Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

Shapley values in higher-order Markov graph-based attribution modelling

Sven Nieuwkerk

Student number: 542549

Date final version: 3 July 2022

Abstract

Companies invest increasing amounts of money in online marketing and this needs to be divided between a vast array of different online marketing channels. Therefore, information about the contribution of every channel to a possible conversion is crucial for companies. In the attribution modelling literature different models for estimating the consumer paths and the contribution of the different channels in those paths are proposed. This research combines higher-order Markov models with Shapley values to solve the attribution problem and adds new insights into attribution in a higher-order Markovian framework.

In order to do this, the "*PathData*" dataset from the "*ChannelAttribution*" package is used. Then, the consumers paths in this dataset are estimated by different order Markov graphs. It is found that higher-order Markov outperform normal Markov graphs in predicting a conversion. Finally the attribution of the channels to a possible conversion is measured using Shapley values, the Removal effect and some heuristic. Heuristics are found to be very inconsistent attribution measures compared to both Shapley values and the Removal effect. Furthermore, the Removal effect assigns less attribution to channels which are more frequent in the dataset compared to the Shapley value for every order Markov graph. Finally, attribution in higher-order Markov graphs seem to better capture the carry- and spillover effects between different channels.

Supervisor: Kathrin Gruber

Second assessor: Luuk van Maasakkers

Contents

1 Introduction

Due to the increased usage of smartphones, tablets and other mobile devices the number of potential interactions between consumers and service providers has grown massively (Gartner Research, 2019). An interaction between a consumer and a company is called a *touchpoint*. Businesses have reacted to this by increasing investments in online marketing over the last years (Kannan, Reinartz, & Verhoef, 2016). In the United States for example, spending on display advertisements alone grew from $39.4 billion to $49.8 billion between 2017 and 2018 according to IAB (2019). Display advertisement, however, is not the only marketing *channel* available for companies to reach out to consumers. Tueanrat, Papagiannidis, and Alamanos (2021) state there exist a vast array of different channels, which can exist either in the physical or the online world. Examples include social media advertising or search engine advertising, where a company pays the search engine to have a higher rank on the search engine results. These are also widely used by companies to reach out to consumers.

Companies invest in these different promotional channels to improve the chance of a consumer making a *conversion*, that is a sign-up, subscription or purchase. Consumers can visit various different channels on their way to a possible conversion. Because these different channels may influence the probability to visit another channel or in the end lead to a conversion indirectly, the attribution of every channel to a possible conversion can be unclear. However, this information can be of critical importance for managers making decisions on how to distribute the marketing budget between different channels (Kannan et al., 2016).

To better measure the attribution of different channels to a conversion, various solutions exist in the industry, known as attribution models. The first attribution models to be developed were the rule-based or heuristic models. These assign attribution to different channels following some predetermined assumptions. One of these is the *last-click* model. This model assigns all the attribution for a conversion to the channel of the last touchpoint of a consumer before the conversion. Other widely used rule-based model consist of models that evenly assign the attribution to the channel of the first touchpoint or assign the attribution to the different channels of all the touchpoints.

However, these heuristics are outperformed by more sophisticated data-driven attribution models. These include a bagged logistic regression model proposed by Shao and Li (2011), a probabilistic model using hazard and survival functions and time elements proposed by Ji, Wang, and Zhang (2016), a Naive Bayes approach proposed by Li and Kannan (2014) and a Deep Neural Network approach by Arava, Dong, Yan, Pani, et al. (2018).

In this research a graph-based Markovian approach will be adopted. Archak, Mirrokni, and Muthukrishnan (2010) first used a first-order Markovian approach to model customer behaviour, when customers were exposed to different channels. Furthermore, six different measures of attribution called ad factors are proposed to measure attribution of the different channels in a Markov graph, one of them being the Removal effect. Anderl, Becker, Von Wangenheim, and Schumann (2016) extended this paper by adopting a higher-order Markovian approach to model the customer paths. To determine the attribution of the different channels the Removal effect was used. Higher-order Markov graphs were found to outperform normal Markov graphs for modelling the consumer paths and attribution based on heuristics was shown to be incorrect compared to the attribution measured in the Markov graphs. Abhishek, Fader, and Hosanagar (2012) propose a Hidden Markov Model, where one state reflects the interest level of a user in the product. Note that this method, where states are assumed to be latent unobserved engagement levels, is different from the method in Anderl et al. (2016), where the states are the channels of the touchpoints along the consumer journey.

To better measure attribution in a multi-touch environment and in a data-driven way, Dalessandro, Perlich, Stitelman, and Provost (2012) propose Shapley Values to estimate the attribution. Shapley Values are widely used in Cooperative Game Theory to fairly distribute value or credit between different players in a game. Singal, Besbes, Desir, Goyal, and Iyengar (2022) implement Shapley Values for Markovian approaches and compare the attribution measured by Shapley Values to attribution measured by the last-click heuristic and by the Removal Effect. It was found that Shapley values correct flaws of both the last-click heuristic and the Removal effect.

However, Shapley values are not compared to other attribution measures for higher-order Markov graphs. As higher-order Markov models are also defined in this framework, further research can be done to compare attribution measured by Shapley Values to other attribution measures for a higher-order Markovian approach. Therefore the research question of this paper is

"*How do different attribution measurements compare for different higher-order Markov models and between higher-order Markov models and normal Markov models?*"

For the purpose of this research, the "*PathData*" dataset from the "*ChannelAttribution*" package in R is used, consisting of 88387 consumer paths. These paths are used to estimate a Markov model for the consumer journey. Using AUC and BIC values as model fit measures, the ideal specification for the Markov models is found. Then, attribution of the different channels is computed using heurstics, the Removal effect and Shapley values. The results of this research are that higher-order Markov

models are found to better estimate the consumer paths and also have a representative performance in estimating the consumer paths compared to non-Markovian models. Furthermore, Shapley values are found to assign more attribution to channels that are more common in the dataset compared to the Removal effect. The attribtion assigned by the Shapley values is more consistent than from the different heuristics. Finally the attributions of higher-order Markov models better capture the interconnectivity of the different channels than attribution in a normal Markov framework.

