4-22-2011

Using Textual Features to Predict Popular Content on Digg

Paul H. Miller
University of Nebraska-Lincoln, pmiller987@gmail.com

USING TEXTUAL FEATURES TO PREDICT POPULAR CONTENT ON DIGG

by

Paul H. Miller

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: English

Under the Supervision of Professor Stephen Ramsay

Lincoln, Nebraska

May, 2011

PREDICTING THE POPULARITY OF ONLINE CONTENT

Paul H. Miller, M.A.

University of Nebraska, 2011

Adviser: Stephen Ramsay

Over the past few years, collaborative rating sites, such as Netflix, Digg and Stumble, have become increasingly prevalent sites for users to find trending content. I used various data mining techniques to study Digg, a social news site, to examine the influence of content on popularity. What influence does content have on popularity, and what influence does content have on users' decisions? Overwhelmingly, prior studies have consistently shown that predicting popularity based on content is difficult and maybe even inherently impossible. The same submission can have multiple outcomes and content neither determines popularity, nor individual user decisions. My results show that content does not determine popularity, but it does influence and limit these outcomes.

Table of Contents

List of Multimedia Objects

Introduction

Most data mining projects studying social networks have focused on network topology, while relatively few studies have used data mining to study art and cultural artifacts to predict popularity. I used various data mining techniques to collect samples from digg.com, a social media site, to try to predict the popularity of news articles submitted to the site. Digg.com is a user-contributory, news-aggregator website, where users post links to news articles, blog posts, YouTube videos, or even just a single image, and then vote them up or down by "digging" or "burying" these submissions. Digg is one of the largest social media sites on the web and is the leading user-contributory news site with an estimated three to four million users posting about 16,000 stories per day to the site (Lerman and Ghosh 5).

One of the most useful results of data mining is not the prediction itself, but the insights that the algorithm can provide and the structural patterns that algorithm can reveal. My study is not the first to apply data mining techniques to popularity and cultural artifacts. Two companies have been applying data mining to music and movies over the past five years. Epagogix Ltd. has had moderate success predicting successful movies, but Platinum Blue Music Intelligence claims to have identified five Grammy winners in 2005. Although the companies have been financially successful, the trends they identified still seem like fundamental information for scholars interested in studying popular culture—how can anyone study pop culture without asking how or why it became popular in the first place? Unfortunately, the algorithms that Platinum Blue and Epagogix found remain proprietary and whatever patterns they found are kept strictly

confidential. Hopefully, my project and the series of recent academic projects that have studied Digg can reveal how it is that cultural artifacts become popular.

Tied to the issue of popularity on Digg is the debate over reading practices on the Web. The argument typically is that pop culture wants the shortest, fastest read that is posted on the internet. At the beginning of my study, I assumed that readers would favor these shorter, simpler articles. I assumed that they would prefer articles with pictures, rather than purely textual articles. However, reading practices on the Web are not that simple. Digg users are much more likely to click on articles with "pic" in the title, but actually significantly less likely to vote for that article. Users are also much less inclined to vote for shorter articles (under 150 words) than longer ones. While users on Digg do not show a preference for particularly long stories or long sentences, they do not discriminate against them either. It seems that reading practices on the Web cannot be reduced to simply favoring the most consumable reading.

Signal from Noise

Kevin Rose founded Digg in 2004 to create a sort of alternative search engine: a site that separated signal from noise, with the idea that humans provide a more informed judgment of the stories than the algorithms of Yahoo! News or Google News. People use Digg for different purposes than Google: no one searches for an address, or looks up the nearest pizza parlor on Digg. However, it functions as a form of bibliographic control in the same way that Google and Yahoo! filter information in the overwhelming environment of the web.

Crowdsourced, user-contributory websites are the newest way to contend with one of the biggest challenges of the information age: too much information. Search engines have become some of the most successful companies on the Web, providing a necessary filter and organization to this overload. The internet is often romanticized for the access it affords, but this notion is ultimately a naïve one: much of humanity's knowledge has been digitized, but most of it people will never see, and it is how search engines rank and organize that knowledge that largely determines which webpages people do see. These search engines control the information we see, the webpages we read, and the stories and trends that go "viral." In an information-driven economy, the control of information has become a central component of contemporary social, political and economic processes. Simply put, there are winners and losers in our society and search engines partly determine the people that go in each pile.

Search engines rankings are so lucrative that the entire Search Engine Optimization industry has developed to manipulating them. SEO developers have successfully reverse engineered the algorithm when they can predict what Google or Yahoo! will return for a given search. When they know what each site will be ranked then they have defined the process that determines the ranking of each story. In a way, this is what I have attempted to do for Digg. A traditional search engine has a central algorithm that organizes the ranking process, but Digg's crowdsourced search engine is controlled by a very different process. My project attempts to identify the social algorithm that ranks these stories.

A number of studies have recently developed models to predict popularity on Digg and other collaborative rating sites. These studies consistently argue that the topology of the Digg social network and the design of the website create vast disparities in the number of users who see each submission. These disparities largely structure popularity on the site, unevenly distributing the visibility of submissions so that stories receive a highly unequal number of views. Differences between submissions' visibilities structure the uneven distribution of votes on the site. Overwhelmingly, these prior studies have consistently shown that predicting popularity based on content is difficult and maybe even inherently impossible. The same submission can have multiple outcomes and content neither determines popularity, nor individual user decisions.

I collected 10,187 stories from Digg between November 18 and November 25, 2010, scraping the URL for each submission and collecting textual metrics about the story. I collected a successive sample of 2,200 stories from March 13 to March 22, 2011. For the second sample, I implemented a model of the social network and site architecture to estimate the number of users that see a story to account for the disparities in visibility. Using this estimate, I was able to examine the effect of content on both popularity and the distribution of individual user decisions. In addition, I scraped the number of exit clicks to each story to see the number of users that clicked through to each submission. These metrics allowed me to calculate the ratio between the number of users that see a title to the number that click on that title, and the ratio of users that click on the title to the number that vote for it. I used these ratios to calculate, and distinguish between, the appeal of the story title and the appeal of the story content.

The Digg Website

The Digg website is divided into three main sections: Top News, Upcoming, and My News. Digg users can follow another user, and see all of the stories that other person votes on or submits, as those stories are posted to each person's My News homepage. Users try to follow people who vote on topics they are interested in, allowing them to create a sort of customized My News homepage. When someone submits a story it is posted to their followers' My News homepages, automatically with one digg (from the submitter). If it gets a second vote then it is posted to the Upcoming page. Stories that get enough votes—typically at least 60—are promoted to the Top News page, where thousands of people see it each hour. The design of the site creates an inherent competition for users as they try to get as many stories promoted as possible. They try to be the first person to submit a story on a breaking news topic and to get followers by digging and submitting high quality stories. Users create their own titles and a brief description for each link—usually they stick with the article's title, but frequently modify it by adding "(pic)" or "(video)" at the end of the title. When they submit a story they also tag it with one of the Digg's ten topic categories: Business, Entertainment, Gaming, Lifestyle, Offbeat, Politics, Science, Sports, Technology, and World News.

Very few stories are ever promoted. Out of 16,000 stories submitted each day, only about 1,200 ever get the second vote that places them on the Upcoming page. Of these, fewer than 300 get ten votes or more. Out of the original 16,000 submitted each day, fewer than 100 are promoted.

