulting Wikipedia activity. We validate this claim in the Timeseries Analysis section.

Furthermore, this usage in general provides documented cases of reader interactions on Wikipedia, a context which we believe is also unstudied. While Wikipedia forums and edit histories are both well documented and well studied in the literature (Welser et al. 2011), the vast majority of Wikipedia activity is passive consumption of content (Antin and Cheshire 2010). Due to the assumed causal nature of our joint Wikipedia-Reddit activity, we have a relatively closed environment in which we may study this browsing activity.

In the current work we present preliminary empirical results detailing how links to Wikipedia propagate through a subsection of Reddit, an aggregation and forum site, and how popularity on that site affects Wikipedia usage. We conduct an analysis of approximately 30,000 Wikipedia links posted on a subsection of the site called "`/r/todayilearned`", also known by its abbreviation "TIL". This represents approximately one year of activity on this subsection, also known as a *subreddit*.

We produce net pageview response curves for the posted links, and provide a short analysis of these curves, including an analysis of their functional principal component decomposition. We fit general count models to investigate the relationship between Wikipedia and Reddit. We then provide a qualitative analysis of the Wikipedia categories of each link. The analysis of passive response curve for social media sites and aggregators we believe to be relatively unstudied in academic literature due to the scarcity with which these data are found.

Wikipedia as a Sensor

For many social media applications and studies, the collection of passive activity time series is both desirable and, unfortunately, unobtainable. While in many cases direct action (link sharing, posting, commenting, etc.) may be tracked (Leskovec, Backstrom, and Kleinberg 2009; Mathioudakis and Koudas 2010; Lerman and Ghosh 2010), for a general post (Tweet, Reddit Post, Digg Post, etc.) the passive viewership life-cycle remains publicly a mystery.

Individual sites generally do not publish user logs with pageview activity. Even for sites that mainly provide links as content (so-called social media aggregators), due to the heterogeneous nature of the linked-to sites recovering pageview

counts is generally difficult to the point of intractable, especially with respect to recovering time series data.

Wikipedia, however, provides somewhat of a remedy for this situation; pageviews for every Wikipedia page are provided by Domas Mituzas and the Wikipedia Analytics team. These are collected every hour on the hour. This means that, for posts that specifically involve a wikipedia page, this count may be used as a proxy of the passive activity of users with respect to the post. In particular, for Reddit, which directly links each user to a given site, we receive an upper bound on the amount of passive activity a post receives. If we further assume that the hourly number of pageviews over the course of a week is roughly equal in expectation to the same time series for either of its adjacent weeks, then, assuming no other "special" events occur, we can recover the passive activity curve of a Reddit post by subtracting the succeeding week from the preceding week. This relies on effective post life cycles being shorter than one week, something enforced by Reddit's default sorting algorithm, which prioritizes new content.

Wiki Background and Terminology: Before providing results we will review the relevant Wikipedia terminology. Wikipedia is composed of *pages* or titles, each with a corresponding URL. Associated with most pages are *categories*, loose tags which associate the page with other, similar pages. Outside of a top layer of categories (of which two kinds exist), little structure is provided. There are 24,739 unique categories observed in our observational window.

Reddit Background and Terminology: We will also review here the relevent terminology particular to Reddit as a social media aggregator. As a site Reddit is comparable to Pinterest or Tumblr (and indeed interesting comparisons can be made), in that content is usually not generated by the site itself but taken from other sites. Reddit is at the extreme in this sense, in that, except for text based "self-posts" and comments, the site does not host any of its own content, and simply links to other sites. All three communities are almost entirely user driven, and Reddit is also usually user moderated.

Reddit's content is focused around the *post*, also referred to as a *submission*. Each post consists of a link and/or lightly formatted text, as well as a title. The posts may then be *voted* on by other users (by default each post is upvoted by its *author*, the user that submitted the post). An *upvote* indicates reader approval, while a *downvote* indicates disapproval. Posts are ordered on Reddit's pages using one of five algorithms. The default algorithm is called *hot* and is predominantly impacted by submission times, quickly cycling out old content. In general it is conjectured that most users do not vote (Van Mierlo 2014), and the *volume* of votes is at best an upper bound of the number of actual voters due to bot participation and Reddit's built-in bot fighting algorithm which artificially inflates vote counts. Voting is Reddit's lowest form of active participation. (Salihefendic 2010)

As well as hosting the links, Reddit hosts *comments*, which facilitate discussion of the main content by users. Though not the focus of this paper, it should be known that the *content* of these comments can become quite complex, and may affect voting.

