Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Bachelor in Economics and Econometrics

Temporal Graph Neural Network with Attention for Multi-Touch Attribution

Enzo Akira Ido (534787)

Supervisor:	Prof. Dr. Kathrin Gruber
Second assessor:	Name of your second assessor
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

The internet has revolutionized the filed of marketing. With the increased use of digital platforms, more information has been available to track the customer journey. However, in order to take advantage of this data, the correct methods have to be used. Multi-touch attribution models have been used to study this type of data but still present some limitations. One of the main models currently studied is the Markov chain and the Shapley value. While possessing a strong theoretical foundation, they can easily become computationally intractable. To solve this issue, this research proposes to use a temporal graph neural network with attention (TempGAN). When comparing these models, it is found that TempGAN is able to outperform the Markov model in terms of Brier score, AUC, and F1 score while still producing realistic attribution scores. However, the model still struggles to distinguish positive and negative cases.

1 Introduction

The internet has revolutionized the field of marketing. With the abundance of data resulting from these developments, marketing analysis has become increasingly quantitative. The adoption of online platforms like Instagram, Amazon, and Netflix, has generated an abundance of data that is of extreme value to marketers. More specifically, according to The World Bank (2021), as of 2021, 60% of the world's population had access to the internet. The cloud software company DOMO estimates that Americans alone use 4,416,720 GB of data per minute (Martin, 2019).

A specific marketing challenge that has gained popularity with the rise of the internet is the multi-touch attribution (MTA) problem. Companies use multiple communication channels to engage customers and influence their purchasing behavior towards specific products or services. Every interaction the consumer has with a given advertisement has an impact on his/her behavior. The objective of MTA is to quantify the level of importance that each channel had in leading the consumer to the conversion (Shao & Li, 2011). What makes this problem relevant is that knowing the added value of each channel allows marketers to optimize the customer journey, leading to more conversions and higher revenues.

Initial literature on the MTA problem focused on the use of rule based heuristics since these are simple to implement and intuitive. Methods that fall under this category include *last touch attribution* (LTA), *uniform weights*, and *customized weights*. While simple and intuitive, these methods cannot fully capture the complexity of the customer's interactions with the multiple channels. Dalessandro, Perlich, Stitelman and Provost (2012) approached this limitation by using game theory, more specifically the Shapley value (SV), to determine the attribution scores. However, Singal, Besbes, Desir, Goyal and Iyengar (2022) argued that computing the SV value is generally intractable for this problem and proposed to use a Markov Chain to model the customer's journey and generate attribution scores using the *counterfactual adjusted Shapley value* (CASV). The main limitation of this method is that the number of channels cannot be significantly expanded since this would cause the Markov chain to also become computationally intractable.

Exploring methods that can deal with a greater number of channels is relevant since, realistically, advertisers use multiple channels to reach the consumer. Ignoring such information could lead to sub-optimal predictions and inaccurate attribution scores. With a limitation on the number of channels, the advertiser cannot accurately describe the customer's journey and hence loses valuable information. A possible way around the issue would be to group similar channels together. However, it can be argued that every channel has unique characteristics that influence the consumer in a different way. Thus, it is important to develop a model which allows advertisers to keep the original number of channels even if this number is very high.

This research proposes to solve this issue by using Graphical Neural Networks (GNNs) with attention. This type of neural network takes a graph as an input, learns from its structure and generates predictions. In the case of the MTA problem, a logical way to represent the consumer's interactions with the channels would be a graph. Being an algorithm that can use information coded in such a structure, GNNs seem like an interesting method to apply to this challenge. Furthermore, GNNs have been proven to effectively deal with complex data structures (Zhou et al., 2020). More specifically, in this research, the TempGAN algorithm proposed by Mohan and Pramod (2021) will be extended and applied to the context of MTA.

When comparing this model's performance to the Markov chain, it is found that TempGAN produces a better Brier score, AUC, and F1 score. When considering more complex data inputs, TempGAN again outperforms the Markov chain when enough training data is available. However, while both models achieve a desirable Brier score that is significantly close to 0, they do not produce outstanding AUC and F1 scores. Regarding MTA, both models produce similar attribution scores and assign Facebook as the most influential channel and Online Video, Online Display, and Paid Search as the least. Overall, TempGAN has shown to be a highly calibrated model that produces realistic attribution scores but still struggles to differentiate positive and negative instances.

The contributions of this paper are twofold. The first is that, to the best of my knowledge, GNNs with attention have not been applied in the context of the MTA problem. The second is that, I extend the work of Mohan and Pramod (2021) by applying their work in the context of graph classification.

The remainder of the paper is structured as follows. Section 2 presents a literature review on the current methodologies applied in the context of MTA as well as on GNNs. Section 3 describes the data used in this research while Section 4 provides a detailed overview of the methods applied. Finally, I present the main outcomes of the research in Section 5 and conclude in Section 6.

2 Literature Review

2.1 Multi-touch attribution

A well established concept in marketing is the 4 *P's* popularized by Borden (1964). Each *P* refers to an essential element for a successful marketing campaign and they are: Product, Price, Place and Promotion. In this research, the focus is drawn to the Promotion element. This *P* refers to the methods utilized to communicate with the consumer about the product/service being sold. Even though many changes have taken place since the popularization of this concept, it still remains very relevant today.

A significant change that took place since the establishment of the 4 *P's* is the rise of digital marketing. With new online platforms like Instagram, Amazon, and Netflix, companies have found alternative ways to interact with the consumers. Even though these platforms were not necessarily developed as a marketplace, they have become one of the main advertising channels for many companies. Noticing this trend, online platforms have also started to profit by charging for such a service. As a matter of fact, the digital advertising industry achieved a revenue of USD$ 125 billion in 2019, 16% higher than in 2018 (Hogan, Bruderle, Silverman & Krasnow, 2020).

Despite the widespread adoption of this new advertising strategy, a few new challenges have surfaced, drawing the focus and attention of industry professionals. One significant challenge pointed out by Gordon et al. (2021) is the ad effect measurement, which concerns "the estimation of incremental effects of advertisements on consumer behaviors". In fact, 75% of brand professionals consider this the biggest threat to digital ad budgets in 2019 (Benes, 2019). One of the reasons behind such a concern is that, without a visibility on the ad effect measurement, marketers cannot optimally allocate their budget to the best performing channels and hence cannot obtain the best Return on Advertising Spend (ROAS). This is even more concerning given that decisions regarding digital marketing campaigns have previously been largely based on trial-and-error (Alhabash, Mundel & Hussain, 2017).

Fortunately, most of these platforms provide a substantial amount of highly granular data that can be leveraged using the correct tools. One of these tools is called a Marketing Mix Model (MMM) in which the researcher regresses the total revenue, on the spend per channel and other independent variables (Gordon et al., 2021). Even though these models provide a good indication of the efficiency of the budget allocation, its dependence on aggregated data cannot fully capture channel-specific ad effects and the influence of the inter-channel relations on consumer behavior. For this type of analysis, multi-touch attribution (MTA) is commonly used. Unlike MMMs, multi-touch attribution models take a bottom-up approach and use individual level data to assign credits to each of the channels. Several MTA models, with varying complexity, have been proposed. More basic models include *last touch attribution*, where all the credit is allocated to the last channel that was interacted with before conversion, and *uniform weights*, where the credit is equally assigned to all channels. Examples of more complex attribution models are Markov chains and neural networks, both of which are detailed in Sections 2.2 and 2.3.

