

Singapore Management University

## Institutional Knowledge at Singapore Management University

---

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

---

6-2024

### Network-based representations and dynamic discrete choice models for multiple discrete choice analysis

Huy Hung TRAN

Singapore Management University, hhtran@smu.edu.sg

Tien MAI

Singapore Management University, atmai@smu.edu.sg

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Operations Research, Systems Engineering and Industrial Engineering Commons](#), [OS and Networks Commons](#), and the [Transportation Commons](#)

---

#### Citation

TRAN, Huy Hung and MAI, Tien. Network-based representations and dynamic discrete choice models for multiple discrete choice analysis. (2024). *Transportation Research Part B: Methodological*. 184, 1-22. Available at: [https://ink.library.smu.edu.sg/sis\\_research/8890](https://ink.library.smu.edu.sg/sis_research/8890)

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [cherylds@smu.edu.sg](mailto:cherylds@smu.edu.sg).

# Network-based Representations and Dynamic Discrete Choice Models for Multiple Discrete Choice Analysis

Hung Tran<sup>1</sup> and Tien Mai<sup>2,\*</sup>

<sup>1</sup>*School of Computing and Information Systems, Singapore Management University, 80 Stamford Rd, Singapore, hhtran@smu.edu.sg*

<sup>2</sup>*School of Computing and Information Systems, Singapore Management University, 80 Stamford Rd, Singapore, atmai@smu.edu.sg*

*\* Corresponding author*

June 8, 2023

## Abstract

In many choice modeling applications, people demand is frequently characterized as multiple discrete, which means that people choose multiple items simultaneously. The analysis and prediction of people behavior in multiple discrete choice situations pose several challenges. In this paper, to address this, we propose a random utility maximization (RUM) based model that considers each subset of choice alternatives as a *composite* alternative, where individuals choose a subset according to the RUM framework. While this approach offers a natural and intuitive modeling approach for multiple-choice analysis, the large number of subsets of choices in the formulation makes its estimation and application intractable. To overcome this challenge, we introduce directed acyclic graph (DAG) based representations of choices where each node of the DAG is associated with an elemental alternative and additional information such that the number of selected elemental alternatives. Our innovation is to show that the multi-choice model is equivalent to a *recursive route choice model* on the DAG, leading to the development of new efficient estimation algorithms based on dynamic programming. In addition, the DAG representations enable us to bring some advanced route choice models to capture the correlation between subset choice alternatives. Numerical experiments based on synthetic and real datasets show many advantages of our modeling approach and the proposed estimation algorithms.

**Keywords:** Multiple discrete choice, network-based representation, recursive route choice model.

# 1 Introduction

The demand of consumers is often characterized as multiple discrete, where they choose several items simultaneously. For example, a household may choose a mix of different vehicle types, an individual may participate in multiple daily activities, a tourist may purchase multiple air tickets for their family, or a consumer may purchase multiple items from a store or borrow several books from a library. In the choice modeling literature, researchers have mentioned various ways to handle multiple discrete choice behaviors. The most natural and intuitive way would be to use the traditional random utility maximization (RUM) (McFadden, 1981) single discrete choice models by identifying all combinations of elemental alternatives and treating each combination as a composite alternative. However, this approach is known to be intractable to use, as the number of composite alternatives grows exponentially as the number of elemental alternatives grows. Other modeling approaches include the multiple discrete-continuous models (MDC) (Bhat, 2008, Bhat et al., 2015). Such models have been proposed to address choice situations that involve a subset of elemental alternatives along with continuous intensities of consumption for the chosen elemental alternatives. While such models use continuous values to present the number of choices, they can get rid of the combinatorial structure of the choice model and the issue of having exponentially many choice alternatives. However, they cannot handle multiple discrete choice situations. Extensions have been proposed to handle the discrete nature of people’s choice behavior (Bhat, 2022, Lee et al., 2022), but these models are still expensive to estimate and perform prediction. In fact, existing approaches are typically based on constrained utility maximization, which assumes that a consumer finds an exact optimal solution that maximizes their overall utility while satisfying some complex constraints. This can be a strong assumption since constrained utility maximization is not an easy problem to the choice-makers. Furthermore, it also conflicts with the importance of bounded rationality in decision-making, stated in several prior cognitive psychology and behavioral economics studies (Campitelli and Gobet, 2010, Sent, 2018).

In this paper, we get back to the RUM-based approach, as it would be one of the most natural and intuitive ways to model multiple discrete choice behaviors. Specifically, we focus on addressing the computational challenge that arises when there is an exponential number of subset choice alternatives to be considered. To overcome this challenge, we draw inspiration from the observation that the multiple discrete choice problem shares some synergies with the route choice modeling context, where the number of paths going from an origin to a destination in a network is exponentially large and it is almost infeasible to enumerate all the paths and use them. We, therefore, bring some advanced route choice models and algorithms to address the challenge. Our approach involves the proposition of Directed Acyclic Graphs (DAG) to represent the multiple choice process, where each node can represent an elemental alternative with some additional information, such as the number of selected alternatives. We then show that there is an equivalence between choosing multiple choices and walking through the DAG. This approach allows us to leverage dynamic programming techniques to efficiently estimate the multiple discrete choice model. Further, we also demonstrate that some advanced route

choice models, such as the nested recursive logit (Mai et al., 2015), can be used to capture complex correlation structures between composite choice alternatives. We summarize our main contributions in the following.

**Contributions:** We make the following contributions.

We introduce a RUM-based model for multiple discrete choices where each individual is assumed to consider a set of composite alternatives and make a choice according to a standard RUM model, such as the multinomial logit (MNL) model (Train, 2003). An important aspect of our model is that the choice-maker is subject to a constraint on the range of the size of the composite alternative. Such a constraint is reasonable and simple enough to be considered in the context.

To handle the estimation problem, which involves an exponential number of composite alternatives, we propose two DAG-based representations for the choices. Each node of the DAG represents an elemental alternative and can include some additional information, such as the number of selected elemental choices. By employing the recursive logit (RL) route choice model (Fosgerau et al., 2013), we show that the probability of selecting a composite alternative is equal to the probability of selecting a path in the DAG. This allows us to use results from the route choice modeling literature to show that the choice probabilities of composite alternatives can be computed by solving systems of linear equations, which can be done in poly-time. Consequently, the computation of the log-likelihood can also be done in poly-time. It is noted that the DAG differs from a real transportation network in that it is cycle-free, so the log-likelihood function is always computable regardless of the values of the choice parameters. In contrast, when it comes to route choice modeling, computing the log-likelihood may necessitate higher values for the choice parameters (Mai and Frejinger, 2022).

It can be observed that the composite alternatives may share a complex correlation structure due to inter-overlappings between subsets of alternatives. By bringing route choice models into the context, we further show that the nested RL (Mai et al., 2015), which was developed to capture the correlation between path alternatives in a transportation network, can be conveniently used.

We evaluate our modeling approach and the proposed estimation algorithms with synthetic datasets of different sizes, and with two real-world datasets. The results demonstrate that our methods are highly scalable for handling large-scale settings and provide better in-sample and out-of-sample results compared to other RUM-based baselines. Moreover, our results based on real datasets also show that the nested RL model outperforms the RL model (on the DAGs) in terms of both in-sample and out-of-sample fits.

Our DAG-based representations are highly general and can incorporate other types of constraints and settings, such as customer budgets or specific constraints on the selected items. These constraints can be transformed into additional information that can be incorporated into nodes of the DAG. This may result in a large-sized DAG, but previous work has shown that recursive route choice models can handle networks of up to half a million nodes and arcs (de Moraes Ramos et al., 2020, Mai et al., 2021), in terms of both estimation and prediction. All these imply the

scalability and flexibility of our modeling approach.

**Paper outline:** Section 2 presents a literature review. Section 3 presents a logit-based model for the multiple discrete choice problem. Section 4 presents the DAG-based presentations and their properties. In Sections 5 and 6 we present the RL and nested RL model on the DAGs and show equivalence properties. Section 7 provides numerical experiments and Section 8 concludes. The appendix includes proofs that are not shown in the main part of the paper, and a discussion on an extension to a multiple discrete-count problem, i.e., individuals can select an elemental alternative multiple times.

**Notation:** Boldface characters represent matrices (or vectors), and  $a_i$  denotes the  $i$ -th element of vector  $\mathbf{a}$ , if  $\mathbf{a}$  is indexable. For any integer number  $m > 0$ , let  $[m]$  denote the set  $\{1, 2, \dots, m\}$ .

## 2 Literature Review

Numerous demand models have been developed to analyze people’s multiple-choice behavior using a constrained utility maximization framework. These models often assume that a choice-maker would maximize an overall utility while satisfying the constraints they face. For instance, [Kim et al. \(2002\)](#) develop a direct utility model that allows for both choices of single goods, and joint choice of multiple goods. [Bhat \(2005, 2008\)](#) propose the Multiple Discrete-Continuous Extreme Value (MDCEV) model, which considers both package choice of multiple elemental alternatives, along with the continuous amount of consumption for the chosen elemental alternatives. The MDCEV model has been used across various fields due to the existence of closed-form choice probabilities (see [Ma and Ye, 2019](#), [Mouter et al., 2021](#), [Varghese and Jana, 2019](#), for instance). These models assume a continuous decision space and rely on Karush-Kuhn-Tucker (KKT) conditions to ensure the optimality of observed demand. Several extensions have been proposed for incorporating important factors of consumer decisions such as incorporating utility complementarity ([Bhat et al., 2015](#)), multiple constraints ([Castro et al., 2012](#), [Satomura et al., 2011](#)) and non-linear pricing ([Howell et al., 2016](#)).

Despite many successes, it remains challenging for the above models to capture the discrete nature of people’s behavior, which motivated [Bhat \(2022\)](#) to propose the Multiple Discrete-Count (MDC) model for multiple discrete-count choice modeling. The MDC model consists of two components: a total count model and an MDCEV model, where the discrete component corresponds to whether or not an individual chooses a specific elemental alternative over the total count, and the continuous component refers to the fractional split of investment in each of the consumed alternatives over the total count. While the MDC model has a closed-form solution and would be useful in multiple discrete choice applications, one possible drawback is that it involves many parameters, which could make it expensive to estimate. Additionally, the probability of zero counts involves enumerating many subsets of the goods, which can be intractable to handle exactly.

Several other structural models for multiple discrete demand have also been proposed. [Hendel \(1999\)](#) proposed a multiple-discrete choice model to examine differentiated products demand and applied it to firm-level survey data on demand for personal computers. [Dubé \(2004\)](#) employed the same modeling framework to analyze household panel data of carbonated soft drinks. [van der Lans \(2018\)](#) built a model based on the assumption that a consumer’s purchase decision maximizes variety while satisfying utility constraints. The solution to the constrained optimization problem is found by a sequential update procedure that reflects the discrete nature of consumer demand. [Lee and Allenby \(2014\)](#) extended models of constrained utility maximization by introducing demand indivisibility as an additional constraint, showing that ignoring the discrete nature of indivisible demand could result in biases in parameter estimates and policy implications. All the aforementioned models are based on constrained utility maximization, i.e., a choice-maker finds an exact optimal solution that maximizes his/her overall utility while satisfying potentially complex constraints. As mentioned earlier, this assumption might conflict with the importance of considering bounded rationality in decision-making and would be too strong because constrained utility maximization is not easy to solve.

Our modeling approach is distinct from the aforementioned works in that we directly extend the RUM framework and classical single discrete choice models (such as the MNL) to account for multiple discrete behaviors without imposing a strong assumption on how choice-makers choose composite alternatives. We assume only that the choice-makers have a constraint on the range of the number of elemental alternatives they will select, noting that our framework is general to incorporate more complex constraints. Furthermore, by demonstrating the equivalence between a multiple discrete choice decision and a path choice in a network of choices, we provide an intuitive interpretation of how individuals make complex multiple choices.

As we bring route choice models, particularly link-based recursive models, into the context of multiple discrete choice analysis, it is relevant to discuss some of the current state-of-the-art models in the literature. Route choice modeling has been approached using both path-based and link-based recursive models. Path-based models, as reviewed by [Prato \(2009\)](#), rely on path sampling, which may lead to parameter estimates that depend on the sampling method. In contrast, link-based recursive models are constructed using the dynamic discrete choice framework ([Rust, 1987](#)), and are essentially equivalent to discrete choice models based on all possible paths. Link-based recursive models have several advantages, including consistent estimation and ease of prediction, which have made them increasingly popular. The recursive logit (RL) model, proposed by [Fosgerau et al. \(2013\)](#), was the first link-based recursive model developed. Since then, several recursive models have been introduced to handle correlations between path utilities ([Mai, 2016](#), [Mai et al., 2015, 2018](#)), dynamic networks ([de Moraes Ramos et al., 2020](#)), stochastic time-dependent networks ([Mai et al., 2021](#)), or discounted behavior ([Oyama and Hato, 2017](#)), with applications in areas such as traffic management ([Baillon and Cominetti, 2008](#), [Melo, 2012](#)) and network pricing ([Zimmermann et al., 2021](#)). For a comprehensive review, we suggest referring to [Zimmermann and Frejinger \(2020\)](#).

### 3 Logit-based Multiple Discrete Choice Model

Let  $[m] = \{1, 2, \dots, m\}$  be the set of the elemental alternatives. It is assumed that each individual  $n$  associated a utility  $u_{ni}$  with each elemental choice alternative  $i \in [m]$ . Under the RUM framework (McFadden, 1981), each utility  $u_{ni}$  is the sum of two parts  $u_{ni} = v_{ni} + \epsilon_{ni}$ , where  $\epsilon_{ni}$  is a random term that cannot be observed by the analyst, and the deterministic term  $v_{ni}$  can include attributes of the alternative as well as socio-economic characteristics of the individual. In general, a linear-in-parameters formula is used, i.e.,  $v_{ni}(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \mathbf{x}_{ni}$ , where  $\boldsymbol{\beta}$  is a vector of parameters to be estimated and  $\mathbf{x}_{ni}$  is a vector of attributes of alternative  $i$  as observed by individual  $n$ . In the classical single-choice setting, the RUM framework assumes that individuals make a choice decision by maximizing their random utilities, yielding the choice probabilities

$$P(i|[m]) = P(u_{ni} \geq u_{nj}; \forall j \in [m]) = P(v_{ni}(\boldsymbol{\beta}) + \epsilon_{ni} \geq v_{nj}(\boldsymbol{\beta}) + \epsilon_{nj}; \forall j \in [m])$$

In the multiple discrete choice setting, each individual will choose a subset of elemental alternatives, instead one single choice alternative  $j \in [m]$ . So, each composite alternative now is a subset  $S$  of  $[m]$ . To extend the single RUM model to address the multiple discrete choice behavior, for each composite alternative  $S \subseteq [m]$ , let  $v^n(S|\boldsymbol{\beta})$  be its deterministic utility. We assume that  $v^n(S|\boldsymbol{\beta})$  can be written as a sum of the utilities of the elemental alternatives in  $S$ , i.e.,  $v^n(S|\boldsymbol{\beta}) = \sum_{i \in S} v_{ni}(\boldsymbol{\beta})$ . Under the RUM framework, we now assume that each individual selects a composite alternative by maximizing its random utility  $u^n(S|\boldsymbol{\beta}) = v^n(S|\boldsymbol{\beta}) + \epsilon_{nS}$ , from a set of composite alternatives that she finds appealing. We further assume that individual  $n$  will not consider all the possible subsets of  $[m]$ . Instead, she is only interested in subsets whose sizes lie in the interval  $[L_n, U_n]$ , where  $L_n, U_n \in [m]$  and  $L_n \leq U_n$ . These bounds represent the minimum and maximum allowable sizes for a composite alternative to be considered by the choice-maker. Let us denote  $\Omega^{[L_n, U_n]}$  as the consideration set of individual  $n$ ,  $\Omega^{[L_n, U_n]} = \{S \subseteq [m] \mid |S| \in [L_n, U_n]\}$ , the choice probability of a composite alternative  $S \in \Omega^{[L_n, U_n]}$  can be written as

$$P(S \mid \Omega^{[L_n, U_n]}, \boldsymbol{\beta}) = P(v_n(S|\boldsymbol{\beta}) + \epsilon_{nS} \geq v(S'|\boldsymbol{\beta}) + \epsilon_{nS'}; \forall S' \in \Omega^{[L_n, U_n]})$$

If the random terms  $\epsilon_{nS}$  are i.i.d Extreme Value Type I (i.e., the choice model over  $\Omega^{[L_n, U_n]}$  is the MNL model), then the choice probabilities become

$$P(S \mid \Omega^{[L_n, U_n]}) = \frac{\exp(v(S|\boldsymbol{\beta}))}{\sum_{S' \in \Omega^{[L_n, U_n]}} \exp(v(S'|\boldsymbol{\beta}))} = \frac{\exp(\sum_{i \in S} v_{ni}(\boldsymbol{\beta}))}{\sum_{S' \in \Omega^{[L_n, U_n]}} \exp(\sum_{i \in S'} v_{ni}(\boldsymbol{\beta}))} \quad (1)$$

We refer to this choice model as the logit-based multiple discrete choice model (LMDC). The model will collapse to the single-choice MNL model when  $L = U = 1$ . The number of elements in  $\Omega^{[L_n, U_n]}$  is  $\sum_{t=L_n}^{U_n} \binom{m}{t}$ , which is generally exponential in  $m$ . So, it is practically not possible to compute  $P(S \mid \Omega^{[L_n, U_n]})$  by enumerating all the elements in the set. This is the main challenge arising from the use of (1). The complex overlapping structure between the composite

alternatives within  $\Omega^{[L_n, U_n]}$  also presents a challenge in capturing the correlation using conventional single choice models such as the classical nested or cross-nested logit model (Ben-Akiva, 1973, Ben-Akiva and Bierlaire, 1999). We will show later that such correlations can be captured naturally using the nested recursive route choice model (Mai et al., 2015).

