Time-weighted multi-touch attribution and channel relevance in the customer journey to online purchase

David A. Wooff[a*] and Jillian M. Anderson[b]

[a] *Durham University, Department of Mathematical Sciences, Stockton Road, Durham DH1 3LE, UK*
[b] *Summit Media Ltd, Albion Mills, Albion Lane, Willerby, Hull, HU10 6DN, UK*

August 5th 2012

Abstract

We address statistical issues in attributing revenue to marketing channels and inferring the importance of individual channels in customer journeys towards an online purchase. We describe the relevant data structures and introduce an example. We suggest an asymmetric bathtub shape as appropriate for time-weighted revenue attribution to the customer journey, provide an algorithm, and illustrate the method. We suggest a modification to this method when there is independent information available on the relative values of the channels. To infer channel importance, we employ sequential data analysis ideas and restrict to data which ends in a purchase. We propose metrics for source, intermediary, and destination channels based on two- and three-step transitions in fragments of the customer journey. We comment on the practicalities of formal hypothesis testing. We illustrate the ideas and computations using data from a major UK online retailer. Finally, we compare the revenue attributions suggested by the methods in this paper with several common attribution methods.

Keywords: Sequential analysis; Metrics; Path to conversion; Clickstream; Digital marketing; E-commerce.

1 Introduction

This paper concerns statistical analysis of the routes to online purchase – known as conversion – by customers at a retail internet site. Prior to conversion, consumers typically visit several websites, including multiple visits to the final retail site, for purposes including searching, browsing and knowledge building (Moe, 2003). A typical example might begin with a customer searching for a product, narrowing down on product details, using shopping comparison sites to compare prices, checking for availability of vouchers, and so forth. This is the customer journey, also known as the *clickstream*. Retailers use a variety of online marketing channels to raise brand awareness and drive conversions; therefore, it is possible for a consumer to interact with multiple marketing channels prior to conversion. The customer journey is recorded via cookies stored on the consumer's computer. Usually, some fraction of the sale revenue is attributed to steps in the journey. Simplistically, these are monetary rewards for sites which funnel customer traffic towards the final retailer. These sites are classified as marketing channels of various kinds: display campaigns, direct email advertisements, social media such as Facebook, and so forth. One fundamental problem is to decide which fraction of revenue should be attributed to each marketing channel: in the UK, this is the weighted attribution problem; in the USA it is better known as the multi-touch attribution problem. More detailed descriptions of the process may be found in Abhishek et al. (2012) and Xu et al. (2012).

In 2012, total spend on digital advertising in the UK alone amounted to £5416 million, with annual growth of around 13% (Internet Advertising Bureau UK, 2013). In the USA, corresponding spend is presently around $40000 million (Dalessandro et al., 2012). Around 58% of UK spend is on pay-per-click (PPC) advertisements via search engines such as Google, Bing, and Yahoo. The remaining spend is on

*Corresponding author. Email: d.a.wooff@durham.ac.uk

other digital marketing channels. This sector of the economy is already of major importance, and growing, but many aspects are poorly understood, including our area of interest, the customer journey. Industry evidence is that around 65% of conversion journeys contain more than one visit to the final retail site, and about 81% contain interactions with more than one marketing channel. There is enormous interest in determining which channels are relevant to the final purchase. One reason is that the different marketing channels might be stages in, or different aspects of, an advertising campaign, and where it is desired to measure the value of each aspect in contributing to the final purchase decision. Understanding the true value of each kind of marketing channel should lead to better budget planning, to identification of crucial steps in the journey, and to improved exploitation of emerging channels.

1.1 The attribution problem

Existing methods for attributing conversions to marketing channels range from the simplistic to detailed algorithms. The most basic methods attribute the conversion to a single step in the journey, typically the first step in the journey ("first click wins") or the last step prior to conversion ("last click wins"). By only acknowledging a single channel within a conversion journey we underestimate the importance of channels which might only appear as intermediate in the journey, but which may in fact be crucial to the conversion. Multi-channel attribution assigns a proportion of the conversion revenue to each step in the journey. A recent survey suggests that 30% of retailers use single-source attribution, 34% use a multiple-source method, and 11% use an algorithm-based approach (Osur et al., 2012), with attribution depending on inferred measure of channel relevance Shao and Li (2011); Abhishek et al. (2012); Xu et al. (2012). Many current multi-channel models are subjective with weights assigned on a marketer's experience rather than data analysis.

There is no industry standard for attributing revenue and no single measure exists for comparing the many different methods available. Dalessandro et al. (2012) recommends these properties of a good attribution model: (1) *fairness* – attribution should be based on the channel's ability to influence conversions; (2) *data-driven* – attribution should be based on statistical principles, but should also utilise a retailer's knowledge of the marketplace; (3) *interpretability* – the attribution model should be transparent and sufficiently simple to be understood and implemented by all. We propose methods which satisfy these criteria, and which also takes into account temporal features in the journey. We propose a method for dealing with attribution when we have no information about the relevance of different channels to conversion behaviour, and a modification of the method when we do have such information.

1.2 The channel relevance problem

A number of algorithm-based methods use converting and non-converting journeys in order to determine the probability of each channel leading to a conversion (Shao and Li, 2011). Abhishek et al. (2012) view the journey as a funnelling process whereby customers are influenced by typically narrower funnels at each step by the marketing material. They address the likelihood to convert at each stage and then derive a valuation based on the increment that each step has on the consumer's probability to convert. They use data from an online campaign for a large car manufacturer and construct a hidden Markov model to relate advertising stages to conversion behaviour. This is useful for tightly-defined advertising campaigns. Xu et al. (2012) view the journey as a Markov process with a special structure – mutually exciting point processes – and so fit models which result in a measure of each channel's value as well as allowing prediction of conversion rate. Both these methods require conversion and non-conversion histories. These arise because each advertising stimulus can be assessed as leading definitely to a conversion or, in a time-censored sense, to a non-conversion.

Our interest is in data which is less clean. We consider only journeys which end in a conversion for a particular retailer, from whatever source. we cannot consider non-converting journeys as we have no data concerning them, as is standard in data of this kind. We cannot analyse journeys which end at a different retailer. We may analyse fragments of journeys in which a customer visits a particular retailer, but does not make purchase, but doing so requires many quite deep assumptions which reflect factors concerning a particular retailers position within the marketplace. In other words, we may analyse only what we have

observed and there is no element of experimental design involved - for that, the methods described in Abhishek et al. (2012); Xu et al. (2012) are more appropriate.

For an introduction to statistical methods to discover statistically surprising patterns in sequences see for example Agrawal and Srikant (1995); Zaki (2000a,b); Wang and Yang (2005); however this is not central to our problem of inferring channel relevance. The main focus in Agrawal and Srikant (1995) is to find customer journeys which have a specified minimum level of support, each such journey being classified as a sequential pattern; a subsidiary focus is on which items are purchased as part of the same journey. See Hahsler et al. (2005) for a more recent discussion of mining of association rules and a computer package providing tools. The problem of predicting the next step in a journey conditional on the observed history is also much studied, but is not relevant here. For predicting from a clickstream history, see for example Gunduz and Ozsu (2003); Gunduz-Oguducu and Ozsu (2006). There is also a literature on exploring web navigation behaviour; these tend to focus on website analytics. Berendt and Spiliopoulou (2000), for example, use knowledge of local web infrastructure with sequential pattern analysis to assess site design. Other researchers have used Markov and Hidden Markov models to construct predictions for customer browsing behaviour; see Jamalzadeh (2012) for an overview.

In Section 2 we describe the relevant data structures and introduce an example. In Section 3 we suggest an asymmetric bathtub shape as appropriate for time-weighted revenue attribution to the customer journey, provide an algorithm, and illustrate the method. In Section 4 we suggest a modification to this method when there is independent information available on channel relevance. In Section 5 we address the problem of inferring channel relevance from data, and suggest metrics in Section 6. We illustrate the methodology in Section 7. In Section 8, we compare the revenue attributions suggested by the methods in this paper with several common attribution methods.

2 Preliminary processing of data

2.1 Data Collection

We suppose that web analytics tools have been used to collect information about a customer's journey subsequent to a conversion. Pixel tracking is used to record each visit a user makes to a website. The marketing channel and time of each visit is recorded, along with conversion details such as sale type, sale ID, and revenue. Visits may be categorised at the marketing channel level (e.g. direct, PPC, organic search) or at a more granular level (e.g. search term, keyword, category). We make no inferences regarding journeys which may be artificially shortened via users either deleting or refusing permission to store cookies. Impressions, or views, of an advertisement may also be included in the journey sequence and assigned a value similarly to visits.

A visit duration window is applied to multiple visits from the same channel: subsequent visits are not recorded if they occur within a given timeframe thereby reducing the influence of click fraud and user behaviour (e.g. page refresh, navigation confusion). Industry standards set the visit duration window at 10 minutes for marketing channels. Furthermore, a maximum time between a visit and conversion is imposed, and will be referred to as the cookie window. The choice of cookie window is subjective, but guided by industry expertise. For retail, 31 days is commonly employed.

We exclude journeys reaching a terminus such as site registration or booking an appointment. We assume where necessary that abandonment of a journey without conversion is final within a given time period. This is an approximation as some customers do continue their journeys after long breaks. Note that some journeys which do not end in online conversion may end in offline conversion, with customers visiting a store to purchase a product identified online. This is presently excluded from our analysis.

Mathematically we will view the different possible visits as nodes in a sequence. The definition of what constitutes a visit source depends on the requirements of the retailing company. Sometimes this will be at a very fine level of detail, such as named weblinks. At other times the sources may have been classified by the retailer into a smaller number of channel categories such as 'direct', 'email', etc, as it deems appropriate. This is the case in the example we discuss in Section 7.

2.2 Data Processing

Suppose we observe a sequence of customer visits to a retail website made at times $T_1, T_2, \ldots, T_k$. We make an assumption that visits that occur further back in time than a specified amount T_{max} are not relevant to the current conversion. Analysis of the journey database allows a retailer-specific T_{max} to be set. The journey lengths, $T_k - T_1$, of all journeys in the database are analysed, with the 90^{th} percentile chosen as T_{max}. Journeys where $T_k - T_1 > T_{max}$ are truncated at the visit T^*, where $T^* \leq T_{max}$.

