Social Media Pulse

Scott Hendrickson

Gnip, Inc.

June 13, 2013

1 Introduction

When a breaking news story emerges on a social media platform, the rate of media activities follows one of three easily identifiable patterns. It is useful to identify and model these patterns so that we can

1. compare the growth, peak and decay of stories
2. compare the behavior of people on different social media platforms
3. better identify emerging stories

The three patterns are summarized in Table 1. Broadly, audience behavior can be separated by whether the audience expected the event or if the event was a complete surprise. Secondarily, if the event was a surprise, whether the event was observed by a large number of people in the social media audience simultaneously or spread virally through the social media platform.

The pattern of surprise events is characterized by a sharp rise in activity as users quickly share information. It is followed by an exponential decay in volume after the information saturates the network. I call this pattern the *Social Media Pulse*. It is described by a Poisson Counting Process, described below. The pattern of expected events is characterized by period of growth of activity that is often much more symmetric with the decay of the story. These patterns may span hours or several days and consequently will be superimposed on the daily and weekly cyclic patterns typically of the social media platform.

Examples of stories that show the Social Media Pulse include the Oaxaca, Mexico earthquake (20 March 2012) experienced and tweeted by many people across southern Mexico and the announcement of Beyonce's pregnancy at the 2012 Grammy Awards. These are surprise events, witnessed by many social media users at the same time and generating social media activity shortly after being observed.

Examples of events not showing the Social Media Pulse are planned events such as the Grammy's themselves, or the arrival of a hurricane, both of which

Event Type	Response	Examples
Expected	Approx. Symmetric	Hurricane, Video Music Awards, Olympics.
Unexpected (Simultaneous Observation)	Social Media Pulse	Beyonce performs pregnant at VMAs; Mexico earthquakes; Steve Jobs passing; new link shared
Unexpected (Breaking on Social Media)	Cascade Models	Osama Bin Laden strike (up to the point news agencies broke the story, then Social Media Pulse), Whitney Houston passing, Syrian dissident actions (often no traditional news coverage).

Table 1: Summary of behavior of social media activity volume with time for expected and unexpected events.

are predicted and covered extensively by many media outlets over the period of days or even weeks.

News stories that are originally broken on social media do not follow this Social Media Pulse, but are better represented by models of information cascading through the network. The news of Bin Laden's death was broken on twitter 18 minutes before conventional news sources and grew slowly until it was released from traditional media sources. (Mathematically, the pattern is a mix of the Unexpected-Breaking and Unexpected-Simultaneous.) In the case of unexpected stories breaking or partially breaking (e.g. the Twitter story gets ahead of the news story on Twitter), the growth and peak are best modeled by exponential growth with saturation.

The Social Media Pulse pattern is general in that it is associated with surprise events in firehoses from Twitter, Tumblr, Wordpress and discussion platforms such as Disqus and Intense Debate. The following section is a deeper look at a model of the Social Media Pulse.

2 Social Media Pulse Model

The motivation to quantify the Social Media Pulse is to standardize comparison and speed detection of emerging stories.

We want to both compare the evolution of different stories as well as compare a given story across the various social media platforms. Each platform has features and modes of conversation that attract more of one type of user than another.

The rapid growth of activities at the beginning of the pulse is due to many of the users interested in a new story posting almost simultaneously. This is

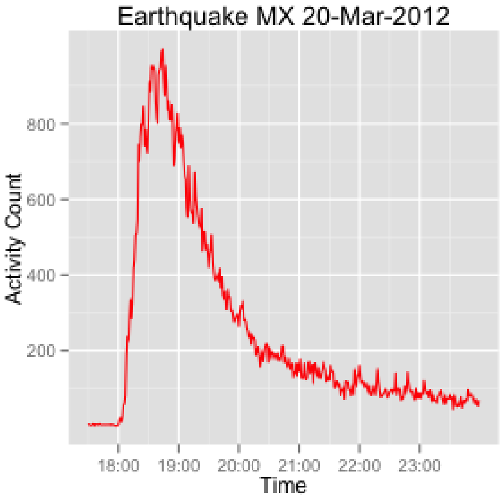

Figure 1: Twitter activity pulse associated with earthquake in southern Mexico.

not viral spreading of a story through the network. Rather, the beginning of the pulse is often very abrupt because everyone posting witnessed the event simultaneously. The event (Earthquake or Beyonce') occurs and many users grab for their phone or keyboard to post about it. Once a story has been passed to most of these users followers, activity falls off. This decaying of a story is characterized by exponential decrease of activity.

A reasonable strategy for building the Social Media Pulse model is to assume the process is a Homogeneous Poisson Counting Process. There are some excellent references for the statistical thinking and mathematics behind this model (`http://en.wikipedia.org/wiki/Poisson_process`, `http://www.math.uconn.edu/~valdez/math288s08/Math288-Weeks7to9.pdf`). I will outline the arguments briefly here.