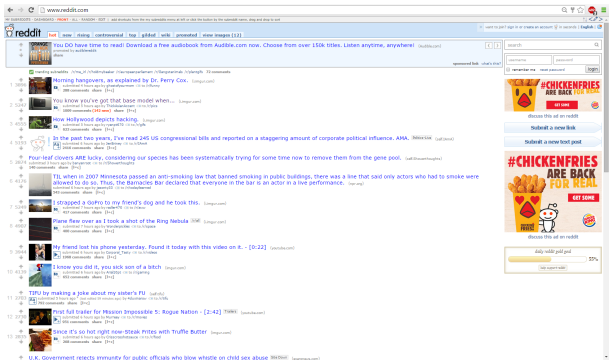

Figure 1: A typical example of Reddit's front page interface. This particular UI has been augmented with a tool called Reddit Enhancement Suite, which adds additional options and information.

Reddit is subdivided into *subreddits*, each of which focuses on a some category or theme. The creation and curation of the subreddits is user motivated. Upon accessing Reddit, each user is presented with the so-called "front page", a landing page which includes a mix of all of his or her chosen (*subscribed* to) subreddits. Users not logged in will be presented with a mix of *default* subreddits. All new users are also automatically subscribed to the *defaults*; "`/r/todayilearned`" is one such default subreddit. In general most subreddits are not viewed by more than a small fraction of the userbase. There are more than $90,000$ subreddits, and individual ones will often establish their own rules and guidelines on submitted content.

In this paper we focus on the behavior of only one subreddit, "`/r/todayilearned`", often abbreviated to "TIL". This particular subreddit focuses on content that contains interesting yet not widely known information, and allows only a single form of submission - a link and a title. Examples include a post titled

> "TIL An American gymnast with a wooden leg won six medals, including three gold, in a single day at the 1904 Olympics. He was the only Olympian to have competed with a prosthetic limb for the next 100 years, until 2008"

linking to *wikipedia.org/wiki/George_Eyser*, and a post titled

> "TIL that a young Pablo Picasso had to burn his own paintings for warmth in his freezing Paris apartment"

linking to *wikipedia.org/wiki/Pablo_Picasso#Before_1900*. The subreddit has a rule for submissions that explicitly states

> "No news or recent sources. News and any sources (blog, article, press release, video, etc.) more recent than two months are not allowed."

During our observational window over 18% of the posts on `/r/todayilearned` were Wikipedia links, and by the end of 2012 the subreddit was one of the 10 largest subreddits with over 2 million subscribers, making it ideal for this study.

Data Structure and Effect Duration: Both the Reddit and Wikipedia data have straightforward count structures for their votes and pageviews respectively.

We model each Reddit submission as a shock to the corresponding Wikipedia pageview count. These shocks may vary in strength. Furthermore, there is no explicit end to the shock period because the submission is never erased, but we observe that even a popular submission will not be able to maintain a high page ranking for more than a few days under the default *hot* algorithm. Ranking within the subreddit corresponds to visibility and thus passive viewing of the link itself. Clearly then, a wiki link submitted on reddit can be analyzed as a temporary shock rather than a permanent one. Because the subreddit does not allow current events or sources less than two months old, we generally do not expect there to be other external shocks to the same page at the same time.

Wikipedia on Reddit

Before analyzing the combined content, we present a short analysis of the performance of Wikipedia links on TIL. Our dataset consists of approximately 50 weeks of content, spanning most of 2012; we recorded every post possible, meaning that, with exceptions for deletions, server errors, and privacy policies, we have *every* post made to the subreddit during that time window.

While not a majority of the content on the subreddit, Wikipedia links account for 17.69% of the total posted content, which rises to 18.89% after removing deleted posts. This is about 30000 individual links, and is the leading domain of content (followed by "youtube.com" and self-posts to TIL). As seen in Figure 1, out of domains with over 100 submissions, wikipedia clearly dominates the submission pool.

