A Measurement of the Social Media Impulse Response Function

David Hardtke, Ph.D.
http://www.stinkyteddy.com
david@stinkyteddy.com

February 12, 2010

1 The Inadvertent Launch

For several months prior October 19, 2009, we had been developing a novel search engine called Stinky Teddy. The idea behind Stinky Teddy is to utilize real-time data (Twitter posts, blogs, blog comments, and other types of user-generated content) to contextualize words and phrases that might be used as part of a web search query. Utilizing this enhanced understanding of the way in which people are using certain words and phrases at a particular moment in time, we built a search engine where the content presented adjusts dynamically to the "collective consciousness" of the web. Our hypothesis is that many search queries are motivated by a universal shared experience, and that we can anticipate the information goal of the search engine query by "listening" to the "gossip" and "chatter" on the real-time web.

The details of Stinky Teddy, however, are rather unimportant for this study. All that is important to know is that a search engine named Stinky Teddy located at http://www.stinkyteddy.com existed on October 19, 2009, and that this search engine was publicly available for use (no password protection or "private beta") but was not being used. The creators had not yet publicized the search engine in any way except by giving a short demo at the ReadWriteWeb Real-Time Summit on October 15, 2009. The search engine was still in the process of development and was not going to be released for several months. The user interface was very simple and standard for search engines (search box in the middle of a mostly empty page).

On October 19, 2009, at 10:51 AM Pacific Daylight Time, something unexpected happened. Suddenly, and without warning, people started using Stinky Teddy. At that moment ReadWriteWeb posted a somewhat favorable story about Stinky Teddy on their blog [1]. This story was not pitched by the company, but instead resulted from conversations at the Real-Time Summit the previous week. If we had had advanced warning of story, we would have asked for a delay in order to fix several known bugs. My first response to the story was, "I can't believe they misspelled my name", followed by "I can't believe they posted a story without warning me", followed by "This is cool, let me email my friends and family", followed by "hey, wait a minute, I'm not going to do anything at all because I have a unique opportunity to measure the impulse response function of the social media ecosystem."

[1] "Stinky Teddy: A Cool Real-Time Search Engine with a Rather Odd Name," Frederic Lardinois, http://www.readwriteweb.com/archives/stinky_teddy_realtime_search_review.php

2 What is an Impulse Response Function?

Suppose that we have some system that takes an input signal and produces an output signal. The impulse response function is simply the output of system when driven by a short input pulse. An example would be the sound that comes out of the public address system when the rock concert roadie says “check” into the microphone. In this case the output sound signal is amplified but has the same shape as the input signal. Another example would be the sound emanating from a large bell when struck with a hammer. In this case, the response is very long lived as the bell can ring for minutes and bears no resemblance to the input pulse shape (the whack of a hammer). Although systems can be complicated, the impulse response function is a simple way to characterize the aggregate behavior of the system.

What does this have to do with an isolated blog post in ReadWriteWeb? Since this article in ReadWriteWeb was completely unique (i.e. was not part of a concerted public relations push), all traffic coming to stinkyteddy.com after the publication of this article can be directly or indirectly attributed to this single blog post. As mentioned above, the response function characterizes the relationship between a short input pulse and the output signal. This single article was the short input pulse, and the traffic arriving at stinkyteddy.com was the output signal. Hence, we can measure the response function of the system. The system under study in this case is what one might call the “social media ecosystem”. It is the web-blog-social media network environment through which much information is now distributed. It is a system of original content producers(few), re-distributors and plagiarizers (lots), sharers(more), and readers(many). People have studied the link structure and influence paths within this system, but I know of no instance where the time response of the system has been examined. I also know of no instances where the absolute numeric response of the system has been cleanly measured and reported, allowing us to see how the social media ecosystem amplifies the input signal.

ReadWriteWeb is a fairly influential original content source within this social media ecosystem. ReadWriteWeb is ranked approximately 20th in the top 100 blogs according to Technorati[2]. They rank blogs not by readership but by their influence and authority in the blogosphere. The ratings change daily, but on October 19, 2009, ReadWriteWeb was ranked 24th. At the time of this writing, they are ranked 14th (the Huffington Post is #1, followed by Techcrunch). The traffic to ReadWriteWeb is not overwhelming. They are the 1400th most popular web site in the United States according to Alexa[3], and they have about 700,000 unique month visitors according to Compete[4].

