A New Approach to Generate Individual Level Data of Walled Garden Platforms: Linear Programming Reconstruction

Yifei Pang
CMU
yifeip@andrew.cmu.edu

Sreenidhi Ganachari
CMU
sganacha@andrew.cmu.edu

Yuan Yuan
Adobe
yuyuan@adobe.com

Steven Wu
CMU
zstevenwu@cmu.edu

Xiaojing Dong
SCU
xdong1@scu.edu

Jin Xu
Adobe
auxu@adobe.com

Zhenyu Yan
Adobe
wyan@adobe.com

Abstract

Understanding customer journeys through their interactions with marketing touchpoints is critical for user's behavior modeling and advertiser's digital marketing strategy optimization, but privacy restrictions from Walled Garden platforms[1] create barriers and challenges to such practices. Although the ideal data an advertiser requires for marketing data analysis would be detailed individual level activity data and ads interaction behaviors, the data an advertiser can obtain from a Walled Garden platform are aggregate level statistics. To overcome this gap, this paper introduces a novel approach, Linear Programming Reconstruction, which primarily relies on the linear programming algorithm that leverages aggregated Walled Garden platform statistics, while integrating detailed event level data from other marketing channels to reconstruct the complete individual level interaction journeys. We detail the proposed approach and provide an initial demonstration of its effectiveness in reconstruction by applying it in building an attribution model through experiments. The approach provides a valuable solution for overcoming data limitations from Walled Garden platforms, making it possible for deeper user behavior modeling and improved marketing strategies in the context of evolving privacy regulations.

1 Introduction

In the rapidly evolving landscape of digital marketing, building customers' journeys through various touchpoints is crucial for analyzing users' behavior and optimizing marketing strategies. Traditional marketing channels such as Paid Search and Email provide detailed individual level data, enabling marketers to construct a comprehensive picture of each user's path to conversion. This detailed visibility supports a range of machine learning models, such as attribution models [1, 2, 3, 4, 5], which rely on complete user paths to measure and optimize marketing performance. However, a significant challenge arises with the Walled Garden platforms like Meta and YouTube due to stringent privacy policies. For example, the EU's General Data Protection Regulation [6, 7] is one such regulation that safeguards sharing of individual data. These policies prohibit companies from directly releasing personal data for potential marketing usage.

[1]Walled Garden platforms refers to the platforms which limit the access to user level data, citing privacy policies, such as Google and Meta.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

To address the challenges of limited data access, researchers in marketing have proposed various approaches for marketing analysis, especially the contribution of different channels on their conversions, based on accessible data. One well-known approach is Match Market Test [8], which selects two non-overlapping geographic regions as the control and test group to measure conversion changes resulting from different advertising. Additionally, some companies also use Last Touch Attribution (LTA) to assign channel contributions with limited data access. However, these methods offer a narrow perspective on channel contributions and fail to account for synergistic effect across channels. Moreover, they fail to make full use of Walled Garden data that are accessible.

In this study, we propose a novel approach: Linear Programming Reconstruction. This method leverages the linear programming algorithm to reconstruct individual level data that aligns with the published Walled Garden statistics, drawing upon insights from prior research on reconstruction attack [9, 10, 11, 12, 13]. The proposed approach expands beyond the linear programming reconstruction by designing post-processing and shuffling techniques to further refine the results, making them more accurate and stable. By integrating the often isolated aggregate statistics from Walled Garden platforms with available event level data from other marketing channels, we can reconstruct complete user paths. Experiments using simulated data suggest that reconstructed data reasonably approximates the hidden data for a selection of attribution models, indicating its potential as a powerful tool for marketing analysis even with privacy constraints and possibly providing deeper insights into customer behavior.

2 Problem Formulation

The core challenge faced in this research is the inability to achieve complete user journeys due to the restricted access to individual level data from Walled Garden platforms. This lack of detailed data makes it challenging to build user level model, which is often the foundation for any marketing and targeting decisions.