This research contributes to the literature by evaluating the Shapley values in a higher-order Markov framework and comparing them to other attribution measures. While higher-order Markov graphs have been used to model the consumer journey (Anderl et al., 2016) and Shapley values have been used to measure attribution in different frameworks (Berman, 2018; Dalessandro et al., 2012; Singal et al., 2022), this researh combines both of these methods to bring new insights to the attribution problem.

The remainder of this paper is structered as follows: section 2 divides the research question in multiple sub-questions and provides an overview of the relevant literature. Section 3 describes the data used in this research. Section 4 describes how the Markov models are estimated and how the different attribution measures are computed. Section 5 evaluates the results of this research and uses the findings of the sub-questions to answer the research question. Finally, section 6 summarises the main findings of this research and states several limitations of it.

2 Theory

To analyse the research question, the attribution values in the models need to be evaluated. However, before attribution values can be computed, first a model for the consumer path needs to be estimated. There are many different approaches in attribution literature for modelling the consumer path (Arava et al., 2018; Ji et al., 2016; Li & Kannan, 2014; Shao & Li, 2011). However, this research will focus on an attribution framework based on Markov graphs. Markov graphs were first proposed to model attribution by Archak et al. (2010). They stated that the heuristic attribution models, which were the industry standard at that time for measuring the effectiveness of an ad-campaign, didn't account enough for the interconnected channel-dependencies. Therefore they propose to use Markov graphs to model the effect of different marketing channels on consumers.

Adopting this, Abhishek et al. (2012) propose a Hidden Markov model to estimate the consumer journey through to conversion funnel. The different states of the Markov graph are the unobserved

levels of interest the consumer has. It is found that different marketing channels have different impacts on an possible conversion depending on where in the conversion funnel the consumer is exposed to the channel. Anderl et al. (2016) extend the work of Archak et al. (2010) by proposing higher-order Markov graphs to model the attribution problem. They state that these higher-order Markov graphs better estimate the carryover and spillover effects. These are the effects of a previous visit influencing the consumer to visit the company again via the same or different channels respectively. Because of this property of higher-order Markov graphs, this research will adopt a Markov model with channels as states and higher-order Markovian effects instead of a Hidden Markov graph.

As a good model for the consumer path is needed before estimating the attribution of different channels, the Markov model and it's order need to be correctly specified. Furthermore, the Markov models need to have a representative performance compared to other models for modelling consumer paths. Therefore the first sub-question is

"*What is the best specification for the Markov model and how does it compare against other models?*"

Although attribution models are build to measure the effect that different channels had on the way to a conversion in the past, Shao and Li (2011) state that attribution models should be able to accurately predict a conversion. Thus, to select the order that is the best fit for the Markov model and to compare the Markov models to other attribution models, predictive performance can be used. However, when comparing predictive performance of different models, generalization should also be taken into account. While more complex models may perform better on the data these models are trained on, simpler models often outperform the more complex models on new data. Therefore, the better generalizability of simpler models should be considered with predictive performance.

Considering the predictive performance of the models, Anderl et al. (2016) find that higher-order Markov models outperform the normal Markov models. Furthermore, the robustness of the attribution measure is low for the second and third order Markov models implying that these results should be generalizable. Furthermore, Anderl et al. (2016) use a 4th order logit model as a benchmark for the performance of modelling the consumer path. It is found that the higher-order Markov models outperform this logit model. As their results are consistent over four different datasets, these results should also hold up in this research. Therefore, the hypothesis for this research is that higher-order Markov model will better estimate the consumer path than normal Markov model and that Markov models will have a representative performance.

To improve on the heuristic attribution models, Archak et al. (2010) propose six different ad factors to measure the attribution of different channels in Markov graphs. One of these ad factors is the Removal effect, which is also used by Anderl et al. (2016) to compute the attribution in higher-order Markov graphs. The Removal effect for a channel is the change in the probability of a path leading to a conversion when removing that channel from the model.

Dalessandro et al. (2012) use Cooperative Game Theory to causally motivate a new attribution measure and show this can be approximated using the Shapley values, motivating them as "fair", "data-driven" and "interpretable". The Shapley value uniquely distribute the attribution generated by all channels between the different channels. The attribution of a channel assigned by the Shapley values is a weighted average over all coalitions with that channel of the value that would be lost if that channel dropped out of the coalitions. Berman (2018) showed in a stylized example that Shapley values can be considered a more fair attribution measure than the last click heuristic. Singal et al. (2022) apply the Shapley values to measure attribution in a Markov graph.

The Removal effect and the Shapley value are similar in that both measures assign attribution to a channel by evaluating how much is lost without that channel. The difference is that the Removal effect evaluates this loss by the difference in probability of a path reaching conversion and the Shapley value evaluates the loss by the difference in total value generated by coalitions with and without the channel. This notion of using loss to evaluate the attribution of a channel is similar to the lift measure often used in data mining as an association measure. The lift, however, measures the gain between the response of a target variable compared to a baseline variable (Vu et al., 2019), instead of measuring the loss. Both the Removal effect and the Shapley value are proposed in literature as attribution measures in a Markov graph. To compare these attribution measures, together with the heuristic attribution measures, the second sub-question is

"*What are the differences between the different attribution measures and how are they explained?*"

Anderl et al. (2016) found the the attribution as measured by the heuristic is less consistent for higher-order Markov graphs and that the heuristics consistently over- or undervalue certain channels. Furthermore, Singal et al. (2022) showed in a framework with a hidden Markov model that also the Removal effect has limitations. It was found that the Removal effect overestimates the attribution for channels with more touchpoints and found that the Shapley value corrects this. Finally, Anderl et al. (2016) found that higher-order Markov graphs better measure the carry- and spillover effects between the different channels. It is expected that this will be reflected in the attributions of higher-order Markov graphs.