The manner in which news and information is distributed in the modern world differs from the days when a single newspaper could claim to have “All the news that is fit to print”. In those days, the editor decided what was important and, placed it on page 1, and the story would become important that day. Presumably, few people read the story the day after it was published, or the day after that. Sharing consisted of cutting out the article with scissors and mailing it to your friend. In some cases, major articles would be syndicated by regional papers, but always with authorization of the original content creator.

Information is now created, distributed, and consumed differently. We still have at the foundation the original content creators. These consist of traditional news organizations as well as newer, online-only organizations such as ReadWriteWeb. The home page of ReadWriteWeb uses a blog format where articles are posted sequentially (newest article at the top) with a long preview and the opportunity to click through to the full article. Many people visit their website one or more times during the day and read the articles that interest them. Other regular readers have subscribed to ReadWriteWeb through a RSS reader or personalized news service (Google Reader, Netvibes, etc.) and directly navigate to the articles that interest them. Alternatively, people share their favorite article links with others using services such as Twitter, Digg, or Facebook. In this consumption mode, irregular readers of ReadWriteWeb are exposed to their content. For all of these consumption scenarios the original article in ReadWriteWeb is made available to readers. In terms of traffic at the target (stinkyteddy.com) there is a direct relationship between readers on

[2] http://technorati.com/blogs/top100/
[3] http://www.alexa.com/siteinfo/readwriteweb.com
[4] http://siteanalytics.compete.com/readwriteweb.com/

ReadWriteWeb and visits to stinkyteddy.com (presumably many people who read an article about a new startup click through to check out the product).

The social media ecosystem also has untraceable, indirect pathways. Some blogs are follower blogs that write original stories about topics they see elsewhere. In this case, http://www.shinyshiny.tv and http://www.killerstartups.com (among others) wrote original pieces that were inspired by but not directly copied from the ReadWriteWeb article. Many other web sites simply plagiarize original content and repost it with or without attribution. On social networks, people spontaneously share information without mentioning the source – someone might tweet "I'm checking out Stinky Teddy." These avenues produce traffic as well, but it is generally impossible to attribute this traffic to its origin. Because of the unique circumstances in this case (no traffic before article, no other simultaneous media, no attempts at promotion), all of this indirect traffic to stinkyteddy.com can also be credited to ReadWriteWeb. This type of spontaneous and/or indirect traffic, by the way, is what entrepreneurs fantasize about when they hope to "go viral."[5]

It is natural that the different pathways from ReadWriteWeb to stinkyteddy.com will have different timescales and lifetimes. We can measure an overall quantity of visitors to the site, but it is generally impossible to track the exact pathway on an individual basis.

3 Data

At the time of the ReadWriteWeb article, stinkyteddy.com was configured to log information about visitors to the site and users of the search engine. Every page accessed within the domain is logged with the exact time of access. In addition, a unique identifier is stored in a cookie whenever a user accesses any page within the domain. This cookie is used to identify repeat visitors and to monitor how each visitor interacts with the page. No personally identifiable information is stored.

At the time of the article, we were not storing the referral page that is sometimes contained in the http header[6]. This can be used to track the page from which a user clicked. We activated this logging at a later date, but found that by that time most of the clicks had "null" as a value for referral page. We did not do any detailed studies of the traffic for which we did have an available referral url.

Traffic to stinkyteddy.com goes primarily to two pages. The home page consists of nothing but a search box and links to other pages. There are also check-boxes that allow people to configure which types of content they want included in their search (web, news, video, real-time, and images). The look and feel is such that every visitor knows exactly how the site operates – it is a standard search engine. Unlike many new products, there are no barriers to usage in the form of registration requirements or learning hurdles. We speculate that the low barriers to usage and high familiarity with the product behavior mean that the ratio of people who visit stinkyteddy.com compared to the number of people that read about stinkyteddy.com should be fairly high.

In terms of total number of visitors, there were approximately 11,500 visits to http://www.stinkyteddy.com in the one month period after the publication of the article in ReadWriteWeb. There were roughly the same number of searches performed. Of course, many people visited the site without doing a search and other people performed multiple searches. We have several other pages accessible, including a preferences page where the user can configure the behavior of the search engine and an "about" page that describes the product. Each of these pages received approximately 1,000 visits in the one month period after the launch. These low visit numbers confirm that people are generally uninterested in configuring a search engine or knowing how it works.