2.1 Walled Garden Data

Individual level data has been treasured by marketing practitioners for its potential in conducting personalization and improving marketing metrics and decision making process. This leads to the popularity of the Walled Garden data. Advertisers purchase ads from platforms, such as Google or Meta, hoping to mine the detailed individual level data to track users interactions and responses to the ads. Due to the privacy regulations, getting those data became impossible. Only aggregate level statistics can be obtained from these platforms, which summarize user interactions and conversions at a macro level without noise. These data are commonly referred to as the Walled Garden data.

The aggregated statistics serves as a critical source of information for commercial analysis and marketing strategy optimization, even though it lacks the granularity needed for user level analysis. Table 1 provides an illustrative example of Walled Garden data, where only some summary statistics (shown in columns) at certain level of time aggregation, such as in days, (shown in columns) are provided. In this example, column 2 indicates the total number of conversions on a daily bases and column 3 lists the total number of touchpoints on any given date (column 1). The other columns list the lagged values of each metric, up to lagged 30 days in this example. The "Lag$i$ Count" in this context represents the number of touchpoints that occurred within a specified time window $(i-1, i]$ days before the conversion date (column 1). These statistics offer a snapshot of overall touchpoints but do not reveal individual user's information.

Without individual level information, these data pose a challenge for applying machine learning models that rely on detailed user paths. To overcome this limitation, we develop the approach to reconstruct a complete user level paths using Walled Garden data, which will be described in section 3. Although reconstructed data may not be identical to the hidden real data, it allows us to obtain a reconstruction result that is both close to the real data and has practical utility.

2.2 Attribution model

In digital marketing, attribution models are essential for analyzing user interactions across various channels and understanding each channel's contribution to conversions. Accurately assigning credit to each channel in a customer's journey allows marketers to fine-tune their strategies and ultimately

Table 1: Walled Garden data schematic diagram

Date	# Conversion	# Touchpoints	Lag1 count	Lag2 count	...	Lag30 count
2018-06-30	470	91	10	8	...	1
2018-06-29	361	122	15	20	...	0
...	...	...	...	...	...	...

maximize their return on investment (ROI) [1, 2, 3, 4, 5]. However, a valid attribution model builds on individual level path data, that become inaccessible due to privacy regulations.

Advertisers still have their first party data that records detailed user level behaviors and interactions with marketing activities after a user landed on the advertiser's tractable sources, such as the website or the mobile app. From the advertiser's perspective, the user's interactions with third party platform are recorded only at aggregate level, while their purchase actions are recorded at the individual level. It is therefore essential to reconstruct the individual level data from Walled Garden data, before combining them to obtain attribution model results.

To illustrate our approach's capability in performing such a task and to evaluate its performance, we adopt a popular attribution model, the Markov model [4, 5, 14]. This model is built based on the assumption that each user's clicking behaviors across channels are Markovian. It then evaluates each channel's contribution by examining its removal effect on conversions based on the transition probabilities between channels. This model is also publicly available[2], making it an accessible benchmark for evaluating the proposed approach.

3 Approach

In this section, we describe the details of our approach designed to solve the challenges due to the data limitations of Walled Garden platforms. Our approach consists of three steps and the workflow is depicted in Figure 1. The detailed description of the algorithm are outlined in Algorithm 1 in the appendix.

In this algorithm, the first step solves a constrained LP problem and obtains the preliminary results that consists of a set of reconstructed data. The second step conducts post-processing, which refines the step 1 solution to better approximate the hidden data. Finally, the third step involves shuffling, which improves performance stability by generating multiple groups of reconstructed data. This proposed approach extends the LP reconstruction approach by adding two additional steps, including post-processing and shuffling. These two steps allow us to ensure the reconstructed data can satisfy the known knowledge of user behaviors in the context of marketing activities and also enhance the stability of the reconstructed data.