After the event, the probability of an observer posting on the topic decays at a constant rate, λ, such that,

$$f(t) = \lambda \exp(-\lambda t), \text{ for } x \geq 0. \tag{1}$$

If the same social media poster were to observe the same event over and over and post on it, the expected value of their average time to post is $E(X) = 1/\lambda$. This distribution has an interesting "memory-less" quality you may want to learn more about later. For now, remember that we are looking at the behavior of *many* social media posters, that is, we have a new random variable that is the sum of many of the X's above, $S = X_1 + X_2 + \ldots + X_{n\text{posters}}$. To calculate the probability of activity posts for this new random variable, we take the convolution of the individual distributions.

The result is the Gamma distribution,

$$\begin{aligned} f_S(t) &= \frac{\beta^{-\alpha}(t-t_0)^{\alpha-1}\exp(\frac{-(t-t_0)}{\beta})}{\Gamma(\alpha)} \quad (2) \\ r(t) &= N_{activities} f_S(t) \quad (3) \\ & \quad (4) \end{aligned}$$

where t_0 is the starting point of the pulse.

The model is motivated by both the desire to compare different stories on a given social media platform and also to compare stories across platforms with analytical consistency. To calculate parameters for comparison (total number of tweets, the half-life of the story, the time to peak, etc.) A reasonable strategy for fitting Eq. 4 is to bucket activity data to estimate the activity rate at regular time intervals. The rate is give by the number of activities per bucket divided by the bucket width, δt_{bucket}. The quality of the estimates depends on the bucket size as related to the timescales of the pulse. That is, we can determine a reliable activity rate as long as the bucket width is small compared to the timescale,

$$\delta t_{bucket} << t_{peak} \tag{5}$$

Practically, it is often adequate to have t_{peak} only a few times t_{bucket}.

The lower limit is set by the requirement that there not be too few activities per period. If the number of activities is too low, the estimate of rate becomes very uncertain (estimates of averages vary as $1/\sqrt{\text{sample size}}$). So, we have the condition,

$$\text{rate} >> 1 \text{ activity per period} \tag{6}$$

In practice, we choose δt_{bucket} to satisfy both Eq 5 and 6.

We can use common curve fitting techniques to compute the parameters of Eq 4. For example, in Python, use the Scipy package, optimize module method `optimize.leastsq(...)` to quickly perform the fit. To make sure we fit the pulse quickly and accurately, it is important to window the data around the pulse.

It is also convenient to provide an initial guess for the fit calculate from the sample data. In the following, the left-hand values are initial values for the fit and the right-hand values are calculated from the pulse sample data.

$$\begin{aligned} \beta &= \frac{t_{range}}{2} \quad (7) \\ \alpha &= \frac{t_{max}}{\beta} + 1 \quad (8) \\ t_0 &= t(r_{max}) \quad (9) \\ N_{activities} &= 4 * r_{max} \quad (10) \end{aligned}$$

where $t_{range} = t_{max} - t_{min}$, the data window size, and t_0 is the start time of the pulse.

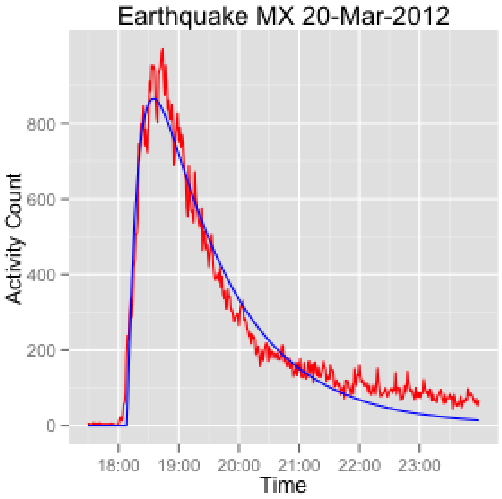

Figure 2: Twitter: Mexico Earthquake Social Media Pulse with fitted Social Media Pulse.

Once we have fit the function, use r_0, α and β to calculate the relevant parameters that characterize Social Media Pulse.

Time-to-Peak. The peak time is the time when the activity rate reaches its peak. This is derived by maximizing Eq 4 . The result is,

$$t_{time-to-peak} = \beta(\alpha - 1) \tag{11}$$

The lower the time-to-peak, the faster the story reached a majority of the interested parties on the social media platform. This is a useful proxy for comparing the speeds at which various stories get out.

Story Activity Volume. The story activity volume is the integral of the rate for the relevant time. The cumulative distribution function of the gamma distribution, the "generalized regularized incomplete gamma function" is,

$$F_S(t) = Q(\alpha, 0, \frac{t - t_0}{\beta}) \tag{12}$$

in terms of the incomplete gamma functions,

$$S_{vol} = \int_0^\infty r(t)dt = N_{activities} F_S(t), \tag{13}$$

This is a useful proxy for comparing story size.

Story Half-life. Attention wanes after the story peaks. The half-life is the point where half of the total number of activities for the story have been observed. Therefore, we define $F_S(t_{\frac{1}{2}life}0 = \frac{1}{2}$, or,

Table 2: Fit parameters for Twitter, earthquake.