MTA has become especially relevant given the additional complexity in the customer journey introduced by new digital platforms. In this new era, the customer decision process has transitioned from a series of discrete activities to a continuous process (Tueanrat, Papagiannidis & Alamanos, 2021). This has led to a decline of the traditional single-channel approach and companies are now seeing the need to adopt a more comprehensive mindset to influence the full customer journey (Faulds, Mangold, Raju & Valsalan, 2018). As a result, customers are enjoying a freedom to shape their own journey like never before and can seemingly switch and integrate different channels (Herhausen, Kleinlercher, Verhoef, Emrich & Rudolph, 2019; Hu & Tracogna, 2020). Now, the brand knowledge acquired from each interaction can be more easily transferred and accumulated to the subsequent channels (Tueanrat et al., 2021). While this could be an effective way to increase brand awareness and potentially lead to more conversions, if not done right, it could also be highly counterproductive. Namely, when implementing an omnichannel approach, the migration effect and channel cannibalisation need to be considered (Fornari, Fornari, Grandi, Menegatti & Hofacker, 2016). While the former refers to the phenomenon where customers change channels altogether instead of using different ones to complement their experience, the latter refers to the tendency of an increased satisfaction towards one channel being accompanied by a decrease in consumer purchases in another (Ansari, Mela & Neslin, 2008; Chiu, Hsieh, Roan, Tseng & Hsieh, 2011). Both of these issues could lead to an increase in operational cost without a similar compensation in revenue, leading to profit losses. Even though these issues can still be found in traditional mass media channels like TV and radio, they are even more accentuated in the digital marketing space due to the reasons presented in the beginning of the paragraph. This makes the use of MTA even more critical when developing a successful marketing strategy in today's society.

MTA can also contribute to marketing literature by helping to identify trends in consumer behavior. This field of economics has been studied extensively and is critical for marketers to develop a successful marketing strategy. The rise of the digital economy has led to significant behavioral changes (Krajnović, Sikirić & Bosna, 2018) and through MTA, researchers can identify the channels that appeal to specific groups and establish correlations between the demographic attributes and the unique qualities of those channels.

Finally, Gordon et al. (2021) emphasizes the benefits of combining experimental findings with observational data such as the ones produced by MTA. Experiments can have several designs but, in general, they compare the differences between two groups: one that receives the treatment and another one that does not. With a large enough sample size, the only considerable difference between the two groups is the variable being studied. Because of this, experiments are very successful in determining causal relationships and have been applied extensively in marketing. However, in order for experiments to produce accurate insights that can be generalized to other settings, the group size must be large enough, and in many situations, this could prove to be a major constraint. In fact, Gordon et al. (2021) mentions that the minimum sample size for ad experiments can exceed 500,000 test subjects and even then, the average confidence interval produced is around 100%. As a way to reduce this uncertainty, one could compare the outcomes of the experiment with the attribution scores generated by a model trained on observational data. If both produce similar results, then more confidence can be attributed to the outcomes of the experiment.

Overall, MTA provides valuable contributions to marketing literature and hence, a model that can produce realistic scores must be developed. Out of all the MTA models already discussed, this research will focus on Markov chains and neural networks.

2.2 Markov Chain

Markov chains are an effective way to model processes where the current state is dependent on the previous ones. Here, a state is defined as the circumstance the modelled process finds itself at a given point in time. The exact definition is context specific and it is up to the researcher to decide. Nevertheless, having defined a state space that characterizes all the possible states that the process can take, the Markov chain then estimates transition probabilities between them.

Based on this, one possible way to model the customer journey is by considering the channels as the state space. More specifically, we can consider the interaction path of a given consumer as an absorbing Markov chain with two absorbing states: quit and conversion. In these types of Markov chains, the absorbing states once entered, can never be left. It makes sense to consider quitting and converting as absorbing states since they mark the end of the journey for that specific consumer and no further interactions will take place. Anderl, Becker, von Wangenheim and Schumann (2016) implemented this structure and were able to obtain significant predictive performance improvements over simple logit models.

Singal et al. (2022) also applied Markov chains in the context of the MTA problem. In their research, the Markov chain was similar to the one described previously. However, the authors also included an action space which was defined as the set of actions an advertiser can take (Singal et al., 2022). Furthermore, the state space was not considered to be the channels but rather the conversion funnel used in marketing literature. This funnel is a theoretical concept that captures the journey of the customer from being unaware of the product/service being offered, to becoming interested in it, and finally converting. The four stages usually described in this funnel are: Awareness, Interest, Desire, and Action. These are exactly the states used in the Markov chain modelled by Singal et al. (2022). While the researchers did not provide empirical results of this implementation, they discussed theoretical results that showed that it is compatible with a modified version of the Shapley value (SV) which they proposed, called *counterfactual adjusted Shapley value*. This metric will be discussed in the following paragraphs.

While these models have been applied in the context of MTA, they can become computationally intractable as the dimension of the input data increases. More specifically, the number of parameters to be estimated grows exponentially and this can quickly grow to be too large. Hence, a model without such a limitation is desirable.

Markov chains are a good way to model the customer journey, however, they do not directly output attribution scores. Hence, many researchers have used it in combination with the Shapley value to fully tackle the MTA problem.

The Shapley value is a concept that originated in game theory and is used to allocate the total value generated by a coalition to its individual players. To compute it, one needs to consider all the possible permutations of the players and determine their average marginal contribution. Dalessandro et al. (2012) applied this concept to generate attribution scores and highlighted that this approach has several advantages. Among them are that the Shapley value has a strong theoretical foundation, provides a fair distribution and has other desirable features like efficiency, symmetry and linearity.

However, Singal et al. (2022) pointed out that computing the SV is usually computationally intractable for MTA and one has to rely on approximations. Furthermore, the authors argue that the Shapley value is not counterfactual in nature, which is desirable for MTA. To account for this, the authors proposed the *counterfactual adjusted Shapley value* (CASV).

This value is actually calculated as the difference between two Shapley values:

$$\psi_s^{a,shap}(M) = \pi_s^{a,shap}(M) - \pi_s^{a,shap}(M_s^a) \tag{1}$$

With s representing a state, a an action and M the Markov model. The way this formula incorporates the counterfactual is through M_s^a which represents the Markov model that replaces the transition probabilities of (s, a) with those of $(s, 1)$, 1 being the no-ad action.

A drawback of the CASV is that one needs data on the value generated by a coalition without any players and many times this is not available. This is especially the case for online multi-touch attribution since it is not always possible to determine how many consumers converted without ever coming in contact with one of the company's channels. Furthermore, a limitation of both the Shapley value and the CASV is that they are computationally intensive as all permutations of the players have to be considered.