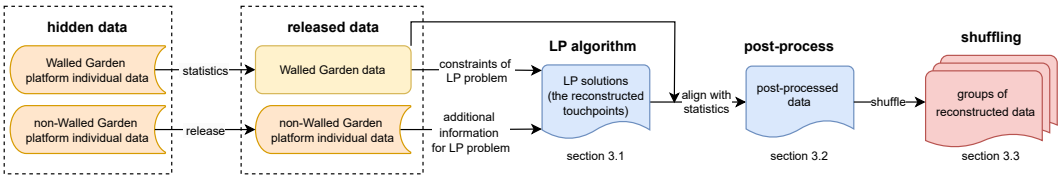

Figure 1: The workflow diagram of Linear Programming Reconstruction

3.1 Linear Programming

Linear programming (LP) is a mathematical method used to optimize linear objective functions, subject to a set of linear equality or inequality constraints [15, 16]. It's also commonly applied as a reconstruction method in the field of privacy [9, 10, 11, 12, 13]. In step 1 of Algorithm 1 in the appendix, an LP is set up and the objective function is to minimize the statistical errors between the

[2]Available at: `https://channelattribution.io/docs/chpro/functions/markov-model`

reconstructed data and the hidden data (line 3). Specifically, we want to minimize the differences d between the statistics of the reconstructed data X and the given statistics D (see Equation 2, 3 in the appendix).

The results from LP tend to be overly simplistic, ignoring the complex features of human behaviors. For example, the LP results can lead to the situation where a really small portion of the users (less than 1% for non-converted users and less than 10% for converted users) are assigned with most of (about 90%) the touchpoints while the majority get none. This may be reasonable in other context, but are not aligned with what marketing data are typically observed. To address this issue, we introduce additional constraints to the algorithm (line 4) to reflect such additional constraints that are not reflected in the LP design. For example, based on common human behaviors, we limit the number of clicks each user can make from a certain channel on a particular day to x_{max} (line 5). This limit can be set according to the specific context, by adding domain expert's knowledge. In our context the x_{max} value is defined based on the combination of the scale of daily clicks of each channel and extensive experience of our domain experts in Marketing.

To better distribute the touchpoints and reduce their concentration, we introduce another metric, path touch percentage P_C, which calculates the percentage of users having touchpoints on the channel C.[3] This constraint ensures that the reconstructed data aligns closely with P_C, with a small error bound[4] (line 6). This helps distribute the touchpoints among nearly the same number of users, largely approximating the original distribution. Using the aforementioned constraints and objective function, we obtain a solution X to the LP problem (line 7). Last in this step, we apply probabilistic rounding in Equation 1, to the fractional touchpoint solutions X, converting the click counts to integers, yielding a preliminary result (line 8).

$$\text{Round}(x) = \begin{cases} \lfloor x \rfloor, & \text{with probability } 1 - (x - \lfloor x \rfloor) \\ \lceil x \rceil, & \text{with probability } x - \lfloor x \rfloor \end{cases} \tag{1}$$

3.2 Post-processing

Due to the deviations from the linear programming solutions and errors introduced by probabilistic rounding, the preliminary results may differ from the hidden data in terms of statistics, making it necessary to further adjust the reconstructed data. We address these discrepancies through post-processing, either by reducing excesses or compensating for shortfalls. Specifically, in Step 2 of Algorithm 1 in the appendix, we calculate the difference d for each statistical entry (line 2). If there are excess touchpoints, we randomly remove them from the corresponding users to mitigate the surplus (lines 3-4). Conversely, if there are fewer touchpoints, we add them to randomly selected users to compensate for the deficit (lines 5-6). Additionally, it is crucial to keep N_C close to nN_P throughout this process to maintain the touchpoints distribution (line 8). This step ensures that the reconstructed data approximates the hidden data in statistical terms.

3.3 Shuffling

Although the statistics above help guide the reconstruction process, it is nearly impossible to derive a single definitive dataset, so relying on a single reconstructed dataset can introduce significant biases. Therefore, to achieve more stable and reliable reconstruction results for evaluation, we apply shuffling to generate multiple datasets with the same statistical properties. Specifically, in Step 3 of Algorithm 1 in the appendix, we first group the data by users within the same statistical groups, specifically those who converted on the same date (line 2). Then, within each group X_d, we randomly shuffle the touchpoints data within each channel, meaning that the click data of all users for that channel is reordered (lines 3-5). This approach preserves the statistics in Equation 3. Each run of the shuffling process yields a different X^* due to the randomization (line 7), enabling the generation of numerous variations of reconstructed data, thereby improving statistical stability during evaluation. While this method is demonstrated in the current context, it serves as a heuristic approach that can be generalized to similar problems with different group configurations.

