DOI: 10.19884/j.1672-5220.202504008

# Investigating Problems Related to Prime Attributes and All Keys in 2NF

LIU Changqi, LIU Guohua*, YU Xiaoxue, XU Yulu, ZHANG Limeng, ZHU Dongyan, ZHENG Xiang
School of Computer Science and Technology, Donghua University, Shanghai 201620, China

**Abstract: In relational database normalization theory, identifying all keys and prime attributes is essential yet highly challenging. It has been shown that the prime attribute problem is nondeterministic polynomial-time complete (NP-complete). Additionally, the maximum number of keys in a relation scheme is exponential in the number of attributes. These conclusions hold when the normal form is unknown, and when the normal form is known they may change. For instance, when a relation scheme is in Boyce-Codd normal form (BCNF), the prime attribute problem falls within the polynomial-time complexity class (P-class), and listing all keys becomes less cumbersome. In the realm of normalization, second normal form (2NF) serves as a foundational stage, and any scheme that is in BCNF or third normal form (3NF) inherently satisfies the requirements of 2NF. Therefore, this paper focuses on the problems related to the prime attribute and all keys for 2NF. First, we present a necessary condition and a sufficient condition for a relation scheme to be in 2NF. Then, we demonstrate that the maximum number of keys is exponential in the number of attributes and functional dependencies of a relation scheme in 2NF. Furthermore, we propose an algorithm for finding all keys of a relation scheme in 2NF. Finally, we demonstrate that the prime attribute problem remains NP-complete in 2NF. This study deepens the theoretical understanding of the computational complexity inherent in 2NF normalization and provides practical references for database scheme design.**

**Keywords**: prime attribute; key; NP-complete; normalization; second normal form (2NF)
**CLC number**: TP311.13 **Document code**: A
**Article ID**: 1672-5220(2026)02-0128-10

Open Science Identity (OSID)

## 0 Introduction

Codd[1-2] laid the foundation for the relational database normalization theory by defining specific normal forms for relation schemes. These normal forms establish clear objectives for database administrators (DBAs). Keys of a relation scheme are crucial in Codd's normal form definitions. However, enumerating all possible keys can be tedious, as the maximum number of keys in a relation scheme is exponential in the number of attributes[3]. Furthermore, determining whether an attribute is a prime attribute becomes impractical due to the complexity implied by the above statements. Lucchesi et al.[4] demonstrated that the prime attribute problem was nondeterministic polynomial-time complete (NP-complete).

A relation scheme $R$ is in second normal form (2NF) if every nonprime attribute is fully functionally dependent on each key. Similarly, it is in third normal form (3NF) if no nonprime attribute is transitively functionally dependent on any key. Thus, $R$ is in Boyce-Codd normal form (BCNF) if and only if, for all disjoint nonempty attribute sets $X$ and $Y$ in $R$, whenever $X$ functionally determines $Y$, $X$ must be a superkey of $R$.

If the normal form of a relation scheme is unknown, problems such as identifying prime attributes and listing all keys become difficult to solve. However, if we know that a relation scheme is in BCNF, we can draw favorable conclusions regarding these two problems.

When a relation scheme $R(U, F)$ is in BCNF, where $U$ is the set of attributes and $F$ is the set of functional dependencies (FDs), we can demonstrate that the problems related to prime attributes and keys are tractable as follows. If $R$ is in BCNF, then the left-hand side of every FD in $F$ is a superkey[1]. If $F$ is LR-minimum[5], the left-hand side of every FD in $F$ is a key. Let $X$ be a subset of $U$ and $A$ be any attribute in $U$. The set of all FDs that are derivable from $F$ by repeatedly applying the inference rules (including the FDs in $F$) is called the closure of $F$ and is denoted by $F^+$. For an FD $X \rightarrow A \in F^+$, where $A \notin X$, there exists a derivation of $X \rightarrow A$ from $F$. This derivation can be constructed by using Armstrong's pseudotransitivity, union and restricted augmentation rules[5].

In the case of an FD $X \rightarrow A \in F^+$ and $X \rightarrow A \notin F$, $X$ must include the left-hand side of an FD in $F$. It is not possible to generate any new key from $X$, so all keys of $R$ are the left-hand sides of FDs in $F$, resulting in the number of keys being equal to the number of FDs in $F$. This demonstrates that listing all keys is not overly cumbersome when a relation scheme is in BCNF. Since all keys of $R$ are the left-hand sides of all FDs in $F$ when $R$ is in BCNF, we can determine whether an attribute $A$ is a prime attribute by checking the left-hand side of every FD

Received date: 2025-04-10
* Correspondence should be addressed to LIU Guohua, email: ghliu@dhu.edu.cn
Citation: LIU C Q, LIU G H, YU X X, et al. Investigating problems related to prime attributes and all keys in 2NF [J]. *Journal of Donghua University (English Edition)*, 2026, 43(2): 128-137.

in $F$. The time complexity of this determination is $O(|U||F|)$, where $|U|$ is the number of attributes in $U$ and $|F|$ is the number of FDs in $F$. Therefore, the prime attribute problem belongs to the polynomial-time complexity class ($P$-class) when $R$ is in BCNF.

In the realm of normalization, 2NF serves as a foundational stage, and a scheme that is in BCNF or 3NF inherently satisfies the requirements of 2NF. Therefore, this paper aims to investigate problems related to prime attributes and keys when a relation scheme is in 2NF.

We aim to provide a new perspective on the problems of prime attributes and keys, focusing on the case where a relation scheme is in 2NF. The contributions are as follows.

1) We present a necessary condition and a sufficient condition for a relation scheme in 2NF.

2) We illustrate that a new key can be generated by replacing an existing key with the left-hand side of a specific FD. Additionally, we show that the number of all keys in relation scheme $R$ is exponential in the number of attributes and functional dependencies when the relation scheme is in 2NF.

3) We present an algorithm that aims to find all keys of a relation scheme in 2NF.

4) We demonstrate that the problem of determining prime attributes remains NP-complete when a relation scheme is in 2NF.

In this study, Section 1 introduces relational model preliminaries; Section 2 illustrates the structure of the keys of a relation scheme which is in 2NF; Section 3 analyzes the related work about keys and prime attributes; Section 4 analyzes the time efficiency of Algorithm 1 and compares it with relevant algorithms; Section 5 discusses how to apply the proposed method to normal-form determination and key-related problems.

## 1 Preliminaries

The basic concepts of the relational model, functional dependencies and Armstrong's inference system used in this paper are referred to Refs. [6-9]. We first present some basic terminology, notation, lemmas and theorems that will be used throughout the paper.

Given a relation scheme $R(U, F)$, we can apply inference rules to the FDs in $F$ to derive additional FDs, $F \models X \rightarrow Y$ denotes that $X \rightarrow Y$ is derivable from $F$. The set of all FDs that are derivable from $F$ by repeatedly applying the inference rules (including the FDs in $F$) is called the closure of $F$ and is denoted by $F^+$. The sequence of FDs for deriving an FD is called a derivation of the FD, where each FD is either a member of $F$ or follows from earlier FDs by applying one of the inference rules A1-A5 in Armstrong's inference system.