Gaming Digg

While most people use the site for fun, the social-media marketing industry often exploits it as a venue to spam with product links. This spam is usually much more subtle, integrated into a blog post or a news article about the product rather than a direct link to the company website. A promoted story translates to real-world profit through what users call the "Digg Effect." A successful story can bring tens of thousands of page views per hour, sometimes a quarter million hits in an afternoon, often crashing the promoted link. Although Digg officially prohibits people from using it for professional marketing, they have not been able to completely control it. A common issue on the site is that a relatively small number of users control a disproportionate amount of the Top News page.[1] Many social media marketers have monetized their profile, spending hours each day digging stories, building an enormous following, and then charging companies to submit material through their profile.

One of the challenges of using data mining to collect Digg is the frequent modifications to the site. In the fall of 2010, when I got my initial data on Digg, the site had removed the "bury" option for users after they found that a large group of Tea Party activists, called the "Digg Patriots," had organized to censor progressive political stories

[1] Digg has repeatedly taken steps to reduce the influence of top users since 2006 (Rose). In the March 2011 sample, there were large discrepancies between the number of votes a story required to be promoted to Top News. Most stories were promoted around 60 votes, while a small number were not promoted until they had over 100; several were not promoted until 190 votes—probably because the Digg promotion algorithm requires a higher number of votes for stories submitted by users with a large number of followers. Despite this, the sample shows that having more friends is still an enormous advantage for promoting stories.

(Schott). However, during the March 2011 sample, the bury button was back after users complained that spam was infiltrating the site without it.

Diggbot

At the beginning of the project, I initially believed that popularity on Digg was about simplicity: short articles with short sentences, short words and short paragraphs would appeal to Digg users, whereas longer, more difficult articles would be less likely to be promoted to the Top News page. However, the results indicate the opposite: stories with comparatively complex syntax—longer sentences, words and articles—actually receive a higher average number of votes.

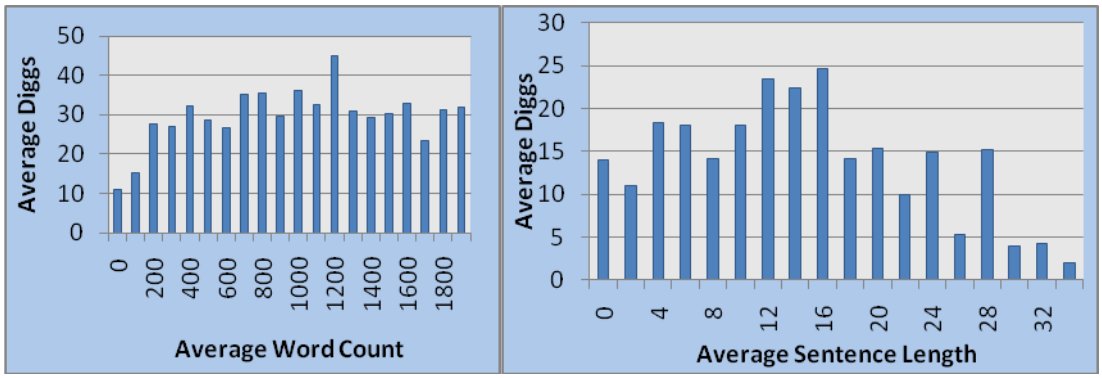
To study popularity on the site, I wrote a Webcrawler to scrape stories as they were submitted to Digg. My program scraped the Upcoming/Recent page every ten minutes to find the newest submissions. For each story, Diggbot pulled down the Story ID number assigned to each story by Digg, as well as the submissions' URL. It then scraped each story's URL to determine its total word count and the average length of each word, sentence and paragraph. The Stanford Part-of-Speech Tagger identified the percentage of nouns, verbs and adjectives in the story. The Digg API allowed me to determine the article's assigned topic, its digg count, the number of followers for the user who submitted the story, and the exact time it was submitted. The final vote count was collected forty eight hours after a story was submitted, which was enough time to determine whether a story had gone viral, since forty eight hours allows submissions to accumulate over 99% of their total votes, with the exception of highly successful stories

that are not only promoted to the Top News page, but are then promoted to Top News/24 Hours and Top News/7 Days.

The program scraped the site from November 18 to November 25, 2010, collecting 10,187 submissions. The sample included stories with a minimum of 2 votes and a maximum of 891. Votes are distributed very unequally, with both the minimum and the mode at 2 votes.

The vast majority of submissions on Digg contain less than 150 words. Out of 10,187 submissions, 70.4% have less than 150 words. However, these stories received an average of 12 diggs, whereas the stories above 150 words received an average of 29.7 diggs. The longest 29.6% of stories received 49.8% of the total votes. Votes seem to be evenly distributed among groups of stories with 150 words or more. That is, stories with 150 words appear to be as likely to succeed as those with 1,000. Sentence length did not

Figure 1. Average diggs by the average word count and sentence length.



show a strong correlation with votes. Users seem to favor stories with 11 to 17 words per sentence. However, votes are much more evenly distributed among stories with different sentence lengths than compared to word count.

Coloring the graphs by topic showed clusters of topics and indicated differences between the average difficulty of each topic. These clusters seemed to show genre expectations for the topic. To see how conformity to these trends affected submissions' popularity, I graphed the average article length for each topic and calculated the distance of each story from the average for that topic. Since different topics have significantly different average votes, I did the same for the vote count—calculating stories' distance from the average vote of their topic. For article length the mean was between 100 and 500 words for each topic, although there were vast differences from the mean, with some stories as high as several thousand words. I grouped stories by their difference from the average to the nearest 100 words. My sample had fewer than 100 stories for each grouping above 500, above which the data becomes increasingly noisy.

Comparing the distance from these two averages shows how genre conformity affects popularity, as shown in Figure 2. Looking at all of the stories shows that users prefer stories *above* the topic's average. Stories that exceed their genre's mean article length receive a higher average number of votes, whereas stories shorter than their genre's mean receive fewer than the average votes. Stories that exceed the average word count by 500 to 700 words receive roughly double the average votes. Most individual topics correspond with this overall trend. Interestingly, two topics show the opposite trend: Sports and Offbeat both lose votes as they become more difficult.

Genre conformity for the average paragraph length, sentence length and word length also showed similar results, although not as pronounced as with article length. As shown in both Figure 2 and Figure 3, the trend line crosses the mean vote count just negative of the average for each content metric. Stories directly at content conformity receive one to three votes higher than average. Stories with an average sentence length more than two words below the mean all consistently receive fewer than the average votes. Stories above the mean consistently receive a higher average vote count (with the exception of five words above the mean), but only by one to ten votes.

Figure 2. Distance from the average word count for each topic compared to the distance from average vote count for each topic.

