Real Time Social Data Sampling

Scott Hendrickson, Brian Lehman, and Joshua Montague

Gnip, Inc.

May 24, 2013

1 Introduction

Occasionally, we want to calculate social activity rates to identify changes in activity rate that signal some change in the data. And we want to understand the uncertainty in our estimates or possibly design for a given level of certainty.

This means we need to calculate the relationship between activity counts, observation time, 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.

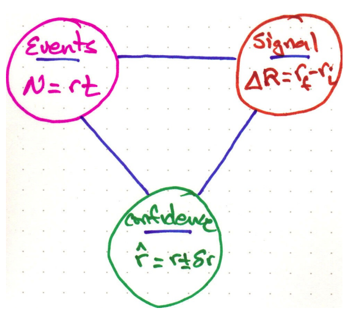

Figure 1: The number of events is a product of the rate and time. Signal is the change in event rate we want to detect. Confidence is the way to characterize the uncertainty in our estimate of rate. These parameters are not independent: choose any two and this paper describes how to calculate the third.

For many topics of interest, social media streams contain data with significant volume and frequency to create reliable, high-resolution signals in the time series over short observation times. But for the less common activities (e.g. blog posts on curling), we need to take some care in identifying the observation time, signal sensitivity and confidence required. These three quantities are related;

typically, we can choose two and the third will be defined for the stream we are watching.

A small complexity is that you may also choose to sample a known fraction of the total social activities to control costs or to scale down analysis. This effectively decreases the rate of activities and so enters into our calculations of event frequencies. There are a handful of related questions you that may come up around sampling and signals. In the next section, we list some of these questions. In the following section, we will look at filtering and sample in a little more detail. The middle of this white paper builds up the calculation method required to answer these questions. Finally, we provide some example calculations.

1.1 Motivating Questions

This white paper addresses how to answer questions related to social media activity time series sampling, signal, and confidence. Here are a few questions you may encounter:

- 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?
- 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 many buckets do I aggregate to optimize the trade-off between signal sensitivity and signal latency?
- 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?

1.2 Sample Operators and Order of Filters

It may be useful to take a detour in to sampling before moving the the core of the calculation. If not, just 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 an intermediate, manageable size for further 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 a subset of the firehose, likely to have explicit relation to the Super Bowl.

In the case of a major event like the Super Bowl, this filtered stream may represent a very large number of tweets. Therefore, 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 matched tweets. For this example, we add "sample:12" to the PowerTrack rule.

It is useful to know a little about how Gnip's sampling algorithms work to inform sampling decisions. Some key features of the Gnip sampling on premium streams including Tumblr, Twitter, Wordpress and Disqus:

1. 1% resolution on the sampling operator
2. Sample is deterministic and stable for near-rule matches. This means that you will get the same tweets for matches to the "super bowl" portion of the rules "super bowl sample:10" and "(super bowl OR superbowl) sample:10"
3. Progressively inclusive (i.e. the 2% stream includes the activities from the 1% stream plus an additional 1%, and so on)
4. Tweets are first selected from the firehose to reach the desired sampling rate and then filtered by keywords

Continuing with our Super Bowl example, combining filtering and sampling, we can calculate that an $x = 12\%$ sample of the firehose activities, $N_f = 500\text{M/day}$, filtered on activities containing "superbowl" "super bowl" and "contains:xlvii" (assume these rules return $y = 5\%$ of the stream to make this a concrete example) will leave us with approximately

$$N_{obs} = xyN_f = 0.12(0.05)500\text{M} = 3\text{M} \tag{1}$$

observed 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, followed by sampling, Equation 1 would also be a reasonable estimate of N_{obs} on the time scale of the sampling calculation. However, this process would require relaxing properties 2 and 3 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 short term behavior because it amplifies the effects of statistical variations in the short term.