Given a set of attributes $X$, we define the closure of $X$ (relative to $F$), denoted by $X_F^+$, as the set of attributes that are functionally dependent on $X$. Formally, $X_F^+$ contains the set of all attributes $A$ such that $F \models X \rightarrow A$.

**Lemma 1**[5] Let $X \rightarrow A$ be in $F^+$, where $A \notin X$. Then there exists a derivation of $X \rightarrow A$ from $F$ in which reflexivity is not used and augmentation is either unused or used only in the last step.

Formally, if $F$ is a set of FDs, a cover of $F$ is any set of FDs with the same closure as $F$. It is important to note that a cover needs not to be a subset of $F$. A cover is nonredundant if it does not contain any proper subset that is also a cover. An FD $f$ in a set of FDs $F$ is redundant if $F - \{f\} \models f$. Thus, a cover is redundant if and only if it contains a redundant FD[5].

Let $|F|$ denote the number of FDs in $F$, specifically the cardinality of $F$. $F$ is minimum if there is no set $G$ where $|G| < |F|$ and $F^+ = G^+$. $F$ is L-minimum if it is minimum, and for every FD $X \rightarrow Y$ in $F$, there is no proper subset $X'$ of $X$ such that $X' \rightarrow Y$ is in $F^+$. $F$ is LR-minimum if it is L-minimum, and replacing FD $X \rightarrow Y$ in $F$ with $X \rightarrow Y'$, where $Y'$ is a proper subset of $Y$, alters the closure of $F$.

Given a set of FDs $F$, if $F \models X \rightarrow Y$, and $Y \rightarrow X$, then the sets of attributes $X$ and $Y$ are equivalent, denoted by $X \leftrightarrow Y$. The set of all FDs in $F$ with left-hand sides equivalent to $X$ is denoted by $E_F(X)$, and $|E_F(X)|$ represents the cardinality of $E_F(X)$. The set of left-hand sides of FDs in $E_F(X)$ is denoted by $e_F(X)$. The family consisting of all $E_F(X)$ is denoted by $E_F$, and $|E_F|$ is the cardinality of $E_F$. For $E_F$, $E_F(X_i) \cap E_F(X_j) = \varnothing$ and $E_F(X_1) \cup E_F(X_2) \cup \cdots \cup E_F(X_i) \cup \cdots \cup E_F(X_j) \cup \cdots \cup E_F(X_{|F|}) = F$, where $1 \leqslant i, j \leqslant |F|$. Therefore, $E_F$ forms a partition of $F$.

If $X$ in an $E_F(X)$ is a key, it is denoted as $E_F(X_K)$. If the attributes in an $E_F(X)$ are all prime attributes, it is denoted as $E_F(X_P)$. If the attributes in an $E_F(X)$ are all nonprime attributes, it is denoted as $E_F(X_N)$. If both nonprime attributes and prime attributes are present together in an $E_F(X)$, it is denoted as $E_F(X_M)$ [10].

**Example 1** Let $R(U, F)$ be a relation scheme, where $U = \{A, B, C, D, G, H\}$ is the set of attributes and $F = \{A \rightarrow BC, BC \rightarrow A, BCD \rightarrow H, H \rightarrow G, G \rightarrow H\}$ is the set of FDs.

The keys in Example 1 are $BCD$ and $AD$. The prime attributes are $A$, $B$, $C$ and $D$, while the nonprime attributes are $G$ and $H$. Therefore, $E_F(A_P) = \{A \rightarrow BC, BC \rightarrow A\}$, $E_F(BCD_K) = \{BCD \rightarrow H\}$, and $E_F(H_N) = \{H \rightarrow G, G \rightarrow H\}$. Let $e_F(A) = \{A, BC\}$, $e_F(BCD) = \{BCD\}$, $e_F(H) = \{H, G\}$, and $E_F = \{E_F(A), E_F(BCD), E_F(H)\}$. The set $F$ is nonredundant, minimum, L-minimum and LR-minimum. However, it is not optimal because there is another set of FDs $G = \{A \rightarrow BC, BC \rightarrow A, AD \rightarrow H, H \rightarrow G, G \rightarrow H\}$ such that $|G|_A < |F|_A$ and $F^+ = G^+$.

Given a set of FDs $G$, $n$ is the length of $G$ (in attribute symbols) and $p$ is the number of FDs in $G$. Since an LR-minimum cover can be found in $O(np^2 + n^2)$ time, in this paper, we will directly use LR-minimum cover as the set of FDs of a relation scheme.

For a relation scheme $R(U, F)$, the attributes in $U$ can be divided into four disjoint subsets: $U_1$, $U_2$, $U_3$ and $U_4$ [10]. These subsets are defined based on the occurrence of attributes on $F$. Specifically, $U_1$ consists of attributes that appear only on the left-hand sides of FDs, $U_2$ consists of attributes that appear only on the right-hand sides of FDs, $U_3$ consists of attributes that appear on both the left- and right-hand sides of FDs, and $U_4$ consists of attributes that do not appear in $F$.

The relationship between the possibility of an attribute becoming a prime attribute and the sets $U_1$, $U_2$, $U_3$ and $U_4$ is established by the following lemmas.

**Lemma 2**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs.

1) Attributes in $U_1$ or $U_4$ are prime attributes, and $U_1 \cup U_4$ is included in every key of $R$.

2) Attributes in $U_2$ are nonprime attributes.

3) $K$ is the unique key of $R$ if and only if $K \cap U_3 = \varnothing$.

**Lemma 3**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs, and $A \in U$. If $A$ is a nonprime attribute, it must appear on the right-hand side of at least one FD in $F$.

**Lemma 4**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. Let $A \in U$. If $A$ is a prime attribute and appears in $F$, it must appear on the left-hand side of at least one FD in $F$.

**Lemma 5**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. $X \rightarrow Y \in F$, and $K$ is a key of $R$. If $X_F^+ \cap K \neq \varnothing$, $X$ must contain prime attributes.

**Lemma 6**[10] Let $R(U, F)$ be a relation scheme where $U$ is the set of attributes and $F$ is the set of FDs. If $X \rightarrow Y \in F$ and $A \in Y$, where $A$ is a prime attribute, then $X$ must contain prime attributes.

**Lemma 7**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. For $E_F(X)$, if $X$ contains prime attributes, all other elements in $e_F(X)$ also contain prime attributes; if $X$ does not contain prime attributes, $E_F(X)$ must be an $E_F(X_N)$.

**Lemma 8**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. If there is an $E_F(X_K)$ in $E_F$, it must be unique.

**Lemma 9**[10] Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. If any attribute in $U_2$ appears in an $E_F(X)$, $E_F(X)$ is not an $E_F(X_P)$.

