New Measures of Clumpiness for Incidence Data

Yao Zhang

Eric T. Bradlow

Dylan S. Small

June 2, 2013

Yao Zhang is an associate at Credit Suisse, Eric T. Bradlow is the K. P. Chao Professor, Professor of Marketing, Statistics and Education, Co-Director of the Wharton Customer Analytics Initiative, Vice-Dean and Director of Wharton Doctoral Programs, and Dylan S. Small is Associate Professor of Statistics, at The Wharton School of the University of Pennsylvania. All correspondence on this manuscript should be addressed to Yao Zhang, yao.zhang@rocketmail.com, 215-833-5022; 11 Madison Avenue, New York NY, 10010.

Abstract

In recent years, growing attention has been placed on the increasing pattern of "clumpy data" in many empirical areas such as financial market microstructure, criminology and seismology, and digital media consumption to name just a few; but a well-defined and careful measurement of clumpiness has remained somewhat elusive. The related "hot hand" effect has long been a widespread belief in sports, and has triggered a branch of interesting research which could shed some light on this domain. However, since many concerns have been raised about the low power of existing "hot hand" significance tests, we propose a new class of clumpiness measures which are shown to have higher statistical power in extensive simulations under a wide variety of statistical models for repeated outcomes. Finally, an empirical study is provided by using a unique dataset obtained from Hulu.com, an increasingly popular video streaming provider. Our results provide evidence that the "clumpiness phenomena" is widely prevalent in digital content consumption, which supports the lore of "bingeability" of online content believed to exist today.

KEYWORDS: Clumpiness, test statistics, media consumption

1 Introduction

In recent years, data clumpiness has been widely observed in many areas. Intra-day stock transaction data are generally irregularly spaced where short durations tend to be followed by short durations, and long durations tend to be followed by long durations (Engle and Russell, 1998). Criminological research has shown that crime spreads through local environments via a contagion-like process (Bernasco and Nieuwbeerta, 2005). Similarly, the occurrence of an earthquake is well known to increase the likelihood of another earthquake nearby in space and time (Holden et al., 2003). The measurement and analysis of clumpy customer behavior has also started to gain more and more attention in marketing contexts (Schwartz et al., 2012), where technology has enabled fast and repeated consumption in a single session.

Data clumpiness is usually considered as irregular cluster(s) of activity gathered together, and many models have been developed in the literature to analyze those "clumpy" data, such as the hidden Markov, self-exciting point process and autoregressive conditional duration models. However, there is a lack of nonparametric, exploratory data analysis measures of clumpiness. In this paper, we provide researchers with a class of new measures to assess clumpiness. This class of measures is useful both as an exploratory data analysis tool and as a way to assess whether a model has fully captured the clumpiness in the data, e.g., by using the measure to carry out a posterior predictive check (Gelman et al., 1996).

Our development of the new class of clumpiness measures has been particularly motivated by practical problems in marketing. Companies care about customer value, investors care about firm value, and a better measurement of clumpiness enables companies to better estimate customer value, thus enabling superior targeting and pricing. In particular, ignoring clumpiness could result in a non-sufficient or even misunderstanding of a customer base. Let us imagine one customer who makes purchases regularly, where it is reasonable to expect that he would continue to make stable purchases in the future. For another customer whose purchasing activity is very clumpy, there is high uncertainty about him. He might either have lost

interest and never come back, or just simply have become "inactive" for a period of time and return. Once he returns, it is very likely that he would make a lot of purchases again. As a result, a nice/valid measure of clumpiness adds a new building block to profiling customers by which companies are able to deliver special promotions to "re-activate" the "right" customers. From the investor perspective, clumpiness means high growth potential but large risk, thus a clumpiness measure should serve as an important factor which should be taken into account when making investment decisions.

To highlight our new approach and its practicality, we utilize data on daily incidence video consumption from an online media company (Hulu.com) to apply our new measures, and assess the statistical power of our clumpiness measures. Since Hulu.com has been described as "binge-able"[1], its data is particularly appropriate for this purpose. Whether customers get hooked on a certain TV series watching episodes day-after-day, or they get initially excited about access to so much content but soon forget about it for a while, binges of activity do appear with regularity at Hulu.com. To illustrate this point in an exploratory way (but without loss of generality), we provide an exploratory plot in the figure below containing a couple of representative paths which describe the activity of actual registered customers of Hulu.com for 120 days starting March 1, 2009. The "•" indicates that the individual visited Hulu.com to view at least one video that day (incidence data):

Customers A and B exhibit relatively steady consumption, but customers C, D, E and F do not; instead, their viewing rates likely change, suggesting the need for a model that can handle non-stationarity. One simple but effective story of changing rates is a "buy till you die" story, which proposes constant repeat transaction rates until attrition (Fader et al., 2005). This process may account for customer C but definitely not for customers D, E and F, because they seem to exhibit clumpiness which can not be captured by a"buy till you die" model. Activity followed by a period of little or no activity may not be a signal of "consumer death" but rather a sign of "hot, then cold, then hot again." As shown in Figure 1, the latter case seems more

[1] "The Year in TV - The 2008 Culture Awards" New York Magazine, December 2008, (accessed 31 July 2011), [available at http://nymag.com/arts/cultureawards/2008/52737/index1.html].

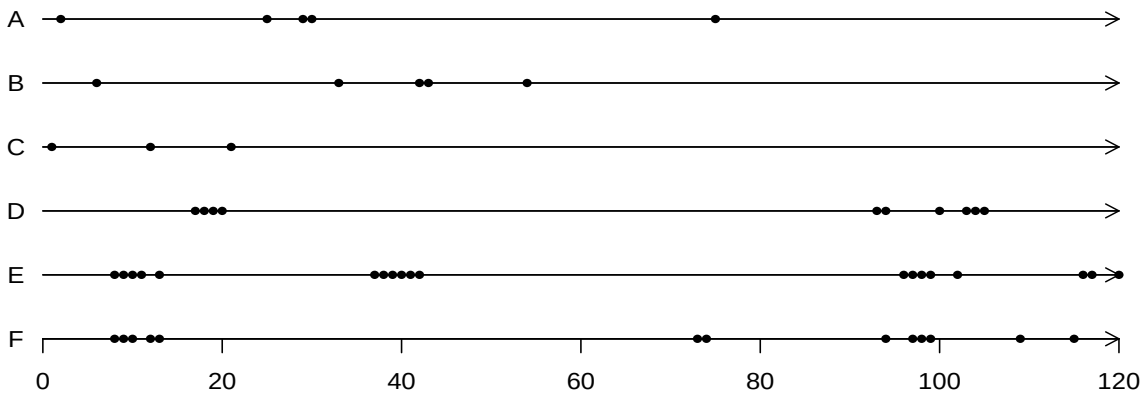


Figure 1: Examples of Representative Customers from Hulu.com

likely for "clumpy" customers. Thus, a careful measurement of clumpiness could serve as an indication of the right model to use, before the model is ever fit. This can be of great value to researchers who value parsimony, but also appropriate model fit.