We also make the assumption that time gaps larger than a specified amount T_Δ imply separate journeys. Thus, if any adjacent times satisfy $T_j - T_i > T_\Delta$, we end one journey at T_i and start another at T_j. All transition times $(T_j - T_i)$ within the journey database are analysed, with T_Δ set at the 90^{th} percentile. For the purposes of this article we consider only one value for T_Δ, however, it is understood that T_Δ may vary depending on the sequence of marketing channels. Journey fragments prior to T_i are not considered in this article.

A maximum number of visits V_{max} might also be imposed, in that journeys with number of visits exceeding V_{max} are assumed to be due to tracking discrepancies and are removed from the analysis.

Imposing a T_{max} and T_Δ results in left-censoring of the data. The main implication is that data concerning the first click is lost. In analysing such data, the implicit assumption is that T_1 is either genuinely the start of a new journey, or a click made in the same journey but with the preceding click so distant in time that it is deemed irrelevant. For analysis of journeys which end in conversion, the use of a time gap threshold may result in early parts of the journey being discarded. For data where conversion behaviour is an outcome, journeys might be separated into non-converting and converting fragments, and the correlation between the two may be lost.

2.3 Example

Table 1: Journeys and conversion revenues for 25 customers, minimum two-step journeys with $S = 11$. Figures given are times of visit rounded to the nearest minute and starting time arbitrarily at $T_1 = 0$ for each customer.

i	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	Revenue
1	0	19	70	106	106							869
2	0	113										309
3	0	0	0	37	37	114						50
4	0	0	118									329
5	0	1	7	7	7	122	122					280
6	0	84	137									322
7	0	53	111	144	144	144						196
8	0	13	13	14	137	142	147	148				100
9	0	0	136	149	149	149	149	149				244
10	0	25	77	79	79	79	167	167	167	167	167	378
11	0	0	50	169								494
12	0	172										205
13	0	178	178	178								340
14	0	52	179	179								370
15	0	0	180	180								136
16	0	0	79	181								1289
17	0	33	39	191								160
18	0	198										213
19	0	14	27	99	115	120	125	204				249
20	0	6	139	145	150	153	153	167	206			163
21	0	77	216	216	218	218	218					330
22	0	90	117	121	151	151	241	243	243	247	247	95
23	0	6	126	251	251	251	251					150
24	0	263	263									270
25	0	20	22	23	23	153	247	272				239

Consider the fragment of data shown in Table 1. Data are taken from a sample of customer conversions made on a leading multichannel retail website. Each journey has a starting time T_1, and a number of visits in sequence with time recorded. Also shown is the amount of conversion, the revenue attributed to each journey. These data are reported to two decimal places, but shown rounded in the table. The focus of analysis for this data set is the route to conversion. A maximum journey length T_{max} of 30 days was used, and visits made before T_{max} are removed. A time gap threshold of $T_\Delta = 14$ days was also used, and fragments of any journey with at least such a time gap were discarded. Each customer journey is analysed separately and only time since start of journey is assumed relevant. As such, we fix $T_1 = 0$ for each journey. A maximum number of visits in the journey was also set at $V_{max} = 11$; journeys with more than 11 visits were removed. More than 95% of journeys in the database contained 11 visits or fewer. It is, of course, possible to explore the implications of different choices of T_{max} and so forth, but this is outside the scope of this paper.

The data subset contains visits from a number of channels which may be split into varying degrees of granularity. Natural search channels may be split by search partner (e.g. Google, Bing) or category (e.g. brand, non-brand). Affiliate channels may be categorised according to type (e.g. cashback, voucher codes); this is particularly important for understanding the value of marketing campaigns within the context of attribution and budget forecasting. Visits via individual comparison sites are also included. Finally, for account optimisation, PPC visits may be split at the keyword level, where keyword can be broadly interpreted as meaning a search word or phrase. Identifying keywords which have a strong influence on likely final conversion is a crucial aspect of digital marketing performance. Visits which are not classified into a specific channel are classed as "unlisted referrers" and could be excluded from the attribution model, or assigned a weight of zero; for discussion see Section 8.

This sample of data exhibits features typical of the problem. Journeys vary in length of time. Significant time can be spent on one visit, or the journey can be relatively time-homogeneous. There are two two-step journeys. Instances where successive visits are within the same minute as the previous click (for example, see journey 3 in Table 1) represent visits either by a different channel or search query and are not to be interpreted as page refresh errors. Single visit journeys are assigned revenue and removed from the attribution database after data cleaning.

3 Naive time-weighted Revenue allocation

Suppose we observe the customer journey $X_{(1)} \to X_{(2)} \to \ldots \to X_{(k)}$, $2 \leq k \leq S$, with conversion at node $X_{(k)}$ resulting in revenue R, and where S is some truncating choice. Suppose we visit node $X_{(i)}$ at time T_i, so that the journey begins at T_1 and ends at T_k. Suppose also that we have no information concerning the relative importance of nodes in the journey. The problem is to attribute the revenue to the nodes in the journey, or equivalently to value each node. There are many views as to how we might do this. One is to attribute all revenue to the last node in the journey, known as *last click wins*. This corresponds to the view that the journey itself is irrelevant and that the customer would have arrived at node $X_{(k)}$ irrespective of starting point. Another view is to attribute all revenue to the first node in the journey, known as *first click wins*. This corresponds to the view that once the journey has started at $X_{(1)}$ the journey will end inexorably with a conversion at node $X_{(k)}$. A third view is that all nodes in the journey count equally towards the final conversion, in which case revenue might be attributed equally to each node. There are many other views which suggest that clicks closer to conversion should have a higher weighting. These lead to weights based on monotonically rising functions, for example positive linear and exponential.

In discussion with digital marketing experts at the collaborating company, none of these views is felt to be reasonable. Instead, they suggest the following plausible structure. We value recent clicks highly, especially the most recent click. We value the initiating click highly, but less highly than the last click. We value intervening clicks not highly if they are quite distant in time, and less than the initiating click. We regard clicks close in time to the last click as being highly relevant. This suggests that the shape of value which we wish to allocate to clicks in the journey might have an asymmetric bathtub shape, with the rim of the bath lower at the left-hand side. Such bathtub shapes are common in survival analysis, through representing hazard functions. We now consider how to construct such a shape for this application. A simple asymmetric bathtub shape, constructed using a beta distribution, is shown in Figure 1 for Journey

10 of Table 1.

3.1 Theory

The beta distribution is of the form $f(x) = kx^{a-1}(1-x)^{b-1}$, $0 < x < 1$, where k is a normalising parameter which is of no interest in this context. The parameter choices $0 < a < b < 1$ lead to asymmetric U-shaped distributions with a higher rim at the right-hand side. Other parameter choices can lead to J-shaped and unimodal distributions. Although the distribution is defined on the interval $(0, 1)$, it is trivial to transform journey time (T_1, T_k) to $(0, 1)$ and back again. In fact we will transform not to $(0, 1)$ but to $(\epsilon, 1-\epsilon)$ to avoid infinities at the asymptotes. Experience shows that a good choice is $\epsilon = 0.01$. The minimum of the distribution occurs at

$$\gamma = \frac{a-1}{a+b-2}, \tag{1}$$

so that $\theta = f(\gamma)$ will be the smallest possible weight given to any click.

We need to make choices about the relative values of clicks. Let θ_L be the relative value of the last click in the journey as compared to the first click in the journey. Let θ_F be the relative value of the first click in the journey as compared to θ, potentially the value assigned to the least valuable click in the journey.

The choices of θ_F and θ_L will depend on context. In discussion with our marketing collaborator, it was felt appropriate to deem the last click as worth about four times as much as the first click, and the first click as worth about twice the minimum value we would wish to assign. That is, $\theta_L = 4$ and $\theta_F = 2$, so that the last click is worth $\theta_L\theta_F = 8\theta$, eight times as much as the least valuable click. Such choices are unavoidable. For example, the judgement that all clicks should be evenly weighted corresponds to $\theta_F = \theta_L = 1$. Similarly, where there is an attribution which rises linearly in value from first click to last click, the underlying choice is $\theta_F = 1$ and θ_L is proportional to the slope of the chosen line.

Given these assumptions, we now generate parameter values for our beta distribution. We have

$$\theta_L = \frac{f(1-\epsilon)}{f(\epsilon)} \Rightarrow a = b + v, \tag{2}$$

where

$$v = \frac{\log \theta_L}{\log (1/\epsilon - 1)}.$$

Note that $v > 0$ in order to obtain a higher rim at the right-hand side. We have also

$$\theta_F = \frac{f(\epsilon)}{f(\gamma)} = (\frac{\epsilon}{\gamma})^{b+v-1}(\frac{1-\epsilon}{1-\gamma})^{b-1}, \tag{3}$$

where we can re-express γ via (1,2) as

$$\gamma = \frac{b+v-1}{2b+v-2}.$$

This gives a highly non-linear equation in b, which may be solved numerically. The constraints of the numerical solution are that $0 < b < 1 - v$. This follows as we require $a < 1$ in order to obtain a U-shape. An algorithm for attributing revenue to a channel is thus as follows.

(i) Choose θ_F and θ_L. Fix $\epsilon = 0.01$. Compute v.

(ii) Solve (3) for b and determine a via (2).