Now, given a set of observations  $\mathcal{D} = \{(S_n, L_n, U_n), n \in [N]\}$ , where each observation comprises an observed composite alternative and two integral values representing the minimum and maximum sizes of the subsets being considered by the corresponding choice-maker. The log-likelihood function of the observation data can be defined as

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\beta}) = \sum_{n \in [N]} \ln P(S_n | \Omega^{[L_n, U_n]}). \quad (2)$$

The model parameters  $\boldsymbol{\beta}$  can be estimated via maximum likelihood estimation (MLE), i.e. solving  $\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$ . It is well-known that, in the context of single choices and the choice model is MNL, if the utility functions are linear-in-parameters, then the log-likelihood function is concave in the model parameters. Despite having a more complex structure, it is possible to extend this result to show that the log-likelihood function in (2) also manifests concavity in  $\boldsymbol{\beta}$ .

**Proposition 1** *If  $v_{ni}(\boldsymbol{\beta})$  are linear in  $\boldsymbol{\beta}$ ,  $\forall n \in [N]$ ,  $i \in [m]$  and the choice model over the composite alternatives is the MNL, then the function  $\mathcal{L}(\boldsymbol{\beta})$  is concave in  $\boldsymbol{\beta}$ .*

**Proof.** We simply write the log-likelihood function under the MNL model as

$$\begin{aligned} \mathcal{L}(\mathcal{D}|\boldsymbol{\beta}) &= \sum_{n \in [N]} \ln P(S_n | \Omega^{[L_n, U_n]}) \\ &= \sum_{n \in [N]} v(S_n | \boldsymbol{\beta}) - \sum_{n \in [N]} \ln \left( \sum_{S' \in \Omega^{[L_n, U_n]}} \exp(v(S' | \boldsymbol{\beta})) \right). \end{aligned} \quad (3)$$

Since  $v_{ni}(\boldsymbol{\beta})$  is linear in  $\boldsymbol{\beta}$ , the first part of (3) is linear in  $\boldsymbol{\beta}$  and the second part of (3) is a log-sum-exp function of linear functions, which is also concave (Boyd et al., 2004). So,  $\mathcal{L}(\boldsymbol{\beta})$  is also concave as desired. ■

The concavity can guarantee global optimums when solving the MLE problem. This property generally does not hold if  $v_{ni}(\boldsymbol{\beta})$  is not linear in  $\boldsymbol{\beta}$ , or the choice model is not the MNL.

The computation of the log-likelihood function  $\mathcal{L}(\boldsymbol{\beta})$  involves handling sets of exponentially many elements  $\Omega^{[L_n, U_n]}$ . Therefore, it would be impractical to compute it directly. We will deal with this challenge in the next section by converting the choices over composite alternatives into a path choice model in a cycle-free network.



## 4 DAG-based Representations

In this section, we present two DAG-based representations for the multiple-choice problem. Our idea is to build a cycle-free directed graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  where  $\mathcal{N}$  is the set of nodes and  $\mathcal{A}$  is the set of arcs. Each node in  $\mathcal{N}$  will contain information about whether an elemental alternative is chosen or not, and the number of chosen elemental alternatives. We construct the DAGs in such a way that there is a one-to-one mapping between a composite alternative in the multiple-choice problem and a path choice in the DAG. It then leads to the result that a recursive route choice model on the DAG will be equivalent to the LMDC model described in the previous section. The DAG representations will be introduced below, with the subscript  $n$  used to index the choice-maker omitted for notational simplicity.

### 4.1 Binary-Choice DAG

The first DAG representation draws inspiration from the notion that, when confronted with a set of discrete choices, an individual may consider each option one by one and decide whether to choose it or not, keeping in mind the current count of the already chosen items. When the number of selected elemental alternatives reaches its upper bound, the choice-maker will stop the selection process. In order to represent this as a DAG, a graph will be constructed with nodes arranged in  $m + 2$  tiers. The first tier will comprise a single node that denotes the sole origin of the graph, while the final tier will likewise comprise a single node representing the only destination of the DAG. Each intermediate tier between the origin and destination will provide details pertaining to one of the  $m$  elemental alternatives. Each node at an intermediate tier between the origin and destination will encompass details regarding whether the corresponding elemental alternative is chosen or not, and the number of selected alternatives during the choice process. The DAG representation described above is referred to as the *Binary-Choice DAG* (or *BiC* for short). The name *Binary-Choice* reflects the fact that two options are presented for each elemental alternative – either being chosen or rejected. We illustrate this DAG in [Figure 1](#) below.

We create arcs in  $\mathcal{A}$  to connect nodes in the  $k$ -th tier to  $(k + 1)$ -th tier, for  $k = 0, 1, \dots, m$ . Each arc of the DAG will represent a binary decision – choosing or not choosing the next elemental alternative. For instance, from  $p_0^0$ , there are two arcs connecting it with  $p_1^0$  and  $p_1^1$ , where the arc  $(p_0^0, p_1^0)$  represents the option of not selecting alternative 1, and the arc  $(p_0^0, p_1^1)$  represents the option of selecting that alternative. All nodes in the  $m$ -th tier (the second last tier) will connect to the destination  $d$ , which is an absorbing node. For any node of form  $p_k^U$ , we create an arc directly connecting it to the destination, as the choice process already reaches the maximum number of elemental alternatives allowed and must be terminated. Furthermore, for a node  $p_m^k$  where  $k \leq L$ , we do not connect it with any other node of the DAG. In this case, the choice process reaches the last stage but has not gotten a sufficient number of choice alternatives to form a feasible composite choice. The following proposition shows that there is a one-to-one

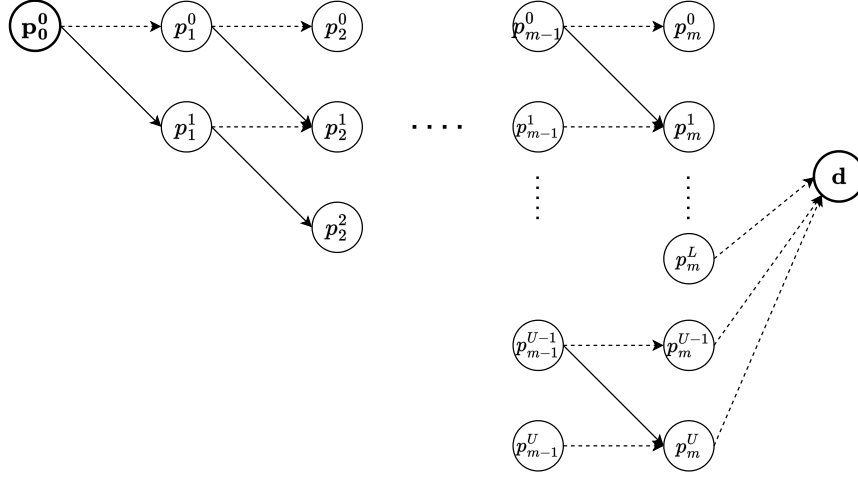


Figure 1: Binary-Choice DAG.

mapping between a path in the BiC DAG with a composite alternative.

**Proposition 2** *There is a one-to-one mapping between any composite alternative  $S \in \Omega^{[L,U]}$  and a path from  $p_0^0$  to  $d$  in the BiC DAG.*

**Proof.** For any composite alternative  $S \in \Omega^{[L,U]}$ , we then create a path  $\{p_0^{j_0}, \dots, p_m^{j_m}, d\}$  from  $p_0^0$  to  $d$  such that, for any  $i \in [m]$ , we follow the arc  $(p_{i-1}^{j_{i-1}}, p_i^{j_{i-1}})$  if  $i \notin S$  (i.e., alternative  $i$  is not chosen), and follow the arc  $(p_{i-1}^{j_{i-1}}, p_i^{j_{i-1}+1})$  if  $i \in S$  ( $i$  is chosen). Since  $S \in [L, U]$ , such a path always exists and can be uniquely created. For the opposite way, let  $\{p_0^{j_0}, p_1^{j_1}, p_2^{j_2}, \dots, p_m^{j_m}, d\}$  be a path connecting  $p_0^0$  and  $d$ . We can create a composite alternative  $S$  by adding item  $i$  to  $S$ , for any  $i \in [m]$ , if arc  $(p_{i-1}^{j_{i-1}}, p_i^{j_i})$  corresponds to a “selecting” option, i.e.,  $j_i = j_{i-1} + 1$ . The way we created the DAG implies that the size of  $S$  should be in the range  $[L, U]$ , thus  $S \in \Omega^{[L,U]}$ . We complete the proof. ■

Proposition 3 below gives the exact numbers of nodes and arcs of the BiC DAG.

**Proposition 3** *The number of nodes and arcs in the BiC DAG are*

$$|\mathcal{N}| = m + mU - \frac{U^2}{2} + \frac{U}{2} + 2$$

$$|\mathcal{A}| = 2m + 2mU - U^2 - L + 1.$$

## 4.2 Multiple-Choice DAG

The second DAG representation takes a different perspective on the multiple-choice behavior. Rather than assuming that individuals consider each option one by one and decide whether to “select” or “reject” it, this modeling approach presumes that the individuals immediately navigate to an alternative of interest, choosing them one by one. We represent this choice process

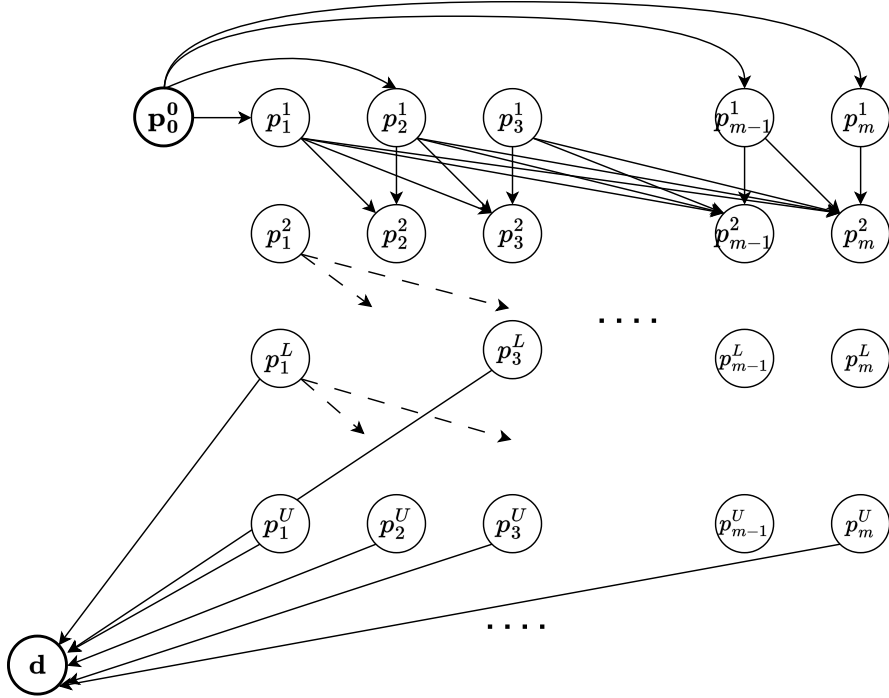


Figure 2: Multi-Choice DAG.

by the DAG in Figure 2 where nodes are also arranged into  $m + 2$  tiers, each corresponds to an elemental alternative indexed from 1 to  $m$ . There is also an original node, denoted as  $p_0^0$ , and a destination  $d$ . Each node of the DAG, except the destination, will be connected to nodes in the subsequent tiers, reflecting the fact that individuals can directly navigate to the next alternatives that they are interested. Each node in the DAG (except the destination) is denoted as  $p_j^c$ , where  $j \in [m]$  represents the index of the corresponding tier, and the superscript  $c$  represents the total number of selected elemental alternatives. As a result, there is no node with  $c > U$ . For each node  $p_j^c$  of the DAG, we create arcs connecting it with nodes  $\{p_{j+1}^{c+1}, \dots, p_m^{c+1}\}$ , reflecting the fact that individuals can select any subsequent alternatives in  $\{j + 1, \dots, m\}$ , and the number of selected items is increased by 1. For nodes of the form  $p_k^c$ , for any  $k \in [m]$  and  $L \leq c \leq U$ , we directly connect them to the destination, as the number of selected items is already in its allowed range  $[L, U]$ . We call this DAG as *Multi-Choice DAG* (or *MuC DAG* for short). The name *Multi-Choice* is used because, in contrast to the *BiC DAG* where each node has two arcs representing the options of selecting or rejecting, there are several arcs connecting each node to other nodes in the *MuC DAG*.

Similar to the *BiC DAG*, we can also show that there is a one-to-one mapping between any composite alternative  $S$  and a path from  $p_0^0$  to  $d$  in the *MuC DAG*.

**Proposition 4** *There is a one-to-one mapping between any composite alternative  $S \in \Omega^{[L,U]}$  and a path from  $p_0^0$  to  $d$  in the *MuC DAG*.*

**Proof.** The proof can be easily verified given the way the DAG is constructed. Let  $S = \{j_1, j_2, \dots, j_K\}$  be a feasible composite alternative, where  $K \in [L, U]$  and  $j_1 < j_2 < \dots < j_K$ .

We map  $S$  with the path  $(p_0^0, p_{j_1}^1, p_{j_2}^2, \dots, p_{j_K}^K, d)$ , where each arc  $(p_{j_k}^k, p_{j_{k+1}}^{k+1})$ ,  $k = 1, \dots, K - 1$ , corresponds to the selection of the elemental alternative  $j_{k+1}$  after  $j_k$ . It can be seen that  $(p_0^0, p_{j_1}^1, p_{j_2}^2, \dots, p_{j_K}^K, d)$  is a feasible path from  $p_0^0$  to  $d$ , and the mapping is unique. For the opposite direction, given a feasible path  $(p_0^0, p_{j_1}^1, \dots, p_{j_K}^K)$ , where  $j_k \in [m]$ ,  $k \in [K]$ , and  $j_1 < j_2 < \dots, j_K$ , we can uniquely map it with the subset  $\{j_1, \dots, j_K\}$ . The way we constructed the MuC DAG implies that this subset should belong to  $\Omega^{[L,U]}$ , which completes the proof. ■

The following proposition gives the exact numbers of nodes and arcs of the MuC DAG.

**Proposition 5** *The numbers of nodes and arcs in the MuC DAG are*

$$|\mathcal{N}| = mU - \frac{U^2}{2} + \frac{U}{2} + 2$$

$$|\mathcal{A}| = \begin{cases} \frac{m^2U}{2} - \frac{mU^2}{2} + mU + 2m + \frac{U^3}{6} - \frac{U^2}{2} + \frac{U}{3} + 2, & L = 0 \\ \frac{m^2U}{2} - \frac{mU^2}{2} - mL + mU + 2m + \frac{L^2}{2} - \frac{3L}{2} + \frac{U^3}{6} - \frac{U^2}{2} + \frac{U}{3} + 1, & L > 0 \end{cases}.$$

Compared to the size of the BiC DAG considered in the previous section, we can see that the numbers of nodes in the BiC and MuC DAG are quite similar – both are in  $\mathcal{O}(mU)$ , while the MuC DAG has significantly more arcs; it has  $\mathcal{O}(m^2U)$  arcs, compared to  $\mathcal{O}(mU)$  arcs in the BiC DAG.

## 5 Recursive Logit Model on the DAGs

In this section, we first present an RL model (Fosgerau et al., 2013) applied to the DAGs previously introduced. We then show that the choice probabilities over the composite alternatives are the same as the path choice probabilities yielded by the RL model, enabling us to utilize the estimation methods developed in prior work (Fosgerau et al., 2013, Mai et al., 2015) to efficiently estimate the LMDC model.

### 5.1 Recursive Logit Model on the DAGs

Consider a DAG  $G = (\mathcal{N}, \mathcal{A})$  where  $\mathcal{N}$  is the set of nodes and  $\mathcal{A}$  is the set of arcs. For each node  $k \in \mathcal{N}$ , let  $N(k)$  be the set of nodes that can be reached directly from  $k$ , i.e.,  $N(k) = \{a \mid (k, a) \in \mathcal{A}\}$ . A path in the DAG is a sequence of nodes  $(k_0, k_1, \dots, k_L)$  in which  $k_0 = p_0^0$  (the origin of the DAG),  $k_L = d$  (the destination of the DAG), and  $k_{i+1} \in N(k_i)$ ,  $\forall i = 0, \dots, L - 1$ . It can be seen that, for both BiC and MuC DAGs, the length of any path is  $m + 2$ . We associate each arc  $(k, a) \in \mathcal{A}$  with a utility  $v(a|k)$ . Specifically, for the BiC DAG, we set

$$\begin{cases} v^{\text{BiC}}(p_{j+1}^c | p_j^c) = 0, & c \in \{0, \dots, U\}, j \in \{0, \dots, m\} \\ v^{\text{BiC}}(p_{j+1}^{c+1} | p_j^c) = v_j & c \in \{0, \dots, U - 1\}, j \in \{0, \dots, m\} \\ v^{\text{BiC}}(k|d) = 0, & \forall k \in \mathcal{N}, d \in N(k), \end{cases} \quad (4)$$

noting that all the utilities  $v^{\text{BiC}}(a|k)$  are functions of the choice parameters  $\beta$ . Intuitively, for each arc  $(k, a)$  of the BiC DAG, we set  $v(a|k) = 0$  if the arc  $(k, a)$  represents a “non-selecting” option. Otherwise, we set  $v(a|k)$  to the utility of the elemental alternative that is associated with  $a$ . In particular, we assign a zero utility for arcs that connects nodes to the destination  $d$ .

For the MuC DAG, we set the arc utilities as follows

$$\begin{cases} v^{\text{MuC}}(p_{j+1}^{c+1}|p_j^c) = v_j, & c \in \{0, \dots, U-1\}, j \in \{0, \dots, m\} \\ v^{\text{MuC}}(k|d) = 0, & \forall k \in \mathcal{N}, d \in N(k). \end{cases} \quad (5)$$

That is, for any arc  $(k, a) \in \mathcal{A}$ , we set  $v^{\text{MuC}}(a|k)$  to the utility of the elemental alternative that is associated with  $a$ . All the arcs that connect nodes to the destination will be set to zero utilities. In the following, we use the notation  $v^{\text{BiC}}(\cdot)$  to present the RL model. The RL model applied to the MuC DAG can be formulated in the same way.