3 Data

The data used in this research is the "*PathData*" dataset contained in the R package "*ChannelAttribution*". This dataset contains 10,000 path instances. These paths can be observed multiple times, so one instance can count as multiple observations, and every instance has 4 attributes. "path" contains the sequence of channels the path visits, "total_conversions" contains the number of users that followed the path and made a conversion, "total_null" contains the number of users that didn't make a conversion and "total_conversion_value" is the total value of all conversions that consumers following this path made. It should be noted that the "total_conversion_value" attribute is not used during this research and that the "path" attribute contains several duplicate paths.

Path data like this is often collected using cookie tracking, meaning the different touchpoints from a device with the site of the service provider can be tracked and stored. However, using cookie data has several disadvantages. Examples are the inability of cookies to track one consumer using multiple devices or a device, where the consumer blocks cookies (Flosi, Fulgoni, & Vollman, 2013).

The total number of channels in the dataset is 12. All channels in the dataset have the names of Greek letter and the original marketing channel can therefore not be recovered. As a consequence for this research, the attribution of the different channels will not be put into a marketing context. However, as this paper is a technical evaluation of the different attribution measures in a Markov model, this does not limit this research.

Table 1: Descriptive data of the *PathData* dataset

Channel name	every	alpha	beta	gamma	delta	epsilon	zeta	eta	theta	iota	kappa	lambda	mi
Percentage of clicks	-	0.4219	0.1085	0.0047	0.0001	0.0164	0.0115	0.1147	0.0604	0.2014	0.0078	0.0526	0.00004
Average clicks per path	4.2790	1.8052	0.4641	0.0202	0.0005	0.0704	0.0493	0.4908	0.2583	0.8617	0.0335	0.2249	0.0002
Rank	-	1	4	10	11	7	8	3	5	2	9	6	12

The total number of observed consumer paths in the dataset is 88387, of which 22.38% (19785 paths) lead to conversions and 77.62% (68602 paths) don't lead to a conversion. Table 1 shows the descriptive statistics of the different channels in the *PathData* dataset. There are large differences in frequency between the different channels. alpha is by far the most visited channel, with more then 40% of the total of interactions. iota (20%), eta (11%) and beta (11%) are also very prevalent in the dataset. gamma (0.5%), delta (0.01%) and mi (0.004%) are almost non-existent. The length of the average consumer journey is 4.27 channels.

4 Methodology

In this section, firstly, different models are introduced to explain the customer journey paths. Additionally, the different measures for model fit are described. Then, different attribution measures are proposed to measure the attribution of the different channels.

4.1 Modelling the customer journey path

To compute the attribution of different marketing channels, first a good estimation of the consumer journey is needed. For this research, as proposed by Anderl et al. (2016) and Archak et al. (2010), Markov and higher-order Markov graphs are adapted to model the customer journey path. Following Anderl et al. (2016), a logit regression model using order effects is proposed as a benchmark to compare the predictive performance of the Markov graphs against. By doing this the model fit of the Markov graphs can be evaluated (Shao & Li, 2011). The Markov graphs are compared to each other using the Bayesian information criterion (BIC) and the receiver operating characteristic (ROC) curves with the area under the curve (AUC) values to compare the predictive performance of the Markov graphs to the predictive performance of the logit regression model.

4.1.1 Markov Graphs

A Markov graph is defined by a set of states, $S = \{s_1, s_2, \ldots, s_n\}$, and a transition matrix, P, which consists of a transition probability $p_{i,j}$ for going from state s_i to s_j. Here, $0 \leq p_{i,j} \leq 1$ for every i and j and $\sum_{j=1}^{n} p_{i,j} = 1$ for every i. In this study, the states of the Markov graphs are the (combination of) channels the consumer is last exposed to combined with two absorbing states for conversion, c_c, and no conversion, c_n.

This implies that for the estimated first order Markov graph the set of states S is the set with all channels combined with the absorbing states and the transition probabilities are

$$p_{i,j} = P(X_t = s_j \mid X_{t-1} = s_i), \tag{1}$$

where X_t is a random variable of the state of the Markov graph at time t and $s_i, s_j \in S$.

For a k th order Markov graph the transition probability of going to state j is

$$P(X_t = s_j \mid X_{t-1} = s_i, \ldots, X_{t-k} = s_l), \tag{2}$$

where X_t is again a random variable of the state of the Markov graph at time and t, $s_i, s_j, s_l \in \mathcal{C}$, where C is the set of all channels and $\mathcal{C} = C \cup c_c \cup c_n$. These higher-order Markov graphs are

equivalent to first order graphs where $S = (\times_{i=1}^{k}\{C, \emptyset\}) \cup c_c \cup c_n$, in which the empty set $\emptyset$ is used when the order of the graph exceeds the number of channels already visited before channel t at the start of a path and c_c and c_n are again the absorbing states, and using the transition probabilities in equation 1. As the order of the Markov graph goes up, the number of states rises exponentially. Therefore, the number of transition probabilities to be estimated goes up rapidly and the model quickly gets more complex for higher-orders.