(iii) For Journey J with revenue R_J to attribute, transform the click times $T_1, T_2, \ldots, T_k$ linearly to $(\epsilon, 1-\epsilon)$. This gives transformed time values $T_1^* = \epsilon, T_2^*, \ldots, T_k^* = 1 - \epsilon$. Evaluate $w_i = f(T_i^*)$ for each transformed time. The proportion of revenue attributed to the channel clicked at time T_i is $w_i^* R_i$, where

$$w_i^* = \frac{w_i}{\sum_{i=1}^k w_i}.$$

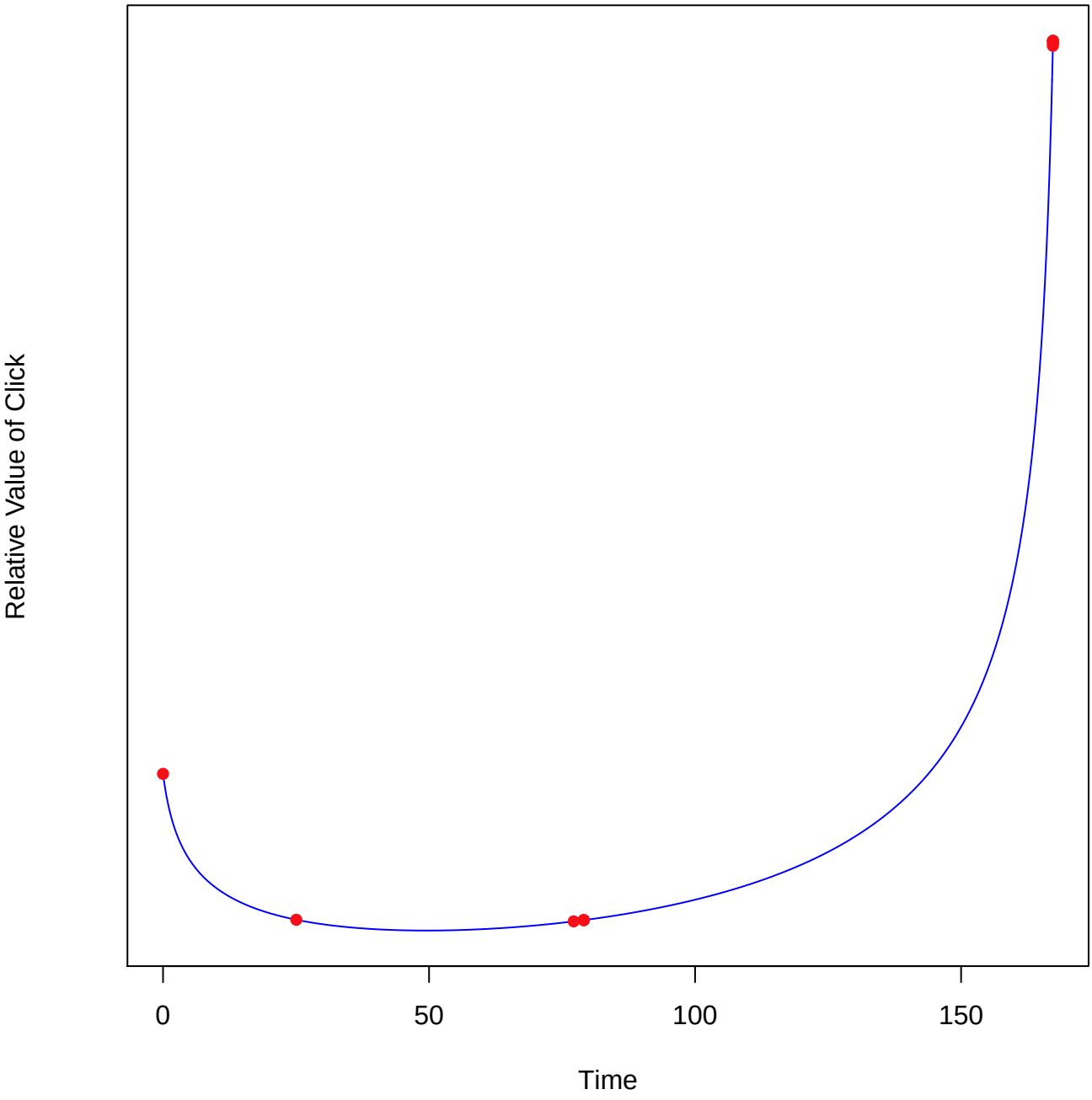

Figure 1: Simple bathtub model for click value, with clicks for Journey 10 marked.

There may be journeys for which all recorded click times are the same, perhaps because of rounding. In this case the rescaling to $(0, 1)$ fails and it is simplest to give equal weight to all clicks in such journeys.

The steepness of the bath rims is governed by ϵ, for which we suggested an appropriate default value of 0.01. Smaller values of ϵ imply steeper behaviour at the asymptotes, with the consequence that the last click will be valued relatively more than the penultimate click, and the first click relatively more than the second click. If this level of detail is deemed worth pursuing, the implications of different values of ϵ can be shown to digital marketing staff and an appropriate alternative value chosen, but this needs to take into account the proximity of clicks in unscaled time.

3.2 Example

Table 2: Weights and revenue attributions for Journey 10 of those shown in Table 1.

i	Time of click, T_i	Weight, w_i^*	Revenue attributed
1	0.00	0.042	16.03
2	25.05	0.023	8.56
3	77.18	0.022	8.31
4	79.05	0.022	8.36
5	79.09	0.022	8.36
6	79.09	0.022	8.36
7	167.25	0.169	63.71
8	167.26	0.169	63.92
9	167.27	0.170	64.13
10	167.27	0.170	64.13
11	167.27	0.170	64.13
		1.000	378.00

For our data set we choose $\theta_F = 2$ and $\theta_L = 4$. Solving with these choices we obtain $a = 0.739$ and $b = 0.437$. The curves obtained are shown in Figure 2 for journeys 2,3,10,25. For journey 10, the weights and revenue attribution are shown in Table 2. The revenue attributions for all journeys are shown in Table 3. Note that attributions must now be accumulated over channels (or at a more granular level depending on purpose); for example the clicks at T_1 and T_2 for a journey could correspond to the same channel. One feature evident in this data set is multiple clicks close in time, and so which attract similar revenues. In principle it is not difficult to provide more sophisticated methods which could take into account subjective judgements concerning clicks close in time. For example, one might wish to discount all but the most recent of a group of clicks occurring in a narrow time range.

4 Informed revenue allocation

In this section we discuss weighted attribution when we also have information about the relative importance of different nodes. Judgements about relative importance may be made directly. For example, in the context of online marketing a company might wish to value PPC channels more highly than natural search or email marketing. A number of researchers, see for example Shao and Li (2011); Abhishek et al. (2012); Xu et al. (2012), have provided measures of channel value relating directly to probability of conversion. This requires data on converting and non-converting journeys. Where we have data only on converting journeys, we provide a method in Section 5 to infer channel relevance based on sequential data analysis of journey fragments.

Whether channel value is inferred or specified, we suppose that the relative values of the n channels are $u_1, u_2, \ldots, u_n$, where $\sum_{i=1}^{n} u_i = 1$. There are different possible ways of merging weights due to time and weights due to channel value. The simplest is to compound the two sets of weights and then re-normalize. Thus, suppose that $a_{(1)}, a_{(2)}, \ldots, a_{(k)}$ are the weights suggested by time of click for a k-step journey. These

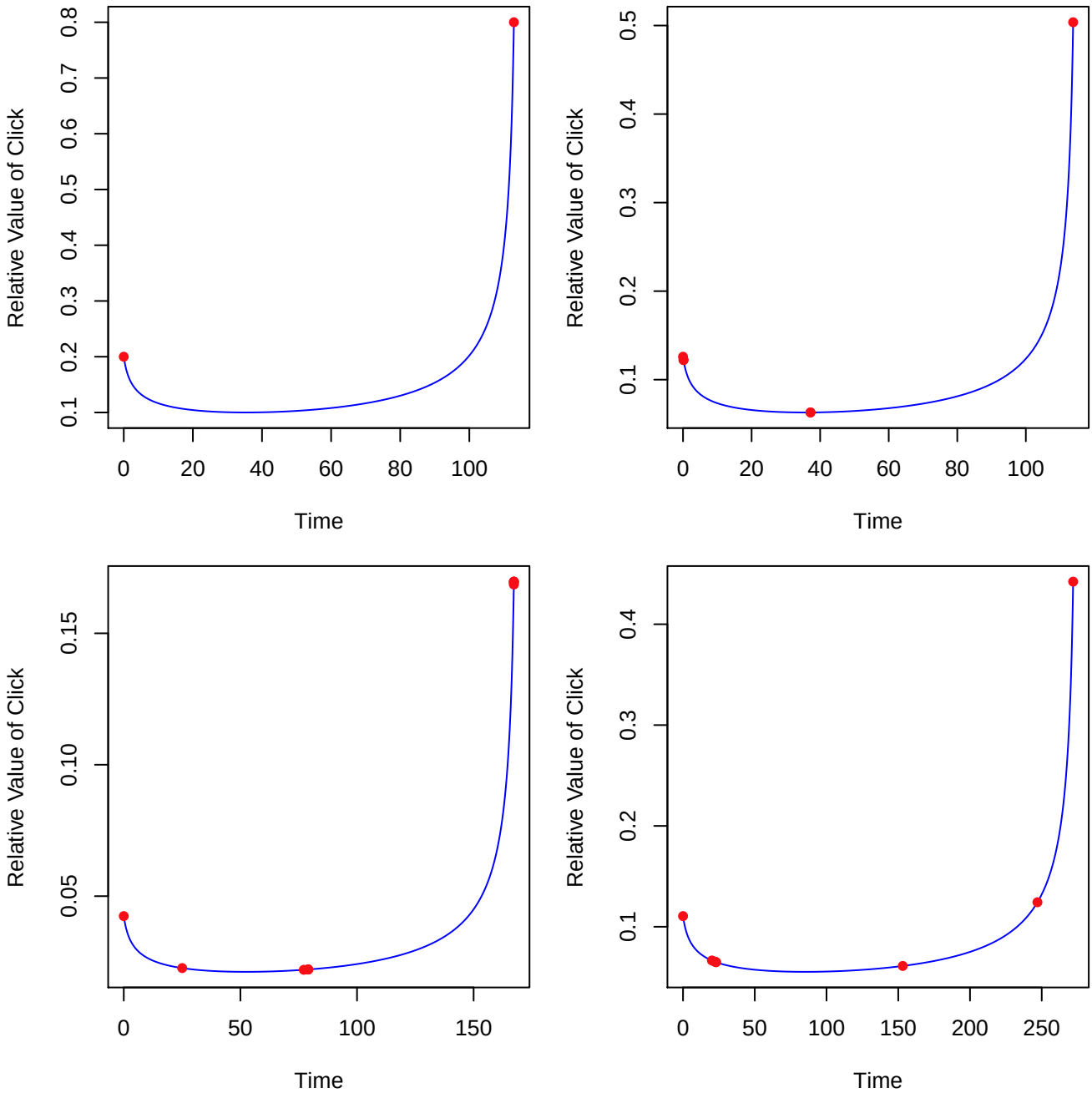

Figure 2: Value of clicks for journeys 2,3,10,25. Beta function parameters are $a = 0.739$, $b = 0.437$. Last click is worth $\theta_L = 4$ times as much as first click. First click is worth $\theta_F = 2$ times as much as the minimum possible.