Domain	Number of Links
en.wikipedia.org	30005
youtube.com	8636
self.todayilearned	5893
imdb.com	2713
imgur.com	1470
reddit.com	1216

Table 1: Top Domains ordered by number of submissions. "en.wikipedia.org" is clearly the leading domain.

Wikipedia links also accumulate more upvotes, downvotes, netvotes, and comments than other links. The distributions of each are obviously skewed, as most posts in general receive only a few (less than 5) votes total (and one comment), while a very small minority of posts receive thousands. We thus report the Wilcoxon Rank Sum test as well as usual summary statistics, found in Table 2.

Medians	All	Wiki	Non-Wiki	WRS p-value
Upvotes	4	17	2	$< 10^{-15}$
Downvotes	3	8	2	$< 10^{-15}$
Comments	1	2	0	$< 10^{-15}$

Table 2: Table of summary statistics for posts submitted to /r/todayilearned, including the Wilcoxon Rank Sum test (WRS).

Timeseries Analysis

It is useful to conduct a preliminary analysis to verify that the appearance of a link to a Wikipedia article in a TIL post is associated with an increase in the pageviews of the relevant Wikipedia article. There are 28,497 posts in 2012 with a link to a Wikipedia article as well as one preceding and one succeeding week of uncorrupted Wikipedia pageview data. Looking at the week prior to such a TIL post, we find that the mean number of pageviews of the relevant Wikipedia articles is 12,249 and the median number of pageviews is 2,479. For the week starting with the TIL post, the mean Wikipedia pageviews is 19,010 with the median being 4,137. The week subsequent to TIL post has the mean pageviews being 12,553 and the median being 2,403.

Visual inspection suggest that pageviews increase considerably in the week of the TIL post. This is confirmed by formally conducting t-tests of the difference in means. The t-test of the difference in means between the second and first weeks has a statistic of 34.30 ($p < .001$). A t-test of the difference in means between the second and third weeks has a statistic of 26.02 ($p < .001$). A signed rank test of the equivalence of the medians in the second and first weeks yields a z-value of 91.30 ($p < .001$) and for the second and third weeks a z-value of 96.53 ($p < .01$). All of these test statistics suggest overwhelming rejection of the equivalence of key summary statistics comparing the relevant Wikipedia pageviews from the week of the TIL post to the week prior and the week following.

If all of the TIL induced Wikipedia activity is constrained to one week and if there was no time trend in the data, we would expect to see mean and median pageviews for the relevant Wikipedia articles to be statistically indistinguishable. We do not find this to be quite the case. The t-test for the difference in mean pageviews between the first and third weeks is -2.11 ($p = .035$). This 2.5% increase in mean pageviews between the first and the third week may reflect a small amount of the TIL-induced pageviews falling into the third week, or just natural variability in the data where our very large sample size provides sufficient power to reject very small differences at conventional confidence levels. Looking at the difference in median pageviews provides some support for this conjecture. In contrast to the pattern for mean pageviews, median pageviews are higher in the first week than in the third week with the test of equality of the medians having a z-statistic of 6.04 ($p < .001$).

Based on this analysis we will obtain a baseline for expected pageviews in the absence of the TIL post by taking the average of the pageviews for the relevant Wikipedia articles in the first and third week. This will generally make the analysis performed in the next section a bit conservative in that we are attributing the small increase in third week pageviews to the background level of pageviews rather than potentially being the result of TIL activity.