In order to build "good" measures of clumpiness, we can borrow some insights from the "hot hand" effect in sports, a widespread belief that success breeds success and failure breeds failure. It is defined as when the performance of a player during a particular period is significantly better than could be expected on the basis of the player's overall record (Gilovich et al. 1985, 1989a, 1989b)(later "GVT" refers to these three articles as a group). Although many statistics have been introduced to investigate the long-lasting debate about whether the "hot hand" effect actually exists, most of them are criticized for their low power of detecting the effect. Therefore we propose a new class of clumpiness measures which we show to have higher power under various familiar non-stationary models that have appeared in the literature. More importantly, the new measures are robust to the "type" of clumpiness, i.e appropriate when there is sequential dependence ("an omnipresent effect"), but also non-stationarity("an occasional effect") which does not hold for extant measures.

Thus, as an overarching goal, we intend to bring greater attention to clumpiness which should be considered as an important feature to both researchers and practitioners as a measure of key strategic importance, and to provide a set of powerful statistics with desirable properties.

The remainder of this article is organized as follows. In Section 2, we begin with a brief literature review of related "hot hand" measures. In Section 3, we propose an outline summarizing the desired properties that a clumpiness measure should have, and then introduce a class of new measures with accompanying statistical test procedures. This is followed in Section 4 by a large simulation study in which we carefully examine the power of the proposed measures, and existing ones, under various alternative model assumptions. In Section 5, we turn to a detailed empirical analysis of the Hulu.com data which provides strong evidence in support of clumpiness in digital content consumption. Finally we conclude with a discussion of several additional issues that arise from this work. The paper also contains a number of empirical appendices that assess the sensitivity of our measures to reversing coding, and timing effects when detecting clumpiness.

2 Hot Hand Measures

The "hot hand" effect in sports has gained much attention in cognitive psychology because it touches on an interesting topic about human perception and cognition of random and non-random events. A long-lasting debate about whether this effect exists was triggered by three articles by GVT[2]. They interpreted the hot hand belief as a manifestation about a statistically significant deviation from what is expected by the simple binomial model, namely, serial dependence or non-stationary, and argued the performance of professional basketball players does not provide persuasive evidence of "hot hands". Their studies have been widely cited, leading many to believe that the perception by fans and players that athletes sometimes get hot is an example of how people erroneously see patterns in random data.

An important question to be asked is whether GVT provide enough evidence to reject the existence of the "hot hand" phenomenon in basketball. Is it possible that the effect does exist, even with a substantial effect size, but traditional statistical tests are generally low in power thus not capable of detecting it? In GVT, there are four different types of statistics to analyze

[2]GVT was introduced in Section 1, which refers to three articles as a group.

each player's sequence of hits and misses: the proportion of successful shots, conditioned by the success or failure of the previous shot(s); the first-order correlation coefficient; the number of runs in the data using the Wald-Wolfowitz runs test; and the number of successful, moderately successful, and less successful series of consecutive shots, in blocks of four. To check stationarity, a chi-square test was performed on the successive blocks of four shots.

A number of researchers decisively raised concerns about the power of these significance tests. Studies by Dorsey-Palmateer and Smith (2004), Miyoshi (2000), Wardrop (1999), and a working paper by Frame et al. (2003), all demonstrate that standard hot hand tests are unable to detect non-stationarity and changes in the success probability. Due to all of the aforementioned concerns about the low power of significance tests, in the present paper we propose a new class of clumpiness measures which are shown to have higher statistical power under models which have appeared in the literature. A good review of the "hot hand" literature is given by Bar-Eli et al. (2006). A complete description of nine existing "hot hand" measures to be used as a benchmark is contained in Appendix A.

3 New Clumpiness Measures: Motivation

What is clumpiness? To the casual viewer, a sequence of (incidence) data is called clumpy when burst(s) of activities or clump(s) of events are observed, but this is not a proper statistical description. In the language of statistics, clumpiness indicates non-constant propensity, specifically temporary elevations of propensity— i.e. periods during which one event is more likely to occur than the average level. Hence, it shares almost the same idea with the "hot hand" effect. As a result, we will use the same operational definition of clumpiness as in the "hot hand" literature—serial dependence or non-stationarity, but interpret them in a different way.

In the "hot hand" literature, two common corresponding standards to identify clumpiness are: (a) that a player should perform in a way where "success breeds success" and (b) there

should be streaks of success in which performance has been raised that stand out from streaks due to luck. (a) is tested by examining conditional probabilities and (b) is tested by examining the number and length of runs. Let us look at them one by one. What does it mean by "success breeds success"? Does it simply mean one outcome of success would enhance propensity of the next trial? What about two outcomes of success or three outcomes? There is always a challenge in choosing how many previous outcomes should be included and which patterns of history should be compared.

Regarding the second standard, the question we raise is whether a clump of events is equivalent to a run? Is clumpiness only associated with consecutive successes? Does "being hot" have to do with the duration and frequency of streaks? Not necessarily. Clumpiness only suggests that the propensity should be larger than the average level over some period(s) of time, but it does not guarantee that successes have to occur consecutively. Thus one failure within a sequence of successes does not indicate that the "hot" period necessarily ended on that day. Two consecutive successes should not be treated too differently from the case when they are separated by one failure.