In the context of the RL model, a path choice is modeled as a sequence of arc choices. It is assumed that, at each node  $k \in \mathcal{N}$ ,  $k \neq d$ , the traveler chooses the next node  $a$  from the set of possible outgoing nodes  $A(k)$  by maximizing the sum of the arc utility  $v^{\text{BiC}}(a|k)$  and an expected downstream utility (or value function)  $V^{\text{BiC}}(a)$  from node  $a$  to the destination. This value function is defined recursively as

$$V^{\text{BiC}}(k) = \begin{cases} 0, & k = d \\ -\infty & \text{if } N(k) = \emptyset \\ E_\epsilon \left[ \max_{a \in A(k)} \{v^{\text{BiC}}(a|k) + \epsilon(a) + V^{\text{BiC}}(a)\} \right], & \forall k \in \mathcal{N} \setminus \{d\} \end{cases}$$

where  $\epsilon(a)$  are i.i.d Extreme Value Type I with scale  $\mu > 0$  and are independent of everything in the network. The value function can be rewritten as

$$V^{\text{BiC}}(k) = \begin{cases} 0, & k = d \\ -\infty & \text{if } N(k) = \emptyset \\ \mu \ln \left( \sum_{a \in N(k)} \exp \left( \frac{1}{\mu} (v^{\text{BiC}}(a|k) + V^{\text{BiC}}(a)) \right) \right), & \forall k \in \mathcal{N} \setminus \{d\}. \end{cases} \quad (6)$$

Moreover, according to the closed form of the logit model, the probability of choosing node  $a$  from node  $k$  is

$$\begin{aligned} P^{\text{BiC}}(a|k) &= \frac{\exp \left( \frac{1}{\mu} (v^{\text{BiC}}(a|k) + V^{\text{BiC}}(a)) \right)}{\sum_{a' \in N(k)} \exp \left( \frac{1}{\mu} (v^{\text{BiC}}(a'|k) + V^{\text{BiC}}(a')) \right)} \\ &= \frac{\exp \left( \frac{1}{\mu} (v^{\text{BiC}}(a|k) + V^{\text{BiC}}(a)) \right)}{\exp \left( \frac{1}{\mu} (V^{\text{BiC}}(k)) \right)}, \quad \forall (k, a) \in \mathcal{A}. \end{aligned}$$

The probability of a path  $\sigma = \{k_0, \dots, k_{m+1}\}$  in the DAG can be computed as the product of

the arc probabilities

$$P^{\text{BiC}}(\sigma|\beta) = \prod_{i=0}^m P^{\text{BiC}}(k_{i+1}|k_i) = \frac{\exp\left(\frac{1}{\mu}v^{\text{BiC}}(\sigma)\right)}{\exp\left(\frac{1}{\mu}V^{\text{BiC}}(k_0)\right)} = \frac{\exp\left(\frac{1}{\mu}v^{\text{BiC}}(\sigma)\right)}{\sum_{\sigma' \in \Omega^{\text{BiC}}} \exp\left(\frac{1}{\mu}v^{\text{BiC}}(\sigma')\right)} \quad (7)$$

where  $v^{\text{BiC}}(\sigma) = \sum_{i=0}^m v^{\text{BiC}}(k_{i+1}|k_i)$  and  $\Omega^{\text{BiC}}$  is the set of all the paths from  $p_0^0$  to  $d$ . Given a set of path observations  $\mathcal{D} = \{\sigma_1, \dots, \sigma_T\}$ , the RL can be estimated by maximizing the log-likelihood function

$$\mathcal{L}^{\text{BiC}}(\mathcal{D}|\beta) = \frac{1}{\mu} \sum_{t \in [T]} \left( v^{\text{BiC}}(\sigma_t) - V^{\text{BiC},t}(k_0^t) \right) \quad (8)$$

where  $k_0^t$  is the first node of path  $\sigma_t$ ,  $t \in [T]$ . Here, we note that the value function  $V^{\text{BiC},t}(\cdot)$  has superscript  $t$  which reflects the fact that it could be observation dependent, as different individuals may have different lower and upper bounds on the number of selected alternatives (i.e.,  $L$  and  $U$ ), leading to different DAG structures, and different value functions.

## 5.2 Equivalence Properties

Having introduced the RL model on the MuC and BiC DAG, we now connect the RL to the LMDC model presented in Section 3. Propositions 2 and 4 above show that there are one-to-one mappings between any subset  $S \in \Omega^{[L,U]}$  to a path in the BiC or Muc DAG. To facilitate the later exposition, for each subset  $S \in \Omega^{[L,U]}$ , let  $\sigma^{\text{BiC}}(S)$  and  $\sigma^{\text{MuC}}(S)$  be the two paths in the BiC and MuC DAGs, respectively, that maps with  $S$ . Theorem 1 below states our main result saying that the choice probability  $P(S|\Omega^{[L,U]})$  are equal to the probabilities of the corresponding paths in the BiC and MuC DAGs, given by the RL model.

**Theorem 1** *For any  $S \in \Omega^{[L,U]}$  we have*

$$P(S|\Omega^{[L,U]}) = P^{\text{BiC}}(\sigma^{\text{BiC}}(S)) = P^{\text{MuC}}(\sigma^{\text{MuC}}(S)).$$

**Proof.** We first prove the claim for the BiC DAG. For a subset  $S \in \Omega^{[L,U]}$ , Proposition 2 tells us that  $\sigma^{\text{BiC}}(S) = \{p_0^{j_0}, \dots, p_m^{j_m}, d\}$  such that  $p_0^{j_0} = p_0^0$  and for any  $i \in [m]$ , if  $i \in S$  then  $j_i = j_{i-1} + 1$  (i.e., alternative  $i$  is selected), and if  $i \notin S$  then  $j_i = j_{i-1}$  (i.e., alternative  $i$  is not selected). The total arc utility of  $\sigma^{\text{BiC}}(S)$  becomes  $\sum_{i=0}^m v^{\text{BiC}}(p_{i+1}^{j_{i+1}}|p_i^{j_i})$ . Moreover, the definition of  $v^{\text{BiC}}(\cdot)$  in (4) implies that  $v^{\text{BiC}}(p_{i+1}^{j_{i+1}}|p_i^{j_i}) = 0$  if  $j_{i+1} = j_i$  and  $v^{\text{BiC}}(p_{i+1}^{j_{i+1}}|p_i^{j_i}) = v_i$  if  $j_{i+1} = j_i + 1$ . Combine this with the above, we have  $\sum_{i=0}^m v^{\text{BiC}}(p_{i+1}^{j_{i+1}}|p_i^{j_i}) = \sum_{i \in S} v_i$ , or  $v(S) = v^{\text{BiC}}(\sigma^{\text{BiC}}(S))$ . Furthermore, from (1) and (7), the choice probabilities of subset  $S$  and

path  $\sigma^{\text{BiC}}(S)$  can be computed as

$$P(S | \Omega^{[L,U]}) = \frac{\exp\left(\frac{1}{\mu}v(S)\right)}{\sum_{S' \in \Omega^{[L,U]}} \exp\left(\frac{1}{\mu}v(S')\right)}$$

$$P^{\text{BiC}}(\sigma^{\text{BiC}}(S)) = \frac{\exp\left(\frac{1}{\mu}v^{\text{BiC}}(\sigma^{\text{BiC}}(S))\right)}{\sum_{\sigma' \in \Omega^{\text{BiC}}} \exp\left(\frac{1}{\mu}v^{\text{BiC}}(\sigma')\right)}.$$

This directly implies that  $P(S | \Omega^{[L,U]}) = P^{\text{BiC}}(\sigma^{\text{BiC}}(S))$  for any  $S \in \Omega^{[L,U]}$ .

In a similar way, for the MuC DAG, given a subset  $S = \{j_1, j_2, \dots, j_K\} \in \Omega^{[L,U]}$ , where  $K \in [L, U]$  and  $j_1 < j_2 < \dots < j_K$ . Proposition 4 tells us that  $S$  can be uniquely mapped to path  $\sigma^{\text{MuC}}(S) = (p_0^0, p_{j_1}^1, p_{j_2}^2, \dots, p_{j_K}^K, d)$ , where each arc  $(p_{j_k}^k, p_{j_{k+1}}^{k+1})$ ,  $k = 1, \dots, K - 1$ , corresponds to the selection of the elemental alternative  $j_{k+1}$  after  $j_k$ . The accumulated utility of  $\sigma^{\text{MuC}}(S)$  becomes

$$v^{\text{MuC}}(\sigma^{\text{MuC}}(S)) = \sum_{k=0}^{K-1} v^{\text{MuC}}(p_{j_{k+1}}^{k+1} | p_{j_k}^k) \stackrel{(a)}{=} \sum_{k=0}^{K-1} v_{j_{k+1}} = \sum_{i \in S} v_i,$$

where (a) is due to the definition of  $v^{\text{MuC}}(\cdot)$  in (5). We then have  $v^{\text{MuC}}(\sigma^{\text{MuC}}(S)) = v(S)$ . Similar to the case of the BiC DAG, the choice probabilities should be the same, i.e.,  $P(S | \Omega^{[L,U]}) = P^{\text{MuC}}(\sigma^{\text{MuC}}(S))$ , which completes the proof. ■

Let  $\mathcal{D} = \{S_1, \dots, S_N\}$  be a set of observed subsets and  $\mathcal{D}^{\text{BiC}} = \{\sigma^{\text{BiC}}(S_1), \dots, \sigma^{\text{BiC}}(S_N)\}$  and  $\mathcal{D}^{\text{MuC}} = \{\sigma^{\text{MuC}}(S_1), \dots, \sigma^{\text{MuC}}(S_N)\}$  be the sets of the mapped paths in the BiC DAG and MuC DAG, respectively. Theorem 1 directly implies that the log-likelihood of  $\mathcal{D}$  is also equal to the log-likelihood functions of  $\mathcal{D}^{\text{BiC}}$  and  $\mathcal{D}^{\text{MuC}}$  defined based on the BiC and MuC DAG, respectively. As a result, the estimation of the LMDC model can be done by estimating the RL with the  $\mathcal{D}^{\text{BiC}}$  or  $\mathcal{D}^{\text{MuC}}$ . We state this result in the following Corollary.

**Corollary 1** *For any set of subset observations  $\mathcal{D} = \{S_1, \dots, S_N\}$ , we have*

$$\mathcal{L}(\mathcal{D} | \beta) = \mathcal{L}^{\text{BiC}}(\mathcal{D}^{\text{BiC}} | \beta) = \mathcal{L}^{\text{MuC}}(\mathcal{D}^{\text{MuC}} | \beta).$$

### 5.3 Illustrative Example

We give a simple example to illustrate the equivalence between the LMDC model introduced in Section 3 and the RL model on the DAGs. We take a simple instance with  $m = 3$  (there are 3 elemental alternatives) and the number of alternatives to be selected is in the range of  $[1, 2]$  (i.e.,  $L = 1, U = 2$ ). The elemental utilities and the choice probabilities of the composite alternatives are provided in Table 1. The probability of each subset is calculated using (1).

The corresponding BiC DAG for this example is illustrated in Figure 3, noting that dashed arcs represent "not-selecting" options and have zero utilities. In Table 2 we report the corresponding

		No.	Subset	Subset utility	$P(S_n)$
		$S_1$	$\{s_1\}$	-1	0.414
		$S_2$	$\{s_2\}$	-1.5	0.251
		$S_3$	$\{s_3\}$	-2	0.152
		$S_4$	$\{s_1, s_2\}$	-2.5	0.093
		$S_5$	$\{s_1, s_3\}$	-3	0.056
		$S_6$	$\{s_2, s_3\}$	-3.5	0.034

Table 1: LMDC model with  $m = 3$ ,  $L = 1$ ,  $U = 2$

value function and path probabilities given by the RL model. It is easy to see the matching path  $\sigma^n$  has the same probability as that of  $S^n$  reported in Table 1.

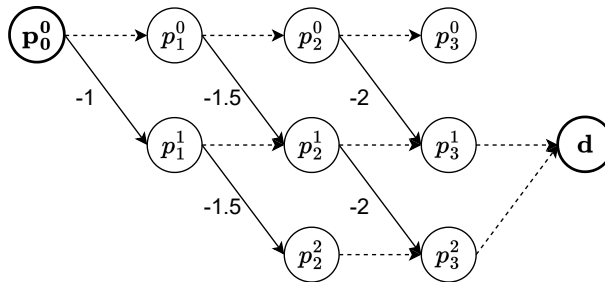


Figure 3: BiC DAG representation

Similar to the BiC DAG, we illustrate the MuC DAG representation of the toy instance in Figure 4 where dashed arcs represent the termination of the choice process, i.e., the choice-maker finishes the selecting process and directly navigates to the destination. We report the value function as well as the path choice probabilities in Table 3 which also shows that path choice probabilities are identical to the choice probabilities of the corresponding subsets.

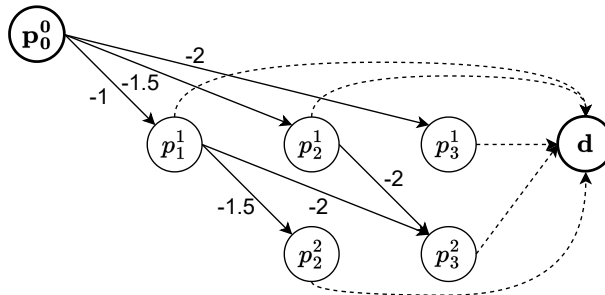


Figure 4: MuC DAG representation

## 5.4 Maximum Likelihood Estimation

In this section, our focus is on estimating the RL model applied to the DAGs. We specifically delve into the calculation of the value functions  $V^{\text{BiC}}(\cdot)$  and  $V^{\text{MuC}}(\cdot)$ , along with their gradients,



Node	$V^{\text{BiC}}(\cdot)$				
$p_0^0$	-0.118				
$p_1^0$	-0.945	No.	Path	Subset	$P(\sigma^n)$
$p_2^0$	-2	$\sigma^1$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^1 \rightarrow p_3^1 \rightarrow d$	$\{s_1\}$	$0.563 \times 0.836 \times 0.881$
$p_3^0$	$-\infty$	$\sigma^2$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^1 \rightarrow d$	$\{s_2\}$	$0.437 \times 0.652 \times 0.881$
$p_1^1$	0.306	$\sigma^3$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^0 \rightarrow p_3^1 \rightarrow d$	$\{s_2\}$	$0.437 \times 0.348 \times 1.000$
$p_2^1$	0.127	$\sigma^4$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^2 \rightarrow p_3^2 \rightarrow d$	$\{s_1, s_2\}$	$0.563 \times 0.164 \times 1.000$
$p_3^1$	0	$\sigma^5$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	$\{s_1, s_3\}$	$0.563 \times 0.836 \times 0.119$
$p_2^2$	0	$\sigma^6$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	$\{s_2, s_3\}$	$0.437 \times 0.652 \times 0.119$
$p_3^2$	0				
$d$	0				

Table 2: Value function and path choice probabilities for the BiC DAG example.

Node	$V^{\text{MuC}}(\cdot)$				
$p_0^0$	-0.118				
$p_1^1$	0.306	No.	Path	Subset	$P(\sigma^n)$
$p_2^1$	0.127	$\sigma^1$	$p_0^0 \rightarrow p_1^1 \rightarrow d$	$\{s_1\}$	$0.563 \times 0.736$
$p_3^1$	0	$\sigma^2$	$p_0^0 \rightarrow p_2^1 \rightarrow d$	$\{s_2\}$	$0.285 \times 0.881$
$p_2^2$	0	$\sigma^3$	$p_0^0 \rightarrow p_3^1 \rightarrow d$	$\{s_3\}$	$0.152 \times 1.000$
$p_3^2$	0	$\sigma^4$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^2 \rightarrow d$	$\{s_1, s_2\}$	$0.563 \times 0.164 \times 1.000$
$d$	0	$\sigma^5$	$p_0^0 \rightarrow p_1^1 \rightarrow p_3^2 \rightarrow d$	$\{s_1, s_3\}$	$0.563 \times 0.100 \times 1.000$
		$\sigma^6$	$p_0^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	$\{s_2, s_3\}$	$0.285 \times 0.119 \times 1.000$

Table 3: Value function and path choice probabilities for the MuC DAG representation.

as they play a crucial role in the maximum likelihood estimation (Mai and Frejinger, 2022). Our analyses primarily revolve around the BiC DAG, with a brief discussion for the MuC DAG case.

According to prior work (Fosgerau et al., 2013, Mai et al., 2015), to compute the value function, it is convenient to use a vector  $\mathbf{z}$  of size  $|\mathcal{N}|$  with elements  $z_k^{\text{BiC}} = \exp\left(\frac{1}{\mu} V^{\text{BiC}}(k)\right)$  for all  $k \in \mathcal{N}$ . From Equation 6,  $\mathbf{z}^{\text{BiC}}$  satisfies the following system of equations

$$z_k^{\text{BiC}} = \begin{cases} 1 & \text{if } k = d \\ \sum_{a \in N(k)} z_a^{\text{BiC}} \exp\left(\frac{1}{\mu} v^{\text{BiC}}(a|k)\right) & \text{otherwise.} \end{cases} \quad (9)$$

By defining a matrix  $\mathbf{M}^{\text{BiC}}$  of size  $|\mathcal{N}| \times |\mathcal{N}|$  with entries

$$M_{ka}^{\text{BiC}} = \begin{cases} \exp\left(\frac{1}{\mu} v^{\text{BiC}}(a|k)\right), & a \in N(k), k \in \mathcal{N} \setminus \{d\} \\ 0 & \text{otherwise.} \end{cases}$$

The linear system in (9) can be written in a matrix form as  $\mathbf{z}^{\text{BiC}} = \mathbf{M}^{\text{BiC}} \mathbf{z}^{\text{BiC}} + \mathbf{b}$  or  $(\mathbf{I} - \mathbf{M}^{\text{BiC}}) \mathbf{z}^{\text{BiC}} = \mathbf{b}$ , where  $\mathbf{b}$  is a vector of size  $|\mathcal{N}|$  with zero elements except  $b_d = 1$  and  $\mathbf{I}$  is the identity matrix of size  $|\mathcal{N}|$ . As a result, one can obtain  $\mathbf{z}^{\text{BiC}}$  by solving the system of linear equations  $(\mathbf{I} - \mathbf{M}^{\text{BiC}}) \mathbf{z}^{\text{BiC}} = \mathbf{b}$ , which will always give a unique solution if  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$  is invertible. In a general setting where the graph may have cycles, it is possible that  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$  is not invertible if the arc utilities  $v^{\text{BiC}}(a|k)$  are not negative and their the magnitudes are not significant (Mai and Frejinger, 2022). In the context of the BiC (or MuC) DAG, the graph does not contain any cycles, implying that the matrix  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$  is consistently invertible, according to Proposition 6 below. Consequently, the vector  $\mathbf{z}^{\text{BiC}}$  can be obtained by utilizing the inverse matrix of  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$ , specifically as  $\mathbf{z}^{\text{BiC}} = (\mathbf{I} - \mathbf{M}^{\text{BiC}})^{-1} \mathbf{b}$ .

