On the complexity of package recommendation problems

Citation for published version:
https://doi.org/10.1145/2213556.2213592

Digital Object Identifier (DOI):
10.1145/2213556.2213592

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
On the Complexity of Package Recommendation Problems

Ting Deng
School of Computer Science
and Engineering
Beihang University
Beijing, China
dengting@act.buaa.edu.cn

Wenfei Fan
Lab. for Foundations of
Computer Science
School of Informatics
University of Edinburgh, UK
wenfei@inf.ed.ac.uk

Floris Geerts
Dept. of Mathematics &
Computer Science
University of Antwerp
Antwerp, Belgium
floris.geerts@ua.ac.be

ABSTRACT
Recommendation systems aim to recommend items that are likely to be of interest to users. This paper investigates several issues fundamental to such systems.

(1) We model recommendation systems for packages of items. We use queries to specify multi-criteria for item selections and express compatibility constraints on items in a package, and use functions to compute the cost and usefulness of items to a user.

(2) We study recommendations of points of interest, to suggest top-k packages. We also investigate recommendations of top-k items, as a special case. In addition, when sensible suggestions cannot be found, we propose query relaxation recommendations to help users revise their selection criteria, or adjustment recommendations to guide vendors to modify their item collections.

(3) We identify several problems, to decide whether a set of packages makes a top-k recommendation, whether a rating bound is maximum for selecting top-k packages, whether we can relax the selection query to find packages that users want, and whether we can update a bounded number of items such that the users' requirements can be satisfied. We also study function problems for computing top-k packages, and counting problems to find how many packages meet the user's criteria.

(4) We establish the upper and lower bounds of these problems, all matching, for combined and data complexity. These results reveal the impact of variable sizes of packages, the presence of compatibility constraints, as well as a variety of query languages for specifying selection criteria and compatibility constraints, on the analyses of these problems.

Categories and Subject Descriptors: H.3.5 [Information Systems]: Online Information Services – Web-based services; F.2.0 [Theory of Computation]: Analysis of Algorithms and Problem Complexity – General

Keywords: Recommendation problems, Complexity, Query relaxation.

1. INTRODUCTION
Recommendation systems, a.k.a. recommender systems, recommendation engines or platforms, aim to identify and suggest information items or social elements that are likely to be of interest to users. Traditional recommendation systems are to select top-k items from a collection of items, e.g., books, music, news, Web sites and research papers [3], which satisfy certain criteria identified for a user, and are ranked by ratings with a utility function. More recently recommendation systems are often used to find top-k packages, i.e., sets of items, such as travel plans [34], teams of players [22] and various course combinations [19, 26, 27]. The items in a package are required not only to meet multi-criteria for selecting individual items, but also to satisfy compatibility constraints defined on all the items in a package taken together, such as team formation [22] and course prerequisites [26]. Packages may have variable sizes subject to a cost budget, and are ranked by overall ratings of their items [34].

Recommendation systems are increasingly becoming an integral part of Web services [34], Web search [4], social networks [4], education software [27] and commerce services [3]. A number of systems have been developed for recommending items or packages, known as points of interest (POI) [34] (see [3, 4] for surveys). These systems use relational queries to specify selection criteria and compatibility constraints [2, 7, 19, 27, 34]. There has also been work on the complexity of computing POI recommendations [22, 26, 27, 34]. However, to understand central issues associated with recommendation systems, there is much more to be done. (1) The previous complexity results were developed for individual applications with specific selection criteria and compatibility constraints. They may not carry over to other settings. This highlights the need for studying recommendation problems in a uniform model. (2) In most cases only lower bounds were given (NP-hard by e.g., [26, 27]). Worse still, among the few upper bounds claimed, some are not quite correct. It is necessary to set the record straight by establishing matching upper and lower bounds. (3) No previous work has studied where high complexity arises. Is it from variable sizes of packages, compatibility constraints or from complex selection criteria? The need for understanding this is evident when developing practical recommendation systems. (4) In practice one often gets no sensible recommendations. When this happens, a system should be able to come up with recommendations for the users to revise selection criteria, or for vendors to adjust their item collections. However, no matter how important these issues are, no previous work has studied recommendations beyond POI.
Example 1.1: Consider a recommendation system for travel plans, which maintains two relations specified by:

flight(#, From, To, DT, DD, AT, AD, Pr),
POI(name, city, type, ticket, time).

Here a flight tuple specifies flight # from From to To that departs at time DT on date DD and arrives at time AT on date AD, with airfare Pr. A POI tuple specifies a place name to visit in the city, its ticket price, type (e.g., museum, theater), and the amount of time needed for the visit.

1. Recommendations of items. A user wants to find top-3 flights from EDT to NYC with at most one stop, departing on 1/1/2012, with lowest possible airfare and duration time. This can be stated as item recommendation: (a) flights are items; (b) the selection criteria are expressed as a union $Q_1 \cup Q_2$ of conjunctive queries, where $Q_1$ and $Q_2$ select direct and one-stop flights from EDT to NYC leaving on 1/1/2012, respectively; and (c) the items selected are ranked by a utility function $f()$ (e.g., allowing To to be a city within 15 miles of NYC, then direct flights are available, e.g., from EDT to EWR. This suggests that the user revise her selection criteria by recommending query relaxations.

2. Recommendations of packages. One is planning a 5-day holiday, by taking a direct flight from EDT to NYC departing on 1/1/2012 and visiting as many places in Dur holiday, by taking a direct flight from Dur to...
These results are also of interest to the study of top-give precise bounds for the problems studied in, when selection criteria are given by an identity query. These FP
problem for top-(d) In the absence of compatibility constraints, the decision complexity in this model, from decision problems to function and counting problems. They tell us where complexity arises, complementing previously stated results.

(a) Query languages dominate the complexity of recommendation problems, e.g., the problem for deciding the maximum bound for top-k package recommendations ranges from \(D^2\text{CQ}\)-complete for CQ, \(\text{PSPACE}\)-complete for FO and \(\text{DATALOG}\), to \(\text{EXPTIME}\)-complete for \(\text{DATALOG}\).

(b) Variable package sizes do not make our lives harder when combined complexity is concerned for all the languages given above. Indeed, when packages may have variable sizes, all these problems have the same combined complexity as their counterparts when packages are restricted to be singleton sets. In fact, variable sizes of packages have impact only on data complexity, or when \(\mathcal{L}_Q\) is a simple language with a \(\text{PTIME}\) complexity for its membership problem. These clarify the impact of package sizes studied in, e.g., [34].

(c) The presence of compatibility constraints does not increase the complexity when the query language \(\mathcal{L}_Q\) is FO, \(\text{DATALOG}\), or \(\text{DATALOG}\). Indeed, for these languages, all the problems for package recommendations and their counterparts for item recommendations have the same complexity. Moreover, these constraints do not complicate the data complexity analyses. However, compatibility constraints increase combined complexity when \(\mathcal{L}_Q\) is contained in \(\exists\text{FO}\).

(d) In the absence of compatibility constraints, the decision problem for top-k package recommendations is \(\text{DP}\)-complete and its function problem is \(\text{FP}^{\text{NP}}\)-complete when \(\mathcal{L}_Q\) is CQ. They are \(\text{coNP}\)-hard and \(\text{FP}^{\text{NP}}\)-hard, respectively, even when selection criteria are given by an identity query. These give precise bounds for the problems studied in, e.g., [34].

These results are also of interest to the study of top-k query answering, among other things. A variety of techniques are used to prove the results, including a wide range of reductions, and constructive proofs with algorithms (e.g., for the function problems). In particular, the proofs demonstrate that the complexity of these problems for CQ, UCQ and \(\exists\text{FO}\) is inherent to top-k package querying itself, rather than a consequence of the complexity of the query languages.

Related work. Traditional recommendation systems aim to find, for each user, items that maximize the user’s utility (see, e.g., [3] for a survey). Selection criteria are decided by content-based, collaborative and hybrid approaches, which consider preferences of each user in isolation, or preferences of similar users [3]. The prior work has mostly focused on how to choose appropriate utility functions, and how to extrapolate such functions when they are not defined on the entire item space, by deriving unknown values from known ones. Our model supports content-based, collaborative and hybrid criteria in terms of various queries. We assume a given utility function that is total, and focus on the computational complexity of recommendation problems.

Recently recommendation systems have been extended to finding packages, which are presented to the user in a ranked order based on some rating function [6, 26, 27, 34]. A number of algorithms have been developed for recommending packages of a fixed size [6, 22] or variable sizes [26, 27, 34]. Compatibility constraints [22, 26, 27, 34] and budget restrictions [34] on packages have also been studied. Instead of considering domain-specific applications, we model recommendations of both items and packages (fixed size or polynomial size) by specifying general selection criteria and compatibility constraints as queries, and supporting aggregate constraints defined with cost budgets and rating bounds.