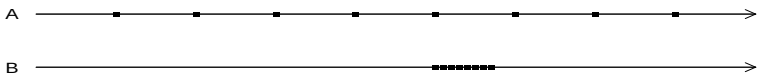
Those drawbacks heuristically explain why existing measures lack power in detecting the "hot hand" effect. However, if one considers the "hot hand" or clumpiness as not being simple sequential dependence between trials, or an extremely unlikely streak—then how else should it be described and measured? We discuss this next.

3.1 Desired Properties

Having stated the problem, we now proceed to build new measures. The usual way is to propose some measure and then present its properties; instead we do the reverse way — come up with a list of properties which a reasonable measure should have, and then present new measures which have those properties. The proposed properties are as follows:

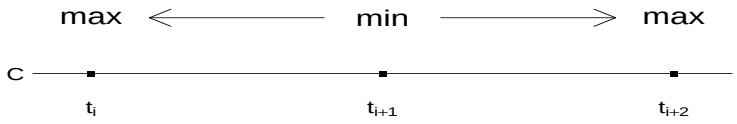
- **Minumum** : The measure should be the minimum if the events are equally spaced.

- **Maximum** : The measure should be the maximum if all the events are gathered together.



- **Continuity** : Shifting event times by a very small amount should only change the measure by a small amount.

- **Convergence** : As events move closer(away), the measure should increase(decrease).



The Minimum and Maximum properties are straightforward and easy to understand. Empirically, regardless of the number of events, it is natural to think of the case of all events clustered together as the most clumpy. On the other hand, one tends to think that the least clumpy case corresponds to that the events are uniformly spaced over the time line. Among the existing "hot hand" measures, most of them have these two properties (see Table 1).

We argue that Continuity is a necessary property. It would "make little sense" if a tiny change in the occurrence of events results in a dramatic move in the degree of clumpiness. As a result, neither runs-like measures which focus on the consecutiveness of events nor any hard threshold method satisfies this property.

The Convergence property, while somewhat imprecise, is the key expression of people's intuitive concept of clumpiness. Unfortunately, no existing hot hand measures have this property. Take the runs test for example, it only takes into account the number of runs, and totally ignores the information contained by the lengths of runs. As long as visits are not consecutive, it is considered not very clumpy even though all the events are already close to each other.

If there are only two events, it is very clear how the Convergence property would work. But for the case of over two events, a careful choice of definition is needed for "movement of

event" to make sure it does not conflict with the other properties. As the occurrence of one event moves towards another visit, it also move away from the other neighbour(if any). It is natural to use the minimal value of the distances to two neighbours as a description of local convergence. The smaller the value, the more convergent it is. Now the Convergence property can be rephrased as follows: given all the other events, the clumpiness measure should move in the same direction with local convergence. As a result, when one event goes from the left endpoint to right endpoint within its local interval, the measure behaves like a U-curve, and vice versa, which is illustrated above. In particular, the Convergence property could be listed as d(Clumpy)/d(IET) < 0 when IETs are made smaller.

For a reasonable measure of clumpiness, the Minimum and Maximum property sets its boundaries, the Continuity property governs its speed, and the Convergence property specifies its direction. These four properties combined provide a comprehensive description of the measure dynamics. Before introducing our new measures, let us summarize which desired properties the "hot hand" measures have. As displayed in Table 1, although most of them satisfy the Minimum and Maximum property, none of them have the Continuity or Convergence property.

Property	Minimum	Maximum	Continuity	Convergence
Runs	$\surd$	$\surd$	$\times$	$\times$
Test of stationary	$\times$	$\times$	$\times$	$\times$
Serial correlation	$\surd$	$\surd$	$\times$	$\times$
AC_1	$\surd$	$\surd$	$\times$	$\times$
AC_2	$\times$	$\surd$	$\times$	$\times$
AC_3	$\times$	$\surd$	$\times$	$\times$
S_0	$\surd$	$\surd$	$\times$	$\times$
S_1	$\surd$	$\surd$	$\times$	$\times$
S_2	$\surd$	$\surd$	$\times$	$\times$
New Measures	$\surd$	$\surd$	$\surd$	$\surd$

Table 1: Summary of properties

3.2 Construction of new measures

The complexity of patterns across a dataset, the complexity of patterns within a clump and the complexity of the background in which clumps are often embedded, make a universal measure of clumpiness difficult to achieve. Although the development of desired properties have added some useful rules to building new measures, it still leaves enough room for researchers to make their own measures case-by-case. Consequently, the present paper will not explore all the possible choices; instead, we only consider a special class by adding the following three conditions:

- **Function of inter-event times(IETs)**

  The discussion of desired properties prompts us to consider measures constructed using IETs, by which we can better control the measure dynamics. IETs not only provide flexible dependence structures, but also relax the reliance on consecutiveness of events. In addition, the measures using IETs are not restricted to the discrete case, and can be easily extended to assess continuous arrival times. Another nice property of choosing a class of clumpiness measures based on IETs is their invariance to forward or backward orderings of the observations, or any transformation that doesn't influence the IETs, or does so (given we scale by the range) in a totally proportional way.

- **Symmetry**

  Due to the Convergence property that the clumpiness measure should reach the minimum value when the event sits in the middle of its interval, we utilize symmetry to avoid a distortion.

- **Convexity**

  The Convergence property dictates the changing direction of the measure, but what about the changing rate? Clumpiness can be considered as a resistance index well known in the engineering literature. Let us imagine an experiment like this, a ball is tied between

two holders with two springs of equal length. As a result, the stable condition is the ball ending up in the middle. As the ball is pushed toward one holder, it requires more and more force to act on the ball. Specifically, the force grows quadratically. Relating this analogy to clumpiness, the changing rate of the measure should be increasing when one event is approaching one another. In other words, convexity is a reasonable assumption, which allows for a more sensitive detection of extremes.

Theorem 1 *Any convex and symmetric function of IETs satisfies all the desired properties.*

The details of the proof are in Appendix B.

