CeADAR – Centre for Applied Data Analytics Research

Enterprise Ireland Data Analytics Technology Centre

SmartAd – Weak Signals and Causal Relationships

Document Type:	Literature Review
Project Title:	CeADAR
DDN Theme:	1 – Intelligent Analytic Interfaces
Theme Leader:	Sarah Jane Delany (DIT)
DDN Sub-Theme:	Ease of Interaction
Authors:	Ingo R. Keck
Document Version:	0.1
Date of Delivery to ISG:	16th December, 2014
Number of pages:	8

CeADAR is a research partnerhip comprising University College Dublin, University College Cork, and Dublin Institute of Technology.

`http://www.ceadar.ie`

Contents

1 Executive Summary

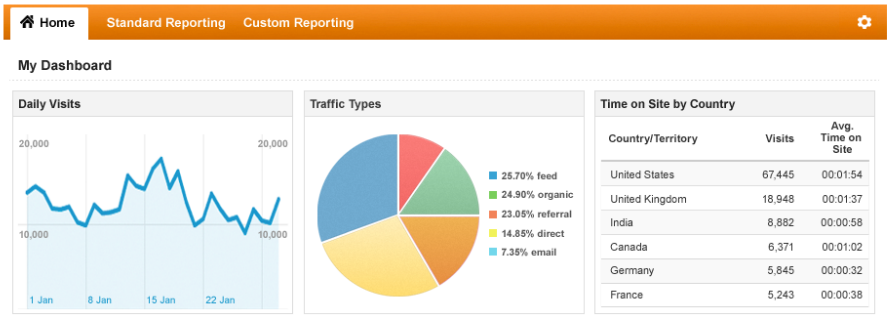
In this review, we discuss the state-of-the-art of both research and development in the analysis of advertisement campaign data for weak signals and causal relationships:

- The question, what parts of an ad campaign work and what parts do not, as well as how they all interact with each other, is a long studied problem in marketing, but has not been solved so far.
- To find causal relationships between events is not a problem unique to marketing. Relevant techniques can be applied to different forms of events, if enough data is available.
- *Weak signals* describe different things depending on the context. In the case of this project, both the interpretation of weak signals as a difference in split testing, as well as the description of it to be a small noisy signal is adequate.
- The combination of data with different temporal granularity from miscellaneous channels is called *data fusion* and is a problem that is still the subject of ongoing investigation in the scientific literature.
- Existing data fusion methods depend on the field of application.
- *Correlation* and *causality* are two different subjects.
- Correlation can be calculated and is a measure of similarity between two signals. Correlation can be causal or random or both. Statistical tests exist that give the probability of an observed level of random correlation.
- Causality requires an informed choice. The interaction of two events must be understood before it is possible to infer from a measured high correlation that a causal relationship exists.
- Existing solutions for advertisement campaign analytics usually measure impact from channels separately and do not provide tools to estimate causality or interaction between channels.
- More advanced analytics are provided as a service by specialised consultancy companies.
- Visualisation of correlation again depends on the field of application. It is usually done by showing connections between correlated signals or events, or by plotting the *correlation matrix*.

2 Introduction

One of the major questions that arose from our discussions with companies involved in the advertising space is how they can determine which advertising initiatives have been successful and which have not. This age-old question in advertising gave rise to the famous John Wanamaker quote: "Half the money I spend on advertising is wasted; the trouble is, I don't know which half".

Today, advertising runs through various channels like television, radio, print, web and social media. Each of these distribution forms has different characteristics regarding form, audience and data about customers interaction with the advertisements. User tracking in web and social media can provide very deep insights into how people react to the different forms, but, naturally, can only link online ads together with the tracked user.



Time on Site by Country

Country/Territory	Visits	Avg. Time on Site
United States	67,445	00:01:54
United Kingdom	18,948	00:01:37
India	8,882	00:00:58
Canada	6,371	00:01:02
Germany	5,845	00:00:32
France	5,243	00:00:38

Figure 1: Google Analytics dashboard example. Source: `http://www.google.com/intl/en_uk/analytics/`

State of the art in marketing research so far is to analyse the direct impact of different advertisement campaigns separately. In reality, consumers are not just subject to the influence of one campaign, instead they are constantly presented with different forms of advertisements via a subset or all of the noted channels. The effects of campaigns in different channels might interact: A customer that has first seen an advertisement on the TV and heard about it on the radio might be more easily influenced to register a corresponding ad on a social network that finally leads him to the offer webpage of the company.

