Real Time Social Data Sampling

Scott Hendrickson, Brian Lehman, and Joshua Montague

Gnip, Inc.

June 3, 2013

1 Introduction

We often wish to calculate social activity rates to identify signals of emerging topics or stories. We want to know what we can detect and understand the certainty of our estimates. This means we need to calculate the relationship between activity counts, signal sensitivity and confidence. This is illustrated in Figure 1. The event rate is the number of activities per time, the signal is the change in rate, and confidence comes from the statistics of the activities and sampling operators. Time enters the calculation because the longer we count, the more activities we count.

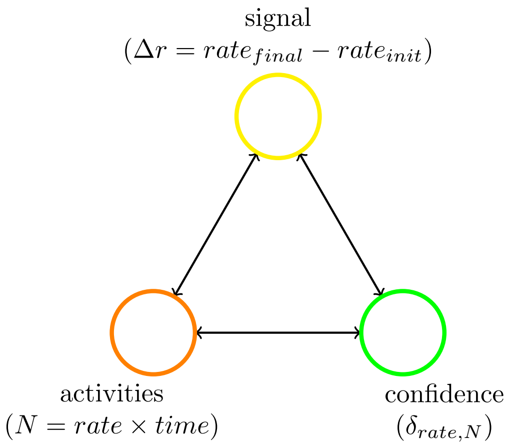

Figure 1: The number of activities is a product of the rate and time. Signal is the change in event rate we want to detect. Confidence characterizes the uncertainty in our estimate of rate. These parameters are not independent: choose any two and calculate the third.

For many topics, social media streams contain data with sufficient frequency to create reliable, high-resolution signals in short observation times. But for the less common topics, the activities (e.g. blog posts or tweets) are less frequent and we need to take care in identifying the observation time, signal sensitivity and confidence level. These three quantities are related; typically, we can choose two and calculate the third.

There are a handful of related questions that may come up regarding sampling and signals. In the next section, we list some of these questions to explore the meaning of the relationships outlined above. After that, we will look at filtering and sampling in a little more detail. The middle of this white paper builds up the calculation method required to answer these questions. Finally, in the last section, we provide example calculations.

1.1 Motivating Questions

Common questions regardind rate, signal and confidence level:

- The activity rate has doubled from five counts to ten counts between two of my buckets. Is this a significant change? Or is this expected variation due to low-frequency events?
- I want to minimize the total number activities I consume, what sampling factor should I use if I want to detect changes of 2x in activity rate in 1-hour?
- How long should I observe and count activities to detect a change in signal of 5%?
- How do I describe the trade-off between signal latency (how long I have to wait) and rate uncertainty (how confidently I can estimate activity rate)?
- How do I define confidence levels on rate estimates for a social activity time series with only 20 events per day?
- I plan to bucket the data in order to estimate activity rate, how big (i.e. what duration) should the buckets be?
- How many activities should I target to collect in each bucket in order to be have a 95% confidence that my rate estimate is accurate for each bucket?

1.2 Sample Operators and Order of Filters

In order to to control costs or to scale analysis, we may choose to sample a known fraction of the total social activities. This presents an additional complexity to the ideas introduced in the introduction. Using a sampling filter effectively decreases the rate of activities. If you plan to use a sampling filter, it may be useful to explore sampling before moving the the core of the calculation. If not, please skip to the next section.

There are two approaches to sampling a firehose of social data. Both involve reducing the number of activities in the real time stream to a manageable size for analysis.

The first step is to use keyword filtering. Gnip provides filtering on keywords to select only the portion of the stream that is relevant to the topic you want to analyze. For example, if you are interested in tracking the Super Bowl, you might start with a broad stream defined by the keywords “superbowl” “super bowl” and “contains:xlvii”, the latter being the Roman numeral of the Super Bowl as might be seen in hashtags or short links. This would limit the social data stream to activities related to the Super Bowl.

In the case of a major event like the Super Bowl, the keyword-filtered stream may represent a very large number of activities. In this case, a second step might be to filter this stream to a known fraction of the total firehose. For example, using Gnip’s sampling operator, we can reduce the stream to only fraction, e.g. 12%, of the activities. The corresponding Powertrack rule would become “(super bowl OR superbowl) sample:12”.