Several decision problems for course package recommendations have been shown \(\text{NP}\)-hard [26, 27]. It was claimed that problems of forming a team with compatibility constraints [22] and the problem of finding packages that satisfy some budget restrictions (without compatibility constraints) [34] are \(\text{NP}\)-complete. In contrast, we establish the precise complexity of a variety of problems associated with POI recommendations (Table 1, Section 7). Moreover, we provide the complexity of query relaxation and adjustment recommendations, which have not been studied by prior work.

There has also been a host of work on recommending items and packages taken from views of the data [2, 7, 19, 23, 34]. Such views are expressed as relational queries, representing preferences or points of interest [2, 7, 19]. Here recommendations often correspond to top-k query answers. Indeed, top-k query answering retrieves the \(k\)-items (tuples) from a query result that are top-ranked by some scoring function [16]. Such queries either simply select tuples, or join and aggregate multiple inputs to find the top-k tuples, by possibly incorporating user preference information [19, 29]. A number of top-k query evaluation algorithms have been developed (e.g., [12, 23, 28]; see [16] for a survey), as well as algorithms for incremental computation of ranked query results [10, 14, 24] that retrieve the top-k query answers one at a time. A central issue there concerns how to combine different ratings of the same item based on multiple criteria. Our work also retrieves tuples from the result of a query. It differs from the previous work in the following. (1) In contrast to top-k query answering, we are to find items and sets of items (packages) provided that a utility or rating function is given. (2) We focus on the complexity of recommendations problems rather than the efficiency or optimization of query evaluation. (3) Beyond recommendations of POI, we also study query relaxation and adjustment recommendations.

Query relaxations have been studied in, e.g., [8, 13, 17, 18]. Several query generalization rules are introduced in [8], assuming that query acceptance conditions are monotonic. Heuristic query relaxation algorithms are developed in [13,
of expressing compatibility constraints commonly found in practice, including those studied in [19, 22, 26, 27, 34].

Aggregate constraints. To specify aggregate constraints, we define a cost function and a rating function over packages, following [34]: (1) \( \text{cost}(N) \) computes a value in \( \mathbb{R} \) as the cost of package \( N \); and (2) \( \text{val}(N) \) computes a value in \( \mathbb{R} \) as the overall rating of \( N \). For instance, \( \text{cost}(N) \) in Example 1.1 is computed from the total time taken for visiting POI, while \( \text{val}(N) \) is defined in terms of airfare and total ticket prices.

We just assume that \( \text{cost()} \) and \( \text{val()} \) are PTIME computable aggregate functions, defined in terms of \( \text{e.g., max, min, sum, avg} \), as commonly found in practice.

We also assume a cost budget \( C \), and specify an aggregate constraint \( \text{cost}(N) \leq C \). For instance, the cost budget \( C \) in Example 1.1 is the total time allowed for visiting POI in 5 days, and the aggregate constraint \( \text{cost}(N) \leq C \) imposes a bound on the number of POI in a package \( N \).

Top-k package selections. For a database \( D \), queries \( Q \) and \( Q_e \) in \( \mathcal{L}_Q \), a natural number \( k \geq 1 \), a cost budget \( C \), and functions \( \text{cost}() \) and \( \text{val}() \), a top-k package selection is a set \( N = \{ N_i | i \in [1,k] \} \) of packages such that for each \( i \in [1,k] \), (1) \( N_i \subseteq Q(D) \), i.e., its items meet the criteria given in \( Q \); (2) \( Q_e(N_i,D) = \emptyset \), i.e., the items in the package satisfy the compatibility constraints specified by query \( Q_e \); (3) \( \text{cost}(N_i) \leq C \), i.e., its cost is below the budget; (4) the number \( |N_i| \) of items in \( N_i \) is no larger than \( p(|D|) \), where \( p \) is a predefined polynomial and \( |D| \) is the size of \( D \); indeed, it is not of much practical use to find a package with exponentially many items; as will be seen in Section 4, we shall also consider a constant bound \( B_p \) for \( |N_i| \); (5) for all packages \( N' \not\in N \) that satisfies conditions (1–4) given above, \( \text{val}(N') \leq \text{val}(N_i) \), i.e., packages in \( N \) have the \( k \) highest overall ratings among all feasible packages; and (6) \( N_i \neq N_j \) if \( i \neq j \), i.e., the packages are pairwise distinct.

Note that packages in \( N \) may have variable sizes. That is, the number of items in each package is not bounded by a constant. We just require that \( N_i \) satisfies the constraint \( \text{cost}(N_i) \leq C \) and \( |N_i| \) does not exceed a polynomial in \( |D| \).

Package recommendation is to find a top-k package selection for \( (Q,D,Q_e,\text{cost}(),\text{val}(),C) \), if there exists one.

As shown in Example 1.1, users may want to find, \( \text{e.g., a top-k travel-plan selection with the minimum price.} \)

Item recommendations. To rank items, we use a utility function \( f() \) to measure the usefulness of items selected by \( Q(D) \) to a user [3]. It is a PTIME-computable function that takes a tuple \( s \) from \( Q(D) \) and returns a real number \( f(s) \) as the rating of item \( s \). The functions may incorporate users’ preference [29], and may be different for different users.

Given a constant \( k \geq 1 \), a top-k selection for \( (Q,D,f) \) is a set \( S = \{ s_i | i \in [1,k] \} \) such that (a) \( S \subseteq Q(D) \), i.e., items in \( S \) satisfy the criteria specified by \( Q \); (b) for all \( s \in Q(D), S \) and \( i \in [1,k], f(s) \leq f(s_i) \), i.e., items in \( S \) have the highest ratings; and (c) \( s_i \neq s_j \) if \( i \neq j \), i.e., items in \( S \) are distinct.

Given \( D,Q,f \) and \( k \), item recommendation is to find a top-k selection for \( (Q,D,f) \) if there exists one.
For instance, a top-3 item selection is described in Example 1.1, where items are flights and the utility function \( f() \) is defined in terms of the airfare and duration of each flight.

The connection between item and package selections. Item selections are a special case of package selections. Indeed, a top-\( k \) selection \( S = \{ s_i \mid i \in [1,k] \} \) for \((Q, D, f)\) is a top-\( k \) package selection \( N \) for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C)\), where \( N = \{ N_i \mid i \in [1,k] \} \), and for each \( i \in [1,k] \), (a) \( N_i = \{ s_i \} \), (b) \( Q_c \) is a constant query that returns 0 on any input, referred to as the empty query; (c) \( \text{cost}(N_i) = \{ N_i \mid N_i \neq \emptyset \} \), and \( \text{cost}(\emptyset) = \infty \); that is, \( \text{cost}(N_i) \) counts the number of items in \( N_i \) if \( N_i \neq \emptyset \), and the empty set is not taken as a recommendation; (d) the cost budget \( C = 1 \), and hence, \( N_i \) consists of a single item by \( \text{cost}(N_i) \leq C \); and (e) \( \text{val}(N_i) = f(s_i) \).

In the sequel, we use top-\( k \) package selection specified in terms of \((Q, D, f)\) to refer to a top-\( k \) selection \( S \) for \((Q, D, f)\), i.e., a top-\( k \) package selection for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C)\) in which \( Q_c, \text{cost}(\cdot), \text{val}(\cdot) \) and \( C \) are defined as above.

We say that compatibility constraints are absent if \( Q_c \) is the empty query; e.g., \( Q_c \) is absent in item selections. One might want to consider general \( \text{PTIME} \) compatibility constraints \( Q_c \). As will be seen in Section 4, the complexity when \( Q_c \) is in \( \text{PTIME} \) remains the same as its counterpart when \( Q_c \) is absent for all the problems studied in this paper.

Query languages. We consider \( Q_c, Q, \) in a query language \( \mathcal{L}_Q \), ranging over the following (see e.g., [1] for details):

(a) conjunctive queries (CQ), built up from atomic formulas with constants and variables, i.e., relation atoms in database schema \( R \) and built-in predicates \((=, \neq, <, \leq, \geq)\), by closing under conjunction \( \land \) and existential quantification \( \exists \);

(b) union of conjunctive queries (UCQ) of the form \( Q_1 \cup \cdots \cup Q_r \), where for each \( i \in [1,r] \), \( Q_i \) is in CQ;

(c) positive existential FO queries (\( \exists \forall \)) built from atomic formulas by closing under \( \land, \lor, \neg \) and universal quantification \( \forall \); and

(d) nonrecursive datalog queries (DATALOG\(_{\text{nonrec}} \)), defined as a collection of rules of the form \( p(x_1) \leftarrow p_1(x_1), \ldots, p_n(x_n) \), where the head \( p \) is an IDB predicate and each \( p_i \) is either an atomic formula or an IDB predicate, such that its dependency graph is acyclic; the dependency graph of a DATALOG query \( Q \) is a directed graph \( G_Q = (V, E) \), where \( V \) includes all the predicates of \( Q \), and \( (p', p) \) is an edge in \( E \) iff \( p' \) is a predicate that appears in a rule with \( p \) as its head \([9]\);

(e) first-order logic queries (FO) built from atomic formulas using \( \land, \lor, \neg, \exists \), and universal quantification \( \forall \); and

(f) datalog queries (DATALOG), defined as a collection of rules \( p(x_1) \leftarrow p_1(x_1), \ldots, p_n(x_n) \), for which the dependency graph may possibly be cyclic, i.e., DATALOG is an extension of DATALOG\(_{\text{nonrec}} \) with an inflationary fixpoint operator.