4.1.2 Logit model

The logit model proposed in Anderl et al. (2016) is

$$\text{logit}(Y_i) = \alpha + \sum_{i=1}^{n} \beta_{(4i-3)} d_{i1} + \beta_{(4i-2)} d_{i2} + \beta_{(4i-1)} d_{i3} + \beta_{(4i)} d_{i4}, \tag{3}$$

where $d_{i,t}$ is a dummy variable, which is 1 if channel i is in position t of the customer path. Here the position is counted from the end of the path. This means the last four contacts of every journey are used to predict if the journey leads to a conversion. Therefore, this logit model is most comparable to a 4th order Markov model, which also uses the last four contacts to predict the next interaction. However, it is still comparable with the other order Markov models.

The α is the intercept and the β_i's are the coefficients for the different dummies. As there is always at least one channel in every path and exactly one channel can be in the final position of a path, it always holds that $\sum_{i=1}^{n} d_{i1} = 1$. Therefore α and $\beta_1, \beta_5, \ldots, \beta_{45}$ are not separately identifiable and α is forced to 0. The interpretation of the different β coefficients in a logit model is complex an therefore these are not used to compute attribution.

4.1.3 Model fit measures

Shao and Li (2011) state that to compare the model fit of Markov graphs to other models for modelling the customer journey, the accuracy of predicting a conversion can be used. As measure for predictive accuracy the AUC value under the ROC curve will be used. Gonçalves, Subtil, Oliveira, and de Zea Bermudez (2014) give formal definitions of the ROC curve and the AUC value. Summarised, the ROC curve plots the True Positive (TP) rate of a classifier against the False Positive (FP) rate and the AUC value is the area under this curve. A straight 45° angle line gives the performance of a completely random classifier. A better classifier has a ROC curve that rises above this line and thus has a higher AUC value. According to Bradley (1997), the AUC values have

a few desirable properties compared to the overall accuracy. Of these properties the independence of prior class distribution is the most important for this research.

As a second measure for model fit, the BIC is used. The BIC widely used in literature to compare different models. Katz (1981) proved that the BIC is consistent for estimating the order of a Markov graph. The BIC is computed as

$$BIC = k * ln(n) - 2ln(L), \tag{4}$$

where k is the number of variables that are estimated, n is the number of observations and L is the likelihood of the Msrkov model. For a Markov model the likelihood can be computed by the accuracy of predicting every step in the path from every user. Then L depends on every step in the path and for a k'th order Markov model is computed as

$$L = \Pi_{i=1}^{n}\Pi_{t=1}^{p_i}\Pi_{s\in S}P(X_t = s \mid X_{t-1} = x_{i,t-1}, \ldots, X_{t-k} = x_{i,t-k})^{\mathbb{1}_{x_{i,t}=s}} \tag{5}$$

$$= \Pi_{i=1}^{n}\Pi_{t=1}^{p_i}P(X_t = x_{i,t} \mid X_{t-1} = x_{i,t-1}, \ldots, X_{t-k} = x_{i,t-k}), \tag{6}$$

where X_t is a random variable of the state of a Markov graph, $x_{i,t}$ is the state of path i in position t, S is the set of states, o is the order of the Markov graph and p_i is the amount of channels in path i. Because the likelihood consists of the product of the probabilities of making a step, the number of observations is the sum of the number of steps from every consumer path. While comparing BIC values, a lower value of the criterion indicates the preferred model. Compared to the AUC values, the BIC also takes generalizibility in account by penalising models that use more parameters.

4.2 Attribution measures

Attribution measures assign attribution to the different channels. Several different attribution measures that assign attribution to different channels in a Markov graph are used in this research. Firstly the heuristic attribution measures will be introduced, then the Removal Effect and finally the Shapley values. Dalessandro et al. (2012) gives three criteria for a good attribution measure. Firstly, an attribution measure should be fair, meaning that it should reflect the contributions of every channel to a possible conversion. Secondly, attribution measures should be data-driven, that is it should be tailored to campaign specific properties of every ad campaign. Finally, a good attribution measure should be interpretable, thus both understandable and based on statistical foundations. The different attribution measures are evaluated on these properties.

4.2.1 Heuristic attribution measures

The first group of methods to assign attribution to different channels were the rule-based or heuristic attribution measures. In this paper the last-touch (LTA), first-touch (FTA) and linear (LIN) heuristics are used. LTA and FTA values for channel r are computed as

$$\pi_r^{LTA} = \frac{l_r}{n}, \tag{7}$$

$$\pi_r^{FTA} = \frac{f_r}{n}, \tag{8}$$

where channel f_r is the number of paths that converge with r as first channel after the start, l_r is the number of paths paths that converge with r as last channel before the conversion and n is the number of paths. These attribution measures assign all the attribution to the last and first channel on the path respectively.

LIN for channel r is computed as

$$\pi_r^{LIN} = \frac{1}{n} \sum_{i \in V} \frac{n_{i,r}}{p_i}, \tag{9}$$

where V is the set of all paths that converge, $n_{i,r}$ is the number of times channel r is present in path i, p_i is the number of channels in path i and n is the number of paths. This attribution measure divides the attribution of a path equally between the channels of every touchpoint of the path.

These heuristic attribution measure are shown to be outperformed by more sophisticated attribution measure in both practical settings (Anderl et al., 2016; Dalessandro et al., 2012; Singal et al., 2022) and in a theoretical setting (Berman, 2018). The heuristics lack fairness as channels get assigned a attribution value without trying to find the actual contribution of the channel to the conversion. They also are not data-driven as the different properties of each ad campaign are not considered, but are assumed to be equal by assuming the same rules for every campaign. The heuristics are however very easily understandable, but lack statistical foundation.