It is useful to know how Gnip’s sampling algorithms work to inform sampling decisions. Sampling is available on Gnip’s premium streams including Tumblr, Twitter, Wordpress and Disqus. Some key features of Gnip sampling:

1. 1% resolution
2. Stable sampling rate (short time resolution)
3. Sample is deterministic and returns the same activites for near-rule matches. For example, this means that you will get the same tweets for matches to the “super bowl” portion of the rules “super bowl sample:12” and “(super bowl OR superbowl) sample:12”
4. Sampling is progressively inclusive (i.e. a 2% stream (e.g. “sample:2”) includes the exact activities from the 1% stream plus an additional 1%, and so on)
5. Activities are first selected from the firehose to reach the desired sampling rate and then filtered by keywords

What happens when we look at the combination of filtering with sampling? Continuing with our Super Bowl example, assume our term filtering rules return $y = 5\%$ of the stream. Assume further that we sample $x = 12\%$ sample of the firehose activities and the total activities for the day are $N_f = 500\text{M}$. Filtering and sampling will leave us with approximately,

$$N_{observed} = xyN_{firehose} = 0.12(0.05)500 \text{ M} = 3 \text{ M} \tag{1}$$

activities in a day.

Once you understand this order, it is natural to ask why Gnip does not filter first, then sample. The difference is not in the final outcome, but how long you have to wait for a reliable estimate of rate. If Gnip were to filter on keywords first, Equation 1 would also be a reasonable estimate of $N_{observed}$ on the time scale of the sampling calculation. However, this process would require relaxing properties 3 and 4 in the list above. Both attributes are desirable for most social data projects. Doing the sample first, followed by the keyword filter, gives a slightly more complex behavior because it requires us to deal the effects of statistical variations in the short term.

2 Signal

In many situations, the main question is: *"How many events must we observe in order to detect a change in activity rate?"* Answering this question requires an understanding of the trade offs between sampling time, activity rates, and signal size. First, we will define signal in terms of the activity rate.

2.1 Activity Rate

The average activity rate is calculated by,

$$\bar{r} = \frac{N}{T} \tag{2}$$

where N is the number of activities and T the time period over which we count activities. Because there will be statistical variations in the number of activities in any given time interval, there will be some uncertainty in our estimate of average rate. The more activities we count, the lower the uncertainty in our estimate as shown in Figure 2.

Below, we will determine how many activities N we need to count to estimate the average activity rate, $\bar{r}$, to a desired level of confidence (e.g. 95%). Or, inverting this question: given a level of confidence, how wide is the range of uncertainty about the rate estimate?

The details of calculating the confidence level can be found in the next section. First, we explore the connection between uncertainty in the rate estimate and the size of the signal we can detect.

2.2 Signal Sensitivity

When activity rates are high, rate estimates will be more certain; when low, rate estimates will be less certain. High uncertainty in the rate estimates may hide the small changes in activity rates we wish to identify. The variation in activity rate do to the statistics associate with infrequent events (or short observation time) must be smaller than the signal we want to detect. Therefore, we observe a valid signal in a time series when the activity rate has changed by more than the rate signal sensitivity, Δr, defined as,

$$|r(t_f) - r(t_i)| \geq \Delta r. \tag{3}$$

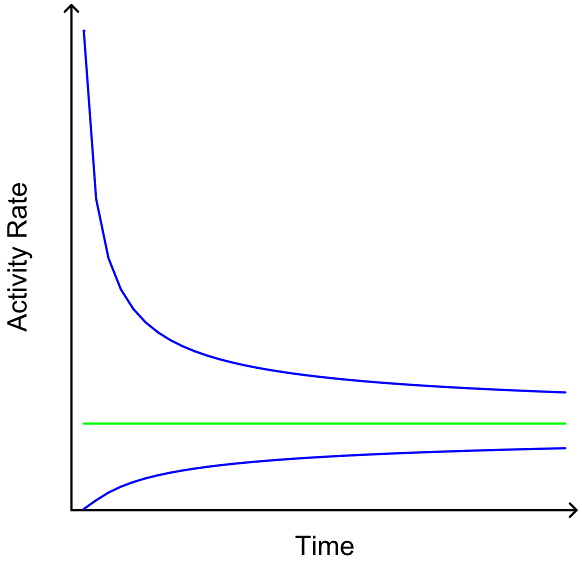

Figure 2: Uncertainty in the estimated rate of activities goes down as we wait longer in order to count more activities. The green line represents the rate while the blue lines show the upper and lower 95% confidence band.