2.3 Temporal Graph Attention Network

Graph Neural Networks (GNNs) are a type of neural network that can learn from information that is structured as a graph (Bronstein, Bruna, LeCun, Szlam & Vandergheynst, 2017). Many variations of this structure have been proposed and one that has gained popularity is the Graph Convolution Network (GCN). These methods also learn from information structured as a graph but "aggregate node information from the neighborhoods in a convolutional fashion"(Zhang, Tong, Xu & Maciejewski, 2019). A further extension of this structure is the Graph Attention Network (GAT). In this structure, an attention layer is included in order to allow the network to focus only on the most relevant parts of the input data (Velickovic et al., 2017).

The methods mentioned in the previous paragraph focus on static networks where the nodes and edges do not change over time. In many cases, the problem requires a non-static structure where time information is incorporated into the graph. This is the case with the problem at hand. Namely, consumers interact with the channels in a specific order that is relevant to their decision making process and it cannot be ignored. Mohan and Pramod (2021) specify two non-static networks: dynamic and temporal. While the former refers to a graph where the nodes and edges change over time, the latter refers to a structure where a timestamp is associated to each edge. Based on this definition, it seems more appropriate to model the MTA problem as a temporal network since the timestamp of each edge can reflect the order in which the consumer interacted with each channel.

Mohan and Pramod (2021) proposed a *temporal graph attention network* (TempGAN) to deal with such graphs. To apply this model, one first needs to structure the problem as a graph using an adjacency matrix A, with node feature matrix F, and a PPMI matrix M. More detail regarding these inputs are provided in Section 4.3.1.

These inputs are then fed into a two layer neural network to produce embeddings for each of the nodes. At each hidden layer, attention and convolution are applied. More details regarding the architecture of TempGAN is provided in Section 4.3.2.

TempGAN efficiently handles complex graph structures and this removes limitations regarding the number of nodes and channels. This sets it apart from other attribution models, such as the Markov chain and the Shapley value that rely on counting a significant number of permutations. Additionally, unlike simpler attribution models like *last touch attribution* and *uniform weights*, TempGAN incorporates the sequence of consumer-channel interactions. While a Markov chain also has this advantage, TempGAN's convolutional structure can capture more complex relations between non-neighboring nodes and channels. All of these advantages are crucial for accurately modeling the customer journey and hence, this research will apply the

TempGAN model to the MTA problem.

3 Data

The data that is going to be used for this research is obtained from a Kaggle project [1]. This data set contains observations from July 2018 regarding the interactions consumers had with a company's advertising channels. Six variables are included in the data set and they are summarized in Table 1. With this data set, it is possible to trace the customer journey through the Cookie and Timestamp variables. Furthermore, the Conversion variable allows me to assign a target binary variable to each of the customer's journey.

Table 1: Description of the variables included in the data set

Variable	Description
Cookie	Anonymous user ID
Timestamp	Date and time of the interaction
Interaction	Describes the type of interaction that occurred. Can either be "impression" or "conversion".
Conversion	Binary variable indicating whether conversion took place or not
Conversion value	Revenue generated by the conversion
Channel	Marketing channel that brought the customer to the website. Channels included are: Facebook, Instagram, Online Display, Online Video, and Paid Search

The data set contains $586,737$ observations, each representing an interaction a consumer had with one of the channels. As can be seen in Table 3, 30% of these interactions were with Facebook, 26% with Paid Search, 19% with Online Video, 13% with Instagram, and 12% with Online Display. Also important to note that more than 50% of the interactions that led to a conversion were with Facebook and Online Video. The number of unique consumers is $240,180$ and the average number of interactions is 2.44. Of the total number of consumers, $17,639$ converted (7.35%). The maximum number of interactions a consumer had was 134 and the minimum was 1. Table 2 summarizes the statistics regarding the data set while Table 3 summarizes the distribution of the number of interactions by channel.

In this research, subsets of the entire data set are used to evaluate the performance of the models on data of varying complexity. More detail regarding this analysis is provided in Section 4.4. The subsets of the data that are used are: 1) paths with a length higher than or equal to 5 and 2) paths with a length higher than or equal to 10. Statistics on these subsets are also presented in Tables 2 and 4. An interesting change to note is that the conversion rate increases significantly for the data set with paths of length higher than 10. Also, while the second most present channel is Paid Search for the full data, on the subsets this changes to Online Video.

While this data provides the necessary variables to perform the analysis, there are some limitations. First, it is not exactly clear how this data was retrieved. The author of the project does not specify whether the data was simulated or extracted from a real business. Furthermore, the data has already been previously cleaned and there was no mention of how this was done.

[1] https://www.kaggle.com/code/hughhuyton/multitouch-attribution-modelling

Table 2: Summary statistics of the full data set and of the two subsets

	Full Data	Length ≥ 5	Length ≥ 10
Number of Interactions	586,737	232,530	106,706
Number of Consumers	240,108	26,805	6,716
Average Journey Length	2.44 (3.10)	8.67 (6.01)	15.89 (8.33)
Average Number of Channels per path	1.29 (0.57)	2.10 (0.84)	2.28 (0.92)
Number of Conversions	17,629	3,767	1,342
Conversion Rate	7.35%	1.62%	19.98%

Note: Standard deviations are shown in paranthesis.

Table 3: Distribution of the number of interactions by channel for the full data

	Full Data		
	Conversion	Non Conversion	All
Facebook	21,464 (33%)	154,277 (30%)	175,741 (30%)
Instagram	9,157 (14%)	66,044 (13%)	75,201 (13%)
Online Video	17,103 (27%)	96,199 (18%)	113,302 (19%)
Paid Search	11,342 (18%)	140,098 (27%)	151,440 (26%)
Online Display	5,380 (8%)	65,673 (13%)	71,053 (12%)

Note: Percentages refer to column total.

Given the scarcity of free data sets on customer journey, this data set proved to be the most complete and hence it was still used for this research.

4 Methodology

4.1 Markov Chain

The structure of the Markov chain presented in Anderl et al. (2016) is favored over the one presented in Singal et al. (2022) as the latter provides strong empirical results. More specifically, with n channels, a first-order Markov chain is defined by a non-absorbing state space $S = \{start, s_1, s_2, ..., s_n\}$ and two absorbing states $\{conversion, quit\}$. Here, each s_i represents a channel and the special state *start* is added to account for the consumer's first interaction. Furthermore, each path is assigned one of the absorbing states depending on the corresponding conversion variable. While Chierichetti, Kumar, Raghavan and Sarlos (2012) establish that click streams do not exactly follow a first order Markov chain, this is still used in this research due to the limitation presented by the data. To estimate a higher degree Markov chain, data on consumer paths with a length higher than the chosen degree is necessary and by filtering such cases, the number of training instances would be significantly reduced.