Generalized Negative Binomial Model						
Generalized negative binomial regression				Number of obs = 24765		
Log pseudolikelihood = -236097.62				Pseudo $R^2 = 0.0841$		
Wiki_2						
	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]	
logWiki_{13}	.7330604	.004519	162.22	0.000	.7242034	.7419175
netscore	.004557	.0003138	14.52	0.000	.0039419	.0051721
netscore^2	-6.03e-06	8.61e-07	-7.00	0.000	-7.72e-06	-4.34e-06
netscore^3	4.13e-09	8.93e-10	4.63	0.000	2.38e-09	5.88e-09
netscore^4	-1.24e-12	3.65e-13	-3.39	0.001	-1.96e-12	-5.24e-13
netscore^5	1.31e-16	5.04e-17	2.60	0.009	3.22e-17	2.30e-16
comments	.005626	.0009502	5.92	0.000	.0037637	.0074883
comments2	-.0000121	2.97e-06	-4.08	0.000	-.0000179	-6.30e-06
comments3	1.18e-08	3.48e-09	3.39	0.001	4.98e-09	1.86e-08
comments4	-5.03e-12	1.62e-12	-3.11	0.002	-8.21e-12	-1.86e-12
comments5	7.57e-16	2.55e-16	2.97	0.003	2.57e-16	1.26e-15
logDayOfYear	-.02738	.0105838	-2.59	0.010	-.0481239	-.006636
Constant	2.285674	.0673804	33.92	0.000	2.153611	2.417737
$\ln \alpha$						
logWiki_{13}	-.2419808	.0112078	-21.59	0.000	-.2639477	-.2200139
netscore	.0008954	.0000488	18.36	0.000	.0007998	.000991
Constant	1.311628	.0829941	15.80	0.000	1.148963	1.474294

Table 3: Generalized Negative Binomial Model co-efficients.

Does the Magnitude of TIL Activity Help Predict the Increase in Wikipedia Pageviews?

Our analysis in this section looks at whether the magnitude of activities within TIL related to the post, votes, and comments helps to predict the increase in relevant Wikipedia page views. It uses the same three week setup as the previous section. This analysis will be very conservative if the influence of the typical TIL post lasts only a day or two, as appears to be the case, because we will then be effectively pooling those days where the TIL post increases relevant Wikipedia pageviews with five or six days where there is no influence. The model to be estimated takes the form of:

$$\text{Wiki}_2 = f(\text{Wiki}_{13}, \text{Netscore}, \text{Comments}, \text{Time Variables})$$

where Wiki_2 is a vector of pageviews of Wikipedia articles associated with TIL posts in the second week, Wiki_13 is the average of these pageviews over the first and third weeks, Netscore is the difference between Upvotes and Downvotes, Comments is the number of comments on the corresponding TIL post, and there are a variety of possible time variables such as a time trend, month indicators, day of week indicators, and hour of the day indicators.

A few of these variables deserve special attention. First, while our 2012 Reddit data includes Upvotes and Downvotes, only Netscore is currently available via Reddit's API. Reddit has since stopped making Upvotes and Downvotes available as part of its efforts to combat vote-manipulation bots. We only use TIL posts in the analysis in this section which have non-negative Netscore. This drops about 10% of the observations. Most of these are consistent with little activity as such posts tend to drop way down the list of available post. There are a few outliers with a sizable negative Netscore values that make modeling negative Netscores problematic. Examination of these cases suggests that what started out as a standard post with reasonable interest degenerated into an organized flame war and downvoting campaign. Second, Netscore and Comments both have some large values. Examination of these for Netscore did not reveal observations that appeared odd in the sense of being inconsistent with the presumed underlying data generating process. This was not the case for some very large values for Comments where there periodically was an intense back and forth between a relatively small number of users which suggested that increasing the number of comments might not always be predicted to increase pageviews of the relevant Wikipedia articles. We will operationalized the influence of both Netscore and Comments in terms of a fifth order polynomial in those variables.

Our dependent variable, Wiki_2, represents count data. The simplest count data model is a poisson regression model which parameterizes the expected count in terms of a matrix, X, of predictor variable. It imposes the restriction that the conditional mean and variance are equal, a characteristic that our data do not have. However, the poisson model has been shown to be the quasi-maximum likelihood data and provides consistent estimates of the regression parameters and of their standard errors if an appropriate robust variance-covariance matrix is used (Wooldridge 2010). If one wants to additionally model the nature of the over dispersion it is typical to move to a negative bionomial model (Hilbe 2011; 2014), where the variance is usually modeled as a poisson-gamma mixing distribution that varies with the condition mean. A more flexible version of model known as a gen-

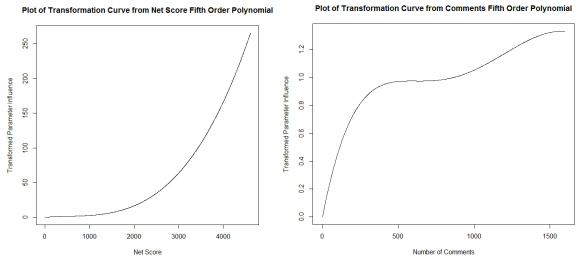

Figure 2: Fifth Order Polynomial transformation fit for Netscore and Comments, respectively.

eralized negative binomial regression model allows for the over-dispersion scaling parameter to be a function of observable covariates. We estimate the parameters of this model in Table 3, where in addition to the fifth order polynomials in terms of Netscore and Comments, we include the log of logWiki_{13} and the log of the time trend.