Consequently, convex and symmetric functions provide a large pool of possible choices, and we choose four meaningful ones as our new measures for further illustration. Let n and N be the number of events and the number of trials for one sequence of incidence data in a given observation period. Denote by t_i and x_i the i_{th} occurrence of events and IETs respectively[3]. Since we want new measures irrespective of the length of the observation period, IETs scaled by N are used. In addition, in order to facilitate relatively easy comparison across the measures, all the new measures are normalized correspondingly by subtracting the minimum and dividing by the range (i.e. $\frac{\text{T - min}}{\text{max - min}}$, where T is the measure). In the paper, all the reported numbers are therefore rescaled values which are between 0 and 1.

1. Second moment: $L_2 = \sum_{i=1}^{n+1} x_i^2$

   It is one of the most widely used descriptors of a probability distribution, describing how far a set of numbers is spread out. Thus, it forms an important part of a systematic approach to distinguishing between probability distributions; and hence clumpiness.

2. Entropy: $H_p = \sum_{i=1}^{n+1} x_i \log(x_i)$.

   In information theory, entropy is a measure of the uncertainty associated with a random variable. It quantifies the expected value of the information contained in a message.

[3]For i=2,. . .,n, $x_i = t_i - t_{i-1}$; $x_1 = t_1$ and $x_{n+1} = N - t_n$. It follows that $\sum_{i=1}^{n+1} x_i = N$.

Although it is designed to describe a different story, its characterization and deep implication of disorder can provide us with a good measure of clumpiness.

3. Log utility[4]: $M = -\sum_{i=1}^{n+1} \log(x_i)$.

Log utility plays an important role in economics. A nice feature of using this measure is that a given percentage change in any of the IETs has the same effect on the measure. Hence, it "normalizes" the ranges being aggregated. A negative sign is added to make the measure move in the right direction.

4. Sum of 3 largest components: $C_3 = \sum_{i=1}^{3} x_{[r]}$,

where $x_{[r]}$ is the r-th largest ordered IET. In this manner, if the IETs are clumpy for some period of times, this measure picks it up.

All new measures have a positive relationship with clumpiness. In other words, a larger measure value means more clumpy. In our simulation study, we will further explain the relationship and distinction between these new measures, and discuss which measure works best under which scenarios. We will also discuss the sensitivity of clumpiness values with respect to the first and last IET values in Appendix C.

3.3 Evolution of Clumpiness: Intuition

Let us take a look at how the new measures behave when applied to real data. Here we select three customers from Hulu.com. All of them have five visits, which makes it easier to compare them. The details are displayed in Figure 2, where the x-axis represents the time line, and the y-axis shows the clumpiness measure (H_p is used here) when we "stand at that time point". Each visit is recorded on the x-axis.

[4]If the last event occurs at the end of observation period, there is a problem of taking the logarithm of zero. We set x_{n+1} equal to 1 instead of 0 to avoid this problem. This adjustment is reasonable since the special case rarely happens (2% of the time in our Hulu.com example), and also it does not make a significant difference after being scaled by N.

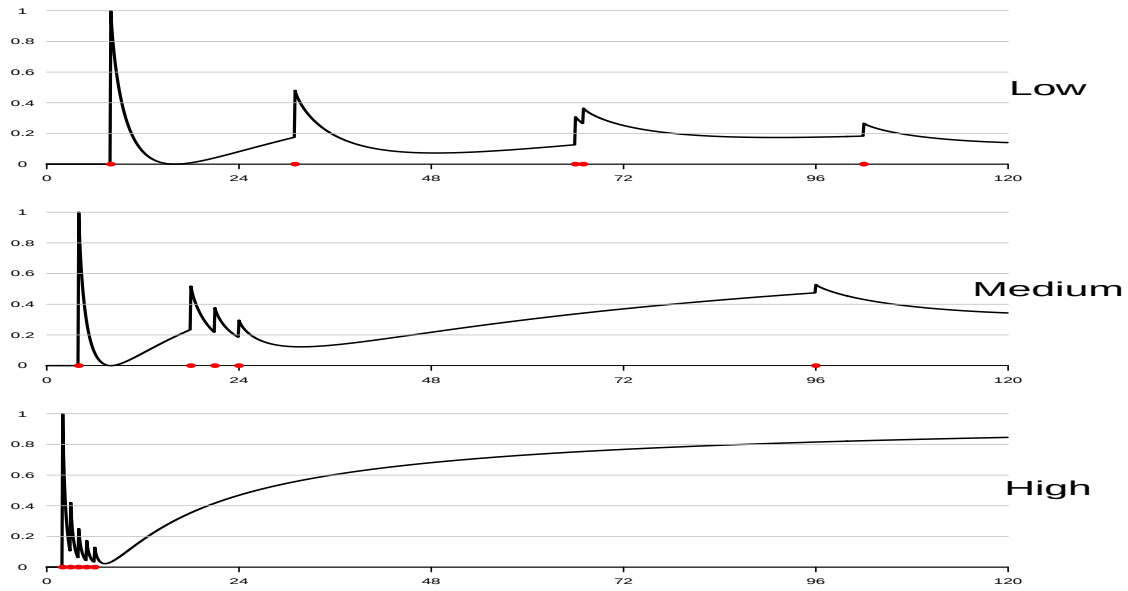


Figure 2: Evolution of clumpiness measure. Three customers all with 5 visits are selected, and observed over time. The black dots on the x axis indicate the observations, and the clumpiness measure (entropy value) as of each day is plotted on the y axis.

Based on the three curves' end points, it can be seen that the three customers are ranked from least to most clumpy at the end of the time period. But, that only tells a little bit of the story. It is more interesting to look at the evolution of the clumpiness measure as time goes on. Take the first customer for example, at the very beginning, there is no activity, thus no value is available for the clumpiness measure. When one visit suddenly occurs, the measure jumps to 1, the maximum value. It makes sense since this is the most clumpy case given only one visit. As time passes and no further activity is observed, we might expect that A is not as clumpy as we thought, thus the measure goes down. But when the period of no activity grows longer and longer, it seems that all activities of A are clustered during the very beginning, so the measure increases again.

In other words, we keep updating our judgment as more data come in. If no activity is observed, the clumpiness measure evolves gradually; if a visit occurs, it behaves like a "shock" since an observation of activity contains more information than no activity. When the number of visits is small, the influence of one shock is large since it changes the story dramatically.

