Can Television Advertising Impact Be Measured on the Web? Web Spike Response as a Possible Conversion Tracking System for Television

Brendan Kitts, Michael Bardaro, Dyng Au, Al Lee, Sawin Lee, Jon Borchardt, Craig Schwartz, John Sobieski, John Wadsworth-Drake
Adap.tv
801 Second Avenue, Suite 800
Seattle WA 98104
bkitts@adap.tv

ABSTRACT
Consumers are increasingly using internet-connected devices while watching television. This paper will show that it is possible to measure web activity bursts that peak about 13 seconds after the end of traditional TV ad broadcasts. By measuring this effect, we propose that it may be possible to deploy a web-based TV conversion tracking system that will work on TV systems.

Categories and Subject Descriptors
C.4 [**Computer Systems Organization**]: Performance of Systems – *Measurement techniques.*

General Terms
Measurement

Keywords
Television, advertisement, conversion, web, targeting, ad, cross-channel, multi-channel

1. INTRODUCTION
TV effects are notoriously difficult to measure. Unlike online advertising, there is no cookie to enable tracking of a user between an ad view and action. This has left television with critical problems with the ability to measure and optimize airings, and we believe is resulting in a large number of irrelevant and poorly targeted ads.

This paper will present experimental findings that suggest that this situation may be changing. We show that by aligning web activity with TV broadcasts in time and space and applying some signal processing techniques, it is possible to measure web activity bursts that peak about 13 seconds after traditional TV ad broadcasts. The implications for computational advertising are profound. Using this effect, we note that it may be possible to deploy a web-based TV conversion tracking system that will both work today on existing TV systems, and could be used for future IP-connected TV systems, thus enabling real-time optimization of television ad targeting for the first time.

ADKDD'14, August 24 – 27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2648584.2648591

2. PREVIOUS WORK
Zigmond and Stipp (2010) is the earliest report of television ads appearing to cause web spikes [16]. They noted the effect was causing spikes in the Google search engine, and that they could match the spikes temporally to TV airings.

Lewis and Reiley (2013) also reported on search spikes occurring due to SuperBowl ads on Yahoo!'s search engine [12]. They believed that this would only be visible on large airings, however we show later that it is possible to see activity on fairly small broadcasts.

Joo et. al. (2011) initially reported no web effect using AOL search data from 2006, queries aggregated to category level, and TV airing data aggregated to the day level [4], but then later reported a positive finding [5]. It is unclear whether Joo et. al.'s initial methodology, which was heavily aggregated, made it impossible to detect TV effects, or whether the magnitude of the effect has grown as second screen devices have grown.

Clipp (2011) noted significant, day-level web effects from television ads [2].

Kitts et. al. (2010) reported on web-spikes using advertiser weblogs and used demographics of the responders, geography, time recency to match to airings – however the method used a supervised training scheme to develop a parametric model [6]. The work we present here is non-parametric which provides significant advantages for faithfully reporting on effects without introducing expectations about how different variables impact web activity which may not hold over time.

2.1 Contribution
Much of the previous work has been concerned with showing that a short latency effect exists – the present paper will attempt to use that effect to provide measurement for real television campaigns. The present paper extends upon the previous research in several ways: (1) We describe algorithms that attribute web activity to television airings without the need for training or parametric assumptions. (2) We show how the attributed web response can be used in a TV ad targeting system to automatically target TV ads to the most responsive media. (3) We demonstrate the above in a live advertiser television campaign (4) We quantify the amount of TV activity that was correctly identified using an in-market experiment. Our general conclusion is that web spike response appears to be workable as a conversion tracking signal for television.

3. ANATOMY OF A WEB SPIKE

3.1 Alignment

We begin with an advertiser who is running television advertising, and who maintains a website.

Let $I(M(t_1,G,z_1))$ be the impressions associated with a media airing M at time t_1, timezone z_1 and geography G. Let $W(t_2,G,z_2)$ be a web traffic metric such as new visitors at time t_2, timezone z_2, and geography G. The timezones are represented as number of hours from GMT. The web server is located in time-zone z_2.

In order to align the TV broadcasts, we have to map the media airings onto the same time-zone, so that we're seeing media activity and web activity that both occurred at exactly the same time. We then bucket the web activity into the same geographic and time buckets for both web traffic and television airings. Let $W(T,G)$ be the sum of all web activity in web-server time-zone z_2, and $I(M(T,G))$ be the impressions from a media airing M within the same time and geography. We can calculate these as follows:

$$W(T,G) = \sum_{t \in T} W(t,G,z_2)$$

$$I\big(M(T,G)\big) = \sum_{t-(z2-z1) \in T} I\big(M(t-(z_2-z_1),G,z_1)\big)$$

Local broadcasts have well-defined times and time-zones. National broadcasts are typically recorded separately from local broadcasts, and require slightly different logic. Firstly, as for what time and time-zone to record for the national broadcast, Live Feed Networks air using East Coast time. Dual feed networks re-broadcast their East Coast programming at the same local time for West Coast. For these we need to create a "Ghost Airing" that is a copy of the national airing, but with impressions scaled to West Coast population and East Coast airing by its proportion of population. A special geography *G=National* sums both local and national broadcasts so that we are capturing all broadcast activity.

After aligning the web activity with airing, it is possible to begin to characterize the shape of the web response curve.

In order to analyze the shape of the web response curve using a sub-minute time-grain, we used Lewis and Reiley (2013)'s Super Bowl Yahoo Search data which shows search queries for a brand-name every 10 seconds after a 30 second Super Bowl commercial aired for the same brand [12]. We used a log-normal distribution to fit the data (KS^*=0.10; υ=5.2322; σ=1.2102; Table 1). From the parameterized curve we find that TV searches occur rapidly after exposure of a TV ad. The peak response activity occurs just 13 seconds after the end of the ad. In comparison, after seeing a display ad for a retailer on a website (also reported in [12]), there is a peak for searching the same retailer's name at 23 seconds after exposure. Therefore, TV ads seem to drive faster response relative to display ads.