The logWiki_{13} parameter is the most important variable in the model and given the specification can be interpreted as an elasticity of .73. An almost identical estimate was obtained in a log-log OLS model while a somewhat higher estimate, .85, was found using a Tukey biweight robust regression estimate. The logDayOfTheYear variable suggests that pageviews are decreasing at a slow rate over the year. We had no strong prior on the sign or magnitude of this coefficient since more Reddit viewers on TIL might be expected to increase pageviews, and more Wikipedia articles and linking to a larger number of them would tend to have the opposite effect.

The null hypothesis that the magnitude of the TIL activity does not influence the magnitude of pageviews of relevant Wikipedia articles is that the Netscore and Comments parameters are all zero. This is obviously not the case as all of these parameters are individually significant and jointly significant at the $p < .001$ level. In Figure 2, we plot the curve implied by the Netscore 5^{th} order polynomials. This shows that the Netscore influence on Wikipedia pageviews is increasing and does so at a steeply increasing rate once Netscore is large. This is not surprising because a very large Netscore typically indicates that the TIL post has jumped to a reasonably high position on the home Reddit page. The influence curve for the 5^{th} order Comment polynomial is plot in Figure 2. This shows steeply rising influence up through about 500 comments and much slower increases beyond that point. Again, this is consistent with prior expectations that a TIL post that draws a large number of comments is likely to be of interest to a sizable number of Reddit users but that after some point the commenting exchanges going on dont send many fresh users to Wikipedia.

The variance of the model is represented by modeling $\ln(\alpha)$ with α being the over-dispersion parameter. The constant term here is sizable, suggesting considerable over-dispersion. Limited exploration with covariates shows that this over-dispersion is decreasing in the log of Wiki_{13}, suggesting that Wikipedia articles with considerable background pageviews are more predictable than infrequently accessed pages, and increasing with Netscore. The linear version of Wiki_{13} and the log version of Netscore resulted in considerably worse log-likelihoods, while the addition of Comment and time variables result in little improvement.

Inclusion of other time-related variables generally result in insignificant or marginally significant parameter estimates. We suspect that this may be because while there appear to be strong day of the week and hour effects, these are adequately capture by the Netscore and Comment variables. A falsification test that substitutes Wiki_1 for Wiki_2 as the dependent variable and uses logWiki_3 along with the polynomials in Netscore and Comments and logDayOfTheYear shows that the coefficients on all of the polynomial terms are zero. Since they did not occur until after Wiki_1 had taken place, this is the expected result unless something in the model being estimated intrinsically produced biased parameter estimates. All of this work points to TIL posts linking to Wikipedia articles causing pageviews of the relevant articles, and that the magnitude of this effect is clearly linked to the magnitude of internal TIL activity related to the post. We now turn to modeling the short run dynamics of how this process works.

Functional Data Analysis

Besides predicting on summary statistics of the timeseries we also conducted an analysis of shape of the response curves. While obviously scaled by the number of viewers, we would also like to know whether responses have different distributions. That is, whether or not some posts pass quickly while others slowly rise, or whether all posts have approximately the same lifecycle.

We are able to directly construct response curves by subtracting an approximation of the background number of pageviews from the signal during and immediately after the stimulus (the post). Here, we subtract the preceding week of counts from the week directly succeeding the post.

We normalize each timeseries in the L^2 sense. Directly averaging these timeseries at each time point, we recover the curve shown in Figure 4. While too brutal to discern differences between response curves, we see that, on average, the response peaks within twenty four hours. Furthermore, the average response curve has two distinct maxima, followed by a sharp decline. Most of the response is contained within the first two days.