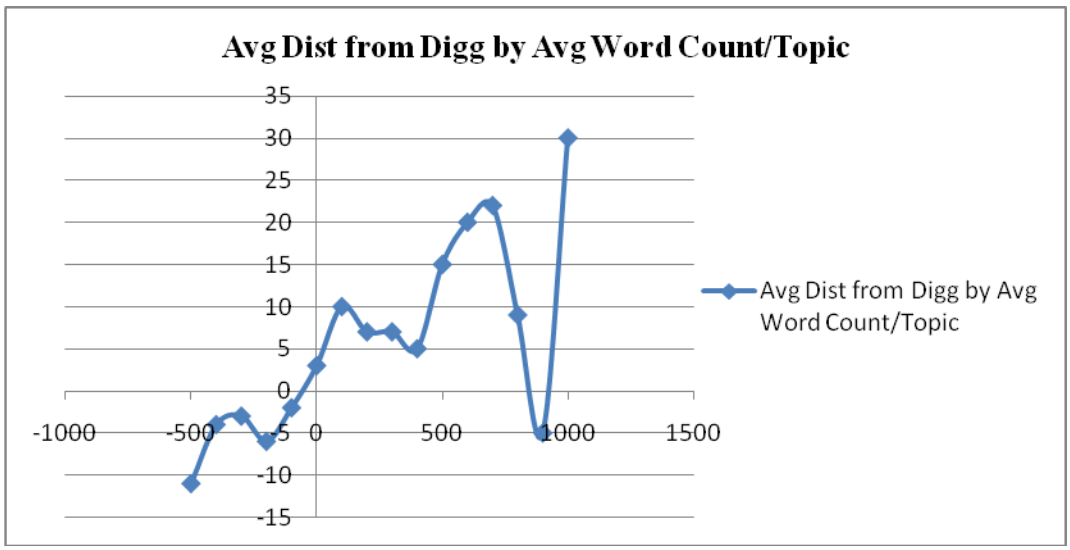
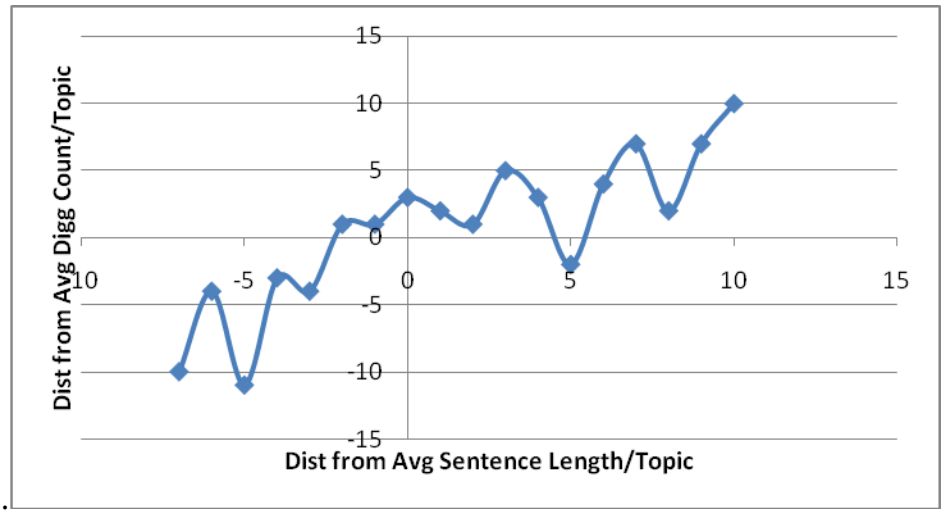


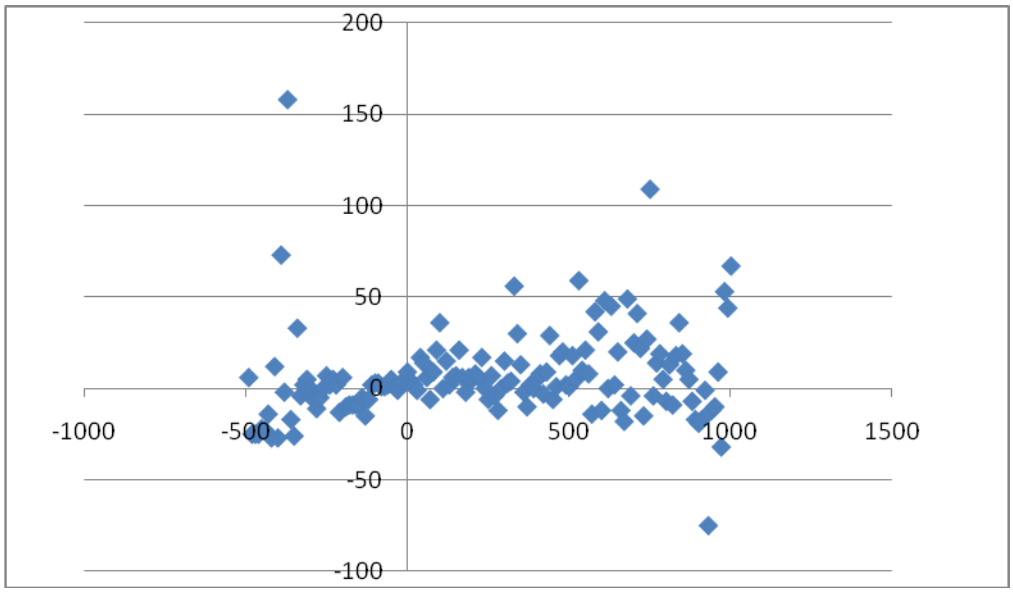
Figure 3. Distance from the average sentence length for each topic compared to the distance from average vote count for each topic.



Genre conformity shows the strongest correlation between content and popularity that has been published about Digg, but despite this trend, the data still indicates that content does not entirely account for success. A scatter chart of these stories by their distance from the mean, but grouped to a more specific number—to the nearest ten—shows the enormous variability between these content metrics and stories' final votes (Figure 4). There are large differences between stories' popularity, but this popularity does not seem to correspond with large differences in stories' content features. With a standard deviation of 47 votes, there was no difference in content that corresponded with even half of one standard deviation. Two learning algorithms from WEKA (Baynesian and J48 Decision Tree) consistently identified the story submitter's follower count as the

most influential single metric for a story's success.[2] Clearly, there is more to popularity than content.

Figure 4. Article length grouped to the nearest 10 by distance from the mean. Groups close to the average contain many more stories, which is why they cluster so consistently. Groups further from the average have much fewer stories and show the extreme variability within the sample.



The Case Against Content

Prior studies have focused exclusively on the social dynamics for predicting popularity on Digg. This trend was at least partially influenced by a sociological paper published in *Science* in 2006, which argued that predicting popularity based on content is fundamentally impossible. This study, "Inequality and Unpredictability in an Artificial Cultural Market," tried to create replica online social "worlds" to find out if success is

[2] WEKA, or the Waikato Environment for Knowledge Acquisition, is a popular open source program for data mining.

deterministic or random. They began the study with the hypothesis that if the participants' choices were determined by the content of the music then the same songs would repeatedly become successful with a similar audience.

The site was large enough to recruit over 14,000 participants, enough to create eight such artificial worlds. They took a control group and had the group rate songs from one to five stars to determine the quality of each song. The research team assumed that users from the same site would have similar tastes in music and randomly selecting roughly 1,800 participants for each world would homogenize the overall preferences across all of the different worlds. In seven of the eight worlds, the team introduced social influence by showing participants the number of times that other participants had downloaded each song. The research team wanted to see if the same set of songs became successful in each world. The experiment was designed to "measure inherent unpredictability: the extent to which two worlds with identical songs, identical initial conditions, and indistinguishable populations generate different outcomes" (Salganik, Dodds and Watts 854). If the user's music choices were subject to, or determined by, the content, then each world should have ended up with the same popular songs. However, the worlds ended up with almost nothing in common as each world created vastly different popularity charts.