These times can be influenced by the amount of content on web pages hosting display ads, auto-completions on search engines, and other factors, and so our numbers are not definitive but merely provide some guidance as to the approximate time-scale involved.

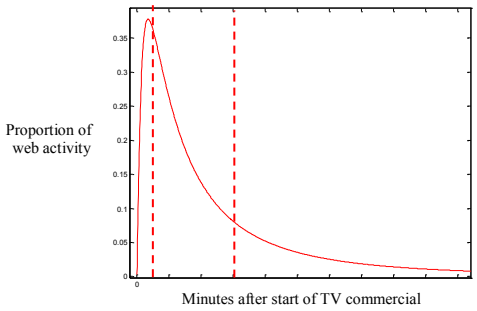


Figure 1b. Log-Normal fit to Yahoo searches received after the start of a Super Bowl commercial [12] with lines showing 1 and 6 minute marks. Measurement regimes should try to capture the peak activity which is within the first minute as this carries the highest signal-to-noise ratio. After 6 minutes 50% of activity has been recorded and signal-to-noise begins to drop significantly.

Table 1. Log-Normal fit to Super Bowl data from [12] Including Descriptive Statistics.

Statistic	Seconds between Display Ad and Search on featured Brand	Seconds between start of Super Bowl Ad and Search on featured Brand
KS*	0.07377	0.10356
Mode	21.62s	43.27s
Mean	136.96s	389.35s
Stdev	213.21s	710.06s
Sigma	1.1094	1.2102
Mu	4.3044	5.2322
Skew	8.4431	11.537
Kurtosis	246.78	562.26

*Kolmogorov-Smirnov statistic showing goodness of fit.

3.2 Techniques for Increasing Signal-to-Noise

Lewis and Reiley (2013) presented a statistical argument that web effects from TV – if present – would only be visible on very large broadcasts such as Super Bowl broadcasts [12]. The crux of their argument was that buying N small airings each at $1/N$ the price, would have to contend with the same degree of noise, but with $1/N$ the media effect strength, and only a square root of N gain in t-test power. Thus it would be a net loss to execute more, smaller airings, even with the same media budget. However Super Bowl airings are economically infeasible for most advertisers. How can we go from a $4 million dollar TV spot to a $400 spot and have a similar level of signal detection?

Several practical steps can be taken to improve the signal-to-noise ratio on small TV campaigns[1]:

The first is the use of geographic areas G – if TV is run in only a small number of geographic areas at higher media weight, then for a far less expensive campaign, a higher weight can be applied per capita without incurring the cost of a national campaign.

[1] Super Bowl ads cost $34.80 per thousand impressions (CPM), where-as the national TV media CPM average was only $6.60 based in 2012. Therefore small TV spots can buy about 5 times more impressions for the same budget, which also increases signal amplitude.

The second is to localize the effects temporally. At 13 seconds after the end of a commercial broadcast, signal-to-noise ratio is at its maximum – at that time, more visitors on the site are newly arrived due to the recent TV airing than background. The time window we use needs to be fine enough to sample this high signal-to-noise region of the curve. For example, the difference between a 30 second sampling window and 1 day is a 1,800x reduction in signal mean. Therefore localization in time with short time windows is critical for achieving a temporary signal-to-noise superiority.

A third method for increasing signal to noise is to eliminate robotic activity. Bots tend to produce large volumes of traffic and can completely mask human activity. Methods for eliminating bot activity vary, but because many bots are designed to avoid detection, one good method is to use a well-supported system such as Google Analytics to capture and extract data, since this is supported by Google's bot filtration systems [11].

The fourth method is to measure the targeting of the television ads. Targetedness is a new concept in television which measures how well an advertisement matches the audience. We have shown that untargeted ads can produce almost no lift at all. Targeted impressions, rather than simply viewers, can be used to estimate more reliable web spike results [8].

The fifth method is to filter to subsets of traffic that have a higher prevalence of the television behavior that we're trying to isolate. Real-time responses to a TV ad are occurring from people watching the broadcast live, and tend to require a tablet or mobile device. Traffic is more likely to visit the homepage, rather than a deep-linked page, and the activity is likely from new visitors who haven't been on the site before. By focusing on this class of traffic we can eliminate more organic background activity, leaving a higher signal to noise ratio for the TV generated traffic. A list of the filters that we test in this paper are below:

New Cookie: Visitors who have been assigned a cookie for the first time. This eliminates traffic that has visited the site before, and so increases the magnitude of the web spike compared to background activity

Homepage Requests: Visitors who are requesting the homepage are also more likely to be those who are navigating to the site for the first time in response to an advertisement.

NULL Referrer Requests: Requests with NULL Referrer are requests where the person is not known to have navigated from a search engine, deep link, or another method. This is often people who have directly typed the URL into their web browser to access the site for the first time.

Mobile and Tablet UserAgents: Requests with a mobile or tablet user agent string have a much greater response to television ads.

4. LIVE TV ADVERTISER EXPERIMENT

We will now analyze a live advertiser to determine if web spike responses can be detected on very small airings. The data are from a live TV campaign that ran from 2/11/2013 to 7/9/2013 with 35,296 airings. The airings were detected using digital watermarks. The average spot cost was $143 per airing –nowhere near Super Bowl airings costs.