Having created a path for every consumer using the previously mentioned states, it is possible to estimate a transition probability matrix $P \in \mathbb{R}^{n+1 \times n+1}$. This matrix is estimated using only the training set. Before making forecasts on the evaluation set, the fundamental matrix $F \in \mathbb{R}^{n+1 \times n+1}$ must be calculated. This can be done using the equation $F = (I_{n+1} - P)^{-1}$ where I_{n+1} is the $n+1$ identity matrix. Note that the (i,j)-th entry of the resulting matrix will equal the expected number of visits to state j, given that the first state is i. Here the vector $p_c = \{p_{s,conversion}\}_{s \in S} \in \mathbb{R}^{n+1}$ is also defined. In other words, vector p_c contains the probabilities

Table 4: Distribution of the number of interactions by channel for the subsets of the data

	Length $\geq$ 5			Length $\geq$ 10		
	Conversion	Non Conversion	All	Conversion	Non Conversion	All
Facebook	14,313 (36%)	64,801 (34%)	79,114 (34%)	9,204 (38%)	30,578 (37%)	39,782 (37%)
Instagram	6,078 (15%)	27,742 (14%)	33,820 (15%)	3,795 (16%)	13,108 (16%)	16,903 (16%)
Online Video	12,772 (32%)	49,819 (26%)	62,591 (27%)	8,409 (35%)	24,711 (30%)	33,120 (31%)
Paid Search	4,406 (11%)	32,043 (17%)	36,449 (16%)	1,891 (8%)	9,254 (11%)	11,145 (10%)
Online Display	2,242 (6%)	18,314 (10%)	20,556 (9%)	881 (4%)	4,875 (6%)	5,756 (5%)

Note: Percentages refer to column total.

of transitioning from each of the non-absorbing states s to the *conversion* state. Using both the fundamental matrix F and the vector p_c, it is possible to determine the eventual conversion probability from each channel by computing $h = Fp_c \in \mathbb{R}^{n+1}$. Vector h will contain the eventual conversion probabilities based on the training data only and the validation path must be incorporated somehow. This can be done by taking the weighted average of the probabilities in h. If $w_i \in \mathbb{R}^{n+1}$ is a vector containing the frequencies of each channel in the validation path i, the weighted sum can be computed as $\frac{1}{W_i} w_i^t h$ with $W_i = \sum_{j=1}^{n+1} w_{ij}$. Put differently, the final forecasted conversion probability will equal the eventual conversion probability of the training data weighted by the frequencies of each channel in the evaluation instance.

4.2 Shapley Value

The attribution scores derived using the Shapley value are used as the benchmark for comparison with the ones generated by the neural network. Using the *counterfactual adjusted Shapley value* proposed in Singal et al. (2022) would be ideal, however in this application it is not possible since there is no data on the value of the coalition with the no-ad action. For this reason, the standard Shapley value used in Dalessandro et al. (2012) is applied.

The data set for this research can be described as $\lambda = \{S = \{s_1, ..., s_n\}, \gamma = \sum Y, m\}$ where S is the set of all n channels, $\gamma = \sum Y$ is the total number of conversions and m is the total number of consumers. Using this notation, Dalessandro et al. (2012) defines the Shapley value of channel i as:

$$V_i = \sum_{C \subseteq S \setminus s_i} \omega_{C,i} * [E[\gamma | C \cup s_i] - E[\gamma | C]] \tag{2}$$

$$\omega_{C,i} = \frac{|C|!(|S| - |C| - 1)}{|S|!} \tag{3}$$

In order to make this computation feasible, the expectation is taken to be the number of conversions achieved by coalition C in the training data.

4.3 Temporal Graph Attention Neural Network (TempGAN)

The following section describes the necessary steps to prepare the inputs for TempGAN and also its neural network architecture.

4.3.1 TempGAN Inputs

In order to apply TempGAN, each consumer's interaction path must be modeled as a temporal graph $G = (V, E_T, T)$. Here, V represents the set of vertices - the channels, E_T represents the set of time-stamped edges - the sequence in which the consumer interacted with the channels, and T represents the set of time stamps. The temporal graph can be represented as a set of triplets (v_i, v_j, t) with $t \in T$ being the time of interaction between vertex v_i and v_j. In this graph, a *start* node is also included to account for the first interaction. This is implemented using the NetworkX package in python.

In addition to this, the graph must also be represented as an adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, initial feature matrix $F \in \mathbb{R}^{|V| \times k}$, and PPMI matrix $M \in \mathbb{R}^{|V| \times |V|}$ as described below. The initial feature matrix is set as a random matrix and the steps to compute the PPMI matrix are described in the following paragraph.

- **Adjacency matrix** A: $|V| \times |V|$ matrix. Element (i, j) of matrix A is equal to 1 if there is an edge between nodes i and j, and 0 otherwise.
- **Feature matrix** H: $|V| \times k$ matrix with k being the number of features of each node.
- **PPMI matrix** M: $|V| \times |V|$ matrix which captures the statistical relationship between the nodes of the graph.

The *positive point wise mutual information* (PPMI) is a statistic that captures the co-occurence pattern of two nodes. In the TempGAN algorithm, this information is crucial for determining the non direct neighboring nodes to consider during convolution and attention operations.

The PPMI is approximated by first calculating the PMI using the co-occurrence statistics of the nodes. The formula below shows how to determine the PMI between nodes v_i and v_j:

$$PMI(v_i, v_j) = log(\frac{\frac{N(v_i, v_j)}{N}}{\frac{N(v_i)}{N} \frac{N(v_j)}{N}}) \tag{4}$$

Here $N(v_i, v_j)$ is the number of time respecting paths from node v_i to v_j, $N(v_i)$ is the number of paths containing node v_i and N is the total number of paths. Note that a time respecting path from v_i to v_j is defined as a set of edges $E = (v_i, v_k, t_1), (v_k, v_l, t_2), ..., (v_n, v_j, t_n)$ such that $t_1 \leq t_2 \leq ... \leq t_n$.

Finally, to obtain the PPMI simply set the negative entries of the PMI matrix to 0:

$$PPMI(v_i, v_j) = max(PMI(v_i, v_j), 0) \tag{5}$$

4.3.2 TempGAN Architecture

The initial part of the TempGAN architecture that creates node embeddings is based on the research of Mohan and Pramod (2021). This is extended by applying another attention layer to generate conversion probability estimates. The first step is to apply a linear transformation to the feature matrix H using the parameter matrix $W_1 \in \mathbb{R}^{|V| \times |V|}$ to generate high-level features:

$$H' = W_1 H \tag{6}$$

Then, an attention coefficient matrix E is computed by multiplying H' with a shared attention weight $A_1 \in \mathbb{R}^{k \times |V|}$:

$$E = H' A_1 \tag{7}$$

Then, to provide non-linearity and normalize the attention coefficients, a leaky relu and softmax function are applied to E:

$$E_{ij} = softmax_j(leakyRelu(E_{ij})) \tag{8}$$

In the next step the concept of temporal neighborhood is utilized. The temporal neighborhood of a node v_i at time t consists of the set of nodes connected to v_i which have an edge with timestamp bigger than t. Since temporal graphs are considered, it is necessary to set the attention weights of the edges which are not in the temporal neighborhood to 0. This can be done by:

$$\hat{E_{ij}} = \begin{cases} E_{ij} & \text{if } M_{ij} + A_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Finally, for each node, the model propagates the high-level features of the nodes in the temporal neighborhood using a relu function and gives more or less importance to each node depending on its learned attention weights:

$$\hat{H} = \sigma(\hat{E} W_1 H) \tag{10}$$

In order to aggregate the node embeddings into a single graph embedding, another attention layer is applied. This is done in a similar fashion as the previous layer where a learnable attention weight $A_2 \in \mathbb{R}^{|V|}$ is multiplied by embedding matrix $\hat{H}$:

$$\vec{h} = \hat{H}^T A_2 \tag{11}$$

This vector $\vec{h}$ is then fed into a fully connected layer and a sigmoid activation function to produce the probability estimate.

