Social Media Activity
Time Series Signals and Sampling

Scott Hendrickson, Gnip, Inc

February 12, 2013

1 Introduction

For many topics, the social media streams contain frequent enough data to create reliable, high-resolution volume-based signals. But for the long tail of topics, we need to take more care in identifying on what time scale an activity time series signal is meaningful. This white paper addresses questions related to social media activity time series sampling, signal and confidence.

1.1 Motivating Questions

- I plan to bucket data to estimate rate, how long should the buckets be?
- How many activities should I target to collect per bucket?
- How many buckets do I aggregate to optimize the trade-off between signal sensitivity and signal latency?
- I want to minimize the activities I consume, what sampling factor should I use if I want my signal to have 1-hour sensitivity?

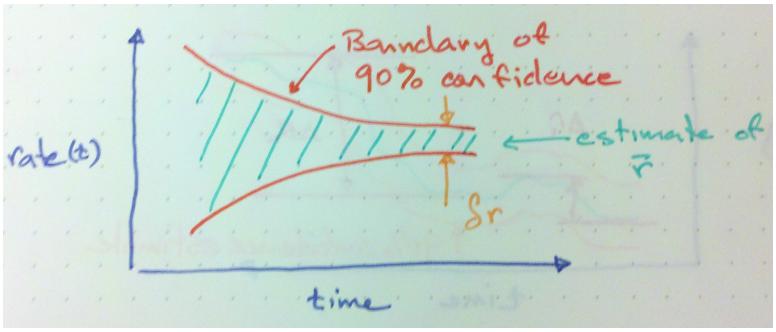

Figure 1: Uncertainty in the estimated rate of activities goes down as we observe more events.

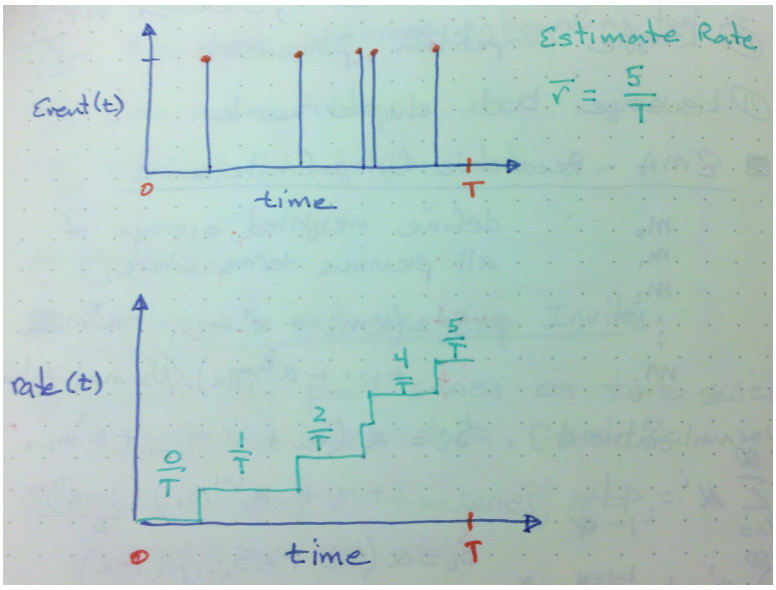

Figure 2: Events and rate estimates

- How do I describe the trade-off between signal latency and rate uncertainty?
- How do I define confidence levels for rate estimates for low-frequency time series?
- ...

1.2 Activity Rate

For all of these questions, it is useful to start toward an answer by defining the activity rate.

$$\bar{r} = \frac{N}{T} \tag{1}$$

where N is the number of activities and T period. Our goals will be defined in terms of how many activities N we can count to estimate $\bar{r}$ to the desired level of confidence. The next step is to define the signal we are trying to detect.

2 Signal

We have a activity time series-based signal when the activity rate changes with time more than the sensitivity threshold. That is,

$$|r(t_f) - r(t_i)| \geq \Delta r \tag{2}$$

The time scale of the change, $T_l = t_f - t_i$, is the Signal Latency. This definition implies that the we must observe activities for a time $T > T_l$ to observe the signal, Δr.

2.1 Signal-Confidence Criteria

We will be estimating the activity rate by Eq. 1 by counting activities for a known time. The number of counts in any given period will be distributed about the mean. As we count more activities, our estimate of rate will improve. If we count thousands of activities per minute, our confidence of the estimate of activity rate will be very high after a short time. For rare activities, we will have to count for a longer time before we have a high level of confidence in our rate estimate.

Referring to the signal definition Eq. 2, the criteria that allows us to calculate confidence in terms of signal is,

$$\delta r << \Delta r \tag{3}$$

The variation of the observed number from the average number will decrease with increasing time or increasing rate (the number of activities counted).