But for a large number of visits, the influence becomes negligible.

Another fact worth mentioning is that the clumpiness measure can vary significantly over time. Take customer C for example, if we stand at day 10, all five visits look evenly distribution, so it is not clumpy at all. But when we move to day 120, all the visits are located during the very beginning, so it is considered instead to be very clumpy.

3.4 Test of Clumpiness

For a chosen clumpiness measure, a formal statistical test is needed to determine the significance of clumpiness. Larger values than would be expected under a model of randomness then provides an indication of clumpiness, with smaller values than expected indicating stability. Although the null hypothesis is random sampling without replacement, where n and N are known, the null distribution for the new clumpiness measures can not be generally derived in closed-form. To facilitate our analysis, Monte Carlo simulation is applied to compute the "Z table", the table of clumpiness critical values. Two different procedures are needed depending on the size of n, N. The detailed procedure is presented as follows for the general case:

1. Initialize the iteration number M and level of significance α.
2. Given n and N, for m in 1 to M :
   (a) Take a sample of size n from N days without replacement.
   (b) Calculate the clumpiness measure of the random sample.
3. Find the α-percentile as the critical value.

Once the "Z table" is generated, a test of clumpiness can be implemented for each sequence of data. The null hypothesis of "randomness" is rejected and this sequence is judged to be clumpy when the clumpiness measure is larger than the corresponding critical value.

For the case that n is much smaller than N, a useful approximation is available for those critical values. Using the well-known law of rare events, a Poisson distribution can be used as

an good approximation of the binomial distribution if n is sufficiently large and p is sufficiently small. Similarly, the Dirichlet distribution is the limiting case of the distribution of IETs in the random draws without replacement (under the Bernoulli distribution conditional on the total number of events). There is a rule of thumb stating that the approximation result is excellent if $N \geq 20$ and $n/N \leq 0.1$.

One major advantage of using a Dirichlet distribution for approximation is that only n is needed; hence a much simpler implementation. To illustrate the accuracy of the Dirichlet approximation, critical values for the measure of H_p are displayed in Figure 3, where N is set to be four different values. The x-axis is n, and the y-axis shows the critical values.

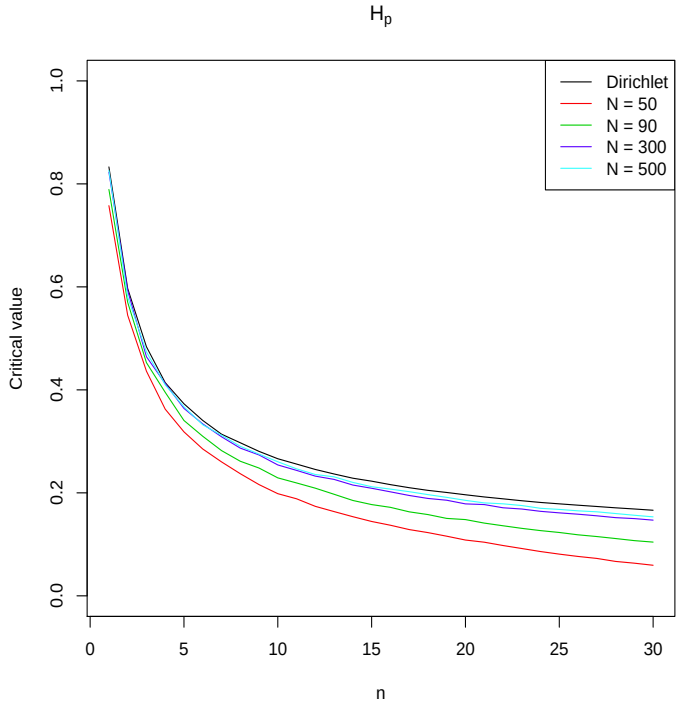


Figure 3: Dirichlet approximation. Critical values are plotted on the y axis when n and N are given. The black curve indicates the values when using Dirichlet approximation, and the four other curves are the "exact" values by varying N.

As is seen in Figure 3, the critical value is decreasing when the number of visits increases. In other words, for a subject with very few events observed, there is little information about

clumpiness, thus a high value of the test statistic is needed to reject. On the other hand, a lower cutoff is enough for a subject with lots of observed events.

Note as well from Figure 3 that it is more conservative i.e. lower type II error, to claim clumpiness if using the Dirichlet approximation as the critical value is lower for all n. Intuitively, since the Dirichlet distribution is continuous, there does not exist a smallest increment. Thus distances between events tend to be smaller under the Dirichlet model, and it tends to be "more clumpy" than its discrete equivalent. That is why higher standards exist for the Dirichlet approximation. In the next section, simulation results of clumpiness tests will be displayed using a variety of alternative models.

4 Simulation Study

In Section 4.1 and 4.2, we study the performance of thirteen test statistics: nine "hot hand" measures (as discussed in Appendix A), and four new measures proposed in Section 3. A careful discussion and comparison between them will be implemented under the null and eight alternative models which serve as a wide set of clumpy models taken from the extant literature. Due to the centrality of Hidden Markov Model (HMM) in the statistics literature in modelling clumpy data, we discuss its relation to our clumpiness measure separately in Section 4.3.

4.1 Size

We begin with a study of size of those tests (Type I error under the null). In this case, the null is the binomial model which assumes the trials are independent and the probability of success (Π) is constant. We set N equal to 120 to correspond to the Hulu.com data, and use 10,000 iterations for each measure and Binomial proportion Π. Percentages of iterations which are rejected at the level of 5% (Type I error) are reported in Table 2. The standard error for the proportion of Type I errors for a .05 level test in 10,000 iterations is .002.

Propensity	0.05	0.1	0.2	0.3	0.4
Runs	0.05	0.04	0.03	0.03	0.03
Test of stationary	0.01	0.06	0.07	0.06	0.05
Serial correlation	0.09	0.05	0.04	0.05	0.05
AC_1	0.01	0.02	0.03	0.04	0.04
AC_2	0.00	0.01	0.01	0.02	0.04
AC_3	0.00	0.00	0.01	0.01	0.02
S_0	0.02	0.02	0.02	0.03	0.03
S_1	0.02	0.02	0.02	0.03	0.03
S_2	0.02	0.02	0.03	0.03	0.03
L_2	0.05	0.05	0.05	0.05	0.05
H_p	0.06	0.05	0.05	0.04	0.05
M	0.06	0.05	0.05	0.05	0.05
C_3	0.06	0.06	0.06	0.06	0.06

Table 2: Size of tests. In each cell, we ran 10,000 iterations under the binomial model, and reported the Type I error rate.