**Lemma 10**[10] $F \models X \rightarrow A$ if and only if there exists an FD $V \rightarrow W$ in $F$ such that $A \in W$ and $V \subseteq X_F^+$.

## 2 Composition of Key and Problems Related to Prime Attributes and All Keys for 2NF

For a relation scheme $R(U, F)$, where $U$ is the set of attributes and $F$ is the set of FDs, according to the definition of the key, there exists an FD $K \rightarrow U \in F^+$ and no FD $K' \rightarrow U \in F^+$, where $K' \subset K$, $K$ denotes a key of $R$ and $K'$ denotes a proper subset of $K$. By Lemma 1, it can be known that a derivation of $K \rightarrow U$ from $F$ exists, in which the reflexivity is not utilized, and augmentation is either unused or only used in the final step. Thus, the composition of $K$ could be either a left-hand side of an FD in $F$ or a set of attributes from different left-hand sides of FDs in $F$. The diversity in the sets of FDs across normal forms leads to differences in the key components. In this section, in order to investigate the problems related to prime attributes and all keys, we will show the feature of the set of FDs, and on this basis, illustrate the composition of the key in a relation scheme that is in 2NF.

**Theorem 1** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs, and $F$ is LR-minimum. If $F = \varnothing$ or $F \neq \varnothing$ and all attributes in $U$ are prime attributes, $R$ is in 2NF.

**Proof** According to Lemma 2, when $F = \varnothing$, none of the attributes in $U$ appear in $F$, so $U = U_4$. Since all attributes in $U$ are prime attributes and there is no nonprime attribute partially dependent on any key of $R$, according to the definition of 2NF, it follows that $R$ is in 2NF.

When $F \neq \varnothing$ and all attributes in $U$ are prime attributes, there is no nonprime attribute that is partially dependent on each key of $R$, according to the definition of 2NF, we can conclude that $R$ is in 2NF.

Theorem 1 is an extreme case. In the general case, $F \neq \varnothing$ and $U$ contains both prime and nonprime attributes.

**Theorem 2** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs and $F$ is LR-minimum. $F \neq \varnothing$ and $U$ contains both prime and nonprime attributes. If there is an $E_F(X_K)$ in $E_F$, and the other $E_F(X)$ is either empty or $E_F(X_P)$ or $E_F(X_N)$, $R$ is in 2NF.

**Proof** When there is an $E_F(X_K)$ and the other $E_F(X)$ is empty, namely, $E_F = \{E_F(X_K)\}$, nonprime attributes only appear in $E_F(X_K)$. Since the left-hand sides of FDs in $E_F(X_K)$ are keys of $R$, there is no nonprime attribute that is partially dependent on each key of $R$. According to the definition of 2NF, we can know that $R$ is in 2NF.

When there is an $E_F(X_K)$ and other nonempty $E_F(X)$ is $E_F(X_P)$ or $E_F(X_N)$, nonprime attributes only appear in $E_F(X_K)$ or $E_F(X_N)$. From the definition of $E_F(X_K)$, we can know that the left-hand side of any FD in $E_F(X_K)$ is a key of $R$, and nonprime attributes only appear on the right-hand sides of FDs in $E_F(X_K)$. It shows that nonprime attributes which appear in $E_F(X_K)$ are all fully dependent on each key of $R$ in $E_F(X_K)$.

For any FD $X \rightarrow A \in F^+$, where $A$ is an attribute that appears in $E_F(X_N)$, consider a derivation of FD $X \rightarrow A$. The first FD $V \rightarrow W (A \in W)$ may be in $E_F(X_K)$ or some $E_F(X_N)$. If $V \rightarrow W (A \in W)$ is in $E_F(X_K)$, there is not

any FD $Z \rightarrow S$ in the derivation of FD $X \rightarrow A$, where $Z$ is a proper subset of a key. It shows that $A$ is fully dependent on each key of $R$. If $V \rightarrow W$ ($A \in W$) is in some $E_F(X_N)$, the second FD $T \rightarrow Y$ in the derivation may be in $E_F(X_K)$ or some $E_F(X_N)$, and $V \cap Y \neq \varnothing$. If FD $T \rightarrow Y$ is in $E_F(X_K)$, from $T \rightarrow Y$ and $V \rightarrow W$, we can infer a new FD $T \cup (V - V \cap Y) \rightarrow W(A \in W)$ by pseudotransitivity. It is obvious that $T \cup (V - V \cap Y)$ contains a key $T$, and is not a proper subset of a key. If FD $T \rightarrow Y$ is in some $E_F(X_N)$, from $T \rightarrow Y$ and $V \rightarrow W$, we can infer a new FD $T \cup (V - V \cap Y) \rightarrow W(A \in W)$ by pseudotransitivity. It is obvious that $T \cup (V - V \cap Y)$ only includes nonprime attributes, and is not a proper subset of a key. In the derivation of FD $X \rightarrow A$, there is no FD whose left-hand side is a proper subset of a key.

As we know that $F$ includes one $E_F(X_K)$ and some $E_F(X_P)$ or $E_F(X_N)$, it is impossible for any nonprime attribute in $E_F(X_K)$ or $E_F(X_N)$ to appear on the right-hand side of an FD whose left-hand side is a proper subset of a key. It shows that nonprime attributes in $E_F(X_K)$ or $E_F(X_N)$ are all fully dependent on each key of $R$. Therefore, there exists no nonprime attribute that is partially dependent on any key of $R$. According to the definition of 2NF, we can know that $R$ is in 2NF.

**Example 2** Let $R(U, F)$ be a relation scheme, where $U = \{A, B, C, D, G, H\}$ is the set of attributes, $F = \{A \rightarrow BC, BC \rightarrow A, BCD \rightarrow G, G \rightarrow H, H \rightarrow G\}$ is the set of FDs, and $F$ is LR-minimum.

In Example 2, first, we decide the relation scheme $R$ by the definition of 2NF. The nonprime attributes of $R$ are $G$ and $H$, and all keys are $AD$ and $BCD$. Since we cannot derive any partial dependence between the nonprime attributes ($G$ and $H$) and all keys ($AD$ and $BCD$), $R$ is in 2NF.

For comparison, we decide the relation scheme $R$ by Theorem 2. The distribution of $E_F(X)$ of $F$ consists of one $E_F(X_K)$, one $E_F(X_P)$ and one $E_F(X_N)$, namely, $E_F(BCD_K)$, $E_F(A_P)$ and $E_F(G_N)$. As it satisfies the condition of Theorem 2, $R$ is in 2NF. This conclusion aligns with the determination made by the definition of 2NF.

Theorem 2 presents only a sufficient condition and not a necessary condition. Example 3 shows that the relation scheme $R$ is in 2NF, but the distribution of $E_F(X)$ of $F$ is one $E_F(X_K)$, one $E_F(X_P)$, one $E_F(X_N)$ and one $E_F(X_M)$.