[5] stinkyteddy.com did not "go viral". Not even close.

[6] When a web page is requested, the user's browser can add information to the web page request that informs the web server how to handle the request.

Figure 1 shows the number of hourly visits to http://www.stinkyteddy.com in the 15 hours preceding and 100 hours after the publication of the article on ReadWriteWeb. As expected there is a sudden surge in traffic immediately upon publication of the ReadWriteWeb article. The peak visitor frequency is in the period 1-2 hours after the publication of the article. The view rate drops quickly and is down by a factor of 7 only five hours after the publication. We suspect that the quick burst is due to people who read the article by navigating to ReadWriteWeb or had a live RSS reader that shows the current posts on ReadWriteWeb. ReadWriteWeb cycles articles down the page (as is conventional with many blogs) and the article probably jumped to page 2 at the point where the visitations bottom out (5 hours after publication). This first burst of traffic, which originates primarily from the home page of the news organization, accounts for slightly less that 10% of the total traffic eventually generated by the article.

We see a nearly equal satellite peak with high usage 19-20 hours after the article. This is part of a broadly distributed peak between 5 and 30 hours after the article. 19 hours after article publication corresponds to 9 A.M. Eastern Daylight Time. We speculate that this broad distribution of visits between 5 and 30 hours after publication represents people who found Stinky Teddy through either a news reader service (such as Google reader) or a sharing service such as Twitter, Facebook. Many people probably check these accounts for interesting articles when then first log in each morning – that would explain the large spike at 9 AM EDT. It appears that perhaps East Coast readers access ReadWriteWeb in a fundamentally different way (more through readers and less through the web site). Depending on where you live and when an article is published, you'll see new information through different channels.

After 30 hours, we see a decaying traffic pattern with a daily cycle peaking each morning (45 hours, 69 hours, 90 hours). After 100 hours, the hourly traffic becomes small and therefore dominated by random fluctuations with no clear patterns.

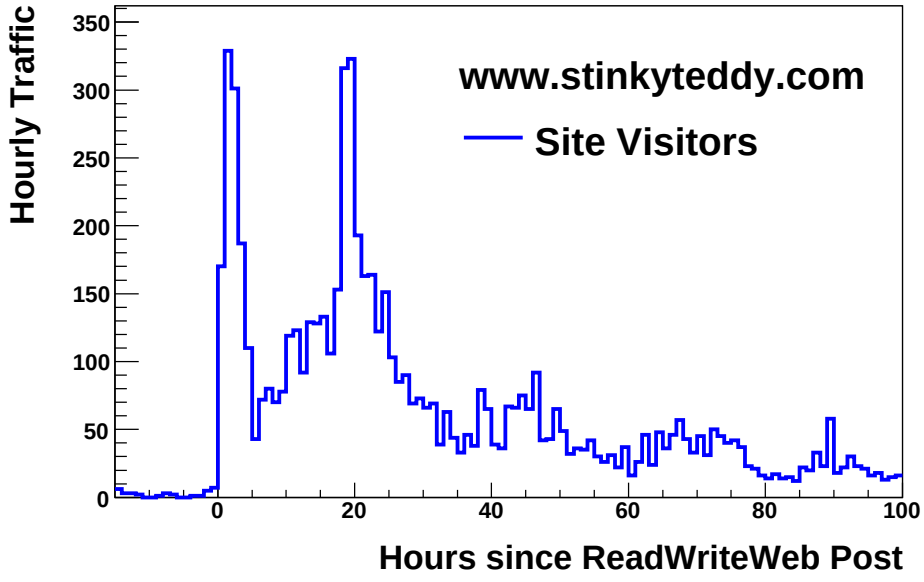

Figure 1: Number of visits per hour to http://www.stinkyteddy.com versus hours since October 19, 10:51 AM PDT.