2 Signal

In many situations, the main question is: *"How many events (or for how long) must we observe in order to detect a signal change with a specified level of certainty?"* Answering this question requires an understanding of the trade offs between sampling time, activity rates, and signal strength. The first step to working out this relationship is to define signal a little more carefully. We will define signal in terms of the activity rate.

2.1 Activity Rate

For all of these questions, it is useful to start by defining the activity rate in the common way.

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

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

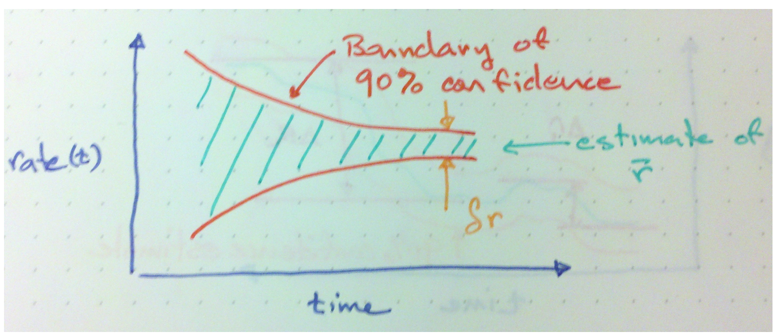

Figure 2: Uncertainty in the estimated rate of activities goes down as we observe more events.

So far, we have worked our way to the point where we are trying to 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%). However, we can invert this question to ask: given a level of confidence, how wide is the range of uncertainty about the rate estimate? This is a useful framing of the question because we are often trying to detect changes on activity rate as our signal. The concept of Signal Sensitivity adds another dimension to our problem–we can now trade Signal Sensitivity for Confidence in our rate estimate.

2.2 Signal Sensitivity

To see how we may do this, consider the trade off between our confidence in the estimated activity rate and the magnitude of the change in activity rate we can detect. If our estimate of activity rate is very uncertain, large variations are merely expected variations for infrequent events. Conversely, if your confidence in our estimate of activity rate is high (say, better than 95%, for example) then we can detect small relative changes in activity rate. Therefore, we say we observe a valid signal in a time series when the activity rate has changed by more than the rate Signal Sensitivity, Δr. That is,

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

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 achieve Signal Sensitivity Δr.

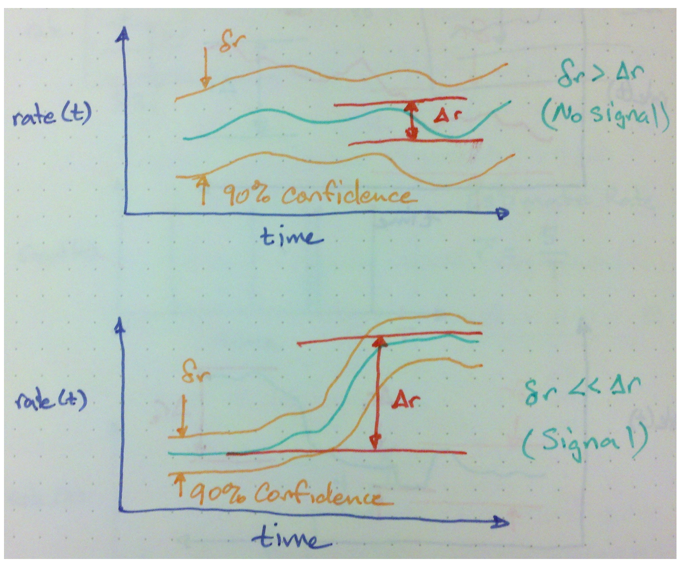

Figure 3: A signal can be detected when the change in rate is greater than the uncertainty in the rate estimates.

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 of counts in any given period will be distributed about the true mean of the distribution. 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 for a longer time 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: ¿¿¿¿¿¿¿ 19d155f19da78b2b44fa653cdc4b6561f9b7dd5e

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

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

The relative variation of the observed number from the average number will decrease with an increasing number events (time or increasing rate (the number of activities counted). The help quantify this inequality, we introduce a criteria factor η that quantifies how much larger the change in rate should be, relative to the uncertainty in the rate estimate:

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

where $\eta \geq 1$.