These languages specify both multi-criteria for item selections and compatibility constraints for package selections.

3. RECOMMENDATIONS OF POI’S

In this section we investigate POI recommendations. We identify four problems for package recommendations (Section 3.1), and establish their complexity (Section 3.2).

3.1 Recommendation Problems

We investigate four problems, stated as follows, which are fundamental to computing package recommendations. We start with a decision problem for package selections. Consider a database \( D \), queries \( Q \) and \( Q_c \), in a query language \( \mathcal{L}_Q \), functions \( \text{val}(\cdot) \) and \( \text{cost}(\cdot) \), a cost budget \( C \), and a natural number \( k \geq 1 \). Given a set \( N \) consisting of \( k \) packages, it is to decide whether \( N \) makes a top-\( k \) package selection. That is, each package \( N_i \) in \( N \) satisfies the selection criteria \( Q \), compatibility constraint \( Q_c \), and aggregate constraints \( \text{cost}(N_i) \leq C \) and \( \text{val}(N_i) \geq \text{val}(N') \) for all \( N' \in N \). As remarked earlier, we assume a predefined polynomial such that \( |N| \leq p(|D|) \) (omitted from the problem statement below for simplicity). Intuitively, this problem is to decide whether a set \( N \) of packages should be recommended.

RPP\( (\mathcal{L}_Q) \): The recommendation problem (package).
INPUT: A database \( D \), two queries \( Q \) and \( Q_c \) in \( \mathcal{L}_Q \), two functions \( \text{cost}(\cdot) \) and \( \text{val}(\cdot) \), natural numbers \( C \) and \( k \geq 1 \), and a set \( N = \{ N_i \mid i \in [1,k] \} \).
QUESTION: Is \( N \) a top-\( k \) package selection for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C)\)?

After all, recommendation systems have to compute top-\( k \) packages, rather than expecting that candidate selections are already in place. This highlights the need for studying the function problem below, to compute top-\( k \) packages.

FRP\( (\mathcal{L}_Q) \): The function rec. problem (packages).
INPUT: \( D, Q, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, k \) as in RPP.
OUTPUT: A top-\( k \) package selection for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C)\) if it exists.

The next question concerns how to find a maximum rating bound for computing top-\( k \) packages. We say that a constant \( B \) is a rating bound for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, k)\) if \( (a) \) there exists a top-\( k \) package selection \( N' = \{ N_i \mid i \in [1,k] \} \) for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C)\) and moreover, \( (b) \) \( \text{val}(N_i) \geq B \) for each \( i \in [1,k] \). That is, \( B \) allows a top-\( k \) package selection. We say that \( B \) is the maximum bound for packages with \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, k)\) if all bounds \( B', B \geq B' \). Obviously \( B \) is unique if it exists. Intuitively, when \( B \) is identified, we can capitalize on \( B \) to compute top-rated packages. Furthermore, vendors could decide, e.g., price for certain items on sale with such a bound, for risk assessment.

MBP\( (\mathcal{L}_Q) \): The maximum bound problem (packages).
INPUT: \( D, Q, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, k \) as in RPP, and a natural number \( B \).
QUESTION: Is \( B \) the maximum bound for packages with \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, k)\) ?

A package \( N \) is valid for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, B)\) if \( (a) \) \( N \subseteq Q(D) \), \( (b) \) \( Q_c(N, D) = \emptyset \), \( (c) \) \( \text{cost}(N) \leq C \), and \( (d) \) \( \text{val}(N) \geq B \), where \( |N| \) is bounded by a polynomial in \(|D|\). Given \( B \), one naturally wants to know how many valid packages are out there, and hence, can be selected. This suggests that we study the following counting problem.

CPP\( (\mathcal{L}_Q) \): The counting problem (packages).
INPUT: \( D, Q, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, B \) as in MBP.
OUTPUT: The number of packages that are valid for \((Q, D, Q_c, \text{cost}(\cdot), \text{val}(\cdot), C, B)\).
3.2 Deciding, Finding and Counting Top-$k$ Packages

We now establish the complexity of $\text{RPP}(L_Q)$, $\text{FRP}(L_Q)$, $\text{MBP}(L_Q)$ and $\text{CPP}(L_Q)$, including their (1) combined complexity, when the query $Q$, compatibility constraint $Q_c$ and database $D$ may vary, and (2) data complexity, when only $D$ varies, while $Q$ and $Q_c$ are predefined and fixed. We study these problems for all the query languages $L_Q$ of Section 2.

Deciding package selections. We start with $\text{RPP}(L_Q)$. The result below tells us that the combined complexity of the problem is mostly determined by what query language $L_Q$ we use to specify selection criteria and compatibility constraints. Indeed, it is $\Sigma^p_2$-complete when $L_Q$ is FO. $\text{PSPACE}$-complete for $\text{DATALOG}_{nr}$ and FO, and it becomes $\text{EXPTIME}$-complete when $L_Q$ is $\text{DATALOG}$. The data complexity is $\text{coNP}$-complete for all the languages considered.

Theorem 3.1: For $\text{RPP}(L_Q)$, the combined complexity is
- $\Pi^p_2$-complete when $L_Q$ is CQ, UCQ, or $\exists \exists \forall^+$;
- $\text{PSPACE}$-complete when $L_Q$ is $\text{DATALOG}_{nr}$ or FO; and
- $\text{EXPTIME}$-complete when $L_Q$ is $\text{DATALOG}$.

The data complexity is $\text{coNP}$-complete for all the languages presented in Section 2, i.e., when $L_Q$ is CQ, UCQ, $\exists \exists \forall^+$, $\text{DATALOG}_{nr}$, FO or $\text{DATALOG}$.

Proof sketch: (1) We show that $\text{RPP}$ is $\Pi^p_2$-hard for CQ by reduction from the complement of the compatibility problem. The latter is to decide whether a given set $N$ of packages satisfies the criteria and constraints, in DP; it then checks whether there exists no package with a higher rating than some $N \in \mathcal{N}$, in $\Pi^p_2$. (2) We show that $\text{RPP}$ is $\text{PSPACE}$-hard for $\text{DATALOG}_{nr}$ by reduction from $\text{Q3SAT}$ (cf. [25]), and for FO by reduction from the membership problem for FO (“given a query $Q$, a database $D$ and a tuple $t$, whether $t \in Q(D)$”) [32], which are $\text{PSPACE}$-complete. We provide an $\text{NPSPACE}$ (= $\text{PSPACE}$) algorithm to check $\text{RPP}$ for these two languages. (3) For $\text{DATALOG}$, we show that $\text{RPP}$ is $\text{EXPTIME}$-hard by reduction from the membership problem for $\text{DATALOG}$, which is $\text{EXPTIME}$-complete [32]. For the upper bound, we give an $\text{EXPTIME}$ algorithm to check $\text{RPP}($DATALOG$)$.

(4) For the data complexity, we first show that the compatibility problem is $\text{NP}$-complete when $Q$ and $Q_c$ are fixed CQ queries, by reduction from $\text{3SAT}$, an $\text{NP}$-complete problem (cf. [25]). From this it follows that $\text{RPP}(\text{CQ})$ is already $\text{coNP}$-hard. We also give a $\text{coNP}$ algorithm for $\text{RPP}$ when $Q$ and $Q_c$ are fixed queries in either FO or $\text{DATALOG}$.

One might think that the absence of compatibility constraints $Q_c$ would make our lives easier. Indeed, $\text{RPP}(\text{CQ})$ becomes $\text{DP}$-complete in the absence of $Q_c$, as opposed to $\Pi^p_2$-complete in the presence of $Q_c$. However, when $L_Q$ is powerful enough to express FO or $\text{DATALOG}_{nr}$ queries, dropping $Q_c$ does not help: $\text{RPP}(L_Q)$ in this case has the same complexity as its counterpart when $Q_c$ is present.

Theorem 3.2: In the absence of $Q_c$, $\text{RPP}(L_Q)$ is
- $\text{DP}$-complete when $L_Q$ is CQ, UCQ, or $\exists \exists \forall^+$;
- $\text{PSPACE}$-complete when $L_Q$ is $\text{DATALOG}_{nr}$ or FO; and
- $\text{EXPTIME}$-complete when $L_Q$ is $\text{DATALOG}$.

Its data complexity remains $\text{coNP}$-complete for all the query languages given in Section 2.

Proof sketch: (1) We show that $\text{RPP}(\text{CQ})$ is $\text{DP}$-hard by reduction from $\text{SAT}$-$\text{UNSAT}$. The latter is to decide, given a pair $(\varphi_1, \varphi_2)$ of $\text{3SAT}$ instances, whether $\varphi_1$ is satisfiable and $\varphi_2$ is not satisfiable. It is $\text{DP}$-complete (cf. [25]).