The help quantify this inequality, introduce a multiplier factor η that describes how much larger the change in rate is compared to the uncertainty of the rate estimate:

Parameterized the criteria,

$$\eta \delta r = \Delta r \tag{4}$$

where $\eta >> 1$ for cases where the criteria is fulfilled.

As will be detailed below, the criteria represents trade-offs of number of the number activities (cost of collection and licensing), time (signal latency–how long we have to wait to know the rate has changed), confidence (reliability of estimates of rate), and size of change we can detect Δr (the signal sensitivity).

These trade offs are summarized in the Table **??**.

3 Statistics of Time Series of Activities

3.1 Time-Between Independent Activities

A workable model of counts of rare activities is that the inter-arrival times are expoentially distributed,

$$p_{activity}(t) = re^{-rt} \tag{5}$$

This assumption leads the Poission distribution of activity counts over time.

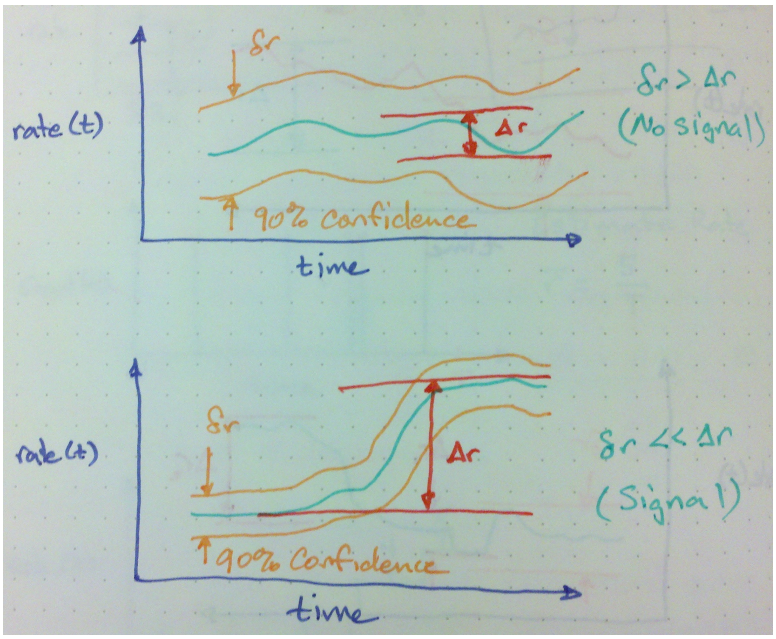

Figure 3: A signal can be detected when the change in rate is greater than the uncertainty in the rate estimates.

Want to...	Actions
Minimize Activities Analyzed	(decrease N) increase ΔR (decrease signal sensitivity) decrease confidence (E.g. from 95% to 90%)
Increase Signal sensitivity	(decrease Δr) increase T (increase number of buckets (k) or increase bucket size (Δt)) increase activity rate (r) by broadening filter or increase PowerTrack sampling
Decrease Signal Latency	(decrease T_l) decrease signal sensitivity Δr decrease confidence factor (α) increase activity rate (r) by broadening filter or increase PowerTrack sampling
Decrease Signal Uncertainty	(decrease δr or increase η) increase T (increase number of buckets (k) or increase bucket size (Δt)) increase activity counts (increase N, r) by broadening filter or increase PowerTrack sampling

Table 1: Summary of model trade-offs.

3.2 Poisson Activity Probability

The probability of observing n activities in time t when the activity rate is r is given by,

$$P(n) = \frac{e^{-rt}(rt)^n}{n!} \tag{6}$$

The expected value is $E[n] = n = \bar{r}t$. The mean and variance of the Poisson distribution are both equal to r.

3.3 Poisson Confidence Intervals

We are counting activities in a defined time interval to estimate the activity rate r. Confidence in the estimate of r goes up as we count more and more activities. Confidence intervals for the Poisson For confidence level $1 - \alpha$,

$$\frac{1}{2}\chi^2(\alpha/2; 2n) \leq \hat{r} \leq \frac{1}{2}\chi^2(1 - \alpha/2; 2n + 2) \tag{7}$$

where χ^2 is the inverse cumlative distrituion function, $CDF^{-1}(p; n)$, of the χ^2 distribution.[1] Note that with this definition of α, the a confidence interval of 90% corresponds to $\alpha = 0.1$.

To determine the parameters of our data collection system, we find the value of n for which the time interval and confidence level match our requirements. Give a set of sensitivity, latency, etc. requirements, we can use the confidence interval to calculate any one of the parameters.

Calculations for various design choices and an unknown are illustrated in the next section. First, let's look at some useful limits and approximations for calculating confidence intervals.