The team also found that enlarging the display that showed the download count (to increase social influence) widened the gap between the lowest and highest downloads. Social influence made success more unpredictable and also more unequal. The more prominently the website presented the download count, the greater the inequality between

songs. Social influence made the rich get richer, and the poor get poorer. The team concluded that "…when individual decisions are subject to social influence, markets do not simply aggregate pre-existing individual preferences. In such a world, there are inherent limits on the predictability of outcomes, irrespective of how much skill or information one has" (Salganik, Dodds and Watts 856).

In the control group, participants individually rated each song with a score of one to five, unaware of how other people had scored them, to make this ranking outside of the social influence. The research team defined these scores as the songs' quality. Salganik's team concluded that "Success was also only partly determined by quality: The best songs rarely did poorly, and the worst rarely did well, but any other result was possible" (854). While song quality did not determine success, it did create limits for popularity.

Since the songs failed to become repeatedly popular, the implication is that popularity is at least partially random, and prediction becomes problematic. The question of popularity goes beyond Digg; predicting success is a time-honored pursuit of capitalism. Outside of purely artistic markets, there are entire industries devoted to predicting the success of advertisements and marketing campaigns. Despite the money invested into these predictions, it's obvious that they often predict incorrectly. The "Cultural Market" study addressed the gap in cultural markets between the "winners" and "losers." If quality determines popularity and the difference in popularity between stories is so unequal, then the difference in quality should be apparent. And if the difference in quality is so apparent, then predicting success should be easy. What is it that

"…accounts for both inequality and unpredictability" (Salganik, Dodds and Watts 854)? Salganik et al. answer this contradiction by arguing that "individuals do not make decisions independently, but rather are influenced by the behavior of others" (854).

On the Digg website, other peoples' decisions are prominently displayed, since the entire concept of the site is to emphasize the behavior of other users. The distribution of downloads in the "Cultural Market" study is similar to the distribution of votes on Digg. They argued that social influence not only makes popularity unpredictable, but also increases the inequality between the most and least popular. Both social influence and the design of the Digg website increase the disparity between successful and unsuccessful stories. On Digg, most stories never receive a single vote outside of the person who submitted it, while popular stories can potentially accumulate thousands of votes. However, it is this very inequality that allowed previous studies to accurately predict popularity on the site. Previous studies have predicted the popularity of stories on Digg through observation of the development of this unequal distribution. They measured the digg count after four or eight hours, and because this inequality becomes prominent after only a few hours they were able to successfully predict stories' final votes 48 or 72 hours later.

The architecture of the Digg website largely structures the number of users who see a particular story and this medium has an enormous impact on the distribution of votes. There are significant differences between the percentages of users that vote on a particular story, but these differences can be overshadowed as relatively small, unpredictable variations are magnified by the website's design. Digg creates a "rich-get-

richer" situation where stories that initially obtain moderately more diggs receive an enormous increase in exposure. The stories that get moderately more diggs on the Upcoming/Recent page are promoted to Upcoming/Trending. Once on this page, their visibility quickly increases: rather than remaining on the page for an average of fifteen minutes, they stay on this page for a few hours. On the Upcoming/Trending page, each digg not only further separates it from similar stories, but also spreads the story to more friends, compounding what was initially a small difference in votes between two stories. Of these Upcoming/Trending stories, the ones that do moderately better than the others are promoted to the Top News section where stories routinely get thousands of views per hour. Through this rich-get-richer design, the architecture and social dynamics structure visibility to allow a story that initially received only moderately more diggs to potentially accumulate several hundred times more votes than stories with a comparable interestingness. Social influence and Digg's design both magnify the disparity between stories, potentially exaggerating the difference between two stories that users would otherwise consider to be of equal quality.

What the "Cultural Market" study means for Digg is that quality is not (entirely) dependent on content but is created through a dynamic social process, and differences in the decisions of just a few users could generate to a substantially different result. If the same content with the same audience can lead to vastly different potential worlds, then on Digg the same content could end up with very different results. To make sure that the same content is not submitted more than once, Digg checks the URL of each submission to ensure that it has not already been posted. However, this mechanism sometimes is not

enough to prevent the same content from being repeatedly submitted. Interestingly, identical content rarely ends up with the same results.

On February 25, the most popular submission of the week was an interview with a member of the radical Westboro Baptist Church and a spokesperson from the hacktavist group known as Anonymous. Both groups had received recent media attention and made the broadcast an appealing interview with the Digg audience. The submission received over 900 votes in the first 24 hours, making it one of the most viral stories of the year. What was interesting about the interview was that the same video was submitted at least fifteen times. The interview was posted and re-posted by dozens of people on YouTube and put onto several different websites, making the same video accessible through different URL addresses. Of the fifteen submissions, seven of them only received a single vote. Those were mostly from users with very few friends, so it is possible that was because so few people saw it. The submission that went viral was submitted by a user with 1,124 followers, which seems like an easy explanation for why his was the most popular. However, the same video was posted by a user with 3,228 followers and a day later had only received 66 votes—not even enough to be promoted to the Top News section.

Content as a Limit

The studies on Digg consistently agree that popularity is a fickle thing: small changes in a user's behavior affect whether a story becomes viral and these random variables make popularity erratic. Salganik et al. write that predictive models of

popularity fail to predict a single outcome, but instead only narrow down the range of possibilities (854). "Models of collective decisions that incorporate social influence can exhibit extreme variation both within and across realizations even for objects of identical quality" (Salganik, Dodds and Watts 854). This means that predictive models return multiple possible outcomes. Gauging from the results of their study, this is not a failure of the model, but a reflection of the reality that popularity is clearly non-deterministic, since small variations in one user's activity could radically alter the system. In this way, success or failure is partially random, so many different results can occur. Prior research emphasizes the importance of social influence and Digg's interface as a medium for structuring popularity. This visibility is undeniably important, but privileging medium and social influence can ignore the fact that if no one ever votes for a story then it will never become popular, regardless of how many people see it. For example, a blank page would never make Top News. Companies often spam Digg by posting advertisements, but these are quickly buried. Saying that social influence largely controls popularity is not to say that content is entirely irrelevant. Salganik's study is important because the results showed that there were many possible outcomes, but *not every outcome was possible*, because popularity was constrained within limits defined by the content.

User Behavior

If success on Digg is so volatile and small differences make success incredibly erratic, then how have other studies been able to predict successful stories? Research on Digg's popularity has found that predicting popularity with content is challenging, but it

is possible by measuring social patterns. It is difficult to predict which story will become popular, but it is certain that some stories will—so rather than focusing on content, these studies have focused on the *process* of popularity. Gabor Szabo and Bernardo A. Huberman studied submissions' popularity growth over time, finding that early and later popularity strongly correlate (83). Measuring the growth of votes in the first two hours after a story's submission, their model was able to predict a submission's success 30 days later with a relative error of 10% (Szabo and Huberman 88). They applied their model to both Digg and YouTube, but it was much more successful on Digg because of the comparative intensity of Digg: stories receive almost all of their votes on the first day of their submission and after 72 hours they have accumulated over 99% of their total votes. In contrast, YouTube videos still receive votes thirty days later. Their model is so successful because viral stories follow a particular pattern of growth, so looking at stories shortly after submission is enough to recognize their momentum. Digg's architecture accentuates popularity, which makes these stories recognizable, since popular stories distinctly separate themselves shortly after submission.