**Example 3** Let $R(U, F)$ be a relation scheme, where $U = \{A, B, C, D, G, H, E\}$ is the set of attributes, $F = \{A \rightarrow BC, BC \rightarrow A, BCD \rightarrow G, G \rightarrow H, HD \rightarrow E\}$ is the set of FDs, and $F$ is LR-minimum.

In Example 3, we decide the relation scheme $R$ by the definition of 2NF. The nonprime attributes of $R$ are $G$, $H$ and $E$, and all keys are $AD$ and $BCD$. Since we cannot derive any partial dependence between the nonprime attributes ($G$, $H$ and $E$) and all keys ($AD$ and $BCD$), $R$ is in 2NF.

The distribution of $E_F(X)$ of $F$ is one $E_F(X_K)$, one $E_F(X_P)$, one $E_F(X_N)$ and one $E_F(X_M)$, namely, $E_F(BCD_K)$, $E_F(A_P)$, $E_F(G_N)$ and $E_F(HD_M)$.

**Theorem 3** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs and $F$ is LR-minimum. $F \neq \varnothing$, and $U$ contains both prime and nonprime attributes. If $R$ is in 2NF, there must be an $E_F(X_K)$ in $E_F$.

**Proof** If $R$ is in 2NF, then for any key $K$ and any nonprime attribute $A$ of $R$, it holds that $K \rightarrow A \in F^+$. Moreover, there does not exist a proper subset $K' \subset K$ such that $K' \rightarrow A \in F^+$. In the derivation of FD $K \rightarrow A$, for all FDs which are selected from $F$, if there is an FD $X \rightarrow Y$ whose left-hand side $X$ is a key, then there is an $E_F(X_K)$ in $E_F$; otherwise, there is no $E_F(X_K)$ in $E_F$. If there is no $E_F(X_K)$ in $E_F$, $A$ may appear in $E_F(X_N)$ or $E_F(X_M)$. Thus, in the derivation of FD $K \rightarrow A$, FDs which are selected from $F$ must be in $E_F(X_N)$ or $E_F(X_M)$. In particular, no left-hand side of these FDs consists solely of prime attributes, the goal of the derivation of FD $K \rightarrow A$ is to replace the nonprime attributes on the left-hand sides of the previous FDs in the derivation and finally derive FD $K \rightarrow A$. Since $R$ is in 2NF, besides FD $K \rightarrow A$, there is no FD whose left-hand side consists solely of prime attributes in the derivation. But if there is no $E_F(X_K)$ in $E_F$, then FD $K \rightarrow A$ cannot be derived by FDs in $F$. Hence, in any derivation that does not contain FD $K \rightarrow A$, it means that FD $K \rightarrow A \notin F^+$, which contradicts the assumption that $K$ is a key. Therefore, there must be an $E_F(X_K)$ in $E_F$.

Theorem 3 presents a necessary condition under which a relation scheme is in 2NF.

**Example 4** Let $R(U, F)$ be a relation scheme, where $U = \{A, B, C, D, G, H\}$ is the set of attributes, $F = \{A \rightarrow BC, BC \rightarrow A, D \rightarrow G, G \rightarrow HD\}$ is the set of FDs, and $F$ is LR-minimum.

Initially, in Example 4, the relation scheme $R$ is determined by using the definition of 2NF. The nonprime attribute of $R$ is $H$, and all keys are $AD$, $AG$, $BCD$ and $BCG$. Since $G \rightarrow H$, the nonprime attribute $H$ is partially dependent on the keys $AG$ and $BCG$. Therefore, $R$ is not in 2NF.

For comparison, we decide the relation scheme $R$ by Theorem 3. The distribution of $E_F(X)$ of $F$ is one $E_F(X_P)$ and one $E_F(X_M)$, namely, $E_F(A_P)$ and $E_F(D_M)$. It violates the necessary condition of Theorem 3. Therefore, $R$ is not in 2NF. This conclusion aligns with determination made by the definition of 2NF.

The above theorems analyze the features of the set of FDs, laying the foundation for the following discussion on the composition of key in 2NF.

**Theorem 4** Let $R(U, F)$ be a relation scheme,

where $U$ is the set of attributes and $F$ is the set of FDs. $R$ is in 2NF. If $F = \varnothing$, attributes in $U$ are all prime. If $F \neq \varnothing$ and $F$ is LR-minimum, a prime attribute must be on the left-hand sides of FDs in $E_F(X_K)$, $E_F(X_M)$ or be in $E_F(X_P)$.

**Proof** When $F = \varnothing$, according to Lemma 2, $U_4 = U$, and thus all attributes in $U$ are prime. When $F \neq \varnothing$ and $F$ is LR-minimum, $E_F$ consists of an $E_F(X_K)$, some $E_F(X_P)$, $E_F(X_M)$ and $E_F(X_N)$. Since all left-hand sides of FDs in $E_F(X_K)$ are keys, the attributes that appear on the left-hand sides of FDs are prime. If a prime attribute appears on the right-hand sides of FDs in $E_F(X_K)$ or $E_F(X_M)$, according to Lemma 4, it must appear on the left-hand side of some FDs in $F$. Since $E_F$ consists of an $E_F(X_K)$, some $E_F(X_P)$, $E_F(X_M)$ and $E_F(X_N)$, a prime attribute that appears on the right-hand sides of FDs in $E_F(X_K)$ or $E_F(X_M)$ may also appear on the left-hand sides of FDs in $E_F(X_K)$, $E_F(X_M)$ or in $E_F(X_P)$. For an FD in $E_F(X_M)$, prime attributes must be included. If the right-hand side contains prime attributes, according to Lemma 6, the left-hand side must contain prime attributes. Therefore, a prime attribute must be on the left-hand sides of FDs in $E_F(X_K)$, $E_F(X_M)$ or be in $E_F(X_P)$.

**Theorem 5** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. $R$ is in 2NF. When $F \neq \varnothing$ and $F$ is LR-minimum, a nonprime attribute may be in $U_2$ or may appear in $E_F(X_M)$ or $E_F(X_N)$.

**Proof** When $F \neq \varnothing$ and $F$ is LR-minimum, $E_F$ consists of an $E_F(X_K)$, some $E_F(X_P)$, $E_F(X_M)$ and $E_F(X_N)$. Since all left-hand sides of FDs in $E_F(X_K)$ are keys, if an attribute that appears in $E_F(X_K)$ is a nonprime attribute, it must appear on the right-hand sides of FDs in $E_F(X_K)$. If it does not appear in $E_F(X_M)$ or $E_F(X_N)$, according to Lemma 2, it must be in $U_2$. Conversely, if a nonprime attribute does not appear on the right-hand sides of FDs in $E_F(X_K)$, it must appear in $E_F(X_M)$ or $E_F(X_N)$. Therefore, a nonprime attribute may be on the right-hand sides of FDs in $E_F(X_K)$ or may appear in $E_F(X_M)$ or $E_F(X_N)$.