3.4 Less-Rare Activities

For large n, the normal approximation makes the interval calculation simpler. When we observe large values of n, the confidence interval can be estimated using the Normal approximation. For example, for 95% confidence interval the interval is symmetric about the mean and given by,

$$\bar{r} - 1.95\sqrt{\bar{r}/n} \leq \hat{r} \leq \bar{r} + 1.95\sqrt{\bar{r}/n} \tag{8}$$

Confidence intervals for activity counts are shown int Table 2.

3.5 Confidence Intervals on Bucketed Time series

For many reasons, counts may be collected in buckets of some pre-defined time length. The rate information may by more naturally calculated by bucket rather

[1] A useful approximation to the exact interval is given by $[n(1 - \frac{1}{9n} - \frac{z_\alpha}{3\sqrt{n}})^3, (n+1)(1 - \frac{1}{9(n+1)} + \frac{z_\alpha}{3\sqrt{n+1}})^3]$.

N	Bounds ($N >> 1$)	Interval ($N >> 1$)	Bounds	Interval
1	-	-	[0.0513, 4.743]	4.692
2	-	-	[0.3554, 6.295]	5.940
3	-	-	[0.8177, 7.753]	6.936
4	-	-	[1.366, 9.153]	7.787
5	-	-	[1.970, 10.51]	8.542
6	-	-	[2.613, 11.84]	9.229
7	-	-	[3.285, 13.14]	9.862
8	-	-	[3.980, 14.43]	10.45
9	-	-	[4.695, 15.70]	11.01
10	-	-	[5.425, 16.96]	11.53
20	-	-	[13.25, 29.06]	15.80
30	[20.99, 39.00]	18.01	[21.59, 40.69]	19.09
40	[29.59, 50.40]	20.80	[30.19, 52.06]	21.87
50	[38.36, 61.63]	23.26	[38.96, 63.28]	24.32
60	[47.25, 72.74]	25.48	[47.85, 74.38]	26.53
70	[56.23, 83.76]	27.52	[56.82, 85.40]	28.57
80	[65.28, 94.71]	29.42	[65.87, 96.35]	30.47
90	[74.39, 105.6]	31.20	[74.98, 107.2]	32.25
100	[83.55, 116.4]	32.89	[84.13, 118.0]	33.94
200	[176.7, 223.2]	46.52	[177.3, 224.8]	47.55
300	[271.5, 328.4]	56.97	[272.0, 330.0]	58.00
400	[367.1, 432.8]	65.79	[367.6, 434.4]	66.81
500	[463.2, 536.7]	73.56	[463.7, 538.3]	74.57
750	[704.9, 795.0]	90.09	[705.5, 796.6]	91.10
1000	[947.9, 1052.]	104.0	[948.5, 1053.]	105.0

Table 2: Confidence intervals for number of counts in time T. Rate confidence range is $\delta N/T$. The large N approximation is shown when the boundaries of within 5% of the exact value.

Parameter	Definition
N	Number of activities in time T
T	Observation time
Δt	Bucket size (for bucketed data where $\Delta t < T$)
k	Observation time measured in buckets ($k = T/\Delta t$)
r	Activity rate
$\bar{r} = N/T$	Estimate of activity rate
δr	Uncertainty of rate estimate
α	Confidence fraction of rate estimate range
Δr	Change in activity rate that defines signal
T_l	Signal latency
η	Rate signal criteria factor

Table 3: Summary of model parameters.

than the total time T required by our confidence requirements. In general, define the relationship between T and the bucket size (constant) as,

$$\Delta t = \frac{T}{k} \tag{9}$$

where k is the number of buckets that we need to aggregate to observe for time T. This parameter can be used to calculate a corresponding signal latency, $k_l = T_l/\Delta t$.

Resolution times are interchangeable with number of buckets k given $\Delta t << T$. In general, the bucket resolution time will not be an even multiple of the bucket size. In this case, imposing the calculation of average rate per bucket $\bar{r} = n/\Delta t$ adds another layer of variability.

See example calculations below for a lookup table of factors.

3.6 Note on Social Media Pulse

In the case where something happens in the real world that many social media users can observe and react to (e.g. an earthquake or a celebrity baby photo link leaked on Twitter). The time series of activities will no longer fulfill independence and constant-rate requirements. In these cases, activities can be characterized by the mathematics of the Social Media Pulse (link!).

These activities are likely to be associated with the change in rate, Δr, that is the signal we are looking for in the stream.

3.7 Model and Parameters

Table 3 summarizes the parameters of the model.

4 Examples

To make this concrete and illustrate the use of the lookup tables, following are some sample calculations.

4.1 Estimate Signal Latency

To do ...

4.2 Estimate the Optimal PowerTrack Sampling Operator Value

To do ...

4.3 Estimate Signal Resolution

To do ...

5 Conclusion and References

This is intended to help you use the Gnip social data streams more effectively. If you find errors or have comments, please email shendrickson@gnip.com. Thank you.