Table 3: Revenue attributions for 25 customer journeys. Figures given are attributions of revenue to the channel clicked at that time, rounded to the nearest integer.

i	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	Revenue
1	86	45	52	342	344							869
2	62	247										309
3	6	6	6	3	3	25						50
4	55	55	219									329
5	24	21	15	15	15	95	95					280
6	58	33	231									322
7	14	7	10	55	55	55						196
8	8	5	5	5	10	13	23	32				100
9	11	11	12	42	42	42	43	43				244
10	16	9	8	8	8	8	64	64	64	64	64	378
11	76	76	38	304								494
12	41	164										205
13	26	104	104	105								340
14	39	20	156	156								370
15	14	14	54	55								136
16	198	197	102	792								1289
17	26	14	14	106								160
18	43	170										213
19	30	18	16	16	16	17	17	119				249
20	17	12	10	11	11	11	11	13	66			163
21	17	9	47	48	69	70	70					330
22	5	3	3	3	3	3	10	12	12	21	21	95
23	8	6	4	33	33	33	33					150
24	30	120	120									270
25	26	16	16	16	16	15	30	106				239

weights are derived using the bathtub method of Section 3, the linear method, or any other desired method. Let $u_{(i)}$ be the value of the i^{th} node clicked. The compounded weight for the node clicked on the i^{th} step of the journey is then

$$a^*_{(i)} = \frac{u_{(i)}a_{(i)}}{\sum_{j=1}^{k} u_{(j)}a_{(j)}}. \tag{4}$$

Thus, an attribution to the node clicked on step i of the journey which is both time-weighted and value-weighted is given by multiplying weight $a^*_{(i)}$ by journey revenue.

5 Inferring node value using sequential analysis

We now address how we can determine the relevance of different channels in a customer journey which ends in a sale. Clearly, the final nodes in the journey are important, but time-weighted attribution of revenue will emphasize these anyway. Therefore in what follows, we will derive relevance of node independently of early or late position in the journey. The proportion of nodes visited across all journeys offers a simple measure of relevance. However, the key is to measure the importance of a node in terms of moving from one to another. Thus, we need to focus on the probabilities of transition. Thus, suppose the customer journey includes the sequence $A \to B \to C$. The questions to answer are: how relevant is the intermediary node B, and would the customer have reached C from A regardless? Ideally we would like to represent customer journeys using probabilistic networks such as Bayesian belief networks; however these are inadequate for the task, partly because they are directed networks and partly because their inherent Markov properties cannot handle multinode histories.

5.1 Principles and notation

We employ a notation based on that of Agrawal and Srikant (1995). Our concern is with journeys which interact with a fixed number, n, of nodes $X_1, X_2, \ldots, X_n$ in some order. In common with the digital marketing community, we call these interactions clicks or visits. The journeys may contain loops, repeated fragments, and so forth. There may or may not be single-click journeys. We described cleaning of the data in Section 2, noting that journeys are typically left-censored to the most recent S steps, so that S is the maximum sequence length. Let $A \to B$ mean the direct transition from node A to node B. Let $A \Rightarrow B$ mean any one-step or two-step transition from A to B. The notation $\bar{B}$ means any node except node B. Let $N\{ij\}$ be the number of times the direct transition $X_i \to X_j$ occurs. We extend the notation to longer sequences, so that $N\{ijk\}$ is the number of times the subsequence $X_i \to X_j \to X_k$ appears.

5.1.1 Cyclic sequences

Ideally we want to deal with uniquely classified nodes, for example a unique landing page within a retail website. In this situation it makes sense to treat a sequence $(A \to A \to B)$ as equivalent to the sequence $A \to B$, such that the sequence then contains no immediate loops, and we do not distinguish between one interaction and more than one interaction with the node. This principal seems to extend naturally to subsequences. That is, $(A \to B \to A \to B)$ might be considered equivalent to $(A \to B)$. Ultimately this is the restriction that the sequence not be cyclic. However, there are difficulties in working with this interpretation. First, checking for cyclicity is non trivial (Wang and Yang, 2005). Secondly, if the journey is actually cyclic, we need to decide which part of the journey to disregard. In determining channel relevance, the possible nodes in many examples happen to be crude bins representing channel type rather than a granular classification. Therefore it is perfectly feasible to observe a journey such as $A \to A$, for example from one shopping comparison site to another. Thus in the remainder of this account we make no sequence restrictions and allow sequences to be cyclic.

6 Metrics based on three-step transitions

We must take into account at least three-step transitions. This is already challenging; dealing with all possible four-step transitions, where we would have to consider all possible intermediary pairs of nodes, is daunting. Thus we restrict attention to two steps and three steps. We will ignore whether fragments of a journey occur early or late. We will remove single-step journeys from consideration as these are not informative for transitions. For each sequence we now construct two-step and three-step fragments as follows. Take as an example the sequence:

$$A \to B \to C \to B \to E \to D.$$

This contains these two-step fragments:

$$A \to B,\ \ B \to C,\ \ C \to B,\ \ B \to E,\ \ E \to D,$$

and these three-step fragments:

$$A \to B \to C,\ \ B \to C \to B,\ \ C \to B \to E,\ \ B \to E \to D.$$

Thus a journey with length s contains $s-1$ two-step fragments and $s-2$ three-step fragments. Clearly by breaking down journeys into fragments we are losing much information, particularly about more complicated journeys.

6.1 A metric for intermediary node value

We now propose a metric for channel relevance. A natural metric for the relevance of a node B in journeys from A to C is the proportion of such journeys which pass through B, which we estimate by the observed proportion:

$$\Lambda_{ABC} = \frac{N\{ABC\}}{N\{AC\} + N\{ABC\} + N\{A\bar{B}C\}}.$$

This is the observed conditional probability that any two- or three-step journey from A to C passes through B, $P(A \to B \to C | A \Rightarrow C)$. If this value is small, it suggests that B is not an important way of reaching C from A. If this value is large, it suggests that B is an important intermediary. More formally, for a (source, intermediary, destination) triple this metric is:

$$\Lambda_{ijk} = \frac{N\{ijk\}}{N\{ik\} + \sum_{j=1}^{n} N\{ijk\}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, \quad k = 1, \ldots, n.$$

A general measure of the value of node X_j is then given by averaging over all source and destination nodes:

$$\lambda_j = \sum_{i=1}^{n} \sum_{k=1}^{n} \Lambda_{ijk}, \quad j = 1, \ldots, n. \tag{5}$$

Note that these measures do not sum to unity:

$$\sum_{j=1}^{n} \lambda_j = \sum_{i=1}^{n} \sum_{k=1}^{n} \frac{v_{ik}}{1 + v_{ik}} \leq 1, \quad v_{ik} = \frac{\sum_{j=1}^{n} N\{ijk\}}{N\{ik\}}, \tag{6}$$

where v_{ik} is the ratio of indirect to direct transitions for node pair (i, k). This sum depends on the total number of direct two-step transitions and the total number of exactly three step transitions for each node pair. Thus, a normalized metric is given by

$$\tilde{\lambda}_j = \lambda_j / \sum_{j=1}^{n} \lambda_j. \tag{7}$$

As a simple average, (7) does not take into account the volumes of journeys between pairs. As such, a refinement is to weight according to volume. Typically we deem the destination node to be more relevant than the source node so that it can be appropriate to weight according to the volume of destination nodes. It is trivial to weight according to other choices of volume. Let z_k be the number of two-step journeys which end at node k, and let z_0 be their sum, i.e. the total number of two-step journeys. That is,

$$z_k = \sum_{i=1}^{n} N\{ik\}; \quad z_0 = \sum_{k=1}^{n} z_k.$$

Then $\tilde{z}_k = z_k / z_0$ is the proportion of two-step journeys which end at node X_k, with $\sum \tilde{z}_k = 1$. This gives a relative measure of the volume of destination node X_k. Now a plausible measure of the value of intermediary node X_j is

$$r_j = \sum_{i=1}^{n} \sum_{k=1}^{n} \tilde{z}_k \Lambda_{ijk}, \quad \tilde{r}_j = \frac{r_j}{\sum_{j=1}^{n} r_j}, \tag{8}$$

where the latter is normalized. If we also wanted to take into account the value of the source node X_i via some weight $\tilde{y}_i$ with $\sum \tilde{y}_i = 1$, then (8) is easily extended to

$$r_j^* = \sum_{i=1}^{n} \sum_{k=1}^{n} \tilde{y}_i \tilde{z}_k \Lambda_{ijk}, \quad \tilde{r}_j^* = \frac{r_j^*}{\sum_{j=1}^{n} r_j^*}. \tag{9}$$

In our later example, we use (8), so that the normed value $\tilde{r}_j$ is our principal metric for determining the relevance of intermediary node j.

6.2 Metrics for the journey relevances of initiating and terminating nodes

We may develop similar metrics to value different features of a journey. The two most useful are as follows. The proportion of journeys from B to C which are preceded by A is estimated by their observed proportion:

$$\Phi_{ABC} = N\{ABC\} / N\{BC\}.$$

If this value is small, it suggests that A is not an important way of starting $B \to C$ journeys. Note that this metric ignores direct AC transitions, and so can't be used as a measure of the importance of A in the journey to C alone. The proportion of journeys from A to B which continue on to C is estimated by:

$$\Psi_{ABC} = N\{ABC\}/N\{AB\}.$$

A high proportion suggests that most customers did not find B a suitable place to stop. A high proportion could also imply that B is a natural way of getting to C. For each of these metrics, we may weight and normalize according to volume as desired.