The analysis of the differing shapes of the curves themselves lies squarely within the span of Functional Data Analysis and its toolset (Ramsay 2006; Viviani, Grön, and Spitzer 2005). Used primarily in setting where over the course of many trials a (random) continuous function is sampled over time, here we may view the response curve as one such random function, and each of our posts as one trial.

In particular, here we apply functional Principal Component Analysis (fPCA) (Ramsay 2006) to the set of regularized response curves. While fPCA is quite similar to its discrete counterpart, it usually involves a projection of sampled functions onto a set of basis functions; this necessitates a choice of basis, which is not unique for finite sample points. We use fourth order b-splines, as our data is generally non-

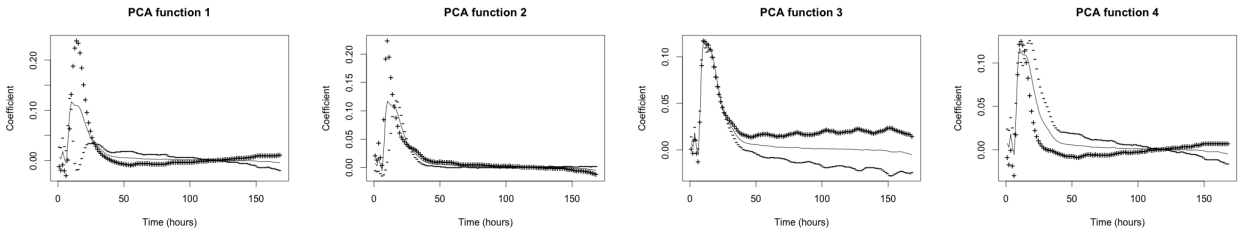

Figure 3: Top 4 PCA functions. "+" symbols denote the positive direction along the component, and "−" denotes the negative direction (note that this need not be a scale, since the component is composed of a mixture of basis functions).

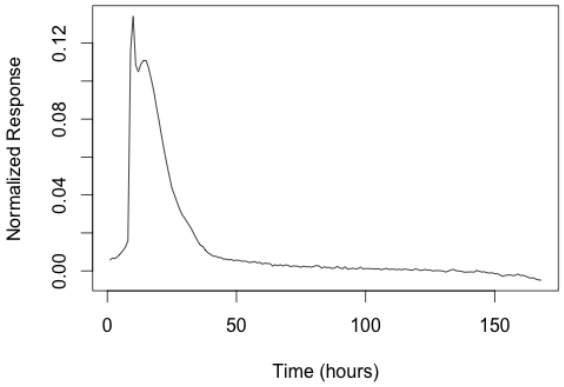

Figure 4: The normalized and averaged response curve.

periodic. This analysis utilizes the `fda` package in `R` (Ramsay et al. 2013).

As seen in Table 3, once the mean is re-added to the components all components prominently feature a spike at about 24 hours. However, looking at component 4, we note that the PC does little to the magnitude of the spike or the resulting tail, but shifts the location of this spike. In particular, as scores decrease for component 4, the peak shifts to the right. We hypothesize that our mean signal's two peaks are the average between slowly shifting peaks, and that, for some number of posts, their popularity is delayed by a lag time.

In order to partially validate this, we cluster the timeseries in the space generated by the components using k-means. From the two clusters produced, we test the differences between the within-cluster distributions of posting-times. In other words, we check using a chi-squared test whether the distributions of posting times over the day were the same in both clusters. As displayed in Table 4 and Figure 5, the clusters exhibit have a significant shift of their posting frequencies over the day. The second cluster posts more frequently during the second half of the day, meaning that posts during the later half of the UTC day.

This corresponds with the second cluster posting during the morning hours of the US continent. Though statistics for the distribution of users by location have not been collected for Reddit, this leads us to believe that posts made early in the morning have a higher propensity to be ignored for several hours.

	χ^2 Results
χ^2 Statistic	85.7432 (df = 23)
p-value	3.64×10^{-9}

Table 4: Results of the χ^2 test for cluster posting-time difference.

