Real-Time Social Data Sampling

Scott Hendrickson, Brian Lehman, and Joshua Montague

Gnip, Inc.

September 3, 2013

1 Introduction

In the world of real-time social data, we are typically observing a series of activities during some period of time and are interested in identifying significant changes in the corresponding activity rate. Such changes may be signals of emerging events or conversations. In this work, we hope to assist current and potential Gnip customers in

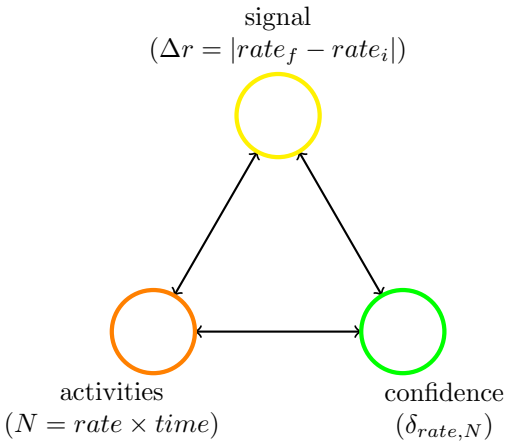


Figure 1: The three parameters used in classifying signal from real-time data: signal size, total activities, and confidence. Signal size is a change in activity rate, the total number of activities observed is a function of observation time, and confidence is that of a reported activity rate. These parameters are not simultaneously independent; we can choose or measure any two and then calculate the third.

we would like to quantify our ability to identify these kinds of signals. Figure 1 illustrates (schematically) the three main parameters involved in such calculations: signal size, total activities, and confidence. The three parameters are not simultaneously independent; we can choose or measure two of them – possibly based on our particular use-case – and they will determine the third. If chosen as *ad hoc* parameters, respectively, the signal size is a difference in activity rates between two observation times, the number of total activities is a function of the observation time and underlying instantaneous activity rate, and the confidence is a measure of statistical uncertainty in the activity rate e.g. a 95% confidence interval.

For popular topics, social media streams contain a sufficient rate of activities e.g. blog posts or Tweets to create reliable, high-resolution signals in short observation times. However, less popular topics with infrequent activities require additional effort in order to adequately determine the number of activities, signal sensitivity and confidence level appropriate for the situation.

In Section 1.1, we begin with examples of questions that may arise regarding the sampling of real-time social data. In Section 1.2, we outline some of the mechanisms by which a user can manage their data collection from Gnip, specifically. In Sections 2 and 3, we outline some of the mathematical framework for calculations of activity rate and sampling statistics. Finally, in Section 4, we work through some example calculations.

1.1 Motivating Questions

Below are examples of questions regarding activity rate, signal, and confidence level that might motivate the use of this whitepaper. The following Sections and example calculations will hopefully help to answer these kinds of questions.

- The activity rate has doubled from five counts to ten counts between two of my measurement buckets. Is this change significant, or is this expected variation e.g. due to low-frequency events?
- I want to minimize the total number activities that I consume (for reasons of cost, storage, etc). How can I do this while still detecting a factor of two change in activity rate in 1 hour?
- How long should I count activities to detect a change in signal of 5%?
- How do I describe the trade-off between signal latency and activity rate uncertainty?
- How do I define confidence levels on activity rate estimates for a time series with only twenty 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 activity rate estimate is accurate for each bucket?

1.2 Filtering and Sampling

Rapid growth in the use of social media has led to a large amount of data becoming available from many different sources; Twitter users alone produce approximately 500 million activities per day. In addition to Twitter, Gnip provides access to data from Tumblr, Foursquare, WordPress, Disqus, IntenseDebate, StockTwits, Estimize, Newsgator, as well as easy access to public API data from more than a dozen additional sources. In order to make this large volume of data more manageable, Gnip customers can take advantage of two approaches to sample from our firehoses, reduce overall data consumption, and focus on activities of interest: PowerTrack filtering[1] and sampling[2].

Gnip's PowerTrack operators allow for filtering of a publisher firehose on keywords or fields that are relevant to the topic in which you are interested. For example, if you are interested in tracking the Super Bowl (the American football event), you might start with a broad stream defined by the keywords "superbowl" "super bowl" and "contains:xlvii", the latter being a substring match of the Roman numeral of the Super Bowl as might be seen in hashtags or short links. This should limit the social data stream to activities that are more closely related to the actual Super Bowl event.