**Proposition 6**  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$  is always invertible.

**Proof.** For any  $n \in \mathbb{N}$ , let us consider  $(\mathbf{M}^{\text{BiC}})^n = \underbrace{\mathbf{M}^{\text{BiC}} \times \mathbf{M}^{\text{BiC}} \times \dots \times \mathbf{M}^{\text{BiC}}}_{n \text{ times}}$  with entries

$$(\mathbf{M}^{\text{BiC}})_{ss'}^n = \sum_{\substack{(t_0, \dots, t_n) \\ t_i \in \mathcal{N}, i=0, \dots, n \\ t_0=s, t_n=s'}} \left( \prod_{i=0}^{n-1} M_{t_i t_{i+1}}^{\text{BiC}} \right).$$

We see that the BiC DAG is cycle-free and any path in the DAG has no more than  $m+2$  nodes. Thus, for any  $n > m+1$ , for any sequence  $(t_0, s_1, \dots, t_n)$  there is at least a pair  $(t_j, t_{j+1})$ ,  $0 \leq j \leq n-1$  such that  $t_{j+1} \notin N(t_j)$ , or  $M_{t_j t_{j+1}}^{\text{BiC}} = 0$ , leading to the fact that  $(\mathbf{M}^{\text{BiC}})_{ss'}^n = 0$  for any  $s, s' \in \mathcal{N}$ . Thus, if  $n > m+1$ , we have  $(\mathbf{M}^{\text{BiC}})^n = \mathbf{0}$  (i.e., the all-zero matrix of size  $|\mathcal{N}| \times |\mathcal{N}|$ ). We then have

$$(\mathbf{I} - \mathbf{M}^{\text{BiC}}) \left( \sum_{t=0}^{n-1} (\mathbf{M}^{\text{BiC}})^t \right) = (\mathbf{I} - (\mathbf{M}^{\text{BiC}})^n) = \mathbf{I},$$

which implies  $\det(\mathbf{I} - \mathbf{M}^{\text{BiC}}) \neq 0$ , or equivalently,  $\mathbf{I} - \mathbf{M}^{\text{BiC}}$  is invertible, as desired. ■

The invertibility of  $\mathbf{I} - \mathbf{M}^{\text{BiC}}$  implies that we can solve the linear system  $(\mathbf{I} - \mathbf{M}^{\text{BiC}}) \mathbf{z}^{\text{BiC}} = \mathbf{b}$  to obtain a unique solution. Directly inverting the matrix  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})$  would be not practically viable if the size of  $\mathbf{M}^{\text{BiC}}$  is large. An alternative method would be to use a simple value iteration, i.e., we iteratively compute  $(\mathbf{z}^{\text{BiC}, t+1}) = \mathbf{M}^{\text{BiC}} \mathbf{z}^{\text{BiC}, t} + \mathbf{b}$  with  $t = 0, 1, \dots$ , starting from an initial vector  $\mathbf{z}^{\text{BiC}, 0}$ , until converging to a fixed point solution. In the context of the BiC DAG, it is possible to show that the value iteration will converge to the desired fixed-point solution after  $m+2$  iterations from any starting vector  $\mathbf{z}^{\text{BiC}, 0}$ .

**Proposition 7** For the BiC DAG, the value iteration solves the linear system  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})\mathbf{z}^{\text{BiC}} = \mathbf{b}$  after  $m + 2$  iterations. As a result, the value function can be computed in  $\mathcal{O}(m^2U)$ .

**Proof.** We leverage the fact that the BiC DAG is cycle-free and its nodes are organized into  $m + 2$  tiers and arcs only connect nodes in consecutive tiers. To facilitate the later arguments, let us denote  $\mathcal{T}(k)$ , for any  $k \in \mathcal{N}$  as the index of the tier of node  $k$ , i.e., for any node  $p_j^i \in \mathcal{N}$ ,  $\mathcal{T}(p_j^i) = j$ , and  $\mathcal{T}(d) = m + 1$ . Let  $\mathbf{z}^{\text{BiC}^*}$  be the unique fixed point solution to the system  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})\mathbf{z}^{\text{BiC}} = \mathbf{b}$ . We will prove, by induction, that after the  $t$ -th iteration of the value iteration,  $z_k^{\text{BiC},t} = z_k^{\text{BiC}^*}$  for all  $k \in \mathcal{N}$  such that  $\mathcal{T}(k) \geq m + 2 - t$ . This, indeed, holds for  $t = 1$  as we see that  $z_d^{\text{BiC},t} = z_d^{\text{BiC}^*}$ . Let's assume that it holds for  $t \geq 1$ , i.e.,  $z_k^{\text{BiC},t} = z_k^{\text{BiC}^*}$  for any  $k$  such that  $\mathcal{T}(k) \geq m + 2 - t$ . In the next iteration, we will update  $\mathbf{z}^{\text{BiC},t+1}$  as

$$z_k^{\text{BiC},t+1} = \sum_{a \in N(k)} z_a^{\text{BiC},t} M_{ka}^{\text{BiC}} + b_k, \quad \forall k \in \mathcal{N} \setminus \{d\}.$$

we have that for any  $k \in \mathcal{N}$  such that  $\mathcal{T}(k) \geq m + 2 - (t + 1)$  ( $k$  in Tier  $m + 2 - (t + 1)$ ), then for any  $a \in N(k)$  we have  $\mathcal{T}(a) \geq m + 2 - t$ . From the induction assumption we have  $z_a^{\text{BiC},t} = z_a^{\text{BiC}^*}$  for any  $a \in N(k)$ . Thus, for any  $k$  such that  $\mathcal{T}(k) \geq m + 2 - (t + 1)$ ,

$$z_k^{\text{BiC},t+1} = \sum_{a \in N(k)} z_a^{\text{BiC},t} M_{ka}^{\text{BiC}} + b_k = \sum_{a \in N(k)} z_a^{\text{BiC}^*} M_{ka}^{\text{BiC}} + b_k = z_k^{\text{BiC}^*},$$

implying that the induction assumption holds for  $t + 1$ . So after  $m + 2$  iterations,  $z_a^{\text{BiC},t} = z_a^{\text{BiC}^*}$  for any  $a \in \mathcal{N}$ .

For the computational complexity, we see that  $\mathbf{M}^{\text{BiC}}$  contains  $|\mathcal{A}| = \mathcal{O}(mU)$  non-zero entries. Thus, each iteration  $(\mathbf{z}^{\text{BiC},t+1}) \leftarrow \mathbf{M}^{\text{BiC}}\mathbf{z}^{\text{BiC},t} + \mathbf{b}$  can be done in  $\mathcal{O}(mU)$ . Consequently, we can obtain the fixed-point solution after  $m + 2$  iterations of the value iteration which can be done in  $\mathcal{O}(m^2U)$ . ■

In the case of the MuC DAG, in a similar way, since the graph is cycle-free, we can prove that the matrix  $\mathbf{I} - \mathbf{M}^{\text{MuC}}$  is always invertible. The fixed-point solution can also be computed by value iteration as  $(\mathbf{z}^{\text{MuC},t+1}) \leftarrow \mathbf{M}^{\text{MuC}}\mathbf{z}^{\text{BiC},t} + \mathbf{b}$ ,  $t = 0, 1, \dots$ . The following proposition states a result regarding the convergence of the value iteration and the computational complexity.

**Proposition 8** For the MuC DAG, the value iteration solves the linear system  $(\mathbf{I} - \mathbf{M}^{\text{MuC}})\mathbf{z}^{\text{MuC}} = \mathbf{b}$  after  $m + 2$  iterations, which can be done in  $\mathcal{O}(m^3U)$ .

**Proof.** The claim can be validated in a similar way as the proof of Proposition 7 as the nodes are also arranged into  $m + 2$  tiers. The only difference is that, in the MuC DAG, we have arcs connecting each node to several nodes in the subsequent tiers. This, however, does not affect the induction proof. That is, if after  $t$  iterations we have  $z_k^{\text{BiC},t} = z_k^{\text{BiC}^*}$  for any  $k$  such that  $\mathcal{T}(k) \geq m + 2 - t$ , then in the next iteration, we will update  $\mathbf{z}_k^{\text{BiC},t+1}$  using  $\mathbf{z}_a^{\text{BiC},t}$  such that  $a \in N(k)$ . The induction assumption also implies that  $\mathbf{z}_a^{\text{BiC},t} = \mathbf{z}_a^{\text{BiC}^*}$  for any  $a \in N(k)$ ,

leading to the result that the induction assumption also holds for  $t + 1$ , which validates the convergence of the value iteration. For the computational complexity, we see that  $\mathbf{M}^{\text{MuC}}$  has  $\mathcal{O}(m^2U)$  non-zero elements. So, each iteration of the value function will take about  $\mathcal{O}(m^2U)$  time, and in total, the value function will take  $\mathcal{O}(m^3U)$  time. ■

Once the linear system  $(\mathbf{I} - \mathbf{M}^{\text{BiC}})\mathbf{z}^{\text{BiC}} = \mathbf{b}$  is resolved, the gradients of the value function w.r.t parameter  $\beta_q$  can be computed as

$$\frac{\partial \mathbf{z}^{\text{BiC}}}{\partial \beta_q} = (\mathbf{I} - \mathbf{M}^{\text{BiC}})^{-1} \left[ \frac{\partial \mathbf{M}^{\text{BiC}}}{\partial \beta_q} \right] \mathbf{z}^{\text{BiC}} \quad \text{and} \quad \frac{\partial V^{\text{BiC}}(k)}{\partial \beta_q} = \mu_{z_k^{\text{BiC}}} \left( \frac{\partial z_k^{\text{BiC}}}{\partial \beta_q} \right),$$

which is another system of linear equations.

## 6 Nested Recursive Logit Model on the DAGs

Mai et al. (2015) propose the nested RL model to capture utility correlation in the network, overcoming the Independence of irrelevant alternatives (IIA) limitation of the RL model (Fosgerau et al., 2013, McFadden, 1981). One notable advantage of the nested RL approach is its ability to naturally capture correlation patterns that arise from the graph structure, similar to the network GEV model (Daly and Bierlaire, 2006, Mai et al., 2017). This is achieved by allowing the scale parameters of the random terms to vary across nodes. In the context of multiple discrete choices, we have shown previously that there is a one-to-one mapping between a composite alternative to a path in the BiC or MuC graphs and the logit choice model over the subsets is equivalent to the RL model applied to either the BiC or MuC graphs. So, it is to be expected that the nested RL model on the BiC or MuC would be useful to naturally capture the correlation between composite alternatives' utilities. In this section, we introduced the nested RL applied to the DAGs and explore its properties through illustrative examples.

### 6.1 Nested Recursive Logit

We use the same notations as in the description of the RL model. In the context of the nested RL, the expected maximum utility of a node  $k \in \mathcal{N}$  is also defined recursively as

$$\tilde{V}^{\text{BiC}}(k) = \begin{cases} 0, & k = d \\ -\infty, & N(k) = \emptyset \\ \mathbb{E} \left[ \max_{a \in N(k)} \left\{ v^{\text{BiC}}(a|k) + \mu_k \epsilon(a) + \tilde{V}^{\text{BiC}}(a) \right\} \right], & \forall k \in \mathcal{N} \end{cases}$$

The difference with regards to the RL model is that the scale parameters  $\mu_k$  are now varying across nodes and can be modeled as functions of the node attributes. Since the choice at each

node is still the logit model, the value function can be computed through closed-form equations:

$$\exp\left(\frac{1}{\mu_k}\tilde{V}^{\text{BiC}}(k)\right) = \begin{cases} 1, & k = d \\ 0, & N(k) = \emptyset \\ \sum_{a \in N(k)} \exp\left(\frac{1}{\mu_k}(v^{\text{BiC}}(a|k) + \tilde{V}^{\text{BiC}}(a))\right), & \forall k \in \mathcal{N} \end{cases} \quad (10)$$

and the node choice probabilities are computed as

$$\begin{aligned} P^{\text{BiC}}(a|k) &= \frac{\exp\left(\frac{1}{\mu_k}(v^{\text{BiC}}(a|k) + \tilde{V}^{\text{BiC}}(a))\right)}{\sum_{a' \in N(k)} \exp\left(\frac{1}{\mu_k}(v^{\text{BiC}}(a'|k) + \tilde{V}^{\text{BiC}}(a'))\right)} \\ &= \exp\left(\frac{1}{\mu_k}(v^{\text{BiC}}(a|k) + \tilde{V}^{\text{BiC}}(a) - \tilde{V}^{\text{BiC}}(k))\right) \end{aligned}$$

To compute the value function, we denote  $\tilde{\mathbf{z}}^{\text{BiC}}$  as a vector of size  $|\mathcal{N}|$  with entries  $\tilde{z}_k^{\text{BiC}} = \frac{1}{\mu_k} \exp\left(\tilde{V}^{\text{BiC}}(k)\right)$  and rewrite [Equation 10](#)

$$\tilde{z}_k^{\text{BiC}} = \sum_{a \in \mathcal{N}} M_{ka}^{\text{BiC}} (\tilde{z}_a^{\text{BiC}})^{\mu_a/\mu_k} + b_k, \quad \forall k \in \mathcal{N}. \quad (11)$$

where matrix  $\mathbf{M}^{\text{BiC}}(|\mathcal{N}| \times |\mathcal{N}|)$  is now defined as  $M_{ka}^{\text{BiC}} = \exp(v^{\text{BiC}}(a|k)/\mu_k)$  for all  $a \in N(k)$  and  $M_{ka}^{\text{BiC}} = 0$  if  $a \notin N(k)$ ,  $\forall k \in \mathcal{N}$ , and  $\mathbf{b}$  is still a vector of size  $|\mathcal{N}|$  with zero elements except  $b_d = 1$ . The system of equations in (11) is highly nonlinear and is more intricate to solve compared to that of the RL model. (Mai et al., 2015) show that this system can be solved conveniently via value iteration for which the convergence cannot be guaranteed if the network contains cycles (Mai and Frejinger, 2022). In our context, it is possible to show that the value iteration always converges to a unique fixed point solution after  $m + 2$  iterations, for both BiC and MuC graphs.

**Proposition 9** *For both BiC and MuC DAGs, the value iteration solves the system (11) after  $m + 2$  iterations with any starting vector  $\tilde{\mathbf{z}}^{\text{BiC},0}$  (or  $\tilde{\mathbf{z}}^{\text{MuC},0}$ ).*

**Proof.** The proof is very similar to the proof of Propositions 7 and 8. That is, since the nodes are organized in  $(m + 2)$  tiers, we can prove, by induction, that after  $t$  iterations, the value iteration will update  $\tilde{\mathbf{z}}^{\text{BiC},t}$  such that  $\tilde{z}_k^{\text{BiC},t} = \tilde{z}_k^{\text{BiC}*}$  for all  $k \in \mathcal{N}$  such that  $\mathcal{T}(k) = m + 2 - t$ , where  $\tilde{z}^{\text{BiC}*}$  the fixed point solution. As a result, the value iteration will return the fixed point solution after  $m + 2$  iterations. The same arguments apply to the MuC graph. ■

The gradients of the value function, however, can be computed by solving systems of linear equations, similar to the case of the RL model. We refer the reader to (Mai et al., 2015) for more details.

## 6.2 Example and Properties

Within this section, we present examples to examine the characteristics of the nested RL model when it is applied to the DAGs. We, however, refrain from discussing how the nested RL model captures the correlation arising from the graph structure, as this has already been exploited in previous work (Mai et al., 2015). In the previous section, we show that for both BiC and MuC graphs, the RL model is equivalent to the LMDC model, regardless of the order of the elemental alternatives that are chosen to construct the DAGs. In other words, the order of the alternatives does not affect the path choice probabilities given by the RL model. In the following, we will show that it might not be the case under the nested RL model, i.e. the path choice probabilities may change if we shuffle the order of the elemental alternatives when constructing the DAGs.

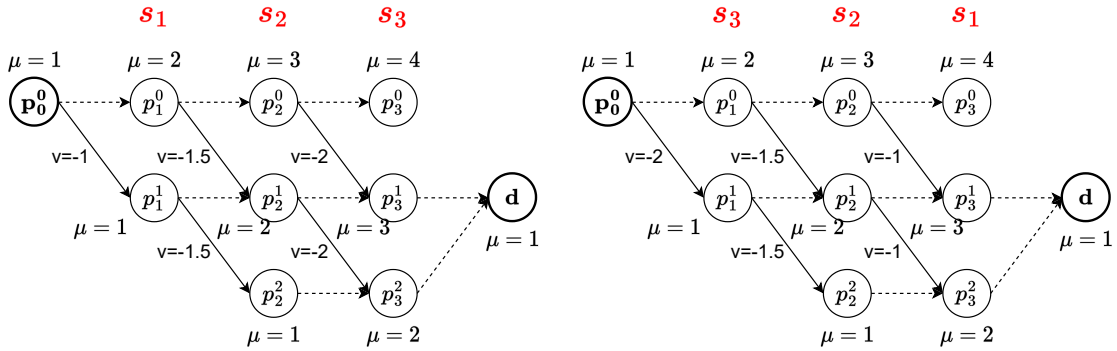


Figure 5: BiC DAG representations based on different alternative orders, for the 3-alternative choice problem.

We take the same example as in Section 5.3. There are 3 alternatives denoted as  $\{s_1, s_2, s_3\}$  and each alternative  $s$  is associated with an utility  $v(s)$  and a scale parameter  $\mu_s$ . In Figure 5 we illustrate two BiC DAG representations of the choices over the three alternatives, where the left-hand-side figure is constructed based on the order  $\{s_1, s_2, s_3\}$  and the right-hand-side one is based on  $\{s_3, s_2, s_1\}$ . We also provide the arc utilities and the scale parameters next to the corresponding arcs and nodes. To facilitate the later analysis, we denote the left-hand-side graph as BiC-1 and the other one as BiC-2.