To find the influence of campaigns in different channels is therefore a difficult task. This project will attempt (in a small way) to answer this question by analysing data to find weak signals that exist between advertisements and subsequent consumer events (e.g. purchases) and to present these to a user in a way that makes them easy to interpret.

Note however that this research question is not unique to the advertising space. Other examples include brain connectivity research, where weak signals from the brain are also used to investigate possible causal relations between activity in different brain regions [5]. The techniques presented in this report can therefore also be applied to other forms of data, though some of the used models may need to be adjusted.

3 Weak Signals

Differences exist in the interpretation of the phrase 'weak signal', depending on the context:

In *economics*, the term *weak signal* is usually used to denominate an early indicator for future changes that might be missed easily [7]. These kind of signals are seen to be not relevant for this project.

In *signal detection theory*, a *weak signal* refers to a small difference in two distributions, and the question is whether this small difference is due to some fundamental reason or just to chance. A common application in marketing is so-called *split testing* or *A/B testing*. In this case the distributions are the click-through rates of both settings A and B and the tester wants to find out if there really is a difference in the effect of A versus B, or if the signal is just due to random fluctuations.

In *signal processing*, a *weak signal* is usually a real signal that is covered by noise and near the limit of being able to be detected and separated from noise [15]. Using multiple, but similar copies of the signal in multivariate analysis is a common method to reduce the effect of noise and extract a more pure signal.

Within the scope of this project, both methods from signal detection theory and from signal processing could be applied. Multivariate analyses will provide a more in-depth analysis of the effects of the ad campaigns, but will also require more data and will have special requirements in order to be able to separate the influence of different forms of advertising. A simple statistical test for difference like the one that is used in A/B testing has far lower initial requirements, but will not be able to separate out detailed effects of different ad campaigns that are run at similar times.

4 Data Fusion

Data from advertisment campaigns is likely to be at very different levels of granularity – for example data from TV or print advertising simply describes the number of potential viewers for an advertisement within a specific period of days, whereas data from web advertisements is much more specific. Click rates and user interactions on web pages can also provide very detailed information on a much finer level of time. It is necessary to bring this information to a common time space for an in depth analysis of the effects of advertisement campaigns on user interactions. This process is called *data fusion* [4][11] and many techniques exist depending on the specific characteristics of the data involved.

In the case of split testing, the common time scale is that of the long term campaigns (print, television, radio), where no detailed interaction data is available. Information on click rates then can simply be summed up for multiple periods within these campaigns and the resulting distributions compared using statistical testing.

In the case of multivariate signal analysis, different techniques exist to bring the data on to a comparable time scale. One common method is to apply pattern extraction on all signals. PCA has been used successfully for this in the case of neuroimaging data [3]. Another common alternative is to generate a higher resolution model of the low granularity signals, estimate the model parameters from the data, and to use the model output for the analysis (see [12] for an example related to customer interactions with movie launches).

5 Correlation And Causality

Correlation in a statistical sense means that a change in one variable, i.e. in the amount of perceived advertisements, is related to a corresponding change in another variable, i.e. sales. Causality is the term used to describe a situation whereby one of these changes is the direct and consequential result of the other.

While correlation often indicates a causal link, it is not a proof of it, only an indication. Correlation also does not infer a direction of information flow between two changes. So-called *spurious correlations* are simply the result of random fluctuations in the data. While, strictly speaking, it is impossible to find out with mathematical methods whether a correlation is just down to noise or whether it is real, statistical methods like t-tests [2], Granger Causality [6] or further specialised tests based on the Bayes statistic framework [13] can provide insight on the probability that a correlation is indeed a real signal and not down to noise or random effects.

It is therefore obvious that, while the process of finding correlations in data can be automated, the decision on whether there is a causal relationship as well must be made by a user who has a deep understanding of the data. One way to ease the work of the user would be to provide visualisations that give insight into the correlations between the data and the statistical viability of these correlations.

6 Existing Solutions

IPA's touchpoints program with its *Hub Survey* provides an in-depth view of customer activities and advertisements consumption over a wide variety of channels [14].

Most analytics systems already make use of direct causal information, i.e. specific links that allow a user to trace the web page visit back to a click on an ad. If this information is available, correlation analysis is straightforward as each ad is directly linked to the outcome. Companies like Google [1], Twitter [10][9], Facebook [8] and many other online ad networks in general provide a basic visual analysis, often for free within their services. Services like mixpanel[1] allow further insights and detailed views on user interactions and pathways.