**Lemma 11** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs, $F \neq \varnothing$ and $F$ is LR-minimum. If $R$ is in 2NF, there exists some $X_{iK}$ in $eF(X_K)$ and an FD $V \rightarrow W$ in some $E_F(X_P)$ or $E_F(X_M)$, such that $W \cap X_{iK} \neq \varnothing$ $(1 \leqslant i \leqslant |eF(X_K)|)$.

**Proof** By way of contradiction. Assume that there does not exist any $X_{iK}$ in $eF(X_K)$ and FD $V \rightarrow W$ in each $E_F(X_P)$ and $E_F(X_M)$, such that $W \cap X_{iK} \neq \varnothing$, where $1 \leqslant i \leqslant |eF(X_K)|$. According to Theorem 4, a prime attribute must be on the left-hand sides of FDs in the $E_F(X_K)$, $E_F(X_M)$ or be in $E_F(X_P)$. Let $K$ be a key which is composed of prime attributes that appear in $E_F(X_P)$, $E_F(X_K)$ and $E_F(X_M)$, $K \notin eF(X_K)$. For $B \in X_{iK}$, $1 \leqslant i \leqslant |eF(X_K)|$, $K \rightarrow B \in F^+$. According to Lemma 1, there exists a derivation of $K \rightarrow B$ from $F$ in which only the pseudotransitivity, union and restricted augmentation rules are used. By the definition, we can know that the derivation of $K \rightarrow B$ from $F$ must be a sequence that begins with FDs in $F$ and ends with $K \rightarrow B$. The first FD in the sequence may be selected from $E_F(X_K)$, $E_F(X_M)$ or $E_F(X_P)$, and the right-hand side of FD must contain $B$. When FDs in $E_F(X_K)$ are selected, according to Lemma 1, the left-hand side of every FD in the sequence includes the left-hand side of the FD selected from $E_F(X_K)$. Since $K \rightarrow B$ is the last FD in the sequence, $K$ includes the left-hand side of the FD selected from $E_F(X_K)$. Because $K$ is a key and the left-hand side of the FD selected from $E_F(X_K)$ is also a key, $K$ and the left-hand side of the FD selected from $E_F(X_K)$ are identical. Namely, $K \in eF(X_K)$. This contradicts the fact that $K \notin eF(X_K)$. Therefore, FDs in $E_F(X_K)$ cannot be selected as the first FD in the sequence. The first FD $V \rightarrow W (B \in W)$ must be selected from some $E_F(X_P)$ or $E_F(X_M)$. Therefore, there exists some $X_{iK}$ in $eF(X_K)$, and FD $V \rightarrow W$ in some $E_F(X_P)$ or $E_F(X_M)$, $W \cap X_{iK} \neq \varnothing$. This contradicts the assumption that there does not exist any $X_{iK}$ in $eF(X_K)$ and FD $V \rightarrow W$ in each $E_F(X_P)$ and $E_F(X_M)$, $W \cap X_{iK} \neq \varnothing$.

**Lemma 12** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs, $F \neq \varnothing$ and $F$ is LR -minimum. If $R$ is in 2 NF, there exists a superkey of $R$ which is obtained from $(X_{iK} - W \cap X_{iK}) \cup V$, where $1 \leqslant i \leqslant |eF(X_K)|$, $X_{iK}$ in $eF(X_K)$ and FD $V \rightarrow W$ in some $E_F(X_P)$ or $E_F(X_M)$ and $W \cap X_{iK} \neq \varnothing$. The key of $R$ which is obtained from $(X_{iK} - W \cap X_{iK}) \cup V$ contains both attributes in $X_{iK}$ and $V$.

**Proof** According to Lemma 1, there exists some $X_{iK}$ in $eF(X_K)$ and FD $V \rightarrow W$ in some $E_F(X_P)$ or $E_F(X_M)$, where $1 \leqslant i \leqslant |eF(X_K)|$ and $W \cap X_{iK} \neq \varnothing$. Since $X_{iK}$ is a key of $R$ and $X_{iK} \rightarrow U \in F^+$. From $V \rightarrow W$ and $X_{iK} \rightarrow U$ by pseudotransitivity, $(X_{iK} - W \cap X_{iK}) \cup V \rightarrow U$ can be derived. Then, $(X_{iK} - W \cap X_{iK}) \cup V$ is a superkey of $R$. By eliminating the redundancy in $(X_{iK} - W \cap X_{iK}) \cup V$, we can obtain a key $K$. If $K$ does not contain attributes in $X_{iK}$, then $K$ is a subset of $V$. As we know that $V \rightarrow W$ in $E_F(X_P)$ or $E_F(X_M)$ and $V$ is not a key, so $K$ is not a key. But it contradicts the assumption that $K$ is a key. Therefore, $K$ must contain attributes in $X_{iK}$. If $K$ does not contain attributes in $V$, then $K$ is a proper subset of $X_{iK}$. According to the definition of the key, any proper subset of $X_{iK}$ is not a key. Thus, $K$ is not a key, but it contradicts the assumption that $K$ is a key. Hence, $K$ must contain attributes in $V$. Therefore, the key of $R$ obtained from $(X_{iK} - W \cap X_{iK}) \cup V$ contains both attributes in $X_{iK}$ and $V$.

**Theorem 6** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs and $F$ is LR-minimum. If $R$ is in 2NF and there is an $E_F(X_K)$ in

$E_F$, the new key can be generated by replacing the attributes in an existing key with the left-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$.

**Proof** Since $eF(X_K)$ is a set of keys, it can be the initial set of existing keys. By Lemma 11, we can know that a new key can be generated by replacing the attributes in the key in $eF(X_K)$ with the left-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$, and the new key that contains both attributes in the key in $eF(X_K)$ and the left-hand sides of FDs in $E_F(X_P)$ or prime attributes in the left-hand sides of FDs in $E_F(X_M)$. Let the key be the first key of induction, and we can prove that the following new keys can be generated by replacing attributes in the existing keys with the left-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$ through mathematical induction. Assume that the $k$th key is solely generated through replacing attributes in the existing keys with the left-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$, where $k$ is the key index. Hence, the $k$th key in the set of existing keys contains attributes that appear in $E_F(X_P)$ or $E_F(X_M)$. For an $E_F(X_P)$ or $E_F(X_M)$, each left-hand side of an FD can functionally determine all attributes that appear in $E_F(X_P)$ or $E_F(X_M)$. Therefore, the $(k+1)$ th key can be generated by replacing the attributes in the $k$th key with the left-hand sides of FDs in $E_F(X_P)$.

**Theorem 7** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes, $F$ is the set of FDs, $F \neq \varnothing$ and $F$ is LR-minimum. If $R$ is in 2NF, the number of all keys of $R$ may be exponential in the number of attributes and FDs.