Web activity was aggregated from advertiser web logs in 5 minute time buckets. Figure 1 shows new visitor web activity using 5 minute time intervals and national geography, from 1/20/2013 to 2/25/2013. The timeseries starts to become more spikey around 2/11/2013. The reason for those spikes can be seen in close-up in Figure 2. Figure 3 shows that when we reveal TV media, that the airings that are producing the spikes line up exactly with the web spikes.

In order to measure webspikes quantitatively, we provided several measurements for web spike magnitude in Table 2:

1. Additional traffic during 5 minute airing is the web traffic in native units that was generated during the TV airing.
2. Percentage increase over base is the during-airing reading divided by base activity. Base activity is equal to the reading in the 30 minutes prior to the airing, and where no other airing occurs within those 30 minutes.
3. tImp Correlation is the correlation coefficient between the metric and TV airing's targeted impressions. This may produce a more accurate measure since the other metrics don't account for targeting.
4. Increase per million impressions (wpmm) is the change in the web traffic multiplied by one million divided by number of television impressions.
5. Mean and Var are the mean and variance of the web traffic in native units. A p-value is shown based on a t-test for the instantaneous spike percentage change, during the time of airing, compared to the percentage changes of differences around baseline activity.
6. % of events shows the web traffic as a percentage of all web traffic. Some web traffic categories that have high spike response from infrequent events.

4.1 Results

4.1.1 Visitors

Figure 2 shows new visitors versus existing visitors. The existing visitor timeseries is almost un-moved when there are TV airings with only a 2.8% spike which is not statistically significant (Table 2). In contrast, new cookied visitors shows distinct spikes that match TV airings that are 14.8% in magnitude and are significant.

4.1.2 The Most Responsive Viewers use Mobile Devices

The highest signal-to-noise ratio for any traffic category is "New Cookie Visitor traffic with Mobile User-agents that reach the Homepage". The web spike here is a dramatic 76% higher than the baseline activity ($p<0.01$; Table 2 and Figure 4). This lift occurs on 2% of the site's traffic.

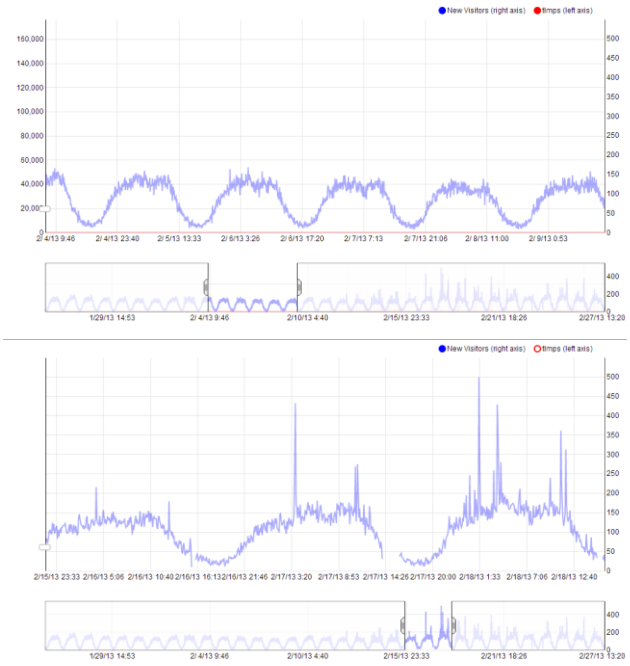
4.1.3 The Least Responsive Traffic to TV are Deep-linked page browsers and Email responders

Most categories of traffic show TV spikes that are statistically significant. However three categories appear to have minimal change from TV. Only 13% of the traffic: "Email", "Product Page", and "Affiliate traffic" are largely not impacted by television (spike magnitude <= 1.7%; not statistically significant). Product pages make sense since these are "deep-linked pages" which would likely be reached after a search for a specific product, where-as the TV airing tends to produce branded searches and first time visitor activity to the home page. Email also makes sense given that email campaigns occur on an episodic basis and are likely to be uncorrelated with television advertising.

4.1.4 Significant Amounts of Paid and Organic Search Traffic is Mis-Attributed and due to Television Broadcasts

It was noteworthy that a considerable amount of Paid Search and Organic Search traffic at the time of a TV airing is actually due to the television broadcast (Table 2; Figure 4). A 9.1% spike was measured on Paid Search (SEM) ($p<0.01$) and 7.3% on Organic Search (SEO) ($p<0.01$), with both increases statistically significant. These effects are important because these are large sources of traffic, and so these lifts translate into a large increase in traffic that can be measured as due to television.

The ability to measure effects on these digital marketing channels is important because most conversion tracking systems on Paid Search will mis-attribute this traffic. Various authors have noted that a problem with current last-click attribution systems is that the search query or paid link receives 100% credit for the conversion, and often results in brand name keywords (eg. for a company called "Physicians Mutual Insurance" a brandname keyword would be "Physicians Mutual") with thousands of conversions and cost per acquisitions of pennies [15]. This makes it seem like these keywords are the most effective advertising vehicles for producing conversions, when in fact the users typing in these brandname keywords already know of the company and are trying to navigate to it. The webspike analysis reported here confirms that a lot of search click-throughs that are being credited to keywords, are in fact occurring due to untracked television broadcasts.



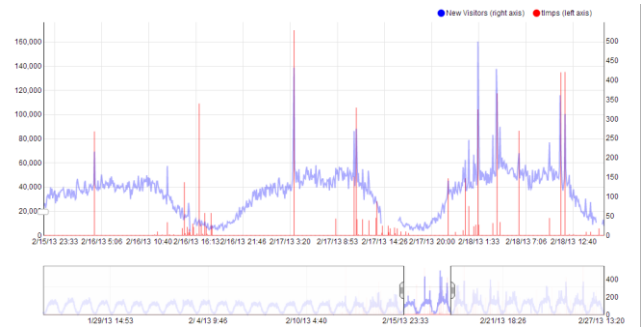


Figure 1. Web activity versus TV airings during a national campaign.