In the absence of $Q_c$, the algorithm given earlier for $\text{RPP}(\exists \exists \forall^+)$ is in DP, and hence so is $\text{RPP}(\exists \exists \forall^+)$. (2-4) The lower bound proofs (2-4) of Theorem 3.1 do not use compatibility constraints and hence remain intact. The upper bounds obviously carry over to this special case.

Computing top-$k$ packages. We give the complexity of the function problem $\text{FRP}(L_Q)$ as follows:

Theorem 3.3: For $\text{FRP}(L_Q)$, the combined complexity is
- $\text{FP}^{\Sigma^p_2}$-complete when $L_Q$ is CQ, UCQ or $\exists \exists \forall^+$;
- $\text{FPSPACE}(\text{poly})$-complete if $L_Q$ is $\text{DATALOG}_{nr}$ or FO;
- $\text{FEXPTIME}(\text{poly})$-complete when $L_Q$ is $\text{DATALOG}$.

In the absence of compatibility constraints, its combined complexity remains unchanged for $\text{DATALOG}_{nr}$, FO and $\text{DATALOG}$, but it is $\text{FP}^\text{NP}$-complete for CQ, UCQ and $\exists \exists \forall^+$. Its data complexity is $\text{FP}^\text{NP}$-complete for all the languages, in the presence or absence of compatibility constraints.

Here $\text{FP}^\text{NP}$ is the class of all functions from strings to strings that can be computed by a $\text{PTIME}$ Turing machine with an $\text{NP}$ oracle (cf. [25]), and $\text{FP}^{\Sigma^p_2}$ is the class of all functions computable by a $\text{PTIME}$ 2-alternating max-min Turing machine [20]. By $\text{FPSPACE}(\text{poly})$ (resp. $\text{FEXPTIME}(\text{poly})$) we mean the class of all functions associated with a two-argument predicate $R$, that satisfies the following conditions: (a) $R_L$ is polynomially balanced, i.e., there is a polynomial $q$ such that for all strings $x$ and $y$, if $R_L(x, y)$ then $|y| \leq q(|x|)$, and (b) the decision problem “given $x$ and $y$, whether $R_L(x, y)$” is in $\text{FPSPACE}$ (resp. $\text{EXPTIME}$) [21].

Given a string $x$, the function associated with $R_L$ is to find a string $y$ such that $R_L(x, y)$ if such a string exists.

These results tell us that it is nontrivial to find top-$k$ packages. Indeed, in order to express compatibility constraints on travel plans given in [34], we need at least CQ; for course combination constraints of [19, 26, 27], we need FO; and for expressivity of flight plans we need $\text{DATALOG}$. These place FRP in $\text{FP}^{\Sigma^p_2}$, $\text{FPSPACE}(\text{poly})$ and $\text{FEXPTIME}(\text{poly})$, respectively.

It was claimed in several earlier papers that when $k=1$, it is $\text{NP}$-complete to find a top-1 package. Unfortunately, it is not the case. Indeed, the proofs of Theorems 3.1, 3.2 and 3.3 tell us that when $k=1$, the function problem $\text{FRP}(L_Q)$ remains $\text{FP}^{\Sigma^p_2}$-complete and the decision problem $\text{RPP}(L_Q)$ is $\Pi^p_2$-complete even when $L_Q$ is FO, not to mention more expressive $L_Q$. Even when $Q$ and $Q_c$ are both fixed, FRP is $\text{FP}^{\Sigma^p_2}$-complete and RPP is $\text{coNP}$-complete when $k=1$.

In the absence of compatibility constraints, only the analyses of the combined complexity of FRP for CQ, UCQ and $\exists \exists \forall^+$ are simplified. This is consistent with Theorem 3.2.

Proof sketch: (1) We show that $\text{FRP}(\text{CQ})$ is $\text{FP}^{\Sigma^p_2}$-hard by reduction from the $\text{MAXIMUM}$ $\Sigma^p_2$ problem, which is $\text{FP}^{\Sigma^p_2}$-complete [20]. The latter is to find, given a formula $\varphi(X) = \forall Y \psi(X, Y)$, the truth assignment $\mu_X$ of $X$ that satisfies $\varphi$ and comes last in the lexicographical ordering if it exists.
where \( \psi \) is a 3SAT instance. We give an FP\(^{\Sigma_2^P} \) algorithm for FRP(3FO\(^+ \)) to compute a top-k package selection if it exists.

(2) We show that FRP(\( \mathcal{L}_Q \)) is FSP\( \text{SPACE} \)poly\()-hard by reducing to it all functions computable by a PSPACE Turning machine in which the output on the working tape is bounded by a polynomial, when \( \mathcal{L}_Q \) is DATALOG\(_m\) or FO; similarly for FRP(DATALOG). For the upper bounds, we give an algorithm in FSP\( \text{SPACE}(\)poly\()) \) resp. EXPTIME\( (\)poly\()) \) for FRP(\( \mathcal{L}_Q \)) when \( \mathcal{L}_Q \) is DATALOG\(_m\) or FO (resp. DATALOG).

(3) When \( Q \) and \( Q_c \) are fixed, we show that FRP(\( CQ \)) in the absence of \( Q_c \) is FP\(^{NP}\)-hard by reduction from MAX-WEIGHT SAT, which is FP\(^{NP}\)-complete (cf. [25]). Given a set \( C \) of clauses with weights, MAX-WEIGHT SAT is to find a truth assignment that satisfies a set of clauses in \( C \) with the most total weight. For the upper bound, we give an FP\(^{NP}\) algorithm for FRP(\( \mathcal{L}_Q \)) when \( \mathcal{L}_Q \) is FO or DATALOG.

(4) When \( Q_c \) is absent, FRP(\( CQ \)) is FP\(^{NP}\)-hard by proof (3) given above, and the algorithm for FRP(3FO\(^+ \)) given in proof (1) is now in FP\(^{NP}\). The proofs for DATALOG\(_m\), FO and DATALOG given in (2) still work in this special case, as no \( Q_c \) is used there when verifying the lower bounds.

Deciding the maximum bound. We show that MBP(\( CQ \)) is D\(^{2\text{-}}\_\text{complete}\). Here D\(^{2\text{-}}\) is the class of languages recognized by oracle machines that make a query to a \( \Sigma_2^p \) oracle and a query to a \( \Pi_2^p \) oracle. That is, \( L \in \text{D}^{2\text{-}} \) if there exist languages \( L_1 \subseteq \Sigma_2^p \) and \( L_2 \subseteq \Pi_2^p \) such that \( L = L_1 \cap L_2 \) [33], analogous to how DP is defined with NP and coNP [25].

When \( \mathcal{L}_Q \) is FO, DATALOG\(_m\) or DATALOG, MBP(\( \mathcal{L}_Q \)) and RPP(\( \mathcal{L}_Q \)) have the same complexity. Moreover, the absence of \( Q_c \) has the same impact on MBP(\( \mathcal{L}_Q \)) as on RPP(\( \mathcal{L}_Q \)).

**Theorem 3.4:** For MBP(\( \mathcal{L}_Q \)), the combined complexity is

- D\(^{2\text{-}}\_\text{complete}\) when \( \mathcal{L}_Q \) \( \in \) CQ, UCQ or 3FO\(^+ \);
- PSPACE\(-\text{complete}\) when \( \mathcal{L}_Q \) \( \in \) DATALOG\(_m\) or FO; and
- \#EXPTIME\(-\text{complete}\) when \( \mathcal{L}_Q \) \( \in \) DATALOG.

When compatibility constraints are absent, its combined complexity remains unchanged for DATALOG\(_m\), FO and DATALOG, but it is DP-complete for CQ, UCQ and 3FO\(^+ \).

Its data complexity is \#P\(-\text{complete}\) for all the languages in the presence or absence of compatibility constraints.

Here we use the framework of predicate-based counting classes introduced in [15]. For a complexity class \( C \) of decision problems, \#-C is the class of all counting problems associated with a predicate \( R_L \) that satisfies the following conditions: (a) \( R_L \) is polynomially balanced (see its definition above); and (b) the decision problem "given \( x \) and \( y \), whether \( R_L(x,y) = 1 \) in \( C \). A counting problem is to compute the cardinality of the set \( \{ y \mid R_L(x,y) \} \), i.e., it is to find how many \( y \) there are such that \( R_L(x,y) = 1 \).

It is known that \#P = \#NP \subseteq \#P_c = \#P^{NP} = \#coNP, but \#NP \neq \#coNP iff NP = coNP, where \#P and \#NP are counting classes in the machine-based framework of [31]. From these we know that the combined complexity of CPP(\( CQ \)) is \#NP\(-\text{complete}\), and the data complexity of CPP(\( \mathcal{L}_Q \)) is \#P\(-\text{complete}\) for all the languages considered.