The Szabo and Huberman study clearly shows that there is a larger structure to popularity on Digg, but I wanted to find if there was a structure to users' individual decisions. The Szabo and Huberman algorithm was a macroscopic model that treated Digg as a single entity, and the rapid decline of attention from the site's users as a property of this entity. To contrast this model, a different study by Kristina Lerman and Tad Hogg used a micro model, which incorporated the decisions of individual users. Rather than treating the site as a single entity, they looked at the percentage of users that

voted for each story. They used this percentage, and an estimate of how many people would see the story, to predict its eventual popularity.

Social dynamics and the site's interface structure affect popularity by determining how many people see a story, but the Lerman and Hogg model shows the percentage of people that liked it enough to vote for that story. Content only loosely determines popularity, but does content determine this percentage? Does this percentage provide a clearer image of what constitutes these limits?

Modeling Digg

I used Kristina Lerman and Tad Hogg's model of user behavior to estimate the number of people that see a story.[3] Their model defines a story's votes as the product of two variables: visibility and interestingness. Visibility is the number of people who have seen a story and interestingness is the percentage that voted for the story. The model sets the rate of change of votes as equal to *v(t)*, the visibility of the story at that time, multiplied by *r*, the interestingness:

$$\frac{dN_{votes}(t)}{d(t)} = rv(t)$$

where $v(t) = v_{lists}(t) + v_{friends}(t)$, and $v_{lists}(t)$ is the sum of users finding the story through various sections of the site. The model needed to be updated to incorporate new additions to the site. In 2009, when Lerman and Hogg collected their sample, Digg had

[3] I provide an overview of the model, because it is a relatively obscure algorithm and to show what had to be modified in the equation to keep pace with the updated version of Digg. For a more detailed description of this model, see Kristina Lerman and Tad Hogg. "Using Stochastic Models to Describe and Predict User-Behavior."

two lists, upcoming and topnews. Currently, within these two sections there are multiple lists (Upcoming/Recent, Upcoming/Trending, etc). A current model needs to incorporate these current sections. Their model used similar rate equations for both of these lists, which I combined into one equation that incorporates the current sections and can also be used for each of the seven current lists; $v_{lists}(t)$ is the sum of this equation solved for each list:

$$v_{lists}(t) = \sum_{j=list}^{n} c_j v f_{page}(p(t))\theta(N_{vote}(t) - h_j)\theta\left(h_{j+1} - N_{vote}(t)\right)\theta(P_j - p(t))$$

The first step function estimates how many users will see a story out of the total number of users visiting the site: v is the rate that all users that go to Digg (600 per hour) and c_j is the fraction of users who proceed to section j. Of these users, f_{page} represents the fraction of users that visit the page on which the story is listed. Most users only look at the first page, with the stories listed at the top receiving the highest visibility. The position of a story is represented by $p(t)$, with "position 1 representing the top of the first page, and 1.5 being halfway down the first page" (Lerman and Hogg 7).

Since each story is usually only shown in one list at a time, the instantaneous rate of change is only dependent on single list. The step functions in the $v_{list}(t)$ rate equation ensure that only current lists are counted towards the sum. Each step function is 0 when it solves to $x < 0$, and the entire rate equation is multiplied by 0. If the step function solves to $x \geq 0$, then the entire equation is multiplied by 1. The last step function accounts for the maximum pages in each list.[4] Digg only maintains a finite number of

[4] Lerman and Hogg used a twenty four hour maximum time for this function for the upcoming section; however, Digg currently maintains the number of past stories based on volume rather than time.

pages for each section, so the position of the story, $p(t)$ is subtracted from the maximum page number for each section, P_j.

When a story is promoted to the next section, it disappears from the previous section. The middle step functions ensure that the equation is only evaluated for the times it was on list j. h_j represents the number of votes a story had when it was promoted to the current list and $h_j + 1$ represents the number of votes the story has when it is promoted to the successive list. I use $j + 1$ to represent the successive list, although it could be misleading in that it indicates that stories are promoted through the lists in a specific order. While most stories are promoted from the Upcoming/Recent to Upcoming/Trending to Topnews/Trending, there is no evidence that they might not skip a list—jumping straight from the Upcoming/Recent page to the Topnews/Trending page. To account for this, Diggbot recorded the count of votes, comments and page views for each new promotion when it scraped the site every five minutes. Stories are always promoted to Upcoming/Recent when they get two votes, so $h_r = 2$, but aside from upcoming/recent, h_j and h_{j+1} are independent variables for each story, with $j + 1$ representing the successive promotion, whichever list that may be.

Out of the users that go to section j, the fraction that see a story is based on two factors, the location of the story and the fraction of users that scroll to that location. The location of a story is determined by the rate of submissions multiplied by the time since the story's promotion to that page, so $p(t) = k_j(t - t(h)) + 1$, where k is the rate of promotions and $t(h)$ is the time the story was promoted. It is always plus one, because stories begin on page one, not page zero.

The fraction of users that see a story is the fraction that travel to the position of the story, *p(t)* on page *m*. The fraction of users that visit *m* pages is based on an inverse Gaussian distribution:

$$e^{-\frac{\lambda(m-\mu)^2}{2m\mu^2}}\sqrt{\frac{\lambda}{2\pi m^3}}$$

The area of the distribution from *m* to ∞ is the number of users that scroll to page *m* and further. When $m = 1$, then $f_{page}(1) = 1$, since all users see the first page. When $m > 1$, the fraction of the distribution is

$$f_{page}(m) = \frac{1}{2}\left(F_m(-\mu) - e^{\frac{2\lambda}{\mu}} F_m(\mu)\right)$$

where $F_m(x) = \mathrm{erfc}(\alpha_m(m - 1 + x)/\mu)$, and $\alpha_m = \sqrt{\lambda/(2(m-1))}$. Erfc($x$) is the complementary error function,

$$erfc(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}\,dt$$

To find $v(t)$, we also need to find the visibility through the MyNews section, $v_{friends}(t)$. In the first few hours after submission stories are mostly discovered through the social network. When users vote for, or comment on a story, their followers see that story in their MyNews page. Estimating this visibility must take account for the submitter's friends, as well as the friends of the people that vote for the story. Users have an average of 51 followers, but these followers begin overlapping through the social network, so even if friends were evenly distributed, each vote would still bring fewer followers because some of the new followers have already seen it from other friends. The number of friends that have not yet seen the story is initially set to the number of

friends of the submitter, as provided by the Digg API. After that, the growth in the number of friends is $\Delta s = 51N_{vote}^{-0.62}$, where *s(t)* is the "number of fans who have not yet seen the story" by time *t* (Lerman and Hogg 7). This number grows with each vote, but also decreases as these fans visit the site. The rate that fans visit the site, *w*, is 0.12 per hour, so *s* is decreasing at a rate of –*w*. The total rate of change in friends is this decrease, plus the number of incoming friends from new votes,

$$\frac{ds}{dt} = -ws + 51N_{vote}^{-0.62}\frac{dNvote}{dt}$$

The rate equation is identical for each page, but uses different parameters for the rate of submissions and the fraction of users that visit that page. The rates are defined in "Digg time," which is the number of votes on the site between two events. In my sample, there was an average of 1509 votes per hour.