We apply the nested RL model to the BiC-1 and BiC-2 graphs and report the path probabilities in Table 4, where each row reports a composite alternative, the mapping paths in BiC-1 and BiC-2 and their corresponding choice probabilities given by the nested RL. It can be observed that the nested RL offers different distributions over composite alternatives (or subsets) when applied to the BiC-1 and BiC-2 graphs, noting that the distributions are the same when using the RL model. This suggests that the chosen order of alternatives used to construct the BiC DAG representations would affect the resulting nested RL choice model.

Using the same example, we can also show that the MuC DAG with different alternative orders can yield different nested RL models. It is more interesting to explore the question that if the BiC and MuC are constructed using the same alternative order, does the probability distribution over paths remains the same? To this end, we simply keep the same order  $\{s_1, s_2, s_3\}$  to construct

Subset	BiC-1		BiC-2	
	Path	Path Probabilities	Path	Path probabilities
$\{s_1\}$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^1 \rightarrow p_3^1 \rightarrow d$	0.28	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^0 \rightarrow p_3^1 \rightarrow d$	0.10
$\{s_2\}$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^1 \rightarrow d$	0.27	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^1 \rightarrow d$	0.29
$\{s_3\}$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^0 \rightarrow p_3^1 \rightarrow d$	0.21	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^0 \rightarrow p_3^1 \rightarrow d$	0.37
$\{s_1, s_2\}$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^2 \rightarrow p_3^2 \rightarrow d$	0.05	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.01
$\{s_1, s_3\}$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.10	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.06
$\{s_2, s_3\}$	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.10	$p_0^0 \rightarrow p_1^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.17

Table 4: Path choice probabilities given by the nested RL on the BiC-1 and BiC-2 graphs.

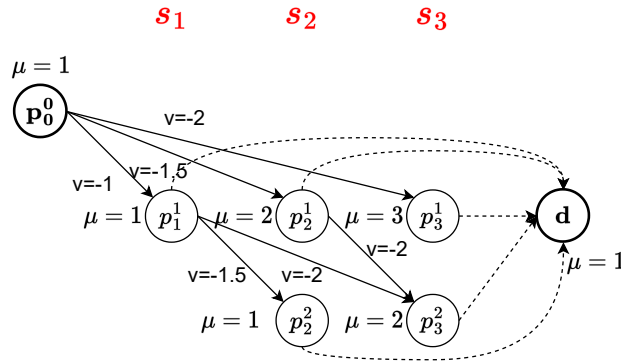


Figure 6: MuC DAG representation for the 3-alternative choice problem.

the MuC graph (as illustrated in Figure 6). The path probabilities are reported in Table 5, which clearly shows that the path distribution given by the nested RL on the MuC DAG is different from those given by the nested RL and BiC reported in Table 4. So, we conclude that, in general, the probability distributions given by the BiC DAG and MuC DAG models, based on the same alternative order, are not necessarily the same. Interestingly, it can be shown that, under some specific settings of the scale function, these two distributions can be completely identical. We state this result in Proposition 10 below.

**Proposition 10** *Assume that the BiC DAG and MuC DAG are constructed based on the same order of elemental alternatives and the scale parameter  $\mu_{p_j^c}$  associated with node  $p_j^c$  in both BiC and MuC graphs is only dependent on  $c$  (i.e., the number of selected alternatives at that node). Then the nested RL models on the BiC and MuC are equivalent, i.e., for any subset  $S \in \Omega^{[L,U]}$  we have*

$$\tilde{P}^{\text{BiC}}(\sigma^{\text{BiC}}(S)) = \tilde{P}^{\text{MuC}}(\sigma^{\text{MuC}}(S)).$$

The proof can be found in the appendix.

Subset	Path	Path Probabilities
$\{s_1\}$	$p_0^0 \rightarrow p_1^1 \rightarrow d$	0.35
$\{s_2\}$	$p_0^0 \rightarrow p_2^1 \rightarrow d$	0.29
$\{s_3\}$	$p_0^0 \rightarrow p_3^1 \rightarrow d$	0.13
$\{s_1, s_2\}$	$p_0^0 \rightarrow p_1^1 \rightarrow p_2^2 \rightarrow d$	0.08
$\{s_1, s_3\}$	$p_0^0 \rightarrow p_1^1 \rightarrow p_3^2 \rightarrow d$	0.05
$\{s_2, s_3\}$	$p_0^0 \rightarrow p_2^1 \rightarrow p_3^2 \rightarrow d$	0.11

Table 5: Path choice probabilities given by the nested RL on the MuC graph.

## 7 Numerical Experiments

To evaluate the performance of the two proposed DAGs for the multiple discrete choice problem, we compare them against two logit-based baseline models. These baseline models will be described in the subsequent sections. In this comparison, we focus solely on evaluating our models and approaches against logit-based models. This choice is made because the other existing choice models developed for multiple choice analysis do not align with the RUM and bounded rationality principles. The experiments are conducted using both synthetic instances and two real-world datasets. In the following subsections, we will begin by describing the baseline models. Subsequently, we will present the numerical results conducted with generated instances under various settings. Finally, we will show our experiments conducted using two real-world datasets. All experiments are performed on a Google Cloud TPU v2-8 server with a 96-core Intel Xeon 2.00 GHz CPU.

### 7.1 Baseline models

In the following, we present some simple logit-based baseline models that can be used for the multiple discrete choice problem.

**Single-Choice model (SC-Base).** The first baseline model is formulated by considering each composite alternative as a sequence of multiple independent single choices. As a result, the choice probability of a composite alternative can be computed as the product of single-choice probabilities. Specifically, the choice probability of a composite alternative  $S$  can be computed as

$$P(S) = \prod_{i \in S} P(i|[m]) = \prod_{i \in S} \frac{e^{\frac{1}{\mu} v_i}}{\sum_{i' \in [m]} e^{\frac{1}{\mu} v_{i'}}$$

It can be seen that the SC-Base model always gives lower choice probabilities, compared to the LMDC model. Furthermore, the SC-Base model is unable to incorporate the lower and



upper constraints on the number of selected elemental alternatives, primarily due to its simple structure.

**Multi-Choice with sampled choice sets (MC-Base).** As mentioned earlier, a challenge raised from the use of the LMDC model is the impracticality of exhaustively enumerating the composite alternatives within the choice set  $\Omega^{[L,U]}$ . A simple way to deal with this challenge is to replace  $\Omega^{[L,U]}$  by a small set of sampled alternatives. This modeling approach would share some similarities with path-based models in route choice modeling – since there are exponentially many paths connecting an origin and a destination, one can sample paths to make the estimation practical (Prato, 2009). One limitation of such a sampling-based approach is that the estimation may be not consistent, as it relies on the specific choice sets that are sampled.

Under the MC-Base model, we can sample composite alternatives from  $\Omega^{[L,U]}$  that satisfy the bound constraints (i.e., the size is within  $[L, U]$ ). We then use this sampled choice set to compute the choice probabilities, and to estimate the choice parameters. Specifically, let  $\Gamma$  be a sampled choice set of feasible composite alternatives. We compute the choice probability of a composite alternative  $S \subset [m]$  as

$$P(S|\boldsymbol{\beta}) = \frac{\exp\left(\frac{1}{\mu} \sum_{i \in S} v_i(\boldsymbol{\beta})\right)}{\sum_{S' \in \Gamma} \exp\left(\frac{1}{\mu} \sum_{i \in S'} v_i(\boldsymbol{\beta})\right)}.$$

It can be seen that the MC-Base model always gives a higher choice probability for each composite alternative, compared to that from the LMDC model, as the sampled choice set  $\Gamma$  is just a subset of  $\Omega^{[L,U]}$ . Consequently, if there is an alternative  $S$  that is not included in  $\Gamma$ , the MC-Base will assign it a zero probability. This would be an issue if  $S$  is an actual choice made by the choice-makers.

## 7.2 Synthetic Data

The objective here is to evaluate the losses incurred by employing the two baseline models, SC-Base and MC-Base, when the true model is the LMDC model. Additionally, we aim to measure the estimation times required by different approaches. To this end, we generate various problem instances by taking  $m \in \{5, 10, 15, 20, 30, 50\}$  and vary the upper bound  $U$  as  $U = 3$  for  $m = 5$ ,  $U \in \{3, 5\}$  for  $m = 10$ ,  $U \in \{5, 10\}$  for  $m = 15$ ,  $U \in \{5, 10\}$  for  $m = 20$ ,  $U \in \{10, 20\}$  for  $m = 30$ , and  $U \in \{10, 30\}$  for  $m = 50$ . We also set  $L = 0$  for all the instances. For each choice of  $m$ , we randomly generate three attributes for each elemental alternative. We also choose a vector  $\boldsymbol{\beta}^0$  and use it to define a “ground-truth” model to generate observations. For each choice model (a baseline or the LMDC model), we generate a set of observations for the estimation and use the parameter estimates to compute the log-likelihood of another set of observations. By comparing these likelihood values, we aim to assess the out-of-sample performances of the considered models. For the MC-Base model, the estimation requires sampled choice sets of composite alternatives. To have this, we simply collect subsets from the observations. In Table 6 below we report the average sizes of the generated choice sets used for estimating the MC-Base

Table 6: Average sizes of the sampled choice sets used for estimating the MC-Base model.

Dataset	$m$	5		10		20		30		50	
	$[L, U]$	[0,3]	[0,3]	[0,5]	[0,5]	[0,10]	[0,10]	[0,20]	[0,10]	[0,30]	
Size of sampled choice set		26	176	630	2799	2993	3000	3000	3000	3000	

model.

Table 7: Parameters used for generating synthetic data.

Parameter	Value
Number of attributes	3
1 <sup>st</sup> attribute	log-normal, mean=0, scale=1
2 <sup>nd</sup> attribute	log-normal, mean=0, scale=1
3 <sup>rd</sup> attribute	constant, value=1
$\beta^0$	$[-0.5, -0.02, -0.1]$
Size of the estimation set for SC-Base and MC-Base	3000
Size of the estimation set for RL models	1000
Size of the prediction set	300

Table 7 gives the parameters used for generating the synthetic instances and provides details regarding the sizes of the estimation and prediction sets. Each problem instance is randomly generated, consisting of three attribute values. The first two attributes follow a log-normal distribution with a mean of 0 and a scale of 1. The third attribute remains constant with a value of one. This attribute aims to reflect a natural phenomenon where the likelihood of people choosing a subset decreases as the number of items in that subset increases. In real-world datasets, it has been observed that a significant proportion of customers only purchase 1-2 products at a time.

We generate the synthetic instances as follows. For each group,  $(m, L, U)$ , 300 sets of sample utilities are randomly generated. Then, for each set of alternative utilities, we generate 40 sets of observations using the true model with  $\beta^0$ . Each choice model is then estimated using the generated sets of observations. Each vector of parameter estimates  $\hat{\beta}$  obtained is then used by the true model to calculate the predicted log-likelihood value of the prediction set. The average predicted log-likelihood values and standard errors over 40 runs are reported in Table 8 where the numbers before “ $\pm$ ” are the means and those after “ $\pm$ ” are the standard errors. The last column reports the average log-likelihood values given by the true model (i.e., the LMDC model) with the true parameters  $\beta^0$ . The results clearly show that the average predicted log-likelihood values given by the RL model applied to the BiC and MuC graphs are almost identical

to those returned by the true model, which numerically confirms the equivalence between the LMDC and RL model on the DAGs, shown in Section 5.2. The SC-Base model performs the worst among the 4 models considered, especially when  $m$  and  $U$  are large. On the other hand, the predicted log-likelihood values obtained from the MC-Base model exhibit a high degree of accuracy, particularly for smaller instances (e.g.  $m \leq 10$ ). However, as the value of  $m$  increases, the discrepancies between the predicted log-likelihood values and the true values become more significant. Intuitively, when  $m$  is small, there is a higher likelihood that the sampled choice set  $\Gamma$  used in the MC-Base model closely resembles the universal choice set  $\Omega^{[L,U]}$ . However, as  $m$  grows larger, the gap between the two sets becomes more pronounced.

Table 8: Predicted average log-likelihood values for synthetic data.

Dataset		Model				True model
$m$	$[L, U]$	RL+BiC	RL+MuC	MC-Base	SC-Base	
5	[0,3]	$-2.95 \pm 0.04$	$-2.96 \pm 0.04$	$-2.95 \pm 0.02$	$-6.30 \pm 0.12$	-2.09
10	[0,3]	$-4.71 \pm 0.06$	$-4.70 \pm 0.06$	$-4.70 \pm 0.03$	$-9.87 \pm 0.27$	-3.18
	[0,5]	$-5.98 \pm 0.06$	$-5.99 \pm 0.06$	$-6.01 \pm 0.03$	$-19.85 \pm 0.76$	-4.40
20	[0,5]	$-9.40 \pm 0.07$	$-9.39 \pm 0.07$	$-9.94 \pm 0.02$	$-30.59 \pm 4.68$	-7.32
	[0,10]	$-11.94 \pm 0.08$	$-11.97 \pm 0.11$	$-13.33 \pm 0.00$	$-97.05 \pm 11.98$	-7.36
30	[0,10]	$-16.92 \pm 0.09$	$-16.92 \pm 0.09$	$-17.79 \pm 0.00$	$-60.54 \pm 4.62$	-13.71
	[0,20]	$-19.05 \pm 0.10$	$-19.04 \pm 0.09$	$-20.77 \pm 0.00$	$-189.22 \pm 23.31$	-13.80
50	[0,10]	$-22.63 \pm 0.08$	$-22.66 \pm 0.09$	$-23.32 \pm 0.00$	$-44.61 \pm 2.63$	-19.98
	[0,30]	$-32.17 \pm 0.11$	$-32.18 \pm 0.14$	$-34.60 \pm 0.00$	$-243.53 \pm 43.35$	-23.86

As claimed earlier, one of the main advantages of our approach is that the LMDC model becomes tractable to estimate via dynamic programming. We numerically assess this advantage by reporting the average estimation times required by the four models in Table 9. The SC-Base model is the fastest, as expected, due to its simple structure. The CPU times required by the MC-Base model strongly depend on the size of the sampled choice set and are quite unstable. Moreover, the estimation times for the two RL models on the BiC and MuC graphs vary compared to the MC-Base model. In certain cases, the RL models exhibit slower estimation times, while in other cases, they are faster. Nevertheless, the overall estimation times of the RL model are surprisingly short, even for larger instances with  $m = 50$ . For instance, in cases where  $m = 50$  and  $[L, U] = [0, 30]$ , it takes approximately 1.5 seconds to estimate the RL with the BiC DAG and around 4 seconds with the MuC DAG. Additionally, we observe that the estimation of the RL model on the BiC DAG is faster compared to the MuC DAG. This discrepancy can be attributed to the fact that, although the number of nodes in both graphs is similar, the MuC graph contains more arcs and is denser in structure.

Table 9: Estimation time (in seconds) for synthetic data.

Dataset		Model			
$m$	$[L, U]$	RL+BiC	RL+MuC	MC-Base	SC-Base
5	[0,3]	$0.10 \pm 0.03$	$0.09 \pm 0.02$	$0.52 \pm 0.12$	$0.01 \pm 0.00$
10	[0,3]	$0.14 \pm 0.01$	$0.08 \pm 0.01$	$0.65 \pm 0.10$	$0.02 \pm 0.00$
	[0,5]	$0.14 \pm 0.01$	$0.10 \pm 0.01$	$0.97 \pm 0.10$	$0.04 \pm 0.00$
20	[0,5]	$0.24 \pm 0.02$	$0.19 \pm 0.02$	$1.79 \pm 0.17$	$0.07 \pm 0.01$
	[0,10]	$0.26 \pm 0.02$	$0.30 \pm 0.03$	$2.17 \pm 0.27$	$0.11 \pm 0.01$
30	[0,10]	$0.40 \pm 0.01$	$0.57 \pm 0.03$	$0.56 \pm 0.30$	$0.12 \pm 0.01$
	[0,20]	$0.61 \pm 0.02$	$0.99 \pm 0.07$	$0.69 \pm 0.02$	$0.18 \pm 0.02$
50	[0,10]	$0.67 \pm 0.04$	$1.58 \pm 0.12$	$0.51 \pm 0.01$	$0.13 \pm 0.01$
	[0,30]	$1.50 \pm 0.81$	$3.97 \pm 0.30$	$1.10 \pm 0.03$	$0.29 \pm 0.04$

### 7.3 Real-world Datasets

In this section, we compare the performances of different logit-based models based on two datasets extracted from real-world data sources. We also consider the two logit-based baselines – the SC-Base and MC-Base models. In-sample results (i.e., estimation results) and out-of-sample results will be reported for each dataset. In the out-of-sample evaluation, the observations are randomly split into two sets with a predetermined ratio. Specifically, 80% of the observations are utilized for parameter estimation, while the remaining 20% serve as a holdout set to assess the predictive performance of the estimated parameters. We generate 40 estimation-holdout pairs and calculate the average log-likelihood values for each model, based on the universal choice set of composite alternatives (i.e., the LMDC model), which will be reported as the evaluation metric.