6.3 Hypothesis tests and tests of uniformity

Conditional on ending at X_k and starting at X_i we have $N\{ik\} + N\{ijk\} + N\{i\bar{j}k\}$ possible journeys of which $N\{ijk\}$ went through X_j. This is like imagining that someone at X_i wants to get to X_k but isn't sure how to get there. We might then assume that the total number who end up at X_k via X_j is binomial $b(N, p)$ with parameters $N = N\{ik\}+N\{ijk\}+N\{i\bar{j}k\}$ and unknown probability p estimated as λ_{ijk}. This leads naturally to a standard error for the estimate as

$$s_{ijk} = \sqrt{\frac{\lambda_{ijk}(1-\lambda_{ijk})}{N\{ik\} + N\{ijk\} + N\{i\bar{j}k\}}}. \tag{10}$$

We can do this for each node separately, and for all n^2 combinations of beginning and ending nodes. However this ignores a degree of correlation between the measures. Instead, conditional on N being fixed, we can treat the outcomes as multinomial for a fixed starting and ending pair. The outcomes then are all routes which pass through an intervening node plus the direct route transitions. Thus, for any pair of nodes X_i, X_k, let $N = N\{ik\} + \sum_j N\{ijk\}$. This is the total number of routes from X_i to X_k either direct or via one intervening node. Now let p_0 be the probability that a route starting at X_i and determined to get to X_k goes directly, and p_j the probability that such a route passes through node X_j. These probabilities may be routinely estimated using the multinomial distribution. A test of uniformity is given by a Chi-squared test. However, the test: $H_0 : p_0 = p_1 = \ldots = p_n$ versus the alternative that at least one p_i differs is not so interesting. This is because we would generally expect a much higher probability p_0 for the direct transition. Therefore, attention could more reasonably focus on the hypotheses such as

$$H_0 : p_1 = p_2 = \ldots = p_n, \qquad \text{or} \qquad H_0 : p_0 = \sum_j p_j, \text{ or} \qquad H_0 : p_i > \delta,$$

i.e. that the indirect transition probabilities are all equal, or that the indirect transitions are as important as the direct, or that individual proportions p_i exceed some threshold δ. Tests on linear contrasts of multinomial proportions are considered in Goodman (1965), who also constructs simultaneous confidence intervals for them. This is summarised as method $S_2(N)$ of Hou et al. (2003), who considers the performance of a number of similar constructions. A problem is the number of tests we would need to carry out: if there are n nodes, we would need to carry out n^2 tests for each set of hypotheses, which would be correlated, and then it is doubtful that we would wish to analyse the results of all of these in detail. Finally, the nature of the data implies very unbalanced sample sizes. Some of the pairs could be associated with such large volumes of data that spuriously small p-values result, whereas for others there may be no or little data. As such, an effect-size approach (Wooff and Jamalzadeh, 2013) may be more useful. A graphic such as the stars plot shown later can also be a useful visual cue as to intermediary node relevance for specific source and destination pairs.

7 Computation and illustration

As an example, we explore data from a major UK online retailer. This records 58667 journeys of which 27420 are single-click and 31247 have at least two clicks. 17841 journeys have at least three clicks. We limit to the most recent $S = 19$ steps of any journey. Each click is classified as belonging to one of nine channels

Channel	Code	Freq	Prob
Affiliates	Aff	3841	0.1401
Banner	Ban	62	0.0023
Price Comparison	Comp	818	0.0298
Listed Referrer	List	96	0.0035
Natural Search (Other)	Nat	1954	0.0713
Natural Search (Brand)	NatB	14081	0.5135
Pay-per-click	PPC	2174	0.0793
Pay-per-click (Brand)	PPCB	2543	0.0927
Unlisted Referrer	Un	1851	0.0675
All		27420	1.0000

Table 4: Single-click-journey probabilities

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	4374	53	289	24	579	1892	467	516	400
Ban	52	40	15	2	24	162	37	28	71
Comp	476	25	342	3	194	528	183	144	62
List	40	3	7	35	18	189	20	30	23
Nat	1052	32	208	26	2199	2239	682	483	277
NatB	3172	174	511	144	2023	29320	1586	1924	1385
PPC	939	44	262	25	839	2161	2719	666	288
PPCB	857	45	135	27	434	2148	499	2955	506
Un	567	87	61	13	222	1122	251	344	6387

Table 5: Bivariate transition counts, $N\{ik\}$

Sender, i	Receiver, k								
	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	274	8	33	7	94	577	68	57	29
Ban	4	6	2	0	0	83	5	6	13
Comp	38	5	34	0	30	177	23	19	4
List	1	0	1	9	8	92	3	5	7
Nat	102	9	29	8	350	829	113	86	50
NatB	701	88	144	82	678	14549	527	555	410
PPC	100	14	34	4	138	789	346	97	21
PPCB	120	7	23	7	81	723	79	323	47
Un	40	7	5	4	36	394	33	34	174

Table 6: Counts of transitions from sender to receiver via the **NatB** node (B), $X_i \to$ **NatB** $\to X_k$.

as shown in Table 4. This shows that a high proportion of single-click journeys for this retailer at this time were branded natural search, coded as **NatB**.

We now take every journey and count all the pair occurrences. The counts are shown in Table 5 and plotted in Figure 3, with area proportional to count. There are 83387 pairs. Again, the **NatB** node dominates, and there are several nodes which carry very little traffic. The conditional bivariate transition matrix, plotted in Figure 4, shows the proportion being received by each receiving node given the sending node, i.e. $P(X_i \to X_j | X_i$ is the sender). Probabilities across rows sum to one. (Interpreting columns is not sensible.) There are two obvious deductions we make from Figure 4. First, there is a high probability of clicking on the same kind of channel, i.e. $A \to A$, regardless of where you start. This is evidenced by a strong diagonal pattern. Secondly, there are high conditional probabilities of ending in nodes **Aff** and **NatB** regardless of starting node, as evidenced by high probabilities in those columns. Understanding of such patterns is useful for marketing design and so forth, but is not our focus here.

We next address journey triples. There are $n = 9$ possible intermediary nodes for each sender and receiver. Table 6 counts the number of triples where the intermediary node is **NatB**. There were overall 94 journeys **Aff** $\to$ **NatB** $\to$ **Nat** and no journeys **List** $\to$ **NatB** $\to$ **Ban**.

We may assess whether the starting and ending nodes of two-step patterns resemble in frequency the starting and ending nodes of three-step patterns. To do this, count for each pair of nodes i, k, the number of direct transitions $N\{ik\}$ and the number of indirect transitions $\sum_j N\{ijk\}$ via any intermediary node. Table 7 shows the number of indirect transitions, and Table 8 shows the proportions v_{ik} of indirect transitions to direct transitions obtained by dividing Table 7 by Table 5. On average this ratio is 66%. We can carry out a Cochran-Mantel-Haenszel test to test for differences between these two tables. This test is strongly significant; we conclude that the two tables have different patterns. However, the statistical significance is partly the result of very large sample sizes. Indirect transitions to **Aff**, **Comp**, **Un** tend to occur relatively less than average, and indirect transitions to **Ban**, **List**, **PPC** tend to occur relatively more than average. Examination of residuals shows that these are weak effects.

7.1 Metrics

We now apply the metrics suggested earlier. We take as an example direct and indirect routes from $A =$ **Aff** to $C =$**Nat**. Table 9 shows the counts and calculations for $A \Rightarrow C$; in all there are $n \times n = 81$ such tables to construct for this data set. A visualization of the flows for this pair is shown in Figure 5. The top node is the source node $A =$ **Aff**. 10.3% of all journeys begin with this node, which is drawn with area proportional to 10.3% as a visual cue to its importance as a starting node. The destination node, $C =$ **Nat**, is drawn with area proportional to 7.8%, reflecting the volume of clicks for this node. Shown are the direct and indirect routes. The area of central nodes is not meaningful, these are simple labels. The widths of lines connecting nodes shows how much traffic is flowing between them. The thickest width is between $A =$ **Aff** and $B =$ **Aff**, representing 2749 clicks from A to B which then proceed to another node. The text at the bottom gives the proportion of journeys reaching the destination directly and indirectly. We see that most

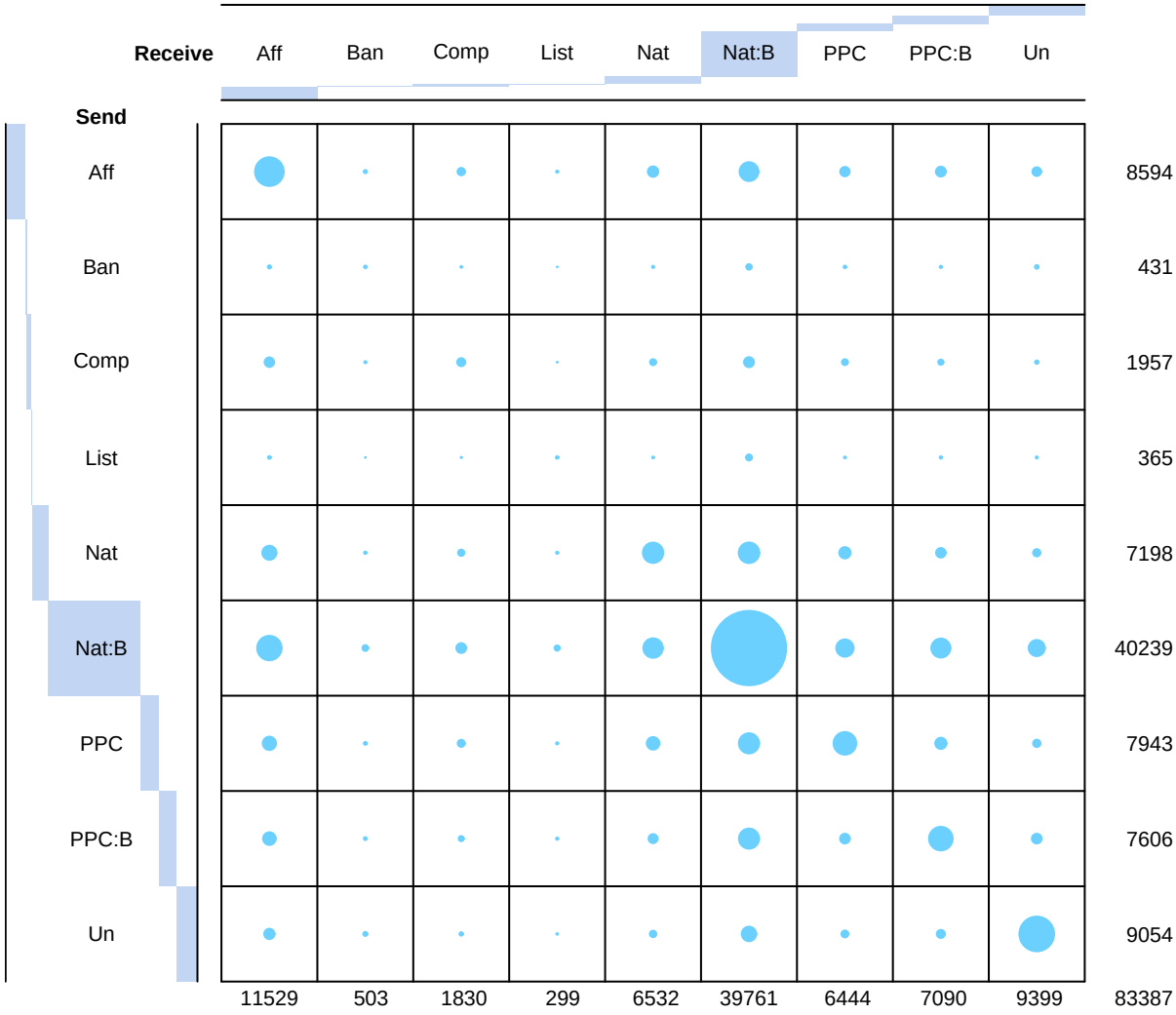

Figure 3: Direct transition frequency as a proportion of all journeys.