In the case of a major event like the Super Bowl, the keyword-filtered firehose may still represent a very large number of activities – possibly more than we can store or process in real-time, or more than our budget allows. In this case, adding an additional sampling operator will reduce the delivered data to a known fraction of the firehose, upstream of any PowerTrack filtering. For example, in order to apply our above Super Bowl PowerTrack rules to a 12% sample of the firehose, we would use a rule such as: "(super bowl OR superbowl OR contains:xlvii) sample:12". Using a sampling filter effectively decreases the number and rate of delivered activities. Some key features of the sampling operator:

1. 1% resolution
2. Stable sampling rate (even on small time scales)
3. Deterministic sampling returns the same activities for near-rule matches. That is, the same Tweets are returned for matches to the "super bowl" portion of the rules "super bowl sample:12" and "(super bowl OR superbowl) sample:12".
4. Progressively inclusive sampling. That is, a 2% stream e.g. "sample:2" includes the exact activities from the 1% stream, plus an additional 1%.

[1] e.g. `http://gnip.com/twitter/power-track/`

[2] Twitter's filtered, rate-limited 1% streaming API provides a non-deterministic combination that is not suitable for many analytic tasks. See [Mor13].

5. Sampling operator precedes PowerTrack filtering. That is, activities are first selected from the full firehose to reach the desired sampling rate, then filtered by keywords. PowerTrack filtering is thus applied to a subset of activities that is still representative of the full firehose.

Consider the use of both the sampling operator and PowerTrack filtering in the case of this (fictitious) Super Bowl example. Assume our PowerTrack filtering rules would return $y = 5\%$ of the full firehose over the course of a day. Assume further that we choose to select an $x = 12\%$ sample of firehose activities to which our PowerTrack filter is applied. Given that the total number of firehose activities (at the time of writing) is about $N_{fh} = 500$ M per day, our filtering and sampling will leave us with approximately

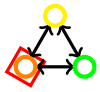
$$N_{observed} = xyN_{fh} = 0.12(0.05)500 \text{ M} = 3 \text{ M} \tag{1}$$

activities in this day.

The order of sampling and filtering is critical to maintaining the deterministic nature of subsequently filtered activities. Additionally, filtering prior to sampling would likely increase the duration of time needed to obtain an estimate of true activity rate, and would also inhibit any experiments that attempt to quantify a topic as a fractional component of the full firehose.

2 Sampling Parameters

In many situations, a simple 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 rate, and signal size.



2.1 Activity Rate

The average activity rate in a time bucket is calculated as

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

where N is the number of activities in a bucket of time length T. Due to the statistical variations in the number of activities in any given time interval, there exists an uncertainty in our estimate of this average rate. Figure 2 illustrates this idea that as we continue to observe additional activities, our confidence in the underlying activity rate grows; the bounds shown are those of a 95% confidence interval.

2.2 Signal Sensitivity

Higher underlying activity rates lead naturally to more certain estimates of said rate than lower ones. For a low activity rate, it is possible that small changes in

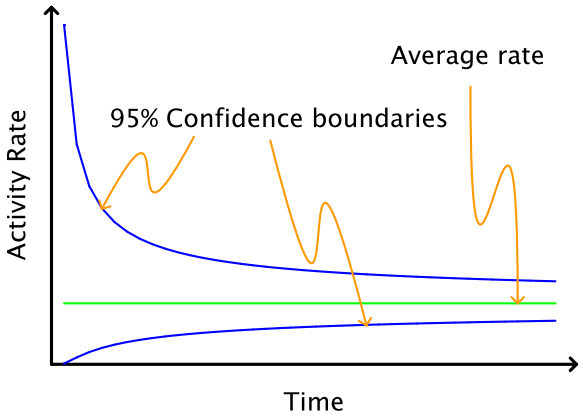
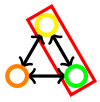


Figure 2: The 95% confidence interval (blue) representing the uncertainty in the estimated activity rate (green) decreases in size as we observe additional activities.

our estimated rate (calculated from one bucket to the next) will be inconclusive; the statistics of infrequent events lead to some amount of variation. In order to declare a valid signal, the variation due to e.g. infrequent events must be smaller than the thing we define as 'signal.' Therefore, we observe a valid signal in a time series when the activity rate between buckets has changed by more than the rate signal sensitivity, Δr, defined as

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

Each bucket size is defined by the difference between the variables t_f and t_i, the times at which the activity rate is measured. The associated time duration $T_l = t_f - t_i$ is the signal latency.



2.3 Signal Sensitivity–Confidence Criteria

If we assume we are interested in estimating the activity rate (c.f. Equation 2) of some form of steady-state process, the observed activities in any given period will be distributed about the long term mean. As for a typical Poisson process (discussed further in Section 3.1), the span of fixed-percentage confidence intervals decreases with an increasing number of observed activities, and as mentioned in Section 2.2, this uncertainty also inversely scales with the underlying activity rate.