**Jewelry-Store dataset.** This dataset<sup>1</sup> contains purchase data from December 2018 to December 2021 (3 years) from a medium-sized jewelry online store, collected by Open CDP project (REES46, 2021). Each jewelry product has three attributes: *Quantity* (total number of that product sold), *Price*, *Popularity* (number of unique customers who purchased that product), and *Gender* (which gender for which the product is designed). We also include a constant attribute which takes values of 1 for all the products. This attribute will capture the number of selected products in each composite observation. Moreover, in order to enhance the accuracy of the simulation and prioritize significant products, we narrow our focus to data collected during the initial two years. Furthermore, we only consider products that have sold a minimum of 50 units. In total, there are 436 items, and the dataset consists of 34,291 observations. It is worth noting that approximately 93% of the orders comprise no more than 5 products, while

<sup>1</sup><https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewelry-store>

any orders exceeding this threshold are treated as anomalous ones. As the dataset does not provide information regarding the lower and upper bounds on the number of selected items that customers consider when making choices, we attempt to infer these bounds based on the observations. That is, we partition the observations into three categories that correspond to customers purchasing 1-2 items, 3-5 items, and more than 5 items. This yields three sets of  $(L, U)$  values, which are used to construct the DAGs. For the LMDC model, we utilize the BiC DAG, noting that the MuC DAG will yield the same estimation results. For the MC-Base baseline, we collect subsets from the observations to form the choice set  $\Gamma$ , giving us a consideration set consisting of 3487 composite alternatives. The estimation outcomes are presented in Table 10. From the results, we can see that the parameter estimates for *Quantity* and *Price* seem to have expected signs. Moreover, parameter estimates for *Quantity* and *Const* are highly significant, while the estimates for *Price* are not. This could be because the products have similar prices to each other, leading customers to prioritize the total quantity of items they can buy based on their budgets. Based on intuition, as the *Const* attribute represents the number of purchased items, it is expected that its estimates would have negative signs. This remark suggests that the parameter estimates of *Const* provided by MC-Base and LMDC have the correct signs, whereas this is not the case for SC-Base. Finally, it is important to highlight that the log-likelihood values, obtained by utilizing the corresponding parameter estimates with the universal choice set  $\Omega^{[L,U]}$ , are presented in the final column of Table 10. These values serve as an indication that the LMDC model significantly outperforms other models in terms of in-sample fit. Here, it is important to note that the final log-likelihood values obtained through Maximum Likelihood Estimation (MLE) are not reported since they are computed based on different choice sets and therefore cannot be directly compared. Instead, to ensure a meaningful in-sample fit comparison, we employ the parameter estimates to calculate the log-likelihood values using the same choice set of composite alternatives, which is the universal choice set.

Table 10: Estimation results for the Jewelry-Store dataset.

Model	Attribute				$\mathcal{L}(\hat{\beta})$	
	Quantity	Price	Gender	Const		
	$\hat{\beta}$	2.352	-0.143	0.005	9.136	
SC-Base	Std. Err.	0.021	0.045	0.016	8.075	-4,967,533
	$t$ -test(0)	111.925	-3.181	0.341	1.131	
	$\hat{\beta}$	1.667	-0.181	-0.059	-3.994	
MC-Base	Std. Err.	0.022	0.062	0.016	0.022	-503,396
	$t$ -test(0)	74.263	-2.897	-3.686	-181.578	
	$\hat{\beta}$	2.363	-0.144	0.005	-6.527	
LMDC	Std. Err.	0.024	0.062	0.018	0.019	-270,061
	$t$ -test(0)	98.704	-2.312	0.280	-343.973	

Table 11 reports the out-of-sample evaluation for the Jewelry-Store dataset. Similar to the synthetic data, we split the observation set into two parts, one part taking 80% of the observations is utilized for parameter estimation, while the second part taking 20% serves as a holdout set to assess the prediction performance of the estimated parameters. We estimate each model considered (LMDC, SC-Base or MC-Base) using the estimation set, and employ the obtained parameters to calculate the log-likelihood value of the holdout set based on the universal choice set of composite alternatives  $\Omega^{[L,U]}$ . Because the data does not provide information about the lower and upper limits of customers’ purchasing items, and these limits may vary for different customers, we categorize the observations based on the actual number of products purchased. For example, in Table 11, the first row presents the prediction results solely based on observations of a single selected item, and the row indicating  $[L, U] = [1, 11]$  reports results for observations where the number of purchased products ranges from 1 to 11. We indicate in bold the best log-likelihood value in each row.

It can be seen, from Table 11, that when  $L = U$ , i.e., the number of purchased items is fixed, the predicted log-likelihood values are similar across the three models. When we increase the range  $[L, U]$ , the LMDC model begins to outperform other models, and the SC-Base becomes the worst. Overall, we see that the LMDC model provides significantly better prediction results compared to the two baselines.

Table 11: Out-of-sample results for the Jewelry-Store dataset

$[L, U]$	Number of Observations	LMDC	MC-Base	SC-Base
[1,1]	30286	<b>-5.97</b>	<b>-5.97</b>	<b>-5.97</b>
[2,2]	3150	<b>-11.22</b>	-11.41	<b>-11.22</b>
[3,3]	615	<b>-16.18</b>	-16.45	<b>-16.18</b>
[4,4]	150	<b>-20.98</b>	-21.12	<b>-20.98</b>
[5,5]	56	<b>-24.92</b>	-25.58	<b>-24.92</b>
[1,2]	33436	<b>-6.78</b>	-7.91	-19.85
[3,5]	821	<b>-18.33</b>	-25.56	-43.85
[6,10]	33	<b>-31.79</b>	-45.59	-87.09
[1,11]	34291	<b>-7.24</b>	-16.03	-149.24

**Book-Crossing dataset.** The dataset<sup>2</sup> were crawled in 4 weeks by Cai-Nicolas Ziegler and coauthors from BookCrossing<sup>3</sup>, surveying books purchased and customer ratings from Amazon Web Services. It is used as the benchmark data in the publication by the same authors (Ziegler et al., 2005). Each book has three attributes: *overall rating* from customers (only count explicit ones), *popularity* (total number of customer ratings, both explicit and implicit), and its *lifespan*

<sup>2</sup><http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

<sup>3</sup><http://www.bookcrossing.com>

(from the first time it is published to the date it was crawled). To capture the number of selected items, we also include an additional attribute that is equal to 1 for every book. We name this attribute as “*Const*”. The observations don’t include some outlier samples where customers were unidentified or had bugged data. There are 100 books, the data consists of 16,989 observations, and the number of selected items varies from 1 to 70. For the MC-Base baseline, we gather subsets from the observations to construct the choice set  $\Gamma$ . This results in a consideration set comprising 4886 composite alternatives.

Table 12: Estimation results for the Book-Crossing dataset

Model		Attribute				$\mathcal{L}(\hat{\beta})$
		Rating	Popularity	Lifespan	Const	
SC-Base	$\hat{\beta}$	0.791	2.677	-0.079	11.496	-
	Std. Err.	0.030	0.028	0.019	12.085	
	$t$ -test(0)	26.502	94.617	-4.095	0.951	
MC-Base	$\hat{\beta}$	-0.145	0.517	0.115	-0.533	-789,434
	Std. Err.	0.037	0.033	0.018	0.032	
	$t$ -test(0)	-3.904	15.706	6.392	-16.775	
LMDC	$\hat{\beta}$	0.777	2.826	-0.078	-4.464	-195,208
	Std. Err.	0.047	0.040	0.022	0.038	
	$t$ -test(0)	16.590	71.192	-3.625	-118.189	

The estimation results are reported in Table 12 which show that most of the parameter estimates are significantly different from zero, except the estimate of *Const* of the SC-Base. It is reasonable to anticipate that the sign of the *Rating* parameter should be negative, as higher ratings should positively affect the attractiveness of a book. Consequently, the parameter estimates for *Rating* exhibit correct signs in the LMDC and SC-Base models, while they display incorrect signs in the MC-Base model. The parameter estimates for *Popularity* are, as expected, positive for all the models. The estimates for the attribute *Const* provided by the LMDC model are negative and highly significant. In contrast, the estimate given by the MC-Base model is less significant, and the estimate given by the SC-Base model is even positive, which contradicts our prior intuition regarding the effect of the number of selected items on the item utilities. The log-likelihood values computed based on the corresponding parameter estimates and the universal choice set are reported in the last column of Table 12, where the value for the SC-Base are too small to be numerically reported, so we indicate it by “-”. Overall, we can see that the LMDC remarkably outperforms the other baselines, in terms of in-sample fit.

For the out-of-sample evaluation, similar to the Jewelry dataset, we split the set of observations to different subsets according to the values of  $[L, U]$ . The predicted log-likelihood values are reported in Table 13 and we indicate in bold the best values across the three models considered. It can be observed that when  $L = U$ , i.e., the number of selected books is fixed, the predicted

Table 13: Out-of-sample results for the Book-Crossing dataset

$[L, U]$	Number of observations	LMDC	MC-Base	SC-Base
[1, 1]	11331	<b>-4.34</b>	<b>-4.34</b>	<b>-4.34</b>
[2, 2]	2275	<b>-8.29</b>	-8.39	<b>-8.29</b>
[3, 3]	1000	<b>-11.69</b>	-12.00	<b>-11.69</b>
[4, 4]	575	<b>-14.83</b>	-15.18	-14.84
[5, 5]	342	<b>-17.68</b>	-18.14	<b>-17.68</b>
[6, 6]	256	<b>-20.47</b>	-20.90	<b>-20.47</b>
[7, 7]	191	<b>-22.99</b>	-23.50	-23.00
[8, 8]	144	<b>-25.51</b>	-25.95	<b>-25.51</b>
[9, 9]	104	<b>-27.83</b>	-28.27	<b>-27.83</b>
[2, 9]	4887	<b>-14.10</b>	-27.84	-99.70
[10, 20]	518	<b>-39.34</b>	-48.01	-126.50
[1, 70]	16986	<b>-11.51</b>	-47.55	–

log-likelihood values are quite similar across the three models. As the range of  $[L, U]$  expands, the LMDC model begins to significantly outperform the other models. In particular, when  $[L, U] = [1, 70]$ , the average predicted log-likelihood value given by LMDC is -11.51, which is much larger than the value returned by the MC-Base. Moreover, in this case, the predicted log-likelihood of the SC-Base is even too small to be numerically reported and we indicate it by “–”. In general, similar to the findings observed in the case of the Jewelry dataset, the LMDC model significantly performs better than the other baselines in our out-of-sample evaluation.

#### 7.4 Nested LMDC with Real-World Datasets

As discussed earlier, the nested RL model can be applied to the DAGs to capture the complex correlation between composite alternatives’ utilities. This section presents experimental evaluations of the nested RL model using the two aforementioned real-world datasets. It is worth mentioning that, unlike the RL model, the application of the nested RL to the BiC and MuC DAGs offers two distinct models. Therefore, the results obtained from both BiC and MuC graphs will be presented and compared. These models are referred to as nested LMDC models.

We employ the Jewelry-Store and Book-Crossing datasets, as in the previous experiments. The attributes used to define the utility functions in the RL model are retained for this experiment as well. The scale functions  $\mu_k$ , for all  $k \in \mathcal{N}$ , are important components of the nested RL model and we specify them as follows. For the Jewelry-Store dataset, for each node  $p_i^c$  of the DAG, we define scale function  $\mu_{p_i^c}$  as exponential functions of three attributes: *Quantity*, *Const*,



and *Count*, where *Quantity* is the quantity of item  $i \in [m]$ , *Const* is a constant attribute taking values of 1, and *Count* is equal to  $c$  to capture the number of selected items at that node. The use of exponential functions to model the scales is to guarantee that  $\mu_k > 0$  for all  $k \in \mathcal{N}$ , as suggested by [Mai et al. \(2015\)](#). Regarding the Book-Crossing data, the scale functions are specified using two attributes: *Const* and *Count*.

Table 14: Estimation results of nested LMDC models for the Jewelry-Store dataset

DAG	Property	Utility function				Scale function			$\mathcal{L}(\hat{\beta})$
		Quantity	Price	Gender	Const	Quantity	Const	Count	
BiC DAG	$\hat{\beta}$	0.785	0.066	0.054	-4.062	0.451	-0.308	-0.294	-257,657
	Std. Err.	0.041	0.078	0.015	0.052	0.006	0.013	0.002	
	$t$ -test(0)	19.088	0.842	3.503	-77.677	71.525	-24.027	-129.013	
MuC DAG	$\hat{\beta}$	1.924	-0.016	0.049	-3.043	-0.146	-1.115	0.074	-249,290
	Std. Err.	0.027	0.044	0.011	0.032	0.010	0.011	0.001	
	$t$ -test(0)	71.153	-0.365	4.546	-94.506	-14.600	-98.057	68.516	

Table 15: Parameter estimation of the nested LMDC models for the Book-Crossing dataset.

Model	Property	Utility function				Scale function		$\mathcal{L}(\hat{\beta})$
		rating	Popularity	Age	const	Const	Count	
BiC DAG	$\hat{\beta}$	-1.392	2.227	0.269	-4.505	0.232	0.034	-181,020
	Std. Err.	0.008	0.008	0.008	0.008	0.002	0.000	
	$t$ -test(0)	-181.955	289.699	35.019	-597.270	105.135	1255.165	
MuC DAG	$\hat{\beta}$	0.727	1.670	-0.121	-2.747	-0.691	0.034	-177,907
	Std. Err.	1.260	0.061	0.270	1.721	0.279	0.001	
	$t$ -test(0)	0.577	27.474	-0.449	-1.596	-2.477	31.051	

The estimation results of the Jewelry-Store and Book-Crossing datasets are reported in Tables 14 and 15. For the Jewelry-Store dataset, similar to the observations we have in the MDLC model, the estimates for *Quantity* and *Count* are significant and have their expected signs, while it is not the case for the estimates of *Price* and *Gender*. The parameter estimates for the three attributes *Quantity*, *Const* and *Count* are also significant. The final log-likelihood values are presented in the last column of Table 14. These values indicate that the nested RL model exhibits superior performance in terms of in-sample fit when applied to the MuC DAG compared to the BiC DAG.

For the Book-Crossing dataset, the results reported in Table 15 show that all the parameter estimates given by the nested RL model with the BiC DAG are all significant, but the sign of the estimate of *Rating* is negative, which is intuitively incorrect. On the other hand, the estimates given by the MuC DAG seem to have their expected signs, but they are only significant for *Popularity* of the utility functions, and *Count* of the scale functions. The MuC graph also

gives a better in-sample log-likelihood value. In general, we can see that the nested LMDC model based on the MuC DAG offers better performance in terms of in-sample fit. Intuitively, this would be due to the fact that each node in the MuC DAG has several outgoing arcs connecting it with other subsequent nodes while each node in the BiC DAG only has 2 outgoing arcs corresponding to a “selecting” or “not-selecting” option, thus the nested RL on the MuC would be better in capturing the correlation between alternatives.

We now switch ourselves to the out-of-sample evaluation. Table 16 reports the average predicted log-likelihood values for both Jewelry-Store and Book-Crossing datasets. For the sake of comparison, we include the average predicted log-likelihood value given by the LMDC model reported previously. The results are computed based on the entire observation sets we have, which range in size from 1 to 70 for the Jewelry-Store dataset and from 1 to 11 for the Book-Crossing dataset. The results clearly indicate that the nested LMDC models, utilizing either the BiC or MuC graphs, outperform the LMDC model for both datasets. Furthermore, the nested LMDC model consistently exhibits better performance when applied to the MuC DAG compared to the BiC DAG.

Table 16: Prediction results of the nested LMDC models for the Jewelry-Store and Book-Crossing datasets.

Dataset	$[L, U]$	LMDC	Nested LMDC	
			BiC DAG	MuC DAG
Jewelry-Store	[1,11]	-7.24	-7.14	-7.15
Book-Crossing	[1,70]	-11.51	-10.37	-10.28

In summary, our numerical results based on the two real-world datasets indicate several advantages of our modeling and estimation approaches. Firstly, the LMDC model performs better than the other simple baselines (SC-Base and MC-Base) in terms of both in-sample and out-of-sample fits. In addition, when the nested RL is brought into the context, we also observe that it significantly improves the LMDC model in both in-sample and out-of-sample fits.

Finally, we report in Table 17 the estimation times of the LMDC and nested LMDC models, for both Jewelry-Store and Book-Crossing datasets. As expected, the computing times of the LMDC model are significantly smaller than those of the nested model. For the LMDC, the estimation times are quite similar across the BiC and MuC graphs. On the other hand, for the nested LMDC models, the MuC DAG requires less computing time, compared to the BiC DAG, even though the MuC is denser. This would be because the nested RL with the MuC fits better with the data, thus the MLE requires fewer iterations to converge. In general, for both real datasets with reasonably large numbers of items and observations, our estimation method manages to estimate the LMDC model in less than 6 minutes and estimate the nested LMDC models in about 1 hour (slightly more or less). This is possible thanks to the use of several techniques and findings employed to estimate the RL and nested RL model mentioned above. In particular, all the computations are formulated as *sparse matrix operations*, which also greatly

speeds up the estimation.

Table 17: Estimation times (in seconds) for the Jewelry-Store and Book-Crossing datasets .

Dataset	$m$	$[L, U]$	No observations	LMDC		Nested LMDC	
				BiC DAG	MuC DAG	BiC DAG	MuC DAG
Jewelry-Store	436	[1,11]	34291	333	332	4134	2009
Book-Crossing	100	[1,70]	16986	53	70	1,985	926

## 8 Conclusion

In this paper, we have conducted a study on models based on RUM for analyzing multiple discrete choices. To address the challenge of having exponentially many alternatives in composite choice sets, we have introduced an innovative approach based on network representations and recursive route choice models. We have proposed two DAGs to capture the multiple discrete choice processes. Furthermore, we have demonstrated that by applying the RL model to these DAGs, we can obtain an equivalent model, implying that the estimation of the LMDC model can be achieved by estimating RL models on the DAGs.

We have also delved into the estimation of the RL model on the DAGs and established important results, such as the polynomial-time computation of the value function using value iteration. Additionally, we have shown that the nested RL approach can be employed to relax the IIA limitation of the LMDC model and capture the correlation between composite alternatives. Theoretical analysis and examples have revealed that applying the nested RL to the BiC and MuC DAGs leads to two distinct models. However, under certain conditions of the scale function, these resulting models can be equivalent.

Our numerical results, based on both synthetic and real datasets, have demonstrated the superior performance of our LMDC model compared to other baselines in terms of in-sample and out-of-sample fits. Remarkably, the estimation of the RL model on the DAGs can be done rapidly, with estimation times within seconds even for instances involving 50 items. We have also provided numerical evaluations using real datasets for the nested LMDC model, which have indicated its improved performance over the LMDC model in both in-sample and out-of-sample fits.

Our research provides innovative and efficient methods to tackle the multi-discrete choice problem through the utilization of route choice models, which opens up several promising avenues for further exploration. For instance, it would be interesting to extend our approach to problems that involve continuous variables or complex behavioral constraints. Additionally, leveraging models and techniques from the route choice literature could prove valuable in addressing challenges related to data availability, such as missing or noisy data.

## Acknowledgments

The work is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2- RP-2020-017).