**Proof** Every element in $eF(X_K)$ is a key. According to Lemma 11, there are keys obtained by replacing attributes of keys in $eF(X_K)$ with attributes that only appear in $E_F(X_P)$ or $E_F(X_M)$. Therefore, the number of all keys of $R$ is $(|eF(X_K))| + C)$, where $1 \leqslant |eF(X_K)| \leqslant |F|$ and $C$ is the number of keys composed of prime attributes that only appear in $E_F(X_P)$, $E_F(X_M)$ and prime attributes in $E_F(X_K)$.

According to Lemma 10, for $X_{iK} \in eF(X_K)$, where $1 \leqslant i \leqslant |eF(X_K)|$, if a key is obtained by replacing attributes of $X_{iK}$ with attributes that only appear in $E_F(X_P)$ or $E_F(X_M)$, then there exists an FD $V \rightarrow W$ in $E_F(X_P)$ or $E_F(X_M)$, where $W \cap X_{iK} = \varnothing$. In the maximum possible case, the intersection of the right-hand side of each FD in $E_F(X_P)$ or $E_F(X_M)$ with $X_{iK}$ is not empty.

The process of replacing attributes to generate keys forms a tree structure. Here, $X_{iK}$ serves as the root, and its children are keys derived by sequentially replacing attributes of $X_{iK}$ with the right-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$. The number of keys in the first level of this tree is equal to the number of FDs in $E_F(X_P)$ and $E_F(X_M)$. For each key in the first level, there is a tree with the key as the root, and the children of the key are keys obtained by replacing attributes of the key with right-hand sides of FDs in $E_F(X_P)$ and $E_F(X_M)$. Since the intersection of every right-hand side of FD in $E_F(X_P)$, $E_F(X_M)$ and the key may not be empty, the maximum number of children for the key is also equal to the number of FDs in $E_F(X_P)$ and $E_F(X_M)$.

Let the number of FDs in $E_F(X_P)$ and $E_F(X_M)$ be $n$. Therefore, the maximum number of keys in the second level is $n^2$. If the height of the tree is $h$, then the maximum number of keys in the tree is $n^h$. Since there are $|E_F(X_K)|$ keys in $eF(X_K)$, the number of all keys of $R$ is $(|eF(X_K)| + |E_F(X_K)| \times n^h)$. When $E_F(X_K)$ only contains one FD and $E_F(X_N)$ is empty, let $m = |F|$. In this case, the number of FDs in $E_F(X_P)$ and $E_F(X_M)$ is maximum, namely $n = m - 1$. Therefore, the number of all keys of $R$ is $(m-1)^h$. When $E_F(X_K)$ only contains one FD, for the tree of keys whose root is $X_K$, the height of the tree is related to cardinality of $X_K$.

When an attribute is replaced once, a level is added. For an attribute, the maximum number of repeated replacements is the maximum number of FDs in $E_F(X_P)$ and $E_F(X_M)$, namely, $m - 1$. Therefore, the maximum number of levels that an attribute can add is $m - 1$. Let $s = |U|$, the cardinality of $X_K$ is $s - a$, where $a$ is the number of attributes in $U - X_K$, namely, $a = |U - X_K|$. Except for $\varnothing$, every subset of $U - X_K$ can be a left-hand side of an FD in $E_F(X_P)$ and $E_F(X_M)$. Therefore, the maximum number of FDs in $E_F(X_P)$ and $E_F(X_M)$ can be $2^a - 1$. Since the maximum number of FDs in $E_F(X_P)$ and $E_F(X_M)$ is $m - 1$, we have $(2^a - 1) = m - 1$; hence $a = \log_2 m$.

Therefore, the cardinality of $X_K$ is $s - \log_2 m$, and the height of the tree with $X_K$ as the root is $(s - \log_2 m) \times (m - 1)$, namely, $h = (s - \log_2 m) \times (m - 1)$. Thus, the number of all keys of $R$ is $|eF(X_K)| + C = 1 + (m - 1)^{(s-\log_2 m)\times(m-1)}$, where $m$ is the number of FDs in $F$ and $s$ is the number of attributes in $U$. Therefore, the number of all keys of $R$ may be exponential in terms of the number of attributes and FDs.

**Theorem 8** If a relation scheme $R(U, F)$ is in 2NF, then Algorithm 1 can exactly find keys.

**Proof** Step 8 in Algorithm 1 generates new keys based on Theorem 6. If the set of keys does not change, Step 4 stops, indicating that keys have been found. Thus, Algorithm 1 can exactly find keys.

**Algorithm 1** Finding keys of a relation scheme in 2NF

**INPUT**: a relation scheme $R(U, F)$, $U$ is the set of attributes, $F$ is the set of FDs, $F \neq \varnothing$ and is LR-minimum, and $R$ is in 2NF.

**OUTPUT**: the set of all keys $K$.

**Begin**

1: $K \leftarrow \varnothing$;

2: $K \leftarrow K \cup eF(X_K)$;

3: $K' \leftarrow \varnothing$;

4: **while** $K' \neq K$ **do** {

5: $S \leftarrow K - K'$; $n \leftarrow |S|$; $K' \leftarrow K$;

6: for $k = 1$ to $n$ {

7: select the $k$th key $K$ from $S$;

8: while (there is an FD $V \rightarrow W$ in $E_F(X_P)$ or $E_F(X_M)$, and $W \cap K \neq \varnothing$ do {

generate a key $K'$ from $(K - W \cap K) \cup V$ by moving redundant attributes;

**if** $K' \notin K$ then $K \leftarrow K \cup \{K'\}$

}

}

}

End

**Theorem 9** The time complexity of Algorithm 1 is $O(m^{(s-\log_2 m)\times m})$, $m$ is the cardinality of $F$, and $s$ is the cardinality of $U$.

**Proof** In Algorithm 1, Step 4 stops when all keys are found. According to Theorem 7, the number of all keys of $R$ is $1 + (m - 1)^{(s-\log_2 m)\times(m-1)}$, and thus the time complexity of Algorithm 1 is $O(m^{(s-\log_2 m)\times m})$.

Except in extreme cases, from Theorem 9, it can be seen that even if we know that a relation scheme is in 2NF, listing all the keys is still a very troublesome task.

Although we show that a prime attribute must be on the left-hand sides of FDs in $E_F(X_K)$, $E_F(X_M)$ or the attribute appears in $E_F(X_P)$ in Theorem 4, so far, we have no other more effective way to distinguish between $E_F(X_M)$ and $E_F(X_P)$ besides that of finding all the keys. Thus, the prime attribute problem for 2NF is still intractable.

Lucchesi et al.[4] proved that the prime attribute problem is NP-complete without knowing which the normal form a relation scheme is in. From the above discussion, we can see that even if we know that a relation scheme is 2NF, the prime attribute problem for 2NF is still intractable. Therefore, their conclusion still applies to the case where a relation scheme is known in 2NF.