Story Interestingness

In some ways, comparing popularity and content does not address a question of subjectivity. The social network and Digg interface never allow submissions a level playing field. Someone with 10,000 friends can submit a story that only 1% of their friends digg, and then that story is promoted, thousands of other people see it. Someone with 30 friends might have 50% of their friends digg a submission, but in terms of total diggs, it is the comparative loser. This disparity between story visibilities is partially the reason that content-based predictions are so problematic. It means that comparing votes and content does not fully address a question of subjectivity, but rather a question of the social network and site architecture.

What if submissions all received equal visibility? If the exact same number of people saw each submission, then a submission's popularity would be an aggregate of a consistent number of individual decisions. Users' votes could be completely random and still form non-random distributions through the inequalities in visibility. Would this distribution of votes still be non-random? Since visibility is rarely equal on Digg, this distribution unfortunately remains hypothetical. However, as an alternative, we can look at votes as a percentage of visibility. Examining votes as a percentage begins to account for the vast disparities created through the social network and site structure. A series of earlier studies, by Lerman and Hogg, developed a model to estimate the number of users that see a story, and with that estimate, they calculated the percentage of users who vote on a submission out of the users who see that submission. Lerman and Hogg used this percentage as part of their equation to predict popularity, but I used it to ask a different question that more directly addresses an issue of subjectivity: rather than predicting total votes, can interestingness itself be predicted?

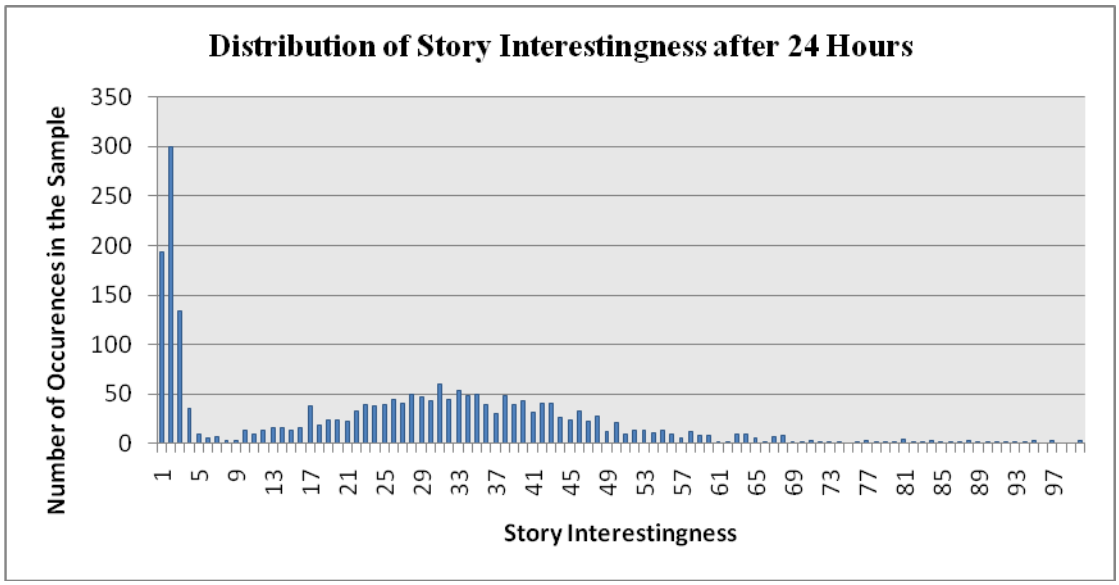
To examine interestingness more closely, I also began tracking view counts to distinguish between title interestingness and story interestingness. In February of 2011, Digg began showing the view count for each submission, which is the sum of exit clicks from Digg and the permalink count for the story. Recording these view counts creates a way to more specifically examine user decisions, since the interestingness ratio combines two different decisions in most users' behavior: users see a title and decide to click on it or not, then they decide to digg, not digg, or bury it. Interestingness clumps these decisions together, but tracking view counts distinguishes these decisions, so we can see

the interestingness of the title/description (the users who click on a title out of the number of users who see that title) as well as the interestingness of the story (the number of users who digg a story out of the users who clicked on it). View counts were not accessible through the Digg API, but scraping the site in five minute intervals provided Diggbot near real-time access to changes in the number of views. Digg only maintains a limited number of pages in each section, which limited access to changes in view counts to the first twenty four hours for most stories, although highly successful stories obviously stay in the Top News/Week and Top News/30 Days sections for much longer.

Lerman and Hogg found that interestingness followed a lognormal distribution (10). Story interestingness is also distributed lognormally. Twenty four hours after submission, submissions have separated into two main groups, promoted and unpromoted stories. Although it seems counter-intuitive, promoted stories actually have a lower story interestingness and unpromoted have a higher story interestingness. After twenty four hours, most promoted stories have a story interestingness of less than 5%. In contrast, most unpromoted stories have a story interestingness between 30-50% after twenty four hours, with many stories actually exceeding 100%, since users occasionally digg stories without clicking on them. This occurs before a story is promoted when some followers are simply trying to help their friend. It also seems to occur with a breaking news headline. For example, several stories on President Mubarak's resignation from the Egyptian Presidency exceeded a story interestingness of 100%. After twenty four hours, one story had a story interestingness as high as 650%, and after eight hours, one story had

a story interestingness of 1,100%. However, when stories reach the Top News page they receive many more views, but typically only a very small ratio of 0.02 votes per view.

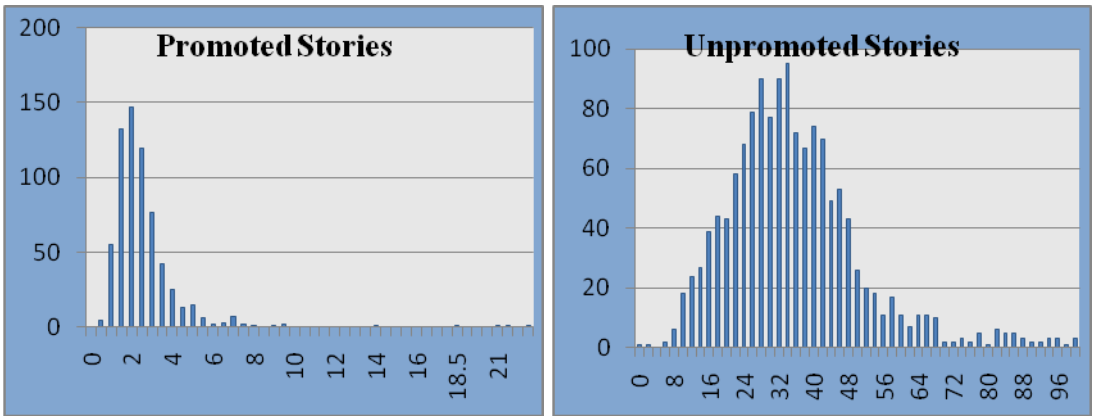
Figure 4. Story interestingness compared with the number of samples at that interestingness.



This distribution is the overlay of the voting patterns of two underlying populations: followers and non-followers. Szabo and Huberman also found that voting is strongly influenced by the social network: shortly after submission, users are three times more likely to vote on their friends' stories (87). However, once stories are promoted to Top News, then they are no more likely to vote on their friends' stories than any other submission. Lerman and Hogg also found that followers are much more likely to vote for a story than non-followers (12). Since unpromoted stories are mostly seen by followers, then story interestingness is mostly composed of by followers for unpromoted submissions. Promoted submissions are seen almost entirely by non-followers. Lerman and Hogg found results for interestingness for the followers population to be much higher than interestingness for non-followers, with modes at 20% and 2%, respectively (12).