## References

- Baillon, J.-B. and Cominetti, R. Markovian traffic equilibrium. *Mathematical Programming*, 111(1-2):33–56, 2008.
- Ben-Akiva, M. *The structure of travel demand models*. PhD thesis, MIT, 1973.
- Ben-Akiva, M. and Bierlaire, M. Discrete choice methods and their applications to short term travel decisions. Chapter for the Transportation Science Handbook, Preliminary Draft, MIT, March 1999.
- Bhat, C. R. A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological*, 39(8): 679–707, 2005.
- Bhat, C. R. The multiple discrete-continuous extreme value (mdcev) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B: Methodological*, 42(3):274–303, 2008.
- Bhat, C. R. A closed-form multiple discrete-count extreme value (mdcntev) model. *Transportation Research Part B: Methodological*, 164:65–86, 2022.
- Bhat, C. R., Castro, M., and Pinjari, A. R. Allowing for complementarity and rich substitution patterns in multiple discrete–continuous models. *Transportation Research Part B: Methodological*, 81:59–77, 2015.
- Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.
- Campitelli, G. and Gobet, F. Herbert simon’s decision-making approach: Investigation of cognitive processes in experts. *Review of general psychology*, 14(4):354–364, 2010.
- Castro, M., Bhat, C. R., Pendyala, R. M., and Jara-Díaz, S. R. Accommodating multiple constraints in the multiple discrete–continuous extreme value (mdcev) choice model. *Transportation Research Part B: Methodological*, 46(6):729–743, 2012.
- Daly, A. and Bierlaire, M. A general and operational representation of generalised extreme value models. *Transportation Research Part B*, 40(4):285 – 305, 2006.
- de Moraes Ramos, G., Mai, T., Daamen, W., Frejinger, E., and Hoogendoorn, S. Route choice behaviour and travel information in a congested network: Static and dynamic recursive models. *Transportation Research Part C: Emerging Technologies*, 114:681–693, 2020.

- Dubé, J.-P. Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Science*, 23(1):66–81, 2004.
- Fosgerau, M., Frejinger, E., and Karlström, A. A link based network route choice model with unrestricted choice set. *Transportation Research Part B*, 56:70–80, 2013.
- Hendel, I. Estimating multiple-discrete choice models: An application to computerization returns. *The Review of Economic Studies*, 66(2):423–446, 1999.
- Howell, J. R., Lee, S., and Allenby, G. M. Price promotions in choice models. *Marketing Science*, 35(2):319–334, 2016.
- Kim, J., Allenby, G. M., and Rossi, P. E. Modeling consumer demand for variety. *Marketing Science*, 21(3):229–250, 2002.
- Lee, S. and Allenby, G. M. Modeling indivisible demand. *Marketing Science*, 33(3):364–381, 2014.
- Lee, S., Kim, S., and Park, S. A sequential choice model for multiple discrete demand. *Quantitative Marketing and Economics*, 20(2):141–178, 2022.
- Ma, J. and Ye, X. Modeling household vehicle ownership in emerging economies. *Journal of the Indian Institute of Science*, 99:647–671, 2019.
- Mai, T. A method of integrating correlation structures for a generalized recursive route choice model. *Transportation Research Part B: Methodological*, 93:146–161, 2016.
- Mai, T. and Frejinger, E. Undiscounted recursive path choice models: Convergence properties and algorithms. *Transportation Science*, 56(6):1469–1482, 2022.
- Mai, T., Fosgerau, M., and Frejinger, E. A nested recursive logit model for route choice analysis. *Transportation Research Part B*, 75(0):100 – 112, 2015. ISSN 0191-2615.
- Mai, T., Frejinger, E., Fosgerau, M., and Bastin, F. A dynamic programming approach for quickly estimating large network-based mev models. *Transportation Research Part B*, 98:179–197, 2017.
- Mai, T., Bastin, F., and Frejinger, E. A decomposition method for estimating recursive logit based route choice models. *EURO Journal on Transportation and Logistics*, 7(3):253–275, 2018.
- Mai, T., Yu, X., Gao, S., and Frejinger, E. Route choice in a stochastic time-dependent network: the recursive model and solution algorithm. *Transportation Research Part B: Methodological*, 151:42, 2021.
- McFadden, D. Econometric models of probabilistic choice. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*, chapter 5, pages 198–272. MIT Press, 1981.

- Melo, E. A representative consumer theorem for discrete choice models in networked markets. *Economics Letters*, 117(3):862–865, 2012.
- Mouter, N., Koster, P., and Dekker, T. Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transportation research part A: policy and practice*, 144:54–73, 2021.
- Oyama, Y. and Hato, E. A discounted recursive logit model for dynamic gridlock network analysis. *Transportation Research Part C: Emerging Technologies*, 85:509–527, 2017.
- Prato, C. G. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2:65–100, 2009.
- REES46. eCommerce purchase history from jewelry store. <https://rees46.com/en/open-cdp>, 2021.
- Rust, J. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033, 1987.
- Satomura, T., Kim, J., and Allenby, G. M. Multiple-constraint choice models with corner and interior solutions. *Marketing Science*, 30(3):481–490, 2011.
- Sent, E.-M. Rationality and bounded rationality: You can’t have one without the other. *The European Journal of the History of Economic Thought*, 25(6):1370–1386, 2018.
- Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.
- van der Lans, R. A simultaneous model of multiple-discrete choices of variety and quantity. *International Journal of Research in Marketing*, 35(2):242–257, 2018.
- Varghese, V. and Jana, A. Multitasking during travel in mumbai, india: Effect of satiation in heterogeneous urban settings. *Journal of Urban Planning and Development*, 145(2):04019002, 2019.
- Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.
- Zimmermann, M. and Frejinger, E. A tutorial on recursive models for analyzing and predicting path choice behavior. *EURO Journal on Transportation and Logistics*, 9(2):100004, 2020.
- Zimmermann, M., Frejinger, E., and Marcotte, P. A strategic markovian traffic equilibrium model for capacitated networks. *Transportation Science*, 55(3):574–591, 2021.

# Appendix

## A Missing Proofs

### A.1 Proof of Proposition 3

The complementary counting approach is used to determine the number of nodes. First, we count nodes in a complete BiC DAG without any restriction on the subset size. Each tier  $j$  in the first  $m + 1$  middle tiers has exactly  $j + 1$  nodes:  $\{p_j^0, p_j^1, \dots, p_j^j\}$ . Add one node  $d$  from the final tier, the total number of nodes is

$$|\mathcal{N}_{full}| = \sum_{j=0}^m (j + 1) + 1 = \frac{m^2}{2} + \frac{3m}{2} + 2$$

Next, we subtract nodes of form  $p_j^c$  that are lost due to violation the constraint  $c \leq U$ :

$$|\mathcal{N}_U| = \sum_{j=U+1}^m (j - U) = \frac{m^2}{2} + \frac{m}{2} - mU + \frac{U^2}{2} - \frac{U}{2}$$

Aggregation of these two components leads to the final node count in the BiC DAG:

$$\begin{aligned} |\mathcal{N}| &= |\mathcal{N}_{full}| - |\mathcal{N}_U| \\ &= m + mU - \frac{U^2}{2} + \frac{U}{2} + 2. \end{aligned}$$

For the number of arcs, we see that most nodes have 2 outgoing arcs, except  $U$  nodes belonging to the  $m$ -th tier and the absorbing node  $d$ . The first  $L - 1$  nodes of the  $m$ -th tier have no arc because they don't satisfy minimum requirement  $L$ . The remaining  $U - L + 1$  nodes have exactly 1 arc leading to  $d$ . Another exception is the absorbing state  $d$  which has no arc. As a result, the number of arcs in a BiC DAG is

$$\begin{aligned} |\mathcal{A}| &= 2(|\mathcal{N}| - U - 1) + (U - L + 1) \\ &= 2m + 2mU - U^2 - L + 1. \end{aligned}$$

### A.2 Proof of Proposition 5

The node structure in MuC DAG is similar to BiC DAG's, except all  $m$  nodes of form  $p_j^0$  are lost:

$$|\mathcal{N}| = mU - \frac{U^2}{2} + \frac{U}{2} + 2$$

For the number of arcs, we first consider  $L = 0$  as a special case when counting the number of arcs in the graph. Another interesting observation is that the total number of arcs in the cases of  $L = 0$  and  $L = 1$  differs by exactly one edge connecting the source node to the destination node. Therefore, in the subsequent computations, we temporarily consider  $L \geq 1$ .

We denote the superscript  $c$  of  $p_j^c$  as their “depths”. Due to the size constraint, the depth of a node cannot exceed  $U$ . Arcs in the MuC DAG can be divided into three categories based on depths: arcs start from the source node  $p_0^0$ , outgoing arcs from nodes at depths  $c < L$ , and arcs from the remaining nodes (from depth  $L$  to  $U$ ). The edge count in each category is determined as follows:

- Arcs from the node  $p_0^0$ : In case  $L \geq 1$ , there are only  $|\mathcal{A}_1| = m$  arcs from  $p_0^0$  to nodes at depth  $c = 1$ :
- Arcs from nodes at depths  $c < L$ : Each depth  $c$  contains  $m - c + 1$  nodes. Each node  $p_j^c$  at depth  $c$  has arcs leading to some nodes of form  $p_{j'}^{c+1}$  where  $j < j' \leq m$ . In conclusion, the edge count in this scenario is

$$|\mathcal{A}_2| = \sum_{c=1}^{L-1} \sum_{j=c}^m \sum_{j'=j+1}^m 1$$

- Arcs from nodes at depths  $c \in [L, U]$ : Each depth  $c$  contains  $m - c + 1$  nodes, and each node  $p_j^c$  has arcs leading to nodes of the form  $p_{j'}^{c+1}$  where  $j < j' \leq m$  plus one arc to the final node  $d$ . The total arc count is

$$|\mathcal{A}_3| = \sum_{c=L}^U \sum_{j=c}^m \left( \sum_{j'=j+1}^m 1 + 1 \right)$$

So, the total number of arcs in the MuC DAG is:

$$\begin{aligned} |\mathcal{A}| &= |\mathcal{A}_1| + |\mathcal{A}_2| + |\mathcal{A}_3| \\ &= \frac{m^2 U}{2} - \frac{m U^2}{2} - mL + mU + 2m + \frac{L^2}{2} - \frac{3L}{2} + \frac{U^3}{6} - \frac{U^2}{2} + \frac{U}{3} + 1. \end{aligned}$$

We complete the proof.

### A.3 Proof of Proposition 10

We first prove the following lemma showing that the value functions for both BiC and MuC are identical.

**Lemma 1** *Under the assumptions of Proposition 5, the value functions are the same for both graphs, i.e., for two nodes  $k \in \mathcal{N}^{\text{BiC}}, k' \in \text{MuC}$  such that  $k$  and  $k'$  are both of the same form  $p_i^c$ ,*



for  $c \in [0, U], i \in 0, \dots, m$ , then

$$\tilde{V}^{\text{BiC}}(k) = \tilde{V}^{\text{MuC}}(k').$$

**Proof.** For notational simplification, we write  $\tilde{V}^{\text{BiC}}(p_j^c)$  to denote the value function at the node of form  $p_i^c$  in the BiC graph, and  $\tilde{V}^{\text{MuC}}(p_j^c)$  denotes the value function at the node of form  $p_j^c$  in the MuC graph. Moreover, from the assumption that the scale function at node  $p_j^c$  is only dependent on  $c$ , to facilitate the later exposition, let us define  $\mu^*(c)$ , for any  $c \in [0, U]$  such that  $\mu_{p_j^c} = \mu^*(c)$ . Moreover, to simplify the proofing process, we assume there is no bounding constraint related to the number of selected alternatives. This means there is no arc removed due to constraint conflicts in both DAGs. Given the stated conditions, we will prove the equality between  $\tilde{V}^{\text{BiC}}(p_j^k)$  and  $\tilde{V}^{\text{MuC}}(p_j^k)$  by induction through all  $m + 2$  tiers of nodes in reversed order (from Tier  $m + 1$  to Tier 0) in both BiC DAG and MuC DAG.

First, the result holds for the absorbing state  $d$  of the final tier since  $\tilde{V}^{\text{BiC}}(d) = \tilde{V}^{\text{MuC}}(d) = 0$ . Afterwards, we assume  $\tilde{V}^{\text{BiC}}(p_{j'}^k) = \tilde{V}^{\text{MuC}}(p_{j'}^k)$  for all nodes lying from tier  $j + 1$  to the final tier  $m + 1$ . Our goal is to prove this statement for tier  $i$ :  $\tilde{V}^{\text{BiC}}(p_j^k) = \tilde{V}^{\text{MuC}}(p_j^k)$ .

In the nested BiC DAG, each node  $p_j^k$  leads to two following nodes  $p_{j+1}^k$  and  $p_{j+1}^{k+1}$  (except nodes only leading to  $d$ ). Based on [Equation 10](#), the expected maximum utility of each node  $p_j^k$  can be recursively derived from the system

$$\exp\left(\frac{1}{\mu_{p_j^k}} \tilde{V}^{\text{BiC}}(p_j^k)\right) = \begin{cases} 1, & j = m, k \in [L, U] \\ 0, & j = m, k \notin [L, U] \\ \exp\left(\frac{1}{\mu_{p_j^k}} (v^{\text{BiC}}(p_{j+1}^k | p_j^k) + \tilde{V}^{\text{BiC}}(p_{j+1}^k))\right) \\ \quad + \exp\left(\frac{1}{\mu_{p_j^k}} (v^{\text{BiC}}(p_{j+1}^{k+1} | p_j^k) + \tilde{V}^{\text{BiC}}(p_{j+1}^{k+1}))\right) & 0 \leq j < m \end{cases}$$

Since the arc  $(p_j^k, p_{j+1}^k)$  represents the “not selecting” option,  $v^{\text{BiC}}(p_{j+1}^k | p_j^k) = 0$ . The other arc provides the equation  $v^{\text{BiC}}(p_{j+1}^k | p_j^k) = v_{j+1}$ . On the other hand,  $\mu_{p_j^k}$  can be replaced with  $\mu^*(k)$  from the given statements. For  $p_j^k$  nodes where  $j < m$ , the equation becomes

$$\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_j^k)\right) = \exp\left(\frac{1}{\mu^*(k)} (v_{j+1} + \tilde{V}^{\text{BiC}}(p_{j+1}^k))\right) \\ + \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j+1}^k)\right)$$

Apply the above formulation on node  $p_{j+1}^k$ , we get

$$\begin{aligned} \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j+1}^k)\right) &= \exp\left(\frac{1}{\mu^*(k)}(v_{j+2} + \tilde{V}^{\text{BiC}}(p_{j+2}^{k+1}))\right) \\ &\quad + \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j+2}^k)\right) \end{aligned}$$

which leads to the following if combined with the previous equation

$$\begin{aligned} \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_j^k)\right) &= \exp\left(\frac{1}{\mu^*(k)}(v_{j+1} + \tilde{V}^{\text{BiC}}(p_{j+1}^{k+1}))\right) \\ &\quad + \exp\left(\frac{1}{\mu^*(k)}(v_{j+2} + \tilde{V}^{\text{BiC}}(p_{j+2}^{k+1}))\right) \\ &\quad + \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j+2}^k)\right) \end{aligned}$$

This replacement action can be recursively applied until a determinable state is reached

$$\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_m^k)\right) = \begin{cases} 1, & k \in [L, U] \\ 0, & k \notin [L, U] \end{cases}$$

Therefore, the final equation achieved is

$$\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_j^k)\right) = \begin{cases} \sum_{j'=j+1}^m \exp\left(\frac{1}{\mu^*(k)}(v_{j'} + \tilde{V}^{\text{BiC}}(p_{j'}^{k+1}))\right) + 1, & k \in [L, U] \\ \sum_{j'=j+1}^m \exp\left(\frac{1}{\mu^*(k)}(v_{j'} + \tilde{V}^{\text{BiC}}(p_{j'}^{k+1}))\right), & k \notin [L, U] \end{cases}$$

Fortunately, according to the DAG topology in [Figure 2](#), expected utility values in the MuC DAG follow a similar system

$$\exp\left(\frac{1}{\mu_{p_j^k}^*} \tilde{V}^{\text{MuC}}(p_j^k)\right) = \begin{cases} 1, & j = m, k \in [L, U] \\ 0, & j = m, k \notin [L, U] \\ \sum_{j'=j+1}^m \exp\left(\frac{1}{\mu^*(k)}(v_{j'} + \tilde{V}^{\text{MuC}}(p_{j'}^{k+1}))\right) + 1, & 0 \leq j < m, k \in [L, U] \\ \sum_{j'=j+1}^m \exp\left(\frac{1}{\mu^*(k)}(v_{j'} + \tilde{V}^{\text{MuC}}(p_{j'}^{k+1}))\right), & 0 \leq j < m, k \notin [L, U] \end{cases}$$

The only exceptions are nodes of form  $p_j^0$ ,  $\forall j | 0 < j \leq m$  which don't exist in MuC DAG. Nevertheless, when  $\tilde{V}^{\text{BiC}}(p_{j'}^k) = \tilde{V}^{\text{MuC}}(p_{j'}^k)$ ,  $\forall j' | j < j' \leq m$ , it is easy to also point out the equality of nodes in tier  $i$  also holds:  $\tilde{V}^{\text{BiC}}(p_j^k) = \tilde{V}^{\text{MuC}}(p_j^k)$ . As such, the hypothesis is successfully proved by induction.