The question remaining to be answered is whether the differences between ad campaigns is due to chance or not. This service is mostly provided by dedicated consulting agencies (for example Marketshare[2]) or can be done with the help of statistics software like SPSS, SAS, octave, Matlab, etc. by an experienced statistician.

Visualisations of correlations are wide-spread and usually consist of either visualising the *correlation matrix*, or drawing lines between correlated events, where the characteristics of the line (e.g. colour, width) corresponds to the level and probability of correlation (see figure 2 for simple examples.) The correlation matrix shows the (cross-) correlation values for all investigated signals/events in the analysis.

[1] https://mixpanel.com

[2] http://marketshare.com

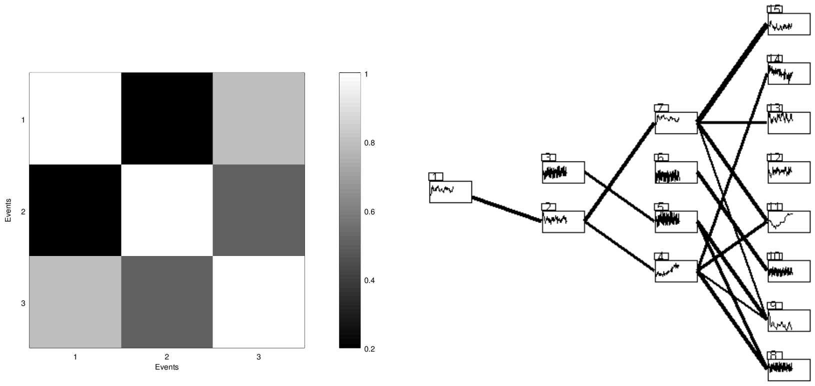


Figure 2: On the left side an example for a correlation matrix visualisation, where the level of correlation between events is encoded in the colour. The right side shows a more complex correlation analysis of activity networks in a human brain imaging experiment.

7 Conclusions

Our conclusion, after carrying out this State-Of-The-Art review, is that the subject of this project has not been explicitly implemented in software so far. It can be separated into three parts: data fusion, correlation between events and visualisation of the results. Data fusion can be either very demanding or straightforward, depending on the level of analysis that is to be performed. In the case of split testing, the channels can be simply defined and used as the split delimiter. For a more detailed analysis of cross-influences, advanced data fusion techniques will have to be used, which may be difficult to implement in the restricted CeADAR six month process circle.

References

[1] Google Analytics. About campaigns, Nov 2014. `https://support.google.com/analytics/answer/1247851`.

[2] Joan Fisher Box. Guinness, gosset, fisher, and small samples. *Statistical Science*, 2(1):45–52, Feb 1987.

[3] V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14:140–151, 2001.

[4] Federico Castanedo. A review of data fusion techniques. *The Scientific World Journal*, 2013:704504, 2013.

[5] Karl Friston, Baojuan Li, Jean Daunizeau, and Klaas Stephan. Network discovery with dcm. *NeuroImage*, 2010.

[6] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, Aug 1969.

[7] Elina Hiltunen. Good sources of weak signals: A global study of where futurists look for weak signals. *Journal of Futures Studies*, 12(4):21–44, May 2008.

[8] Facebook Inc. Tracking facebook ads in google analytics. `https://www.facebook.com/business/google-analytics`.

[9] Twitter Inc. Conversion tracking for websites. https://support.twitter.com/groups/58-advertising/topics/247-measure-analyze/articles/20170807-conversion-tracking-for-websites.

[10] Twitter Inc. Tweet activity dashboard. https://support.twitter.com/groups/58-advertising/topics/247-measure-analyze/articles/320043-tweet-activity-dashboard.

[11] Bahador Khaleghi, Alaa Khamis, and Fakhreddine O. Karray. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(28-44), 2013.

[12] Ho Kim and Dominique M. Hanssens. Paid and earned media, consumer interest and motion picture revenue. http://www.anderson.ucla.edu/faculty/dominique.hanssens/Website/Kim%20&%20Hanssens%20%20June%202014.pdf, 2014.

[13] John Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142(2):573–603, 2013.

[14] Institute of Practitioners in Advertising. Touchpoints methodology. http://www.ipa.co.uk/page/touchpoints-methodology.

[15] Rahul Tandra and Anant Sahai. SNR walls for signal detection. *IEEE Journal Of Selected Topics In Signal Processing*, 2(1), Feb 2008.