$$\hat{y} = sigmoid(\vec{h}^T W_2) \tag{12}$$

To determine the loss and perform back propagation, the binary cross entropy loss function is used. Furthermore, given that A_2 is a learnable parameter that assigns more/less weight to each of the node embeddings when determining the final graph representation, it is possible to interpret it as the attribution score of each channel. Finally, a pseudo-algorithm for TempGAN can be seen in Algorithm 1.

```
Algorithm 1: TempGAN Algorithm
  Data: Adjacency Matrix A, PPMI Matrix M, Initial Feature matrix F, observed binary
        variable y, epochs
  Result: Predicted conversion probability $\hat{y}$, Attribution scores A$_2$
  for 1 to epochs do
      $H' = W_1 H$;
      $E = H' A_1$;
      $\hat{E_{ij}} = \begin{cases} E_{ij} & \text{if } M_{ij} + A_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$ ;
      $\hat{H} = \sigma(\hat{E} W_1 H)$ ;
      $\vec{h} = \hat{H}^T A_2$;
      $\hat{y} = sigmoid(\vec{h}^T W_2)$;
      Objective Function L = binary cross entropy loss$(\hat{y}, y)$;
      Adam(L)
  end
  return $\hat{y}, A_2$;
```

4.4 Model comparison

Given the unobservable nature of the attribution scores, it is not clear how to objectively determine whether the scores generated by one method are better than the other. Hence, the analysis of this output will mainly revolve around a subjective comparison and the study of their implications in a marketing strategy.

Nonetheless, Li and Kannan (2014) mention that attribution models must also have a strong predictive power. This provides an opportunity to more objectively analyse the accuracy of the attribution scores. More specifically, it is possible to assume that the model that produces a better predictive accuracy also produces better attribution scores.

To assess the predictive accuracy, several metrics will be used. These include the Brier score, AUC, and the F1 score which are all standard measures when assessing binary outcome predictions. Furthermore, to evaluate the change in performance of the different models when changing the complexity of the data, the models will also be implemented on three sets of data: 1) full data set, 2) only paths with a minimum length of 5, 3) only paths with a minimum length of 10. Here it is assumed that a longer path is more complex.

5 Results

In the following section, the results of the models are analyzed and compared. Note that all models were trained on 70% of the data and evaluated on the remaining 30%. Furthermore, to prevent look ahead bias, the consumers were ordered in chronological order based on their last interaction and then the train-test split was made. This ensures that the models' out-of-sample predictions are based only on the available information at the time the forecast was made.

5.1 Model Performance

In this section, the performance of the two proposed models are compared to each other along with a Naive model. The latter model consists of simply predicting a 0 for all observations. Given that the data set is highly imbalanced, this model serves as a benchmark for the other two.

Given the definition of the Brier score as the mean squared error of the predicted probabilities, a well calibrated model would achieve a value close to 0. From Table 5, it is possible to see that all three models achieve desirable values with scores less than 0.05. However, when looking at the AUC and F1 scores, they are both considerably small for all models. Both of these findings imply that the proposed models are well calibrated but cannot properly distinguish between positive and negative outcomes. A possible explanation for this could be due to the highly imbalanced data set. Given that only 7.35% of the instances are positive, both models mostly predict low probabilities. This guarantees a low Brier score but also provides a weak discriminatory power. Furthermore, when using a threshold of 0.5 to classify instances based on the predicted probabilities, no instances are assigned as a conversion. In order to be able to compute the F1 score, the threshold was reduced to 0.1. This shows that indeed the models are producing very small probabilities.

Given these limitations, it is still possible to see that TempGAN achieves improvements compared to the Markov model. First, TempGAN is able to achieve a smaller Brier score compared to the Naive model (predicts 0 for all instances) and the Markov chain. Furthermore, TempGAN outperforms the Naive model in terms of AUC while the Markov chain does not. Finally, the F1 score for TempGAN is also higher than the one obtained by the Markov Chain. All of these findings show that TempGAN is more calibrated and has a higher predictive capacity compared to the Markov chain. With this in mind, TempGAN still seems like a promising method to model the customer journey and predict conversions.

Table 5: Evaluation metrics for all models on the full data set

	Full Data		
	Naive	Markov Chain	TempGAN
Brier Score	0.0487	0.0480	0.0467
AUC	0.5000	0.4915	0.5431
F1	0.0	0.0596	0.0834

5.2 Data complexity

To understand how the models compare to each other when dealing with data of varying complexity, their performance is also compared using a data set containing only paths with a length higher than 5 and 10. The assumption is that as the path length increases, the complexity of the data also increases.

The data set with only the most complex paths is the one which contains paths longer than or equal to 10. When analyzing the performance on this data set, it is possible to see that both models produce very similar results. The only area TempGAN is able to achieve a significant improvement is with the Brier score. A possible explanation for TempGAN not being able to outperform the Markov chain is that the number of training instances decreases significantly. Neural networks have been found to perform better when more data is available. When changing the data set to include only paths with length grater than 10, the number of total consumers is decreased to only 6, 716. So while the complexity of the data increases, the number of training instances decreases. Since the performance of TempGAN on this data set can be traced to the trade-off between data complexity and number of training instances, it is not possible to conclude that TempGAN underperforms with data of higher complexity. It simply highlights the dependence of such algorithms on the availability of training data.

In light of the data availability constraint for paths with a length higher than 10, the data set with paths of length higher than or equal to 5 seems to achieve a better balance in the previously mentioned trade-off. This data set contains 26, 805 consumers and a minimum path length of 5 is relatively complex. In Table 6 it is possible to see that TempGAN outperforms the Markov chain in all metrics. While it does not produce the lowest Brier score, it still provides a significantly lower value as compared to the Markov chain. Furthermore, TempGAN produces the highest AUC and F1 scores out of all models. This provides an indication that TempGAN might provide better results in more complex data sets given that it contains enough data instances to train on. This is of great value to a real world application where most companies have a data set fulfilling these characteristics.

Table 6: Evaluation metrics for all models on the limited data set

	Length ≥ 5			Length ≥ 10		
	Naive	Markov Chain	TempGAN	Naive	Markov Chain	TempGAN
Brier Score	0.0471	0.06256	0.0478	0.0452	0.0921	0.0558
AUC	0.5000	0.4985	0.5265	0.5000	0.5000	0.4967
F1	0.0	0.0900	0.0936	0.0	0.0865	0.0863

5.3 Attribution Scores

In this section, the attribution scores produced using the Shapley value and TempGAN are analyzed.