Figure 2 shows the traffic to http://www.stinkyteddy.com on a daily basis for 25 days after the publication of the article. We see that traffic to the site died off gradually in the month after the article. Nonetheless, the total amount of traffic in the first 3 days (6,200 visits) is similar to the amount of traffic in days 4 through 25 (5,400 visits). We have also modeled the number of visitors per day as an exponential function, and the result is shown by the dotted line on the plot. The result is:

$$\text{Traffic}(d) = e^{6.17-0.059d}, \tag{1}$$

where d is the number of days since the article. This function begins to describe the data several days after the article was published – in this case day 4. There is some structure related to weekend effects (Saturdays are the slowest

day on the Internet, day 11) and a presentation given at a small gathering (day 15), but otherwise the exponential fit describes the data well.

Exponential decay patterns arise when the instantaneous loss rate is a time-independent fixed fraction of the remaining amount. This is a reasonable model of a sharing-driven traffic pattern – if each share yields an average of less than one subsequent share, the traffic will die out exponentially. If this ratio (subsequent shares/share) is greater than one, traffic will increase over time, i.e. "going viral". In our case, the half-life (time that it takes for the amount of traffic to be reduced to 1/2 the original value) is approximately 10 days.

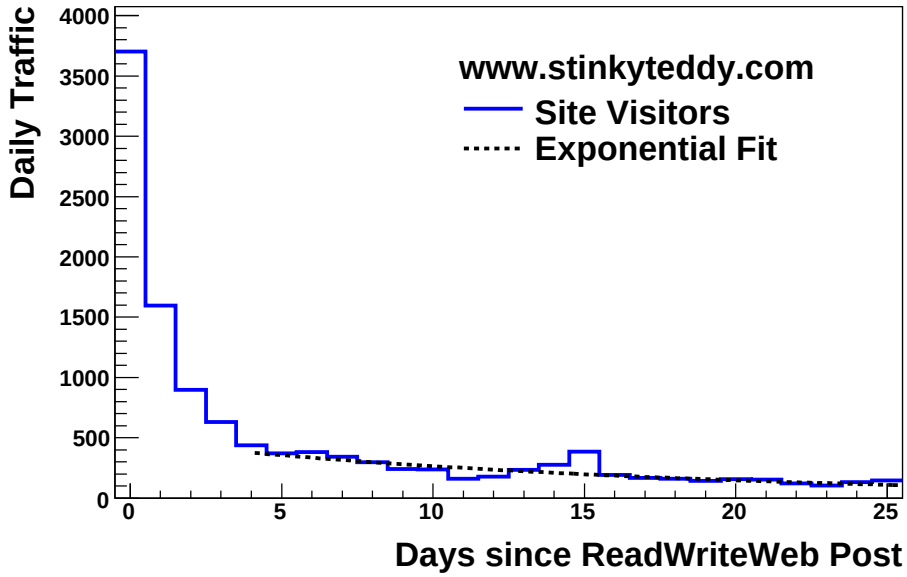

Figure 2: Number of visits per day to http://www.stinkyteddy.com versus day since October 19, 10:51 AM PDT.

4 Conclusion

Due to a unique set of circumstances, we were able to make a measurement of the social media impulse response function. After a single publication in a fairly influential blog (ReadWriteWeb) we observed traffic to our search engine (http://www.stinkyteddy.com). As we made no attempts to generate additional publicity, all traffic to the website can be attributed either directly or indirectly to the ReadWriteWeb post.

The time histories of the site visitors show three primary components to the traffic. First, there is short lived component (several hours) due to direct navigation from the ReadWriteWeb article to the target web site. There is a second broad peak that we suspect arises primarily from news readers and sharing services. This component has a rather sharp peak at 9 AM EDT the next day, so this component is akin to the "morning newspaper over coffee".

Finally, there is a slowly decaying component starting several days after the original post that can be modeled by an exponential function with half-life of approximately 10 days.

To gain a better understanding of news consumption in the era of social media, this study can be repeated and improved. First, traffic to the primary article needs to be monitored and compared with traffic to the destination web site. It would be interesting to see if social media amplified the signal (more visitors to Stinky Teddy than to the

ReadWriteWeb post). Additionally, accurate referrer logging should be kept. We did not activate referral logging until several days later, and in most cases the referrer page is not available in the http request. It would be interesting to see how much traffic comes from copy cat blogs that plagiarize or repackage the original content. Most importantly, this study should be repeated simply to see if the time patterns observed here are universal.