[3]These statistics can be provided through a company-authorized clean room, which is increasingly available from Walled Garden platforms, such as Meta.

[4]The error bound can be tuned, smaller value may lead to longer processing time. Empirically, values in 0.001~0.01 could be suitable.

4 Experiments

To evaluate the effectiveness of our reconstruction approach, we conduct experiments by applying the reconstructed data to a marketing attribution model to benchmark its performance relative to the ground truth data, which is generated as the hidden data for the purpose of the experiment. The ground truth data provides the detailed information of each user's path, in an ideal format for an advertiser who may be interested in building an individual level attribution model for marketing analytics. These data are then aggregated into the format that is likely obtained from a Walled Garden platform. The proposed approach is applied to these aggregated statistics to execute the reconstruction process. Although the proposed approach allows us to obtain the reconstructed individual level data, getting these data is not the end purpose. To evaluate its value to an advertiser, these reconstructed data are then applied to a commonly used model, attribution model. Its performance is then evaluated against the results when applying the ground truth data to the same model. For this purpose, the Markov model discussed in section 2.2 is adopted, as it is one of the most popular approach currently adopted in many attribution practice and the model is also readily available.

4.1 Ground truth data

Since real-world data cannot be disclosed due to privacy policies, we generate synthetic data based on key statistics derived from the real data. Specifically, we use three statistics: the average number of channels touched per user, path touch percentages, and the average number of touchpoints per channel for both converted and non-converted users. To create this synthetic dataset, we first use a Poisson distribution to determine the number of channels touched per user. We then select specific channels based on normalized path touch percentages as probabilities. The number of touchpoints per channel is generated using a Poisson distribution, and conversion and touchpoint dates are randomly assigned within a specified range. This synthetic dataset maintains the core statistical characteristics of the real data and serves as ground truth for evaluating our reconstruction approach.

With the ground truth data ready, we calculate similar Walled Garden statistics from these ground truth data. The proposed approach is applied to these generated statistics to reconstruct the individual level data (see Figure 1).

4.2 Evaluation results

We reconstructed multiple groups of data using the released aggregated statistics through the three steps outlined in Section 3. To evaluate the performance of these reconstructed groups, we applied them on the first-order Markov attribution model and compared the channel contribution results between the two datasets [14]. Specifically, we randomly selected 100 groups of reconstructed data for each ground truth dataset and computed the average contributions for each channel. Our evaluation focuses on "META" and "YouTube (YT)" as Walled Garden channels, and "EMAIL", "DISPLAY", and "Paid Search (PS)" as non-Walled Garden channels. Figure 2 illustrates a representative experiment, demonstrating only minor differences in channel contributions between the ground truth data and the reconstructed data. Additionally, Figure 3 presents the KDE distribution of relative errors after each step across 100 experiments, each based on a distinct 3000-user ground truth dataset. As the three steps progress, the reconstruction errors become smaller and more concentrated. Notably, after all three steps, 61.0% of the channels exhibited an absolute relative error margin within 5%, and 93.0% were within a 10% margin, underscoring the effectiveness of the reconstructed data.

5 Conclusion

In this study, we proposed a novel approach, Linear Programming Reconstruction, to reconstruct individual level data from Walled Garden platforms. This method effectively addresses the challenge of limited access to such data, delivering useful and reliable results based on our evaluation. Furthermore, it has the potential to facilitate deeper insights into user behavior, with possible applications in other areas facing similar data limitations. Despite its effectiveness in compensating for inaccessible data, our study has certain limitations, including a limited variety of evaluation models and insufficient testing on large-scale datasets. Future work could explore the performance of this approach under more diverse and industry-oriented statistical conditions, along with evaluations using different models and large-scale data, to further enhance its utility and applicability.

References

[1] Vibhanshu Abhishek, Peter Fader, and Kartik Hosanagar. Media exposure through the funnel: A model of multi-stage attribution. *Available at SSRN 2158421*, 2012.

[2] Anindya Ghose and Vilma Todri-Adamopoulos. Toward a digital attribution model. *MIS quarterly*, 40(4):889–910, 2016.