The associated 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 achieve signal sensitivity Δr.

2.3 Signal Sensitivity–Confidence Criteria

We will be estimating the activity rate in Equation 2 by counting activities for a determined time period. The number activities in any given period will be distributed about the long term mean. As we count more activities, our estimate of rate will converge toward the true value. 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 longer before we have a high level of confidence in our rate estimate.

Referring to the signal sensitivity definition in Equation 3, we can establish a rough criteria for confidence in terms of signal uncertainty, δr:

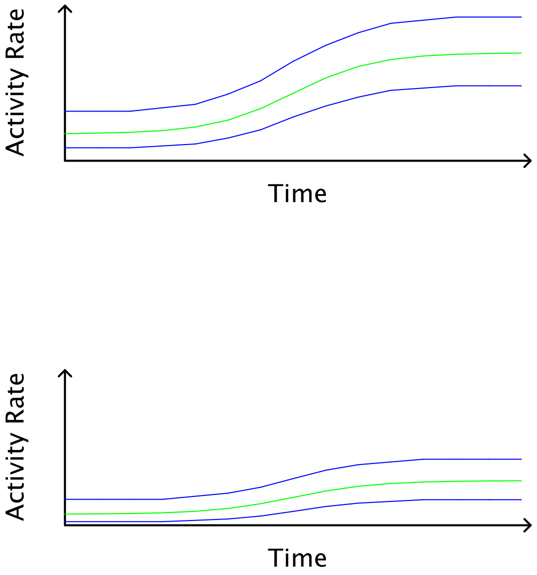

Figure 3: A signal can be detected when the change in rate is greater than the uncertainty in the rate estimates.

$$\frac{\delta r}{\bar{r}_{max}} << \frac{\Delta r}{\bar{r}_{max}}. \tag{4}$$

where $\bar{r}_{max}$ will be the lower rate estimate (initial if detecting rising activity rate but final if detecting falling activity rate).

To help quantify this inequality, we introduce a criteria factor, η, that tells how much larger than the relative uncertainty the change in rate must be.

$$\frac{\delta r}{\bar{r}_{max}} = \frac{\eta \Delta r}{\bar{r}_{max}} \tag{5}$$

where $\eta \geq 1$.

This criteria represents the trade off between signal size and confidence.

3 Statistics of Time Series of Activities

In this section, we examine the underlying statistics in order to calculate confidence intervals and confidence levels.

3.1 Poisson Activity Probability

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

$$p_{activity}(t) = re^{-rt} \tag{6}$$

This assumption leads to a Poission distribution of activity counts over time.

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

$$P(n) = \frac{e^{-rt}(rt)^n}{n!} \tag{7}$$

The expected value is $E[n] = n = rt$. The mean and variance of the Poisson distribution are both equal to r.

3.2 Poisson Confidence Intervals

We are counting activities in a defined time interval to estimate the activity rate $\bar{r}$. Confidence intervals for the Poisson distribution with confidence level conf% $= 100\%(1 - \alpha)$ are given by,

$$\frac{1}{2T}\chi^2(\alpha/2; 2n) \leq r \leq \frac{1}{2T}\chi^2(1 - \alpha/2; 2n + 2) \tag{8}$$

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

Confidence interval sizes for confidence levels of 90% are shown in 1

To determine the parameters satisfying our data collection goals, we find the value of n for which the time interval and confidence level match our requirements for signal detection. That is, we can now calculate any one of signal sensitivity, signal latency, activity rate, and confidence level given the other parameters. Calculations for various design choices are illustrated in the last section of this paper.

3.3 Confidence Interval Approximations and Bucketed Activity Counts

This section deals with approximations to the Poisson confidence interval for large numbers of activities. In addition, we it contains some observations about bucketed activity counts. You can skip this section and move to calculations if these ideas don't apply to your system.