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

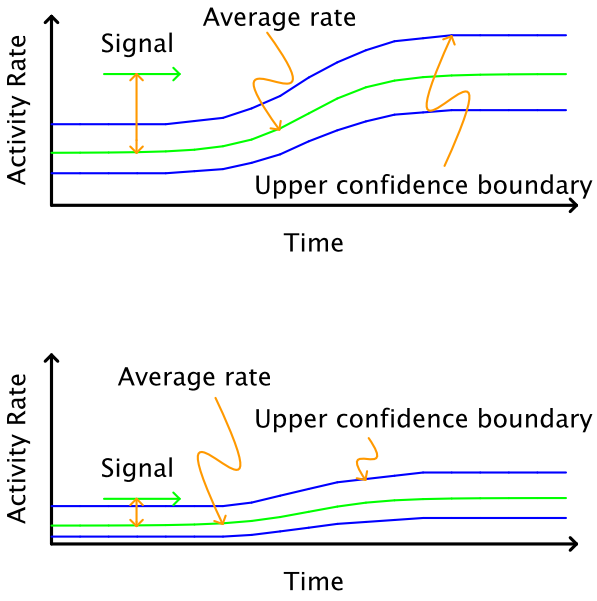


Figure 3: A change in rate from the start time (left) to the end time (right) is established when the change is equal to – or greater than – the uncertainty in the rate earlier estimate. The upper image shows the point at which we cross this threshold. Before this criteria is met, the change in rate remains within the uncertainty and observation of a signal is indeterminate. The lower image shows this situation.

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

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

To make this inequality a bit more concrete, we can introduce a criteria factor, η, which determines the relative size difference between our observed rate change (i.e. potential signal) and the uncertainty.

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

Flexibility in the choice of η allows us to prioritize high-certainty classification of signal (larger separation of observed signal and uncertainty bounds, typically requires longer observation time), or lower-certainty classification (smaller signal-uncertainty separation, typically faster).

% 2013-09-03

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 Poisson 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 [Geo12] 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.[3] 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 Table 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, it contains some observations about

[3] 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 (δn)	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 interval size is $\delta r = \delta N/T$. Note that the relative uncertainty goes down while the absolute size of the interval increases.

bucketed activity counts. You can skip this section and move to calculations if these ideas don't apply to your system.

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 T
Avg. activity rate	$\bar{r} = N/T$	Our estimate of average activity rate
Rate variability	δr	Uncertainty of rate estimate
Confidence	α	Confidence level is $1 - \alpha$
Signal sensitivity	$\Delta r = r_f - r_i$	Detectable change in activity rate
Signal latency	T_l	Time required to detect Δr
Signal confidence	η	Rate signal criteria multiplier factor (i.e. $\eta = 3$ means relative signal is $3\times$ random variations in sample)
Bucket	Δt	Predetermined time scale for estimating rate (probably already determined in your system)
Number of buckets	$k = T/\Delta t$	Duration expressed in number of buckets
Sampling rate	S	PowerTrack sampling operator (e.g. "sample:S")

Table 2: Summary of model parameters.

Goal	Possible Actions	Example
Minimize activities (i.e. decrease N)	increase Δr (decrease signal sensitivity); decrease confidence (E.g. from 95% to 90%); increase T_l (wait longer for the signal)	See example in Section 4.1 that illustrates long signal latency
Increase signal sensitivity (i.e. decrease Δr)	increase T (increase number of buckets (k); increase bucket size (Δt)); increase activity rate (r) by broadening filter or increase PowerTrack sampling	See example in Section 4.3 that illustrates sensitivity with high rate
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	See example in Section 4.2 that illustrates long signal latency
Decrease signal uncertainty (decrease η)	increase T (increase number of buckets (k); or increase bucket size (Δt)); increase activity counts (increase N, r) by broadening filter; increase PowerTrack sampling	See example in Section 4.2 that illustrates a calculation for small $\eta \leq 1$

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 activities per minute or 14400 activities per day, we need to meet our signal sensitivity criteria,

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

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:

$$\frac{\delta r}{\bar{r}} = \frac{(11.54)}{10} \approx 115.4\% \not\leq 33\% \qquad (12)$$

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 relative 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 $\bar{r} = \frac{10 \text{ activities}}{1 \text{ min}}$. 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 $\bar{r} = 167 \text{ s}^{-1}$. 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 $\delta r = 5 \text{ s}^{-1}$. 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 $167 \text{ s}^{-1} + 50 \text{ s}^{-1} = 217 \text{ s}^{-1}$. 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.

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.