Table 2. Web Metrics that are more filtered are able to better identify the impact of the television airing

Metric	Mean Units	Var Units	% of events	tImp Corr.	Inc per mill. Imps wpmm	% 5 min. inc. over base w 30 min. excl.	addtl. traffic during 5 min of TV airing
New home mobile	5.1	34	2%	0.069	9.64	75.6%	3.9
New home direct	34.6	446	11%	0.057	2.45	34.4%	11.9
Mobile	23.4	188	7%	0.055	2.95	34.1%	8.0
New home	58.7	1,035	18%	0.055	2.14	30.7%	18.0
Home Page	58.7	1,035	18%	0.055	2.14	30.7%	18.0
Direct To Site	53.2	617	17%	0.052	1.65	25.3%	13.5
No Referrer	53.2	617	17%	0.052	1.65	25.3%	13.5
Tablet	14.9	101	5%	0.035	1.64	24.0%	3.6
New Visitors	164.3	5,889	52%	0.035	0.89	14.8%	24.3
Unique Visitors	262.2	15,565	83%	0.033	0.62	10.9%	28.6
Visits	291.6	19,930	92%	0.028	0.59	10.2%	29.7
SEM	54.6	866	17%	0.023	0.67	9.1%	5.0
Desktop	126.0	3,663	40%	0.027	0.46	7.8%	9.8
SEO	36.8	458	12%	0.016	0.46	7.3%	2.7
Existing Visitors	97.9	2,569	31%	0.028	0.16	2.8%*	2.7
Category Page	60.6	980	19%	0.009	0.25	2.3%*	1.7
Email	4.8	61	1%	0.017	0.36	1.7% (ns)	0.1
Product Page	33.1	310	10%	0.024	0.07	0.7% (ns)	0.2
Affiliates	7.3	20	2%	0.028	0.19	0.0% (ns)	0
Display	0.7	1	0%	0.008	0.23	na	0

* Significant $p<0.05$; (ns) Not significant; All other metrics are significant at $p<0.01$

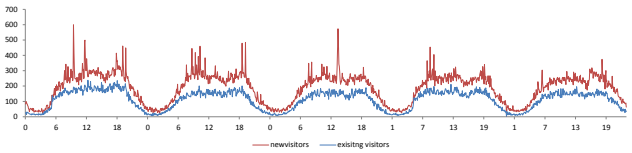


Figure 2. Existing visitors versus New Cookied Visitors: Web spikes tend not to show up in Existing Visitor traffic.

5. WEB RESPONSE ATTRIBUTION

5.1 Problem Description

"To Learn what TV is, you must learn what TV is not" – Yoda Speak Generator [17]

The durability of web spikes introduces the intriguing possibility that we could use these web effects to build a *television conversion tracking system* that we could use for TV effectiveness reporting and optimization. In order to build such a system, we can make use of a useful property of the above models – specifically, our observed web activity $W(T,G)$ can divided into two parts: (a) the web activity due to the television airing $f(I(M(T,G)))$ plus (b) background web activity that would have occurred anyway $W_{NoTV}(T,G)$:

$$f(I(M(T,G))) = W(T,G) - W_{NoTV}(T,G)$$

When we are trying to estimate the total effects of television, a key insight is that we don't have to identify the exact parametric functional form of f – indeed this could be a complex function, require a large number of factors, involve complex interactions, and may take a great deal of insight to understand. It would be easy to get the model wrong. Instead, our approach is to estimate the *background* $W_{NoTV}(T,G)$, and remove it from observed activity to infer the activity due to TV: in other words learn what TV is not. In practice, background traffic tends to be more reliable to estimate than activity during TV airings, as there are many more observations of web activity, and it is often periodic and predictable. TV estimates then become the observed residuals after subtraction of background, in the narrow window during TV airings when there is high signal-to-noise. We now turn to two algorithms for removal of background activity which introduce minimal assumptions.

5.2 Algorithm I: Instantaneous Minute Treatment-Control

The Instantaneous Treatment-Control algorithm works by estimating the web traffic without TV as being equal to the web traffic in the time period before the airing. In the formula below, the web activity control period *y* indicates how many periods of time prior to the present will be used to create an average for web activity (eg. *y*=1), and the algorithm requires there to have been no media airings during z time-periods prior to the present (eg. *z*=1), the exclusion window. This ensures that the previous web activity isn't being elevated by an earlier media event.

$$W_{NoTV}(T,G) = \begin{cases} \frac{1}{y}\sum_{t=1}^{y} W(T-t,G) : if \sum_{t=1}^{z} I(M(T-t,G)) = 0 \\ UNDEF: otherwise \end{cases}$$

By using the web activity prior to the airing as the noTV baseline, the treatment when TV was applied is close in time to the control, and the environmental conditions – including day-hour-minute traffic levels - of the website should be very similar in these two cases. Therefore, any change in web visits is likely to have been caused by the TV airing.

This algorithm has an advantage of being robust when faced with real-world problems such as bots. Bots are notorious for crawling various pages, and then may not be seen again for some time. This can cause significant problems when estimating models of web activity, since global models such as regression can be severely disrupted by what appears to be large amounts of web traffic.

However, because web activity immediately prior to the airing is close in time to time of airing, it is likely that the bot will be present and generating a similar inflation during both time buckets *T* and *T*-1. Therefore if we have:

$$f\big(I(T,G)\big) = W(T,G) + BOT(T,G) - \big(W(T-1,G) + BOT(T-1,G)\big)$$

If $BOT(T-1,G) = BOT(T,G)$ then the estimate will not be affected. This is likely to be true in many cases because bot activity is correlated in time.