This criteria represents trade offs between the number of 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 Signal Sensitivity (the size of the rate change we can detect), Δr. These trade offs are summarized in the Table 2.

2.4 Summary of Parameters and Trade offs

For reference, we assamble the parameters definitions and a table summarizing trade off in this section before moving into the details of calculating confidence intervals for rare events in the next section. Table 1 summarizes the parameters of the model. Table 2 summarizes the trade offs in parameters for a given target.

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 a Poission distribution of activity counts over time.

Parameter	Definition
N	Number of activities in time T
T	Duration of observation
r	Activity rate
$\bar{r} = N/T$	Estimate of average activity rate
δr	Uncertainty of rate estimate
α	Confidence
Δr	Signal Sensitivity: Change in activity rate
T_l	Signal latency, time to detect Δr
η	Rate signal criteria factor
Δt	Bucket size (for bucketed data where $\Delta t < T$)
k	Duration in number of buckets ($k = T/\Delta t$)

Table 1: Summary of model parameters.

Target	Actions
Minimize Activities	(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 2: Summary of model trade offs.

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.

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 3: Confidence intervals given the number of events counted N in unit time T. Rate confidence range is $\delta N/T$.

3.2 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 distribution with confidence level $1-\alpha$ are given by

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

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 3

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

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. That is, we can now calculate any one of Signal Sensitivity, Signal Latency, activity rate, and confidence level given all of 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 confidenc interval for large numbers of actvities and has some observations about bucketed activity counts. You can skip this section and move to calculations in many cases.

3.3.1 Less-Rare 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,

$$\bar{r} - 1.95\sqrt{\bar{r}/n} \leq \hat{r} \leq \bar{r} + 1.95\sqrt{\bar{r}/n} \tag{9}$$

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.

4 Example Calculations

We present some example calculations to make this concrete and illustrate the use of the lookup tables.

4.1 Estimate the Optimal Powertrack Sampling Operator Value

To do ...

4.2 Estimate Signal Latency

Imagine we observe rate of 10 activities per minute and we want to detect a change of activity rate to 20 activitities 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 3

- Calculating T_l
- Signal criteria factor $\eta = 3$ – In this case we choose a criteria that reflects our wish to see fewer false positives.
- Signal Sensitvity $\frac{\eta \Delta r}{\bar{r}} = \frac{20-10}{min} \frac{1min}{10activities} = 100\%$
- Confidence Interval Size at $\bar{r} = 10$ is 11.54.

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

$$\frac{\eta \delta r}{\bar{r}} = \frac{3(11.54)}{10} \approx 346\% \nleq 100\% \tag{11}$$

The time T_l that it takes to observe this signal Δr depends 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 for small numbers of activities this up in Table 3. As N_t increases, the relative 90% confidence Interval Size narrows around the average rate. We need to find the value for N at which:

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

$$\frac{\eta \delta r}{\bar{r}} = 3 \text{ Relative Interval Size} = 100\% \tag{12}$$

You can look up the required Relative Interval Size in Table 3, 100%/3 = 33% to see that we need to observe 100 events on average to reach our criteria. Therefore, $T_l = 10$ min.

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 $\eta = 10$, or $\Delta r > 10\delta r$. 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 $\bar{r} = 167\ \mathrm{s}^{-1}$. 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 $\delta r = 5\ \mathrm{s}^{-1}$. 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 $\Delta r = \eta\delta r = 10(5\ \mathrm{s}^{-1}) = 50\ \mathrm{s}^{-1}$. For an increasing activity rate, this corresponds to a total activity rate of $167\ \mathrm{s}^{-1} + 50\ \mathrm{s}^{-1} = 217\ \mathrm{s}^{-1}$. For a decreasing rate, $117\ \mathrm{s}^{-1}$.

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.