**Proof sketch:** (1) We show that CPP(\( CQ \)) is \#NP\(-\text{hard}\) by reduction from \( \Pi_1 \text{SAT} \). Given \( \varphi(X,Y) = \forall X \psi(X,Y) \), where \( \psi \) is a 3DNF, \( \Pi_1 \text{SAT} \) is to count the number of truth assignments of \( Y \) that satisfy \( \varphi \), and is \#coNP-complete [11]. The reduction is an 1-1 mapping from the solutions to CPP(\( CQ \)) to the truth assignments for \( \varphi(X,Y) \), and hence is parsimonious. We also show that CPP(\( 3\text{FO}^+ \)) is in \#NP.

(2) We show that CPP is \#PSPACE\(-\text{hard}\) for DATALOG\(_m\) and FO by parsimonious reductions from \#QBF, which is \#PSPACE\(-\text{complete}\) [21]. Given \( \varphi = \exists X_1 \forall X_2 P_1 \psi_1 \cdots P_n \psi_n \psi \), where \( \psi \) is a 3SAT instance over \( X \) and \( \{ y_i | i \in [1,n] \} \), and \( P_i \) is \( \forall \) or \( \exists \), \#QBF is to count the number of truth assignments of \( X \) that satisfy \( \varphi \). For DATALOG, we verify that CPP is \#EXPTIME\(-\text{hard}\) by parsimonious reduction from all counting problems in \#EXPTIME. We also show that CPP is in \#PSPACE\( (\)resp. \#EXPTIME\) for DATALOG\(_m\) and FO (resp. DATALOG) by the definition of \#C classes.

(3) When \( Q \) and \( Q_c \) are fixed, we show that CPP(\( CQ \)) is \#P\(-\text{complete}\) by parsimonious reduction from \#SAT, which is \#P\(-\text{complete}\) (cf. [25], by \#P \( = \#P \)). Given an instance \( \psi \) of 3SAT, \#SAT is to count truth assignments that satisfy \( \psi \). We also show that CPP is in \#P for all the languages.

(4) When \( Q_c \) is absent, we show that CPP(\( CQ \)) is \#NP\(-\text{hard}\) by parsimonious reduction from \#\( \Sigma_1 \text{SAT} \), which is \#NP\(-\text{complete}\) [11]. Given \( \varphi(X,Y) = \exists X \psi(X,Y) \), \#\( \Sigma_1 \text{SAT} \) is to count truth assignments of \( Y \) that satisfy \( \varphi \), where \( \psi \) is a 3DNF. We also show that CPP(\( 3\text{FO}^+ \)) is in \#NP. When \( \mathcal{L}_Q \) is DATALOG\(_m\), FO or DATALOG, the proofs of (2) given above carry over (its lower bound proofs do not use \( Q_c \)).

---

4. **SPECIAL CASES OF POI RECOMMENDATIONS**

The results of Section 3 tell us that RPP, FRP, MBP and CPP have rather high complexity. In this section we revisit these problems for special cases of package recommendations, to explore the impact of various parameters of these problems on their complexity. We consider the settings when packages are bounded by a constant instead of a polynomial, when \( \mathcal{L}_Q \) is a language for which the membership problem...
is in \( \text{PTIME} \), and when compatibility constraints are simply \( \text{PTIME} \) functions. We also study item recommendations, for which each package has a single item, and compatibility constraints are absent. Our main conclusion of this section is that the complexity of these problems is rather robust: these restrictions simplify the analyses, but not much.

**Packages with a fixed bound.** One might be tempted to think that fixing package size would simplify the analyses. Below we study the impact of fixing package sizes on package selections, in the presence of compatibility constraints \( Q_c \), by considering packages \( N \) such that \( |N| \leq B_p \), where \( B_p \) is a predefined constant rather than a polynomial.

We show that fixing package sizes does not make our lives easier when combined complexity is concerned. In contrast, this does simplify the analyses of data complexity.

**Corollary 4.1:** For packages with a constant bound \( B_p \), the combined complexity bounds of \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) are the same as given in Theorems 3.1, 3.3, 3.4 and 3.5, respectively; and the data complexity is

- in \( \text{PTIME} \) for \( \text{RPP} \),
- in \( \text{FP} \) for \( \text{FRP} \),
- in \( \text{PTIME} \) for \( \text{MBP} \), and
- in \( \text{FP} \) for \( \text{CPP} \),

for all the languages of Section 2. The complexity remains unchanged even when \( B_p \) is fixed to be 1.

**Proof sketch:** (1) The lower bounds of \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) in the presence of \( Q_c \) hold here, since their proofs (Theorems 3.1, 3.3, 3.4 and 3.5) use only top-1 package with one item, and all the upper bounds carry over here. (2) For fixed \( Q \) and \( Q_c \), we give algorithms in \( \text{PTIME}, \text{FP}, \text{PTIME} \) and \( \text{FP} \) for \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \), respectively.

**SP queries.** In contrast, for queries that have a \( \text{PTIME} \) complexity for their membership problem, variable package sizes lead to higher complexity of \( \text{RPP}, \text{FRP}, \text{MBP} \) than their counterparts for packages with a fixed bound.

To illustrate this, we consider SP queries, a simple fragment of CQ queries that support projection and selection operators only. An SP query is of the form

\[
Q(\vec{x}) = \exists \vec{x}, \vec{y} \, (R(\vec{x}, \vec{y}) \land \psi(\vec{x}, \vec{y})),
\]

where \( \psi \) is a conjunction of predicates \( =, \neq, <, \leq, > \) and \( \geq \).

The result below holds for all query languages with a \( \text{PTIME} \) membership problem, including but not limited to SP. In fact the lower bounds remain intact even when the selection criteria are specified by an *identity query*, when \( |\vec{y}| = 0 \) and \( \psi \) is a tautology in an SP query.

**Corollary 4.2:** For SP queries, the combined complexity and data complexity are

- coNP-complete for \( \text{RPP} \), but in \( \text{PTIME} \) for packages with a fixed (constant) bound \( B_p \);
- \( \text{FP}^{\text{NP}} \)-complete for \( \text{FRP} \), but in \( \text{FP} \) for fixed \( B_p \);
- \( \text{DP} \)-complete for \( \text{MBP} \), but in \( \text{PTIME} \) for fixed \( B_p \) and
- \( \#\text{-P} \)-complete for \( \text{CPP} \), but in \( \text{FP} \) for fixed \( B_p \),

when compatibility constraints are present or absent.

**Proof sketch:** (1) For packages of variable sizes, the lower bounds of \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) with fixed \( Q \) in \( CQ \) hold for SP. Indeed, their proofs of Theorems 3.1, 3.3, 3.4 and 3.5 use an identity query as \( Q \), which is in SP. For the upper bounds, the algorithms given there for \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) with a fixed \( Q \) apply to arbitrary SP queries.

(2) For packages with a constant bound, the algorithms for fixed \( Q \) of Corollary 4.1 apply to SP queries, fixed or not.

**PTIME compatibility constraints.** One might also think that we would get lower complexity with \( \text{PTIME} \) compatibility constraints. That is, we simply treat compatibility constraints as \( \text{PTIME} \) functions rather than queries in \( L_Q \). In this setting, the complexity remains the same as its counterpart when \( Q_c \) is absent, no better and no worse.

**Corollary 4.3:** With \( \text{PTIME} \) compatibility constraints \( Q_c \), the combined complexity and data complexity of \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) remain the same as their counterparts in the absence of \( Q_c \), as given in Theorems 3.2, 3.3, 3.4 and 3.5, respectively, for all the languages of Section 2.

**Proof sketch:** The lower bounds of \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) in the absence of \( Q_c \) obviously carry over to this setting, since when \( Q_c \) is empty (see Section 2), \( Q_c \) is in \( \text{PTIME} \). The upper bound proofs for Theorems 3.2, 3.3, 3.4 and 3.5 in the absence of \( Q_c \) also remain intact here. Indeed, adding an extra \( \text{PTIME} \) step for checking \( Q_c(N,D)=\emptyset \) does not increase the complexity of the algorithms given there.

**Item recommendations.** As remarked in Section 2, item recommendations are a special case of package recommendations when (a) compatibility constraints \( Q_c \) are absent, and (b) each package consists of a single item, i.e., with a fixed size 1. Given a database \( D \), a query \( Q \in L_Q \), a utility function \( f() \) and a natural number \( k \geq 1 \), a top-\( k \) item selection is a top-\( k \) package selection specified in terms of \( (Q,D,f) \).

When \( Q_c \) is absent and packages have a fixed size 1, one might expect that the recommendation analyses would become much simpler. Unfortunately, this is not the case.

**Theorem 4.4:** For items, \( \text{RPP}, \text{FRP}, \text{MBP} \) and \( \text{CPP} \) have

- the same combined complexity as their counterparts in the absence of \( Q_c \) (Theorems 3.2, 3.3, 3.4, 3.5), and
- the same data complexity as their counterparts for packages with a constant bound (Corollary 4.1),

for all the query languages given in Section 2.