4.2.2 Removal effect

The Removal effect was first proposed by Archak et al. (2010) as attribution measure in Markov graphs and were adapted by Anderl et al. (2016) in higher-order Markov graphs. The removal effect of state s_i is the change in probability of reaching a conversion if state s_i is removed from the graph. The removal effect can be computed as

$$RE(s_i) = \text{Visit}(s_i) \times \text{Conversion}(s_i), \tag{10}$$

where Visit(s_i) is the probability of a path passing through state s_i and Conversion(s_i) is the probability that a path at state s_i reaches conversion. As states resemble channels in first-order Markov graphs, the Removal effect of a channel is the is the Removal effect of the state resembling the channel. The Removal effects are computed by sampling 100000 paths from the Markov graph and computing the Removal effects for those paths.

Singal et al. (2022) state several reasons why the Removal effect can be unfair in certain situations. However, the Removal effect is still a lot fairer than the heuristics (Anderl et al., 2016). The Removal effect is also data-driven as it uses the Markov graph that is estimated and is therefore developed specifically for this ad campaign. The Removal effect is also fairly interpretable as it has a clear meaning and is based on statistical properties.

4.2.3 Shapley values

The Shapley value is first proposed in the attribution framework by Dalessandro et al. (2012). It is a well known solution concept in game theory introduced by Shapley (1953). The Shapley value can very broadly be seen as the average marginal contribution of a channel to a coalition. The Shapley value of channel r is computed as

$$\pi_r^{Shapley} = \sum_{\chi \in C \setminus \{r\}} \frac{|\chi|!\,(\,|C| - |\chi| - 1)!}{|C|!} \times (v(\chi \cup \{r\}) - v(\chi)), \tag{11}$$

where C is the set of all channels and $v(K)$ is the value, in this case conversions, that is generated by set K. The generated value is by set $K \subseteq C$ is the number of paths that converge only passing through channels in set K. Singal et al. (2022) prove that $\pi_r^{Shapley} = E_{P \sim M}(\mathbb{1}_{P \in V} * \frac{1}{u_P})$, where P is a path in Markov graph M, V is the set of all the converted paths and u_P is the amount of unique channels in path P. This means the Shapley value of channel r can be estimated as

$$\pi_r^{Shapley} = E_{P \sim M}(\mathbb{1}_{P \in} * \frac{1}{u_p}) \tag{12}$$

$$\approx \frac{1}{|W|} \sum_{P \in W} \mathbb{1}_{p \in V} * \frac{1}{u_P}, \tag{13}$$

where W is the set of all paths. It should be noted that in this way the Shapley values have a strong correlation with the LIN heuristic. The difference is that the LIN heuristic gives the attribution of one path multiple times to a channel if it occurs multiple times in the path, while the Shapley values only give the attribution once. The Shapley values are acquired by sampling 100000 paths from the Markov graph and computing the Shapley values from this sample.

Shapley values are widely considered as a fair manner to distribute attribution. Firstly Dalessandro et al. (2012) showed that there is a causal motivation to use Shapley values as a measure for attribution. Furthermore, Shapley values have four desirable probabilities:

- Efficiency: $\sum_{r \in C} \pi_r^{Shapley} = v(C)$. This property says that all value that is generated is exactly distributed over all channels,
- Symmetry: if $r, r' \in C$, $\chi \subseteq C/\{r, r'\}$ and $v(\chi \cup r) = v(\chi \cup r')$ then $\pi_r^{Shapley} = \pi_{r'}^{Shapley}$. This property says that if two channels are identical in that the same value is generated with either of them, then the Shapley values of both of them will be equal,
- Linearity: if $v_1(\cdot)$ and $v_2(\cdot)$ are value generating functions for two games then $\pi_r^{Shapley}(v_1 + v_2) = \pi_r^{Shapley}(v_1) + \pi_r^{Shapley}(v_2)$ and $\pi_r^{Shapley}(\alpha v_1) = \alpha \pi_r^{Shapley}(v_1)$. This property says that for different value generating functions the shapley values are linear.
- Null player: if $r \in C$ and for every $\chi \in C/r$ $v(\chi \cup r) = v(\chi)$, then $\pi_r^{Shapley}$=0. This property says that a player that doesn't add any value to a coalition, the player doesn't get any attribution.

Shapley values are the only attribution measure to combine all four of these probabilities and can therefore be considered a fair attribution measure. As the Shapley values are computed using the Markov graph, just like the Removal effect, they are also data-driven. Finally Shapley values have very strong statistical foundations and therefore have good interpretability.

5 Results

In this section firstly the results for model fit are discussed using the measures introduced in section 4. Then, the different attribution are computed and the differences between these attribution measures for the different Markov models are discussed.

5.1 Model fit

For evaluating the model fit Markov graphs from order 1 to order 8 are considered together with the Logit mode of section 4. These models are compared using the BIC values and the AUC values. Table 2 gives the AUC and BIC values for these models. ROC curves and thus AUC values are computed using 10-fold cross-validation. The Markov models and the AUC values are estimated using the "*ChannelAttribution*" package.

Table 2: AUC values and BIC values of the different Markov models and of the Logit model

Markov order	1	2	3	4	5	6	7	8	Logit
AUC	0.5087	0.5174	0.5278	0.5428	0.5538	0.5589	0.5619	0.5632	0.5091
BIC	1523503	1407676	1555712	1637678	1788343	1883650	1942508	1989616	-

Firstly, it is observed that all higher-order Markov models outperform the Logit model in predicting a conversion measured by the AUC values. Even a 2nd order Markov model outperforms the Logit model, which takes 4th order information in consideration. This means that higher-order Markov models seem an appropriate way to mode the consumer paths as they outperform the benchmark model. This finding agrees with the hypothesis that higher-order Markov models outperform other attribution models in predicting conversions and thus have a representative performance. However, it should be noted that the predictive accuracy of all models is low. The AUC values barely reach above 0.5, which is the AUC value of a completely random prediction.