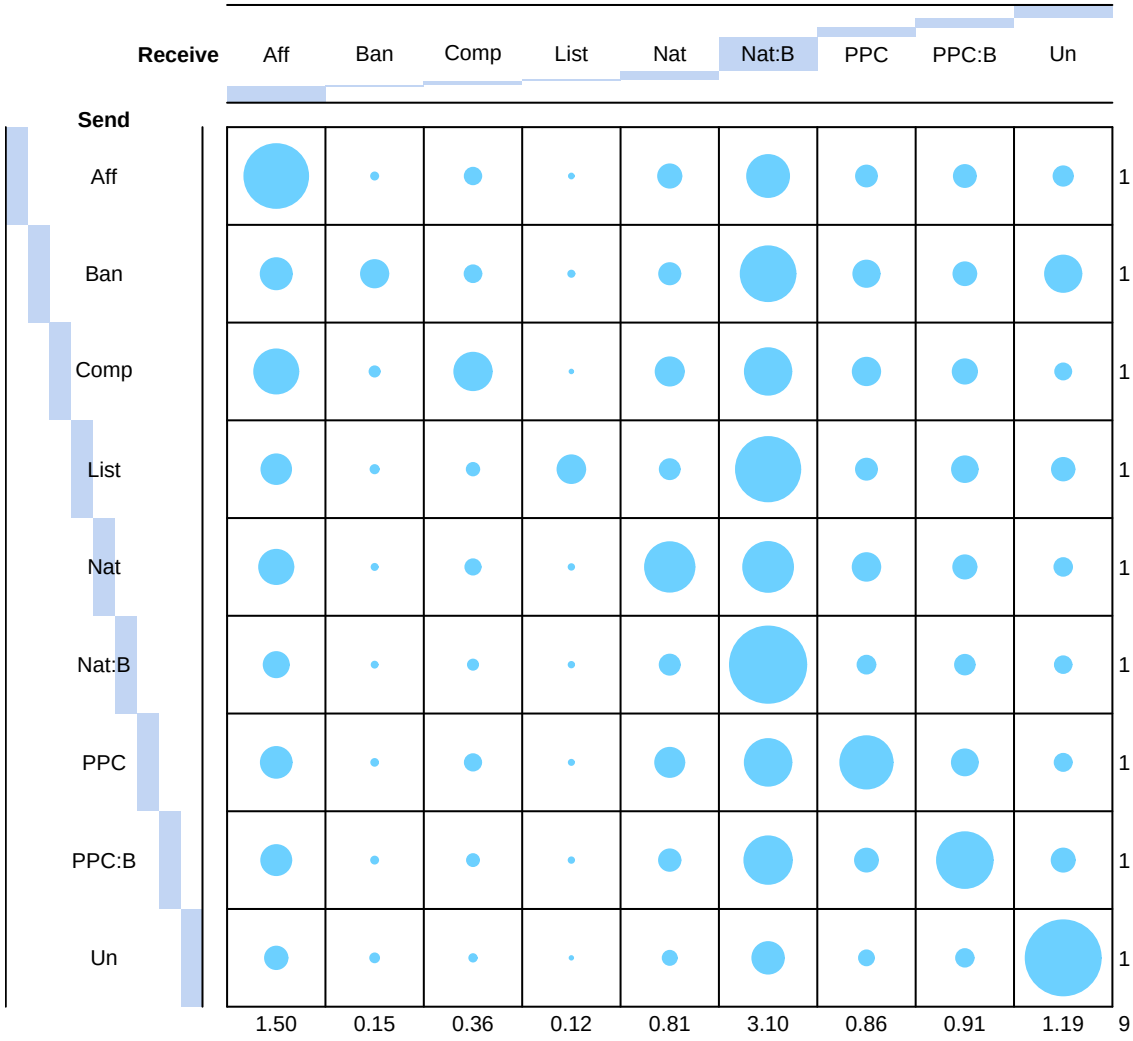

Figure 4: Transition probability given that current state is the sending node.

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	2400	44	187	21	405	1249	339	333	231
Ban	30	29	12	1	14	115	30	21	52
Comp	226	17	200	2	134	305	124	83	38
List	18	3	3	25	13	132	15	24	14
Nat	529	23	130	19	1470	1367	512	321	161
NatB	1551	134	314	117	1399	18188	1136	1209	775
PPC	501	36	164	13	584	1326	1742	447	173
PPCB	471	33	90	22	281	1240	362	1828	289
Un	292	60	36	9	146	710	177	238	4626

Table 7: Counts of transitions from sender to receiver via any intermediary node, $\sum_j N\{ijk\}$.

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	0.55	0.83	0.65	0.88	0.70	0.66	0.73	0.65	0.58
Ban	0.58	0.72	0.80	0.50	0.58	0.71	0.81	0.75	0.73
Comp	0.47	0.68	0.58	0.67	0.69	0.58	0.68	0.58	0.61
List	0.45	1.00	0.43	0.71	0.72	0.70	0.75	0.80	0.61
Nat	0.50	0.72	0.62	0.73	0.67	0.61	0.75	0.66	0.58
NatB	0.49	0.77	0.61	0.81	0.69	0.62	0.72	0.63	0.56
PPC	0.53	0.82	0.63	0.52	0.70	0.61	0.64	0.67	0.60
PPCB	0.55	0.73	0.67	0.81	0.65	0.58	0.73	0.62	0.57
Un	0.51	0.69	0.59	0.69	0.66	0.63	0.71	0.69	0.72

Table 8: Proportion of indirect transitions to direct transitions, v_{ik}, for each node pair.

Node	Transition count				Metric, %		
j	$N\{ij\}$	$N\{jk\}$	$N\{ijk\}$	$\sum_k N\{ijk\}$	Λ_{ijk}	Φ_{ijk}	Ψ_{ijk}
Aff	4374	579	145	2749	14.7	25.0	3.3
Ban	53	24	1	25	0.1	4.2	1.9
Comp	289	194	16	142	1.6	8.2	5.5
List	24	18	1	15	0.1	5.6	4.2
Nat	579	2199	89	332	9.0	4.0	15.4
NatB	1892	2023	94	1147	9.6	4.6	5.0
PPC	467	839	21	264	2.1	2.5	4.5
PPCB	516	434	27	326	2.7	6.2	5.2
Un	400	222	11	209	1.1	5.0	2.8

Table 9: Metric calculations for the relevance of intermediary nodes, for source node $i =$ **Aff** and destination node $k =$ **Nat**.

of the routes from **Aff** to **Nat** are direct, with smaller contributions via **Aff**, **Nat**, **NatB**, and **PPCB**. 145 of the three-step transitions from **Aff** via **Aff** went on to **Nat**, and these 145 clicks represented 14.74% of the direct and indirect transitions from **Aff** to **Nat**. To avoid cluttering the graphic, we avoid drawing flow from intermediary routes if it is less than 2.5% (as an arbitrary threshold) of the number of all routes from **Aff** to **Nat**. One immediate conclusion is that although there are many routes from **Aff** to an intermediary, few of these then continue to **Nat**.

Figure 6 concerns the same two nodes, but reversed so that we are exploring routes from **Nat** to **Aff**. For these routes, the great majority are direct and a large proportion of the remainder are via $B =$ **Aff**. We can try to explore several such graphs in parallel; however the task becomes daunting as we need to explore n^2 graphs in all.

Figure 7 summarises the more important journeys via intermediary nodes. For each (source,destination) pair, a stars plot is shown. This shows the proportion of journeys via each kind of intermediate node. To avoid clutter, we show only intermediary nodes accounting for at least 10% of the journey, and bear in mind that we do not show directly the proportion of direct transitions, which can be inferred by the absence of segments showing indirect transitions. The colour and angle of segments is the same for each intermediary. From such a plot we may discern a number of features, depending on the particular example. Here, for example, we note that **PPC** appears to be an important intermediary for destination **PPC**; **NatB** is an important intermediary whenever the source or destination node is **NatB**; and **Aff** is an important intermediary whenever the source or destination node is **Aff**.

For specified (source, destination) pairs we may compute simultaneous confidence intervals (Goodman, 1965) for the multinomial proportions of direct and indirect journeys. These intervals correct a chosen level of significance (here, 10%) depending on the number of intervals constructed. Table 10 shows such intervals for the journeys from **Aff** to **Nat**, with indirect journey counts $N\{1, j, 5\}$ given by column 4 of Table 7 and direct journey count $N\{1, 5\} = 579$ from Table 5. As we pointed out earlier, we would need to compute 81 such tables to generate confidence intervals for all possible two-step and three-step journeys for this example, and this is partly why a summary measure for overall channel relevance is useful.

7.2 Channel relevance

Table 11 shows the relative value of nodes from three perspectives. The first represents the volume of two-step journeys starting at a node. The second represents the volume of two-step journeys ending at a node. The third shows the relative value of a node as an intermediary using the formulae derived to (8). The main interpretation is that **NatB** is very important in the journey, but not quite so much as would be believed simply looking at source and destination information. On the other hand, channel **Aff** has a slightly more important role than source and destination information suggests. Otherwise there are few major differences between channels for this retailer.