**Corollary 1** Let $R(U, F)$ be a relation scheme, where $U$ is the set of attributes and $F$ is the set of FDs. If $R$ is in 2NF, the prime attribute problem is NP-complete.

## 3 Related Work

So far, a lot of algorithms for finding keys of a relation scheme have been presented[11-23]. Keys of a relation scheme contain prime attributes. If all keys are listed, the prime attribute problem is solved. Since the number of all keys of a relation scheme may be factorial in the number of functional dependencies, factorial in the number of attributes or exponential in the number of attributes and FDs[3, 4, 20], it is tedious to list all keys. Thus, the way of recognizing prime attributes by listing all keys is not practical. Lucchesi et al.[4] have proved the prime attribute problem to be NP-complete, and the difficulty of this problem has a clear conclusion. Kundu[6] presented a sufficient condition for determining whether an attribute is a prime attribute. To test an attribute $A$ based on the sufficient condition, we do not have to list all keys, but we must construct each subset $W$ that does not contain $A$, and the time requirement is $O(2^{|U|-l}|F|)$. To avoid testing all possible subsets, Mannila et al.[7] restricted the subsets to the maximum sets and obtained a corresponding improvement in the time bound. They showed a sufficient and necessary condition for an attribute to be a prime attribute, namely, given a relation scheme $R(U, F)$, where $U$ is the set of attributes, $F$ is the set of FDs, $W \subseteq U$ and $A \in U$. Then $\max(U, A) = \{Y \subseteq U \mid Y$ is a maximum set (with respect to $\subseteq$) such that $Y \not\rightarrow A\}$, and $A$ is a prime attribute if and only if for some $W \in \max(U, A)$, $(WA)^+ F = U$. The time requirement for testing an attribute $A$ based on the sufficient and necessary condition is $O(m_A|U|)$, where $m_A = |\max(U, A)|$. Neither does the sufficient condition in Ref. [6] nor the sufficient and necessary condition in Ref. [7] involve the appearance of attributes in $F$. In fact, the appearance plays an important role for the prime attribute problem. In some

cases, it is easy to discriminate prime attributes and nonprime attributes. Feng et al. [24] illustrated the relationship between a prime attribute and its appearance in $F$, but they only investigated three situations of the appearance of an attribute in $F$. It is shown that an attribute $A$ is a prime attribute when it only appears on the left-hand sides of FDs in $F$ or does not appear in $F$; $A$ is a nonprime attribute when it only appears on the right-hand sides of FDs in $F$. In these cases, the prime attribute problem is in P-class. Hao et al. [17] presented sufficient and necessary conditions for determining whether an attribute appearing in both the left- and right-hand sides of FDs in $F$ to be a prime attribute. These works lay the foundation for this paper.

Graphs are effective tools for finding candidate keys of a relation scheme. Feng[16] presented a graph-based algorithm for finding candidate keys of a relation scheme in which the left-hand side of every FD contained only one attribute. Saiedian et al. [18] used directed graphs to compute candidate keys of a relation in polynomial time, represented the FDs among the attributes in a relation scheme as a graph and presented an algorithm for finding all minimum keys of a relation scheme. Bahmani et al.[25] proposed an automatic method based on the graph theory for key generation. Bordoloi et al. [26] proposed an approach to find all minimal keys by the candidate graph. Fernandez et al. [14] defined port graph rewriting rules and a strategy for finding minimal candidate keys of a relation schema and proposed a method for finding all minimum keys by using strategic port graph rewriting. In a recent extension of the key identification problem, Nakos et al.[27] introduced the targeted candidate key model which focused on finding a minimum set of attributes to determine a specific target subset. This formulation highlights the complexity of attribute selection in reasoning and aligns with efforts to reduce the overhead of key enumeration.

## 4 Time Efficiency Analysis

Throughout this section, $s = |U|$ and $m = |F|$. Comparisons are asymptotic and parameter-dependent. Based on the previous discussion, when testing whether an attribute $A$ is a prime attribute, the algorithm in Ref. [18] requires the construction of every subset $W$ that does not contain $A$, and its time complexity is $O(2^{|U|-l}|F|)$. Here, $|U|$ is the cardinality of the attribute set $U$, $|F|$ is the cardinality of the functional dependency set $F$, and $l$ is usually related to the attribute set $U$, which can be considered a constant or a relatively small quantity compared to $|U|$ in the current discussion. To facilitate comparison with other algorithms, we mainly focus on the order of magnitude of its complexity. The complexity of this algorithm is mainly determined by $2^{|U|}$, because the growth rate of $2^{|U|}$ is higher than that of $|F|$. Therefore, the time complexity of Kundu's algorithm can be approximated as $O(2^{|U|})$. Mannila et al.[7] restricted the subsets to the maximum sets, and on this basis, they tested whether an attribute $A$ was a prime attribute with a time complexity of $O(m_A|U|)$ In the worst-case scenario, $m_A$ may approach $|U|$, so the time complexity of the algorithm in Ref. [7] in the worst-case scenario can be approximated as $O(|U|^2)$.

Now we compare the time complexity of Algorithm 1 with that in Refs. [6] and [7]. The algorithm in Ref. [6] has a time complexity close to $O(2^{|U|})$, while Algorithm 1 has a time complexity of $O(m^{(s-\log_2 m)\times m})$ where $m$ is the cardinality of $F$, and $s$ is the cardinality of $U$. In general, for $m \geqslant 2$ and $s \geqslant 2$, $m^{(s-\log_2 m)\times m}$ grows faster than $2^s$. Thus, Algorithm 1 is not asymptotically faster than the algorithm in Ref. [6]. We present Algorithm 1 mainly as a constructive procedure for enumerating keys in the 2NF setting, rather than as an optimized method in terms of the time complexity. The algorithm in Ref. [7] has a time complexity close to $O(|U|^2)$. Since $m^{(s-\log_2 m)\times m}$ is super-polynomial in $m$, it dominates $|U|^2$. Hence, Algorithm 1 has a higher asymptotic time complexity than the algorithm in Ref. [7]. These comparisons do not affect our main theoretical results (Theorems 1–9 and NP-completeness). Algorithm 1 serves to demonstrate constructibility and to match the exponential bound on the number of keys in 2NF.

In summary, the time complexity of Algorithm 1, the algorithm in Ref. [6] and the algorithm in Ref. [7] are related to the specific values of $m$ and $s$. In large-scale data scenarios, Algorithm 1 provides a constructive baseline for enumeration in 2NF.