5.3 Algorithm II: Day-Hour-Minute Subtraction

Day-Hour-Minute Subtraction works by building an expectation for the Day-Hour-Minute bucket for web activity on the web site when TV is not present. A simple method is to take the average for the day-of-week, hour-of-day and minute-bucket when a media airing has not occurred within the last *z* time-periods prior to the day-hour observation.

$$W_{NoTV}(T,G) = E[W(Y,G)] : \sum_{t=1}^{z} I(M(Y-t,G)) = 0$$
$$\wedge\, DD(Y) = DD(T) \wedge HH(Y) = HH(T) \wedge\, MM(Y) = MM(T)$$
$$W_{NoTV}(T,G) = UNDEF : otherwise$$

where *DD*(*Y*) is the day-of-week for time bucket *Y*, *HH*(*Y*) is the hour of day of time-bucket *Y*, *MM*(*Y*) is the minute bucket for *Y*, and E[] is the mean. For a 7 days, 24 hours and 0 minute buckets, this is equivalent to defining 168 day-hour 1-0 dummy variables in a regression model that is trying to predict the non-TV background web activity.

The Day-Hour-Minute Subtraction algorithm is a global model and so is susceptible to problems including (a) in-time trends in web activity – for example if a website's traffic is growing then the method may incorrectly begin to estimate more activity due to television, (b) bot activity – this will cause outliers to pull the numbers out, (c) transient background changes due to interference from other TV airings from the same campaign. Never-the-less, for well-spaced airings it can be effective.

5.4 Simultaneous Airings

When two or more airings simultaneously occur within the same time period, we have two potential airings that could have caused the spike - which should take credit? There are two methods of handling this case: (a) set the airing attribution to missing because attribution is unclear, or (b) attempt to apportion credit. A useful heuristic for partial attribution is to apply credit in proportion to each airing's television impressions as a percentage of total. For example, say that we have three airings with 100, 700, and 200 impressions each. The middle airing would receive 70% of the credit. This method is problematic if the airings vary widely in targeting quality – for example the first airing may reach the ideal target, where-as the second might be running in the middle of the wee hours of the morning when none of the target are watching. However the benefit of the method is that it enables more airings to be attributed and minimizes the number of factors introduced.

5.5 Attribution Results

5.5.1 Algorithm Comparison

We ran the attribution algorithms on the TV campaign and associated web activity from February 11 2013 to July 2013. The television campaign comprised 35,296 airings. We calculated both attribution methods, and then compared the web response per impression calculated using each algorithm to the targeting score of each airing as calculated using an algorithm published in [8]. The comparison measures the number of buyers per million in the viewing audience. The Instantaneous model has a fit of R=0.64, day-hour subtraction with 3 hour exclusion has a fit of R=0.42, and day-hour subtraction without exclusion produces a fit of R=0.33.

5.5.2 TV Media Performance Analysis

We can also report on the performance of different networks, day-parts and programs in generating web spikes by dividing TV-generated web activity by TV impressions.

$$wpi(M) = \frac{\sum\ W(T,G) - W_{NoTV}(T,G)}{I\big(M(T,G)\big)}$$

The highest web-spike per impression network-programs and network-day-parts for our live TV campaign – as measured by Algorithm 1 - are shown in Tables 3 and 4. The most responsive networks were Discovery Health and Fitness, SOAP and Comedy. The most responsive programs were "Veronica Mars", "One Tree Hill" and "Gilmore Girls". The product being advertised was one that appealed to higher income women who were just married or were renovating, so these intuitively made sense [9],[10].

Table 3. Network Day Hours with highest web response per impression measured in the TV campaign.

Network-Day-Hour	WPI
SOAP - Su - 3 pm	0.00322
COM - Tu - 1 pm	0.00302
DFH - Tu - 11 am	0.00289
DFH - W - 2 pm	0.00273
DFH - M - 7 am	0.00259
COM - W - 1 pm	0.00253
DFH - M - 1 pm	0.00229
COM - Th - 1 pm	0.00214
COM - Tu - 12 pm	0.00211
DFH - Th - 3 pm	0.00206

Table 4. Network Programs with highest web response per impression in the TV campaign.

Network-Program	WPI
LMN – Movie	0.0203
SOAP - Veronica Mars	0.0147
SOAP - One Tree Hill	0.0107
AMC - AMC Movie	0.0085
SOAP - Gilmore Girls	0.0076
OWN - Dr. Phil	0.0073
SOAP - General Hospital	0.0067
WGNA - Law & Order: Criminal Intent	0.0065
SOAP - Beverly Hills, 90210	0.0054
COM - South Park	0.0046

6. EXPERIMENTAL COMPARISON

The web response alignment graphs are suggestive. However, TV is known to have complex and long-lasting effects [3],[13],[14]. Could web spike response be used as a proxy to measure total TV effect?

We firstly need to be able to measure total TV lift. A classic method for measuring total TV effect is to run a controlled experiment. Media is applied to certain geographies, and not to control geographies. The difference in web activity between treatment and control is then measured. This is called a Matched Market Experiment and it has been used in many previous studies to measure television effects [3],[13],[14],[1],[7].