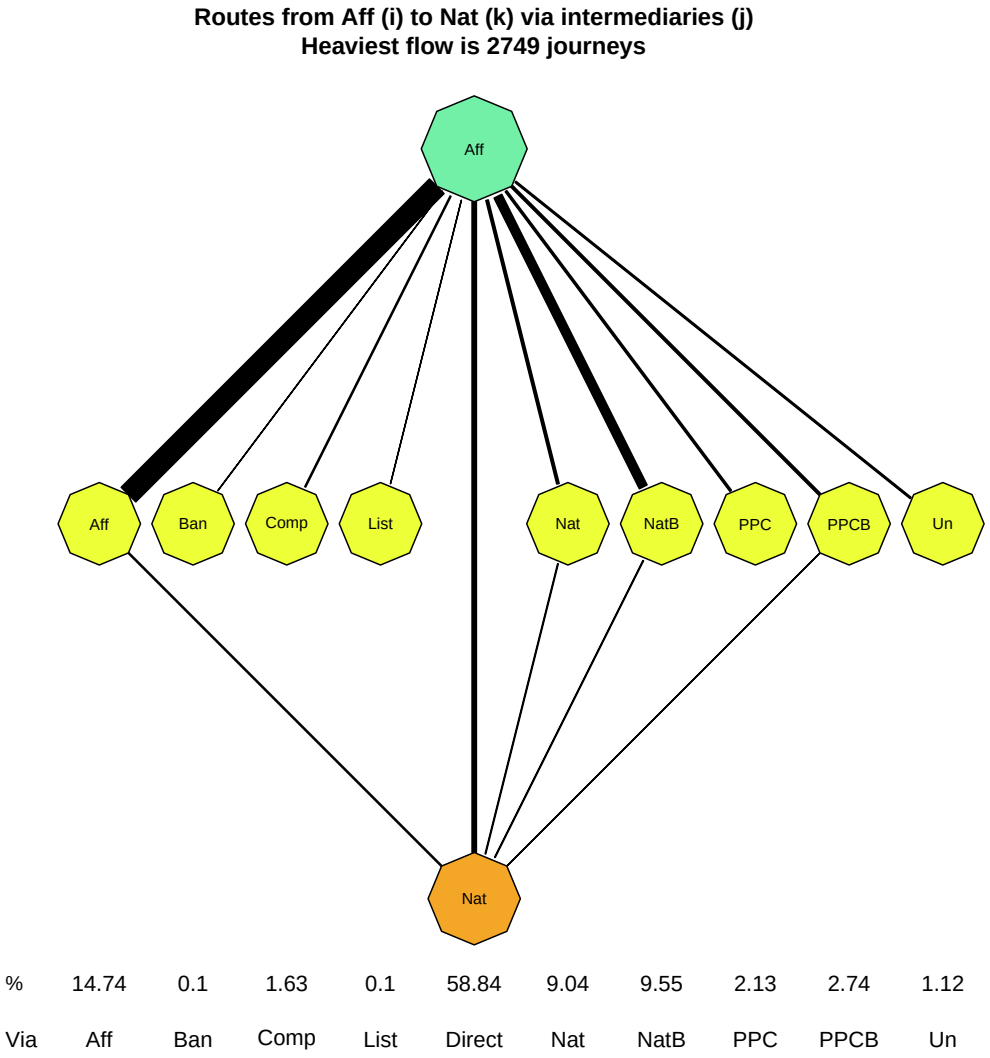

Figure 5: Relevance of intermediate nodes in journeys from **Aff** to **Nat**. Line widths indicate volume. Low volumes are omitted.

Journey via	Mean	Lower	Upper
Aff	0.1474	0.1206	0.1788
Ban	0.0010	0.0001	0.0086
Comp	0.0163	0.0087	0.0303
List	0.0010	0.0001	0.0086
Nat	0.0904	0.0696	0.1168
NatB	0.0955	0.0740	0.1224
PPC	0.0213	0.0123	0.0368
PPCB	0.0274	0.0169	0.0443
Un	0.0112	0.0052	0.0237
Direct	0.5884	0.5475	0.6281

Table 10: Simultaneous 90% confidence intervals for the proportion of journeys from **Aff** to **Nat** via intermediary nodes, and directly.

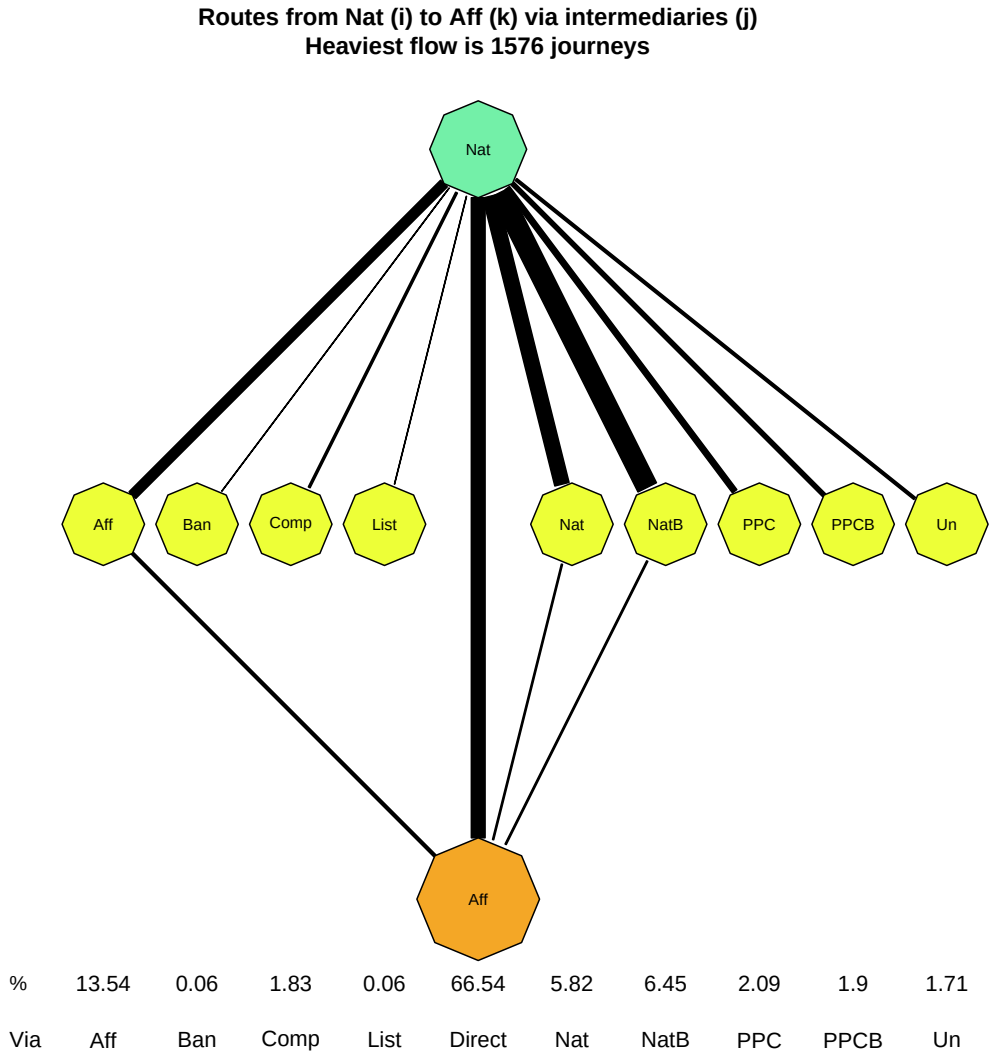

Figure 6: Relevance of intermediate nodes in journeys from **Nat** to **Aff**. Line widths indicate volume. Low volumes are omitted.

	Source	Destination	Intermediary
Aff	0.1031	0.1383	0.1694
Ban	0.0052	0.0060	0.0156
Comp	0.0235	0.0219	0.0349
List	0.0044	0.0036	0.0090
Nat	0.0863	0.0783	0.0861
NatB	0.4826	0.4768	0.4204
PPC	0.0953	0.0773	0.0754
PPCB	0.0912	0.0850	0.0947
Un	0.1086	0.1127	0.0945

Table 11: The value of intermediary nodes

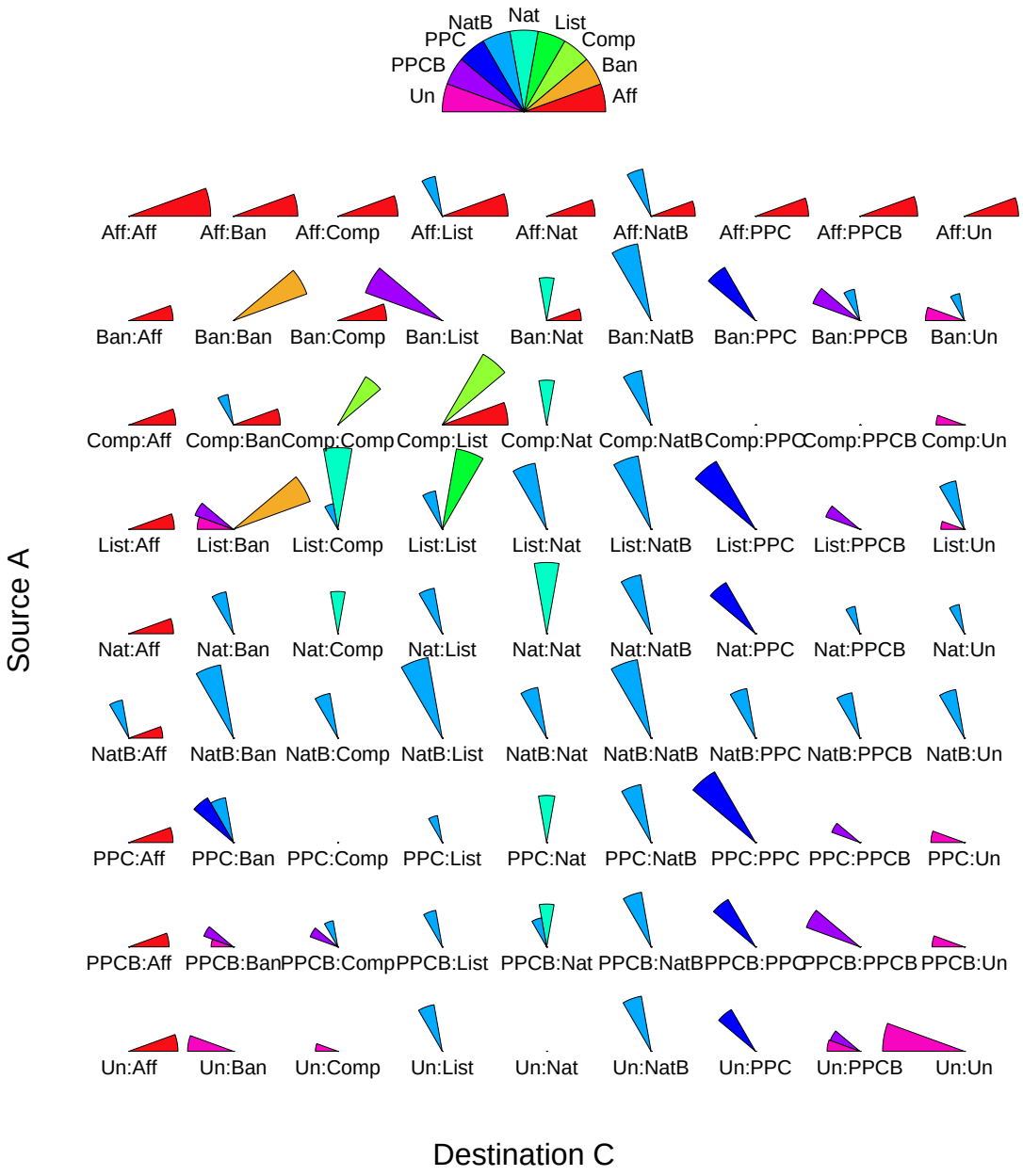

Figure 7: Relevance of intermediate nodes in all journeys. Journeys less than 10% as a proportion are omitted. The colour and angle of segments is the same for each intermediary node B.