The first noticeable difference between the attribution scores seen in Tables 7 and 8 is that the dimension of the values are different. This is due to the fact that while the Shapley values count the number of conversions, the TempGAN scores are given as attention weights. Nonetheless, it is possible to compare the rankings of the channels.

Table 7 shows the attribution scores produced based on the full data set. Here, both methods show that Facebook is the most important channel in leading consumers to conversions. Dehghani and Tumer (2015) conducted an experiment that showed that in 2013, Facebook advertisement produced significant impacts to brand image and consequently conversions. This supports the strong attribution score associated to Facebook by both methods. Another similarity of both methods is that they assign negative scores to Online Video, Online Display, and Paid Search. This would imply that these channels are decreasing the likelihood of a consumer converting and their presence in the customer journey should be significantly reduced and maybe even completely removed. With the lack of visibility on the type of business this data refers to, it is difficult to establish if this conclusion would be reasonable or not. For example, Schlangenotto, Kundisch and Wünderlich (2018) found that Paid Search is generally not effective in increasing conversions for brick-and-mortar stores while Dinner, Heerde Van and Neslin (2014) found that for high-end clothing and apparel retailers it is effective. Nonetheless, these examples show that the negative attribution scores produced by both approaches could still be reasonable. Finally, there seems to be a divergence regarding the importance of Instagram. While the Shapley value approach classifies it as a significant driver in consumer conversion, TempGAN classifies it as the opposite. Again, it is hard to establish the more appropriate score without knowledge of the business, however, it seems more likely that if Facebook provides a positive contribution, Instagram would follow, given their similarities. This would be better reflected in the Shapley value attribution score. This is not to say that TempGAN's attribution score is not realistic. For example, Čuić Tanković, Perišić Prodan and Tomljanović (2022) noted that for small hospitality businesses, Instagram was more effective in converting younger audiences while Facebook performed better for a middle-aged audience. Hence, if this business is targeted towards older people, it could be reasonable to have a pattern as the one described by TempGAN.

Table 7: Attribution scores generated by each method on the full data set

	Full Data			
	Shapley Value	Rank	TempGAN	Rank
Facebook	3,517.67	1	0.363	1
Instagram	1,671.50	2	-2.057	3
Online Video	-1,056.17	4	-2.246	4
Online Display	-3,488.50	5	-1.300	2
Paid Search	-436.50	3	-2.961	5

It is also possible to compare the attribution scores produced on the different data sets. Table 8 shows the attribution scores produced on the data sets including only paths with length greater than or equal to 5 and 10. Similar to the attribution values produced on the full data set, Online Video and Online Display still receive a negative score from all approaches. This shows that even with consumers that take more time to come to a decision, these channels do not seem to contribute to a conversion. In general, the attribution scores generated by the Shapley value produce similar results across all three data sets. Namely, Facebook is the highest contributor while Instagram is the second highest and all the rest have negative contributions. This makes sense since the Shapley value is largely determined by the count of conversions by each coalition. Hence, filtering some entries should not affect the proportion of conversions between the coalitions. This is supported by the statistics shown in Tables 3 and 4 as the channel distribution for conversion cases across all three data sets is very similar. One noticeable difference can be seen when TempGAN is applied on the data set with paths of length greater than 5. Here, Instagram and Paid Search actually receive a positive score while all the others, including Facebook, receive a negative one. Also, TempGAN attributes the most value to Paid Search which, in all other cases, is the channel that receives one of the lowest scores.

Table 8: Attribution scores generated by each method on the limited data set

	Length ≥ 5				Length ≥ 10			
	Shapley Value	Rank	TempGAN	Rank	Shapley Value	Rank	TempGAN	Rank
Facebook	3,506.23	1	-0.223	4	2,477.42	1	0.162	1
Instagram	2,339.07	2	0.340	2	1,910.58	2	-0.887	3
Online Video	-870.02	3	-0.199	3	-568.75	3	-0.946	4
Online Display	-3,501.10	5	-0.316	5	-2,587.08	5	-0.855	2
Paid Search	-1,226.18	4	0.812	1	-1,009.17	4	-1.065	5

5.4 Economic Interpretation

Neural networks have often been labeled as a black box, meaning that it is difficult to interpret what happens in the processing steps. This is an advantage that the Markov model possesses over TempGAN as its output provides significantly better economic interpretations. In this section, the insights provided by the Markov chain are explored further.

Table 9 shows the eventual conversion probabilities for each channel. Here it is possible to see that paths starting with an Online Video are the most likely to convert when considering all consumers. For more indecisive consumers, the best starting channels are the Online Display and Instagram. This is somewhat surprising as it is possible to see that in two cases, a channel that was associated to a negative attribution score has the highest eventual conversion probability. Namely, Online Video received an attribution score of -1,056 and -2.246 in the full data but has the highest eventual conversion probability according to the Markov model. The same happens for Online Display on the data set containing only paths with a minimum length of 5. Even though these conversion probabilities do not account for the intricate interactions between the channels, it would be expected that these would at least receive a positive attribution score.

Table 9: Eventual conversion probabilities estimated on each data set

	Full Data	Length ≥ 5	Length ≥ 10
Facebook	0.0930	0.1892	0.2710
Instagram	0.0930	0.1635	0.2723
Online Video	0.1090	0.1888	0.2518
Paid Search	0.0700	0.1527	0.2707
Online Display	0.0720	0.1928	0.2387

Another interesting output from the Markov model that can be analyzed is the transition matrix. Table 10 shows the transition matrix estimated on the full data set and the cells highlighted in yellow show the highest probability between non-absorbing states for each row. The first thing that stands out is the fact that most channels have the highest probability to transition back to themselves. This is true for Facebook, Online Video, Online Display, and Paid Search. For the last channel, this could potentially highlight an inefficiency. Usually, advertisers have to pay-per-click in Paid Searches and if this channel is consistently drawing consumers back to it without increasing the likelihood of conversion, the result is an increase in cost without an increase in revenue. Note that Paid Search generally receives one of the lowest attribution scores by both methods and hence, it is very likely that it indeed is not the most influential channel in leading consumers to conversion. Furthermore, Table 10 shows that consumers are likely to have their first interaction also with a Paid Search. However, Table 9 shows that paths beginning with this channel receive the lowest eventual conversion probability. Both of these findings once again highlight the inefficiency of using a Paid Search.

Table 10: Transition matrix for the Markov model estimated on the full data set

	Start	Facebook	Instagram	Online Video	Online Display	Paid Search	Conversion	Quit
Start	0	0.247	0.105	0.144	0.143	0.361	0	0
Facebook	0	0.320	0.134	0.020	0.016	0.035	0.045	0.430
Instagram	0	0.317	0.136	0.021	0.016	0.034	0.045	0.432
Online Video	0	0.020	0.008	0.586	0.008	0.017	0.041	0.320
Online Display	0	0.030	0.013	0.014	0.306	0.071	0.039	0.526
Paid Search	0	0.032	0.014	0.017	0.029	0.382	0.035	0.490
Conversion	0	0	0	0	0	0	1	
Quit	0	0	0	0	0	0	0	1

Tables 11 and 12 show the transition matrices estimates on the data set with paths of length 5 and 10, respectively. From Table 12 it is possible to see the same pattern as the transition matrix of the full data set. However, for consumers with path of length higher than or equal to 5, there is a change. Namely, Table 11 shows that the most likely transition for Instagram is not Facebook anymore, but rather back to itself. Also, now consumers interacting with the Online Video are more likely to transition to Facebook and the most likely starting channel is Facebook instead of Paid Search. This pattern is in line with the attribution scores found previously as Facebook has the highest attribution score and is the channel that consumers coming from other channels are most likely to transition to.