We implemented a Matched Market Experiment by purchasing 9,748,347 impressions and 296 airings ($483 per airing) of media on the week of February 11 and March 4 2013 in treatment market $G_j = Seattle$. This purchased approximately 281 Gross Rating Points per week[2] in the targeted area. We selected an aggregated control $W(T, G_{j,CON})$ matched to this treatment and not running media as follows[3]:

$$W(T, G_{j,CON}) = \sum_{i \neq j} w_i \cdot W(T, G_i) + w_0 : \sum I(M(T, G_i)) = 0$$

where w_i were trained using data from times T_0 that were prior to the start of television, selected using stepwise regression to avoid over-fitting, and the model was validated against a test set that was held out in time. The parameters are shown in Table 7.

$$\min w_i : \sum_{T \in TRAIN} \left(W(T_0, G_j) - \sum_i w_i \cdot W(T_0, G_i) + w_0 \right)^2$$

Difference of Differences can now be used to calculate the activity due to the treatment in this kind of design [1]. The method measures the change in treatment area minus change in control area:

$$f\left(I\left(M(T, G_j)\right)\right) = \left(W(T, G_j) - W(T_0, G_j)\right) - \left(W(T, G_{j,CON}) - W(T_0, G_{j,CON})\right)$$

Because we are using an explicit, time-varying control which minimizes difference between $W(T_0, G_{j,CON})$ and $W(T_0, G_j)$, the treatment and control starting terms cancel, and the Difference of Difference formula becomes the formula below. The results are shown in Table 5 and Figure 5.

$$f\left(I\left(M(T, G_j)\right)\right) = W(T, G_j) - W(T, G_{j,CON})$$

$$lift = \frac{W(T, G_j)}{W(T, G_{j,CON})} - 1$$

The first surprising result is that web spike lift readings appear to predict total TV effect. Web Spike analysis reported 30.7% lift for Homepage, 14.8% for new visitors and 10.2% for visits. Experimental measurement exhibits the same relationship: 58%, 27%, 18% (Table 5; Figure 5).

The second result is that the amount of lift measurable by web spike is small relative to the total effect of TV. We experimentally measured an additional 3.5 conversions for every conversion generated during the campaign in the 6 months after the campaign because of elevated lift in treatment area [9],[10]. Web spike is unable to detect this lift as it works on short-term effects. In addition, web spike measurements only observe a narrow time window around each airing when signal-to-noise is maximum. Based on algorithm 1 we measured only 0.69% of total web effect including residuals after 6 months. Never-the-less, despite measuring only a small amount of TV's total effect, the measured signal appears to be correlated with overall TV effect.

Table 5. Experimental Lift versus Web Spike Lift

Traffic	Exp % lift	web spike % lift / base	Exp Corr metric, conversions	web spike timp corr.
Homepage	57.8%	30.7%	0.321	0.079
New visitors	27.4%	14.8%	0.162	0.064
Visits	18.3%	10.2%	0.130	0.054

[2] A Gross Rating Point (GRP) is equal to 100 * impressions per TV Household per area per unit time. For example, 281 GRPs per week is equal to 2.81 impressions per household per week.

[3] The control areas were actually subjected to approximately 20 GRPs of advertising weight due to some national advertising that was unavoidable, so the comparison was 281 GRPs in treatment versus 20 GRPs in control.

7. TARGETING APPLICATIONS

We have discussed how we can perform robust measurement of web-spikes, We will now show how we can use this information to automatically optimize a television campaign.

A simple statement of the television ad targeting problem is to select media in order of value per dollar

$$M_i: \mathbf{max} \frac{wpi(M_i)}{CPI(M_i)}$$

descending until the budget is filled, where $CPI(M_i)$ and $wpi(M_i)$ are clearing prices and media observations. $CPI(M_i)$ can be obtained from the TV stations, so we will focus on estimating the web spike of an upcoming airing $wpi(M_i)$.

In order to calculate the above, we will first assume that we have some historical media airings *M* and web-spike measurements *W*. We calculate *wpi*(*M*) using the method we introduced earlier:

$$wpi(M) = \frac{W(T,G) - W_{NoTV}(T,G)}{I(M(T,G))}$$

In order to estimate the performance of an upcoming/future airing M_i, we will break the airing into a series of features including station, program, and so on. For example consider a future media instance: $M_i =$ (CNN, 8pm, "Piers Morgan", Tuesday, 12/12/2012, Pod1, Pos2, 60s). The following features could be used to predict its performance: Station $m_{i1} =$ (CNN), Station-Hour-Pod $m_{i2} =$ (CNN, 8pm, Pod1), Geography-Station $m_{i3} =$ (National-CNN), and so on. The prediction of performance for media then becomes:

$$wpi(M_i)^* = \sum_q v_{jq} \cdot wpi(m_{jq}) : m_{jq} \in M_i$$

Where each airing feature is weighted by v_{jq} to create an overall estimate. As new airings and web spikes occur, we update our *wpi* statistics for media per above. We can now re-calculate our ranking function and buy better targeted media:

$$M_i^*: \mathbf{max} \frac{wpi(M_i)^*}{CPI(M_i)}$$

8. CONCLUSION

The methods described in this paper measure anonymous web activity after TV broadcasts and so are privacy-friendly. They can also be used on TV systems today. Although we haven't presented the results in this paper, we have also employed the technique for a range of different TV advertisers including large branded websites, e-commerce advertisers, and high consideration purchase websites, who are executing television campaigns, all with good results. The economic impact of a conversion tracking system for television would be significant. Direct Response television is a 20 billion dollar industry. Perhaps web spike might be as important.

9. REFERENCES

[1] Angrist, J. and Pischke, J. 2010, *Mostly Harmless Econometrics*, Princeton University Press.

[2] Clipp, C. 2011, An Exploration of Multimedia Multitasking: How Television Advertising Impacts Google Search, Honors thesis, Economics, Duke University, Durham, North Carolina.

[3] Hu, Y., Lodish, L., Krieger, M. 2007. An Analysis of Real World TV Advertising Tests: a 15 year update, *Journal of Advertising Research*, Vol. 47, No. 3, pp. 341-353.