■

We are now to prove that the probability distribution over paths given by both BiC and MuC are also the same. To this end, we see that the probability of each path is the product of all arcs' probabilities. In both nested BiC and MuC models, the probability of an arc from state  $k$  to state  $a$  can be rewritten as:

$$\tilde{P}(a|k) = \frac{\exp\left(\frac{1}{\mu_k}v(a|k)\right) \exp\left(\frac{1}{\mu_k}\tilde{V}(a)\right)}{\exp\left(\frac{1}{\mu_k}\tilde{V}(k)\right)}.$$

Next, we investigate the probability of a path  $\sigma^{\text{BiC}}(S)$  corresponding to subset  $S = \{j_1, j_2, \dots, j_K\}$  where  $j_1 < j_2 < \dots < j_K$ . Proposition 2 denotes  $\sigma^{\text{BiC}}(S)$  as a path consisting of  $K + 1$  segments. The  $k^{\text{th}}$  segment runs from  $p_{j_k}^k$  to  $p_{j_{k+1}-1}^k$ :  $\sigma_k^{\text{BiC}}(S) = \{p_{j_k}^k, p_{j_{k+1}-1}^k, \dots, p_{j_{k+1}-1}^k\}$  such that the  $0^{\text{th}}$  one has  $j_0 = 0$ . The last  $K^{\text{th}}$  segment starts at  $p_{j_K}^K$  and ends at  $d$ :  $\sigma_K^{\text{BiC}}(S) = \{p_{j_K}^K, p_{j_K+1}^K, \dots, p_m^K, d\}$ . The probability of the  $k^{\text{th}}$  segment ( $0 \leq k < K$ ) is

$$\tilde{P}^{\text{BiC}}(\sigma_k^{\text{BiC}}(S)) = \prod_{j=j_k}^{j_{k+1}-2} \frac{\exp\left(\frac{1}{\mu_{p_j^k}}v(p_{j+1}^k|p_j^k)\right) \exp\left(\frac{1}{\mu_{p_j^k}}\tilde{V}^{\text{BiC}}(p_{j+1}^k)\right)}{\exp\left(\frac{1}{\mu_{p_j^k}}\tilde{V}^{\text{BiC}}(p_j^k)\right)}$$

Since all arcs in a segment corresponds to an action of not selecting an alternative (and doesn't increase the selection count  $k$ ), all utility values are zero:  $\exp\left(\frac{1}{\mu_k}v(p_{j+1}^k|p_j^k)\right) = 1$ . The considered scenario also provide a formula for the scale parameters:  $\mu_{p_j^k} = \mu^*(k)$ . Replace these alternative forms and reduce, we have

$$\tilde{P}^{\text{BiC}}(\sigma_k^{\text{BiC}}(S)) = \prod_{j=j_k}^{j_{k+1}-2} \frac{\exp\left(\frac{1}{\mu^*(k)}\tilde{V}^{\text{BiC}}(p_{j+1}^k)\right)}{\exp\left(\frac{1}{\mu^*(k)}\tilde{V}^{\text{BiC}}(p_j^k)\right)} = \frac{\exp\left(\frac{1}{\mu^*(k)}\tilde{V}^{\text{BiC}}(p_{j_{k+1}-1}^k)\right)}{\exp\left(\frac{1}{\mu^*(k)}\tilde{V}^{\text{BiC}}(p_{j_k}^k)\right)}$$

In the case of the  $K^{\text{th}}$  segment, the last arc from  $p_m^K$  to  $d$  also has zero utility. Moreover,  $\tilde{V}^{\text{BiC}}(d) = 0$ . Therefore:

$$\tilde{P}^{\text{BiC}}(\sigma_K^{\text{BiC}}(S)) = \left( \prod_{j=j_K}^{m-1} \frac{\exp\left(\frac{1}{\mu^*(K)}\tilde{V}^{\text{BiC}}(p_{j+1}^K)\right)}{\exp\left(\frac{1}{\mu^*(K)}\tilde{V}^{\text{BiC}}(p_j^K)\right)} \right) \frac{1}{\exp\left(\frac{1}{\mu^*(K)}\tilde{V}^{\text{BiC}}(p_m^K)\right)} = \frac{1}{\exp\left(\frac{1}{\mu^*(K)}\tilde{V}^{\text{BiC}}(p_{j_K}^K)\right)}$$

We multiply the probability of all segments to achieve the total likelihood of the complete path. Note that between two consecutive segments  $k$  and  $k + 1$  there exists an intermediate arc  $(p_{j_{k+1}-1}^k, p_{j_{k+1}}^{k+1})$  with utility  $v_{k+1}$  and scale parameters  $\mu_{p_{j_{k+1}-1}^k} = \mu^*(k)$ .

$$\begin{aligned}
\tilde{P}^{\text{BiC}}(\sigma^{\text{BiC}}(S)) &= \prod_{k=0}^{K-1} \left( \tilde{P}^{\text{BiC}}(\sigma_k^{\text{BiC}}(S)) \frac{\exp\left(\frac{1}{\mu^*(k)} v_{k+1}\right) \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j_{k+1}}^{k+1})\right)}{\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j_{k+1}-1}^k)\right)} \right) \frac{1}{\exp\left(\frac{1}{\mu^*(K)} \tilde{V}^{\text{BiC}}(p_{j_K}^K)\right)} \\
&= \prod_{k=0}^{K-1} \frac{\exp\left(\frac{1}{\mu^*(k)} v_{k+1}\right) \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j_{k+1}}^{k+1})\right)}{\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{BiC}}(p_{j_k}^k)\right)} \frac{1}{\exp\left(\frac{1}{\mu^*(K)} \tilde{V}^{\text{BiC}}(p_{j_K}^K)\right)}
\end{aligned} \tag{12}$$

Finally, we evaluate the corresponding path  $\sigma^{\text{MuC}}(S) = (p_0^0, p_{j_1}^1, p_{j_2}^2, \dots, p_{j_K}^K, d)$  for the subset  $S = \{j_1, j_2, \dots, j_K\}$  in MuC DAG. Its probability is also the product of all arcs:

$$\begin{aligned}
\tilde{P}^{\text{MuC}}(\sigma^{\text{MuC}}(S)) &= \left( \prod_{k=0}^{K-1} \tilde{P}^{\text{MuC}}(p_{j_{k+1}}^{k+1} | p_{j_k}^k) \right) \tilde{P}^{\text{MuC}}(d | p_{j_K}^K) \\
&= \prod_{k=0}^{K-1} \frac{\exp\left(\frac{1}{\mu_{p_{j_k}^k}} v(p_{j_{k+1}}^{k+1} | p_{j_k}^k)\right) \exp\left(\frac{1}{\mu_{p_{j_k}^k}} \tilde{V}^{\text{MuC}}(p_{j_{k+1}}^{k+1})\right)}{\exp\left(\frac{1}{\mu_{p_{j_k}^k}} \tilde{V}^{\text{MuC}}(p_{j_k}^k)\right)} \frac{\exp\left(\frac{1}{\mu_{p_{j_K}^K}} (v(d | p_{j_K}^K) + \tilde{V}^{\text{MuC}}(d))\right)}{\exp\left(\frac{1}{\mu_{p_{j_K}^K}} \tilde{V}^{\text{MuC}}(p_{j_K}^K)\right)}
\end{aligned}$$

Each arc has an utility  $v(p_{j_{k+1}}^{k+1} | p_{j_k}^k) = v_{k+1}$ , except  $v(d | p_{j_K}^K) = \tilde{V}^{\text{MuC}}(d) = 0$ . The condition statements give  $\mu_{p_{j_k}^k} = \mu^*(k)$ . Consequently, the above equation becomes

$$\tilde{P}^{\text{MuC}}(\sigma^{\text{BiC}}(S)) = \prod_{k=0}^{K-1} \frac{\exp\left(\frac{1}{\mu^*(k)} v_{k+1}\right) \exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{MuC}}(p_{j_{k+1}}^{k+1})\right)}{\exp\left(\frac{1}{\mu^*(k)} \tilde{V}^{\text{MuC}}(p_{j_k}^k)\right)} \frac{1}{\exp\left(\frac{1}{\mu^*(K)} \tilde{V}^{\text{MuC}}(p_{j_K}^K)\right)}. \tag{13}$$

From Proposition 10, we have  $\tilde{V}^{\text{BiC}}(p_j^k) = \tilde{V}^{\text{MuC}}(p_j^k)$  for all  $p_j^k$  nodes that exist in both DAGs. Apply this equality to Equation 12 and Equation 13, we have

$$\tilde{P}^{\text{BiC}}(\sigma^{\text{BiC}}(S)) = \tilde{P}^{\text{MuC}}(\sigma^{\text{MuC}}(S)),$$

as desired.

## B Extension to Multiple Discrete-Count Choice Problem

In the main body of the paper, our assumption was that each individual had only two choices: either selecting or not selecting an elemental alternative. In this section, we will explore a scenario where individuals can choose each elemental alternative multiple times. In this particular setup, every observation consists of a set of pairs  $(i, c_i)$ , where  $i \in [m]$  is an elemental

alternative, and  $c_i$  is a non-negative integer representing the number of times the alternative  $i$  is chosen. The total number of selected alternatives must also fall within the range of  $[L, U]$ , i.e.,  $\sum_{i \in [m]} c_i \in [L, U]$ . In this section, we will demonstrate how the BiC and MuC DAGs can be adjusted to accommodate this extended scenario.

We first describe the LMDC model adapted for the multiple discrete-count setting. Each composite alternative  $S$  now is a set of pairs of elemental alternatives and their counts, i.e.,  $S = \{(i, c_i) \mid i \in [m]\}$ . Let us denote by  $\bar{\Omega}^{[L, U]}$  the set of all possible composite alternatives,

$$\bar{\Omega}^{[L, U]} = \left\{ S = \{(i, c_i) \mid i \in [m]\} \mid \sum_{i \in [m]} c_i \in [L, U] \right\}.$$

Under the logit setting, the choice probability of an alternative  $S \in \bar{\Omega}^{[L, U]}$  becomes

$$P(S \mid \bar{\Omega}^{[L, U]}) = \frac{\exp(v(S \mid \boldsymbol{\beta}))}{\sum_{S' \in \bar{\Omega}^{[L, U]}} \exp(v(S' \mid \boldsymbol{\beta}))}, \quad (14)$$

where  $V(S \mid \boldsymbol{\beta})$  is the total utility of the composite alternative  $S$ , computed as  $V(S \mid \boldsymbol{\beta}) = \sum_{(i, c_i) \in S} c_i v_i(\boldsymbol{\beta})$ . It can be seen that the choice probability formulation given in (14) is a generalization of (1) when  $c_i$  only takes values of 0 or 1. The log-likelihood function can be defined in the same way as in Section 3. It can also be proved that the log-likelihood function, in the extended context, is also concave in  $\boldsymbol{\beta}$ , guaranteeing global optimality when solving the maximum likelihood estimation. We will describe, in the following, BiC and MuC DAGs for this choice model.

**Adapted BiC DAG.** We also create a DAG whose nodes are denoted as  $p_i^c$  where the superscript  $c$  represents the accumulated number of selected elemental alternatives, and  $i \in [m]$  is the index of the corresponding alternative. Nodes are also arranged in  $(m + 2)$  tiers where each tier will consist of  $(U + 1)$  nodes representing the fact that each elemental alternative can be selected at most  $U$  times, from 0 to  $U$ . In each tier, we create arcs connecting nodes  $p_i^c$  with  $p_i^{c+1}$ , which represents an additional selection of alternative  $i$ . Across different tiers, we create arcs connecting nodes  $p_i^c$  with  $p_{i+1}^c$ , representing the completion of selecting alternative  $i$  and the transition to the next alternative  $i + 1$ . So, from each node, there are two outgoing arcs representing an additional selection of the same alternative or a transition to the next alternative in the sequence. We illustrate the adapted BiC DAG in Figure 7 below.

To apply the RL to the adapted BiC DAG, we define arc utilities as follows

$$\begin{cases} (a) : v^{\text{BiC}}(p_{i+1}^c \mid p_i^c) = 0, \forall c \in \{0, \dots, U\}, i \in \{0, \dots, m\} \\ (b) : v^{\text{BiC}}(p_i^{c+1} \mid p_i^c) = v_i, \forall c \in \{0, \dots, U - 1\}, i \in \{0, \dots, m\} \\ (c) : v^{\text{BiC}}(k \mid d) = 0, \forall k \in \mathcal{N}, d \in N(k), \end{cases} \quad (15)$$

where the assignment (a) is for a transition to the next alternative in the sequence, which does not generate any utilities, and (b) is for a selection of alternative  $i$ , generating a utility of  $v_i$ .

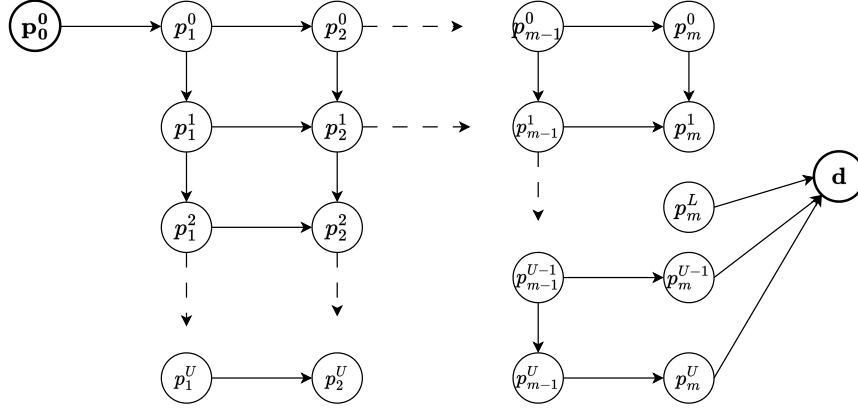


Figure 7: Adapted BiC DAG for multiple discrete-count choices.

**Adapted MuC DAG.** We do the same for the MuC DAG. That is, nodes are also organized into  $(m + 2)$  tiers where the first tier consists of only one node  $p_0^0$  representing the start of the choice process, and the last tier consists of one node  $d$  representing the termination of the choice process. In Tier  $i \in [m]$ , we include  $(U + 1)$  nodes of the form  $p_i^c$ , where  $c = 1, \dots, U$ , representing the fact that each elemental alternative can be selected at most  $U$  times. Within each tier  $i \in [m]$ , we create arcs between nodes  $p_i^c$  and  $p_i^{c+1}$ , representing an additional selection of alternative  $i$ . Across tiers, we connect nodes  $p_i^c$  with  $p_{i+1}^{c+1}, p_{i+2}^{c+1}, \dots, p_m^{c+1}$ , representing selections of next alternatives in the sequence, noting that we only connect nodes  $p_i^c$ , for  $c \geq L, i \in [m]$ , with the destination  $d$ . The adapted MuC DAG is illustrated in Figure 8.

For the use of the RL model on the adapted MuC DAG, we define arc utilities as follows:

$$\begin{cases} (a') : v^{\text{MuC}}(p_i^c | p_i^{c+1}) = v_i, \forall c \in \{0, \dots, U - 1\}, i \in \{1, \dots, m\} \\ (b') : v^{\text{MuC}}(p_{i+1}^{c+1} | p_i^c) = v_{i+1}, \forall c \in \{0, \dots, U - 1\}, i \in \{0, \dots, m - 1\}, \\ (c') : v^{\text{MuC}}(k | d) = 0, \forall k \in \mathcal{N}, d \in N(k), \end{cases} \quad (16)$$

where  $(a')$  are the assignments for arcs that represent a repeated selection of alternative  $i \in [m]$ , and  $(b')$  are for arcs that represent transitions to subsequent alternatives.

Similar to the multiple discrete choice problem considered in the main body of the paper, it can be proved that there is a one-to-one mapping between each discrete-count composite alternative in  $\bar{\Omega}^{[L,U]}$  and a path in the adapted BiC (or adapted MuC) DAG, and when applying the RL model to the adapted graphs, the path choice distributions over paths in the adapted BiC (or adapted MuC) are the same as the distribution over composite alternatives given by (14). We state these results in the following proposition:

**Proposition 11** *The following holds:*

- (i) *There is a one-to-one mapping between each composite alternative in  $\bar{\Omega}^{[L,U]}$  and a path in the adapted BiC DAG, and a path in the adapted MuC DAG.*

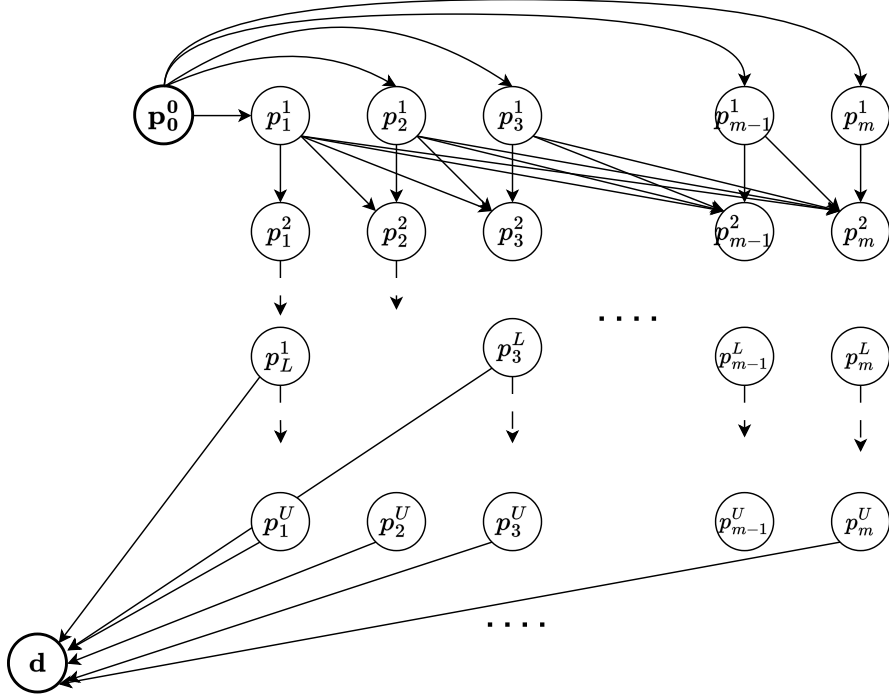


Figure 8: Adapted MuC DAG for multiple discrete-count choices.

- (ii) For each  $S \in \bar{\Omega}^{[L,U]}$ , let  $\bar{\sigma}^{\text{BiC}}(S)$  and  $\bar{\sigma}^{\text{MuC}}(S)$  be its mapped paths in the adapted BiC and MuC DAGs, respectively. In addition, let  $\bar{P}^{\text{BiC}}(\sigma)$  and  $\bar{P}^{\text{MuC}}(\sigma)$  be the probabilities of a given path  $\sigma$  given by the RL model applied to the adapted BiC and MuC graphs, respectively, we have

$$P(S | \bar{\Omega}^{[L,U]}) = \bar{P}^{\text{BiC}}(\bar{\sigma}^{\text{BiC}}(S)) = \bar{P}^{\text{MuC}}(\bar{\sigma}^{\text{MuC}}(S)).$$

As a result, the multiple discrete-count choice model can be estimated via estimating the RL model on the DAGs.

- (iii) When computing the value function required during the estimation of the RL or nested RL model, for both adapted BiC and MuC, the value iteration always returns a fixed-point solution after  $(m + 2)$  iterations.

The proof is very similar to the proofs of similar results presented in the previous sections. We, therefore, omit it.