Table 11: Transition matrix for the Markov model estimated on the data set containing paths of length higher than or equal to 5

	Start	Facebook	Instagram	Online Video	Online Display	Paid Search	Conversion	Quit
Start	0	0.282	0.213	0.122	0.271	0.112	0	0
Facebook	0	0.551	0.041	0.232	0.034	0.023	0.024	0.095
Instagram	0	0.081	0.641	0.035	0.049	0.077	0.015	0.101
Online Video	0	0.549	0.042	0.233	0.035	0.024	0.024	0.093
Online Display	0	0.029	0.021	0.012	0.820	0.010	0.022	0.086
Paid Search	0	0.076	0.139	0.031	0.036	0.554	0.018	0.146
Conversion	0	0	0	0	0	0	1	0
Quit	0	0	0	0	0	0	0	1

Table 12: Transition matrix for the Makrov model estimated on the data set containing paths of lenght higher than or equal to 10

	Start	Facebook	Instagram	Online Video	Online Display	Paid Search	Conversion	Quit
Start	0	0.314	0.141	0.135	0.074	0.337	0	0
Facebook	0	0.612	0.256	0.025	0.014	0.029	0.018	0.046
Instagram	0	0.608	0.257	0.027	0.016	0.030	0.019	0.043
Online Video	0	0.091	0.038	0.687	0.065	0.058	0.013	0.049
Online Display	0	0.100	0.041	0.142	0.564	0.060	0.014	0.079
Paid Search	0	0.027	0.011	0.016	0.009	0.877	0.017	0.043
Conversion	0	0	0	0	0	0	1	0
Quit	0	0	0	0	0	0	0	1

6 Conclusion and Discussion

6.1 Conclusion

With the increasing abundance of data, new techniques have been developed to study the multi-touch attribution problem and the customer journey. Two methods that have been thoroughly researched in this application are the Shapley value and the Markov chain. While intuitive and founded on strong theoretical concepts, these models can sometimes be difficult to implement. This is the case as they rely on estimating and computing several permutations of the data and when these get too high, it might become computationally intractable. This is specially relevant to MTA since, realistically, the number of channels can quickly grow to such a number. Bearing this in mind, it is necessary to develop a model that does not posses this limitation while, at the same time, provides accurate conversion predictions and attribution scores.

This research tackled this problem by implementing a temporal graph attention neural network. The customer journey was modeled as a directional temporal graph with the nodes being the channels and the edges, the time-stamped interactions of the consumer. These graphs where fed into the neural network which provided conversion predictions. The attention scores learned by the network served as the attribution score for each channel.

This architecture was implemented on a data set and the performance of the models was compared. TempGAN showed improvements by producing a lower Brier score and higher AUC and F1 scores. However, both models did present a relatively low AUC and F1 scores, showing that while the models are well calibrated, they cannot properly distinguish between positive and negative instances. Furthermore, when applying TempGAN and the Markov chain to a filtered data set with more complex journeys, TempGAN again was able to outperform. However, as the number of training instances decreased, the performance of TempGAN also decreased, highlighting the dependency of this method on the availability of training data. Overall, TempGAN showed to be an interesting model for further investigation.

The attribution scores produced by the Shapley value and TempGAN were mostly similar. Namely, Online Video, Online Display, and Paid Search all received negative scores, showing that they were actually counterproductive in leading the consumer to conversion. Both methods also agreed on the importance of Facebook in driving conversion and classified it with the highest score. The only point of disagreement was regarding Instagram where the Shapley value assigned a positive score while TempGAN assigned a negative one. Without the knowledge of the business to which this data refers to, it is unclear which value seems more appropriate.

6.2 Limitations and Further Research

Despite the results presented in this research, there are some limitations. The first one regards the data. Customer journey data is highly protected by companies and only a limited option can be found online for free. Because of this, the training data was not complete and prevented a couple of analysis. Examples include the validation of the attribution scores with business logic and the understanding of the models' performance on data of varying complexity. Another limitation is regarding computational power. Neural networks demand a significant amount of computational power to train and with the instruments at hand it was not possible to run hyperparameter optimisation to determine the best number of features and epochs to use in TempGAN. This could have increased the performance of the model. Finally, a limitation of TempGAN is its interpretability. As an attribution model, it would be interesting to have more interpretable parameters, however, this is very limited for neural networks.

Given that conversion data is also imbalanced in real life and that it greatly hindered the performance of TempGAN, further research could focus on techniques to make the algorithm robust to such a limitation. One way this could be done is by using resampling techniques that either oversample the conversion instances or undersample the non-conversion instances. A drawback of this method is that TempGAN's performance will then also rely on the random sampling technique applied. To prevent this, class weighting could be used instead. In this technique, the minority class receives a higher weight when computing its loss during the training phase. This ensures that the model cannot achieve a low score by simply predicting low probabilities and hence should improve its discriminatory power.

A final extension to this research would be to add some interpretability to TempGAN. This could be done by providing a non-random initial feature matrix where each channel has features with clear meanings. By clearly defining the interpretation of each of the k features, it would then be possible to interpret the learned embedding matrix produced by the algorithm.

References

Alhabash, S., Mundel, J. & Hussain, S. A. (2017). Social media advertising: Unraveling the mystery box. In *Digital advertising* (pp. 285–299). Routledge.

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. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167811616300349 doi: https://doi.org/10.1016/j.ijresmar.2016.03.001

Ansari, A., Mela, C. F. & Neslin, S. A. (2008). Customer channel migration. *Journal of marketing research*, *45*(1), 60–76.

Benes, R. (2019, 22nd February). *Agency pros say fraud is biggest threat to their budgets.* Insider Intelligence. Retrieved 17/06/2023, from https://www.insiderintelligence.com/content/agency-pros-say-fraud-is-biggest-threat-to-their-budgets

Borden, N. H. (1964). The concept of the marketing mix. *Journal of advertising research*, *4*(2), 2–7.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Chierichetti, F., Kumar, R., Raghavan, P. & Sarlos, T. (2012). Are web users really markovian? In *Proceedings of the 21st international conference on world wide web* (p. 609–618). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2187836.2187919 doi: 10.1145/2187836.2187919

Chiu, H.-C., Hsieh, Y.-C., Roan, J., Tseng, K.-J. & Hsieh, J.-K. (2011). The challenge for multichannel services: Cross-channel free-riding behavior. *Electronic Commerce Research and Applications*, *10*(2), 268–277.

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).