[4] Joo, M., Wilbur, K., & Zhu, Y. 2011. Does Television Advertising Influence Online Search? Working Paper, Duke University.

[5] Joo, M., Wilber, K., Zhu, Y. 2013, Effects of TV Advertising on Keyword Search in the AOL Dataset, http://ssrn.com/abstract=1720713

[6] Kitts, B., Wei, L., Au, D., Powter, A., Burdick, B. 2010, Attribution of Conversion Events to Multi-Channel Media, *Proceedings of the Tenth IEEE International Conference on Data Mining*, December 14-17, 2010. IEEE Computer Society Press.

[7] Kitts, B., Au, D., Burdick, B. 2013, Real-time Television ROI Tracking using Mirrored Experimental Designs, Trends and Applications in Knowledge Discovery and Data Mining, *Lecture Notes in Computer Science*, Volume 7867, 2013, pp 95-108, Springer.

[8] Kitts, B., Au, D. and Burdick, B. 2013, A High-Dimensional Set Top Box Ad Targeting Algorithm including Experimental Comparisons to Traditional TV Algorithms, *Proceedings of the Thirteenth IEEE International Conference on Data Mining*, Dec 7-10, Dallas, TX. IEEE Press.

[9] Kitts, B. 2014, Inside Art.com's Data-Driven Television Campaign, *Journal of Advertising Research*, in press.

[10] Kitts, B., Wadsworth-Drake, J., Vollmann, W., Ross, I., Martin, G., Tjen, D., Au, D., Zlomek, S., Chun, A., Sobieski, J., Giusti, S., Lyons, M., Harris, J., Kovalik, I., Perkins, B., Smith, S., Hill, M., Boyarsky, A., Morse, E. 2014, Find Your Art. Love Your Space, Ogilvy Award Winning Case Study, Advertising Research Foundation. http://thearf-org-aux-assets.s3.amazonaws.com/ogilvy/14/art-case-study.pdf

[11] Kitts, B., Zhang, J., Wu, G., Brandi, W., Beasley, J., Morrill, K., Ettedgui1, J., Siddhartha, S., Yuan, H., Gao, F., Azo, P., Mahato, R. 2014, Click Fraud Detection, in Mahmoud Abou-Nasr, Robert Stahlbock, Stefan Lessmann, Gary M. Weiss (eds), *Annals of Information Systems*, Special issue on Real World Data Mining Applications, Springer.

[12] Lewis, A. and Reiley, D. 2013, Down-to-the-Minute Effects of Super Bowl Advertising on Online Search Behavior, *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pp. 639-656

[13] Lodish, L., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., Stevens, M. 1995. How T.V. Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments, *Journal of Marketing Research*, Vol. 32, No. 2, pp. 125-139. 1995.

[14] Lodish, L., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., Stevens, M. 1995. A Summary of Fifty-five In-Market Experiments on the Long-term Effect of TV Advertising, Marketing Science, Vol. 14, No. 3, pp. 133-140. 1995.

[15] Blake, T., Nosko, C., Tadelis, S. 2013. Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment, Haas School of Busines, University of

California Berkeley, Working Paper. http://faculty.haas.berkeley.edu/stadelis/Tadelis.pdf

[16] Zigmond, D. and Stipp, H. 2010. Assessing a new advertising Effect: Measurement of the Impact of television commercials on Internet Search Queries, *Journal Of Advertising Research*, June, 2010, pp. 1-7.

[17] Yoda Speak Generator, http://www.yodaspeak.co.uk

Table 6. Minutes away from airing versus web activity in percent of baseline.

Minutes from airing	New visitors	Existing visitors	New Home Mobile	SEM	SEO	Mobile User Agent	Tablet User Agent	Desktop User Agent	Null Referrer	Home Page	Category Page	Product Page
-60	-2%	-1%	6%	-3%	-6%	1%	0%	-2%	1%	-1%	-2%	-3%
-55	-1%	-1%	14%	-3%	-1%	1%	0%	-2%	1%	2%	-2%	-3%
-50	-1%	-1%	1%	-1%	-3%	1%	0%	-1%	0%	-1%	-2%	-3%
-45	0%	0%	1%	-1%	-1%	1%	0%	-1%	0%	1%	-1%	1%
-40	0%	1%	9%	-1%	-1%	1%	5%	-1%	0%	1%	1%	1%
-35	0%	-1%	1%	-1%	-1%	1%	0%	-1%	0%	1%	-1%	1%
-30	0%	-1%	1%	-1%	-1%	1%	0%	0%	0%	-1%	-1%	1%
-25	-1%	-1%	-8%	-1%	-1%	-2%	0%	0%	-1%	-2%	1%	-3%
-20	-1%	-2%	-8%	-1%	-1%	-5%	0%	0%	-1%	-2%	1%	1%
-15	-1%	-1%	-8%	-1%	-1%	-2%	-5%	-1%	-3%	-3%	1%	1%
-10	-2%	0%	-8%	-1%	-1%	-2%	-5%	-2%	-3%	-3%	1%	-3%
-5	0%	1%	1%	1%	-1%	1%	5%	0%	2%	2%	-1%	1%
0	**15%**	**3%**	**76%**	**9%**	**7%**	**34%**	**24%**	**8%**	**25%**	**31%**	**2%**	**1%**
5	0%	2%	1%	1%	-1%	1%	0%	0%	0%	1%	1%	1%
10	0%	1%	-8%	1%	-1%	-2%	0%	1%	0%	-1%	1%	1%
15	0%	1%	-8%	1%	-1%	-2%	0%	0%	-1%	-2%	1%	1%
20	0%	0%	-8%	1%	2%	-2%	0%	0%	-3%	-3%	1%	1%
25	1%	0%	1%	1%	2%	1%	0%	1%	0%	2%	-1%	1%
30	1%	1%	1%	2%	2%	1%	0%	2%	0%	0%	1%	1%
35	1%	1%	9%	1%	2%	4%	0%	1%	2%	2%	1%	1%
40	1%	1%	1%	-1%	2%	1%	0%	1%	0%	0%	1%	1%
45	2%	2%	9%	2%	2%	4%	0%	2%	2%	4%	1%	1%
50	1%	2%	1%	2%	2%	1%	0%	1%	0%	2%	1%	1%
55	1%	2%	1%	2%	2%	1%	-5%	2%	0%	0%	1%	1%
60	1%	0%	1%	1%	5%	-2%	0%	2%	0%	0%	1%	1%

Table 7. Regressed Control weights.