As the table shows, most hot hand tests are biased, i.e., the true Type I error rate does not equal the correct .05 error rate. Some of the bias is due to the continuity problems, and some due to normal approximation. AC and S tend to make too few Type I errors, which indicates that they are not sufficiently calibrated. On the other hand, the new clumpiness measures are all stable and unbiased regardless of the underlying value of Π.

4.2 Power Comparison

A power analysis requires the specification of an alternative model that incorporates failure of the i.i.d assumption. As discussed in GVT, there are two types of violations of the binomial model that might be interpreted as a "hot hand" or clumpiness:

- Non-stationarity (violation of constant propensity)

  The probability of success is not constant and fluctuates over trials. For example, the visiting propensity of a user is 0.5 for 15 days, then suddenly increases, to 0.8 for 5 days, and then returns, to 0.5 for the next 10 days.

- Autocorrelation (violation of independence)

  Another violation of the binomial model is the sequential dependence of outcomes when the current propensity depends on the previous outcome. For example, many people believe that the rate of hitting a shot is higher after a hit than a miss.

We consider eight alternative models here, four for non-stationarity and four for autocorrelation. We select values of the parameters without loss of generality, to both span measurable values and to also provide insight to the empirical application in the Hulu.com data.

The four models of non-stationarity are as follows:

- One "Hot" spell, Wardrop (1999)

  The data are N Bernoulli trials with probability of success equal to p_B. At a random point, the probability of success increases to p_H for D more trials, and then returns to p_B for the rest of time. Here $N = 120, D = 10, p_B = 0.05$, and $p_H = 0.1$.

- Alternative "Hot/Cold" spells, Dorsey-Palmateer and Smith (2004)

  A sequence of trials has alternative "Hot/Cold" spells with size 10. "Hot" and "Cold" periods have respective probability of success. $p_H = 0.5$ and $p_C = 0.05$.

- Regime-shifting, Frame et al. (2003)

  The probability of success has a fixed probability of switching back and forth between "Hot" and "Cold" regimes. $p_{\text{switch}} = 0.1, p_H = 0.5$ and $p_C = 0.05$.

- Regime-shifting with changing switching, Sun (2004)

  A similar regime-shifting model is used but with a changing switching probability which

is randomly selected from (0.95 and 0.05) with equal chance for every 10 trials. $p_H = 0.5$ and $p_C = 0.05$.

10,000 iterations are used for each model, and the percentage of rejections are reported in Table 3. In column 2, the three best performing statistics are S_1, S_2 and M. It is reasonable since local measures S, allowing one or two failures embedded in a run, are specifically designed for this one "hot" spell model. But importantly, the newly constructed M is just as powerful as them. In column 3, H_p and M are both performing very well in detecting alternative "Hot/Cold" spells, as is the runs test. For the regime-shifting/HMM setting, the new measures are all powerful and better performing than "hot hand" statistics, as is seen in column 4. It is a similar case for column 5, except that none of the measures are very powerful, but our new measures are better performing.

Simulation	Wardrop (1999)	Dorsey-Palmateer(2004)	Frame et al. (2003)	Sun (2004)
Runs	0.45	**0.62**	0.54	0.06
Test of stationary	0.34	0.40	0.35	0.10
Serial correlation	0.45	0.59	0.51	0.12
AC_1	0.34	0.54	0.47	0.10
AC_2	0.29	0.45	0.39	0.14
AC_3	0.22	0.28	0.25	0.08
S_0	0.37	0.24	0.24	0.12
S_1	**0.49**	0.28	0.28	0.15
S_2	**0.58**	0.24	0.34	0.16
L_2	0.33	0.58	**0.74**	**0.27**
H_p	0.45	**0.78**	**0.79**	0.26
M	**0.55**	**0.84**	**0.79**	0.21
C_2	0.43	0.45	**0.74**	**0.31**

Table 3: Statistical power of clumpiness measures under a variety of non-stationary models

Regarding the autocorrelation case, Wardrop (1999) simulated data which are 100 dichotomous trials (we used 120 for consistency with the Hulu.com data). The first trial is a Bernoulli trial with propensity of p_B. After L or more consecutive successes, the probability of success increases to p_H; After L or more consecutive failures, the probability of success decreases to p_C; otherwise it remains p_B. In addition, we propose a slightly different model to enrich the

dependence structure. Among the previous 3 outcomes, if there are more than 2 successes, the probability of success increases to p_H; if there are all failures, the probability of success decreases to p_C; otherwise it remains p_B. $p_B = 0.5$, $p_C = 0.3$, and $p_H = 0.7$.

The simulation results are shown in Table 4. When $L = 1$, Serial correlation and AC_1 perform very well since they are exactly designed to detect order one dependence. The new measures, which are generally powerful, are also powerful in this case. For others cases, "hot hand" statistics are not as powerful as the new clumpiness measures with the exception of AC when (and only when) the right order is specified. This is critical as it indicates that even if AC statistics could be powerful for detecting autocorrelation, they perform well only when the order of dependence is correctly specified. If the dependence structure is not exactly correct, AC is misleading and much worse than our clumpiness measures.

L	1	2	3	new
Runs	0.98	0.71	0.35	0.44
Test of stationary	0.46	0.41	0.13	0.41
Serial correlation	**0.99**	0.68	0.31	0.41
AC_1	**0.99**	0.66	0.29	0.38
AC_2	0.90	**0.84**	0.39	0.56
AC_3	0.72	0.63	0.52	0.51
S_0	0.49	0.43	0.36	0.20
S_1	0.45	0.44	0.33	0.26
S_2	0.42	0.43	0.32	0.30
L_2	0.94	0.81	**0.58**	0.70
H_p	0.98	**0.85**	**0.59**	**0.71**
M	**0.99**	**0.85**	0.54	0.65
C_3	0.85	0.77	**0.62**	**0.77**

Table 4: Statistical power of clumpiness measures under autocorrelation of length L

Thus, our simulation revealed that the new class of clumpiness measures are more "powerful" in general to detect clumpiness. Moreover, the improvement should be very significant under many cases. Finally, the simulation shows that H_p and M are both robust measures.