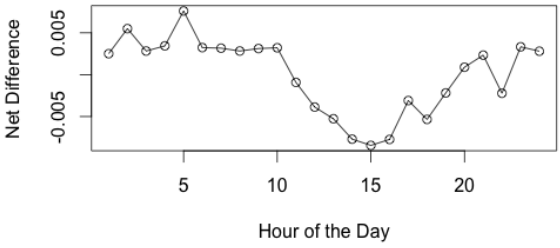

Figure 5: Net difference in normalized distributions (Cluster 1 - Cluster 2).

Articles That Attract Reddit's Interest

We now turn our attention to the types of pages which attract interest. While this is not directly relevant to the shape or size of response curves, it provides insight into what types of pages are chosen by post authors and which of the chosen pages perform well. This information is accessible in raw form via the categories of each Wikipedia page; however, while a topological ordering of article categories exists, it is less than informative and quite complex (Nastase and Strube 2008).

Towards this end we instead apply a topic model to amalgamated category data. For each page we combine each of its categories into one single "document". Each document corresponds with a post on Reddit, and thus has associated upvotes, downvotes, etc.. We then remove stopwords and

Rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	treaties	albums	canadian	recipients	heads	singlechart
2	united	certification	21^{st} century	grand	russian	single
3	convention	recording	male	knights	soviet	numberone
4	carbon	englishlanguage	20^{th} century	united	heros	certification
5	laws	grammy	living	members	cold	songs
% Weight	0.0184	0.0285	0.1268	0.1200	0.0346	0.0515
% Score	0.0242	0.0277	0.1224	0.1499	0.0311	0.0322
Rank	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	treaties	english	treaties	treaties	languages	films
2	convention	20^{th} century	1944	peace	subjectverbobject	englishlanguage
3	laws	lgbt	1967	space	fusional	screenplays
4	children	british	international	history	iso	film
5	1980	21^{st} century	aviation	cold	analytic	best
% Weight	0.0179	0.4411	0.0225	0.0171	0.0119	0.1096
% Score	0.0260	0.3990	0.0268	0.0206	0.0073	0.1327

Table 5: Topics with top words per topic as well as the total percent of the document weights and weighted net score.

general Wikipedia editing related terms. Due to the short length of most of these documents, we choose not to use Latent Dirichlet Allocation (Tang et al. 2014), as it historically has poor performance on sparse documents. Instead we here use Non-negative Matrix Factorization (Lee and Seung 1999; Saha and Sindhwani 2012).

These topic models are by no means definitive; they only serve to provide qualitative insight into the interests of "`/r/todayilearned`". We show the top 5 words for each topic in Table 5, along with both the percentage of the total document weights and the percentage of the weighted scores. This second statistic is computed by multiplying each post's net score by its document's weight in the given topic. Documents can contribute to multiple topic scores.

Topic 8 is clearly dominant, but upon inspection appears to encompass a wide range of actual articles. However, when weighted by scores, it performs considerably worse than its proportion of weights would suggest, indicating that it is less popular. Topic 12 on the other hand is clearly about films, and has a higher proportion of the score.

The presence of these topics in general provides insight into the selection biases of "TIL" users. In particular, music, people, politics, films, and history seem to be common topics to submit.

Previous Work

Little research has been done on social media viewer trends. While sharing on large social media sites or across multiple sites has been a common focus of study (Leskovec, Backstrom, and Kleinberg 2009; Mathioudakis and Koudas 2010; Lerman and Ghosh 2010), studies of the so-called lurkers usually rely on specialized sites with fully accessible data (Panciera et al. 2010; Muller et al. 2010; Shami, Muller, and Millen 2011). These studies also generally focus on differences between lurkers and contributors.

Of the studies of sharing on large social media sites, (Lerman and Ghosh 2010) similarly found that the majority of activity on a link takes place within 48 hours. In particular, the authors found that on Digg, a site at the time comparable in size to Reddit, most stories would be buried within 20 minutes, with a similar result being shown for Twitter. With a small number of votes, however, the story could be pushed to the front page; this provided some lag between submission and "jump". Reddit's dynamics and method for post/story display differ slightly, but we find similar results in the pageview statistics.

In a much larger study, (Leskovec, Backstrom, and Kleinberg 2009) tracks the evolution of phrase clusters across a large number of sites. While in a slightly different setting and measuring the active rather than passive response, the authors' results also show a short, 48-hour-level response for most stimuli.