DMA	w_i	TVHH	DMA	w_i	TVHH
Intercept	2.6494		HOUSTON,TX	-0.1806	2,123,460
ALBUQUERQUE,NM	0.5695	694,040	KNOXVILLE,TN	0.572	552,380
ALEXANDRIA,LA	-0.6613	90,740	LAREDO,TX	-1.2988	69,790
BANGOR,ME	0.919	144,230	LASVEGAS,NV	0.614	721,780
BATON ROUGE,LA	-1.0384	326,890	LEXINGTON,KY	-0.8404	506,340
BILOXI,MS	-1.3307	122,740	LOS ANGELES,CA	0.1774	5,659,170
BRIST/JOHN,VA-TN	1.0413	334,620	MIAMI/FTLAUD,FL	-0.4226	1,538,090
BUFFALO,NY	0.4897	633,220	NEWORLEANS(LA)	0.4283	633,930
BURLINGTON,VT-PQ	0.47	330,650	ROCKFORD,IL	1.249	189,160
CHARLOTTE,NC	0.1617	1,147,910	SALISBURY,MD	-0.6211	158,340
CHEYENNE,WY	1.3214	54,710	SHREVEPORT,LA	-1.4458	386,180
CLEVELAND,OH	0.0004	1,520,750	SPRINGFIELD,MA	0.7213	262,960
DULUTH,MN-WI	1.2876	174,360	TOLEDO,OH-ON	0.7759	423,100
FARGO,ND	1.5819	240,330	MADISON,WI	0.2814	377,260
GREEN BAY,WI	-0.6902	443,420	SALINAS-MONTEREY,CA	0.7138	227,390

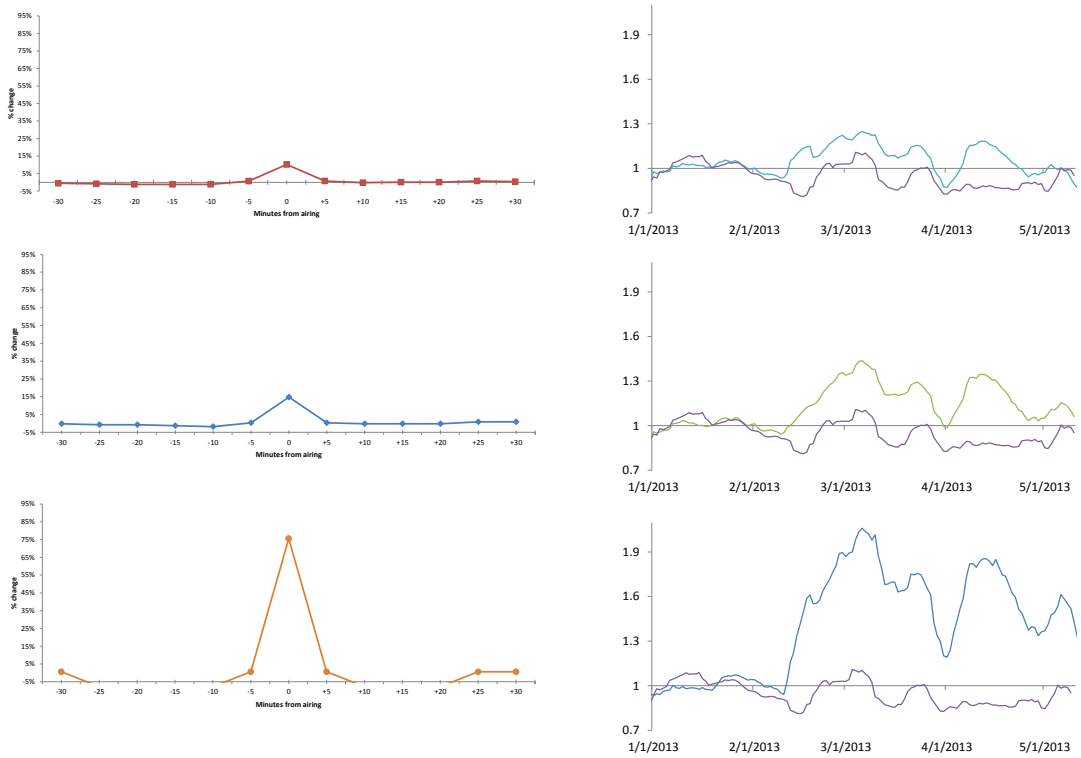


Figure 5. (left) Web Spike Percentage Lift for 60 minutes before through 60 minutes after a TV airing (at minute 0) with spike magnitude expressed in terms of percentage lift over average baseline activity. (right) Experimental Lift Measurement measured using Difference of differences calculated over a treatment and control geographic area for three web metrics. The timeseries is shown as a 7 day moving average. The control area with web activity is normalized to 1.0, and treatment area shows activity in units of percentage over control. The top graphs are Visits, middle are New Visitors, and bottom are New Visitors to the Homepage on Mobile devices. As Web Spike Percentage lift increases, so too does Experimentally measured lift.