Finally, it can be seen that the AUC values increase when the order of the model goes up. This is to be expected as more information is used. However due to the increasing number of parameters needing to be estimated, higher-order Markov models are more prone to overfitting and are therefore less generalizable. Therefore the Markov model with order 4 is chosen based on the AUC values, as the increase in AUC values after this is small compared to the increase in parameters.

Looking at the BIC values, the Markov model with order 2 is the best performing model in predicting the consumer path. From the 3rd order onwards the BIC values again rise above the score of the 1st order model and these BIC values keep increasing when the order of the models go up. As 2nd order Markov models are found to be optimal, this finding agrees with our hypothesis that higher-order Markov models better estimate the consumer journey. However, the 1st order Markov model still outperforms every other higher-order Markov model.

Comparing the outcomes from the AUC measure to the BIC measure, different conclusions are found for the optimal order of the Markov model, as for the AUC values the results keep improving for a higher-order, while for the BIC values the result decrease. This is, however, expected, as the BIC values also consider the generazibility of the models while, the AUC values do not.

Overall it can be concluded that higher-order Markov models outperform normal Markov models and that Markov models have a representative performance in estimating the consumer path compared to other models. These findings agree with the hypotheses of the first sub-question. The different

measures for model fit, however, choose different specifications of the model as optimal, because these measure weigh predictive accuracy and generalizibility differently. Therefore, the attributions in both 2nd and 4th order Markov models will be further inspected in the following subsection.

5.2 Attribution

Now the different attribution measures are compared. In order to do this, firstly a global image of the attribution will be made by looking at all the different attribution values. Then the differences between the Removal effect, the Shapley values and the heuristics will be compared in a higher-order Markov model. For this a 4th order Markov graph will be used, as this was found to be the optimal specification of the model with AUC values as measure for model fit. Finally the difference in attribution between the normal Markov graph, the 2nd order Markov model, as was found to be optimal by the BIC values, and the 4th order Markov model is investigated.

To make the attribution measures more comparable, all attribution results are scaled as percentages by dividing them by the sum of the respective attribution measure. Both Removal effect and Shapley values are computed by simulating 100000 runs through the estimated Markov graphs and are computed from that sample. It should be noted the "*ChannelAttribution*" package was not used to simulate the Removal effects as there is an error in the package in computing these. This error comprises that the package does not allow transitions from a channel to itself while simulating. As the package can estimate the transition matrices with these transitions, the transition matrices

Table 3: Attribution measured by the different attribution measures in percentages

Attribution measure	LTA	FTA	LIN	Removal effect (1st order)	Shapley (1st order)	Removal effect (2nd order)	Shapley (2nd order)	Removal effect (4th order)	Shapley (4th order)
alpha	31.88	42.69	38.29	27.09	34.18	28.21	36.38	28.91	38.13
beta	14.31	5.00	10.53	12.70	10.24	12.16	10.30	9.04	9.77
gamma	0.83	0.47	0.62	0.85	0.63	0.89	0.70	0.92	0.71
delta	0.01	0.03	0.01	0.02	0.01	0.02	0.01	0.03	0.02
epsilon	0.5	2.68	1.38	3.01	2.03	3.25	2.03	3.25	1.95
zeta	0.14	0.54	0.69	1.97	1.06	1.79	1.02	1.89	1.03
eta	15.99	21.06	17.89	17.10	18.14	17.33	18.08	17.52	18.32
theta	8.12	3.3	5.17	10.11	7.74	9.65	6.74	9.04	6.00
iota	23.28	16.96	19.50	19.41	19.45	19.35	18.54	19.68	18.00
kappa	0.37	1.16	0.70	1.42	0.96	1.30	0.83	1.34	0.84
lambda	4.56	6.1	5.23	6.34	5.55	6.05	5.38	5.57	5.21
mi	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02

estimated by the package can still be used while simulating the Removal effect.

The attributions computed for all the attribution measures are shown in table 3. It shows that channel alpha is considered as most influential channel by all attribution measures. This however is not surprising as almost half of the touchpoints (42.19%, see table 1) in the dataset are with alpha. However, for most attribution measures alpha earns around 37% of the total attribution, with the Removal effect dropping even to around 28%. The eta and iota channels are the second and third most influential channels, both having around 18% attribution on most attribution measure. This is notable because iota has almost twice as many touchpoint in the dataset as eta. Then beta, theta and lambda (around 10,8 and 6% respectively) contribute most to a conversion. The attribution of beta is around its total percentage of touchpoints and theta and lambda are slightly above theirs. Finally the epsilon, zeta, kappa, gamma, delta and mi (around 2, 1, 1, 0.8, 0.02 and 0.01%) channels barely have any contribution. However all also had very low percentages of the total touchpoints. All of them actually seem to attribute a little more then would be expected only based on the percentage of touchpoints in the dataset with these channel.

Overall, the attributions of every channel seem to resemble their respective percentage of touchpoints in the dataset. This could possibly be explained by the results about the model fit. It was found there that the predictive accuracy of the models was barely better then a random predictor. This means that the estimated ad campaign didn't have a huge influence on the possible conversion and therefore could implicate that the ad campaign wasn't all that effective. This means that the marketing channels didn't have a large impact on the conversion and therefore none of the channels should have attribution scores that far exceed their appearance rate.