4.3 Clumpiness Within a HMM

HMM, a statistical Markov model in which the system is assumed to be a Markov process with unobserved (hidden) states, is generally considered as an appropriate parametric model for clumpiness. We next study the performance of the new measures of clumpiness as a function of the HMM parameters. This analysis provides deep insight, we believe, into when one should expect the new set of clumpiness measures to be powerful.

We model the sequence of outcomes as 120 repeated Bernoulli trials. In particular,

$$y_t \sim \begin{cases} \text{Bernoulli}(p_1) & \text{if in State 1,} \quad Z_t = 1, \\ \\ \text{Bernoulli}(p_2) & \text{if in State 2,} \quad Z_t = 2. \end{cases} \tag{1}$$

where the latent-state variable, Z_t, indicates which state a subject is on each day. The vector of within-state propensities p_i contains the probabilities of success in States 1 and 2, respectively, and the full model parameterization is given by:

$$\mathbf{p} = (p_1,\ p_2), \quad \Theta = \begin{pmatrix} 1-\theta_{12} & \theta_{12} \\ \\ \theta_{21} & 1-\theta_{21} \end{pmatrix} \tag{2}$$

where the entry θ_{ij} represents the transition probability from state i to state j.

We set the baseline HMM with $p_1 = 0.5,\ p_2 = 0.1,\ \theta_{12} = 0.1,\ \theta_{21} = 0.1$, and we then vary the parameters one at one time in an "experimental design" way. In other words, by changing one parameter while fixing the other three, we can see the direct impact on clumpiness. The rejection rates using H_p (the most powerful statistics amongst our new class found via simulation) are displayed in Figure 4:

As discussed in the introduction section, when the propensities in the two HMM states are close, a HMM is approximately a simple Bernoulli process and no clumpy patterns exist as the

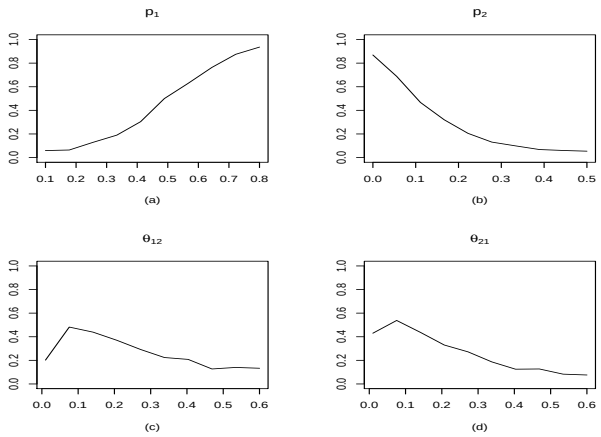


Figure 4: Power under various parameter values of a hidden Markov model. The power of the test (rejection rate) is plotted on the y axis by varying one parameter at a time.

measure H_p shows (panel 4(a), H_p at $p_1 = 0.1$). But as the difference increases between states 1 and 2, more and more clumpiness is identified. In addition, θ_{12} and θ_{21} are both inversely correlated with the lengths of staying in states 1 and 2; either too long or too short periods can essentially reduce a HMM to simple Bernoulli model, which is demonstrated by the simulation results. This is why we see an inverted-U shape in the bottom two panels (c) and (d) of Figure 4.

5 Empirical analysis: Hulu.com

As mentioned throughout, Hulu.com is an online media company, along with YouTube, that are the top two online video content properties in the U.S. Accounting for over 15 billion and nearly 50% of all videos viewed in the U.S. (comScore, 2010), understanding behavior and the degree to which it shows clumpiness, is a large and strategic problem. We analyze a cohort of 563 randomly selected customers who registered for Hulu.com during the last two weeks of February 2009, and this cohort is representative of the entire population. It is important to consider a cohort of customers so that other time-varying factors that could drive clumpiness are controlled for/constant across viewers. The data structure is therefore a matrix

of individuals-by-days (of dimension 563 by 120), where the entries are 1s or 0s indicating whether that individual visited Hulu and watched at least one video on that day. Across the 563 customers in 120 days, there were 4840 total views (i.e. days of viewing), and they watched videos on an average of 8.7 out of 120 days. We note that due to sporadic viewing, i.e. an incidence rate of 7.2%, having powerful clumpiness measures is even more crucial.

We then applied the aforementioned thirteen measures to the cohort of viewers. Both hot hand measures and new measures show there are a lot of "clumpy" customers at Hulu.com. The percentage of clumpy customers by each measure is shown below in Figure 5[5]:

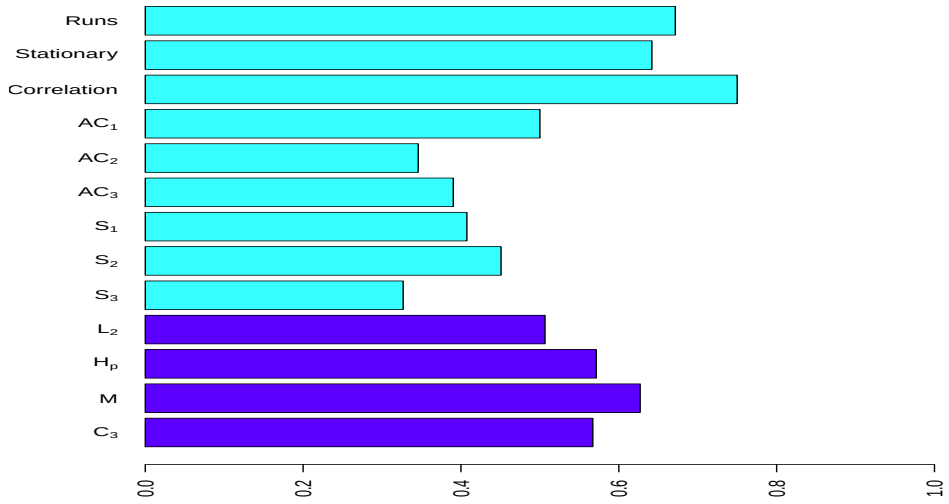


Figure 5: Percentages of clumpy users for a 2009 cohort of customers at Hulu.com

In order to compare the results between the "hot hand" measures and our new measures, we chose the runs test and H_p as the representative old and new measures respectively. Based on the two measures, 76.6% of the customers get the same "label"; among them, 50.5% are identified as being clumpy, and 26.1% as not. As for the customers with different "labels", 16.7% are defined to be clumpy by the runs test, and not by H_p; 6.8% have the opposite.