[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	Interval Bounds	Interval Size	Relative Interval
1	[0.0513, 4.744]	4.693	4.693
2	[0.3554, 6.296]	5.940	2.970
3	[0.8177, 7.754]	6.936	2.312
4	[1.366, 9.154]	7.787	1.947
5	[1.970, 10.51]	8.543	1.709
6	[2.613, 11.84]	9.229	1.538
7	[3.285, 13.15]	9.863	1.409
8	[3.981, 14.43]	10.45	1.307
9	[4.695, 15.71]	11.01	1.223
10	[5.426, 16.96]	11.54	1.154
20	[13.25, 29.06]	15.81	0.7904
30	[21.59, 40.69]	19.10	0.6366
40	[30.20, 52.07]	21.87	0.5468
50	[38.96, 63.29]	24.32	0.4864
60	[47.85, 74.39]	26.54	0.4423
70	[56.83, 85.40]	28.57	0.4082
80	[65.88, 96.35]	30.47	0.3809
90	[74.98, 107.2]	32.25	0.3584
100	[84.14, 118.1]	33.94	0.3394
200	[177.3, 224.9]	47.55	0.2378
300	[272.1, 330.1]	58.00	0.1933
400	[367.7, 434.5]	66.82	0.1670
500	[463.8, 538.4]	74.58	0.1492
750	[705.5, 796.6]	91.11	0.1215
1000	[948.6, 1054.]	105.0	0.1050

Table 1: Confidence intervals given the number of events counted N in unit time T. Rate confidence range is $\delta N/T$. Note that the relative uncertainty goes down while the absolute size of the interval increases.

3.3.1 Frequent Activities

When we observe large numbers of activities, 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,

$$\bar{r} - 1.96\sqrt{\bar{r}/n} \leq \hat{r} \leq \bar{r} + 1.96\sqrt{\bar{r}/n} \tag{9}$$

3.3.2 Bucketed Activity Counts

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

$$\Delta t = \frac{T}{k} \tag{10}$$

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.

3.4 Summary of Parameters and Trade offs

When activities are common, we can estimate the activity rate to a high level of certainty in a short time. With lower uncertainty in our estimate of activity rate, we can detect small changes in activity rate–we have high Signal Sensitivity. For rare activities, we have to wait longer to count enough activities to estimate the activity rate to the desired level of confidence to detect a small signal. These trade offs are summarized in the Table 3.

For reference, we assemble the parameter definitions and a table summarizing trade offs. Table 2 summarizes the parameters of the model. Table 3 summarizes the trade offs in parameters for a given target.

4 Example Calculations

Below are example calculations to make these ideas concrete and illustrate the use of the lookup tables.

Parameter	Symbol	Definition
Activity count	N	Number of activities in time T
Sample time	T	Duration of observation
Activity rate	r	Number of activities per time
Ave. activity rate	$\bar{r} = N/T$	Our estimate of average activity rate
Rate variability	δr	Uncertainty of rate estimate
Confidence parameter	α	Confidence level is $1 - \alpha$
Signal sensitivity	Δr	Change in activity rate
Signal latency	T_l	Time required to detect Δr
Signal confidence	η	Rate signal criteria factor
Bucket size	Δt	Predetermined time for estimating rate
Number of buckets	k	Duration in number of buckets ($k = T/\Delta t$)
Sampling rate	S	Powertrack sampling operator (e.g. "sample:S")

Table 2: Summary of model parameters.

Goal	Possible Actions
Minimize activities	(i.e. decrease N) increase Δr (decrease signal sensitivity) decrease confidence (E.g. from 95% to 90%)
Increase signal sensitivity	(i.e. 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	(i.e. 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	(i.e. decrease 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 3: Summary of model trade offs.

4.1 Estimate the Optimal Powertrack Sampling Operator Value

The Sampling Operator, S, is the percent sample size extracted from the firehose. Selecting S is a process that often starts at $S = 100\%$. By carefully monitoring the number of activities, N, that are filtered through the rules, we get an estimate for $\bar{r}$.

Using 100% of the firehose for one minute, imagine that we observe $\bar{r} = 10$ activities. Further, say that we want to detect a change in activity rate of 20 activities per minute using $\eta = \frac{1}{3}$. What sample size should we extract from the firehose?