[3] Tahir M Nisar and Man Yeung. Attribution modeling in digital advertising: An empirical investigation of the impact of digital sales channels. *Journal of Advertising Research*, 58(4):399–413, 2018.

[4] Lukáš Kakalejčík, Jozef Bucko, PA Resende, and Martina Ferencova. Multichannel marketing attribution using markov chains. *Journal of Applied Management and Investments*, 7(1):49–60, 2018.

[5] Kunal Mehta and Ekta Singhal. Marketing channel attribution modelling: Markov chain analysis. *International Journal of Indian Culture and Business Management*, 21(1):63–77, 2020.

[6] General Data Protection Regulation GDPR. General data protection regulation. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016.

[7] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[8] Jon Vaver and Jim Koehler. Measuring ad effectiveness using geo experiments. *Google Inc*, 2011.

[9] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[10] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017.

[11] Aloni Cohen and Kobbi Nissim. Linear program reconstruction in practice. *Journal of Privacy and Confidentiality*, 10(1), Jan. 2020.

[12] Aloni Cohen, Aleksandar Nikolov, Zachary Schutzman, and Jonathan Ullman. The theory of reconstruction attacks. DifferentialPrivacy.org, 10 2020. `https://differentialprivacy.org/reconstruction-theory/`.

[13] Aloni Cohen, Aleksandar Nikolov, Zachary Schutzman, and Jonathan Ullman. Reconstruction attacks in practice. DifferentialPrivacy.org, 10 2020. `https://differentialprivacy.org/diffix-attack/`.

[14] Davide Altomare and David Loris. Channelattribution whitepaper. ChannelAttribution.io, 02 2022. `https://channelattribution.io/docs/chopen/resources/whitepaper`.

[15] George B Dantzig. Linear programming. *Operations research*, 50(1):42–47, 2002.

[16] Rajendra Kunwar and HP Sapkota. An introduction to linear programming problems with some real-life applications. *European Journal of Mathematics and Statistics*, 3(2):21–27, 2022.

A Algorithm Illustration

In the appendix, we present the complete reconstruction algorithm for reference.

Algorithm 1 Linear Programming Reconstruction

Input:

- D: Walled Garden data (with the same format in Table 1)
- Chs: Array of all Walled Garden Channels
- Y: Individual level data indexed by user_id, containing exact click information for non-Walled Garden channels
- n: Total number of users
- P_C: Path touch percentage for each channel C in Chs

Output:

- X^*: Reconstructed data indexed by user_id, containing the number of touchpoints for each channel on each day

Step 1: Linear Programming

X takes the information about user_id and the date of conversion from Y
for each channel C in Chs **do**
    Define LP objective function:

$$\min_X \ \gamma = \max(d[\text{date}, \text{lag}]) \quad \text{(where lag refers to the lag in Table 1)} \tag{2}$$

$$d[\text{date}, \text{lag}] = \left| D[\text{date}, \text{lag}] - \sum (X[\text{Conversion Date} = \text{date}][C][\text{date} - \text{lag}]) \right| \tag{3}$$

    Add constraints:
        For each date d: $0 \leq X[\text{user\_id}][C][d] \leq x_{\max}$
        Count the number of users with touchpoints on C as N_C: $|N_C/n - P_C| \leq$ error_bound
    Solve the LP problem, solution: X
    Use probabilistic rounding (see Equation 1) to convert each $X[\text{user\_id}][C][\text{date}]$ to integer
end for

Step 2: Post-Processing

for each channel C in Chs **do**
    Recalculate $d[\text{date}, \text{lag}]$ and N_C
    if $d[\text{date}, \text{lag}] > 0$ **then**
        Select random $X[\text{Conversion Date} = \text{date}][C][\text{date} - \text{lag}] > 0$ to reduce, s.t. $d[\text{date}, \text{lag}] = 0$
    else if $d[\text{date}, \text{lag}] < 0$ **then**
        Select random $X[\text{Conversion Date} = \text{date}][C][\text{date} - \text{lag}]$ to add touchpoints, s.t. $d[\text{date}, \text{lag}] = 0$
    end if
    Maintain $N_C \approx nP_C$ at the same time
end for

Step 3: Shuffling

for each conversion date d **do**
    $X_d \leftarrow X[\text{Conversion Date} = d]$
    for each channel C in Chs **do**
        Shuffle $X_d[C]$
    end for
end for
Output X_{shuffled} as X^*

B Additional Experimental Results

We also provide additional visualizations of the experimental results referenced in Section 4.2. These figures illustrate the performances of the reconstructed data on the Markov attribution model, complementing the summary provided earlier.