A concern that could be raised is that the measures of clumpiness should not depend on how one labels successes and failures; what is coded as a 0/1 is somewhat arbitrary. This

[5]Here 0.05 was chosen as threshold.

is a valid point in some cases, especially when the 0/1 variable are coin flips with fair coins. Although the values would change if we flip the labels, the look-up critical values would also change correspondingly with the new number of successes, so that our new measures still holds. On the other hand, we believe that label switching is not very relevant for media consumption (e.g. Hulu data), because visitation (incidence) is a rare event and recoding it differently would yield a markedly different dataset. However, to assess the sensitivity to coding, in Table 5 we computed the entropy values H_p by using the two labelling alternatives, and compared their test results. They produce the same result for 75% of the customers, and both estimate that the Hulu population consists of half clumpy users.

Percentage	Non-clumpy original	Clumpy original	Total
Non-clumpy flip	32%	13%	45%
Clumpy flip	12%	43%	55%
Total	44%	56%	100%

Table 5: Clumpiness classification comparison between label switching. The clumpiness measure is first computed by using the original label. Then we flip the label, and use the new label to calculate the measure. The resulting comparison is reported.

From Hulu's point of view, identifying approximately $\frac{1}{2}$ of the viewing population as clumpy regardless of the measure, suggests that this is likely to be a robust phenomenon. Furthermore, it suggests the potential importance of re-activating customers or stemming churn from those customers who are identified to be at risk as their future value to the firm due to advertising revenue, one of Hulu's primary revenue drivers, may be potentially significant. We discuss this and other issues in the conclusion section.

6 Conclusions and future research

Incidence data often have episodes with burst(s) of activity or large clumps of events. The extant literature has found clumpiness difficult to quantify and detect. As demonstrated in this research, the masking between clumpiness (potentially a good thing) and attrition (customer "death") make separating them an important research and practical problem.

This paper has proposed four properties a "reasonable" clumpiness measure should have, and then presented a new class of measures, and associated formal statistical tests. In a large simulation we demonstrate that the new clumpiness measures have higher statistical power, in general than existing "hot hand" measures for a wide class of statistical models. Moreover, two of the new measures H_p and M appear very robust and powerful in a wide class of settings. The nice thing about our measure is that once the data on inter-arrival times are obtained, the computation of clumpiness can even be done in a simple formula in standard software like Excel. It is one aspect, we believe, that makes our approach appealing and has the potential for it to be utilized broadly.

Finally, the empirical study on a sample data from Hulu.com indicated that there exists a significant amount of customers who show clumpiness. With that said, it is an interesting question for future research to apply our new clumpiness measures to a wide variety of data sets to assess whether the bingeable nature of the data for Hulu.com is widespread. Furthermore, the role that our clumpiness measures can play in sequentially identifying valuable customers, a broad topic of theoretical and applied research, is also a totally open avenue. In particular, what information is added above and beyond standard classification variables such as recency, frequency, and monetary value (RFM, a corner stone of classification for many businesses). Regardless of what is found in other empirical studies, we believe that our research is one necessary step in that one needs robust and powerful methods first, before considering their application to specific inferential problems.

References

Bar-Eli, M., Avugos, S., and Raab, M. (2006), "Twenty years of hot hand research: Review and critique," *Psychology of Sport and Exercise*, 7, 525–553.

Bernasco, W. and Nieuwbeerta, P. (2005), "How do residential burglars select target areas? A new approach to the analysis of criminal location choice," *British Journal of Criminology*,

44, 296–315.

comScore (2010), "U.S. online video rankings," *comScore Releases*.

Dorsey-Palmateer, R. and Smith, G. (2004), "Bowlers hot hands," *The American Statistician*, 58, 38–45.

Engle, R. F. and Russell, J. R. (1998), "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 66, 1127–1162.

Fader, P. S., Bruce, G. S. H., and Ka, L. L. (2005), "RFM and CLV: using iso-value curves for customer base analysis," *Journal of Marketing Research*, 42, 415–430.

Frame, D., Hughson, E., and Leach, J. C. (2003), "Runs, regimes, and rationality: The hot hand strikes back," .

Gelman, A., Meng, X. L., and Stern, H. S. (1996), "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, 6, 733–807.

Gilovich, T., Vallone, R., and Tversky, A. (1985), "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17, 295–314.

Holden, L., Sannan, S., and Bungum, H. (2003), "A stochastic marked point process model for earthquakes," *Natural Hazards and Earth System Sciences*, 3.

Miyoshi, H. (2000), "Is the hot-hands phenomenon a misperception of random events?" *Japanese Psychological Research*, 42, 128–133.

Schwartz, E. M., Bradlow, E. T., Fader, P. S., and Zhang, Y. (2012), "'Children of the HMM': Modeling longitudinal customer behavior at Hulu.com," *Wharton Marketing Department Working Paper*.

Sun, Y. (2004), "Detecting the hot hand: An alternative model," in *Proceedings of the 26th annual conference of the cognitive science society*, pp. 1279–1284.

Tversky, A. and Gilovich, T. (1989a), "The cold facts about the" hot hand" in basketball," *Chance*, 2, 16–21.

— (1989b), "The "hot hand": Statistical reality or cognitive illusion?" *Chance*, 2, 31–34.

Wardrop, R. (1999), "Statistical tests for the hot-hand in basketball in a controlled setting," *American Statistician*, 1, 1–20.