**Proof sketch:** (1) Combined complexity. The upper bounds of these problems in the absence of \( Q_c \) (Theorems 3.2, 3.3, 3.4, 3.5) obviously remain intact here. The lower bound proofs for \( \text{RPP} \) and \( \text{CPP} \) given there are still valid for item recommendations, since they require only top-1 packages with a single item. For \( \text{FRP} \) and \( \text{MBP} \), however, new lower bound proofs are required for item recommendations.

More specifically, we show that \( \text{FRP}(CQ) \) is \( \text{FP}^{\text{NP}} \)-hard by reduction from \text{MAX-WEIGHT SAT}, and that \( \text{MBP}(CQ) \) is \( \text{DP} \)-hard by reduction from \text{SAT-UNSAT}, for item recommendations. For other languages \( L_Q \), the proofs for \( \text{FRP}(L_Q) \) and \( \text{MBP}(L_Q) \) are given along the same lines as their counterparts for Theorems 3.3 and 3.4, respectively.

(2) Data complexity. The algorithms developed for Corollary 4.1 suffice for item selections when \( Q \) is fixed.

**Summary.** From these results we find the following.

**Variable sizes of packages.** (1) For simple queries that have a \( \text{PTIME} \) membership problem, such as SP, the problems with variable package sizes have higher combined and data complexity than their counterparts with a fixed (constant) package size. This is in line with the claim of [34]. (2) In contrast, for any query language that subsumes CQ, variable
sizes of packages have no impact on the combined complexity of these problems. This is consistent with the observation of [26]. (3) When it comes to the data complexity, however, variable (polynomially) package sizes make our lives harder: RPP, FRP, MBP and CPP in this setting have a higher data complexity than their counterparts with a fixed package size.

**Compatibility constraints.** (1) For CQ, UCQ and $\exists FO^*$, the presence of $Q_c$ increases the combined complexity of the analyses. (2) In contrast, for more powerful languages such as DATALOG$_{eq}$, FO and DATALOG, neither $Q_c$ nor variable sizes make any difference. Indeed, RPP, FRP, MBP and CPP have exactly the same combined complexity as their counterparts for item recommendations, in the presence or absence of $Q_c$. (3) For data complexity, the presence of $Q_c$ has no impact. Indeed, when $Q_c$ is fixed, it is in PTIME to check $Q_c(N, D) = \emptyset$ for all $L_Q$ in which $Q_c$ is expressed; hence $Q_c$ can be encoded in the cost() function, and no longer needs to be treated separately. (4) To simplify the discussion we use $L_Q$ to specify $Q_c$. Nonetheless, all the complexity results remain intact for any class $C$ of $Q_c$ whose satisfiability problem has the same complexity as the membership problem for $L_Q$. In particular, when $C$ is a class of PTIME functions, the presence of $Q_c$ has no impact on the complexity.

The number $k$ of packages. All the lower bounds of RPP, FRP and MBP remain intact when $k = 1$ ($k$ is irrelevant to CPP), i.e., they carry over to top-1 package selections.

5. RECOMMENDATIONS OF QUERY RELAXATIONS

We next study query relaxation recommendations. In practice a selection query $Q$ often finds no sensible packages. When this happens, the users naturally want the recommendation system to suggest how to revise their selection criteria by relaxing the query $Q$. We are not aware of any recommendation systems that support this functionality.

Below we first present query relaxations (Section 5.1). We then identify two query relaxation recommendation problems, and establish their complexity bounds (Section 5.2).

5.1 Query Relaxations

Consider a query $Q$, in which a set $X$ of variables (free or bound) and a set $E$ of constants are parameters that can be modified, e.g., variables or constants indicating departure date and time of flights. Following [8], we relax $Q$ by replacing constants in $E$ with variables, and replacing repeated variables in $X$ with distinct variables, as follows.

(1) For each constant $c \in E$, we associate a variable $w_c$ with $c$. We denote the tuple consisting of all such variables as $\vec{w}$.

(2) For each variable $x \in X$ that appears at least twice in atoms of $Q$, we introduce a new variable $u_x$ and substitute $u_x$ for one of the occurrences of $x$. For instance, an equijoin $Q_1(\vec{u}, \vec{y}) \land Q_2(\vec{y}, \vec{z})$ is converted to $Q_1(\vec{u}, \vec{z}) \land Q_2(u_x, \vec{z}^\prime)$, a Cartesian product. This is repeated until no variable has multiple occurrences. Let $\vec{u}$ be the tuple of all such variables.

We denote the domain of $w_c$ (resp. $u_x$) as dom$(R.A)$ if $c$ (resp. $x$) appears in $Q$ as an $A$-attribute value in relation $R$.

To prevent relaxations that are too general, we constrain variables in $\vec{u}$ and $\vec{w}$ with certain ranges, by means of techniques developed for query relaxations [8,18] and preference queries [29]. To simplify the discussion, we assume that for each attribute $A$ in a relation $R$, a distance function $\text{dist}_{R,.A}(a,b)$ is defined. Intuitively, if $\text{dist}_{R,.A}(a,b)$ is within a bound, then $b$ is close enough to $a$, and we can relax $Q$ by replacing $a$ with its “neighbor” $b$. For instance, DB can be generalized to CS if $\text{dist}(DB, CS)$ is small enough [8]. We denote by $\Gamma$ the set of all such distance functions.

Given $\Gamma$, we define a relaxed query $Q_{\Gamma}$ of $Q(\vec{x})$ as:

$$Q_{\Gamma}(\vec{x}) = \exists \vec{w} \exists \vec{u} (Q'(\vec{x}, \vec{w}, \vec{u}) \land \psi_w(\vec{w}) \land \psi_u(\vec{u})), $$

where $Q'$ is obtained from $Q$ by substituting $w_c$ for constant $c$, and $u_x$ for a repeated occurrence of $x$. Here $\psi_w(\vec{w})$ is a conjunction of predicates of either (a) $\text{dist}_{R,.A}(w_c, c) \leq d$, where the domain of $w_c$ is dom$(R.A)$, and $d$ is a constant, or (b) $u_x = c$, i.e., the constant $c$ is unchanged. Query $\psi_u(\vec{u})$ includes such a conjunct for each $w_c \in \vec{w}$; similarly for $\psi_u(\vec{u})$.

We define the level gap($\gamma$) of relaxation of a predicate $\gamma$ in $\psi_w(\vec{w})$ as follows: $\text{gap}(\gamma) = d$ if $\gamma$ is $\text{dist}_{R,.A}(w_c,c) \leq d$, and $\text{gap}(\gamma) = 0$ if $\gamma$ is $u_x = c$; similarly for a predicate in $\psi_u(\vec{u})$.

We define the level of relaxation of query $Q_{\Gamma}$, denoted by $\text{gap}(Q_{\Gamma})$, to be $\sum_{\gamma \in (\psi_w(\vec{w}) \cup \psi_u(\vec{u}))} \text{gap}(\gamma)$.

Example 5.1: Recall query $Q$ defined on flight and POI in Example 1.1. The query finds no items, as there is no direct flight from ETD to NYC. Suppose that $E$ has constants ETD, NYC, 1/1/2012 and $X = \{x_{tkt}\}$, and that the user accepts a city within 15 miles of the original departure city (resp. destination) as From (resp. To), where dist() measures the distances between cities. Then we can relax $Q$ as:

$$Q_1(f\#, Pr, nm, tp, tkt, tm) = \exists DT, AD, uto, w_{\text{E6}}, \text{wNYC}, \text{wDD} (\text{flight}(f\#, w_{\text{E6}}, x_{\text{E6}}, DT, \text{WDD}, AD, Pr) \land x_{tkt} = \text{wNYC} \land \text{POI}(nm, uto, tp, tkt, tm) \land \text{wDD} = 1/1/2012 \land \text{dist}(\text{wNYC}, \text{NYC}) \leq 15 \land \text{dist}(\text{wDD}, \text{EDT}) \leq 15 \land x_{tkt} = uto).$$

The relaxed $Q_1$ finds direct flights from ETD to EWR, since the distance between NYC to EWR is within 15 miles.

We can relax $Q_1$ by allowing $w_{\text{DD}}$ to be within 3 days of 1/12/11, where the distance function for dates is $\text{dist}_{\text{DD}}(\cdot)$. Then $Q_2$ may find more available direct flights than $Q_1$, with possibly cheaper airfare. One can further relax $Q_2$ by allowing $ut_{\text{DD}}$ and $x_{tkt}$ to match different cities nearby, i.e., we convert the equijoin to a Cartesian product.

We consider simple query relaxation rules here just to illustrate the main idea of query relaxation recommendations, and defer a full treatment of this issue to future work.

5.2 Query Relaxation Recommendations

We now study recommendation problems for query relaxations, for package selections and for item selections.