Dehghani, M. & Tumer, M. (2015). A research on effectiveness of facebook advertising on enhancing purchase intention of consumers. *Computers in Human Behavior*, *49*, 597-600. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0747563215002411` doi: https://doi.org/10.1016/j.chb.2015.03.051

Dinner, I. M., Heerde Van, H. J. & Neslin, S. A. (2014). Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. *Journal of marketing research*, *51*(5), 527–545.

Faulds, D. J., Mangold, W. G., Raju, P. & Valsalan, S. (2018). The mobile shopping revolution: Redefining the consumer decision process. *Business Horizons*, *61*(2), 323-338. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0007681317301672` doi: https://doi.org/10.1016/j.bushor.2017.11.012

Fornari, E., Fornari, D., Grandi, S., Menegatti, M. & Hofacker, C. F. (2016). Adding store to web: migration and synergy effects in multi-channel retailing. *International Journal of Retail & Distribution Management*, *44*(6), 658–674.

Gordon, B. R., Jerath, K., Katona, Z., Narayanan, S., Shin, J. & Wilbur, K. C. (2021). Inefficiencies in digital advertising markets. *Journal of Marketing*, *85*(1), 7–25.

Herhausen, D., Kleinlercher, K., Verhoef, P. C., Emrich, O. & Rudolph, T. (2019). Loyalty formation for different customer journey segments. *Journal of Retailing*, *95*(3), 9–29.

Hogan, S. S., Bruderle, C., Silverman, D. & Krasnow, S. (2020, May). *Internet advertising revenue report* (Tech. Rep.). New York: Interactive Advertising Bueareu (IAB).

Hu, T.-I. & Tracogna, A. (2020). Multichannel customer journeys and their determinants: Evidence from motor insurance. *Journal of Retailing and Consumer Services*, *54*, 102022. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0969698919309087` doi: https://doi.org/10.1016/j.jretconser.2019.102022

Krajnović, A., Sikirić, D. & Bosna, J. (2018). Digital marketing and behavioral economics. *CroDiM: International Journal of Marketing Science*, *1*(1), 33–46.

Li, H. A. & 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. Retrieved from `https://doi.org/10.1509/jmr.13.0050` doi: 10.1509/jmr.13.0050

Martin, N. (2019). *How much data is collected every minute of the day.* Retrieved from `https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/` (Accessed on 05/06/2023)

Mohan, A. & Pramod, K. (2021). Temporal network embedding using graph attention network. *Complex & Intelligent Systems*, 1–15.

Schlangenotto, D., Kundisch, D. & Wünderlich, N. V. (2018). Is paid search overrated? when bricks-and-mortar-only retailers should not use paid search. *Electronic Markets*, *28*. doi: https://doi.org/10.1007/s12525-018-0287-4

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).

Singal, R., Besbes, O., Desir, A., Goyal, V. & Iyengar, G. (2022). Shapley meets uniform: An axiomatic framework for attribution in online advertising. *Management Science*, *68*(10), 7457-7479. Retrieved from `https://doi.org/10.1287/mnsc.2021.4263` doi: 10.1287/mnsc.2021.4263

The World Bank. (2021). *International telecommunication union (itu) world telecommunication/ict indicators database.* The World Bank. Retrieved from `https://data.worldbank.org/indicator/IT.NET.USER.ZS`

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. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0148296320308584` doi: https://doi.org/10.1016/j.jbusres.2020.12.028

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. et al. (2017). Graph attention networks. *stat*, *1050*(20), 10–48550.

Zhang, S., Tong, H., Xu, J. & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, *6*(1), 1–23.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57-81. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2666651021000012` doi: https://doi.org/10.1016/j.aiopen.2021.01.001

Čuić Tanković, A., Perišić Prodan, M. & Tomljanović, D. (2022, Nov.). Differences between instagram and facebook sponsored posts for small hospitality businesses. *ENTRENOVA - ENTerprise REsearch InNOVAtion*, *8*(1), 287–298. Retrieved from `https://hrcak.srce.hr/ojs/index.php/entrenova/article/view/23862` doi: 10.54820/entrenova-2022-0025

A Code Description

The zip file containing the code used in this research contains 6 python notebooks:

- Data cleaning Markov
- Markov chain
- Shapley value
- TempGAN
- Analysis - model performance
- Analysis - attribution scores

The code in the "Data cleaning Markov" notebook, converts the raw data obtained from the Kaggle project into the format that is needed for the Markov chain estimation. The main change applied is the grouping of the entries of the raw data by consumer. This implies that while multiple lines could be associated to a single consumer id in the raw data, only one line is present for each consumer id after processing the data. Furthermore the paths of each consumer are given by the channel name followed by the string ">". An example of the path after processing the data is "Facebook > Instagram > Paid Search". The resulting dataframe is exported to a csv file.

In the "Markov chain" notebook, the package Channel Attribution is used to estimate the transition matrix. This algorithm takes the csv file produced by the "Data cleaning Markov" notebook and first filters the data according to the three data subsets used in the research. Then, the data is split into training and evaluation sets. A Markov chain is estimated on the training set and the procedure described in Section 4.1 is applied. Some additional methods are used to format the data frame into an acceptable format.

In the "Shapley value" notebook, the csv file generated by the Data cleaning Markov notebook is again used. Once again the data is filtered according to the subset being studied. A function called calculate shapley is created. This function first determines the powerset of the channels in the data. Then, the function counts the number of conversions associated to each element in the powerset. This corresponds to the characteristic function of the Shapley value. Finally, the formulas in Equation 2 and 3 are used to calcualte the Shapley value.

In the "TempGAN" notebook, the raw data extracted from the Kaggle project is used. First, the NetworkX package is used to create a multi-directional graph for each consumer. Then, the data is filtered according to the subset being studied. The function "adjacency matrix" is used to determine the adjacency matrix for each consumer. I then create a function called "count unique time respecting paths" which uses the "all simple paths" methods from the NetworkX package to count the number of unique time respecting paths between two distinct nodes. A method called "count unique time respecting paths itself" is also created to count the number of time respecting paths from a node to itself. The function "count self loops" counts the number of loops from a channel to itself. All of these three functions are used in the "co occurrence" method to determine the co-occurrence matrix. The output of this method is then used together with the "PPMI" method to determine the PPMI matrix for each consumer. The necessary inputs to TempGAN are converted to tensors since the neural network architecture was developed using pytorch. The architecure was coded in the same way as seen in 1. Finally, the data was ordered according to the last interaction of each consumer and the train-test split was made. For every instance of the training set, the gradient was set to zero and back propagation was performed. The results were then saved to a data frame which was exported to a csv file.

In the notebook called "Analysis - model performance", all the results obtained from the Markov chain and TempGAN were loaded. Then, because the data was saved as tensors for TempGAN, some strings needed to be removed from the input data. Namely, the input csv file had values in the form of "tensor([[0.5]])" which needed to be converted to the format "0.5". The F1 and AUC scores were computed using existing packages from sklearn. The brier score was calculated using a function that I created.

In the notebook called ”Analysis - attribution scores” the attribution scores of each channel were printed.