## 5 Conclusions and Future Work

The set of FDs is the foundation for finding keys in a relation scheme. However, the set of FDs exhibits different features depending on which normal form the relation scheme is in. We have shown the feature of the set of FDs based on the distribution of $E_F(X)$ of $F$, and on this basis, have demonstrated that a new key can be generated by replacing the attributes in the existing key with the left-hand sides of FDs in $E_F(X_P)$ or $E_F(X_M)$. Furthermore, we have determined that the number of keys for $R$ increases exponentially with the number of attributes and the number of FDs. Our research findings indicate that although the problems related to the prime attribute and all keys have optimistic solutions when $R$ is in BCNF, the prime attribute problem is still NP-complete, and listing all keys remains cumbersome when $R$ is in 2NF. In future studies, we will investigate the problems related to prime attributes and all keys when $R$ is in other normal forms, such as 3NF.

## References

[1] CODD E F. Further normalization of the data

base relational model[J]. *Data Base Systems*, 1972, 6(1972): 33-64.
[2] CODD E F. Recent investigations in relational data base systems[M]. New York: IBM Thomas J. Watson Research Division, 1974.
[3] DEMETROVICS J. On the number of candidate keys[J]. *Information Processing Letters*, 1978, 7(6): 266-269.
[4] LUCCHESI C, OSBORN S L. Candidate keys for relations[J]. *Journal of Computer and System Sciences*, 1978, 17(2): 270-279.
[5] BEERI C, BERNSTEIN P A. Computational problems related to the design of normal form relational schemas[J]. *ACM Transactions on Database Systems*, 1979, 4(1): 30-59.
[6] KUNDU S. An improved algorithm for finding a key of a relation[C]//Proceedings of the Fourth ACM SIGACT-SIGMOD Symposium on Principles of Database Systems. New York: ACM, 1985: 189-192.
[7] MANNILA H, RAIHA K J. Practical algorithms for finding prime attributes and testing normal forms[C]//Proceedings of the Eighth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York: ACM, 1989: 128-133.
[8] MAIER D. Minimum covers in the relational database model (extended abstract)[C]// Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing. New York: ACM, 1979: 330-337.
[9] OSBORN S L. Testing for existence of a covering Boyce-Codd normal form[J]. *Information Processing Letters*, 1979, 8(1): 11-14.
[10] XU Y L, LIU G H, YU X X, et al. Features of prime attributes in a relation scheme[J]. *Journal of Donghua University (English Edition)*, 2025, 42(6): 689-698.
[11] CORDERO P, ENCISO M, MORA A. Automated reasoning to infer all minimal keys[C]//IJCAI. Palo Alto: IJCAI, 2013: 817-823.
[12] DEMBA M. KeyFinder: an efficient minimal keys finding algorithm for relational databases[J]. *Inteligencia Artificial*, 2021, 24(68): 37-52.
[13] DEMETROVICS J, THI V D. Relations and minimal keys[J]. *Acta Cybernetica*, 1988, 8(3): 279-285.
[14] FERNANDEZ M, VARGA J. Finding candidate keys and 3NF via strategic port graph rewriting[C]//Proceedings of the 22nd International Symposium on Principles and Practice of Declarative Programming. New York: ACM, 2020: 1-14.
[15] FADOUS R, FORSYTH J. Finding candidate keys for relational data bases[C]//Proceedings of the 1975 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1975: 203-210.
[16] FENG Y C. Algorithm for solving candidate key by graph theory[J]. *Chinese Journal of Computers*, 1988, 11(9): 556-558. (in Chinese)
[17] HAO Z X, LIU G H. An algorithm to find out all candidate keys of relation schema[J]. *Chinese Journal of Computers*, 1991, 14(4): 300-307. (in Chinese)
[18] SAIEDIAN H, SPENCER T. An efficient algorithm to compute the candidate keys of a relational database schema[J]. *The Computer Journal*, 1996, 39(2): 124-132.
[19] WASTL R. Linear derivations for keys of a database relation schema[J]. *Journal of Universal Computer Science*, 1998, 4(11): 883-897.
[20] YU C T, JOHNSON D T. On the complexity of finding the set of candidate keys for a given set of functional dependencies[J]. *Information Processing Letters*, 1976, 5(4): 100-101.
[21] LIU G H, HAO Z X, CHEN Z J. A quick replacement algorithm for finding all candidate keys of a relation scheme[J]. *Chinese Journal of Computers*, 1998, 21(10): 890-895. (in Chinese)
[22] HAO Z X, GUO J F. A hypergraph based method for finding out all candidate keys of relation schema[J]. *Chinese Journal of Computers*, 1992, 15(4): 264-270. (in Chinese)
[23] HAO Z X, LIU G H, LIU C L. A method for finding all candidate keys of relational schema based on attributes relative table[J]. *Journal of Computer Research and Development*, 1994, 31(6): 6-13. (in Chinese)
[24] FENG Y C, PANG C Y. Algorithm for finding candidate key word[J]. *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, 1986, 14(5): 655-658. (in Chinese)
[25] BAHMANI A H, NAGHIBZADEH M, BAHMANI B. Automatic database normalization and primary key generation[C]//2008 Canadian Conference on Electrical and Computer Engineering. New York: IEEE, 2008: 000011-000016.
[26] BORDOLOI S, KALITA B. Designing graph database models from existing relational databases[J]. *International Journal of Computer Applications*, 2013, 74(1): 25-31.
[27] NAKOS V, NGO H Q, TSOURAKAKIS C E. Targeted least cardinality candidate key for relational databases[EB/OL]. (2024-08-24)[2025-01-20]. https://arxiv.org/pdf/2408.13540.

# 2NF 中主属性和所有键的相关问题研究

刘长奇，刘国华*，于晓雪，徐雨露，张丽萌，朱东燕，郑　翔
东华大学 计算机科学与技术学院，上海 201620

**摘　要：**在关系数据库规范化理论中，识别所有键和主属性至关重要，但也极具挑战性。研究表明，主属性问题属于非确定性多项式时间完全(nondeterministic polynomial-time complete，NP-complete)问题，简记NP 完全问题。此外，关系模式中键的最大数量随属性数量呈指数增长。这些结论是在范式未知的前提下得出的，若范式已知，结论可能会有所不同。例如，当关系模式满足 Boyce-Codd 范式(Boyce-Codd normal form，BCNF)时，主属性问题属于多项式时间复杂度类(polynomial-time complexity class，P-class)的问题，而列出所有键也不那么繁琐。在规范化领域，第二范式(second normal form，2NF)是一个基础阶段，且任何满足 BCNF 或第三范式(third normal form，3NF)的关系模式都必然符合 2NF 的要求。因此，本文聚焦于2NF 下的主属性和所有键的相关问题。首先，我们提出了判断关系模式是否满足 2NF 的必要与充分条件。其次，我们证明了 2NF 关系模式中键的最大数量随属性数量及函数依赖数量呈指数增长。再次，我们提出了一种用于找出 2NF 关系模式所有键的算法。最后，我们证明了主属性问题在 2NF 下仍属于 NP 完全问题。该研究深化了对 2NF 规范化中计算复杂性的理论理解，并为数据库模式设计提供了实用参考。

**关键词：**主属性；键；NP 完全；规范化；第二范式（2NF）