Imagine that for this problem, we are comfortable with a signal latency of 2 days–i.e. our system needs to react to signals in about 2 days. Given that we expect 10 activites per minute or $14,400$ activities per day, we need to meet our signal sensitivity criteria,

$$\frac{\delta r}{\bar{r}} = \text{ Relative Interval Size} = \frac{1}{3}\frac{(20-10)}{min}\frac{1min}{10activities} = 33\% \tag{11}$$

over this 2-day period. Table 1 requires 100 activities for a Relative Interval Size of 33%. Hence, instead of using 100% of the firehose, we could use $S = \frac{100}{28{,}000} << 1\%$.

4.2 Estimate Signal Latency

Imagine we observe rate of 10 activities per minute and we want to detect a change in activity rate of 20 activities per minute. How long does it take to identify a change in the activity rate as a signal with 90% confidence level? To calculate an answer, we will be using the signal sensitivity–Confidence Criteria, 5 and Confidence Interval Sizes from Table 1

- Calculating T_l
- Signal criteria factor $\eta = \frac{1}{3}$ – In this case we choose a criteria that reflects our wish to see fewer false positives.
- Signal Sensitvity $\frac{\eta \Delta r}{\bar{r}} = \frac{1}{3}\frac{(20-10)}{min}\frac{1min}{10activities} = 33\%$
- Confidence Interval Size at $N = 10$ is 11.54.

It is clear that we cannot detect a change in activity rate of 10 activities/minute by measuring for only 1 min. Notice that our criteria is not fulfilled:

$$\frac{\delta r}{\bar{r}} = \frac{(11.54)}{10} \approx 115.4\% \nleq 33\% \tag{12}$$

The time T_l that it takes to observe this signal $\Delta r = 20$ with signal criteria factor of $\eta = \frac{1}{3}$ depends on the total number of activities N_t that we must observe to have a credible estimate of the activity rate. Because activities are infrequent, we will look up the confidence interval size, synonymous to δr, for small numbers of activities in Table 1. As N_t increases, the relative 90% confidence interval size narrows around the average rate, which can be seen through the decreasing realive interval value in Table 1. We need to find the value for N_t.

We can only detect a signal Δr when our signal criteria is fulfilled:

$$\frac{\delta r}{\bar{r}} = \text{ Relative Interval Size} = 33\% = \frac{\eta \Delta r}{\bar{r}} \tag{13}$$

You can look up the required Relative Interval Size in Table $1, 100\%/3 = 33\%$ to see that we need to observe at least 100 events on average to reach our criteria. Therefore, $T_l = 10 minutes$ because we will have observed 100 activities in 10 minutes given $r_b ar = \frac{10 activities}{1 min}$. That is, we must observe 10 minutes of activities to detect our desired signal.

4.3 Estimate signal sensitivity

Suppose we would like to determine the magnitude of a signal change needed to classify it as significant. As shown in Equation 5, classifying a signal Δr as significant depends on the choice of criteria factor η and the observation parameters that determine the uncertainty δr. Specifically, we will need to choose a criteria factor η and confidence level $(1 - \alpha)$, and our observation will be characterized by total activity count N and total time T.

Let us assume we have decided to classify as significant a signal with $\eta = \frac{1}{10}$, or $\delta r = \frac{1}{10}\Delta r$. Furthermore, we have chosen a 90% confidence interval ($\alpha = 0.1$), and observed N=10,000 activities over a period of T=1 minute (60 seconds) for an estimated activity rate of $\bar{r}$ = 167 s^{-1}. We next use Equation 8 to calculate the interval of activities for our 90% confidence level, and divide by observation period T to obtain the corresponding minimum significant activity rate δr = 5 s^{-1}. Recall, however, that we have also specified a criteria factor $\eta = 10$. Therefore, in this example, in order to classify the change in rate as significant, we must observe a change at the level of $\Delta r = \frac{1}{\eta}\delta r = 10(5 \text{ s}^{-1}) = 50 \text{ s}^{-1}$. For an increasing activity rate, this corresponds to a total activity rate of 167 s^{-1} + 50 s^{-1} = 217 s^{-1}. For a decreasing rate, 117 s^{-1}.

5 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.

6 References