This distribution does not perfectly match theirs, likely because promoted and unpromoted stories are voted on primarily, but not exclusively, by non-friends and friends, respectively. The addition of the upcoming/trending page also adds a new intermediate stage, where users are more likely to be exposed to submissions of non-followers. Most importantly, this is the distribution of story interestingness, as opposed to overall interestingness.

Figure 5. Distribution of story interestingness twenty four hours after submission.

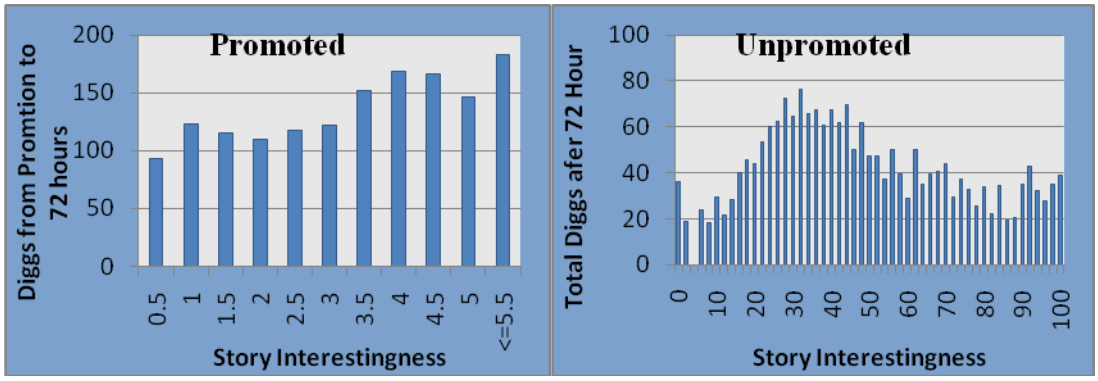


Initially, I expected that there would be stronger correlations between content and interestingness than between content and popularity. However, textual metrics did not indicate any greater correlations than from looking at total votes. Users do not seem to show a preference towards more complex or more simplistic language.

In the overall sample, story interestingness does not correlate with a higher number of votes, since promoted stories end up with more votes from non-followers. However, within the subsample of promoted stories, story interestingness does correlate with a higher number of votes. However, within the subsample of unpromoted stories, the stories with submissions at 30% story interestingness receive an average of over 60

votes. This may be because these are stories that were selected for the Upcoming/Trending page and did not make it to Top News, but still would have been exposed to a higher number of non-followers.

Figure 6. Story interestingness by diggs.



Stories are typically promoted to Top News with at least 60 votes, but some are not promoted until as many as 190 votes, probably because the Digg promotion algorithm requires more votes from a story submitted from a user with more followers. To account for this, Figure 6 shows story interestingness compared to the growth in votes from the time of promotion to 72 hours after submission. At least for promoted submissions, story interestingness positively correlates with popularity.

The sample contains only 671 promoted stories, so such a small sample size might not indicate weak correlations between content and story interestingness. The one strong correlation, however, was between story interestingness and pictures. It is a common assumption on Digg that pictures are a guaranteed way for a story to get more votes. However, the effect of pictures on interestingness is much more complex than that. Prior to promotion, it is unclear what effect pictures have on interestingness. They may correlate with a higher story interestingness, but the distribution seems erratic enough

that there may be no relationship. After promotion, pictures correlate negatively with story interestingness.

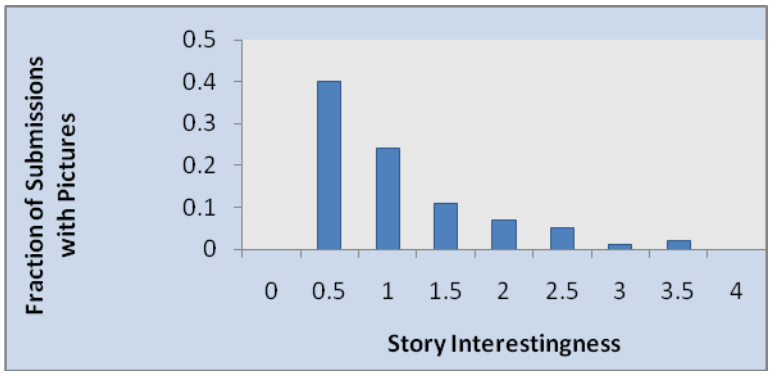
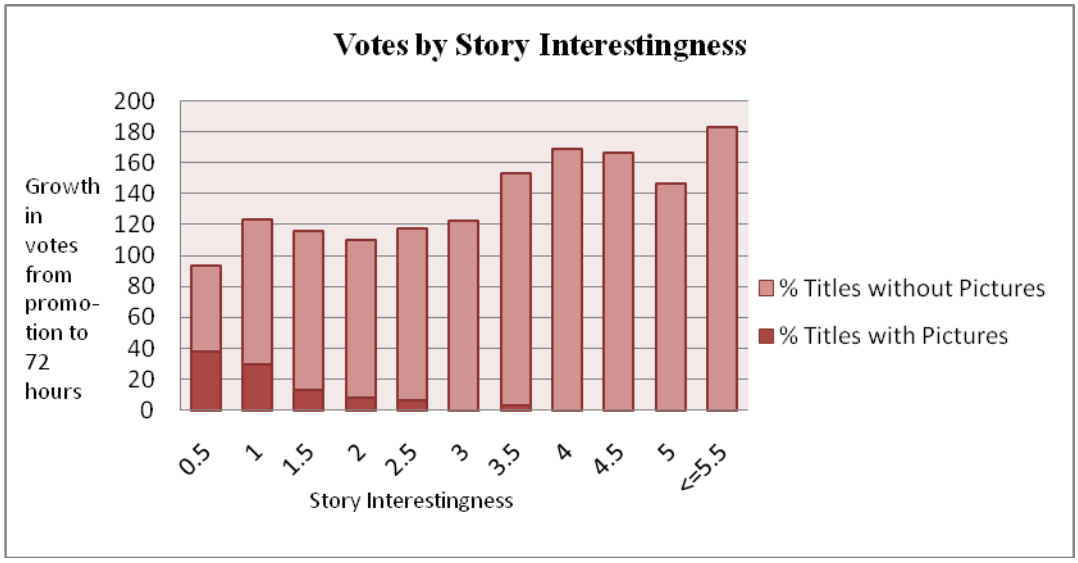


Figure 7. Fraction of promoted stories with pictures by story interestingness.

Figure 8. The growth in votes in promoted stories from the time of their promotion to 72 hours. Story interestingness compared with the growth in votes for promoted stories. Story interestingness correlates positively with the growth in votes.



In the sample of promoted stories, not one submission with a picture had above an story interestingness of 3.5%. However, stories with pictures received an almost identical number of average votes, 160.79, as stories without pictures, 160.77. If their story interestingness was so low, then how did they receive an equal number of votes? While the story interestingness was lower for submissions with pictures, the title

interestingness was actually significantly higher among submissions with "pic" in the title.[5] The stories with "pic" in their titles received an average of 10,000 views, compared with 7,000 for titles without "pic." It seems that Digg users want to see submissions with pictures, but do not seem to vote for them.