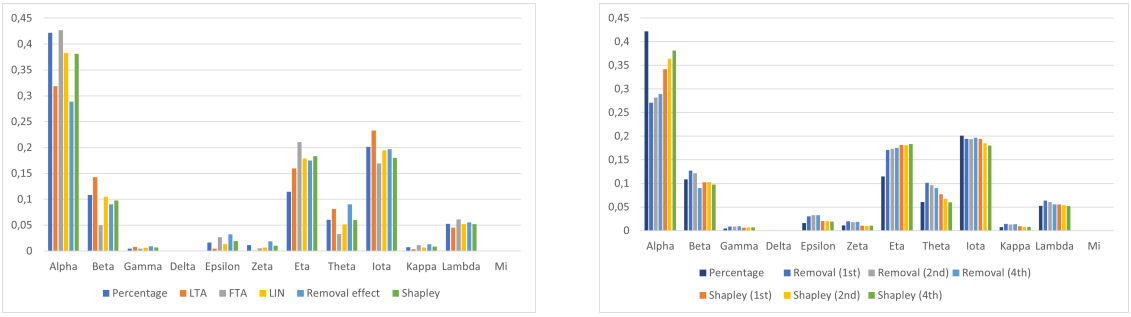

Figure 1: Attribution of the different attribution measures of the 4th order Markov graph (left) and comparison of Removal effects and Shapley values for 1st, 2nd and 4th order Markov graph and the percentage of total clicks (right)

In the left side of figure 1 a visualization of table 3 for the attribution measures in the 4th order

Markov graph is given, together with the percentage of clicks for every channel from table 1. This will be used to compare the different attribution measures in a higher-order Markov graph. Firstly, it is noted the the LIN heuristic has a very close resemblance to the Shapley value heuristic. This could again be explained by the ad campaign not having a huge influence on the eventual conversion. As the computation of the two measures is similar, the expected difference between the simple heuristic and the more Sophisticated Shapley value should be caused by the influence of the ad campaign. And because the influence of the ad campaign is very small the difference between the LIN heuristic and the Shapley value also is.

The LTA and FTA heuristics are very inconsistent attribution measures, as was expected based on Anderl et al. (2016). For almost all channels there seems to be no correlation between having a high attribution score on these heuristics compared to both the overall appearance and having a high score on the other attribution scores. This means that whether a channel is often the first or last channel on a consumer path doesn't seem to have an important influence on the overall attribution. The only noteable exception to this is eta. This channel has both a higher LTA (16%) and FTA (21%) attribution than it has touchpoints (11%) and this could explain that eta was the channel with the most significant difference between it's attribution and it's percentage of clicks.

When comparing the Removal effect and the Shapley values it stands out the the Removal effect often differs more from the percentage of touchpoints than the Shapley values. Singal et al. (2022) found that Removal effects favoured channels that had a higher appearance rate. Here it is found that the Removal effect for alpha, the channel that appears most in the dataset, is far lower (29%) than both the Shapley value for that channel (38%) and its percentage of touchpoints (42%). For the second biggest channel, iota, the Removal effect also is lower than its percentage of touchpoint (19% against 20%) but the Removal effect is higher than the Shapley value (18%). Furthermore, eta and beta, the 3rd and 4th biggest channels, also have a higher attribution measured by Shapley values than Removal effect, while almost all the smaller channels (gamma, delta, epsilon, zeta, kappa and lambda) get more attribution from the Removal effect than from the Shapley values. This is opposite to the hypothesis that Removal effect would be higher for the channels that appear more frequent in the paths.

The right side of figure 1 gives a visualization of the Removal effects and the Shapley values of the 1st, 2nd and 4th order Markov graphs from table 3. Comparing the attribution results from the the different order graphs, it can be seen that as the order of the graphs goes up, the attribution values change for the Removal effect and the Shapley value. For all of the channels, the direction of

change of an attribution measure is constant between going from a 1st to 2nd order graph as going from a 2nd to a 4th order graph. Furthermore, the direction of this change is also almost always the same for the Removal effect and the Shapley value. To illustrate, the alpha channel gets more attribution from a 2nd order graph than from a 1st order graph and more from a 4th order graph than from a 2nd order graph. Further, these directions are the same for the Shapley value. The iota channel is a notable example from this as the Removal effect for this channel goes sligtly up for higher-order graphs, but the Shapley values decreases. The direction of the change in attribution seems to be independent from the percentage of total clicks, as the attribution measures seem to converge to the percentage of clicks for some channels (alpha, theta, lamda), but diverge from the percentage of clicks for other channels (gamma, epsilon).

Overall, it can be concluded that the order of the Markov graph has a significant effect on the attribution as measured by both the Shapley values as the Removal effect. Because the effect on attribution is present for both attribution measures, is bigger for higher-order Markov graphs and seems to be independent of the percentage of clicks a channel has, this effect could be explained by the higher-order Markov graphs better estimating the carry- and spillover effects. The channels with high carry- and spillover effects get more attribution assigned to them in higher-order models and the channels with low carry- and spillover effects get assigned less.

Now the findings of this research about attributions are summarised. Firstly the attributions of every channel resembles the percentage of touchpoints of each channel. This could be caused by the low predictive power of the models. Furthermore the FTA and LTA heuristics seem very inconsistent, which is in line with the hypothesis based on the findings of Anderl et al. (2016). Furthermre, the LIN heuristic closely resembles the Shapley values which can be explained by their similar computation. The attribution given to channel by the Removal effect is to higher for the channels that appear less frequent in the dataset. This finding contradicts the finding of Singal et al. (2022) that the Removal effect is higher for channels that appear more often in the dataset. Finally, the direction in the change of the Removal effect and the Shapley values is the same when the order of Markov mode goes up. This seems to resemble the higher-order Markov models ability to better capture the carry- and spillover effects.

6 Conclusion

This research investigated the research question "*how do different attribution measurements compare for higher-order Markov models and between higher-order Markov models and normal Markov models?*" In order to do this, Markov models for the consumer paths with different order specifications were computed and compared to a benchmark logit model. Higher-order Markov models were found to be a better model for the consumer paths than normal Markov models and also had a representative performance compared against the logit model.