8 Comparison of weighted attribution mechanisms

Using the data of Section 5, Figure 8 shows the total revenue attributions to eight channels for 58667 journeys for seven attribution methods: (1) the bathtub method described in Section 3 with $\theta_L = 4$ and $\theta_F = 2$; (2) first click wins; (3) last click wins; (4) equal weighting of all clicks – this corresponds to $\theta_L = 1$ and $\theta_F = 1$; (5) linear with last click valued at four times first click – this corresponds to $\theta_L = 4$ and $\theta_F = 1$; (6) exponential with last click valued at four times first click; and (7) the bathtub method additionally weighted according to channel value using metric $\tilde{r}_j$ (8), and with weights shown in Table 11. These weights are then compounding with time using (4). For this online retailer, all attribution methods yield similar results. Of note is that first-click-wins (2) tends to undervalue the **Aff** channel, whereas last-click-wins (3) tends to overvalue it; this is expected as the the nature of affiliate sites is to target consumers at the end of their journey that have already made the decision to buy and to provide a reward (e.g. cashback) for the purchase. Natural search (**Nat** and **NatB**) and PPC (**PPC** and **PPCB**) clicks can be assumed to be part of all stages of the buying journey (browsing, researching and buying) and therefore are expected to be rewarded similarly independent of the attribution model. It should be noted that an exception to this is that the bathub/value method (7) tends more highly to reward the **NatB** channel as it was found to be the most important intermediary channel in a typical journey: see the final column of Table 11, suggesting that **NatB** is perhaps more a navigational click rather than a conversion driver.

9 Discussion

In this paper we offer a sensible revenue attribution mechanism based on appropriate time-weighting of clicks. We have also shown how the method may be modified when there is separate information available on the quality of visitable channels. There is unavoidably a subjective element in choosing an appropriate shape for time-weighted attribution. This is the same problem faced by Bayesian statisticians in choosing an appropriate prior. This is an uncomfortable fact for major retailers, who often naively expect that there is a single "right" answer. The choice of attribution shape and parameters such as θ_L, the ratio of last click to first click value, depend on the aims of the attribution. If a retailer wishes only to prioritize last-click-wins, then that is the "right" answer for them. Ultimately, the right attribution scheme is the one which produces the most traffic or revenue to the retailer. This can be tested in principle by designing experiments in which groups of search terms are allocated to different attribution schemes and the subsequent effect on traffic and revenue measured.

The deep question is whether a specified channel actually matters. We have provided a metric based on three-step transitions in order to measure this importance. Statistical sequential pattern analysis of this kind is highly challenging: one aim of future work is to examine longer journey fragments. A second theme of future work is to explore the roles of intermediary nodes in determining conversion behaviour; however we would need to collect meaningful data about non-converting journeys in order to do this, and this would require being careful about the assumptions of non-converting journeys. We have not taken into account the value of conversion; for example it may be that some nodes are relevant only for low-revenue conversions.

Acknowledgements

Part of this research was funded by Knowledge Transfer Partnership KTP007499, funded by Summit Media Ltd. and by the UK Technology Strategy Board. We are grateful to Summit Media Ltd for providing data and expertise.

References

Abhishek, V., P. Fader, and K. Hosanagar (2012). The long road to online conversion: A model of multi-channel attribution. http://dx.doi.org/10.2139/ssrn.2158421.

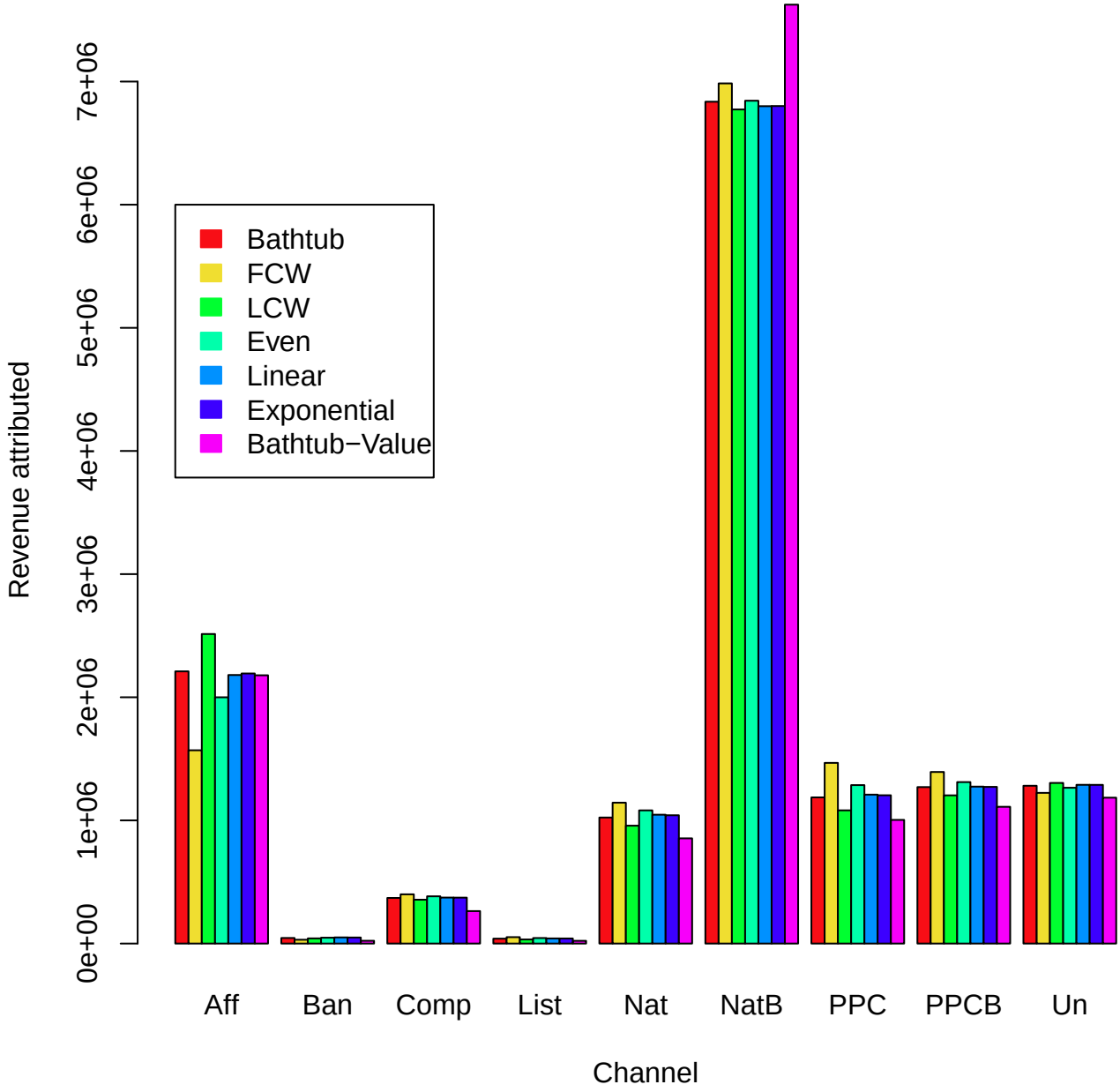

Figure 8: Comparison of total revenue attributions to eight channels for 58667 journeys for seven attribution methods.

Agrawal, R. and R. Srikant (1995). Mining sequential patterns. Technical report, IBM Research Division, Almaden Research Center.

Berendt, B. and M. Spiliopoulou (2000). Analysis of navigation behaviour in web sites integrating multiple information systems. *VLDB J. 9*(1), 56–75.

Dalessandro, B., C. Perlich, O. Stitelman, and F. Provost (2012). Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, ADKDD '12, New York, NY, USA, pp. 7:1–7:9. ACM.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics 7*(2), 247–254.

Gunduz, S. and M. T. Ozsu (2003). A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 535–540.

Gunduz-Oguducu, S. and M. T. Ozsu (2006). Incremental click-stream tree model: Learning from new users for web page prediction. *Distributed and Parallel Databases 19*(1), 5–27.

Hahsler, M., B. Grün, and K. Hornik (2005, October). arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software 14*(15), 1–25.

Hou, C.-D., J. Chiang, and J. J. Tai (2003). A family of simultaneous confidence intervals for multinomial proportions. *Computational Statistics and Data Analysis 43*(1), 29–45.

Internet Advertising Bureau UK (2013). 2012 Online Adspend Full Year Results. `http://www.iabuk.net/research/library/2012-full-year-digital-adspend-results`.

Jamalzadeh, A. (2012). Statistical methods for ecommerce. Phd thesis, Durham University.

Moe, W. W. (2003). Buying, searching or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology 13*(1&2), 29–39.

Osur, A., E. Riley, T. Moffett, S. Glass, and E. Komar (2012). The Forrester Wave Interactive Attribution Vendors Q2 2012. Technical report, Forrester Research, Inc.

Shao, X. and L. Li (2011). Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 258–264.

Wang, W. and J. Yang (2005). *Mining Sequential Patterns from Large Data Sets*. New York: Springer.

Wooff, D. A. and A. Jamalzadeh (2013). Robust and scale-free effect sizes for non-normal two-sample comparisons, with applications in e-commerce. *Journal of Applied Statistics To appear*.

Xu, L., J. A. Duan, and A. B. Whinston (2012). Path to purchase: A mutually exciting point process model for online advertising and conversion. `http://dx.doi.org/10.2139/ssrn.2149920`.

Zaki, M. J. (2000a). Sequence mining in categorical domains: Algorithms and applications. In R. Sun and L. Giles (Eds.), *Sequence Learning: Paradigms, Algorithms, and Applications*, Volume 1828 of *LNAI State-of-the-Art-Survey*, pp. 162–187. Springer-Verlag, Heidelberg, Germany.

Zaki, M. J. (2000b). Sequence mining in categorical domains: Incorporating constraints. In *Proceedings of the 9th International Conference on Information and Knowledge Management, Washington D.C*, pp. 422–429.