The query relaxation problem for packages. Consider a database $D$, queries $Q$ and $Q_c$ in $L_Q$, functions cost() and val(), a cost budget $C$, a rating bound $B$, and a natural number $k \geq 1$. When there exists no top-$k$ package selection for $(Q,D,Q_c,\text{cost}(),\text{val}(),C)$, we need to relax $Q$ to find more packages for the users. More specifically, let $\Gamma$ be a collection of distance functions, and $X$ and $E$ be sets of variables and constants in $Q$, respectively, which are parameters that can be modified. We want to find a relaxed query $Q_{\Gamma}$ of $Q$ such that there exists a set $N$ of $k$ valid packages for $(Q_{\Gamma}, D, Q_c, \text{cost}(), \text{val}(), C, B)$, i.e., for each $N \in N$, $N \subseteq Q_{\Gamma}(D)$, $Q_{\Gamma}(N,D) = \emptyset$, cost$(N) \leq C$, val$(N) \geq B$, and $|N|$.
is bounded by a polynomial in \(|D|\). Moreover, we want \(Q_t\) to minimally differ from the original \(Q\), stated as follows.

For a constant \(g\), a relaxed query \(Q_t\) of \(Q\) is called a relaxation of \(Q\) for \((Q,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B,k,g)\) if (a) there exists a set \(N\) of \(k\) distinct valid packages for \((Q_t,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B)\), and (b) gap\((Q_t)\) \(\leq g\).

### QRPP\((\mathcal{L}_Q)\): The query relaxation rec. problem (packages)

**INPUT:** A database \(D\), a query \(Q\in\mathcal{L}_Q\) with sets \(X\) and \(E\) identified, a query \(Q_c\in\mathcal{L}_Q\), two functions \(\text{cost}(\cdot)\) and \(\text{val}(\cdot)\), natural numbers \(C,B,g\) and \(k\geq 1\), and a collection \(\Gamma\) of distance functions.

**QUESTION:** Does there exist a relaxation \(Q_t\) of \(Q\) for \((Q,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B,k,g)\)?

No matter how important, QRPP is nontrivial: it is \(\Sigma_2^P\)-complete for CQ, PSPACE-complete for DATALOGnr and FO, and EXPTIME-complete for DATALOG. It is NP-complete when selection criteria \(Q\) and compatibility constraints \(Q_c\) are both fixed. Fixing \(Q\), alone reduces the combined complexity of QRPP\((\mathcal{L}_Q)\) when \(L_Q\) is CQ, UCQ or \(\oplus\Sigma^P\), but it does not help when it comes to DATALOGnr, FO and DATALOG, or when the data complexity is concerned.

**Theorem 5.1:** For QRPP\((\mathcal{L}_Q)\), the combined complexity is

- \(\Sigma_2^P\)-complete when \(L_Q\) is CQ, UCQ or \(\oplus\Sigma^P\);
- PSPACE-complete when \(L_Q\) is DATALOGnr or FO; and
- EXPTIME-complete when \(L_Q\) is DATALOG.

In the absence of compatibility constraints, its combined complexity remains unchanged for DATALOGnr, FO and DATALOG, and it is NP-complete for CQ, UCQ and \(\oplus\Sigma^P\).

Its data complexity is NP-complete for all the languages, in the presence or absence of compatibility constraints.

**Proof sketch:**

1. We verify that QRPP\((\mathcal{L}_Q)\) is \(\Sigma_2^P\)-hard by reduction from the \(\exists^*\forall^*\exists^*\exists^*\text{DNF}\) problem, by using relaxed queries. We show that QRPP\((\exists\exists^*\forall)\) is in \(\Sigma_2^P\) by giving a nondeterministic PTIME algorithm that calls an NP oracle.
2. For DATALOGnr and FO, we show that QRPP is PSPACE-hard by reductions from \(\exists\forall\exists\forall\text{SAT}\) and the membership problem for FO, respectively, by using relaxed queries. We give a PSPACE algorithm to check QRPP. Along the same lines we show that QRPP\((\text{DATALOG})\) is EXPTIME-complete.
3. When \(Q\) is fixed, we show that QRPP\((\mathcal{L}_Q)\) is already NP-hard in the absence of \(Q_c\), by reduction from 3SAT. We show the upper bound by giving an NP algorithm to check QRPP\((\mathcal{L}_Q)\) for fixed queries \(Q\) and \(Q_c\) in FO or DATALOG.
4. In the absence of \(Q_c\), it has been shown that QRPP\((\mathcal{L}_Q)\) is NP-hard by the proof of (3) above, and the algorithm for QRPP\((\exists\exists^*\forall)\) given in (1) is now in NP. For DATALOGnr, FO and DATALOG, the proofs given in (2) above can be applied here, since their lower bound proofs do not use \(Q_c\), and their upper bounds still hold in this special case.

The query relaxation problem for items. We also study a special case of QRPP, for item selections. Given \(D, Q, \Gamma, B, k, g\), and a utility function \(f(\cdot)\), QRPP for items is to decide whether there exist a relaxation \(Q_t\) of \(Q\) for \((Q,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B,k,g)\) when \(Q_c\) is empty, and \(\text{cost}(\cdot),\text{val}(\cdot)\) and \(C\) are derived from \(f(\cdot)\) as given in Section 2.

Compared to its package counterpart, item selections simplify the data complexity analyses of query relaxation recommendations. However, it gets no better than QRPP in the absence of \(Q_c\) when the combined complexity is concerned.

**Corollary 5.2:** For all the query languages \(\mathcal{L}_Q\) given in Section 2, QRPP\((\mathcal{L}_Q)\) for items (1) has the same combined complexity as QRPP\((\mathcal{L}_Q)\) in the absence of compatibility constraints; and (2) its data complexity is in PTIME.

**Proof sketch:** For items, (1) we show that QRPP\((\mathcal{L}_Q)\) is NP-hard by reduction from 3SAT. To check QRPP\((\exists\exists^*\forall)\) we give an NP algorithm. For DATALOGnr, FO or DATALOG, the lower bounds of Theorem 5.1 hold here since their proofs use top-1 items only. The upper bounds of the combined complexity also carry over. (2) We give a PTIME algorithm to check QRPP for fixed \(Q\) in FO or DATALOG.

**Remarks.**

1. All the lower bounds of this section remains intact when \(k=1\), i.e., for top-1 package or item selections.
2. The proofs of Theorem 5.1 and Corollary 5.2 also tell us that for packages with a constant bound, QRPP\((\mathcal{L}_Q)\) has the same combined complexity as its counterpart for packages with variable sizes, and it has the same data complexity as its counterpart for items.

### 6. ADJUSTMENT RECOMMENDATIONS

We next study adjustment recommendations. In practice the collection \(D\) of items maintained by a recommendation system may fail to provide items that most users want. When this happens, the vendors of the system would want the system to recommend how to “minimally” modify \(D\) such that users’ requests could be satisfied. Below we first present adjustments to \(D\) (Section 6.1). We then study adjustment recommendations problems (Section 6.2).

#### 6.1 Adjustments to Item Collections

Consider a database \(D\) consisting of items provided by a system, and a collection \(D'\) of additional items. We use \(\Delta(D,D')\) to denote \textit{adjustments to} \(D\), which is a set consisting of (a) tuples to be deleted from \(D\), and (b) tuples from \(D'\) to be inserted into \(D\). We use \(D\oplus\Delta(D,D')\) to denote the database obtained by modifying \(D\) with \(\Delta(D,D')\).

Consider queries \(Q,Q_c\) in \(\mathcal{L}_Q\), functions \(\text{cost}(\cdot)\) and \(\text{val}(\cdot)\), a cost budget \(C\), a rating bound \(B\), and a natural number \(k\geq 1\), such that there exists no top-\(k\) package selection for \((Q,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C)\). We want to find a set \(\Delta(D,D')\) of adjustments to \(D\) such that there exists a set \(N\) of \(k\) valid packages for \((Q,D\oplus\Delta(D,D'),Q_c,\text{cost}(\cdot),\text{val}(\cdot),C)\), i.e., \(D\oplus\Delta(D,D')\) yields \(k\) packages \(N\) that are rated above \(B\), and satisfy the selection criteria \(Q_c\), compatibility constraints \(Q\), as well as aggregate constraints \(\text{cost}(\cdot)\leq C\).

One naturally wants to find a “minimum” \(\Delta(D,D')\) to adjust \(D\). For a constant \(k'\geq 1\), we call \(\Delta(D,D')\) a package adjustment for \((Q,D,Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B,k,k')\) if (a) \(|\Delta(D,D')|\leq k'\), and (b) there exist \(k\) distinct valid packages for \((Q,D\oplus\Delta(D,D'),Q_c,\text{cost}(\cdot),\text{val}(\cdot),C,B)\).

#### 6.2 Deciding Adjustment Recommendations

These suggest that we study the following problem.