Parameter	Value
t_0	fixxx sec
	2012-03-20 18:08:22 UTC
r_0	fixxx act/min
α	0.001152 sec^{-1}
β	0.000220 sec^{-1}
$t_{time_t o_p eak}$	26 min 30 sec
S_{vol}	5,568,800 tweets
$t_{1/2}$	92 min 01 sec
t_{avg}	87 min 49 sec

$$t_{\frac{1}{2}life} = F_S^{-1}(\frac{1}{2}). \tag{14}$$

Average Response Time. Average response time is the balance point between the volume in the tall peak and activity long after the event, in the skinny tail. The average response time is given by,

$$t_{avg} = \alpha\beta \tag{15}$$

Python scripts for fitting Eq **??** and calculating all of the parameters defined above are available at `https://github.com/DrSkippy27/Social-Activity-Pulse-Function`.

3 Notes on the Oaxaca, MX Earthquake

We have been looking at twitter volume matching the terms “quake” and “terremotto” immediately following the earthquake in Oaxaca, MX on 20 March 2012. The Social Media Pulse along with the fitted line is shown in Fig. 2. The parameters of the fit are shown in Table 2.

The USGS web site which monitors earthquakes worldwide (`http://earthquake.usgs.gov/earthquakes/eqinthenews/2012/usc0008m6h/`) reports the earthquake occurred at 18:02:48 UTC, so we can infer Twitter volume starts a few minutes after the event.

Notice the divergence of the model from the data toward the right of the curve. This starts a little before 21:00 and is due to Tweets from users who heard about the earthquake through news sources. Fitting the Social Media Pulse give the queue to look at Tweets around 21:30 to see what had changed from earlier. (This effect also implies that the actual tweet volume is likely greater than the Social Media Pulse volume of 5.67M tweets.)

4 Sigmoid Model

When a story breaks on social media, the volume shows a period of exponential growth starting from a single activity. As the message travels, people stop repeating it–i.e. the message saturates the network. A classic model for exponential growth with saturation yields a curve familiar to many machine learning practitioners from Logistic Regression and Neural Network activation functions. For more information on the model, see `http://en.wikipedia.org/wiki/Logistic_function`.

A convenient model for social media is to center "S-curve" around the time the pulse reaches 1/2 the full height since the tails are asymptotic. In the following, the growth rate is characterized by γ to avoid confusion with the growth rates in the Social Media Pulse model above.

$$r(t) = \frac{r_0}{1 - \exp(-\gamma(t - t_{1/2}))} \tag{16}$$

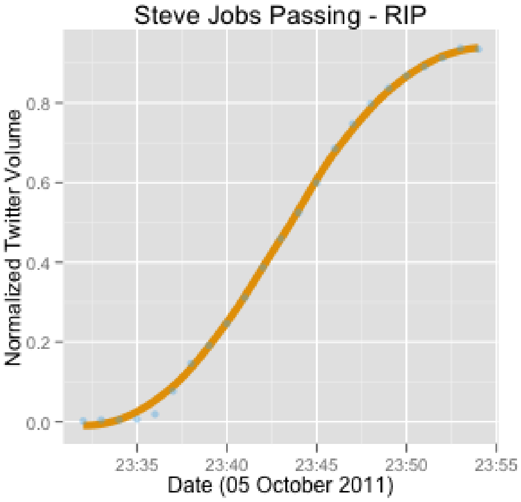

Figure 3: The passing of Steve Jobs breaks on Twitter. Plot shows the normalized volume of tweets mentioning Steve Jobs and "RIP".

As an example, the sad passing of Steve Jobs was first announce through a traditional news source, then tweeted a few seconds later. The news spread from the first few tweets took on a live of its own on Twitter, where many people heard the news for the first time. See Fig. 3 showing the growth of activity up to the peak of the pulse.

The total volume of activity to the peak is given by,

$$S_{peak} = \int_{-\infty}^{T_{peak}} r(t)dt \tag{17}$$

$$= \frac{r_0}{\gamma} \log[1 - \exp(\gamma(T_{peak} - t_{1/2}))] \tag{18}$$

5 Trend Detection

Coming soon - examples of using "Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series" by Stanislav Nikolov `http://web.mit.edu/snikolov/Public/trend.pdf`

6 Conclusion

This is intended to help you use the Gnip social data streams more effectively. The latest version of this document and supporting code for creating figures and tables can be found at:
`https://github.com/DrSkippy27/Gnip-Realtime-Social-Data-Sampling`.

If you find errors or have comments, please email shendrickson@gnip.com. Thank you for using Gnip.

References

[Mor13] F. Morstatter, J. Pfeffer, J. Liu, K. Carley, *Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose*, `http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf` 2013.

[Geo12] F. George B. Golam Kibria, *Confidence Intervals for Signal to Noise Ratio of a Poisson Distribution*, `http://thescipub.com/abstract/10.3844/amjbsp.2011.44.55` 2013.