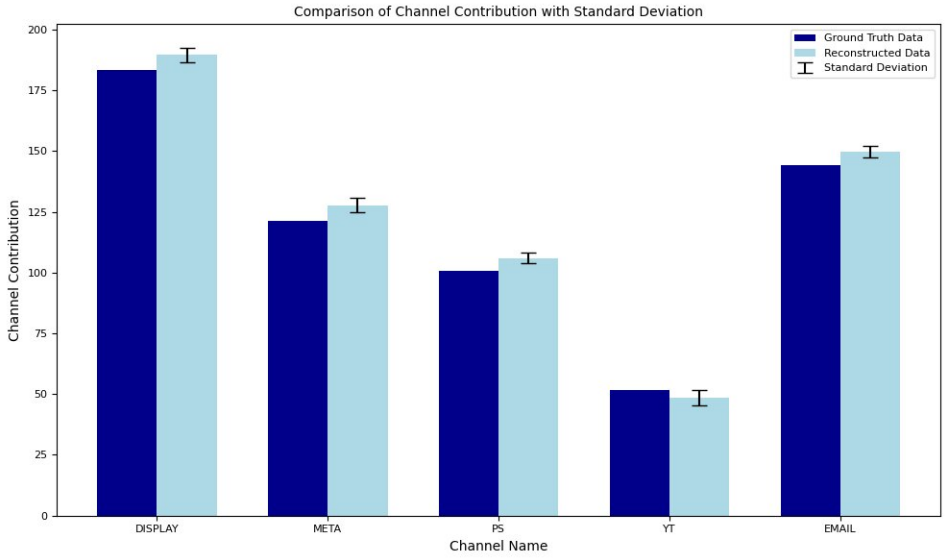

Figure 2: Comparison of channel contribution on the Markov model between the ground truth data and the reconstructed data

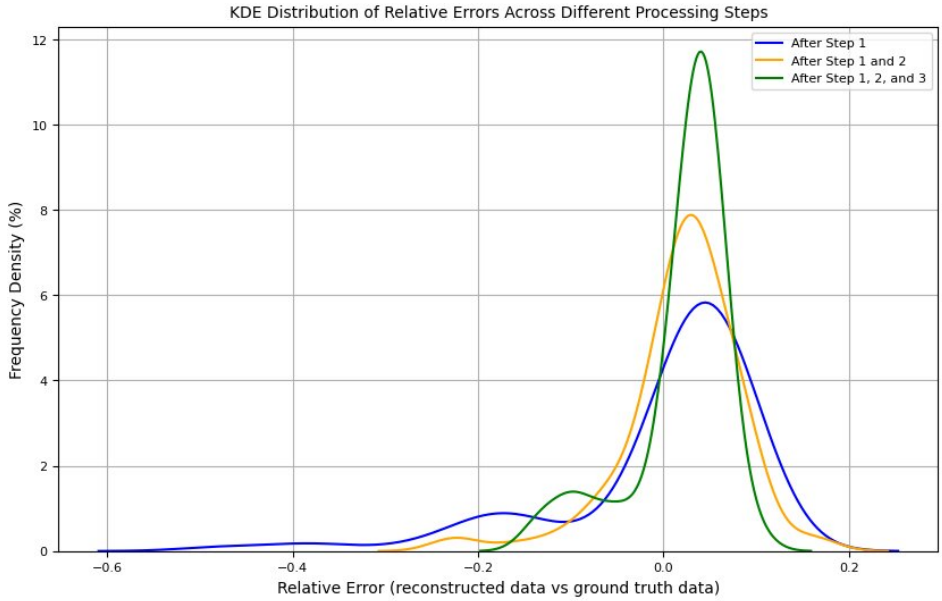

Figure 3: Overall frequency density plot of relative errors in channel contribution between the reconstructed data and the ground truth data on the Markov model