**The adjustment recommendation problem.** Given \(D, D', Q, Q_c, \text{cost}(\cdot),\text{val}(\cdot), k\) and \(k'\), the adjustment recommendation problem for packages, ARPP, is to decide whether there is a package adjustment \(\Delta(D,D')\) for
Proof sketch: (1) We show that ARPP(CQ) is $\Sigma_2^p$-hard by reduction from the $3\text{DNF}$ problem. The reduction here makes use of updates $\Delta(D, D')$, and is different from the one for QRPP(CQ). To verify the upper bound, we give an NP algorithm that uses an NP oracle to check ARPP($\exists FO^+\Sigma^0$).

(2) For DATALOG$^*\text{nr}$ and FO, we show that ARPP is $\text{PSpace}$-hard by reductions from $\text{Q3SAT}$ and the membership problem for FO, respectively, which again use $\Delta(D, D')$ and are different from the ones used earlier. We give a PSPACE algorithm to check ARPP for DATALOG$^*\text{nr}$ and FO. Similarly we show that ARPP(DATALOG) is EXPTIME-complete.

(3) When Q and Q$_c$ are fixed, we show that ARPP(CQ) is NP-hard without Q$_c$ by reduction from 3SAT. We also show an NP algorithm to check ARPP for FO and DATALOG$^*\text{nr}$.

(4) When Q$_c$ is absent, ARPP(CQ) is NP-hard even when Q is fixed by the proof of (3), and the algorithm for $\exists FO^+\Sigma^0$ given in (1) is now in NP. For the languages considered in (2), their lower bound proofs do not use Q$_c$, and their upper bound proofs cover this special case. Moreover, the algorithm developed in (3) and the lower bound of ARPP(CQ) verify that the data complexity is NP-complete.

The adjustment recommendation problem for items.

Given $D, D', Q, k, k'$ and a utility function $f()$, ARPP for items is to decide whether there is an adjustment $\Delta(D, D')$ for $(Q, D, Q_c, \text{cost}(), \text{val}(), C, B, k, k')$, where $Q_c$ is empty, and $\text{cost}(), \text{val}(), C$ are derived from $f()$ (see Section 2).

One might expect that fixing package sizes in item selections would simplify the analyses of adjustment recommendations. Recall that all the problems we have studied so far have a lower data complexity for item selections than their counterparts for packages. For instance, the data complexity of QRPP for items is in PTIME while it is NP-complete for packages; similarly for RPP, FRP, MBP and CPP. In contrast, we show below that the data complexity of ARPP for packages is robust: it remains intact for items. In other words, fixing package sizes does not help here.

Corollary 6.2: For all the languages $L_Q$ given in Section 2, ARPP in the absence of compatibility constraints and ARPP for items have the same combined and data complexity.

Proof sketch: All the lower bound proofs for ARPP($L_Q$) in the absence of $Q_c$ use top-k item selections, and ARPP(CQ) is NP-hard for fixed $Q$. Hence for all $L_Q$, ARPP without $Q_c$ has the same combined complexity as ARPP for items, and the data complexity of CRPP carries over to item selections. Note that when proving ARPP(CQ) is NP-hard for fixed $Q$, we do not use $k = 1$, the only case in the entire paper.

Remarks. One can find the following from the proofs of Theorem 6.1 and Corollary 6.2. (1) For packages with a constant bound, ARPP($L_Q$) has the same combined complexity as ARPP($L_Q$) for packages with variable sizes, and it has the same data complexity as ARPP($L_Q$) for items. (3) When $Q_c$ is in PTIME, ARPP($L_Q$) has the same combined and data complexity as ARPP($L_Q$) in the absence of $Q_c$.

7. CONCLUSIONS

We have studied a general model for recommendation systems, and investigated several fundamental problems in the model, from decision problems RPP, MBP to function problem FRP and counting problem CPP. Beyond POI recommendations, we have proposed and studied QRPP for query relaxation recommendations, and ARPP for adjustment recommendations. We have also investigated special cases of these problems, when compatibility constraints $Q_c$ are absent or in PTIME, when all packages are bounded by a constant $B_p$, and when both $Q_c$ is absent and $B_p$ is fixed to be 1 for item selections. We have provided a complete picture of the lower and upper bounds of these problems, all matching, for both their data complexity and combined complexity, when $L_Q$ ranges over a variety of query languages. These results tell us where complexity of these problems arises.

The main complexity results are summarized in Table 1, annotated with their corresponding theorems (the results for SP (Corollary 4.2) are excluded). As remarked earlier, (1) the data complexity is independent of query languages, and remains unchanged in the presence of compatibility constraints $Q_c$ or not. However, it varies when packages have variable sizes or a constant bound, as shown in Table 1. (2) The complexity bounds of these problems for CQ, UCQ and $\exists FO^+\Sigma^0$ vary when $Q_c$ is present or not, and when packages have a constant bound or not. In contrast, the bounds for FO, DATALOG$^*\text{nr}$ and DATALOG are robust regardless of the presence of $Q_c$ and package sizes. (3) When $Q_c$ is a PTIME function, these problems have the same complexity as their counterparts in the absence of $Q_c$. (4) Item selections do not come with $Q_c$, and have a fixed package size (see Table 1).

The study of recommendation problems is still preliminary. First, we have only considered simple rules for query relaxations and adjustment recommendations, to focus on the main ideas. These issues deserve a full investigation. Second, this work aims to study a general model that subsumes previous models developed for various applications, and hence adopts generic functions $\text{cost}(), \text{val}()$ and $f()$. These need to be fine tuned by incorporating information about users, collaborative filtering and specific aggregate functions. Third, to simplify the discussion we assume that selection criteria $Q$ and compatibility constraints $Q_c$ are expressed in the same language (albeit PTIME $Q_c$). It is worth studying different languages for $Q$ and $Q_c$. Fourth, the recommendation problems are mostly intractable. An interesting topic is to identify practical and tractable cases. Another issue to consider are group recommendations [5], to a group of users instead of a single user.

Acknowledgments. Fan and Geerts are supported in part by an IBM scalable data analytics award, and the RSE-NSFC Joint Project Scheme. Fan is also supported in part by the 973 Program 2012CB316200 and NSFC 61133002 of China. Deng is supported in part by 863 2012AA011203, NSFC 61103031 and CPSF 2011M500208 of China.
<table>
<thead>
<tr>
<th>Problems</th>
<th>Languages</th>
<th>With $Q_c$</th>
<th>Without $Q_c$</th>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>$H_2$-complete$^\dagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 3.1)</td>
<td>DP-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 3.2)</td>
<td>coNP-complete$^\ddagger$, PTIME$^\ddagger$, PTIME$^\ddagger$ (Cor. 4.1)</td>
</tr>
<tr>
<td>FRP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>FP$^{NP}$-complete$^\ddagger$, FPSPACE(poly)-complete$^\ddagger$, FEXPTIME(poly)-complete$^\ddagger$ (Th. 3.3)</td>
<td>FP$^{NP}$-complete$^\ddagger$, FPSPACE(poly)-complete$^\ddagger$, FEXPTIME(poly)-complete$^\ddagger$ (Th. 3.3)</td>
<td>FP$^\ddagger$, FP$^\ddagger$, FP$^\ddagger$ (Cor. 4.1)</td>
</tr>
<tr>
<td>MBP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>$#$-coNP-complete$^\ddagger$, $#$-PSPACE-complete$^\ddagger$, $#$-EXPTIME-complete$^\ddagger$ (Th. 3.5)</td>
<td>$#$-NP-complete$^\ddagger$, $#$-PSPACE-complete$^\ddagger$, $#$-EXPTIME-complete$^\ddagger$ (Th. 3.5)</td>
<td>PTIME$^\ddagger$, PTIME$^\ddagger$, PTIME$^\ddagger$ (Cor. 4.1)</td>
</tr>
<tr>
<td>CPP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>$\Sigma_2^p$-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 5.1)</td>
<td>NP-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 5.1)</td>
<td>PTIME$^\ddagger$, PTIME$^\ddagger$, PTIME$^\ddagger$ (Cor. 5.2)</td>
</tr>
<tr>
<td>QRPP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>$\Sigma_2^p$-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 6.1)</td>
<td>NP-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 6.1)</td>
<td>PTIME$^\ddagger$, PTIME$^\ddagger$, PTIME$^\ddagger$ (Cor. 6.2)</td>
</tr>
<tr>
<td>ARPP</td>
<td>CQ, UCQ, $\exists$FO$^+$, DATALOG$_{in}$, FO DATALOG</td>
<td>$\Sigma_2^p$-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 6.1)</td>
<td>NP-complete$^\ddagger$, PSPACE-complete$^\ddagger$, EXPTIME-complete$^\ddagger$ (Th. 6.1)</td>
<td>PTIME$^\ddagger$, PTIME$^\ddagger$, PTIME$^\ddagger$ (Cor. 6.3)</td>
</tr>
</tbody>
</table>

Table 1: Complexity results ($^\dagger$: items (Th. 4.4); $^\ddagger$: constant bound (Cor. 4.1); $^\ddagger$: PTIME $Q_c$ (Cor. 4.3))

8. REFERENCES