Title Interestingness

To study title interestingness, I used a blog post written by a popular social media marketer. Neal Rodriguez has made a name for himself on Digg as a social media marketing expert and his formula was simple enough that it was easy to test. He wrote that popular submissions on Digg need a title with his formula:

> "The" + (Number) + "Most" + (Over the top adjective) + (Subject) + Of All Time (Synonyms like "in History" or "Ever" will also be accepted) = Popularity (Rodriguez)

Rodriguez provided the terms that were specific enough to test. In terms of title interestingness, he was correct for most of this. I checked each title for each portion of his formula, and for the formula as a whole. In the entire sample, there were no titles that specifically followed the whole formula exactly. Titles with pictures received an average of 43% more views, or 10,454 per submission, than titles without pictures, which received an average of 7,323 views per story. There were very few stories that had an exaggeration in them as specific as Rodriguez suggests ("Of All Time," "Most", "in History," "Ever"), but the 34 titles that did have one of these four exaggerations received

[5] My regular expression identified any case insensitive, singular or plural version of pic, picture, photo, or photograph. "pic" was the most common.

15% more views than titles without. The other suggestion, to include a number in the title, seemed to have no difference in views.

Predicting Interestingness

Like popularity, interestingness seems to be difficult to predict based on content as well. There do not seem to be clear indicators of interestingness either for titles or stories. Story and title interestingness are not evenly distributed, but like popularity as a whole, even when these distributions do correlate with content, it is not enough to explain the vast differences between best and the worst. Digg is organized around the principle of social influence, and as the "Cultural Market" study showed, it seems that not only is the total popularity affected by social dynamics, but interestingness as well. Earlier studies on Digg are able to predict final votes so accurately because the number of votes a story receives in the first few hours after submission correlate strongly with the total number of votes it will receive. Although I have not tested this, it appears that view counts could be predicted in the same way, by measuring voters' early reactions to content.

What the "Cultural Market" study seems to indicate is that the interestingness is not determined solely by content. One answer for why the early-reaction models of other studies work so well is that the algorithm simply waits for the humans to gauge the more semantic aspects of the content that are beyond the computational models. This is the answer that underlies the justification for crowdsourced, collaborative rating sites. However, it does not explain the results of the "Cultural Market" study, and it does not

explain how identical content can be submitted to thousands of Digg users and receive vastly different reactions. Part of the answer to different reactions was different visibilities between the stories, but narrowing in on interestingness and then title and story interestingness shows that even when visibility is accounted for, there are still vast differences between submissions. And it also shows that content does not fully account for these differences. Perhaps the real reason that this early-reaction method of prediction works so well is that interestingness itself is constituted by those early votes. The distribution of users' individual votes "do not simply aggregate pre-existing individual preferences," but rather, are constantly constructing those preferences (Salganik, Watts and Dodds 856). In other words, the digg count that reads next to each title is as much a part of the content as the article itself.

But the article is not irrelevant. The results showed that there are correlations between content and interestingness. However, distinguishing visibility and interestingness in the way that the model does threatens to ignore the fact that content quality is itself a product of social construction. My results show that content creates broad trends on the site, but of course these trends are not intrinsic in any way. Just as voting decisions seem to be a product of social influence, these broader trends certainly change as well. Hypothetically, if Digg existed in a historically different period, the broad preferences towards article length and sentence length would certainly be different. These processes are happening at very different speeds: preferences on individual stories are being generated with each vote, while shifts in the broad preferences of content metrics (sentence length, article length, etc.) are unfolding at a very different rate. And

there are certainly intermediate shifts occurring as well. While a story has a lifespan of 48 hours, a semantic event often stays around longer. For example, with the interview between Anonymous and the Westboro Baptist Church, users posted submissions of the video itself and other articles about the video for several days afterwards.

Part of what my study means to the humanities is a caution against arguments of causality. Scholars often argue that a film or book was popular because it symbolized an emotional undercurrent of the period (e.g. *King Kong* represented racial anxieties or zombie films represent postmodern frustration), but I think that the studies of Digg show that there is not a direct link between content and popularity. It is not a question of whether it was a symbol or was not. The issue is that in terms of popularity, representation might not matter. I want to be cautious in applying the mechanisms of online popularity to offline popularity, but the impact of social influence seems quite clear. As Salganik et al. note, social influence is probably heightened offline than compared to their study (857). Offline, this influence is occurring through a very different medium, but it is there, nonetheless.

Reading Practices on the Web

What my study also means to the humanities is that scholars may have misjudged reading practices on the Web. There is a widespread tendency in academia to disdain online reading practices as favoring the easiest, most consumable reading. My study shows that reading practices on the Web are more complex than that. While users do not prefer articles with long sentences or long words, they do not necessarily seem to

discriminate against them either. In fact, users seem to prefer textual articles to non-textual submissions, since users do not digg stories with pictures as often as they digg stories without. Probably the most distinct finding was that out of the promoted stories, not one with a picture had a story interestingness above 3.5%. The site's users clearly do not discriminate against more textual submissions.

Of course, my results can show nothing about whether the users that dugg an article ever actually read that article. In fact, story interestingness shows that there are some users who do not even click on a title before digging it, but this occurs almost entirely among unpromoted stories that are seen mostly by followers of the story's submitter, who are voting simply to help their friend. Story interestingness also shows that the vast majority of users do click on a title before voting for it, and since Digg users show a proclivity to debate and comment on articles, then presumably, they have read what they are arguing about.

Digg is only one site, and it would be a mistake to over-generalize the tendencies on Digg to the rest of the web. However, according to Alexa's demographics of the site, the four million Digg users are mostly between the ages of 13 and 35—precisely the population that is accused of cursory reading practices because they read predominantly on the web. While my results cannot show whether these users ever actually read the stories they voted for, they do show that users are more inclined to vote for stories with 150 words or more. It seems that textuality is not entirely dead on the internet.

Works Cited

Jamali, Salman. "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis." Thesis. George Mason University, 2009. May 2009. Web. 22 Mar. 2011.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. “The WEKA Data Mining Software: An Update.” *SIGKDD Explorations* 11.1 (2011). Print.

Lerman, Kristina, and Tad Hogg. "Using Stochastic Models to Describe and Predict Social Dynamics of Web Users." (2010). Web. <http://arxiv.org/PS_cache/arxiv/pdf/1010/1010.0237v1.pdf>.

Lerman, Kristina, and Rumi Ghosh. "Information Contagion." (2010). Web. <http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.2664v1.pdf>.

Rodriguez, Neal. "4 Ways I Compose Posts to Drive Millions of Pageviews to Blogs Through Digg."*Blog Tips to Help You Make Money Blogging - ProBlogger*. Web. 01 Feb. 2011.

Rose, Kevin. "Digg Friends." *Digg Blog*. Sept. 2006. Web. 31 Mar. 2011.

Salganik, M. J., Peter S. Dodds, and Duncan J. Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311.5762 (2006): 854-56. Print.

Schott, Ben. “Digg Patriots.” *New York Times*. August 13, 2010. Web.

Szabo, Gabor, and Bernardo A. Huberman. "Predicting the Popularity of Online Content." *Communications of the ACM* 53.2 (2010): 80-88. Print.

Toutanova, Kristina and Christopher D. Manning. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger.” Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000). Print.