Numerous studies have also been undertaken directly relating to Wikipedia. A particular area of study has been the role of Wikipedia editors and their behaviors on article content and quality (Kittur and Kraut 2008; Kittur et al. 2007b; 2007a). Other studies have analyzed the structure of Wikipedia and its relationship with user activity and content development (Blumenstock 2008; Capocci et al. 2006; Voss 2005).

Wikipedia is also a well studied site, so much so that it has its own Wikipedia Research Network[1]. Numerous papers study the network generated by the site's pages (Voss 2005; Capocci et al. 2006), as well as the editor social networks that develop on the associated forums (Kittur et al. 2007a; Niederer and Van Dijck 2010). These usual tracks are rich in data, thus few papers have studied Wikipedia readers.

One important paper that does focus on readers, (Zhang and Zhu 2010), tracks the number of edits made to the Chinese Wikipedia after a large number of readers were blocked; surprisingly the number of edits dropped even from non-blocked users. Another, (Antin and Cheshire 2010), posits that readers are not actually free-riders but contribute value to editors and readers.

[1] http://meta.wikimedia.org/wiki/Wikimedia_Research_Network

Conclusion and Future Work

In this paper we have provided strong statistical evidence suggesting Reddit threads affect Wikipedia viewership levels in a non-trivial manner. We then explored some of the more complex short term dynamics, as well as qualitative analysis of the types of articles submitted. We have demonstrated the use of Wikipedia pageview statistics as a tool to recover counts of otherwise unobservable user activities.

In future works we hope to expand both the functional analysis of the generated response curves, as well as explore implications for the editing of the pages posted to Reddit. Finally, we hope to broaden our current study to include Wikipedia links used in Reddit comments.

References

Antin, J., and Cheshire, C. 2010. Readers are not free-riders: reading as a form of participation on wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 127–130. ACM.

Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, 1095–1096. ACM.

Capocci, A.; Servedio, V. D.; Colaiori, F.; Buriol, L. S.; Donato, D.; Leonardi, S.; and Caldarelli, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E* 74(3):036116.

Hilbe, J. 2011. *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. 2014. *Modeling Count Data*. Cambridge University Press.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37–46. ACM.

Kittur, A.; Chi, E.; Pendleton, B.; Suh, B.; and Mytkowicz, T. 2007a. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web* 1(2):19.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007b. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 453–462. ACM.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM* 10:90–97.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. 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.

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 1155–1158. ACM.

Muller, M.; Shami, N. S.; Millen, D. R.; and Feinberg, J. 2010. We are all lurkers: consuming behaviors among authors and readers in an enterprise file-sharing service. In *Proceedings of the 16th ACM international conference on Supporting group work*, 201–210. ACM.

Nastase, V., and Strube, M. 2008. Decoding wikipedia categories for knowledge acquisition. In *AAAI*, 1219–1224.

Niederer, S., and Van Dijck, J. 2010. Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New Media & Society* 12(8):1368–1387.

Panciera, K.; Priedhorsky, R.; Erickson, T.; and Terveen, L. 2010. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1917–1926. ACM.

Ramsay, J. O.; Wickham, H.; Graves, S.; and Hooker, G. 2013. *fda: Functional Data Analysis*. R package version 2.4.0.

Ramsay, J. O. 2006. *Functional data analysis*. Wiley Online Library.

Saha, A., and Sindhwani, V. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 693–702. ACM.

Salihefendic, A. 2010. How reddit ranking algorithms work. http://amix.dk/blog/post/19588.

Shami, N. S.; Muller, M.; and Millen, D. 2011. Browse and discover: social file sharing in the enterprise. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 295–304. ACM.

Tang, J.; Meng, Z.; Nguyen, X.; Mei, Q.; and Zhang, M. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, 190–198.

Van Mierlo, T. 2014. The 1% rule in four digital health social networks: an observational study. *Journal of Medical Internet Research* 16(2).

Viviani, R.; Grön, G.; and Spitzer, M. 2005. Functional principal component analysis of fmri data. *Human brain mapping* 24(2):109–129.

Voss, J. 2005. Measuring wikipedia.

Welser, H. T.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; and Smith, M. 2011. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, 122–129. ACM.

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Zhang, X. M., and Zhu, F. 2010. Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review, Forthcoming* 07–22.