After the Markov models were specified, attribution was computed in the Markov model. The Removal effect and Shapley value were used as data-driven attribution measures and were compared to some heuristic attribution models. It was found that the attribution assigned to the channels by the heuristic models is very inconsistent. Furthermore, the Removal effect assigns less attribution to the channels most frequently present in the consumer paths compared to the Shapley values. This is in contrast with earlier findings from the existing literature. Finally, attribution in higher-order Markov models was found to better reflect the carry- and spillover effects between the different channels.

This research has several limitations. The information about the consumer paths is probably not complete for multiple reasons. The consumer path data is only collected from online marketing channels. Therefore, possible interactions with offline marketing channels and the influence of these to a possible conversion, are ignored. Furthermore, the data used is collected using cookie tracking. This means that the consumer paths comprise the sequence of contacts a user had with a company on one device. However, many people have multiple devices and thus the consumer paths of one person can be split up between different devices. Consumer paths can also be broken by a consumer deleting their cookie information. Because of these reasons incorrect consumer paths can be recorded.

Secondly, the ability of the models to explain the consumer paths on this dataset was poor. The attribution of the different channels only reflects the part of the total conversions that can be explained by the Markov model. Therefore, the total part of the conversions that is explained and distributed between the different channels by the attribution measures is very small. Thus, further research on more impactful ad campaigns is needed to verify the results.

Another limitation of this research is that the findings are company or even campaign specific. Before the attribution is measured, managers have already made decisions on how the budget is

allocated between the different channels. This could have an important effect on the effectiveness of the different marketing channels on consumers and therefore affect the measured attribution. Furthermore, also decisions about which consumers are targeted by the the campaign, are made beforehand. The decision between targeting recurring or new customers heavily impacts which channels are most effective, and thus get the most attribution. Overall, endogeneity is by definition a property of the attribution problem. Therefore, further research could be made to verify the findings of this research for other companies or campaigns.

Finally, further research could also use time elements to account for the time difference between different touchpoints. Ji et al. (2016) previously used time elements in attribution modelling. Extending the Markov model framework using time elements could help to better estimate the consumer journey.

References

Abhishek, V., Fader, P., & Hosanagar, K. (2012). Media exposure through the funnel: A model of multi-stage attribution. *Available at SSRN 2158421*.

Anderl, E., Becker, I., Von Wangenheim, F., & Schumann, J. H. (2016). Mapping the customer journey: Lessons learned from graph-based online attribution modeling. *International Journal of Research in Marketing*, *33*(3), 457–474.

Arava, S. K., Dong, C., Yan, Z., Pani, A., et al. (2018). Deep neural net with attention for multi-channel multi-touch attribution. *arXiv preprint arXiv:1809.02230*.

Archak, N., Mirrokni, V. S., & Muthukrishnan, S. (2010). Mining advertiser-specific user behavior using adfactors. In *Proceedings of the 19th international conference on world wide web* (pp. 31–40).

Berman, R. (2018). Beyond the last touch: Attribution in online advertising. *Marketing Science*, *37*(5), 771–792.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Dalessandro, B., Perlich, C., Stitelman, O., & Provost, F. (2012). Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy* (pp. 1–9).

Flosi, S., Fulgoni, G., & Vollman, A. (2013). If an advertisement runs online and no one sees it, is

it still an ad?: Empirical generalizations in digital advertising. *Journal of Advertising Research*, *53*(2), 192–199.

Gartner Research. (2019). *Hidden forces that will shape marketing in 2019.* (`https://www.gartner.com/smarterwithgartner/4-hidden-forces-that-will-shape-marketing-in-2019/`)

Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, *12*(1), 1–20.

IAB. (2019, may). *Iab internet advertising revenue report.* (`https://www.iab.com/wp-content/uploads/2019/05/Full-Year-2018-IAB-Internet-Advertising-Revenue-Report.pdf`)

Ji, W., Wang, X., & Zhang, D. (2016). A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th acm international on conference on information and knowledge management* (pp. 1373–1382).

Kannan, P., Reinartz, W., & Verhoef, P. C. (2016). *The path to purchase and attribution modeling: Introduction to special section* (Vol. 33) (No. 3). Elsevier.

Katz, R. W. (1981). On some criteria for estimating the order of a markov chain. *Technometrics*, *23*(3), 243–249.

Li, H., & Kannan, P. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, *51*(1), 40–56.

Shao, X., & Li, L. (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).

Shapley, L. (1953). Quota solutions op n-person games1. *Edited by Emil Artin and Marston Morse*, 343.

Singal, R., Besbes, O., Desir, A., Goyal, V., & Iyengar, G. (2022). Shapley meets uniform: An axiomatic framework for attribution in online advertising. *Management Science*.

Tueanrat, Y., Papagiannidis, S., & Alamanos, E. (2021). Going on a journey: A review of the customer journey literature. *Journal of Business Research*, *125*, 336–353.

Vu, K., Clark, R. A., Bellinger, C., Erickson, G., Osornio-Vargas, A., Zaïane, O. R., & Yuan, Y. (2019). The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies. *BMC medical informatics and decision making*, *19*(1), 1–8.

A Description programming files

The file "thesis r descriptive statistics" is used to compute the descriptive statistics in table 1. The file "thesis r auc values" is used to compute the AUC values in table 2. The AUC values for the markov graphs are computed using the "*ChannelAttribution*" package and are in the markov_auc object, while the AUC value for the logit model is printed with the last line. The BIC values for table 2 are computed using the file "thesis r bic values". The maximum order for which the BIC values are computed can be selected and all values are storing in the BIC object. The 1st number in the vector belongs to the 1st order Markov graph, the 2nd number to the 2nd order Markov graph, etc. The attributions in table 3 are computed using "thesis r file attributions". The attributions for Shapley values and the Removal effect are in the percentage_shapley and percentage_removal objects respectively. The places in the vector don't resemble the alphabetical order of the channels, but the order is printed (line 84) right before the percentage_shapley and percentage_removal are computed.